INTERMEDIATE AND ADVANCED ECONOMETRICS

Problems and Solutions

Stanislav Anatolyev

February 7, 2002

Contents

Preface

This manuscript is a collection of problems that I have been using in teaching intermediate and advanced level econometrics courses at the New Economic School (NES), Moscow, during last several years. All problems are accompanied by sample solutions that may be viewed "canonical" within the philosophy of NES econometrics courses.

Approximately, Chapters 1 through 5 of the collection belong to a course in intermediate level econometrics ("Econometrics III" in the NES internal course structure); Chapters 6 through 9 – to a course in advanced level econometrics ("Econometrics IV", respectively). The problems in Chapters 10 through 12 require knowledge of advanced and special material. They have been used in the courses "Topics in Econometrics" and "Topics in Cross-Sectional Econometrics".

Most of the problems are not new. Many are inspired by my former teachers of econometrics in different years: Hyungtaik Ahn, Mahmoud El-Gamal, Bruce Hansen, Yuichi Kitamura, Charles Manski, Gautam Tripathi, and my dissertation supervisor Kenneth West. Many problems are borrowed from their problem sets, as well as problem sets of other leading econometrics scholars.

Release of this collection would be hard, if not to say impossible, without valuable help of my teaching assistants during various years: Andrey Vasnev, Viktor Subbotin, Semyon Polbennikov, Alexandr Vaschilko and Stanislav Kolenikov, to whom go my deepest thanks. I wish all of them success in further studying the exciting science of econometrics.

Preparation of this manuscript was supported in part by a fellowship from the Economics Education and Research Consortium, with funds provided by the Government of Sweden through the Eurasia Foundation. The opinions expressed in the manuscript are those of the author and do not necessarily reflect the views of the Government of Sweden, the Eurasia Foundation, or any other member of the Economics Education and Research Consortium.

I would be grateful to everyone who finds errors, mistakes and typos in this collection and reports them to sanatoly@mail.nes.ru.

Part I

Problems

1. ASYMPTOTIC THEORY

1.1 Asymptotics of t-ratios

Let X_i, $i = 1, \cdots, n$, be an IID sample of scalar random variables with $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \sigma^2$, $\mathbb{E}\left[(X_i - \mu)^3\right] = 0$, $\mathbb{E}\left[(X_i - \mu)^4\right] = \tau$, all parameters being finite.

(a) Define $T_n \equiv \dfrac{\overline{X}}{\hat{\sigma}}$, where, as usual,

$$\overline{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Derive the limiting distribution of $\sqrt{n}T_n$ under the assumption $\mu = 0$.

(b) Now suppose it is not assumed that $\mu = 0$. Derive the limiting distribution of

$$\sqrt{n}\left(T_n - \operatorname*{plim}_{n\to\infty} T_n\right).$$

Be sure your answer reduces to the result of part (a), when $\mu = 0$.

(c) Define $R_n \equiv \dfrac{\overline{X}}{\overline{\sigma}}$, where

$$\overline{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

is the constrained estimator of σ^2 under the (possibly incorrect) assumption $\mu = 0$. Derive the limiting distribution of

$$\sqrt{n}\left(R_n - \operatorname*{plim}_{n\to\infty} R_n\right)$$

for arbitrary μ and $\sigma^2 > 0$. Under what conditions on μ and σ^2 will this asymptotic distribution be the same as in part (b)?

1.2 Asymptotics with shrinking regressor

Suppose that

$$y_i = \alpha + \beta x_i + u_i,$$

where $\{u_i\}$ are IID with $\mathbb{E}[u_i] = 0$, $\mathbb{E}\left[u_i^2\right] = \sigma^2$ and $\mathbb{E}\left[u_i^3\right] = \nu$, while the regressor x_i is deterministic: $x_i = \rho^i$, $\rho \in (0, 1)$. Let the sample size be n. Discuss as fully as you can the asymptotic behavior of the usual least-squares estimates $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ of $(\alpha, \beta, \sigma^2)$ as $n \to \infty$.

1.3 Creeping bug on simplex

Consider a positive (x, y) orthant, i.e. $\mathbb{R}^2_+$, and the unit simplex on it, i.e. the line segment $x+y = 1$, $x \geq 0$, $y \geq 0$. Take an arbitrary natural number $k \in \mathbb{N}$. Imagine a bug starting creeping from the origin $(x, y) = (0, 0)$. Each second the bug goes either in the positive x direction with probability p, or in the positive y direction with probability $1 - p$, each time covering distance $\frac{1}{k}$. Evidently, this way the bug reaches the unit simplex in k seconds. Let it arrive there at point (x_k, y_k). Now let $k \to \infty$, i.e. as if the bug shrinks in size and physical abilities per second. Determine:

(a) the probability limit of (x_k, y_k);

(b) the rate of convergence;

(c) the asymptotic distribution of (x_k, y_k).

1.4 Asymptotics of rotated logarithms

Let the positive random vector $(U_n, V_n)'$ be such that

$$\sqrt{n}\left(\left(\begin{array}{c} U_n \\ V_n \end{array}\right) - \left(\begin{array}{c} \mu_u \\ \mu_v \end{array}\right)\right) \xrightarrow{d} \mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} \omega_{uu} & \omega_{uv} \\ \omega_{uv} & \omega_{vv} \end{array}\right)\right)$$

as $n \to \infty$. Find the joint asymptotic distribution of

$$\left(\begin{array}{c} \ln U_n - \ln V_n \\ \ln U_n + \ln V_n \end{array}\right).$$

What is the condition under which $\ln U_n - \ln V_n$ and $\ln U_n + \ln V_n$ are asymptotically independent?

1.5 Trended vs. differenced regression

Consider a linear model with a linearly trending regressor:

$$y_t = \alpha + \beta t + \varepsilon_t,$$

where the sequence ε_t is independently and identically distributed according to some distribution $\mathcal{D}$ with mean zero and variance σ^2. The object of interest is β.

1. Write out the OLS estimator $\hat{\beta}$ of β in deviations form. Find the asymptotic distribution of $\hat{\beta}$.

2. An investigator suggests getting rid of the trending regressor by taking differences to obtain

$$y_t - y_{t-1} = \beta + \varepsilon_t - \varepsilon_{t-1}$$

and estimating β by OLS. Write out the OLS estimator $\check{\beta}$ of β and find its asymptotic distribution.

3. Compare the estimators $\hat{\beta}$ and $\check{\beta}$ in terms of asymptotic efficiency.

1.6 Second-order Delta-Method

Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i, $i = 1, \cdots, n$, is an IID sample of scalar random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = 1$. It is easy to show that $\sqrt{n}(S_n^2 - \mu^2) \xrightarrow{d} \mathcal{N}(0, 4\mu^2)$ when $\mu \neq 0$.

(a) Find the asymptotic distribution of S_n^2 when $\mu = 0$, by taking a square of the asymptotic distribution of S_n.

(b) Find the asymptotic distribution of $\cos(S_n)$. Hint: take a higher order Taylor expansion applied to $\cos(S_n)$.

(c) Using the technique of part (b), formulate and prove an analog of the Delta-Method for the case when the function is scalar-valued, has zero first derivative and nonzero second derivative, when the derivatives are evaluated at the probability limit. For simplicity, let all the random variables be scalars.

1.7 Brief and exhaustive

Give brief but exhaustive answers to the following short questions.

1. Suppose that x_t is generated by $x_t = \rho x_{t-1} + e_t$, where $e_t = \varepsilon_t + \theta\varepsilon_{t-1}$ and ε_t is white noise. Is the OLS estimator of ρ consistent?
2. The process for the scalar random variable x_t is covariance stationary with the following autocovariances: 3 at lag 0; 2 at lag 1; 1 at lag 2; and zero for all higher lags. Let T denote the sample size. What is the long-run variance of x_t, i.e. $\lim_{T\to\infty} \mathbb{V}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\right)$?
3. Often one needs to estimate the long-run variance V_{ze} of the stationary sequence $z_t e_t$ that satisfies the restriction $\mathbb{E}[e_t|z_t] = 0$. Derive a compact expression for V_{ze} in the case when e_t and z_t follow independent scalar $AR(1)$ processes. For this example, propose a way to consistently estimate V_{ze} and show your estimator's consistency.

1.8 Asymptotics of averages of AR(1) and MA(1)

Let x_t be a martingale difference sequence relative to its own past, and let all conditions for the CLT be satisfied: $\sqrt{T}\overline{x}_T = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Let now $y_t = \rho y_{t-1} + x_t$ and $z_t = x_t + \theta x_{t-1}$, where $|\rho| < 1$ and $|\theta| < 1$. Consider time averages $\overline{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $\overline{z}_T = \frac{1}{T}\sum_{t=1}^{T} z_t$.

1. Are y_t and z_t martingale difference sequences relative to their own past?
2. Find the asymptotic distributions of $\overline{y}_T$ and $\overline{z}_T$.
3. How would you estimate the asymptotic variances of $\overline{y}_T$ and $\overline{z}_T$?

4. Repeat what you did in parts 1–3 when $\mathbf{x}_t$ is a $k \times 1$ vector, and we have $\sqrt{T}\overline{\mathbf{x}}_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{x}_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$, $\mathbf{y}_t = \mathrm{P}\mathbf{y}_{t-1} + \mathbf{x}_t$, $\mathbf{z}_t = \mathbf{x}_t + \Theta\mathbf{x}_{t-1}$, and P and Θ are $k \times k$ matrices with eigenvalues inside the unit circle.

2. BOOTSTRAP

2.1 Brief and exhaustive

Give brief but exhaustive answers to the following short questions.

1. Comment on: "The only difference between Monte-Carlo and the bootstrap is possibility and impossibility, respectively, of sampling from the true population."
2. Comment on: "When one does bootstrap, there is no reason to raise B too high: there is a level when increasing B does not give any increase in precision".
3. Comment on: "The bootstrap estimator of the parameter of interest is preferable to the asymptotic one, since its rate of convergence to the true parameter is often larger".
4. Suppose that one got in an application $\hat{\theta} = 1.2$ and $s(\hat{\theta}) = .2$. By the nonparametric bootstrap procedure, the 2.5% and 97.5% bootstrap critical values for the bootstrap distribution of $\hat{\theta}$ turned out to be .75 and 1.3. Find: (a) 95% Efron percentile interval for θ, (b) 95% Hall percentile interval for θ, (c) 95% percentile-t interval for θ.

2.2 Bootstrapping t-ratio

Consider the following bootstrap procedure. Using the nonparametric bootstrap, generate pseudosamples and calculate $\dfrac{\hat{\theta}_b^* - \hat{\theta}}{s(\hat{\theta})}$ at each bootstrap repetition. Find the quantiles $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$ from this bootstrap distribution, and construct

$$CI = [\hat{\theta} - s(\hat{\theta})q^*_{1-\alpha/2}, \hat{\theta} - s(\hat{\theta})q^*_{\alpha/2}].$$

Show that CI is exactly the same as Hall's percentile interval, and *not* the t-percentile interval.

2.3 Bootstrap correcting mean and its square

Consider a random variable x with mean μ. A random sample $\{x_i\}_{i=1}^n$ is available. One estimates μ by $\bar{x}_n$ and μ^2 by $\bar{x}_n^2$. Find out what the bootstrap bias corrected estimators of μ and μ^2 are.

2.4 Bootstrapping conditional mean

Take the linear regression

$$y_i = x_i'\beta + e_i,$$

with $\mathbb{E}[e_i|x_i] = 0$. For a particular value of x, the object of interest is the conditional mean $g(x) = \mathbb{E}[y_i|x]$. Describe how you would use the percentile-t bootstrap to construct a confidence interval for $g(x)$.

2.5 Bootstrap adjustment for endogeneity?

Let the model be

$$y_i = x_i'\beta + e_i,$$

but $\mathbb{E}[e_i x_i] \neq 0$, i.e. the regressors are endogenous. Then the OLS estimator $\hat{\beta}$ is biased for the parameter β. We know that the bootstrap is a good way to estimate bias, so the idea is to estimate the bias of $\hat{\beta}$ and construct a bias-adjusted estimate of β. Explain whether or not the non-parametric bootstrap can be used to implement this idea.

3. REGRESSION IN GENERAL

3.1 Property of conditional distribution

Consider a random pair (Y, X). Prove that the correlation coefficient

$$\rho(Y, f(X)),$$

where f is any measurable function, is maximized in absolute value when $f(X)$ is linear in $\mathbb{E}\left[Y|X\right]$.

3.2 Unobservables among regressors

Consider the following situation. The vector (y, x, z, w) is a random quadruple. It is known that

$$\mathbb{E}\left[y|x, z, w\right] = \alpha + \beta x + \gamma z.$$

It is also known that $\mathbb{C}\left[x, z\right] = 0$ and that $\mathbb{C}\left[w, z\right] > 0$. The parameters α, β and γ are not known. A random sample of observations on (y, x, w) is available; z is not observable.

In this setting, a researcher weighs two options for estimating β. One is a linear least squares fit of y on x. The other is a linear least squares fit of y on (x, w). Compare these options.

3.3 Consistency of OLS in presence of lagged dependent variable and serially correlated errors

[1]Let $\{y_t\}_{t=-\infty}^{+\infty}$ be a strictly stationary and ergodic stochastic process with zero mean and finite variance.

(i) Define

$$\beta = \frac{\mathbb{C}\left[y_t, y_{t-1}\right]}{\mathbb{V}\left[y_t\right]}, \quad u_t = y_t - \beta y_{t-1},$$

so that we can write

$$y_t = \beta y_{t-1} + u_t.$$

Show that the error u_t satisfies $\mathbb{E}\left[u_t\right] = 0$ and $\mathbb{C}\left[u_t, y_{t-1}\right] = 0$.

(ii) Show that the OLS estimator $\hat{\beta}$ from the regression of y_t on y_{t-1} is consistent for β.

(iii) Show that, without further assumptions, u_t is serially correlated. Construct an example with serially correlated u_t.

[1]This problem closely follows J.M. Wooldridge (1998) Consistency of OLS in the Presence of Lagged Dependent Variable and Serially Correlated Errors. *Econometric Theory* 14, Problem 98.2.1.

(iv) A 1994 paper in the *Journal of Econometrics* leads with the statement: "It is well known that in linear regression models with lagged dependent variables, ordinary least squares (OLS) estimators are inconsistent if the errors are autocorrelated". This statement, or a slight variation on it, appears in virtually all econometrics textbooks. Reconcile this statement with your findings from parts (ii) and (iii).

3.4 Incomplete regression

Consider the linear regression

$$y_i = x_i' \underset{k_1\times 1}{\beta} + e_i, \quad \mathbb{E}\left[e_i|x_i\right] = 0, \quad \mathbb{E}\left[e_i^2|x_i\right] = \sigma^2.$$

Suppose that some component of the error e_i is observable, so that

$$e_i = z_i' \underset{k_2\times 1}{\gamma} + \eta_i,$$

where z_i is a vector of observables such that $\mathbb{E}\left[\eta_i|z_i\right] = 0$ and $\mathbb{E}\left[x_i z_i'\right] \neq 0$. The researcher wants to estimate β and γ and considers two alternatives:

1. Run the regression of y_i on x_i and z_i to find the OLS estimates $\hat{\beta}$ and $\hat{\gamma}$ of β and γ.

2. Run the regression of y_i on x_i to get the OLS estimate $\hat{\beta}$ of β, compute the OLS residuals $\hat{e}_i = y_i - x_i'\hat{\beta}$ and run the regression of $\hat{e}_i$ on z_i to retrieve the OLS estimate $\hat{\gamma}$ of γ.

Which of the two methods would you recommend from the point of view of consistency of $\hat{\beta}$ and $\hat{\gamma}$? For the method(s) that yield(s) consistent estimates, find the limiting distribution of $\sqrt{n}\left(\hat{\gamma} - \gamma\right)$.

3.5 Brief and exhaustive

Give brief but exhaustive answers to the following short questions.

1. Comment on: "Treating regressors x in a linear mean regression $y = x'\beta + e$ as random variables rather than fixed numbers simplifies further analysis, since then the observations (x_i, y_i) may be treated as IID across i".

2. A labor economist argues: "It is more plausible to think of my regressors as random rather than fixed. Look at *education*, for example. A person chooses her level of education, thus it is random. *Age* may be misreported, so it is random too. Even *gender* is random, because one can get a sex change operation done." Comment on this pearl.

3. Let (x, y, z) be a random triple. For a given real constant γ a researcher wants to estimate $\mathbb{E}\left[y|\mathbb{E}\left[x|z\right] = \gamma\right]$. The researcher knows that $\mathbb{E}\left[x|z\right]$ and $\mathbb{E}\left[y|z\right]$ are strictly increasing and continuous functions of z, and is given consistent estimates of these functions. Show how the researcher can use them to obtain a consistent estimate of the quantity of interest.

4. Comment on: "When one suspects heteroskedasticity, one should use White's formula

$$Q_{xx}^{-1} Q_{xxe^2} Q_{xx}^{-1}$$

instead of conventional $\sigma^2 Q_{xx}^{-1}$, since under heteroskedasticity the latter does not make sense, because σ^2 is different for each observation".

4. OLS AND GLS ESTIMATORS

4.1 Brief and exhaustive

Give brief but exhaustive answers to the following short questions.

1. Consider a linear mean regression $y_i = x_i'\beta + e_i$, $\mathbb{E}[e_i|x_i] = 0$, where x_i, instead of being IID across i, depends on i through an unknown function φ as $x_i = \varphi(i) + u_i$, where u_i are IID independent of e_i. Show that the OLS estimator of β is still unbiased.
2. Consider a model $y = (\alpha + \beta x)e$, where y and x are scalar observables, e is unobservable. Let $\mathbb{E}[e|x] = 1$ and $\mathbb{V}[e|x] = 1$. How would you estimate (α, β) by OLS? What standard errors (conventional or White's) would you construct?

4.2 Estimation of linear combination

Suppose one has an IID random sample of n observations from the linear regression model

$$y_i = \alpha + \beta x_i + \gamma z_i + e_i,$$

where e_i has mean zero and variance σ^2 and is independent of (x_i, z_i).

1. What is the conditional variance of the best linear conditionally (on the x_i and z_i observations) unbiased estimator $\hat{\theta}$ of
$$\theta = \alpha + \beta c_x + \gamma c_z,$$
where c_x and c_z are some given constants?
2. Obtain the limiting distribution of
$$\sqrt{n}\left(\hat{\theta} - \theta\right).$$
Write your answer as a function of the means, variances and correlations of x_i, z_i and e_i and of the constants $\alpha, \beta, \gamma, c_x, c_z$, assuming that all moments are finite.
3. For what value of the correlation coefficient between x_i and z_i is the asymptotic variance minimized for given variances of e_i and x_i?
4. Discuss the relationship of the result of part 3 with the problem of multicollinearity.

4.3 Long and short regressions

Take the true model $Y = X_1\beta_1 + X_2\beta_2 + e$, $\mathbb{E}[e|X_1, X_2] = 0$. Suppose that β_1 is estimated only by regressing Y on X_1 only. Find the probability limit of this estimator. What are the conditions when it is consistent for β_1?

4.4 Ridge regression

In the standard linear mean regression model, one estimates $k \times 1$ parameter β by

$$\tilde{\beta} = \left(X'X + \lambda I_k\right)^{-1} X'Y,$$

where $\lambda > 0$ is a fixed scalar, I_k is a $k \times k$ identity matrix, X is $n \times k$ and Y is $n \times 1$ matrices of data.

1. Find $\mathbb{E}\left[\tilde{\beta}|X\right]$. Is $\tilde{\beta}$ conditionally unbiased? Is it unbiased?

2. Find $\underset{n\to\infty}{\text{plim}}\ \tilde{\beta}$. Is $\tilde{\beta}$ consistent?

3. Find the asymptotic distribution of $\tilde{\beta}$.

4. From your viewpoint, why may one want to use $\tilde{\beta}$ instead of the OLS estimator $\hat{\beta}$? Give conditions under which $\tilde{\beta}$ is preferable to $\hat{\beta}$ according to your criterion, and vice versa.

4.5 Exponential heteroskedasticity

Let y be scalar and x be $k \times 1$ vector random variables. Observations (y_i, x_i) are drawn at random from the population of (y, x). You are told that $\mathbb{E}[y|x] = x'\beta$ and that $\mathbb{V}[y|x] = \exp(x'\beta + \alpha)$, with (β, α) unknown. You are asked to estimate β.

1. Propose an estimation method that is asymptotically equivalent to GLS that would be computable were $\mathbb{V}[y|x]$ fully known.

2. In what sense is the feasible GLS estimator of Part 1 efficient? In which sense is it inefficient?

4.6 OLS and GLS are identical

Let $Y = X(\beta + v) + u$, where X is $n \times k$, Y and u are $n \times 1$, and β and v are $k \times 1$. The parameter of interest is β. The properties of (Y, X, u, v) are: $\mathbb{E}[u|X] = \mathbb{E}[v|X] = 0$, $\mathbb{E}[uu'|X] = \sigma^2 I_n$, $\mathbb{E}[vv'|X] = \Gamma$, $\mathbb{E}[uv'|X] = 0$. Y and X are observable, while u and v are not.

1. What are $\mathbb{E}[Y|X]$ and $\mathbb{V}[Y|X]$? Denote the latter by Σ. Is the environment homo- or heteroskedastic?

2. Write out the OLS and GLS estimators $\hat{\beta}$ and $\tilde{\beta}$ of β. Prove that in this model they are identical. Hint: First prove that $X'\hat{e} = 0$, where $\hat{e}$ is the $n \times 1$ vector of OLS residuals. Next prove that $X'\Sigma^{-1}\hat{e} = 0$. Then conclude. Alternatively, use formulae for the inverse of a sum of two matrices. The first method is preferable, being more "econometric".

3. Discuss benefits of using both estimators in this model.

4.7 OLS and GLS are equivalent

Let us have a regression written in a matrix form: $Y = X\beta + u$, where X is $n \times k$, Y and u are $n \times 1$, and β is $k \times 1$. The parameter of interest is β. The properties of u are: $\mathbb{E}[u|X] = 0$, $\mathbb{E}[uu'|X] = \Sigma$. Let it be also known that $\Sigma X = X\Theta$ for some $k \times k$ nonsingular matrix Θ.

1. Prove that in this model the OLS and GLS estimators $\hat{\beta}$ and $\tilde{\beta}$ of β have the same finite sample conditional variance.

2. Apply this result to the following regression on a constant:

$$y_i = \alpha + u_i,$$

where the disturbances are equicorrelated, that is, $\mathbb{E}[u_i] = 0$, $\mathbb{V}[u_i] = \sigma^2$ and $\mathbb{C}[u_i, u_j] = \rho\sigma^2$ for $i \neq j$.

4.8 Equicorrelated observations

Suppose $x_i = \theta + u_i$, where $\mathbb{E}[u_i] = 0$ and

$$\mathbb{E}[u_i u_j] = \begin{cases} 1 & \text{if } i = j \\ \gamma & \text{if } i \neq j \end{cases}$$

with $i, j = 1, \cdots, n$. Is $\bar{x}_n = \frac{1}{n}(x_1 + \cdots + x_n)$ the best linear unbiased estimator of θ? Investigate $\bar{x}_n$ for consistency.

5. IV AND 2SLS ESTIMATORS

5.1 Instrumental variables in ARMA models

1. Consider an $AR(1)$ model $x_t = \rho x_{t-1} + e_t$ with $\mathbb{E}\left[e_t|I_{t-1}\right] = 0$, $\mathbb{E}\left[e_t^2|I_{t-1}\right] = \sigma^2$, and $|\rho| < 1$. We can look at this as an instrumental variables regression that implies, among others, instruments x_{t-1}, x_{t-2}, $\cdots$. Find the asymptotic variance of the instrumental variables estimator that uses instrument x_{t-j}, where $j = 1, 2, \cdots$. What does your result suggest on what the optimal instrument must be?

2. Consider an $ARMA(1,1)$ model $y_t = \alpha y_{t-1} + e_t - \theta e_{t-1}$ with $|\alpha| < 1$, $|\theta| < 1$ and $\mathbb{E}\left[e_t|I_{t-1}\right] = 0$. Suppose you want to estimate α by just-identifying IV. What instrument would you use and why?

5.2 Inappropriate 2SLS

Consider the model

$$y_i = \alpha z_i^2 + u_i, \quad z_i = \pi x_i + v_i,$$

where (x_i, u_i, v_i) are IID, $\mathbb{E}\left[u_i|x_i\right] = \mathbb{E}\left[v_i|x_i\right] = 0$ and $\mathbb{V}\left[\begin{pmatrix} u_i \\ v_i \end{pmatrix} | x_i\right] = \Sigma$, with Σ unknown.

1. Show that α, π and Σ are identified. Suggest analog estimators for these parameters.

2. Consider the following two stage estimation method. In the first stage, regress z_i on x_i and define $\hat{z}_i = \hat{\pi} x_i$, where $\hat{\pi}$ is the OLS estimator. In the second stage, regress y_i in $\hat{z}_i^2$ to obtain the least squares estimate of α. Show that the resulting estimator of α is inconsistent.

3. Suggest a method in the spirit of 2SLS for estimating α consistently.

5.3 Inconsistency under alternative

Suppose that

$$y = \alpha + \beta x + u,$$

where u is distributed $\mathcal{N}(0, \sigma^2)$ independently of x. The variable x is unobserved. Instead we observe $z = x + v$, where v is distributed $\mathcal{N}(0, \eta^2)$ independently of x and u. Given a sample of size n, it is proposed to run the linear regression of y on z and use a conventional t-test to test the null hypothesis $\beta = 0$. Critically evaluate this proposal.

5.4 Trade and growth

In the paper "Does Trade Cause Growth?" (*American Economic Review*, June 1999), Jeffrey Frankel and David Romer study the effect of trade on income. Their simple specification is

$$\log Y_i = \alpha + \beta T_i + \gamma W_i + \varepsilon_i, \tag{5.1}$$

where Y_i is per capita income, T_i is international trade, W_i is within-country trade, and ε_i reflects other influences on income. Since the latter is likely to be correlated with the trade variables, Frankel and Romer decide to use instrumental variables to estimate the coefficients in (5.1). As instruments, they use a country's proximity to other countries P_i and its size S_i, so that

$$T_i = \psi + \phi P_i + \delta_i \tag{5.2}$$

and

$$W_i = \eta + \lambda S_i + \nu_i, \tag{5.3}$$

where δ_i and ν_i are the best linear prediction errors.

1. As the key identifying assumption, Frankel and Romer use the fact that countries' geographical characteristics P_i and S_i are uncorrelated with the error term in (5.1). Provide an economic rationale for this assumption and a detailed explanation how to estimate (5.1) when one has data on Y, T, W, P and S for a list of countries.

2. Unfortunately, data on within-country trade are not available. Determine if it is possible to estimate any of the coefficients in (5.1) without further assumptions. If it is, provide all the details on how to do it.

3. In order to be able to estimate key coefficients in (5.1), Frankel and Romer add another identifying assumption that P_i is uncorrelated with the error term in (5.3). Provide a detailed explanation how to estimate (5.1) when one has data on Y, T, P and S for a list of countries.

4. Frankel and Romer estimated an equation similar to (5.1) by OLS and IV and found out that the IV estimates are greater than the OLS estimates. One explanation may be that the discrepancy is due to a sampling error. Provide another, more econometric, explanation why there is a discrepancy and what the reason is that the IV estimates are larger.

6. EXTREMUM ESTIMATORS

6.1 Extremum estimators

Consider the following class of estimators called *Extremum Estimators*. Let the true parameter β be the unique solution of the following optimization problem:

$$\beta = \arg\max_{b \in \mathrm{B}} \mathbb{E}\left[f(z,b)\right], \tag{6.1}$$

where $z \in \mathbb{R}^l$ is a random vector on which the data are available, $b \in \mathbb{R}^k$ is a parameter, f is a known function, B is a parameter space. The latter is assumed to be compact, so that there are no problems with existence of the optimizer. The data z_i, $i = 1, ..., n$, are IID.

1. Construct the extremum estimator $\hat{\beta}$ of β by using the analogy principle applied to (6.1). Assuming that consistency holds, derive the asymptotic distribution of $\hat{\beta}$. Explicitly state all assumptions that you made to derive it.

2. Verify that your answer to part 1 reconciles with the results we obtained in class during the last module for the NLLS and WNLLS estimators, by appropriately choosing the form of function f.

6.2 Regression on constant

Apply the results of the previous problem to the following model:

$$y_i = \beta + e_i, \quad i = 1, \cdots, n,$$

where all variables are scalars. Assume that $\{e_i\}$ are IID with $\mathbb{E}[e_i] = 0$, $\mathbb{E}[e_i^2] = \beta^2$, $\mathbb{E}[e_i^3] = 0$ and $\mathbb{E}[e_i^4] = \kappa$. Consider the following three estimators of β:

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

$$\hat{\beta}_2 = \arg\min_b \left\{ \log b^2 + \frac{1}{nb^2} \sum_{i=1}^{n} (y_i - b)^2 \right\},$$

$$\hat{\beta}_3 = \frac{1}{2} \arg\min_b \sum_{i=1}^{n} \left(\frac{y_i}{b} - 1 \right)^2.$$

Derive the asymptotic distributions of these three estimators. Which of them would you prefer most on the asymptotic basis? Bonus question: what was the idea behind each of the three estimators?

6.3 Quadratic regression

Consider a nonlinear regression model

$$y_i = (\beta_0 + x_i)^2 + u_i,$$

where we assume:

(A) Parameter space is $B = \left[-\frac{1}{2}, +\frac{1}{2}\right]$.

(B) $\{u_i\}$ are IID with $\mathbb{E}[u_i] = 0$, $\mathbb{V}[u_i] = \sigma_0^2$.

(C) $\{x_i\}$ are IID with uniform distribution over $[1, 2]$, distributed independently of $\{u_i\}$. In particular, this implies $\mathbb{E}\left[x_i^{-1}\right] = \ln 2$ and $\mathbb{E}[x_i^r] = \frac{1}{1+r}(2^{r+1} - 1)$ for integer $r \neq -1$.

Define two estimators of β_0:

1. $\hat{\beta}$ minimizes $S_n(\beta) = \sum_{i=1}^n \left[y_i - (\beta + x_i)^2\right]^2$ over B.
2. $\tilde{\beta}$ minimizes $W_n(\beta) = \sum_{i=1}^n \left\{\frac{y_i}{(\beta + x_i)^2} + \ln(\beta + x_i)^2\right\}$ over B.

For the case $\beta_0 = 0$, obtain asymptotic distributions of $\hat{\beta}$ and $\tilde{\beta}$. Which one of the two do you prefer on the asymptotic basis?

7. MAXIMUM LIKELIHOOD ESTIMATION

7.1 MLE for three distributions

1. A random variable X is said to have a Pareto distribution with parameter λ, denoted $X \sim$ Pareto(λ), if it is continuously distributed with density

$$f_X(x|\lambda) = \begin{cases} \lambda x^{-(\lambda+1)}, & \text{if } x > 1, \\ 0, & \text{otherwise.} \end{cases}$$

   A random sample $x_1, \cdots, x_n$ from the Pareto(λ) population is available.

   (i) Derive the ML estimator $\hat{\lambda}$ of λ, prove its consistency and find its asymptotic distribution.

   (ii) Derive the Wald, Likelihood Ratio and Lagrange Multiplier test statistics for testing the null hypothesis $H_0 : \lambda = \lambda_0$ against the alternative hypothesis $H_a : \lambda \neq \lambda_0$. Do any of these statistics coincide?

2. Let $x_1, \cdots, x_n$ be a random sample from $\mathcal{N}(\mu, \mu^2)$. Derive the ML estimator $\hat{\mu}$ of μ and prove its consistency.

3. Let $x_1, \cdots, x_n$ be a random sample from a population of x distributed uniformly on $[0, \theta]$. Construct an asymptotic confidence interval for θ with significance level 5% by employing a maximum likelihood approach.

7.2 Comparison of ML tests

[1]Berndt and Savin in 1977 showed that $\mathcal{W} \geq \mathcal{LR} \geq \mathcal{LM}$ for the case of a multivariate regression model with normal disturbances. Ullah and Zinde-Walsh in 1984 showed that this inequality is not robust to non-normality of the disturbances. In the spirit of the latter article, this problem considers simple examples from non-normal distributions and illustrates how this conflict among criteria is affected.

1. Consider a random sample $x_1, \cdots, x_n$ from a Poisson distribution with parameter λ. Show that testing $\lambda = 3$ versus $\lambda \neq 3$ yields $\mathcal{W} \geq \mathcal{LM}$ for $\bar{x} \leq 3$ and $\mathcal{W} \leq \mathcal{LM}$ for $\bar{x} \geq 3$.

2. Consider a random sample $x_1, \cdots, x_n$ from an exponential distribution with parameter θ. Show that testing $\theta = 3$ versus $\theta \neq 3$ yields $\mathcal{W} \geq \mathcal{LM}$ for $0 < \bar{x} \leq 3$ and $\mathcal{W} \leq \mathcal{LM}$ for $\bar{x} \geq 3$.

3. Consider a random sample $x_1, \cdots, x_n$ from a Bernoulli distribution with parameter θ. Show that for testing $\theta = \frac{1}{2}$ versus $\theta \neq \frac{1}{2}$, we always get $\mathcal{W} \geq \mathcal{LM}$. Show also that for testing $\theta = \frac{2}{3}$ versus $\theta \neq \frac{2}{3}$, we get $\mathcal{W} \leq \mathcal{LM}$ for $\frac{1}{3} \leq \bar{x} \leq \frac{2}{3}$ and $\mathcal{W} \geq \mathcal{LM}$ for $0 < \bar{x} \leq \frac{1}{3}$ or $\frac{2}{3} \leq \bar{x} \leq 1$.

[1]This problem closely follows Badi H. Baltagi (2000) Conflict Among Criteria for Testing Hypotheses: Examples from Non-Normal Distributions. *Econometric Theory* 16, Problem 00.2.4.

7.3 Individual effects

Suppose $\{(x_i, y_i)\}_{i=1}^{n}$ is a serially independent sample from a sequence of jointly normal distributions with $\mathbb{E}[x_i] = \mathbb{E}[y_i] = \mu_i$, $\mathbb{V}[x_i] = \mathbb{V}[y_i] = \sigma^2$, and $\mathbb{C}[x_i, y_i] = 0$ (i.e., x_i and y_i are independent with common but varying means and a constant common variance). All parameters are unknown. Derive the maximum likelihood estimate of σ^2 and show that it is inconsistent. Explain why. Find an estimator of σ^2 which would be consistent.

7.4 Does the link matter?

[2]Consider a binary random variable y and a scalar random variable x such that

$$\mathbb{P}\{y = 1|x\} = F(\alpha + \beta x),$$

where the link $F(\cdot)$ is a continuous distribution function. Show that when x assumes only two different values, the value of the log-likelihood function evaluated at the maximum likelihood estimates of α and β is independent of the form of the link function. What are the maximum likelihood estimates of α and β?

7.5 Nuisance parameter in density

Let $z_i \equiv (y_i, x_i')'$ have a joint density of the form

$$f(Z|\theta_0) = f_c(Y|X, \gamma_0, \delta_0) f_m(X|\delta_0),$$

where $\theta_0 \equiv (\gamma_0, \delta_0)$, both γ_0 and δ_0 are scalar parameters, and f_c and f_m denote the conditional and marginal distributions, respectively. Let $\hat{\theta}_c \equiv (\hat{\gamma}_c, \hat{\delta}_c)$ be the conditional ML estimators of γ_0 and δ_0, and $\hat{\delta}_m$ be the marginal ML estimator of δ_0. Now define

$$\tilde{\gamma} \equiv \arg\max_{\gamma} \sum_i \ln f_c(y_i|x_i, \gamma, \hat{\delta}_m),$$

a two-step estimator of subparameter γ_0 which uses marginal ML to obtain a preliminary estimator of the "nuisance parameter" δ_0. Find the asymptotic distribution of $\tilde{\gamma}$. How does it compare to that for $\hat{\gamma}_c$? You may assume all the needed regularity conditions for consistency and asymptotic normality to hold.

Hint: You need to apply the Taylor's expansion twice, i.e. for both stages of estimation.

7.6 MLE versus OLS

Consider the model where y_i is regressed only on a constant:

$$y_i = \alpha + e_i, \quad i = 1, \ldots, n,$$

[2]This problem closely follows Joao M.C. Santos Silva (1999) Does the link matter? *Econometric Theory* 15, Problem 99.5.3.

where e_i conditioned on x_i is distributed as $\mathcal{N}(0, x_i^2\sigma^2)$; x_i's are drawn from a population of some random variable x that is not present in the regression; σ^2 is unknown; y_i's and x_i's are observable, e_i's are unobservable; the pairs (y_i, x_i) are IID.

1. Find the OLS estimator $\hat{\alpha}_{OLS}$ of α. Is it unbiased? Consistent? Obtain its asymptotic distribution. Is $\hat{\alpha}_{OLS}$ the best linear unbiased estimator for α?

2. Find the ML estimator $\hat{\alpha}_{ML}$ of α and derive its asymptotic distribution. Is $\hat{\alpha}_{ML}$ unbiased? Is $\hat{\alpha}_{ML}$ asymptotically more efficient than $\hat{\alpha}_{OLS}$? Does your conclusion contradicts your answer to the last question of part 1? Why or why not?

7.7 MLE in heteroskedastic time series regression

Assume that data (y_t, x_t), $t = 1, 2, \cdots, T$, are stationary and ergodic and generated by

$$y_t = \alpha + \beta x_t + u_t,$$

where $u_t|x_t \sim \mathcal{N}(0, \sigma_t^2)$, $x_t \sim \mathcal{N}(0, v)$, $\mathbb{E}[u_t u_s|x_t, x_s] = 0$, $t \neq s$. Explain, without going into deep math, how to find estimates and their standard errors *for all parameters* when:

1. The entire σ_t^2 as a function of x_t is fully known.

2. The values of σ_t^2 at $t = 1, 2, \cdots, T$ are known.

3. It is known that $\sigma_t^2 = (\theta + \delta x_t)^2$, but the parameters θ and δ are unknown.

4. It is known that $\sigma_t^2 = \theta + \delta u_{t-1}^2$, but the parameters θ and δ are unknown.

5. It is only known that σ_t^2 is stationary.

7.8 Maximum likelihood and binary variables

Suppose Z and Y are discrete random variables taking values 0 or 1. The distribution of Z and Y is given by

$$\mathbb{P}\{Z = 1\} = \alpha, \quad \mathbb{P}\{Y = 1|Z\} = \frac{e^{\gamma Z}}{1 + e^{\gamma Z}}, \quad Z = 0, 1.$$

Here α and γ are scalar parameters of interest.

1. Find the ML estimator of (α, γ) (giving an explicit formula whenever possible) and derive its asymptotic distribution.

2. Suppose we want to test $H_0 : \alpha = \gamma$ using the asymptotic approach. Derive the t test statistic and describe in detail how you would perform the test.

3. Suppose we want to test $H_0 : \alpha = \frac{1}{2}$ using the bootstrap approach. Derive the LR (likelihood ratio) test statistic and describe in detail how you would perform the test.

7.9 Maximum likelihood and binary dependent variable

Suppose y is a discrete random variable taking values 0 or 1 representing some choice of an individual. The distribution of y given the individual's characteristic x is

$$\mathbb{P}\{y = 1|x\} = \frac{e^{\gamma x}}{1 + e^{\gamma x}},$$

where γ is the scalar parameter of interest. The data $\{y_i, x_i\}$, $i = 1, ..., n$, are IID. When deriving various estimators, try to make the formulas as explicit as possible.

1. Derive the ML estimator of γ and its asymptotic distribution.
2. Find the (nonlinear) regression function by regressing y on x. Derive the NLLS estimator of γ and its asymptotic distribution.
3. Show that the regression you obtained in Part 2 is heteroskedastic. Setting weights $\omega(x)$ equal to the variance of y conditional on x, derive the WNLLS estimator of γ and its asymptotic distribution.
4. Write out the systems of moment conditions implied by the ML, NLLS and WNLLS problems of Parts 1–3.
5. Rank the three estimators in terms of asymptotic efficiency. Do any of your findings appear unexpected? Give intuitive explanation for anything unusual.

7.10 Bootstrapping ML tests

1. For the likelihood ratio test of $H_0 : g(\theta) = 0$, we use the statistic

   $$\mathcal{LR} = 2\left(\max_{q\in\Theta} \ell_n(q) - \max_{q\in\Theta, g(q)=0} \ell_n(q)\right).$$

   Write out the formula (no need to describe the entire algorithm) for the bootstrap pseudo-statistic $\mathcal{LR}^*$.
2. For the Lagrange Multiplier test of $H_0 : g(\theta) = 0$, we use the statistic

   $$\mathcal{LM} = \frac{1}{n}\sum_i s\left(z_i, \hat{\theta}^R_{ML}\right)' \hat{J}^{-1} \sum_i s\left(z_i, \hat{\theta}^R_{ML}\right).$$

   Write out the formula (no need to describe the entire algorithm) for the bootstrap pseudo-statistic $\mathcal{LM}^*$.

7.11 Trivial parameter space

Consider a parametric model with density $f(X|\theta_0)$, known up to a parameter θ_0, but with $\Theta = \{\theta_1\}$, i.e. the parameter space is reduced to only one element. What is an ML estimator of θ_0, and what are its asymptotic properties?

8. GENERALIZED METHOD OF MOMENTS

8.1 GMM and chi-squared

Let z be distributed as $\chi^2(1)$. Then the moment function

$$m(z,q) = \begin{pmatrix} z-q \\ z^2-q^2-2q \end{pmatrix}$$

has mean zero for $q = 1$. Describe efficient GMM estimation of $\theta = 1$ in details.

8.2 Improved GMM

Consider GMM estimation with the use of the moment function

$$m(x,y,q) = \begin{pmatrix} x-q \\ y \end{pmatrix}.$$

Determine under what conditions the second restriction helps in reducing the asymptotic variance of the GMM estimator of θ.

8.3 Nonlinear simultaneous equations

Let

$$y_i = \beta x_i + u_i, \qquad x_i = \gamma y_i^2 + v_i, \qquad i = 1, \ldots, n,$$

where x_i's and y_i's are observable, but u_i's and v_i's are not. The data are IID across i.

1. Suppose we know that $\mathbb{E}[u_i] = \mathbb{E}[v_i] = 0$. When are β and γ identified? Propose analog estimators for these parameters.
2. Let also be known that $\mathbb{E}[u_i v_i] = 0$.

(a) Propose a method to estimate β and γ as efficiently as possible given the above information. Your estimator should be fully implementable given the data $\{x_i, y_i\}_{i=1}^n$. What is the asymptotic distribution of your estimator?

(b) Describe in detail how to test $H_0 : \beta = \gamma = 0$ using the bootstrap approach and the Wald test statistic.

(c) Describe in detail how to test $H_0 : \mathbb{E}[u_i] = \mathbb{E}[v_i] = \mathbb{E}[u_i v_i] = 0$ using the asymptotic approach.

8.4 Trinity for GMM

Derive the three classical tests ($\mathcal{W}$, $\mathcal{LR}$, $\mathcal{LM}$) for the composite null

$$H_0 : \theta \in \Theta_0 \equiv \{\theta : h(\theta) = 0\},$$

where $h : \mathbb{R}^k \to \mathbb{R}^q$, for the efficient GMM case. The analog for the Likelihood Ratio test will be called the *Distance Difference test*. Hint: treat the GMM objective function as the "normalized loglikelihood", and its derivative as the "sample score".

8.5 Testing moment conditions

In the linear model

$$y_i = x_i'\beta + u_i$$

under random sampling and the unconditional moment restriction $\mathbb{E}[x_i u_i] = 0$, suppose you wanted to test the additional moment restriction $\mathbb{E}\left[x_i u_i^3\right] = 0$, which might be implied by conditional symmetry of the error terms u_i.

A natural way to test for the validity of this extra moment condition would be to efficiently estimate the parameter vector β both with and without the additional restriction, and then to check whether the corresponding estimates differ significantly. Devise such a test and give step-by-step instructions for carrying it out.

8.6 Interest rates and future inflation

Frederic Mishkin in early 90's investigated whether the term structure of current nominal interest rates can give information about future path of inflation. He specified the following econometric model:

$$\pi_t^m - \pi_t^n = \alpha_{m,n} + \beta_{m,n}\left(i_t^m - i_t^n\right) + \eta_t^{m,n}, \quad \mathbb{E}_t\left[\eta_t^{m,n}\right] = 0, \tag{8.1}$$

where π_t^k is k-periods-into-the-future inflation rate, i_t^k is the current nominal interest rate for k-periods-ahead maturity, and $\eta_t^{m,n}$ is the prediction error.

1. Show how (8.1) can be obtained from the conventional econometric model that tests the hypothesis of conditional unbiasedness of interest rates as predictors of inflation. What restriction on the parameters in (8.1) implies that the term structure provides *no* information about future shifts in inflation? Determine the autocorrelation structure of $\eta_t^{m,n}$.

2. Describe in detail how you would test the hypothesis that the term structure provides no information about future shifts in inflation, by using overidentifying GMM and asymptotic theory. Make sure that you discuss such issues as selection of instruments, construction of the optimal weighting matrix, construction of the GMM objective function, estimation of asymptotic variance, etc.

3. Describe in detail how you would test for overidentifying restrictions that arose from your set of instruments, using the nonoverlapping blocks bootstrap approach.

4. Mishkin obtained the following results (standard errors in parentheses):

m, n (months)	$\alpha_{m,n}$	$\beta_{m,n}$	t-test of $\beta_{m,n}=0$	t-test of $\beta_{m,n}=1$
$3,1$	0.1421 (0.1851)	−0.3127 (0.4498)	−0.70	2.92
$6,3$	0.0379 (0.1427)	0.1813 (0.5499)	0.33	1.49
$9,6$	0.0826 (0.0647)	0.0014 (0.2695)	0.01	3.71

Discuss and interpret the estimates and results of hypotheses tests.

8.7 Spot and forward exchange rates

Consider a simple problem of prediction of spot exchange rates by forward rates:

$$s_{t+1} - s_t = \alpha + \beta\,(f_t - s_t) + e_{t+1}, \quad \mathbb{E}_t\,[e_{t+1}] = 0, \quad \mathbb{E}_t\left[e_{t+1}^2\right] = \sigma^2,$$

where s_t is the spot rate at t, f_t is the forward rate for one-month forwards at t, and $\mathbb{E}_t$ denotes expectation conditional on time t information. The current spot rate is subtracted to achieve stationarity. Suppose the researcher decides to use ordinary least squares to estimate α and β. Recall that the moment conditions used by the OLS estimator are

$$\mathbb{E}\,[e_{t+1}] = 0, \quad \mathbb{E}\,[(f_t - s_t)\,e_{t+1}] = 0. \tag{8.2}$$

1. Beside (8.2), there are other moment conditions that can be used in estimation:

$$\mathbb{E}\,[(f_{t-k} - s_{t-k})\,e_{t+1}] = 0,$$

because $f_{t-k} - s_{t-k}$ belongs to information at time t for any $k \geq 1$. Consider the case $k = 1$ and show that such moment condition is redundant.

2. Beside (8.2), there is another moment condition that can be used in estimation:

$$\mathbb{E}\,[(f_t - s_t)\,(f_{t+1} - f_t)] = 0,$$

because information at time t should be unable to predict future movements in forward rates. Although this moment condition does not involve α or β, its use may improve efficiency of estimation. Under what condition is the efficient GMM estimator using both moment conditions as efficient as the OLS estimator? Is this condition likely to be satisfied in practice?

8.8 Brief and exhaustive

Give concise but exhaustive answers to the following unrelated questions.

1. Let it be known that the scalar random variable w has mean μ and that its fourth central moment equals three times its squared variance (like for a normal random variable). Formulate a system of moment conditions for GMM estimation of μ.

2. Suppose an econometrician estimates parameters of a time series regression by GMM after having chosen an overidentifying vector of instrumental variables. He performs the overidentification test and claims: "A big value of the J-statistic is an evidence against validity of the chosen instruments". Comment on this claim.

3. We know that one should use recentering when bootstrapping a GMM estimator. We also know that the OLS estimator is one of GMM estimators. However, when we bootstrap the OLS estimator, we calculate $\hat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$ at each bootstrap repetition, and do not recenter. Resolve the contradiction.

8.9 Efficiency of MLE in GMM class

We proved that the ML estimator of a parameter is efficient in the class of extremum estimators of the same parameter. Prove that it is also efficient in the class of GMM estimators of the same parameter.

9. PANEL DATA

9.1 Alternating individual effects

Suppose that the unobservable individual effects in a one-way error component model are different across odd and even periods:

$$\begin{array}{ll} y_{it} = \mu_i^O + x_{it}'\beta + v_{it} & \text{for odd } t, \\ y_{it} = \mu_i^E + x_{it}'\beta + v_{it} & \text{for even } t, \end{array} \qquad (*)$$

where $t = 1, 2, \cdots, 2T$, $i = 1, \cdots n$. Note that there are $2T$ observations for each individual. We will call (9.1) "alternating effects" specification. As usual, we assume that v_{it} are $IID(0, \sigma_v^2)$ independent of x's.

1. There are two ways to arrange the observations: (a) in the usual way, first by individual, then by time for each individual; (b) first all "odd" observations in the usual order, then all "even" observations, so it is as though there are $2N$ "individuals" each having T observations. Find out the Q-matrices that wipe out individual effects for both arrangements and explain how they transform the original equations. For the rest of the problem, choose the Q-matrix to your liking.

2. Treating individual effects as fixed, describe the Within estimator and its properties. Develop an F-test for individual effects, allowing heterogeneity across odd and even periods.

3. Treating individual effects as random and assuming their independence of x's, v's and each other, propose a feasible GLS procedure. Consider two cases: (a) when the variance of "alternating effects" is the same: $\mathbb{V}\left[\mu_i^O\right] = \mathbb{V}\left[\mu_i^E\right] = \sigma_\mu^2$, (b) when the variance of "alternating effects" is different: $\mathbb{V}\left[\mu_i^O\right] = \sigma_O^2$, $\mathbb{V}\left[\mu_i^E\right] = \sigma_E^2$, $\sigma_O^2 \neq \sigma_E^2$.

9.2 Time invariant regressors

Consider a panel data model

$$y_{it} = x_{it}'\beta + z_i\gamma + \mu_i + v_{it}, \quad i = 1, 2, \cdots, n, \quad t = 1, 2, \cdots, T,$$

where n is large and T is small. One wants to estimate β and γ.

1. Explain how to efficiently estimate β and γ under (a) fixed effects, (b) random effects, whenever it is possible. State clearly all assumptions that you will need.

2. Consider the following proposal to estimate γ. At the first step, estimate the model $y_{it} = x_{it}'\beta + \pi_i + v_{it}$ by the least squares dummy variables approach. At the second step, take these estimates $\hat{\pi}_i$ and estimate the coefficient of the regression of $\hat{\pi}_i$ on z_i. Investigate the resulting estimator of γ for consistency. Can you suggest a better estimator of γ?

9.3 First differencing transformation

In a one-way error component model with fixed effects, instead of using individual dummies, one can alternatively eliminate individual effects by taking the first differencing (FD) transformation. After this procedure one has $n(T-1)$ equations without individual effects, so the vector β of structural parameters can be estimated by OLS. Evaluate this proposal.

10. NONPARAMETRIC ESTIMATION

10.1 Nonparametric regression with discrete regressor

Let (x_i, y_i), $i = 1, \cdots, n$ be an IID sample from the population of (x, y), where x has a discrete distribution with the support $a_{(1)}, \cdots, a_{(k)}$, where $a_{(1)} < \cdots < a_{(k)}$. Having written the conditional expectation $\mathbb{E}\left[y|x = a_{(j)}\right]$ in the form that allows to apply the analogy principle, propose an analog estimator $\hat{g}_j$ of $g_j = \mathbb{E}\left[y|x = a_{(j)}\right]$ and derive its asymptotic distribution.

10.2 Nonparametric density estimation

Suppose we have an IID sample $\{x_i\}_{i=1}^n$ and let

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[x_i \leq x\right]$$

denote the empirical distribution function if x_i, where $\mathbb{I}(\cdot)$ is an indicator function. Consider two density estimators:

- one-sided estimator:

$$\hat{f}_1(x) = \frac{\hat{F}(x+h) - \hat{F}(x)}{h}$$

- two-sided estimator:

$$\hat{f}_2(x) = \frac{\hat{F}(x+h/2) - \hat{F}(x-h/2)}{h}$$

Show that:

(a) $\hat{F}(x)$ is an unbiased estimator of $F(x)$. Hint: recall that $F(x) = \mathbb{P}\{x_i \leq x\} = \mathbb{E}\left[\mathbb{I}\left[x_i \leq x\right]\right]$.

(b) The bias of $\hat{f}_1(x)$ is $O\left(h^a\right)$. Find the value of a. Hint: take a second-order Taylor series expansion of $F\left(x+h\right)$ around x.

(c) The bias of $\hat{f}_2(x)$ is $O\left(h^b\right)$. Find the value of b. Hint: take a second-order Taylor series expansion of $F\left(x+\frac{h}{2}\right)$ and $F\left(x+\frac{h}{2}\right)$ around x.

Now suppose that we want to estimate the density at the sample mean $\bar{x}_n$, the sample minimum $x_{(1)}$ and the sample maximum $x_{(n)}$. Given the results in (b) and (c), what can we expect from the estimates at these points?

10.3 First difference transformation and nonparametric regression

This problem illustrates the use of a difference operator in nonparametric estimation with IID data. Suppose that there is a scalar variable z that takes values on a bounded support. For simplicity, let z be deterministic and compose a uniform grid on the unit interval $[0, 1]$. The other variables are IID. Assume that for the function $g(\cdot)$ below the following *Lipschitz condition* is satisfied:

$$|g(u) - g(v)| \leq G|u - v|$$

for some constant G.

1. Consider a nonparametric regression of y on z:

$$y_i = g(z_i) + e_i, \quad i = 1, \cdots, n, \tag{10.1}$$

   where $\mathbb{E}[e_i|z_i] = 0$. Let the data $\{(z_i, y_i)\}_{i=1}^n$ be ordered so that the z's are in increasing order. A *first difference transformation* results in the following set of equations:

$$y_i - y_{i-1} = g(z_i) - g(z_{i-1}) + e_i - e_{i-1}, \quad i = 2, \cdots, n. \tag{10.2}$$

   The target is to estimate $\sigma^2 \equiv \mathbb{E}\left[e_i^2\right]$. Propose its consistent estimator based on the FD-transformed regression (2). Prove consistency of your estimator.

2. Consider the following partially linear regression of y on x and z:

$$y_i = x_i'\beta + g(z_i) + e_i, \quad i = 1, \cdots, n, \tag{10.3}$$

   where $\mathbb{E}[e_i|x_i, z_i] = 0$. Let the data $\{(x_i, z_i, y_i)\}_{i=1}^n$ be ordered so that the z's are in increasing order. The target is to nonparametrically estimate g. Propose its consistent estimator based on the FD-transformation of (3). [Hint: on the first step, consistently estimate β from the FD-transformed regression.] Prove consistency of your estimator.

11. CONDITIONAL MOMENT RESTRICTIONS

11.1 Usefulness of skedastic function

Suppose that for the following linear regression model

$$y_i = x_i'\beta + e_i, \quad \mathbb{E}\left[e_i|x_i\right] = 0$$

the form of a skedastic function is

$$\mathbb{E}\left[e_i^2|x_i\right] = h(x_i, \beta, \pi),$$

where $h(\cdot)$ is a known smooth function, and π is an additional parameter vector. Compare asymptotic variances of optimal GMM estimators of β when only the first restriction or both restrictions are employed. Under what conditions does including the second restriction into a set of moment restrictions reduce asymptotic variance? Try to answer these questions in the general case, then specialize to the following cases:

1. the function $h(\cdot)$ does not depend on β;
2. the function $h(\cdot)$ does not depend on β and the distribution of e_i conditional on x_i is symmetric.

11.2 Symmetric regression error

Suppose that it is known that the equation

$$y = \alpha x + e$$

is a regression of y on x, i.e. that $\mathbb{E}\left[e|x\right] = 0$. All variables are scalars. The random sample $\{y_i, x_i\}_{i=1}^n$ is available.

1. The investigator also suspects that y, conditional on x, is distributed symmetrically around the conditional mean. Devise a Hausman specification test for this symmetry. Be specific and give all details at all stages when constructing the test.
2. Suppose that even though the Hausman test rejects symmetry, the investigator uses the assumption that $e|x \sim \mathcal{N}(0, \sigma^2)$. Derive the asymptotic properties of the QML estimator of α.

11.3 Optimal instrument in AR-ARCH model

Consider an $AR(1) - ARCH(1)$ model: $x_t = \rho x_{t-1} + \varepsilon_t$ where the distribution of ε_t conditional on I_{t-1} is symmetric around 0 with $\mathbb{E}\left[\varepsilon_t^2|I_{t-1}\right] = (1-\alpha) + \alpha\varepsilon_{t-1}^2$, where $0 < \rho, \alpha < 1$ and $I_t = \{x_t, x_{t-1}, \cdots\}$.

1. Let the space of admissible instruments for estimation of the $AR(1)$ part be

$$\mathcal{Z}_t = \left\{ \sum_{i=1}^{\infty} \phi_i x_{t-i}, \text{ s.t. } \sum_{i=1}^{\infty} \phi_i^2 < \infty \right\}.$$

   Using the optimality condition, find the optimal instrument as a function of the model parameters ρ and α. Outline how to construct its feasible version.

2. Use your intuition to speculate on relative efficiency of the optimal instrument you found in Part 1 versus the optimal instrument based on the conditional moment restriction $\mathbb{E}\left[\varepsilon_t | I_{t-1}\right] = 0$.

11.4 Modified Poisson regression and PML estimators

[1]Let the observable random variable y be distributed, conditionally on observable x and unobservable ε as Poisson with the parameter $\lambda(x) = \exp(x'\beta+\varepsilon)$, where $\mathbb{E}[\exp\varepsilon|x] = 1$ and $\mathbb{V}[\exp\varepsilon|x] = \sigma^2$. Suppose that vector x is distributed as multivariate standard normal.

1. Find the regression and skedastic functions, where the conditional information involves only x.

2. Find the asymptotic variances of the Nonlinear Least Squares (NLLS) and Weighted Nonlinear Least Squares (WNLLS) estimators of β.

3. Find the asymptotic variances of the Pseudo-Maximum Likelihood (PML) estimators of β based on

   (a) the normal distribution;

   (b) the Poisson distribution;

   (c) the Gamma distribution.

4. Rank the five estimators in terms of asymptotic efficiency.

11.5 Optimal instrument and regression on constant

Consider the following model:

$$y_i = \alpha + e_i, \quad i = 1, \ldots, n,$$

where unobservable e_i conditionally on x_i is distributed *symmetrically* with mean zero and variance $x_i^2\sigma^2$ with unknown σ^2. The data (y_i, x_i) are IID.

1. Construct a pair of conditional moment restrictions from the information about the conditional mean and conditional variance. Derive the optimal unconditional moment restrictions, corresponding to (a) the conditional restriction associated with the conditional mean; (b) the conditional restrictions associated with both the conditional mean and conditional variance.

[1]The idea of this problem is borrowed from Gourieroux, C. and Monfort, A. "Statistics and Econometric Models", Cambridge University Press, 1995.

2. Describe in detail the GMM estimators that correspond to the two optimal sets of unconditional moment restrictions of part 1. Note that in part 1(a) the parameter σ^2 is not identified, therefore propose your own estimator of σ^2 that differs from the one implied by part 1(b). All estimators that you construct should be fully feasible. If you use nonparametric estimation, give all the details. Your description should also contain estimation of asymptotic variances.

3. Compare the asymptotic properties of the GMM estimators that you designed.

4. Derive the Pseudo-Maximum Likelihood estimator of α and σ^2 of order 2 (PML2) that is based on the normal distribution. Derive its asymptotic properties. How does this estimator relate to the GMM estimators you obtained in part 2?

12. EMPIRICAL LIKELIHOOD

12.1 Common mean

Suppose we have the following moment restrictions: $\mathbb{E}[x] = \mathbb{E}[y] = \theta$.

1. Find the system of equations that yield the maximum empirical likelihood (MEL) estimator $\hat{\theta}$ of θ, the associated Lagrange multipliers $\hat{\lambda}$ and the implied probabilities $\hat{p}_i$. Derive the asymptotic variances of $\hat{\theta}$ and $\hat{\lambda}$ and show how to estimate them.
2. Reduce the number of parameters by eliminating the redundant ones. Then linearize the system of equations with respect to the Lagrange multipliers that are left, around their population counterparts of zero. This will help to find an approximate, but explicit solution for $\hat{\theta}$, $\hat{\lambda}$ and $\hat{p}_i$. Derive that solution and interpret it.
3. Instead of defining the objective function
$$\frac{1}{n}\sum_{i=1}^{n}\log p_i$$
as in the EL approach, let the objective function be
$$-\frac{1}{n}\sum_{i=1}^{n}p_i\log p_i.$$
This gives rise to the *exponential tilting* (ET) estimator of θ. Find the system of equations that yields the ET estimator of $\hat{\theta}$, the associated Lagrange multipliers $\hat{\lambda}$ and the implied probabilities $\hat{p}_i$. Derive the asymptotic variances of $\hat{\theta}$ and $\hat{\lambda}$ and show how to estimate them.

12.2 Kullback–Leibler Information Criterion

The Kullback–Leibler Information Criterion (KLIC) measures the distance between distributions, say $g(z)$ and $h(z)$:
$$KLIC(g:h) = \mathbb{E}_g\left[\log\frac{g(z)}{h(z)}\right],$$
where $\mathbb{E}_g[\cdot]$ denotes mathematical expectation according to $g(z)$.

Suppose we have the following moment condition:
$$\mathbb{E}\left[m(z_i, \underset{k\times 1}{\theta_0})\right] = \underset{\ell\times 1}{0}, \quad \ell \geq k,$$
and an IID sample $z_1, \cdots, z_n$ with no elements equal to each other. Denote by e the empirical distribution function (EDF), i.e. the one that assigns probability $\frac{1}{n}$ to each sample point. Denote by π a discrete distribution that assigns probability π_i to the sample point z_i, $i = 1, \cdots, n$.

1. Show that minimization of $KLIC(e : \pi)$ subject to $\sum_{i=1}^{n} \pi_i = 1$ and $\sum_{i=1}^{n} \pi_i m(z_i, \theta) = 0$ yields the Maximum Empirical Likelihood (MEL) value of θ and corresponding implied probabilities.

2. Now we switch the roles of e and π and consider minimization of $KLIC(\pi : e)$ subject to the same constraints. What familiar estimator emerges as the solution to this optimization problem?

3. Now suppose that we have *a priori* knowledge about the distribution of the data. So, instead of using the EDF, we use the distribution p that assigns known probability p_i to the sample point z_i, $i = 1, \cdots, n$ (of course, $\sum_{i=1}^{n} p_i = 1$). Analyze how the solutions to the optimization problems in parts 1 and 2 change.

4. Now suppose that we have postulated a family of densities $f(z, \theta)$ which is compatible with the moment condition. Interpret the value of θ that minimizes $KLIC(e : f)$.

Part II

Solutions

1. ASYMPTOTIC THEORY

1.1 Asymptotics of t-ratios

The solution is straightforward, once we determine to what vector to apply LLN and CLT.

(a) When $\mu = 0$, we have: $\overline{X} \xrightarrow{p} 0$, $\sqrt{n}\overline{X} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, therefore

$$\sqrt{n}T_n = \frac{\sqrt{n}\overline{X}}{\hat{\sigma}} \xrightarrow{d} \frac{1}{\sigma}\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 1)$$

(b) Consider the vector

$$W_n \equiv \begin{pmatrix} \overline{X} \\ \hat{\sigma}^2 \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_i \\ (X_i - \mu)^2 \end{pmatrix} - \begin{pmatrix} 0 \\ (\overline{X} - \mu)^2 \end{pmatrix}.$$

Due to the LLN, the last term goes in probability to the zero vector, and the first term, and thus the whole W_n, converges in probability to

$$\operatorname*{plim}_{n\to\infty} W_n = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

Moreover, since $\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, we have $\sqrt{n}\left(\overline{X} - \mu\right)^2 \xrightarrow{d} 0$.

Next, let $W_i \equiv \left(X_i\ (X_i - \mu)^2\right)'$. Then $\sqrt{n}\left(W_n - \operatorname*{plim}_{n\to\infty} W_n\right) \xrightarrow{d} \mathcal{N}(0, V)$, where $V \equiv \mathbb{V}[W_i]$. Let us calculate V. First, $\mathbb{V}[X_i] = \sigma^2$ and $\mathbb{V}\left[(X_i - \mu)^2\right] = \mathbb{E}\left[((X_i - \mu)^2 - \sigma^2)^2\right] = \tau - \sigma^4$. Second, $\mathbb{C}\left[X_i, (X_i - \mu)^2\right] = \mathbb{E}\left[(X_i - \mu)((X_i - \mu)^2 - \sigma^2)\right] = 0$. Therefore,

$$\sqrt{n}\left(W_n - \operatorname*{plim}_{n\to\infty} W_n\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau - \sigma^4 \end{pmatrix}\right)$$

Now use the Delta-Method with function

$$g\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \equiv \frac{t_1}{\sqrt{t_2}} \Rightarrow g'\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \frac{1}{\sqrt{t_2}}\begin{pmatrix} 1 \\ -\dfrac{t_1}{2t_2} \end{pmatrix}$$

to get

$$\sqrt{n}\left(T_n - \operatorname*{plim}_{n\to\infty} T_n\right) \xrightarrow{d} \mathcal{N}\left(0, 1 + \frac{\mu^2(\tau - \sigma^4)}{4\sigma^6}\right).$$

Indeed, the answer reduces to $\mathcal{N}(0, 1)$ when $\mu = 0$.

(c) Similarly we solve this part. Consider the vector

$$W_n \equiv \begin{pmatrix} \overline{X} \\ \overline{\sigma}^2 \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_i \\ X_i^2 \end{pmatrix}.$$

Due to the LLN, W_n, converges in probability to

$$\operatorname*{plim}_{n\to\infty} W_n = \begin{pmatrix} \mu \\ \mu^2+\sigma^2 \end{pmatrix}.$$

Next, $\sqrt{n}\left(W_n - \operatorname*{plim}_{n\to\infty} W_n\right) \xrightarrow{d} \mathcal{N}(0,V)$, where $V \equiv \mathbb{V}[W_i]$, $W_i \equiv \left(X_i\ X_i^2\right)'$. Let us calculate V. First, $\mathbb{V}[X_i] = \sigma^2$ and $\mathbb{V}\left[X_i^2\right] = \mathbb{E}\left[(X_i^2-\mu^2-\sigma^2)^2\right] = \tau + 4\mu^2\sigma^2 - \sigma^4$. Second, $\mathbb{C}\left[X_i, X_i^2\right] = \mathbb{E}\left[(X_i-\mu)(X_i^2-\mu^2-\sigma^2)\right] = 2\mu\sigma^2$. Therefore,

$$\sqrt{n}\left(W_n - \operatorname*{plim}_{n\to\infty} W_n\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & \tau+4\mu^2\sigma^2-\sigma^4 \end{pmatrix}\right)$$

Now use the Delta-Method with $g\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \dfrac{t_1}{\sqrt{t_2}}$ to get

$$\sqrt{n}\left(R_n - \operatorname*{plim}_{n\to\infty} R_n\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mu^2\tau - \mu^2\sigma^4 + 4\sigma^6}{4(\mu^2+\sigma^2)^3}\right).$$

The answer reduces to that of Part (b) iff $\mu = 0$. Under this condition, T_n and R_n are *asymptotically equivalent*.

1.2 Asymptotics with shrinking regressor

The formulae for the OLS estimators are

$$\hat\beta = \frac{\frac{1}{n}\sum_i y_i x_i - \frac{1}{n^2}\sum_i y_i \sum_i x_i}{\frac{1}{n}\sum_i x_i^2 - \left(\frac{1}{n}\sum_i x_i\right)^2}, \quad \hat\alpha = \bar y - \hat\beta \bar x, \quad \hat\sigma^2 = \frac{1}{n}\sum_i \hat e_i{}^2. \tag{1.1}$$

Let us talk about $\hat\beta$ first. From (1.1) it follows that

$$\begin{aligned}\hat\beta &= \frac{\frac{1}{n}\sum_i(\alpha+\beta x_i+u_i)x_i - \frac{1}{n^2}\sum_i(\alpha+\beta x_i+u_i)\sum_i x_i}{\frac{1}{n}\sum_i x_i^2 - (\frac{1}{n}\sum_i x_i)^2} \\ &= \beta + \frac{\frac{1}{n}\sum_i \rho^i u_i - \frac{1}{n^2}\sum_i u_i \sum_i \rho^i}{\frac{1}{n}\sum_i \rho^{2i} - \frac{1}{n^2}(\sum_i \rho^i)^2} = \beta + \frac{\sum_i \rho^i u_i - \frac{\rho(1-\rho^{1+n})}{1-\rho}\left(\frac{1}{n}\sum_i u_i\right)}{\frac{\rho^2(1-\rho^{2(n+1)})}{1-\rho^2} - \frac{1}{n}\left(\frac{\rho(1-\rho^{(n+1)})}{1-\rho}\right)^2}\end{aligned}$$

which converges to

$$\beta + \frac{1-\rho^2}{\rho^2}\operatorname*{plim}_{n\to\infty}\sum_{i=1}^n \rho^i u_i,$$

if $\xi \equiv \operatorname{plim}\sum_i \rho^i u_i$ exists and is a well-defined random variable. It has $\mathbb{E}[\xi] = 0$, $\mathbb{E}\left[\xi^2\right] = \sigma^2\frac{\rho^2}{1-\rho^2}$ and $\mathbb{E}\left[\xi^3\right] = \nu\frac{\rho^3}{1-\rho^3}$. Hence

$$\hat\beta - \beta \xrightarrow{d} \frac{1-\rho^2}{\rho^2}\xi. \tag{1.2}$$

Now let us look at $\hat\alpha$. Again, from (1.1) we see that

$$\hat\alpha = \alpha + (\beta - \hat\beta)\cdot\frac{1}{n}\sum_i \rho^i + \frac{1}{n}\sum_i u_i \xrightarrow{p} \alpha,$$

where we used (1.2) and the LLN for u_i. Next,

$$\sqrt{n}(\hat{\alpha}-\alpha) = \frac{1}{\sqrt{n}}(\beta-\hat{\beta})\frac{\rho(1-\rho^{1+n})}{1-\rho} + \frac{1}{\sqrt{n}}\sum_i u_i = U_n + V_n.$$

Because of (1.2), $U_n \xrightarrow{p} 0$. From the CLT it follows that $V_n \xrightarrow{d} \mathcal{N}(0,\sigma^2)$. Together,

$$\sqrt{n}(\hat{\alpha}-\alpha) \xrightarrow{d} \mathcal{N}(0,\sigma^2).$$

Lastly, let us look at $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i \hat{e}_i^2 = \frac{1}{n}\sum_i \Big((\alpha-\hat{\alpha}) + (\beta-\hat{\beta})x_i + u_i\Big)^2. \tag{1.3}$$

Using the facts that: (1) $(\alpha-\hat{\alpha})^2 \xrightarrow{p} 0$, (2) $(\beta-\hat{\beta})^2/n \xrightarrow{p} 0$, (3) $\frac{1}{n}\sum_i u_i^2 \xrightarrow{p} \sigma^2$, (4) $\frac{1}{n}\sum_i u_i \xrightarrow{p} 0$, (5) $\frac{1}{\sqrt{n}}\sum_i \rho^i u_i \xrightarrow{p} 0$, we can derive that

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

The rest of this solution is optional and is usually not meant when the asymptotics of $\hat{\sigma}^2$ is concerned. Before proceeding to deriving its asymptotic distribution, we would like to mark out that $(\beta-\hat{\beta})/n^\delta \xrightarrow{p} 0$ and $(\sum_i \rho^i u_i)/n^\delta \xrightarrow{p} 0$ for any $\delta > 0$. Using the same algebra as before we have

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2) \overset{A}{\sim} \frac{1}{\sqrt{n}}\sum_i (u_i^2-\sigma^2),$$

since the other terms converge in probability to zero. Using the CLT, we get

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2) \xrightarrow{d} \mathcal{N}(0, m_4),$$

where $m_4 = \mathbb{E}\left[u_i^4\right] - \sigma^4$, provided that it is finite.

1.3 Creeping bug on simplex

Since x_k and y_k are perfectly correlated, it suffices to consider either one, say, x_k. Note that at each step x_k increases by $\frac{1}{k}$ with probability p, or stays the same. That is, $x_k = x_{k-1} + \frac{1}{k}\xi_k$, where ξ_k is IID Bernoulli(p). This means that $x_k = \frac{1}{k}\sum_{i=1}^k \xi_i$ which by the LLN converges in probability to $\mathbb{E}[\xi_i] = p$ as $k \to \infty$. Therefore, $\operatorname{plim}(x_k, y_k) = (p, 1-p)$. Next, due to the CLT,

$$\sqrt{n}\left(x_k - \operatorname{plim} x_k\right) \xrightarrow{d} \mathcal{N}\left(0, p(1-p)\right).$$

Therefore, the rate of convergence is $\sqrt{n}$, as usual, and

$$\sqrt{n}\left(\begin{pmatrix} x_k \\ y_k \end{pmatrix} - \operatorname{plim}\begin{pmatrix} x_k \\ y_k \end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} p(1-p) & -p(1-p) \\ -p(1-p) & p(1-p) \end{pmatrix}\right).$$

1.4 Asymptotics of rotated logarithms

Use the Delta-Method for

$$\sqrt{n}\left(\begin{pmatrix}U_n\\ V_n\end{pmatrix}-\begin{pmatrix}\mu_u\\ \mu_v\end{pmatrix}\right)\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix},\Sigma\right)$$

and $g\begin{pmatrix}x\\ y\end{pmatrix}=\begin{pmatrix}\ln x-\ln y\\ \ln x+\ln y\end{pmatrix}$. We have

$$\frac{\partial g}{\partial(x\ y)}\begin{pmatrix}x\\ y\end{pmatrix}=\begin{pmatrix}1/x & -1/y\\ 1/x & 1/y\end{pmatrix},\quad G=\frac{\partial g}{\partial(x\ y)}\begin{pmatrix}\mu_u\\ \mu_v\end{pmatrix}=\begin{pmatrix}1/\mu_u & -1/\mu_v\\ 1/\mu_u & 1/\mu_v\end{pmatrix},$$

so

$$\sqrt{n}\left(\begin{pmatrix}\ln U_n-\ln V_n\\ \ln U_n+\ln V_n\end{pmatrix}-\begin{pmatrix}\ln\mu_u-\ln\mu_v\\ \ln\mu_u+\ln\mu_v\end{pmatrix}\right)\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix},G\Sigma G'\right),$$

where

$$G\Sigma G'=\begin{pmatrix}\dfrac{\omega_{uu}}{\mu_u^2}-\dfrac{2\omega_{uv}}{\mu_u\mu_v}+\dfrac{\omega_{vv}}{\mu_v^2} & \dfrac{\omega_{uu}}{\mu_u^2}-\dfrac{\omega_{vv}}{\mu_v^2}\\ \dfrac{\omega_{uu}}{\mu_u^2}-\dfrac{\omega_{vv}}{\mu_v^2} & \dfrac{\omega_{uu}}{\mu_u^2}+\dfrac{2\omega_{uv}}{\mu_u\mu_v}+\dfrac{\omega_{vv}}{\mu_v^2}\end{pmatrix}.$$

It follows that $\ln U_n-\ln V_n$ and $\ln U_n+\ln V_n$ are asymptotically independent when $\dfrac{\omega_{uu}}{\mu_u^2}=\dfrac{\omega_{vv}}{\mu_v^2}$.

1.5 Trended vs. differenced regression

1. The OLS estimator $\hat\beta$ in that case is

$$\hat\beta=\frac{\sum_{t=1}^T(y_t-\frac{1}{T}\sum_{t=1}^T y_t)(t-\frac{1}{T}\sum_{t=1}^T y_t)}{\sum_{t=1}^T\left(t-\frac{1}{T}\sum_{t=1}^T t\right)^2}.$$

Then

$$\hat\beta-\beta=\left(\frac{1}{\frac{1}{T^3}\sum_{t=1}^T t^2-\left(\frac{1}{T^2}\sum_{t=1}^T t\right)^2},-\frac{\frac{1}{T^2}\sum_{t=1}^T t}{\frac{1}{T^3}\sum t^2-\left(\frac{1}{T^2}\sum_{t=1}^T t\right)^2}\right)\left[\begin{array}{c}\frac{1}{T^3}\sum_{t=1}^T\varepsilon_t t\\ \frac{1}{T^2}\sum_{t=1}^T\varepsilon_t\end{array}\right].$$

Now,

$$T^{3/2}(\hat\beta-\beta)=\left(\frac{1}{\frac{1}{T^3}\sum_{t=1}^T t^2-\left(\frac{1}{T^2}\sum_{t=1}^T t\right)^2},-\frac{\frac{1}{T^2}\sum_{t=1}^T t}{\frac{1}{T^3}\sum_{t=1}^T t^2-\left(\frac{1}{T^2}\sum_{t=1}^T t\right)^2}\right)\frac{1}{\sqrt{T}}\sum_{t=1}^T\left[\begin{array}{c}\frac{t}{T}\varepsilon_t\\ \varepsilon_t\end{array}\right].$$

Since

$$\sum_{t=1}^T t=\frac{T(T+1)}{2},\quad\sum_{t=1}^T t^2=\frac{T(T+1)(2T+1)}{6},$$

it is easy to see that the first vector converges to $(12, -6)$. Assuming that all conditions for the CLT for heterogenous martingale difference sequences (e.g., Potscher and Prucha, Theorem 4.12; Hamilton, Proposition 7.8) hold, we find that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix}\frac{t}{T}\varepsilon_t\\ \varepsilon_t\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\sigma^2\begin{pmatrix}\frac{1}{3} & \frac{1}{2}\\ \frac{1}{2} & 1\end{pmatrix}\right),$$

since

$$\lim\frac{1}{T}\sum_{t=1}^{T}\mathbb{V}\left[\frac{t}{T}\varepsilon_t\right]=\sigma^2\lim\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t}{T}\right)^2=\frac{1}{3},$$

$$\lim\frac{1}{T}\sum_{t=1}^{T}\mathbb{V}\left[\varepsilon_t\right]=\sigma^2,$$

$$\lim\frac{1}{T}\sum_{t=1}^{T}\mathbb{C}\left[\frac{t}{T}\varepsilon_t,\varepsilon_t\right]=\sigma^2\lim\frac{1}{T}\sum_{t=1}^{T}\frac{t}{T}=\frac{1}{2}.$$

Consequently,

$$T^{3/2}(\hat{\beta}-\beta)\to(12,-6)\cdot\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\sigma^2\begin{pmatrix}\frac{1}{3} & \frac{1}{2}\\ \frac{1}{2} & 1\end{pmatrix}\right)=\mathcal{N}(0,12\sigma^2).$$

2. Clearly, that for regression $y_t - y_{t-1} = \beta + \varepsilon_t - \varepsilon_{t-1}$ OLS estimator is

$$\check{\beta}=\frac{1}{T}\sum_{t=1}^{T}(y_t-y_{t-1})=\beta+\frac{\varepsilon_T-\varepsilon_0}{T}.$$

So, $T(\check{\beta}-\beta)=\varepsilon_T-\varepsilon_0\sim\mathcal{D}(0,2\sigma^2)$.

3. When T is sufficiently large, $\hat{\beta}\overset{A}{\sim}\mathcal{N}\left(\beta,\frac{12\sigma^2}{T^3}\right)$, and $\check{\beta}\sim\mathcal{D}\left(\beta,\frac{2\sigma^2}{T^2}\right)$. It is easy to see that for large T, the (approximate) variance of the first estimator is less than that of the second.

1.6 Second-order Delta-Method

(a) From CLT, $\sqrt{n}S_n \xrightarrow{d} \mathcal{N}(0,1)$. Using the Mann–Wald theorem for $g(x) = x^2$, we have $nS_n^2 \xrightarrow{d} \chi^2(1)$.

(b) The Taylor expansion around $\cos(0) = 1$ yields $\cos(S_n) = 1 - \frac{1}{2}\cos(S_n^*)S_n^2$, where $S_n^* \in [0, S_n]$. From LLN and the Mann–Wald theorem, $\cos(S_n^*) \xrightarrow{p} 1$, and from the Slutsky theorem, $2n(1-\cos(S_n)) \xrightarrow{d} \chi^2(1)$.

(c) Let $z_n \xrightarrow{p} z = \text{const}$ and $\sqrt{n}(z_n - z) \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\right)$. Let g be twice continuously differentiable at z with $g'(z) = 0$ and $g''(z) \neq 0$. Then

$$\frac{2n}{\sigma^2}\frac{g(z_n)-g(z)}{g''(z)}\xrightarrow{d}\chi^2(1).$$

Proof. Indeed, as $g'(z) = 0$, from the second-order Taylor expansion,

$$g(z_n) = g(z) + \frac{1}{2} g''(z^*)(z_n - z)^2,$$

and, since $g''(z^*) \xrightarrow{p} g''(z)$ and $\dfrac{\sqrt{n}(z_n - z)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$, we have

$$\frac{2n}{\sigma^2} \frac{g(z_n) - g(z)}{g''(z)} = \left[\frac{\sqrt{n}(z_n - z)}{\sigma} \right]^2 \xrightarrow{d} \chi^2(1).$$

QED

1.7 Brief and exhaustive

1. Observe that $x_t = \rho x_{t-1} + e_t$, $e_t = \varepsilon_t + \theta\varepsilon_{t-1}$ is an ARMA(1,1) process. Since $\mathbb{C}[x_{t-1}, e_t] = \theta\sigma^2 \neq 0$, the OLS estimator is inconsistent:

$$\hat{\rho}_{OLS} = \frac{\sum x_t x_{t-1}}{\sum x_{t-1}^2} = \rho + \frac{\sum e_t x_{t-1}}{\sum x_{t-1}^2} \xrightarrow{p} \rho + \frac{\mathbb{E}[e_t x_{t-1}]}{\mathbb{E}[x_{t-1}^2]} \neq \rho.$$

   Note an interesting thing. We do not consider explosive processes where $|\rho| > 1$, but the case $\rho = 1$ deserves attention. From the above it may seem that then $\hat{\rho}_{OLS} \xrightarrow{p} \rho$, since $\mathbb{E}[x_{t-1}^2]$ has $1 - \rho^2$ in the denominator. There is a fallacy in this argument, since the usual asymptotic fails when $\rho = 1$. However, the unit root asymptotics says that we do have $\hat{\rho}_{OLS} \xrightarrow{p} \rho$ in that case even though the error term is correlated with the regressor.

2. The long-run variance is just a two-sided infinite sum of all covariances, i.e. $3 + 2 \cdot 2 + 2 \cdot 1 = 9$.

3. The long-run variance is $V_{ze} = \sum_{j=-\infty}^{+\infty} \mathbb{C}[z_t e_t, z_{t-j} e_{t-j}]$. Since e_t and z_t are scalars, independent and $\mathbb{E}[e_t] = 0$, we have $\mathbb{C}[z_t e_t, z_{t-j} e_{t-j}] = \mathbb{E}[z_t z_{t-j}]\,\mathbb{E}[e_t e_{t-j}]$. Let for simplicity z_t also have zero mean. Then $\mathbb{E}[z_t z_{t-j}] = \rho_z^j \left(1 - \rho_z^2\right)^{-1} \sigma_z^2$ and $\mathbb{E}[e_t e_{t-j}] = \rho_e^j \left(1 - \rho_e^2\right)^{-1} \sigma_e^2$, where ρ_z, σ_z^2, ρ_e, σ_e^2 are $AR(1)$ parameters. To sum up,

$$V_{ze} = \frac{\sigma_z^2}{1 - \rho_z^2} \frac{\sigma_e^2}{1 - \rho_e^2} \sum_{j=-\infty}^{+\infty} \rho_z^j \rho_e^j = \frac{1 + \rho_z \rho_e}{(1 - \rho_z \rho_e)(1 - \rho_z^2)(1 - \rho_e^2)} \sigma_z^2 \sigma_e^2..$$

   To estimate V_{ze}, find the OLS estimates $\hat{\rho}_z$, $\hat{\sigma}_z^2$, $\hat{\rho}_e$, $\hat{\sigma}_e^2$ of the $AR(1)$ regressions and plug them in. The resulting $\hat{V}_{ze}$ will be consistent by the Continuous Mapping Theorem.

1.8 Asymptotics of averages of AR(1) and MA(1)

Note that y_t can be rewritten as $y_t = \sum_{j=0}^{+\infty} \rho^j x_{t-j}$

1. (i) y_t is not an MDS relative to own past $\{y_{t-1}, y_{t-2}, ...\}$, because it is correlated with older y_t's; (ii) z_t is an MDS relative to $\{x_{t-2}, x_{t-3}, ...\}$, but is not an MDS relative to own past $\{z_{t-1}, z_{t-2}, ...\}$, because z_t and z_{t-1} are correlated through x_{t-1}.

2. (i) By the CLT for the general stationary and ergodic case, $\sqrt{T}\overline{y}_T \xrightarrow{d} \mathcal{N}(0, q_{yy})$, where $q_{yy} = \sum_{j=-\infty}^{+\infty} \underbrace{\mathbb{C}[y_t, y_{t-j}]}_{\gamma_j}$. It can be shown that for an AR(1) process, $\gamma_0 = \frac{\sigma^2}{1-\rho^2}$, $\gamma_j = \gamma_{-j} = \frac{\sigma^2}{1-\rho^2}\rho^j$. Therefore, $q_{yy} = \sum_{j=-\infty}^{+\infty} \gamma_j = \frac{\sigma^2}{(1-\rho)^2}$. (ii) By the CLT for the general stationary and ergodic case, $\sqrt{T}\overline{z}_T \xrightarrow{d} \mathcal{N}(0, q_{zz})$, where $q_{zz} = \gamma_0 + 2\gamma_1 + 2\sum_{j=2}^{+\infty} \underbrace{\gamma_1}_{=0} = (1+\theta^2)\sigma^2 + 2\theta\sigma^2 = \sigma^2(1+\theta)^2$.

3. If we have consistent estimates $\hat{\sigma}^2$, $\hat{\rho}$, $\hat{\theta}$ of σ^2, ρ, θ, we can estimate q_{yy} and q_{zz} consistently by $\frac{\hat{\sigma}^2}{(1-\hat{\rho})^2}$ and $\hat{\sigma}^2(1+\hat{\theta})^2$, respectively. Note that these are positive numbers by construction. Alternatively, we could use robust estimators, like the Newey–West nonparametric estimator, ignoring additional information that we have. But under the circumstances this seems to be less efficient.

4. For vectors, (i) $\sqrt{T}\overline{\mathbf{y}}_T \xrightarrow{d} \mathcal{N}(0, Q_{yy})$, where $Q_{yy} = \sum_{j=-\infty}^{+\infty} \underbrace{\mathbb{C}[\mathbf{y}_t, \mathbf{y}_{t-j}]}_{\Gamma_j}$. But $\Gamma_0 = \sum_{j=0}^{+\infty} \mathrm{P}^j \Sigma \mathrm{P}'^j$, $\Gamma_j = \mathrm{P}^j \Gamma_0$ if $j > 0$, and $\Gamma_j = \Gamma'_{-j} = \Gamma_0 \mathrm{P}'^{|j|}$ if $j < 0$. Hence $Q_{yy} = \Gamma_0 + \sum_{j=1}^{+\infty} \mathrm{P}^j \Gamma_0 + \sum_{j=1}^{+\infty} \Gamma_0 \mathrm{P}'^j = \Gamma_0 + (I-\mathrm{P})^{-1}\mathrm{P}\Gamma_0 + \Gamma_0 \mathrm{P}'(I-\mathrm{P}')^{-1}$; (ii) $\sqrt{T}\overline{\mathbf{z}}_T \xrightarrow{d} \mathcal{N}(0, Q_{zz})$, where $Q_{zz} = \Gamma_0 + \Gamma_1 + \Gamma_{-1} = \Sigma + \Theta\Sigma\Theta' + \Theta\Sigma + \Sigma\Theta' = (I+\Theta)\Sigma(I+\Theta)'$. As for estimation of asymptotic variances, it is evidently possible to construct a consistent estimator of Q_{zz} that is positive definite by construction, but it is not clear if Q_{yy} is positive definite after appropriate estimates of Γ_0 and P are plugged in (constructing a consistent estimator of even Γ_0 is not straightforward). In the latter case it may be better to use the Newey–West estimator.

2. BOOTSTRAP

2.1 Brief and exhaustive

1. The mentioned difference indeed exists, but it is not the principal one. The two methods have some common features like computer simulations, sampling, etc., but they serve completely different goals. The bootstrap is an alternative to analytical asymptotic theory for making inferences, while Monte-Carlo is used for studying small-sample properties of the estimators.

2. After some point raising B does not help since the bootstrap distribution is intrinsically discrete, and raising B cannot smooth things out. Even more than that if we're interested in quantiles, and we usually are: the quantile for a discrete distribution is a whole interval, and the uncertainty about which point to choose to be a quantile doesn't disappear when we raise B.

3. There is no such thing as a "bootstrap estimator". Bootstrapping is a method of inference, not of estimation. The same goes for an "asymptotic estimator".

4. (a) $C_E = [q_n^*(2.5\%), q_n^*(97.5\%)] = [.75, 1.3]$. (b) $C_H = [\hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}(97.5\%), \hat{\theta} - q_{\hat{\theta}^*-\hat{\theta}}(2.5\%)] = [2\hat{\theta} - q_n^*(97.5\%), 2\hat{\theta} - q_n^*(2.5\%)] = [1.1, 1.65]$.
(c) We cannot compute this from the given information, since we need the quantiles of the t-statistic, which are unavailable.

2.2 Bootstrapping t-ratio

The Hall percentile interval is $CI_H = [\hat{\theta} - \tilde{q}^*_{1-\alpha/2}, \hat{\theta} - \tilde{q}^*_{\alpha/2}]$, where $\tilde{q}^*_\alpha$ is the bootstrap α-quantile of $\hat{\theta}^* - \hat{\theta}$, i.e. $\alpha = \mathbb{P}\{\hat{\theta}^* - \hat{\theta} \leq \tilde{q}^*_\alpha\}$. But then $\dfrac{\tilde{q}^*_\alpha}{s(\hat{\theta})}$ is the α-quantile of $\dfrac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta})} = T_n^*$, since $\mathbb{P}\left\{\dfrac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta})} \leq \dfrac{\tilde{q}^*_\alpha}{s(\hat{\theta})}\right\} = \alpha$. But by construction, the α-quantile of T_n^* is q^*_α, hence $\tilde{q}^*_\alpha = s(\hat{\theta})q^*_\alpha$. Substituting this into CI_H, we get the CI as in the problem.

2.3 Bootstrap correcting mean and its square

The bootstrap version $\bar{x}_n^*$ of $\bar{x}_n$ has mean $\bar{x}_n$ with respect to the EDF: $\mathbb{E}^*\left[\bar{x}_n^*\right] = \bar{x}_n$. Thus the bootstrap version of the bias (which is itself zero) is $\text{Bias}^*(\bar{x}_n) = \mathbb{E}^*\left[\bar{x}_n^*\right] - \bar{x}_n = 0$. Therefore, the bootstrap bias corrected estimator of μ is $\bar{x}_n - \text{Bias}^*(\bar{x}_n) = \bar{x}_n$. Now consider the bias of $\bar{x}_n^2$:

$$\text{Bias}(\bar{x}_n^2) = \mathbb{E}\left[\bar{x}_n^2\right] - \mu^2 = \mathbb{V}\left[\bar{x}_n^2\right] = \frac{1}{n}\mathbb{V}\left[x\right].$$

Thus the bootstrap version of the bias is the sample analog of this quantity:

$$\text{Bias}^*(\bar{x}_n^2) = \frac{1}{n}\mathbb{V}^*[x] = \frac{1}{n}\left(\frac{1}{n}\sum x_i^2 - \bar{x}_n^2\right)$$

. Therefore, the bootstrap bias corrected estimator of μ^2 is

$$\bar{x}_n^2 - \text{Bias}^*(\bar{x}_n^2) = \frac{n+1}{n}\bar{x}_n^2 - \frac{1}{n^2}\sum x_i^2.$$

2.4 Bootstrapping conditional mean

We are interested in $g(x) = \mathbb{E}[x'\beta + e|x] = x'\beta$, and as the point estimate we take $\hat{g}(x) = x'\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator for β. To pivotize $\hat{g}(x)$, we observe that

$$x'(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, x\left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1}\mathbb{E}\left[e_i^2 x_i x_i'\right]\left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1} x'),$$

so the appropriate statistic to bootstrap is

$$t_g = \frac{x'(\hat{\beta} - \beta)}{s\left(\hat{g}(x)\right)},$$

where $s\left(\hat{g}(x)\right) = \sqrt{x\left(\sum x_i x_i'\right)^{-1}\left(\sum \hat{e}_i^2 x_i x_i'\right)\left(\sum x_i x_i'\right)^{-1} x'}$. The bootstrap version is

$$t_g^* = \frac{x'(\hat{\beta}^* - \hat{\beta})}{s\left(\hat{g}^*(x)\right)},$$

where $s\left(\hat{g}^*(x)\right) = \sqrt{x\left(\sum x_i^* x_i^{*\prime}\right)^{-1}\left(\sum \hat{e}_i^{*2} x_i^* x_i^{*\prime}\right)\left(\sum x_i^* x_i^{*\prime}\right)^{-1} x'}$. The rest is standard, and the confidence interval is

$$CI_t = \left[x'\hat{\beta} - q^*_{1-\frac{\alpha}{2}} s\left(\hat{g}(x)\right); x'\hat{\beta} - q^*_{\frac{\alpha}{2}} s\left(\hat{g}(x)\right)\right],$$

where $q^*_{\frac{\alpha}{2}}$ and $q^*_{1-\frac{\alpha}{2}}$ are appropriate bootstrap quantiles for t_g^*.

2.5 Bootstrap adjustment for endogeneity?

When we bootstrap an inconsistent estimator, its bootstrap analogs are concentrated more and more around the probability limit of the estimator, and thus the estimate of the bias becomes smaller and smaller as the sample size grows. That is, bootstrapping is able to correct the bias caused by finite sample nonsymmetry of the distribution, but not the asymptotic bias (difference between the probability limit of the estimator and the true parameter value). Rigorously, assume for simplicity that x and β are scalars. Then $\hat{\beta} \xrightarrow{p} \beta + a$, where

$$a = \underset{n\to\infty}{\text{plim}} \frac{\sum_{i=1}^n x_i e_i}{\sum_{i=1}^n x_i^2} = \frac{\mathbb{E}[x_i e_i]}{\mathbb{E}[x_i^2]}.$$

Denote $g\binom{u}{v} = \beta + \frac{u}{v}$, and let G_1 be the vector of the first derivatives of g, and G_2 be the matrix of its second derivatives. Also let us define: $\bar{x} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i e_i \\ \frac{1}{n}\sum_{i=1}^n x_i^2 \end{pmatrix}$ and $\mu = \begin{pmatrix} \mathbb{E}\left[x_i e_i\right] \\ \mathbb{E}\left[x_i^2\right] \end{pmatrix}$. Then

$$\hat{\beta} - (\beta + a) = G_1(\mu)(\bar{x} - \mu) + \frac{1}{2}(\bar{x} - \mu)' G_1(\mu)(\bar{x} - \mu) + R_n,$$

and so

$$\mathbb{E}\left[\hat{\beta} - \beta\right] = a + \frac{1}{2}\mathbb{E}\left[(\bar{x} - \mu)' G_1(\mu)(\bar{x} - \mu)\right] + O\left(\frac{1}{n^2}\right).$$

Thus we see that $\hat{\beta}$ is biased from β by

$$B_n = a + \frac{1}{2}\mathbb{E}\left[(\bar{x} - \mu)' G_1(\mu)(\bar{x} - \mu)\right] + O\left(n^{-2}\right).$$

For the bootstrap analogs,

$$\hat{\beta}^* - \hat{\beta} = G_1(\bar{x})(\bar{x}^* - \bar{x}) + \frac{1}{2}(\bar{x}^* - \bar{x})' G_1(\bar{x})(\bar{x}^* - \bar{x}) + R_n^*,$$

and

$$\mathbb{E}^*\left[\hat{\beta}^* - \hat{\beta}\right] = \frac{1}{2}\mathbb{E}^*\left[(\bar{x}^* - \bar{x})' G_1(\bar{x})(\bar{x}^* - \bar{x})\right] + O\left(\frac{1}{n^2}\right).$$

so the bootstrap bias is

$$B_n^* = \mathbb{E}^*\left[(\bar{x}^* - \bar{x})' G_1(\bar{x})(\bar{x}^* - \bar{x})\right] + O\left(n^{-2}\right).$$

Moreover, $a + \mathbb{E}\left[B_n^*\right] = B_n + O(n^{-2})$. We see that $\mathbb{E}\left[\hat{\beta} - \beta\right] = a + O\left(n^{-1}\right)$ and $\mathbb{E}\left[\hat{\beta} - B_n^*\right] = \beta + B_n - \mathbb{E}\left[B_n^*\right] = \beta + a + O(n^{-2})$, which means that by adjusting the estimator $\hat{\beta}$ by changing it to $\hat{\beta} - B_n^*$, we get rid of the bias due to finiteness of sample, but do not put away the asymptotic bias a.

3. REGRESSION IN GENERAL

3.1 Property of conditional distribution

By definition,

$$|\rho(Y, f(X))| = \frac{|\mathbb{C}[Y, f(X)]|}{\sqrt{\mathbb{V}[Y]}\sqrt{\mathbb{V}[f(X)]}}.$$

Now,

$$\begin{aligned}
|\mathbb{C}[Y, f(X)]| &= |\mathbb{E}[(Y - \mathbb{E}[Y])(f(X) - \mathbb{E}[f(X)])]| = |\mathbb{E}[\mathbb{E}[Y - \mathbb{E}[Y]|X](f(X) - \mathbb{E}[f(X)])]| \\
&= |\mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y])(f(X) - \mathbb{E}[f(X)])]| \\
&\le \mathbb{E}|(\mathbb{E}[Y|X] - \mathbb{E}[Y])(f(X) - \mathbb{E}[f(X)])| \\
&\le \sqrt{\mathbb{E}\left[(\mathbb{E}[Y|X] - \mathbb{E}[Y])^2\right]}\sqrt{\mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right]} = \sqrt{\mathbb{V}[\mathbb{E}[Y|X]]}\sqrt{\mathbb{V}[f(X)]},
\end{aligned}$$

therefore

$$|\rho(Y, f(X))| \le \sqrt{\frac{\mathbb{V}[\mathbb{E}[Y|X]]}{\mathbb{V}[Y]}}.$$

Note that this bound does not depend on $f(X)$. We will now see that $a + b\mathbb{E}(Y|X)$ attains this bound, and therefore is a maximizer. Indeed,

$$\begin{aligned}
|\rho(Y, a + b\mathbb{E}[Y|X])| &= \frac{|\mathbb{C}[Y, a + b\mathbb{E}[Y|X]]|}{\sqrt{\mathbb{V}[Y]}\sqrt{\mathbb{V}[a + b\mathbb{E}[Y|X]]}} = \frac{|b|\,|\mathbb{C}[Y, \mathbb{E}[Y|X]]|}{\sqrt{\mathbb{V}[Y]}\sqrt{b^2\mathbb{V}[\mathbb{E}[Y|X]]}} \\
&= \frac{|\mathbb{E}[(Y - \mathbb{E}[Y])(\mathbb{E}[Y|X] - \mathbb{E}[Y])]|}{\sqrt{\mathbb{V}[Y]}\sqrt{\mathbb{V}[\mathbb{E}[Y|X]]}} \\
&= \frac{|\mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[Y])(\mathbb{E}[Y|X] - \mathbb{E}[Y])]|}{\sqrt{\mathbb{V}[Y]}\sqrt{\mathbb{V}[\mathbb{E}[Y|X]]}} \\
&= \frac{\mathbb{V}[\mathbb{E}[Y|X]]}{\sqrt{\mathbb{V}[Y]}\sqrt{\mathbb{V}[\mathbb{E}[Y|X]]}} = \sqrt{\frac{\mathbb{V}[\mathbb{E}[Y|X]]}{\mathbb{V}[Y]}}.
\end{aligned}$$

3.2 Unobservables among regressors

By the Law of Iterated Expectations, $\mathbb{E}[y|x, z] = \alpha + \beta x + \gamma z$. Thus we know that in the linear prediction $y = \alpha + \beta x + \gamma z + e_y$, the prediction error e_y is uncorrelated with the predictors, i.e. $\mathbb{C}[e_y, x] = \mathbb{C}[e_y, z] = 0$. Consider the linear prediction of z by x: $z = \zeta + \delta x + e_z$, $\mathbb{C}[e_z, x] = 0$. But since $\mathbb{C}[z, x] = 0$, we know that $\delta = 0$. Now, if we linearly predict y only by x, we will have $y = \alpha + \beta x + \gamma(\zeta + e_z) + e_y = \alpha + \gamma\zeta + \beta x + \gamma e_z + e_y$. Here the composite error $\gamma e_z + e_y$ is uncorrelated with x and thus is the best linear prediction error. As a result, the OLS estimator of β is consistent.

Checking the properties of the second option is more involved. Notice that the OLS coefficients in the linear prediction of y by x and w converge in probability to

$$\operatorname{plim}\begin{pmatrix}\hat{\beta}\\ \hat{\omega}\end{pmatrix}=\begin{pmatrix}\sigma_x^2 & \sigma_{xw}\\ \sigma_{xw} & \sigma_w^2\end{pmatrix}^{-1}\begin{pmatrix}\sigma_{xy}\\ \sigma_{wy}\end{pmatrix}=\begin{pmatrix}\sigma_x^2 & \sigma_{xw}\\ \sigma_{xw} & \sigma_w^2\end{pmatrix}^{-1}\begin{pmatrix}\beta\sigma_x^2\\ \beta\sigma_{xw}+\gamma\sigma_{wz}\end{pmatrix},$$

so we can see that

$$\operatorname{plim}\hat{\beta}=\beta+\frac{\sigma_{xw}\sigma_{wz}}{\sigma_x^2\sigma_w^2-\sigma_{xw}^2}\gamma.$$

Thus in general the second option gives an inconsistent estimator.

3.3 Consistency of OLS in presence of lagged dependent variable and serially correlated errors

1. Indeed,

$$\mathbb{E}[u_t]=\mathbb{E}[y_t-\beta y_{t-1}]=\mathbb{E}[y_t]-\beta\mathbb{E}[y_{t-1}]=0-\beta\cdot 0=0$$

and

$$\mathbb{C}\left[u_t,y_{t-1}\right]=\mathbb{C}\left[y_t-\beta y_{t-1},y_{t-1}\right]=\mathbb{C}\left[y_t,y_{t-1}\right]-\beta\mathbb{V}\left[y_{t-1}\right]=0.$$

(ii) Now let us show that $\hat{\beta}$ is consistent. Since $\mathbb{E}[y_t]=0$, it immediately follows that

$$\hat{\beta}=\frac{\frac{1}{T}\sum_{t=2}^{T}y_ty_{t-1}}{\frac{1}{T}\sum_{t=2}^{T}y_{t-1}^2}=\beta+\frac{\frac{1}{T}\sum_{t=2}^{T}u_ty_{t-1}}{\frac{1}{T}\sum_{t=2}^{T}y_{t-1}^2}\xrightarrow{p}\beta+\frac{\mathbb{E}\left[u_ty_{t-1}\right]}{\mathbb{E}\left[y_t^2\right]}=\beta.$$

(iii) To show that u_t is serially correlated, consider

$$\mathbb{C}[u_t,u_{t-1}]=\mathbb{C}\left[y_t-\beta y_{t-1},y_{t-1}-\beta y_{t-2}\right]=\beta\left(\beta\mathbb{C}\left[y_t,y_{t-1}\right]-\mathbb{C}\left[y_t,y_{t-2}\right]\right),$$

which is generally not zero unless $\beta=0$ or $\beta=\dfrac{\mathbb{C}\left[y_t,y_{t-2}\right]}{\mathbb{C}\left[y_t,y_{t-1}\right]}$. As an example of a serially correlated u_t take the AR(2) process

$$y_t=\alpha y_{t-2}+\varepsilon_t,$$

where ε_t are IID. Then $\beta=0$ and thus $u_t=y_t$, serially correlated.

(iv) The OLS estimator is inconsistent if the error term is correlated with the right-hand-side variables. This latter is not necessarily the same as serial correlatedness of the error term.

3.4 Incomplete regression

1. Note that

$$y_i=x_i'\beta+z_i'\gamma+\eta_i.$$

We know that $\mathbb{E}[\eta_i|z_i] = 0$, so $\mathbb{E}[z_i\eta_i] = 0$. However, $\mathbb{E}[z_i\eta_i] \neq 0$ unless $\gamma = 0$, because $0 = \mathbb{E}[x_ie_i] = \mathbb{E}[x_i(z_i'\gamma + \eta_i)] = \mathbb{E}[x_iz_i']\gamma + \mathbb{E}[x_i\eta_i]$, and we know that $\mathbb{E}[x_iz_i'] \neq 0$. The regression of y_i on x_i and z_i yields the OLS estimates with the probability limit

$$p\lim\begin{pmatrix}\hat\beta\\ \hat\gamma\end{pmatrix} = \begin{pmatrix}\beta\\ \gamma\end{pmatrix} + Q^{-1}\begin{pmatrix}\mathbb{E}[x_i\eta_i]\\ 0\end{pmatrix},$$

where

$$Q = \begin{pmatrix}\mathbb{E}[x_ix_i'] & \mathbb{E}[x_iz_i']\\ \mathbb{E}[z_ix_i'] & \mathbb{E}[z_iz_i']\end{pmatrix}.$$

We can see that the estimators $\hat\beta$ and $\hat\gamma$ are in general inconsistent. To be more precise, the inconsistency of *both* $\hat\beta$ and $\hat\gamma$ is proportional to $\mathbb{E}[x_i\eta_i]$, so that unless $\gamma = 0$ (or, more subtly, unless γ lies in the null space of $\mathbb{E}[x_iz_i']$), the estimators are inconsistent.

2. The first step yields a consistent OLS estimate $\hat\beta$ of β because of the OLS estimator is consistent in a linear mean regression. At the second step, we get the OLS estimate

$$\begin{aligned}\hat\gamma &= \left(\sum z_iz_i'\right)^{-1}\sum z_i\hat e_i = \left(\sum z_iz_i'\right)^{-1}\left(\sum z_ie_i - \sum z_ix_i'\left(\hat\beta-\beta\right)\right) = \\ &= \gamma + \left(\frac{1}{n}\sum z_iz_i'\right)^{-1}\left(\frac{1}{n}\sum z_i\eta_i - \frac{1}{n}\sum z_ix_i'\left(\hat\beta-\beta\right)\right).\end{aligned}$$

Since $\frac{1}{n}\sum z_iz_i' \xrightarrow{p} \mathbb{E}[z_iz_i']$, $\frac{1}{n}\sum z_ix_i' \xrightarrow{p} \mathbb{E}[z_ix_i']$, $\frac{1}{n}\sum z_i\eta_i \xrightarrow{p} \mathbb{E}[z_i\eta_i] = 0$, $\hat\beta - \beta \xrightarrow{p} 0$, we have that $\hat\gamma$ is consistent for γ.

Therefore, from the point of view of consistency of $\hat\beta$ and $\hat\gamma$, we recommend the second method. The limiting distribution of $\sqrt{n}(\hat\gamma-\gamma)$ can be deduced by using the Delta-Method. Observe that

$$\sqrt{n}(\hat\gamma-\gamma) = \left(\frac{1}{n}\sum z_iz_i'\right)^{-1}\left(\frac{1}{\sqrt n}\sum z_i\eta_i - \frac{1}{n}\sum z_ix_i'\left(\frac{1}{n}\sum x_ix_i'\right)^{-1}\frac{1}{\sqrt n}\sum x_ie_i\right)$$

and

$$\frac{1}{\sqrt n}\sum\begin{pmatrix}z_i\eta_i\\ x_ie_i\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbb{E}[z_iz_i'\eta_i^2] & \mathbb{E}[z_ix_i'\eta_ie_i]\\ \mathbb{E}[x_iz_i'\eta_ie_i] & \sigma^2\mathbb{E}[x_ix_i']\end{pmatrix}\right).$$

Having applied the Delta-Method and the Continuous Mapping Theorems, we get

$$\sqrt{n}(\hat\gamma-\gamma) \xrightarrow{d} \mathcal{N}\left(0, \left(\mathbb{E}[z_iz_i']\right)^{-1}V\left(\mathbb{E}[z_iz_i']\right)^{-1}\right),$$

where

$$\begin{aligned}V &= \mathbb{E}[z_iz_i'\eta_i^2] + \sigma^2\mathbb{E}[z_ix_i']\left(\mathbb{E}[x_ix_i']\right)^{-1}\mathbb{E}[x_iz_i'] \\ &\quad -\mathbb{E}[z_ix_i']\left(\mathbb{E}[x_ix_i']\right)^{-1}\mathbb{E}[x_iz_i'\eta_ie_i] - \mathbb{E}[z_ix_i'\eta_ie_i]\left(\mathbb{E}[x_ix_i']\right)^{-1}\mathbb{E}[x_iz_i'].\end{aligned}$$

3.5 Brief and exhaustive

1. It simplifies a lot. First, we can use simpler versions of LLNs and CLTs; second, we do not need additional conditions beside existence of some moments. For example, for consistency of the OLS estimator in the linear mean regression model $y_i = x_i\beta + e_i$, $\mathbb{E}[e_i|x_i] = 0$, only existence of moments is needed, while in the case of fixed regressors we (1) have to use the LLN for heterogeneous sequences, (2) have to add the condition $\frac{1}{n}\sum_{i=1}^n x_i^2 \underset{n\to\infty}{\longrightarrow} M$.

2. The economist is probably right about treating the regressors as random if he has a random sampling experiment. But his reasoning is completely ridiculous. For a sampled individual, his/her characteristics (whether true or false) are fixed; randomness arises from the fact that this individual is *randomly selected*.

3. $\hat{E}\left[x|z\right] = g(z)$ is a strictly increasing and continuous function, therefore $g^{-1}(\cdot)$ exists and $\mathbb{E}\left[x|z\right] = \gamma$ is equivalent to $z = g^{-1}(\gamma)$. If $\hat{E}[y|z] = f(z)$, then $\hat{E}\left[y|\mathbb{E}\left[x|z\right] = \gamma\right] = f(g^{-1}(\gamma))$.

4. Yes, one should use White's formula, but not because $\sigma^2 Q_{xx}^{-1}$ does not make sense. It does make sense, but is irrelevant to calculation of the asymptotic variance of the OLS estimator, which in general takes the "sandwich" form. It is not true that σ^2 varies from observation to observation, if by σ^2 we mean unconditional variance of the error term.

4. OLS AND GLS ESTIMATORS

4.1 Brief and exhaustive

1. The OLS estimator is unbiased conditional on all x_i-variables, irrespective of how x_i's are generated. The conditional unbiasedness property implied unbiasedness.

2. Observe that $\mathbb{E}[y|x] = \alpha + \beta x$, $\mathbb{V}[y|x] = (\alpha + \beta x)^2$. Consequently, we can use the usual OLS estimator and White's standard error. By the way, the model $y = (\alpha + \beta x)e$ can be viewed as $y = \alpha + \beta x + u$, where $u = (\alpha + \beta x)(e - 1)$, $\mathbb{E}[u|x] = 0$, $\mathbb{V}[u|x] = (\alpha + \beta x)^2$.

4.2 Estimation of linear combination

1. Consider the class of linear estimators, i.e. one having the form $\tilde{\theta} = \mathcal{A}\mathcal{Y}$, where $\mathcal{A}$ depends only on data $\mathcal{X} = ((1, x_1, z_1)' \cdots (1, x_n, z_n)')'$. The conditional unbiasedness requirement yields the condition $\mathcal{A}\mathcal{X} = (1, c_x, c_y)$, where $\delta = (\alpha, \beta, \gamma)'$. The best linear unbiased estimator is $\hat{\theta} = (1, c_x, c_y)\hat{\delta}$, where $\hat{\delta}$ is the OLS estimator. Indeed, this estimator belongs to the class considered, since $\hat{\theta} = (1, c_x, c_y)(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y} = \mathcal{A}^*\mathcal{Y}$ for $\mathcal{A}^* = (1, c_x, c_y)(\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'$ and $\mathcal{A}^*\mathcal{X} = (1, c_x, c_y)$. Besides,

$$\mathbb{V}\left[\hat{\theta}|\mathcal{X}\right] = \sigma^2(1, c_x, c_y)\left(\mathcal{X}'\mathcal{X}\right)^{-1}(1, c_x, c_y)'$$

and is minimal in the class because the key relationship $(\mathcal{A} - \mathcal{A}^*)\mathcal{A}^* = 0$ holds.

2. Observe that $\sqrt{n}\left(\hat{\theta} - \theta\right) = (1, c_x, c_y)\sqrt{n}\left(\hat{\delta} - \delta\right) \xrightarrow{d} \mathcal{N}\left(0, V_{\hat{\theta}}\right)$, where

$$V_{\hat{\theta}} = \sigma^2\left(1 + \frac{\phi_x^2 + \phi_z^2 - 2\rho\phi_x\phi_z}{1 - \rho^2}\right),$$

$\phi_x = \frac{\mathbb{E}[x] - c_x}{\sqrt{\mathbb{V}[x]}}$, $\phi_z = \frac{\mathbb{E}[z] - c_z}{\sqrt{\mathbb{V}[z]}}$, and ρ is the correlation coefficient between x and z.

3. Minimization of $V_{\hat{\theta}}$ with respect to ρ yields

$$\rho^{opt} = \begin{cases} \dfrac{\phi_x}{\phi_z} & \text{if } \left|\dfrac{\phi_x}{\phi_z}\right| < 1, \\ \dfrac{\phi_z}{\phi_x} & \text{if } \left|\dfrac{\phi_x}{\phi_z}\right| \geq 1. \end{cases}$$

4. Multicollinearity between x and z means that $\rho = 1$ and δ and θ are unidentified. An implication is that the asymptotic variance of $\hat{\theta}$ is infinite.

4.3 Long and short regressions

Let us denote this estimator by $\check{\beta}_1$. We have

$$\begin{aligned}\check{\beta}_1 &= \left(X_1'X_1\right)^{-1} X_1'Y = \left(X_1'X_1\right)^{-1} X_1' \left(X_1\beta_1 + X_2\beta_2 + e\right) = \\ &= \beta_1 + \left(\frac{1}{n}X_1'X_1\right)^{-1}\left(\frac{1}{n}X_1'X_2\right)\beta_2 + \left(\frac{1}{n}X_1'X_1\right)^{-1}\left(\frac{1}{n}X_1'e\right).\end{aligned}$$

Since $\mathbb{E}\left[e_i x_{1i}\right] = 0$, we have that $\frac{1}{n}X_1'e \xrightarrow{p} 0$ by the LLN. Also, by the LLN, $\frac{1}{n}X_1'X_1 \xrightarrow{p} \mathbb{E}\left[x_{1i}x_{1i}'\right]$ and $\frac{1}{n}X_1'X_2 \xrightarrow{p} \mathbb{E}\left[x_{1i}x_{2i}'\right]$. Therefore,

$$\check{\beta}_1 \xrightarrow{p} \beta_1 + \left(\mathbb{E}\left[x_{1i}x_{1i}'\right]\right)^{-1}\mathbb{E}\left[x_{1i}x_{2i}'\right]\beta_2.$$

So, in general, $\check{\beta}_1$ is inconsistent. It will be consistent if β_2 lies in the null space of $\mathbb{E}\left[x_{1i}x_{2i}'\right]$. Two special cases of this are: (1) when $\beta_2 = 0$, i.e. when the true model is $Y = X_1\beta_1 + e$; (2) when $\mathbb{E}\left[x_{1i}x_{2i}'\right] = 0$.

4.4 Ridge regression

1. There is conditional bias: $\mathbb{E}\left[\tilde{\beta}|X\right] = (X'X+\lambda I_k)^{-1}X'\mathbb{E}\left[Y|X\right] = \beta - (X'X+\lambda I_k)^{-1}\lambda\beta$, unless $\beta = 0$. Next $\mathbb{E}\left[\tilde{\beta}\right] = \beta - \mathbb{E}\left[(X'X+\lambda I_k)^{-1}\right]\lambda\beta \neq \beta$ unless $\beta = 0$. Therefore, estimator is in general biased.

2. Observe that

$$\begin{aligned}\tilde{\beta} &= (X'X+\lambda I_k)^{-1}X'X\beta + (X'X+\lambda I_k)^{-1}X'\varepsilon \\ &= \left(\frac{1}{n}\sum_i x_ix_i' + \frac{\lambda}{n}I_k\right)^{-1}\frac{1}{n}\sum_i x_ix_i'\beta + \left(\frac{1}{n}\sum_i x_ix_i' + \frac{\lambda}{n}I_k\right)^{-1}\frac{1}{n}\sum_i x_i\varepsilon_i.\end{aligned}$$

   Since $\frac{1}{n}\sum x_ix_i' \xrightarrow{p} \mathbb{E}\left[x_ix_i'\right]$, $\frac{1}{n}\sum x_i\varepsilon_i \xrightarrow{p} \mathbb{E}\left[x_i\varepsilon_i\right] = 0$, $\frac{\lambda}{n} \xrightarrow{p} 0$, we have:

$$\tilde{\beta} \xrightarrow{p} \left(\mathbb{E}\left[x_ix_i'\right]\right)^{-1}\mathbb{E}\left[x_ix_i'\right]\beta + \left(\mathbb{E}\left[x_ix_i'\right]\right)^{-1}0 = \beta,$$

   that is, $\tilde{\beta}$ is consistent.

3. The math is straightforward:

$$\begin{aligned}\sqrt{n}(\tilde{\beta}-\beta) &= \underbrace{\left(\frac{1}{n}\sum_i x_ix_i' + \frac{\lambda}{n}I_k\right)^{-1}}_{\substack{\downarrow^p \\ \left(\mathbb{E}\left[x_ix_i'\right]\right)^{-1}}}\underbrace{\frac{-\lambda}{\sqrt{n}}}_{\substack{\downarrow \\ 0}}\beta + \underbrace{\left(\frac{1}{n}\sum_i x_ix_i' + \frac{\lambda}{n}I_k\right)^{-1}}_{\substack{\downarrow^p \\ \left(\mathbb{E}\left[x_ix_i'\right]\right)^{-1}}}\underbrace{\frac{1}{\sqrt{n}}\sum_i x_i\varepsilon_i}_{\substack{\downarrow^d \\ \mathcal{N}\left(0,\mathbb{E}\left[x_ix_i'\varepsilon_i^2\right]\right)}} \\ &\xrightarrow{p} \mathcal{N}\left(0, Q_{xx}^{-1}Q_{xxe^2}Q_{xx}^{-1}\right).\end{aligned}$$

4. The conditional mean squared error criterion $\mathbb{E}\left[\left(\tilde{\beta}-\beta\right)^2|X\right]$ can be used. For the OLS estimator,
$$\mathbb{E}\left[\left(\hat{\beta}-\beta\right)^2|X\right]=\mathbb{V}\left[\hat{\beta}\right]=(X'X)^{-1}X'\Omega X(X'X)^{-1}.$$
For the ridge estimator,
$$\mathbb{E}\left[\left(\tilde{\beta}-\beta\right)^2|X\right]=\left(X'X+\lambda I_k\right)^{-1}\left(X'\Omega X+\lambda^2\beta\beta'\right)\left(X'X+\lambda I_k\right)^{-1}$$
By the first order approximation, if λ is small, $(X'X+\lambda I_k)^{-1}\approx(X'X)^{-1}(I_k-\lambda(X'X)^{-1})$. Hence,
$$\begin{aligned}\mathbb{E}\left[\left(\tilde{\beta}-\beta\right)^2|X\right]&\approx(X'X)^{-1}(I-\lambda(X'X)^{-1})(X'\Omega X)(I-\lambda(X'X)^{-1})(X'X)^{-1}\\&\approx\mathbb{E}[(\hat{\beta}-\beta)^2]-\lambda(X'X)^{-1}[X'\Omega X(X'X)^{-1}+(X'X)^{-1}X'\Omega X](X'X)^{-1}.\end{aligned}$$
That is $\mathbb{E}\left[\left(\hat{\beta}-\beta\right)^2|X\right]-\mathbb{E}\left[\left(\tilde{\beta}-\beta\right)^2|X\right]=A$, where A is likely to be positive definite. Thus for small λ, $\tilde{\beta}$ may be a preferable estimator to $\hat{\beta}$ according to the mean squared error criterion, despite its biasedness.

4.5 Exponential heteroskedasticity

1. At the first step, estimate consistently α and β. This can be done using the relationship
$$\mathbb{E}\left[y^2|x\right]=(x'\beta)^2+\exp(x'\beta+\alpha),$$
by using NLLS on this nonlinear mean regression, to get $\hat{\beta}$. Then construct $\hat{\sigma}_i^2\equiv\exp(x_i'\hat{\beta})$ for all i (we don't need $\exp(\alpha)$ since it is just a multiplicative scalar that eventually cancels out) and use these weights at the second step to construct a feasible GLS estimator of β:
$$\tilde{\beta}=\left(\frac{1}{n}\sum_i\hat{\sigma}_i^{-2}x_ix_i'\right)^{-1}\frac{1}{n}\sum_i\hat{\sigma}_i^{-2}x_iy_i.$$

2. The feasible GLS estimator is asymptotically efficient, since it is asymptotically equivalent to GLS. It is finite-sample inefficient, since we changed the weights from what GLS presumes.

4.6 OLS and GLS are identical

1. Evidently, $\mathbb{E}\left[Y|X\right]=X\beta$ and $\Sigma=\mathbb{V}\left[Y|X\right]=X\Gamma X'+\sigma^2I_n$. Since the latter depends on X, we are in the heteroskedastic environment.

2. The OLS estimator is
$$\hat{\beta}=\left(X'X\right)^{-1}X'Y,$$

and the GLS estimator is

$$\tilde{\beta} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1}Y.$$

First, $X'\hat{e} = X'\left(Y - X\left(X'X\right)^{-1}X'Y\right) = X'Y - X'X\left(X'X\right)^{-1}X'Y = X'Y - X'Y = 0$. Premultiply this by $X\Gamma$: $X\Gamma X'\hat{e} = 0$. Add $\sigma^2\hat{e}$ to both sides and combine the terms on the left-hand side: $\left(X\Gamma X' + \sigma^2 I_n\right)\hat{e} \equiv \Sigma\hat{e} = \sigma^2\hat{e}$. Now predividing by matrix Σ gives $\hat{e} = \sigma^2\Sigma^{-1}\hat{e}$. Premultiply once gain by X' to get $0 = X'\hat{e} = \sigma^2 X'\Sigma^{-1}\hat{e}$, or just $X'\Sigma^{-1}\hat{e} = 0$. Recall now what $\hat{e}$ is: $X'\Sigma^{-1}Y = X'\Sigma^{-1}X\left(X'X\right)^{-1}X'Y$ which implies $\hat{\beta} = \tilde{\beta}$.

The fact that the two estimators are identical implies that all the statistics based on the two will be identical and thus have the same distribution.

3. Evidently, in this model the coincidence of the two estimators gives unambiguous superiority of the OLS estimator. In spite of heteroskedasticity, it is efficient in the class of linear unbiased estimators, since it coincides with GLS. The GLS estimator is worse since its feasible version requires estimation of Σ, while the OLS estimator does not. Additional estimation of Σ adds noise which may spoil finite sample performance of the GLS estimator. But all this is not typical for ranking OLS and GLS estimators and is a result of a special form of matrix Σ.

4.7 OLS and GLS are equivalent

1. When $\Sigma X = X\Theta$, we have $X'\Sigma X = X'X\Theta$ and $\Sigma^{-1}X = X\Theta^{-1}$, so that

$$\mathbb{V}\left[\hat{\beta}|X\right] = \left(X'X\right)^{-1} X'\Sigma X\left(X'X\right)^{-1} = \Theta\left(X'X\right)^{-1}$$

and

$$\mathbb{V}\left[\tilde{\beta}|X\right] = \left(X'\Sigma^{-1}X\right)^{-1} = \left(X'X\Theta^{-1}\right)^{-1} = \Theta\left(X'X\right)^{-1}.$$

2. In this example,

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix},$$

and $\Sigma X = \sigma^2(1 + \rho(n-1)) \cdot (1, 1, \cdots, 1)' = X\Theta$, where

$$\Theta = \sigma^2(1 + \rho(n-1)).$$

Thus one does not need to use GLS but instead do OLS to achieve the same finite-sample efficiency.

4.8 Equicorrelated observations

This is essentially a repetition of the second part of the previous problem, from which it follows that under the circumstances $\bar{x}_n$ the best linear conditionally (on a constant which is the same as unconditionally) unbiased estimator of θ because of coincidence of its variance with that of the GLS estimator. Appealing to the case when $|\gamma| > 1$ (which is tempting because then the variance of $\bar{x}_n$ is larger than that of, say, x_1) is invalid, because it is ruled out by the Cauchy-Schwartz inequality.

One cannot appeal to the usual LLNs because x is non-ergodic. The variance of $\bar{x}_n$ is $\mathbb{V}[\bar{x}_n] = \frac{1}{n} \cdot 1 + \frac{n-1}{n} \cdot \gamma \to \gamma$ as $n \to \infty$, so the estimator $\bar{x}_n$ is in general inconsistent (except in the case when $\gamma = 0$). For an example of inconsistent $\bar{x}_n$, assume that $\gamma > 0$ and consider the following construct: $u_i = \varepsilon + \varsigma_i$, where $\varsigma_i \sim IID(0, 1-\gamma)$ and $\varepsilon \sim (0, \gamma)$ independent of ς_i for all i. Then the correlation structure is exactly as in the problem, and $\frac{1}{n}\sum u_i \xrightarrow{p} \varepsilon$, a random nonzero limit.

5. IV AND 2SLS ESTIMATORS

5.1 Instrumental variables in ARMA models

Give brief but exhaustive answers to the following short questions.

1. The instrument x_{t-j} is scalar, the parameter is scalar, so there is exact identification. The instrument is obviously valid. The asymptotic variance of the just identifying IV estimator of a scalar parameter under homoskedasticity is $V_{x_{t-j}} = \sigma^2 Q_{xz}^{-2} Q_{zz}$. Let us calculate all pieces: $Q_{zz} = \mathbb{E}\left[x_{t-j}^2\right] = \mathbb{V}\left[x_t\right] = \sigma^2\left(1-\rho^2\right)^{-1}$; $Q_{xz} = \mathbb{E}\left[x_{t-1}x_{t-j}\right] = \mathbb{C}\left[x_{t-1}, x_{t-j}\right] = \rho^{j-1}\mathbb{V}\left[x_t\right] = \sigma^2\rho^{j-1}\left(1-\rho^2\right)^{-1}$. Thus, $V_{x_{t-j}} = \rho^{2-2j}\left(1-\rho^2\right)$. It is monotonically declining in j, so this suggests that the optimal instrument must be x_{t-1}. Although this is not a proof of the fact, the optimal instrument is indeed $x_{t-1.}$. The result makes sense, since the last observation is most informative and embeds all information in all the other instruments.

2. It is possible to use as instruments lags of y_t starting from y_{t-2} back to the past. The regressor y_{t-1} will not do as it is correlated with the error term through e_{t-1}. Among y_{t-2}, y_{t-3}, $\cdots$ the first one deserves more attention, since, intuitively, it contains more information than older values of y_t.

5.2 Inappropriate 2SLS

1. Since $\mathbb{E}[u]=0$, we have $\mathbb{E}\left[y\right] = \alpha\mathbb{E}\left[z^2\right]$, so α is identified as long as z is not deterministic zero. The analog estimator is

$$\hat{\alpha} = \left(\frac{1}{n}\sum_i z_i^2\right)^{-1}\frac{1}{n}\sum_i y_i.$$

Since $\mathbb{E}[v]=0$, we have $\mathbb{E}[z]=\pi\mathbb{E}[x]$, so π is identified as long as x is not centered around zero. The analog estimator is

$$\hat{\pi} = \left(\frac{1}{n}\sum_i x_i\right)^{-1}\frac{1}{n}\sum_i z_i.$$

Since Σ does not depend on x_i, we have $\Sigma = \mathbb{V}\begin{pmatrix} u_i \\ v_i \end{pmatrix}$, so Σ is identified since both u and v are identified. The analog estimator is

$$\hat{\Sigma} = \frac{1}{n}\sum_i \begin{pmatrix} \hat{u}_i \\ \hat{v}_i \end{pmatrix}\begin{pmatrix} \hat{u}_i \\ \hat{v}_i \end{pmatrix}',$$

where $\hat{u}_i = y_i - \hat{\alpha}z_i^2$ and $\hat{v}_i = z_i - \hat{\pi}x_i$.

2. The estimator satisfies

$$\tilde{\alpha} = \left(\frac{1}{n}\sum_i \hat{z}_i^4\right)^{-1} \frac{1}{n}\sum_i \hat{z}_i^2 y_i = \left(\hat{\pi}^4 \frac{1}{n}\sum_i x_i^4\right)^{-1} \hat{\pi}^2 \frac{1}{n}\sum_i x_i^2 y_i.$$

We know that $\frac{1}{n}\sum_i x_i^4 \xrightarrow{p} \mathbb{E}\left[x^4\right]$, $\frac{1}{n}\sum_i x_i^2 y_i = \alpha\pi^2 \frac{1}{n}\sum_i x_i^4 + 2\alpha\pi \frac{1}{n}\sum_i x_i^3 v_i + \alpha \frac{1}{n}\sum_i x_i^2 v_i^2 + \frac{1}{n}\sum_i x_i^2 u_i \xrightarrow{p} \alpha\pi^2 \mathbb{E}\left[x^4\right] + \alpha\mathbb{E}\left[x^2 v^2\right]$, and $\hat{\pi} \xrightarrow{p} \pi$. Therefore,

$$\tilde{\alpha} \xrightarrow{p} \alpha + \frac{\alpha}{\pi^2} \frac{\mathbb{E}\left[x^2 v^2\right]}{\mathbb{E}\left[x^4\right]} \neq \alpha.$$

3. Evidently, we should fit the estimate of the square of z_i, instead of the square of the estimate. To do this, note that the second equation and properties of the model imply

$$\mathbb{E}\left[z_i^2 | x_i\right] = \mathbb{E}\left[(\pi x_i + v_i)^2 | x_i\right] = \pi^2 x_i^2 + 2\mathbb{E}\left[\pi x_i v_i | x_i\right] + \mathbb{E}\left[v_i^2 | x_i\right] = \pi^2 x_i^2 + \sigma_v^2.$$

That is, we have a linear mean regression of z^2 on x^2 and a constant. Therefore, in the first stage we should regress z^2 on x^2 and a constant and construct $\widehat{z_i^2} = \widehat{\pi^2} x_i^2 + \widehat{\sigma_v^2}$, and in the second stage, we should regress y_i on $\widehat{z_i^2}$. Consistency of this estimator follows from the theory of 2SLS, when we treat z^2 as a right hand side variable, not z.

5.3 Inconsistency under alternative

We are interesting in the question whether the t-statistics can be used to check $H_0 : \beta = 0$. In order to answer this question we have to investigate the asymptotic properties of $\hat{\beta}$. First of all, $\hat{\beta} = \beta + \left(\sum_{i=1}^n z_i^2\right)^{-1} \sum_{i=1}^n z_i \varepsilon_i \to^p 0$, since under the null,

$$\sum_{i=1}^n z_i \varepsilon_i \xrightarrow{p} \mathbb{E}\left[(x+v)(u - \beta v)\right] = -\beta\eta^2 = 0.$$

It is straightforward to show that under the null the conventional standard error correctly estimates (i.e. if correctly normalized, is consistent for) the asymptotic variance of $\hat{\beta}$. That is, under the null, $t_\beta \xrightarrow{d} \mathcal{N}(0,1)$, which means that we can use the conventional t-statistics for testing H_0.

5.4 Trade and growth

1. The economic rationale for uncorrelatedness is that the variables P_i and S_i are exogenous and are unaffected by what's going on in the economy, and on the other hand, hardly can they affect the income in other ways than through the trade. To estimate (5.1), we can use just-identifying IV estimation, where the vector of right-hand-side variables is $x = (1, T, W)'$ and the instrument vector is $z = (1, P, S)'$. (Note: the full answer should include the details of performing the estimation up to getting the standard errors).

2. When data on within-country trade are not available, none of the coefficients in (5.1) is identifiable without further assumptions. In general, neither of the available variables can serve as instruments for T in (5.1) where the composite error term is $\gamma W_i + \varepsilon_i$.

3. We can exploit the assumption that P_i is uncorrelated with the error term in (5.3). Substitute (5.3) into (5.1) to get

$$\log Y_i = (\alpha + \gamma\eta) + \beta T_i + \gamma\lambda S_i + (\gamma\nu_i + \varepsilon_i) .$$

Now we see that S_i and P_i are uncorrelated with the composite error term $\gamma\nu_i + \varepsilon_i$ due to their exogeneity and due to their uncorrelatedness with ν_i which follows from the additional assumption and ν_i being the best linear prediction error in (5.3). (Note: again, the full answer should include the details of performing the estimation up to getting the standard errors, at least). As for the coefficients of (5.1), only β will be consistently estimated, but not α or γ.

4. In general, for this model the OLS is inconsistent, and the IV method is consistent. Thus, the discrepancy may be due to the different probability limits of the two estimators. The fact that the IV estimates are larger says that probably Let $\theta_{IV} \xrightarrow{p} \theta$ and $\theta_{OLS} \xrightarrow{p} \theta + a$, $a < 0$. Then for large samples, $\theta_{IV} \approx \theta$ and $\theta_{OLS} \approx \theta + a$. The difference is a which is $\left(\mathbb{E}\left[xx'\right]\right)^{-1} \mathbb{E}[xe]$. Since $\left(\mathbb{E}\left[xx'\right]\right)^{-1}$ is positive definite, $a < 0$ means that the regressors tend to be negatively correlated with the error term. In the present context this means that the trade variables are negatively correlated with other influences on income.

6. EXTREMUM ESTIMATORS

6.1 Extremum estimators

Part 1. The analog estimator is

$$\hat{\beta} = \arg\max_{b \in B} \frac{1}{n} \sum_{i=1}^{n} f(z_i, b). \tag{6.1}$$

Assume that:

1. $\hat{\beta} \xrightarrow{p} \beta$;
2. β is an interior point of the compact set B (i. e. β is contained in B with its open neighborhood);
3. $f(z, b)$ is a twice continuously differentiable function of b for almost all z;
4. the derivatives $\frac{\partial f(z,b)}{\partial b}$ and $\frac{\partial^2 f(z,b)}{\partial b \partial b'}$ satisfy the ULLN condition in B;
5. $\forall b \in B$ there exist and are finite the moments $\mathbb{E}\left[|f(z,b)|\right]$, $\mathbb{E}\left[\left\|\frac{\partial f(z,b)}{\partial b}\right\|\right]$ and $\mathbb{E}\left[\left\|\frac{\partial^2 f(z,b)}{\partial b \partial b'}\right\|\right]$;
6. there exists finite $\mathbb{E}\left[f_b f_b'\right] \equiv \mathbb{E}\left[\frac{\partial f(z,\beta)}{\partial b}\frac{\partial f(z,\beta)}{\partial b'}\right]$, and matrix $\mathbb{E}\left[f_{bb}\right] \equiv \mathbb{E}\left[\frac{\partial^2 f(z,\beta)}{\partial b \partial b'}\right]$ is non-degenerate.

Then $\hat{\beta}$ is asymptotically normal with asymptotic variance given below in (6.3). To see this write the FOC for an interior solution of problem (6.1) and expand the first derivative around $b = \beta$:

$$0 = \frac{\partial}{\partial b}\left(\frac{1}{n}\sum_{i=1}^{n} f(z_i, \hat{\beta})\right) = \frac{\partial}{\partial b}\left(\frac{1}{n}\sum_{i=1}^{n} f(z_i, \beta)\right) + \frac{\partial^2}{\partial b \partial b'}\left(\frac{1}{n}\sum_{i=1}^{n} f(z_i, \beta^*)\right)(\hat{\beta} - \beta), \tag{6.2}$$

where $\beta_s^* \in [\beta_s, \hat{\beta}_s]$ and β^* may be different in different equations of (6.2). In particular, $\beta^* \xrightarrow{p} \beta$. From here,

$$\sqrt{n}(\hat{\beta} - \beta) = -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 f(z_i, \beta^*)}{\partial b \partial b'}\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial f(z_i, \beta)}{\partial b}.$$

By the ULLN condition for $\frac{\partial^2 f(z,b)}{\partial b \partial b'}$ one has

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 f(z_i, \beta^*)}{\partial b \partial b'} \xrightarrow{p} \mathbb{E}\left[f_{bb}\right].$$

By the CLT, applicable to independent random variables,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial f(z_i, \beta)}{\partial b} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[f_b f_b'\right]\right).$$

The last equality uses the FOC for the extremum problem for β: $\mathbb{E}\left[\frac{\partial f(z_i,\beta)}{\partial b}\right]=0$. Thus,

$$\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{d}\mathcal{N}\left(0,\left(\left(\mathbb{E}\left[f_{bb}\right]\right)^{-1}\mathbb{E}\left[f_b f_b'\right]\left(\mathbb{E}\left[f_{bb}\right]\right)^{-1}\right)^{-1}\right). \tag{6.3}$$

Part 2. Let $\mathbb{E}\left[y|x\right]=g(x,\beta)$, $\sigma^2(x)\equiv\mathbb{E}\left[\left(y-g(x,\beta)\right)^2\right]$. The NLLS and WNLLS estimators are extremum estimators with

$$f_1(x,y,b)=-\frac{1}{2}\left(y-g(x,b)\right)^2 \text{ for NLLS,}$$

and

$$f_2(x,y,b)=-\frac{1}{2}\frac{\left(y-g(x,b)\right)^2}{\sigma^2(x)}=\frac{f_1(x,y,b)}{\sigma^2(x)} \text{ for WNLLS.}$$

It was shown in class that

$$\sqrt{n}(\hat{\beta}_{NLLS}-\beta)\xrightarrow{d}\mathcal{N}\left(0,Q_{gg}^{-1}Q_{gge^2}Q_{gg}^{-1}\right),\quad \sqrt{n}(\hat{\beta}_{WNLLS}-\beta)\xrightarrow{d}\mathcal{N}\left(0,Q_{\frac{gg}{\sigma^2}}\right),$$

where

$$\begin{aligned}
Q_{gg} &= \mathbb{E}\left[g_b(x,\beta)g_b(x,\beta)'\right],\\
Q_{gge^2} &= \mathbb{E}\left[g_b(x,\beta)g_b(x,\beta)'(y-g(x,\beta))^2\right],\\
Q_{\frac{gg}{\sigma^2}} &= \mathbb{E}\left[g_b(x,\beta)g_b(x,\beta)'/\sigma^2(x)\right].
\end{aligned}$$

This coincides with (6.3) since

$$\frac{\partial f_1(x,\beta)}{\partial b}=(y-g(x,\beta))g_b(x,\beta)$$

and

$$\frac{\partial^2 f_1(x,\beta)}{\partial b\partial b'}=(y-g(x,\beta))g_{bb}(x,\beta)-g_b(x,\beta)g_b(x,\beta)',$$

and $\mathbb{E}\left[(y-g(x,\beta))g_{bb}(x,\beta)\right]=\mathbb{E}\left[\mathbb{E}\left[y-g(x,\beta)|x\right]g_{bb}(x,\beta)\right]=0$. In fact, we have $\mathbb{E}\left[f_{1b}f_{1b}'\right]=Q_{gge^2}$, $-\mathbb{E}\left[f_{1bb}\right]=Q_{gg}$, $\mathbb{E}\left[f_{2b}f_{2b}'\right]=-\mathbb{E}\left[f_{2bb}\right]=Q_{\frac{gg}{\sigma^2}}$.

It is worth noting that under the ID-condition for NLLS/WNLLS estimators the solution of corresponding extremum problem in population is unique and equals β.

6.2 Regression on constant

For the first estimator use standard LLN and CLT:

$$\hat{\beta}_1=\frac{1}{n}\sum_{i=1}^n y_i\xrightarrow{p}\mathbb{E}[y_i]=\beta \text{ (consistency),}$$

$$\sqrt{n}(\hat{\beta}_1-\beta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i\xrightarrow{d}\mathcal{N}\left(0,\mathbb{V}[e_i]\right)=\mathcal{N}\left(0,\beta^2\right) \text{ (asymptotic normality).}$$

Consider

$$\hat{\beta}_2=\arg\min_b\left\{\log b^2+\frac{1}{nb^2}\sum_{i=1}^n\left(y_i-b\right)^2\right\}. \tag{6.4}$$

Denote $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$. The FOC for this problem gives after rearrangement:

$$\hat{\beta}^2 + \hat{\beta}\bar{y} - \overline{y^2} = 0 \Leftrightarrow \hat{\beta}_{\pm} = -\frac{\bar{y}}{2} \pm \frac{\sqrt{\bar{y}^2 + 4\overline{y^2}}}{2}.$$

The two values $\hat{\beta}_{\pm}$ correspond to the two different solutions of *local* minimization problem in population:

$$\mathbb{E}\left[\log b^2 + \frac{1}{b^2}(y-b)^2\right] \to \min_{\beta} \Leftrightarrow b_{\pm} = -\frac{\mathbb{E}[y]}{2} \pm \frac{\sqrt{\mathbb{E}[y]^2 + 4\mathbb{E}[y^2]}}{2} = -\frac{\beta}{2} \pm \frac{3|\beta|}{2}. \qquad (6.5)$$

Note that $\hat{\beta}_+ \xrightarrow{p} b_+$ and $\hat{\beta}_- \xrightarrow{p} b_-$. If $\beta > 0$, then $b_+ = \beta$ and the consistent estimate is $\hat{\beta}_2 = \hat{\beta}_+$. If, on the contrary, $\beta < 0$, then $b_- = \beta$ and $\hat{\beta}_2 = \hat{\beta}_-$ is a consistent estimate of β. Alternatively, one can easily prove that the unique global solution of (6.5) is always β. It follows from general theory that the global solution $\hat{\beta}_2$ of (6.4) (which is $\hat{\beta}_+$ or $\hat{\beta}_-$ depending on the sign of $\bar{y}$) is then a consistent estimator of β. The asymptotics of $\hat{\beta}_2$ can be found by formula (6.3). For $f(y,b) = \log b^2 + \frac{1}{b^2}(y-b)^2$,

$$\frac{\partial f(y,b)}{\partial b} = \frac{2}{b} - \frac{2(y-b)^2}{b^3} - \frac{2(y-b)}{b^2} \Rightarrow \mathbb{E}\left[\left(\frac{\partial f(y,\beta)}{\partial b}\right)^2\right] = \frac{4\kappa}{\beta^6},$$

$$\frac{\partial^2 f(y,b)}{\partial b^2} = \frac{6(y-b)^2}{b^4} + \frac{8(y-b)}{b^3} \Rightarrow \mathbb{E}\left[\frac{\partial^2 f(y,\beta)}{\partial b^2}\right] = \frac{6}{\beta^2}.$$

Consequently,

$$\sqrt{n}(\hat{\beta}_2 - \beta) \xrightarrow{d} \mathcal{N}(0, \frac{\kappa}{9\beta^2}).$$

Consider now $\hat{\beta}_3 = \frac{1}{2}\arg\min_b \sum_{i=1}^{n} f(y_i, b)$, where $f(y,b) = \left(\frac{y}{b} - 1\right)^2$. Note that

$$\frac{\partial f(y,b)}{\partial b} = -\frac{2y^2}{b^3} + \frac{2y}{b^2}, \quad \frac{\partial^2 f(y,b)}{\partial b^2} = \frac{6y^2}{b^4} - \frac{4y}{b^3}.$$

The FOC is $\sum_{i=1}^{n} \frac{\partial f(y_i,\hat{b})}{\partial b} = 0 \Leftrightarrow \hat{b} = \frac{\overline{y^2}}{\bar{y}}$ and the estimate is $\hat{\beta}_3 = \frac{\hat{b}}{2} = \frac{1}{2}\frac{\overline{y^2}}{\bar{y}} \xrightarrow{p} \frac{1}{2}\frac{\mathbb{E}[y^2]}{\mathbb{E}[y]} = \beta$. To find the asymptotic variance calculate

$$\mathbb{E}\left[\left(\frac{\partial f(y,2\beta)}{\partial b}\right)^2\right] = \frac{\kappa - \beta^4}{16\beta^6}, \quad \mathbb{E}\left[\frac{\partial^2 f(y,2\beta)}{\partial b^2}\right] = \frac{1}{4\beta^2}.$$

The derivatives are taken at point $b = 2\beta$ because 2β, and not β, is the solution of the extremum problem $\mathbb{E}[f(y,b)] \to \min_b$, which we discussed in part 1. As follows from our discussion,

$$\sqrt{n}(\hat{b} - 2\beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa - \beta^4}{\beta^2}\right) \Leftrightarrow \sqrt{n}(\hat{\beta}_3 - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa - \beta^4}{4\beta^2}\right).$$

A safer way to obtain this asymptotics is probably to change variable in the minimization problem from the beginning: $\hat{\beta}_3 = \arg\min_b \sum_{i=1}^{n}\left(\frac{y}{2b} - 1\right)^2$, and proceed as above.

No one of these estimators is *a priori* asymptotically better than the others. The idea behind these estimators is: $\hat{\beta}_1$ is just the usual OLS estimator, $\hat{\beta}_2$ is the ML estimator for conditional distribution $y|x \sim \mathcal{N}(\beta, \beta^2)$. The third estimator may be thought of as the WNLLS estimator for conditional variance function $\sigma^2(x,b) = b^2$, though it is not completely that (we should divide by $\sigma^2(x,\beta)$ in the WNLLS).

6.3 Quadratic regression

Note that we have conditional homoskedasticity. The regression function is $g(x,\beta) = (\beta + x)^2$. Estimator $\hat{\beta}$ is NLLS, with $\dfrac{\partial g(x,\beta)}{\partial \beta} = 2(\beta + x)$. Then $Q_{xx} = \mathbb{E}\left[\left(\frac{\partial g(x,0)}{\partial\beta}\right)^2\right] = \frac{28}{3}$. Therefore, $\sqrt{n}\hat{\beta} \xrightarrow{d} \mathcal{N}(0, \frac{3}{28}\sigma_0^2)$.

Estimator $\tilde{\beta}$ is an extremum one, with

$$h(x,Y,\beta) = -\frac{Y}{(\beta+x)^2} - \ln[(\beta+x)^2].$$

First we check the ID condition. Indeed,

$$\frac{\partial h(x,Y,\beta)}{\partial\beta} = \frac{2Y}{(\beta+x)^3} - \frac{2}{\beta+x},$$

so the FOC to the population problem is $\mathbb{E}\left[\frac{\partial h(x,Y,\beta)}{\partial\beta}\right] = -2\beta\mathbb{E}\left[\frac{\beta+2x}{(\beta+x)^3}\right]$, which equals zero iff $\beta = 0$. As can be checked, the Hessian is negative on all $\mathcal{B}$, therefore we have a global maximum. Note that the ID condition would not be satisfied if the true parameter was different from zero. Thus, $\tilde{\beta}$ works only for $\beta_0 = 0$.

Next,

$$\frac{\partial^2 h(x,Y,\beta)}{\partial\beta^2} = -\frac{6Y}{(\beta+x)^4} + \frac{2}{(\beta+x)^2}.$$

Then $\Sigma = \mathbb{E}\left[\left(\frac{2Y}{x^3} - \frac{2}{x}\right)^2\right] = \frac{31}{40}\sigma_0^2$ and $\Omega = \mathbb{E}\left[-\frac{6Y}{x^4} + \frac{2}{x^2}\right] = -2$. Therefore, $\sqrt{n}\tilde{\beta} \xrightarrow{d} \mathcal{N}(0, \frac{31}{160}\sigma_0^2)$.

We can see that $\hat{\beta}$ asymptotically dominates $\tilde{\beta}$. In fact, this follows from asymptotic efficiency of NLLS estimator under homoskedasticity (see your homework problem on extremum estimators).

7. MAXIMUM LIKELIHOOD ESTIMATION

7.1 MLE for three distributions

1. For the Pareto distribution with parameter λ the density is

$$f_X(x|\lambda) = \begin{cases} \lambda x^{-(\lambda+1)}, & \text{if } x > 1, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore the likelihood function is $L = \lambda^n \prod_{i=1}^n x_i^{-(\lambda+1)}$ and the loglikelihood is $\ell_n = n \ln \lambda - (\lambda + 1) \sum_{i=1}^n \ln x_i$

(i) The ML estimator $\hat{\lambda}$ of λ is the solution of $\partial \ell_n / \partial \lambda = 0$. That is, $\hat{\lambda}_{ML} = 1/\overline{\ln x}$, which is consistent for λ, since $1/\overline{\ln x} \xrightarrow{p} 1/\mathbb{E}[\ln x] = \lambda$. The asymptotic distribution is $\sqrt{n}\left(\hat{\lambda}_{ML} - \lambda\right) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}\right)$, where the information matrix is $I = -\mathbb{E}[\partial s/\partial \lambda] = -\mathbb{E}\left[-1/\lambda^2\right] = 1/\lambda^2$

(ii) The Wald test for a simple hypothesis is

$$\mathcal{W} = n(\hat{\lambda} - \lambda)' I(\hat{\lambda})(\hat{\lambda} - \lambda) = n \frac{(\hat{\lambda} - \lambda_0)^2}{\hat{\lambda}^2} \xrightarrow{d} \chi^2(1)$$

The Likelihood Ratio test statistic for a simple hypothesis is

$$\begin{aligned} \mathcal{LR} &= 2\left[\ell_n(\hat{\lambda}) - \ell_n(\lambda_0)\right] \\ &= 2\left[n \ln \hat{\lambda} - (\hat{\lambda} + 1) \sum_{i=1}^n \ln x_i - \left(n \ln \lambda_0 - (\lambda_0 + 1) \sum_{i=1}^n \ln x_i\right)\right] \\ &= 2\left[n \ln \frac{\hat{\lambda}}{\lambda_0} - (\hat{\lambda} - \lambda_0) \sum_{i=1}^n \ln x_i\right] \xrightarrow{d} \chi^2(1). \end{aligned}$$

The Lagrange Multiplier test statistic for a simple hypothesis is

$$\begin{aligned} \mathcal{LM} &= \frac{1}{n} \sum_{i=1}^n s(x_i, \lambda_0)' I(\lambda_0)^{-1} \sum_{i=1}^n s(x_i, \lambda_0) = \frac{1}{n} \left[\sum_{i=1}^n \left(\frac{1}{\lambda_0} - \ln x_i\right)\right]^2 \lambda_0^2 \\ &= n \frac{(\hat{\lambda} - \lambda_0)^2}{\hat{\lambda}^2} \xrightarrow{d} \chi^2(1). \end{aligned}$$

$\mathcal{W}$ and $\mathcal{LM}$ are numerically equal.

2. Since $x_1, \cdots, x_n$ are from $\mathcal{N}(\mu, \mu^2)$, the loglikelihood function is

$$\ell_n = \text{const} - n \ln |\mu| - \frac{1}{2\mu^2} \sum_{i=1}^n (x_i - \mu)^2 = \text{const} - n \ln |\mu| - \frac{1}{2\mu^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2\right).$$

The equation for the ML estimator is $\mu^2 + \bar{x}\mu - \overline{x^2} = 0$. The equation has two solutions $\mu_1 > 0$, $\mu_2 < 0$:

$$\mu_1 = \frac{1}{2}\left(-\bar{x} + \sqrt{\bar{x}^2 + 4\overline{x^2}}\right), \quad \mu_2 = \frac{1}{2}\left(-\bar{x} - \sqrt{\bar{x}^2 + 4\overline{x^2}}\right).$$

Note that ℓ_n is a symmetric function of μ except for the term $\frac{1}{\mu}\sum_{i=1}^n x_i$. This term determines the solution. If $\bar{x} > 0$ then the global maximum of ℓ_n will be in μ_1, otherwise in μ_2. That is, the ML estimator is

$$\hat{\mu}_{ML} = \frac{1}{2}\left(-\bar{x} + \operatorname{sgn}(\bar{x})\sqrt{\bar{x}^2 + 4\overline{x^2}}\right).$$

It is consistent because, if $\mu \neq 0$, $\operatorname{sgn}(\bar{x}) \xrightarrow{p} \operatorname{sgn}(\mu)$ and

$$\hat{\mu}_{ML} \xrightarrow{p} \frac{1}{2}\left(-\mathbb{E}x + \operatorname{sgn}(\mathbb{E}x)\sqrt{(\mathbb{E}x)^2 + 4\mathbb{E}x^2}\right) = \frac{1}{2}\left(-\mu + \operatorname{sgn}(\mu)\sqrt{\mu^2 + 8\mu^2}\right) = \mu.$$

3. We derived in class that the maximum likelihood estimator of θ is

$$\hat{\theta}_{ML} = x_{(n)} \equiv \max\{x_1, \cdots, x_n\}$$

and its asymptotic distribution is exponential:

$$F_{n(\hat{\theta}_{ML}-\theta)}(t) \to \exp(t/\theta) \cdot \mathbb{I}[t \le 0] + \mathbb{I}[t > 0].$$

The most elegant way to proceed is by pivotizing this distribution first:

$$F_{n(\hat{\theta}_{ML}-\theta)/\theta}(t) \to \exp(t) \cdot \mathbb{I}[t \le 0] + \mathbb{I}[t > 0].$$

The left 5%-quantile for the limiting distribution is $\log(.05)$. Thus, with probability 95%, $\log(.05) \le n(\hat{\theta}_{ML} - \theta)/\theta \le 0$, so the confidence interval for θ is

$$\left[x_{(n)}, x_{(n)}/(1 + \log(.05)/n)\right].$$

7.2 Comparison of ML tests

1. Recall that for the ML estimator $\hat{\lambda}$ and the simple hypothesis $H_0 : \lambda = \lambda_0$,

$$\mathcal{W} = n(\hat{\lambda} - \lambda_0)' I(\hat{\lambda})(\hat{\lambda} - \lambda_0),$$

$$\mathcal{LM} = \frac{1}{n}\sum_i s(x_i, \lambda_0)' I(\lambda_0)^{-1} \sum_i s(x_i, \lambda_0).$$

2. The density of a Poisson distribution with parameter λ is

$$f(x_i|\lambda) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda},$$

so $\hat{\lambda}_{ML} = \bar{x}$, $I(\lambda) = 1/\lambda$. For the simple hypothesis with $\lambda_0 = 3$ the test statistics are

$$\mathcal{W} = \frac{n(\bar{x} - 3)^2}{\bar{x}}, \quad \mathcal{LM} = \frac{1}{n}\left(\sum x_i/3 - n\right)^2 3 = \frac{n(\bar{x} - 3)^2}{3},$$

and $\mathcal{W} \ge \mathcal{LM}$ for $\bar{x} \le 3$ and $\mathcal{W} \le \mathcal{LM}$ for $\bar{x} \ge 3$.

3. The density of an exponential distribution with parameter θ is

$$f(x_i) = \frac{1}{\theta} e^{-\frac{x_i}{\theta}},$$

so $\hat\theta_{ML} = \bar x$, $I(\theta) = 1/\theta^2$. For the simple hypothesis with $\theta_0 = 3$ the test statistics are

$$\mathcal{W} = \frac{n(\bar x - 3)^2}{\bar x^2}, \quad \mathcal{LM} = \frac{1}{n}\left(\sum_i \frac{x_i}{3^2} - \frac{n}{3}\right)^2 3^2 = \frac{n(\bar x - 3)^2}{9},$$

and $\mathcal{W} \geq \mathcal{LM}$ for $0 < \bar x \leq 3$ and $\mathcal{W} \leq \mathcal{LM}$ for $\bar x \geq 3$.

4. The density of a Bernoulli distribution with parameter θ is

$$f(x_i) = \theta^{x_i}(1-\theta)^{1-x_i},$$

so $\hat\theta_{ML} = \bar x$, $I(\theta) = \frac{1}{\theta(1-\theta)}$. For the simple hypothesis with $\theta_0 = \frac{1}{2}$ the test statistics are

$$\mathcal{W} = n\frac{\left(\bar x - \frac{1}{2}\right)^2}{\bar x(1-\bar x)}, \quad \mathcal{LM} = \frac{1}{n}\left(\frac{\sum_i x_i}{\frac{1}{2}} - \frac{n - \sum_i x_i}{\frac{1}{2}}\right)^2 \frac{1}{2}\frac{1}{2} = 4n\left(\bar x - \frac{1}{2}\right)^2,$$

and $\mathcal{W} \geq \mathcal{LM}$ (since $\bar x(1-\bar x) \leq 1/4$). For the simple hypothesis with $\theta_0 = \frac{2}{3}$ the test statistics are

$$\mathcal{W} = n\frac{\left(\bar x - \frac{2}{3}\right)^2}{\bar x(1-\bar x)}, \quad \mathcal{LM} = \frac{1}{n}\left(\frac{\sum_i x_i}{\frac{2}{3}} - \frac{n - \sum_i x_i}{\frac{1}{3}}\right)^2 \frac{2}{3}\frac{1}{3} = \frac{9}{2}n\left(\bar x - \frac{2}{3}\right)^2,$$

therefore $\mathcal{W} \leq \mathcal{LM}$ when $2/9 \leq \bar x(1-\bar x)$ and $\mathcal{W} \geq \mathcal{LM}$ when $2/9 \geq \bar x(1-\bar x)$. Equivalently, $\mathcal{W} \leq \mathcal{LM}$ for $\frac{1}{3} \leq \bar x \leq \frac{2}{3}$ and $\mathcal{W} \geq \mathcal{LM}$ for $0 < \bar x \leq \frac{1}{3}$ or $\frac{2}{3} \leq \bar x \leq 1$.

7.3 Individual effects

The loglikelihood is

$$\ell_n\left(\mu_1, \cdots, \mu_n, \sigma^2\right) = \text{const} - n\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left\{(x_i - \mu_i)^2 + (y_i - \mu_i)^2\right\}.$$

FOC give

$$\hat\mu_{iML} = \frac{x_i + y_i}{2}, \quad \sigma^2_{ML} = \frac{1}{2n}\sum_{i=1}^n \left\{(x_i - \hat\mu_{iML})^2 + (y_i - \hat\mu_{iML})^2\right\},$$

so that

$$\hat\sigma^2_{ML} = \frac{1}{4n}\sum_{i=1}^n (x_i - y_i)^2.$$

Since $\hat\sigma^2_{ML} = \frac{1}{4n}\sum_{i=1}^n \left\{(x_i - \mu_i)^2 + (y_i - \mu_i)^2 - 2(x_i - \mu_i)(y_i - \mu_i)\right\} \xrightarrow{p} \frac{\sigma^2}{4} + \frac{\sigma^2}{4} - 0 = \frac{\sigma^2}{2}$, the ML estimator is inconsistent. Why? The Maximum Likelihood method (and all others that we are studying) presumes a parameter vector of fixed dimension. In our case the dimension instead increases with an increase in the number of observations. Information from new observations goes to estimation of new parameters instead of increasing precision of the old ones. To construct a consistent estimator, just multiply $\hat\sigma^2_{ML}$ by 2. There are also other possibilities.

7.4 Does the link matter?

Let the x variable assume two different values x^0 and x^1, $u^a = \alpha+\beta x^a$ and $n_{ab} = \#\{x_i = x^a, y_i = b\}$, for $a, b = 0, 1$ (i.e., $n_{a,b}$ is the number of observations for which $x_i = x^a, y_i = b$). The log-likelihood function is

$$\begin{aligned} l(x_1,..x_n, y_1, ..., y_n; \alpha, \beta) &= \log \left[\prod_{i=1}^n F(\alpha+\beta x_i)^{y_i}(1-F(\alpha+\beta x_i))^{1-y_i}\right] = \\ &= n_{01}\log F(u^0) + n_{00}\log(1-F(u^0)) + n_{11}\log F(u^1) + n_{10}\log(1-F(u^1)). \end{aligned} \tag{7.1}$$

The FOC for the problem of maximization of $l(...; \alpha, \beta)$ w.r.t. α and β are:

$$\begin{aligned} \left[n_{01}\frac{F'(\hat u^0)}{F(\hat u^0)} - n_{00}\frac{F'(\hat u^0)}{1-F(\hat u^0)}\right] + \left[n_{11}\frac{F'(\hat u^1)}{F(\hat u^1)} - n_{10}\frac{F'(\hat u^1)}{1-F(\hat u^1)}\right] &= 0, \\ x^0\left[n_{01}\frac{F'(\hat u^0)}{F(\hat u^0)} - n_{00}\frac{F'(\hat u^0)}{1-F(\hat u^0)}\right] + x^1\left[n_{11}\frac{F'(\hat u^1)}{F(\hat u^1)} - n_{10}\frac{F'(\hat u^1)}{1-F(\hat u^1)}\right] &= 0 \end{aligned}$$

As $x^0 \neq x^1$, one obtains for $a = 0, 1$

$$\frac{n_{a1}}{F(\hat u^a)} - \frac{n_{a0}}{1-F(\hat u^a)} = 0 \Leftrightarrow F(\hat u^a) = \frac{n_{a1}}{n_{a1}+n_{a0}} \Leftrightarrow \hat u^a \equiv \hat\alpha + \hat\beta x^a = F^{-1}\left(\frac{n_{a1}}{n_{a1}+n_{a0}}\right) \tag{7.2}$$

under the assumption that $F'(\hat u^a) \neq 0$. Comparing (7.1) and (7.2) one sees that $l(..., \hat\alpha, \hat\beta)$ does not depend on the form of the link function $F(\cdot)$. The estimates $\hat\alpha$ and $\hat\beta$ can be found from (7.2):

$$\hat\alpha = \frac{x^1F^{-1}\left(\frac{n_{01}}{n_{01}+n_{00}}\right) - x^0F^{-1}\left(\frac{n_{11}}{n_{11}+n_{10}}\right)}{x^1-x^0}, \quad \hat\beta = \frac{F^{-1}\left(\frac{n_{11}}{n_{11}+n_{10}}\right) - F^{-1}\left(\frac{n_{01}}{n_{01}+n_{00}}\right)}{x^1-x^0}.$$

7.5 Nuisance parameter in density

The FOC for the second stage of estimation is

$$\frac{1}{n}\sum_{i=1}^n s_c(y_i, x_i, \tilde\gamma, \hat\delta_m) = 0,$$

where $s_c(y, x, \gamma, \delta) \equiv \dfrac{\partial \log f_c(y|x, \gamma, \delta)}{\partial\gamma}$ is the conditional score. Taylor's expansion with respect to the γ-argument around γ_0 yields

$$\frac{1}{n}\sum_{i=1}^n s_c(y_i, x_i, \gamma_0, \hat\delta_m) + \frac{1}{n}\sum_{i=1}^n \frac{\partial s_c(y_i, x_i, \gamma^*, \hat\delta_m)}{\partial\gamma'}(\tilde\gamma-\gamma_0) = 0,$$

where γ^* lies between $\tilde\gamma$ and γ_0 componentwise.

Now Taylor-expand the first term around δ_0:

$$\frac{1}{n}\sum_{i=1}^n s_c(y_i, x_i, \gamma_0, \hat\delta_m) = \frac{1}{n}\sum_{i=1}^n s_c(y_i, x_i, \gamma_0, \delta_0) + \frac{1}{n}\sum_{i=1}^n \frac{\partial s_c(y_i, x_i, \gamma_0, \delta^*)}{\partial\delta'}(\hat\delta_m - \delta_0),$$

where δ^* lies between $\hat\delta_m$ and δ_0 componentwise.

Combining the two pieces, we get:

$$\begin{aligned}\sqrt{n}(\tilde{\gamma}-\gamma_0) &= -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial s_c(y_i,x_i,\gamma^*,\hat{\delta}_m)}{\partial\gamma'}\right)^{-1}\times \\ &\quad\times\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_c(y_i,x_i,\gamma_0,\delta_0)+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial s_c(y_i,x_i,\gamma_0,\delta^*)}{\partial\delta'}\sqrt{n}(\hat{\delta}_m-\delta_0)\right).\end{aligned}$$

Now let $n\to\infty$. Under ULLN for the second derivative of the log of the conditional density, the first factor converges in probability to $-(I_c^{\gamma\gamma})^{-1}$, where $I_c^{\gamma\gamma}\equiv-\mathbb{E}\left[\dfrac{\partial^2\log f_c(y|x,\gamma_0,\delta_0)}{\partial\gamma\partial\gamma'}\right]$. There are two terms inside the brackets that have nontrivial distributions. We will compute asymptotic variance of each and asymptotic covariance between them. The first term behaves as follows:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_c(y_i,x_i,\gamma_0,\delta_0)\xrightarrow{d}\mathcal{N}\left(0,I_c^{\gamma\gamma}\right)$$

due to the CLT (recall that the score has zero expectation and the information matrix equality). Turn to the second term. Under the ULLN, $\frac{1}{n}\sum_{i=1}^{n}\dfrac{\partial s_c(y_i,x_i,\gamma_0,\delta^*)}{\partial\delta'}$ converges to $-I_c^{\gamma\delta}=\mathbb{E}\left[\dfrac{\partial^2\log f_c(y|x,\gamma_0,\delta_0)}{\partial\gamma\partial\delta'}\right]$. Next, we know from the MLE theory that $\sqrt{n}(\hat{\delta}_m-\delta_0)\xrightarrow{d}\mathcal{N}\left(0,(I_m^{\delta\delta})^{-1}\right)$, where $I_m^{\delta\delta}\equiv-\mathbb{E}\left[\dfrac{\partial^2\log f_m(x|\delta_0)}{\partial\delta\partial\delta'}\right]$. Finally, the asymptotic covariance term is zero because of the "marginal"/"conditional" relationship between the two terms, the Law of Iterated Expectations and zero expected score.

Collecting the pieces, we find:

$$\sqrt{n}(\tilde{\gamma}-\gamma_0)\xrightarrow{d}\mathcal{N}\left(0,(I_c^{\gamma\gamma})^{-1}\left(I_c^{\gamma\gamma}+I_c^{\gamma\delta}(I_m^{\delta\delta})^{-1}I_c^{\gamma\delta\prime}\right)(I_c^{\gamma\gamma})^{-1}\right).$$

It is easy to see that the asymptotic variance is larger (in matrix sense) than $(I_c^{\gamma\gamma})^{-1}$ that would be the asymptotic variance if we new the nuisance parameter δ_0. But it is impossible to compare to the asymptotic variance for $\hat{\gamma}_c$, which is *not* $(I_c^{\gamma\gamma})^{-1}$.

7.6 MLE versus OLS

1. $\hat{\alpha}_{OLS}=\frac{1}{n}\sum_{i=1}^{n}y_i$, $\mathbb{E}\left[\hat{\alpha}_{OLS}\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[y\right]=\alpha$, so $\hat{\alpha}_{OLS}$ is unbiased. Next, $\frac{1}{n}\sum_{i=1}^{n}y_i\xrightarrow{p}\mathbb{E}\left[y\right]=\alpha$, so $\hat{\alpha}_{OLS}$ is consistent. Yes, as we know from the theory, $\hat{\alpha}_{OLS}$ is the best linear unbiased estimator. Note that the members of this class are allowed to be of the form $\{\mathcal{A}Y,\mathcal{A}X=I\}$, where $\mathcal{A}$ is a *constant* matrix, since there are no regressors beside the constant. There is no heteroskedasticity, since there are no regressors to condition on (more precisely, we should condition on a constant, i.e. the trivial σ-field, which gives just an unconditional variance which is constant by the IID assumption). The asymptotic distribution is

$$\sqrt{n}(\hat{\alpha}_{OLS}-\alpha)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}e_i\xrightarrow{d}N\left(0,\sigma^2\mathbb{E}\left[x^2\right]\right),$$

since the variance of e_i is $\mathbb{E}\left[e^2\right]=\mathbb{E}\left[\mathbb{E}\left[e^2|x\right]\right]=\sigma^2\mathbb{E}\left[x^2\right].$

2. The conditional likelihood function is

$$\mathcal{L}(y_1,...,y_n,x_1,...,x_n,\alpha,\sigma^2)=\prod_{i=1}^n \frac{1}{\sqrt{2\pi x_i^2\sigma^2}}\exp\left\{-\frac{(y_i-\alpha)^2}{2x_i^2\sigma^2}\right\}.$$

The conditional loglikelihood is

$$\ell_n(y_1,...,y_n,x_1,...,x_n,\alpha,\sigma^2)=\text{const}-\sum_{i=1}^n \frac{(y_i-\alpha)^2}{2x_i^2\sigma^2}-\frac{1}{2}\log\sigma^2 \to \max_{\alpha,\sigma^2}.$$

From the first order condition $\frac{\partial \ell_n}{\partial\alpha}=\sum_{i=1}^n \frac{y_i-\alpha}{x_i^2\sigma^2}=0$, the ML estimator is

$$\hat{\alpha}_{ML}=\frac{\sum_{i=1}^n y_i/x_i^2}{\sum_{i=1}^n 1/x_i^2}.$$

Note: it as equal to the OLS estimator in

$$\frac{y_i}{x_i}=\alpha\frac{1}{x_i}+\frac{e_i}{x_i}.$$

The asymptotic distribution is

$$\sqrt{n}(\hat{\alpha}_{ML}-\alpha)=\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i/x_i^2}{\frac{1}{n}\sum_{i=1}^n 1/x_i^2}\xrightarrow{d}\left(\mathbb{E}\left[\frac{1}{x^2}\right]\right)^{-1}\mathcal{N}\left(0,\sigma^2\mathbb{E}\left[\frac{1}{x^2}\right]\right)=\mathcal{N}\left(0,\sigma^2\left(\mathbb{E}\left[\frac{1}{x^2}\right]\right)^{-1}\right).$$

Note that $\hat{\alpha}_{ML}$ is unbiased and more efficient than $\hat{\alpha}_{OLS}$ since

$$\left(\mathbb{E}\left[\frac{1}{x^2}\right]\right)^{-1}<\mathbb{E}\left[x^2\right],$$

but it is not in the class of linear unbiased estimators, since the weights in $\mathcal{A}_{ML}$ depend on extraneous x_i's. The $\hat{\alpha}_{ML}$ is efficient in a much larger class. Thus there is no contradiction.

7.7 MLE in heteroskedastic time series regression

Since the parameter v is never involved in the conditional distribution $y_t|x_t$, it can be efficiently estimated from the marginal distribution of x_t, which yields

$$\hat{v}=\frac{1}{T}\sum_{t=1}^T x_t^2.$$

If x_t is serially uncorrelated, then x_t is IID due to normality, so $\hat{v}$ is a ML estimator. If x_t is serially correlated, a ML estimator is unavailable due to lack of information, but $\hat{v}$ still consistently estimates v. The standard error may be constructed via

$$\hat{V}=\frac{1}{T}\sum_{t=1}^T x_t^4-\hat{v}^2$$

if x_t is serially uncorrelated, and via a corresponding Newey-West estimator if x_t is serially correlated.

1. If the entire function $\sigma_t^2 = \sigma^2(x_t)$ is fully known, the conditional ML estimator of α and β is the same as the GLS estimator:

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}_{ML} = \left(\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\right)^{-1}\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\begin{pmatrix}1\\ x_t\end{pmatrix}y_t.$$

The standard errors may be constructed via

$$\hat{V}_{ML} = T\left(\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\right)^{-1}.$$

2. If the values of σ_t^2 at $t = 1, 2, \cdots, T$ are known, we can use the same procedure as in Part 1, since it does not use values of $\sigma^2(x_t)$ other than those at $x_1, x_2, \cdots, x_T$.

3. If it is known that $\sigma_t^2 = (\theta + \delta x_t)^2$, we have in addition parameters θ and δ to be estimated jointly from the conditional distribution

$$y_t|x_t \sim \mathcal{N}(\alpha + \beta x_t, (\theta + \delta x_t)^2).$$

The loglikelihood function is

$$\ell_n(\alpha, \beta, \theta, \delta) = \text{const} - \frac{n}{2}\log(\theta + \delta x_t)^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{(y_t - \alpha - \beta x_t)^2}{(\theta + \delta x_t)^2},$$

and $\left(\hat{\alpha}\ \hat{\beta}\ \hat{\theta}\ \hat{\delta}\right)'_{ML} = \underset{(\alpha,\beta,\theta,\delta)}{\arg\max}\ \ell_n(\alpha, \beta, \theta, \delta)$. Note that

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}_{ML} = \left(\sum_{t=1}^{T}\frac{1}{(\hat{\theta} + \hat{\delta}x_t)^2}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\right)^{-1}\sum_{t=1}^{T}\frac{y_t}{(\hat{\theta} + \hat{\delta}x_t)^2}\begin{pmatrix}1\\ x_t\end{pmatrix},$$

i. e. the ML estimator of α and β is a feasible GLS estimator that uses $\left(\hat{\theta}\ \hat{\delta}\right)'_{ML}$ as the preliminary estimator. The standard errors may be constructed via

$$\hat{V}_{ML} = T\left(\sum_{t=1}^{T}\frac{\partial\ell_n\left(\hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\delta}\right)}{\partial(\alpha, \beta, \theta, \delta)'}\frac{\partial\ell_n\left(\hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\delta}\right)}{\partial(\alpha, \beta, \theta, \delta)}\right)^{-1}.$$

4. Similarly to Part 1, if it is known that $\sigma_t^2 = \theta + \delta u_{t-1}^2$, we have in addition parameters θ and δ to be estimated jointly from the conditional distribution

$$y_t|x_t, y_{t-1}, x_{t-1} \sim \mathcal{N}\left(\alpha + \beta x_t, \theta + \delta(y_{t-1} - \alpha - \beta x_{t-1})^2\right).$$

5. If it is only known that σ_t^2 is stationary, conditional maximum likelihood function is unavailable, so we have to use subefficient methods, for example, OLS estimation

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}_{OLS} = \left(\sum_{t=1}^{T}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\right)^{-1}\sum_{t=1}^{T}\begin{pmatrix}1\\ x_t\end{pmatrix}y_t.$$

The standard errors may be constructed via

$$\hat{V}_{OLS} = T\left(\sum_{t=1}^{T}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\right)^{-1}\cdot\sum_{t=1}^{T}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\hat{e}_t^2\cdot\left(\sum_{t=1}^{T}\begin{pmatrix}1 & x_t\\ x_t & x_t^2\end{pmatrix}\right)^{-1},$$

where $\hat{e}_t = y_t - \hat{\alpha}_{OLS} - \hat{\beta}_{OLS}x_t$. Alternatively, one may use a feasible GLS estimator after having assumed a form of the skedastic function $\sigma^2(x_t)$ and standard errors *robust to its misspecification*.

7.8 Maximum likelihood and binary variables

1. Since the parameters in the conditional and marginal densities do not overlap, we can separate the problem. The conditional likelihood function is

$$\mathcal{L}(y_1,...,y_n,z_1,...,z_n,\gamma)=\prod_{i=1}^{n}\left(\frac{e^{\gamma z_i}}{1+e^{\gamma z_i}}\right)^{y_i}\left(1-\frac{e^{\gamma z_i}}{1+e^{\gamma z_i}}\right)^{1-y_i},$$

and the conditional loglikelihood –

$$\ell_n(y_1,...,y_n,z_1,...,z_n,\gamma)=\sum_{i=1}^{n}\left[y_i\gamma z_i-\ln(1+e^{\gamma z_i})\right]$$

The first order condition

$$\frac{\partial \ell_n}{\partial \gamma}=\sum_{i=1}^{n}\left[y_i z_i-\frac{z_i e^{\gamma z_i}}{1+e^{\gamma z_i}}\right]=0$$

gives the solution $\hat{\gamma}=\log\frac{n_{11}}{n_{10}}$, where $n_{11}=\#\{z_i=1,y_i=1\}$, $n_{10}=\#\{z_i=1,y_i=0\}$. The marginal likelihood function is

$$\mathcal{L}(z_1,...,z_n,\alpha)=\prod_{i=1}^{n}\alpha^{z_i}(1-\alpha)^{1-z_i},$$

and the marginal loglikelihood –

$$\ell_n(z_1,...,z_n,\alpha)=\sum_{i=1}^{n}\left[z_i\ln\alpha+(1-z_i)\ln(1-\alpha)\right]$$

The first order condition

$$\frac{\partial \ell_n}{\partial \alpha}=\frac{\sum_{i=1}^{n}z_i}{\alpha}-\frac{\sum_{i=1}^{n}(1-z_i)}{1-\alpha}=0$$

gives the solution $\hat{\alpha}=\frac{1}{n}\sum_{i=1}^{n}z_i$. From the asymptotic theory for ML,

$$\sqrt{n}\left(\begin{pmatrix}\hat{\alpha}\\ \hat{\gamma}\end{pmatrix}-\begin{pmatrix}\alpha\\ \gamma\end{pmatrix}\right)\xrightarrow{d}\mathcal{N}\left(0,\begin{pmatrix}\alpha(1-\alpha) & 0\\ 0 & \dfrac{(1+e^{\gamma})^2}{\alpha e^{\gamma}}\end{pmatrix}\right).$$

2. The test statistic is

$$t=\frac{\hat{\alpha}-\hat{\gamma}}{s(\hat{\alpha}-\hat{\gamma})}\xrightarrow{d}\mathcal{N}(0,1)$$

where $s(\hat{\alpha}-\hat{\gamma})=\sqrt{\hat{\alpha}(1-\hat{\alpha})+\dfrac{(1+e^{\hat{\gamma}})^2}{\hat{\alpha}e^{\hat{\gamma}}}}$ is the standard error. The rest is standard (you are supposed to describe this standard procedure).

3. For $H_0:\alpha=\frac{1}{2}$, the $\mathcal{LR}$ test statistic is

$$\mathcal{LR}=2\left(\ell_n(z_1,...,z_n,\hat{\alpha})-\ell_n(z_1,...,z_n,\tfrac{1}{2})\right).$$

Therefore,

$$\mathcal{LR}^*=2\left(\ell_n\left(z_1^*,...,z_n^*,\frac{1}{n}\sum\nolimits_{i=1}^{n}z_i^*\right)-\ell_n(z_1^*,...,z_n^*,\hat{\alpha})\right),$$

where the marginal (or, equivalently, joint) loglikelihood is used, should be calculated at each bootstrap repetition. The rest is standard (you are supposed to describe this standard procedure).

7.9 Maximum likelihood and binary dependent variable

1. The conditional ML estimator is

$$\begin{aligned}\hat{\gamma}_{ML} &= \arg\max_c \sum_{i=1}^n \left\{ y_i \log \frac{e^{cx_i}}{1+e^{cx_i}} + (1-y_i)\log\frac{1}{1+e^{cx_i}} \right\} \\ &= \arg\max_c \sum_{i=1}^n \left\{ cy_i x_i - \log\left(1+e^{cx_i}\right)\right\}.\end{aligned}$$

The score is

$$s(y,x,\gamma) = \frac{\partial}{\partial\gamma}\left(\gamma yx - \log\left(1+e^{\gamma x}\right)\right) = \left(y - \frac{e^{\gamma x}}{1+e^{\gamma x}}\right)x,$$

and the information matrix is

$$\mathcal{J} = -\mathbb{E}\left[\frac{\partial s(y,x,\gamma)}{\partial\gamma}\right] = \mathbb{E}\left[\frac{e^{\gamma x}}{(1+e^{\gamma x})^2}x^2\right],$$

so the asymptotic distribution of $\hat{\gamma}_{ML}$ is $N\left(0,\mathcal{J}^{-1}\right)$.

2. The regression is $\mathbb{E}[y|x] = 1\cdot\mathbb{P}\{y=1|x\} + 0\cdot\mathbb{P}\{y=0|x\} = \frac{e^{\gamma x}}{1+e^{\gamma x}}$. The NLLS estimator is

$$\hat{\gamma}_{NLLS} = \arg\min_c \sum_{i=1}^n \left(y_i - \frac{e^{cx_i}}{1+e^{cx_i}}\right)^2.$$

The asymptotic distribution of $\hat{\gamma}_{NLLS}$ is $\mathcal{N}\left(0, Q_{gg}^{-1}Q_{gge^2}Q_{gg}^{-1}\right)$. Now, since $\mathbb{E}\left[e^2|x\right] = \mathbb{V}[y|x] = \frac{e^{\gamma x}}{(1+e^{\gamma x})^2}$, we have

$$Q_{gg} = \mathbb{E}\left[\frac{e^{2\gamma x}}{(1+e^{\gamma x})^4}x^2\right], \quad Q_{gge^2} = \mathbb{E}\left[\frac{e^{2\gamma x}}{(1+e^{\gamma x})^4}x^2\mathbb{E}\left[e^2|x\right]\right] = \mathbb{E}\left[\frac{e^{3\gamma x}}{(1+e^{\gamma x})^6}x^2\right].$$

3. We know that $\mathbb{V}[y|x] = \frac{e^{\gamma x}}{(1+e^{\gamma x})^2}$, which is a function of x. The WNLLS estimator of γ is

$$\hat{\gamma}_{WNLLS} = \arg\min_c \sum_{i=1}^n \frac{(1+e^{\gamma x_i})^2}{e^{\gamma x_i}}\left(y_i - \frac{e^{cx_i}}{1+e^{cx_i}}\right)^2.$$

Note that there should be the *true* γ in the weighting function (or its consistent estimate in a feasible version), but *not* the parameter of choice c! The asymptotic distribution is $\mathcal{N}\left(0, Q_{gg/\sigma^2}^{-1}\right)$, where

$$Q_{gg/\sigma^2} = \mathbb{E}\left[\frac{1}{\mathbb{V}[y|x]}\frac{e^{2\gamma x}}{(1+e^{\gamma x})^4}x^2\right] = \left(\frac{e^{\gamma x}}{(1+e^{\gamma x})^2}x^2\right).$$

4. For the ML problem, the moment condition is "zero expected score"

$$\mathbb{E}\left[\left(y - \frac{e^{\gamma x}}{1+e^{\gamma x}}\right)x\right] = 0.$$

For the NLLS problem, the moment condition is the FOC (or "no correlation between the error and the pseudoregressor")

$$\mathbb{E}\left[\left(y - \frac{e^{\gamma x}}{1+e^{\gamma x}}\right)\frac{e^{\gamma x}}{(1+e^{\gamma x})^2}x\right] = 0.$$

For the WNLLS problem, the moment condition is similar:

$$\mathbb{E}\left[\left(y - \frac{e^{\gamma x}}{1+e^{\gamma x}}\right)x\right] = 0,$$

which is magically the same as for the ML problem. No wonder that the two estimators are asymptotically equivalent (see Part 5).

5. Of course, from the general theory we have $V_{MLE} \le V_{WNLLS} \le V_{NLLS}$. We see a strict inequality $V_{WNLLS} < V_{NLLS}$, except maybe for special cases of the distribution of x, and this is not surprising. Surprising may seem the fact that $V_{MLE} = V_{WNLLS}$. It may be surprising because usually the MLE uses distributional assumptions, and the NLLSE does not, so usually we have $V_{MLE} < V_{WNLLS}$. In this problem, however, the distributional information is used by all estimators, that is, it is *not* an additional assumption made exclusively for ML estimation.

7.10 Bootstrapping ML tests

1. In the bootstrap world, the constraint is $g(q) = g(\hat{\theta}_{ML})$, so

$$\mathcal{LR}^* = 2\left(\max_{q\in\Theta} \ell_n^*(q) - \max_{q\in\Theta, g(q)=g(\hat{\theta}_{ML})} \ell_n^*(q)\right),$$

where ℓ_n^* is the loglikelihood calculated on the bootstrap pseudosample.

2. In the bootstrap world, the constraint is $g(q) = g(\hat{\theta}_{ML})$, so

$$\mathcal{LM}^* = n\left(\frac{1}{n}\sum_{i=1}^n s(z_i^*, \hat{\theta}_{ML}^{*R})\right)'\left(\hat{J}^*\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n s(z_i^*, \hat{\theta}_{ML}^{*R})\right),$$

where $\hat{\theta}_{ML}^{*R}$ is the restricted (subject to $g(q) = g(\hat{\theta}_{ML})$) ML pseudoestimate and $\hat{J}^*$ is the pseudoestimate of the information matrix, both calculated on the bootstrap pseudosample. No additional recentering is needed, since the ZES rule is exactly satisfied at the sample.

7.11 Trivial parameter space

Since the parameter space contains only one point, the latter is the optimizer. If $\theta_1 = \theta_0$, then the estimator $\hat{\theta}_{ML} = \theta_1$ is consistent for θ_0 and has infinite rate of convergence. If $\theta_1 \neq \theta_0$, then the ML estimator is inconsistent.

8. GENERALIZED METHOD OF MOMENTS

8.1 GMM and chi-squared

The feasible GMM estimation procedure for the moment function

$$m(z,q) = \begin{pmatrix} z-q \\ z^2-q^2-2q \end{pmatrix}$$

is the following:

1. Construct a consistent estimator $\hat{\theta}$. For example, set $\hat{\theta} = \bar{z}$ which is a GMM estimator calculated from only the first moment restriction. Calculate a consistent estimator for Q_{mm} as, for example,
$$\hat{Q}_{mm} = \frac{1}{n}\sum_{i=1}^{n} m(z_i,\hat{\theta})m(z_i,\hat{\theta})'$$

2. Find a feasible efficient GMM estimate from the following optimization problem
$$\hat{\theta}_{GMM} = \arg\min_{q} \frac{1}{n}\sum_{i=1}^{n} m(z_i,q)' \cdot \hat{Q}_{mm}^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} m(z_i,q)$$

The asymptotic distribution of the solution is $\sqrt{n}(\hat{\theta}_{GMM}-\theta) \xrightarrow{d} \mathcal{N}\left(0,\frac{3}{2}\right)$, where the asymptotic variance is calculated as

$$V_{\hat{\theta}_{GMM}} = (Q'_{\partial m}Q_{mm}^{-1}Q_{\partial m})^{-1}$$

with

$$Q_{\partial m} = \mathbb{E}\left[\frac{\partial m(z,1)}{\partial q'}\right] = \begin{pmatrix} -1 \\ -4 \end{pmatrix} \text{ and } Q_{mm} = \mathbb{E}\left[m(z,1)m(z,1)'\right] = \begin{pmatrix} 2 & 12 \\ 12 & 96 \end{pmatrix}.$$

A consistent estimator of the asymptotic variance can be calculated as

$$\hat{V}_{\hat{\theta}_{GMM}} = (\hat{Q}'_{\partial m}\hat{Q}_{mm}^{-1}\hat{Q}_{\partial m})^{-1},$$

where

$$\hat{Q}_{\partial m} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(z_i,\hat{\theta}_{GMM})}{\partial q'} \text{ and } \hat{Q}_{mm} = \frac{1}{n}\sum_{i=1}^{n} m(z_i,\hat{\theta}_{GMM})m(z_i,\hat{\theta}_{GMM})'$$

are corresponding analog estimators.

We can also run the J-test to verify the validity of the model:

$$J = \frac{1}{n}\sum_{i=1}^{n} m(z_i,\hat{\theta}_{GMM})' \cdot \hat{Q}_{mm}^{-1} \cdot \sum_{i=1}^{n} m(z_i,\hat{\theta}_{GMM}) \xrightarrow{d} \chi^2(1).$$

8.2 Improved GMM

The first moment restriction gives GMM estimator $\hat{\theta} = \bar{x}$ with asymptotic variance $A\mathbb{V}(\hat{\theta}) = \mathbb{V}[x]$. The GMM estimation of the full set of moment conditions gives estimator $\hat{\theta}_{GMM}$ with asymptotic variance $A\mathbb{V}(\hat{\theta}_{GMM}) = (Q'_{\partial m} Q_{mm}^{-1} Q_{\partial m})^{-1}$, where

$$Q_{\partial m} = \mathbb{E}\left[\frac{\partial m(x, y, \theta)}{\partial q}\right] = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

and

$$Q_{mm} = \mathbb{E}\left[m(x, y, \theta)m(x, y, \theta)'\right] = \begin{pmatrix} \mathbb{V}(x) & \mathbb{C}(x, y) \\ \mathbb{C}(x, y) & \mathbb{V}(y) \end{pmatrix}.$$

Hence,

$$A\mathbb{V}(\hat{\theta}_{GMM}) = \mathbb{V}[x] - \frac{(\mathbb{C}[x, y])^2}{\mathbb{V}[y]}$$

and thus efficient GMM estimation reduces the asymptotic variance when

$$\mathbb{C}[x, y] \neq 0.$$

8.3 Nonlinear simultaneous equations

1. Since $\mathbb{E}[u_i] = \mathbb{E}[v_i] = 0$, $m(w, \theta) = \begin{pmatrix} y - \beta x \\ x - \gamma y^2 \end{pmatrix}$, where $w = \begin{pmatrix} x \\ y \end{pmatrix}$, $\theta = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$, can be used as a moment function. The true β and γ solve $\mathbb{E}[m(w, \theta)] = 0$, therefore $\mathbb{E}[y] = \beta\mathbb{E}[x]$ and $\mathbb{E}[x] = \gamma\mathbb{E}\left[y^2\right]$, and they are identified as long as $\mathbb{E}[x] \neq 0$ and $\mathbb{E}\left[y^2\right] \neq 0$. The analog of the population mean is the sample mean, so the analog estimators are

$$\hat{\beta} = \frac{\frac{1}{n}\sum y_i}{\frac{1}{n}\sum x_i}, \quad \hat{\gamma} = \frac{\frac{1}{n}\sum x_i}{\frac{1}{n}\sum y_i^2}.$$

2. (a) If we add $\mathbb{E}[u_i v_i] = 0$, the moment function is

$$m(w, \theta) = \begin{pmatrix} y - \beta x \\ x - \gamma y^2 \\ (y - \beta x)(x - \gamma y^2) \end{pmatrix}$$

and GMM can be used. The feasible efficient GMM estimator is

$$\hat{\theta}_{GMM} = \arg\min_{q \in \Theta} \left(\frac{1}{n}\sum_{i=1}^{n} m(w_i, q)\right)' \hat{Q}_{mm}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} m(w_i, q)\right),$$

where $\hat{Q}_{mm} = \frac{1}{n}\sum_{i=1}^{n} m(w_i, \hat{\theta})m(w_i, \hat{\theta})'$ and $\hat{\theta}$ is consistent estimator of θ (it can be calculated, from part 1). The asymptotic distribution of this estimator is

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta) \xrightarrow{d} \mathcal{N}(0, V_{GMM}),$$

where $V_{GMM} = (Q'_m Q_{mm}^{-1} Q_m)^{-1}$. The complete answer presumes expressing this matrix in terms of moments of observable variables.

(b) For $H_0 : \beta = \gamma = 0$, the Wald test statistic is $\mathcal{W} = \hat{\theta}'_{GMM}\hat{V}^{-1}_{GMM}\hat{\theta}_{GMM}$. In order to build the bootstrap distribution of this statistic, one should perform the standard bootstrap algorithm, where pseudo-estimators should be constructed as

$$
\begin{aligned}
\hat{\theta}^*_{GMM} &= \arg\min_{q\in\Theta} \left(\frac{1}{n}\sum_{i=1}^n m(w_i^*, q) - \frac{1}{n}\sum_{i=1}^n m(w_i, \hat{\theta}_{GMM})\right)' \hat{Q}^{*-1}_{mm} \times \\
&\quad \times \left(\frac{1}{n}\sum_{i=1}^n m(w_i^*, q) - \frac{1}{n}\sum_{i=1}^n m(w_i, \hat{\theta}_{GMM})\right),
\end{aligned}
$$

and Wald pseudo statistic is calculated as $(\hat{\theta}^*_{GMM} - \hat{\theta}_{GMM})'\hat{V}^{*-1}_{GMM}(\hat{\theta}^*_{GMM} - \hat{\theta}_{GMM})$.

(c) H_0 is $\mathbb{E}\left[m(w,\theta)\right] = 0$, so the test of overidentifying restriction should be performed:

$$
\mathcal{J} = n\left(\frac{1}{n}\sum_{i=1}^n m(w_i, \hat{\theta}_{GMM})\right)' \hat{Q}^{-1}_{mm}\left(\frac{1}{n}\sum_{i=1}^n m(w_i, \hat{\theta}_{GMM})\right),
$$

where $\mathcal{J}$ has asymptotic distribution χ^2_1. So, H_0 is rejected if $\mathcal{J} > q_{0.95}$.

8.4 Trinity for GMM

The Wald test is the same up to a change in the variance matrix:

$$
\mathcal{W} = nh(\hat{\theta}_{GMM})'\left[H(\hat{\theta}_{GMM})(\hat{\Omega}'\hat{\Sigma}^{-1}\hat{\Omega})^{-1}H(\hat{\theta}_{GMM})'\right]^{-1} h(\hat{\theta}_{GMM}) \xrightarrow{d} \chi^2_q,
$$

where $\hat{\theta}_{GMM}$ is the unrestricted GMM estimator, $\hat{\Omega}$ and $\hat{\Sigma}$ are consistent estimators of Ω and Σ, relatively, and $H(\theta) = \dfrac{\partial h(\theta)}{\partial \theta'}$.

The Distance Difference test is similar to the $\mathcal{LR}$ test, but without factor 2, since $\dfrac{\partial^2 \hat{Q}_n}{\partial\theta\partial\theta'} \xrightarrow{p} 2\Omega'\Sigma^{-1}\Omega$:

$$
DD = n\left[Q_n(\hat{\theta}^R_{GMM}) - Q_n(\hat{\theta}_{GMM})\right] \xrightarrow{d} \chi^2_q.
$$

The LM test is a little bit harder, since the analog of the average score is

$$
\lambda(\theta) = 2\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial m(z_i,\theta)}{\partial\theta'}\right)' \hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^n m(z_i,\theta)\right).
$$

It is straightforward to find that

$$
\mathcal{LM} = \frac{n}{4}\lambda(\hat{\theta}^R_{GMM})'(\hat{\Omega}'\hat{\Sigma}^{-1}\hat{\Omega})^{-1}\lambda(\hat{\theta}^R_{GMM}) \xrightarrow{d} \chi^2_q.
$$

In the middle one may use either restricted or unrestricted estimators of Ω and Σ.

8.5 Testing moment conditions

Consider the unrestricted ($\hat{\beta}_u$) and restricted ($\hat{\beta}_r$) estimates of parameter $\beta \in R^k$. The first is the CMM estimate (i. e. a GMM estimate for the case of just identification):

$$\sum_{i=1}^n x_i(y_i - x_i'\hat{\beta}_u) = 0 \Rightarrow \hat{\beta}_u = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i y_i$$

The second is a feasible efficient GMM estimate:

$$\hat{\beta}_r = \arg\min_b \frac{1}{n}\sum_{i=1}^n m_i(b)' \cdot \hat{Q}_{mm}^{-1} \cdot \frac{1}{n}\sum_{i=1}^n m_i(b), \tag{8.1}$$

where $m_i(b) = \begin{pmatrix} x_i u_i(b) \\ x_i u_i(b)^3 \end{pmatrix}$, $u_i(b) = y_i - x_i b$, $u_i \equiv u_i(\beta)$, and $\hat{Q}_{mm}^{-1}$ is an efficient estimator of

$$Q_{mm} = \mathbb{E}\left[m_i(\beta)m_i'(\beta)\right] = \mathbb{E}\left[\begin{pmatrix} x_i x_i' u_i^2 & x_i x_i' u_i^4 \\ x_i x_i' u_i^4 & x_i x_i' u_i^6 \end{pmatrix}\right].$$

Denote also $Q_{\partial m} = \mathbb{E}\left[\dfrac{\partial m_i(\beta)}{\partial b'}\right] = \mathbb{E}\left[\begin{pmatrix} -x_i x_i' \\ -3x_i x_i' u_i^2 \end{pmatrix}\right]$. Writing out the FOC for (8.1) and expanding $m_i(\hat{\beta}_r)$ around β gives after rearrangement

$$\sqrt{n}(\hat{\beta}_r - \beta) \stackrel{A}{=} -\left(Q_{\partial m}' Q_{mm}^{-1} Q_{\partial m}\right)^{-1} Q_{\partial m}' Q_{mm}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n m_i(\beta).$$

Here $\stackrel{A}{=}$ means that we substitute the probability limits for their sample analogues. The last equation holds under the null hypothesis $H_0 : \mathbb{E}\left[x_i u_i^3\right] = 0$.

Note that the unrestricted estimate can be rewritten as

$$\sqrt{n}(\hat{\beta}_u - \beta) \stackrel{A}{=} \mathbb{E}\left[x_i x_i'\right]^{-1} \begin{pmatrix} I_k & O_k \end{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^n m_i(\beta).$$

Therefore,

$$\sqrt{n}(\hat{\beta}_u - \beta_r) \stackrel{A}{=} \left[\left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1} \begin{pmatrix} I_k & O_k \end{pmatrix} + \left(Q_{\partial m}' Q_{mm}^{-1} Q_{\partial m}\right)^{-1} Q_{\partial m}' Q_{mm}^{-1}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^n m_i(\beta) \xrightarrow{d} \mathcal{N}(0, V),$$

where (after some algebra)

$$V = \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1} \mathbb{E}\left[x_i x_i' u_i^2\right] \left(\mathbb{E}\left[x_i x_i'\right]\right)^{-1} - \left(Q_{\partial m}' Q_{mm}^{-1} Q_{\partial m}\right)^{-1}.$$

Note that V is $k \times k$. matrix. It can be shown that this matrix is non-degenerate (and thus has a full rank k). Let $\hat{V}$ be a consistent estimate of V. By Slutsky and Wald-Mann Theorems

$$\mathcal{W} \equiv n(\hat{\beta}_u - \hat{\beta}_r)' \hat{V}^{-1} (\hat{\beta}_u - \beta_r) \xrightarrow{d} \chi_k^2.$$

The test may be implemented as follows. First find the (consistent) estimate $\hat{\beta}_u$ given x_i and y_i. Then compute $\hat{Q}_{mm} = \frac{1}{n}\sum_{i=1}^n m_i(\hat{\beta}_u) m_i(\hat{\beta}_u)'$, use it to carry out feasible GMM and obtain $\hat{\beta}_r$. Use $\hat{\beta}_u$ or $\hat{\beta}_r$ to find $\hat{V}$ (the sample analog of V). Finally, compute the Wald statistic $\mathcal{W}$, compare it with 95% quantile of $\chi^2(k)$ distribution $q_{0.95}$, and reject the null hypothesis if $\mathcal{W} > q_{0.95}$, or accept otherwise.

8.6 Interest rates and future inflation

1. The conventional econometric model that tests the hypothesis of conditional unbiasedness of interest rates as predictors of inflation, is

$$\pi_t^k = \alpha_k + \beta_k i_t^k + \eta_t^k, \quad \mathbb{E}_t\left[\eta_t^k\right] = 0.$$

Under the null, $\alpha_k = 0$, $\beta_k = 1$. Setting $k = m$ in one case, $k = n$ in the other case, and subtracting one equation from another, we can get

$$\pi_t^m - \pi_t^n = \alpha_m - \alpha_n + \beta_m i_t^m - \beta_n i_t^n + \eta_t^n - \eta_t^m, \quad \mathbb{E}_t\left[\eta_t^n - \eta_t^m\right] = 0.$$

Under the null $\alpha_m = \alpha_n = 0$, $\beta_m = \beta_n = 1$, this specification coincides with Mishkin's under the null $\alpha_{m,n} = 0$, $\beta_{m,n} = 1$. The restriction $\beta_{m,n} = 0$ implies that the term structure provides no information about future shifts in inflation. The prediction error $\eta_t^{m,n}$ is serially correlated of the order that is the farthest prediction horizon, i.e., $\max(m, n)$.

2. Selection of instruments: there is a variety of choices, for instance,

$$\left(1, i_t^m - i_t^n, i_{t-1}^m - i_{t-1}^n, i_{t-2}^m - i_{t-2}^n, \pi_{t-\max(m,n)}^m - \pi_{t-\max(m,n)}^n\right)',$$

or

$$\left(1, i_t^m, i_t^n, i_{t-1}^m, i_{t-1}^n, \pi_{t-\max(m,n)}^m, \pi_{t-\max(m,n)}^n\right)',$$

etc. Construction of the optimal weighting matrix demands Newey-West (or similar robust) procedure, and so does estimation of asymptotic variance. The rest is more or less standard.

3. This is more or less standard. There are two subtle points: recentering when getting a pseudoestimator, and recentering when getting a pseudo-J-statistic.

4. Most interesting are the results of the test $\beta_{m,n} = 0$ which tell us that there is no information in the term structure about future path of inflation. Testing $\beta_{m,n} = 1$ then seems excessive. This hypothesis would correspond to the conditional bias containing only a systematic component (i.e. a constant unpredictable by the term structure). It also looks like there is no systematic component in inflation ($\alpha_{m,n} = 0$ is accepted).

8.7 Spot and forward exchange rates

1. This is not the only way to proceed, but it is straightforward. The OLS estimator uses the instrument $z_t^{OLS} = (1\ x_t)'$, where $x_t = f_t - s_t$. The additional moment condition adds $f_{t-1} - s_{t-1}$ to the list of instruments: $z_t = (1\ x_t\ x_{t-1})'$. Let us look at the optimal instrument. If it is proportional to z_t^{OLS}, then the instrument x_{t-1}, and hence the additional moment condition, is redundant. The optimal instrument takes the form $\zeta_t = Q_{\partial m}' Q_{mm}^{-1} z_t$. But

$$Q_{\partial m} = -\begin{pmatrix} 1 & \mathbb{E}[x_t] \\ \mathbb{E}[x_t] & \mathbb{E}[x_t^2] \\ \mathbb{E}[x_{t-1}] & \mathbb{E}[x_t x_{t-1}] \end{pmatrix}, \quad Q_{mm} = \sigma^2 \begin{pmatrix} 1 & \mathbb{E}[x_t] & \mathbb{E}[x_{t-1}] \\ \mathbb{E}[x_t] & \mathbb{E}[x_t^2] & \mathbb{E}[x_t x_{t-1}] \\ \mathbb{E}[x_{t-1}] & \mathbb{E}[x_t x_{t-1}] & \mathbb{E}[x_{t-1}^2] \end{pmatrix}.$$

It is easy to see that

$$Q'_{\partial m}Q_{mm}^{-1} = \sigma^{-2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

which can verified by postmultiplying this equation by Q_{mm}. Hence, $\zeta_t = \sigma^{-2}z_t^{OLS}$. But the most elegant way to solve this problem goes as follows. Under conditional homoskedasticity, the GMM estimator is asymptotically equivalent to the 2SLS estimator, if both use the same vector of instruments. But if the instrumental vector includes the regressors (z_t does include z_t^{OLS}), the 2SLS estimator is identical to the OLS estimator (for an example, see Problem #2 in Assignment #5). In total, GMM is asymptotically equivalent to OLS and thus the additional moment condition is redundant.

2. This problem is similar to Problem #1(2) of Assignment #3, so we can expect asymptotic equivalence of the OLS and efficient GMM estimators when the additional moment function is uncorrelated with the main moment function. Indeed, let us compare the 2×2 northwest block of $V_{GMM} = \left(Q'_{\partial m}Q_{mm}^{-1}Q_{\partial m}\right)^{-1}$ with asymptotic variance of the OLS estimator

$$V_{OLS} = \sigma^2\begin{pmatrix} 1 & \mathbb{E}[x_t] \\ \mathbb{E}[x_t] & \mathbb{E}[x_t^2] \end{pmatrix}^{-1}.$$

Denote $\Delta f_{t+1} = f_{t+1} - f_t$. For the full set of moment conditions,

$$Q_{\partial m} = -\begin{pmatrix} 1 & \mathbb{E}[x_t] \\ \mathbb{E}[x_t] & \mathbb{E}[x_t^2] \\ 0 & 0 \end{pmatrix}, \quad Q_{mm} = \begin{pmatrix} \sigma^2 & \sigma^2\mathbb{E}[x_t] & \mathbb{E}[e_{t+1}\Delta f_{t+1}] \\ \sigma^2\mathbb{E}[x_t] & \sigma^2\mathbb{E}[x_t^2] & \mathbb{E}[x_te_{t+1}\Delta f_{t+1}] \\ \mathbb{E}[e_{t+1}\Delta f_{t+1}] & \mathbb{E}[x_te_{t+1}\Delta f_{t+1}] & \mathbb{E}[e_{t+1}^2(\Delta f_{t+1})^2] \end{pmatrix}.$$

It is easy to see that when $\mathbb{E}[e_{t+1}\Delta f_{t+1}] = \mathbb{E}[x_te_{t+1}\Delta f_{t+1}] = 0$, Q_{mm} is block-diagonal and the 2×2 northwest block of V_{GMM} is the same as V_{OLS}. A sufficient condition for these two equalities is $\mathbb{E}[e_{t+1}\Delta f_{t+1}|I_t] = 0$, i. e. that conditionally on the past, unexpected movements in spot rates are uncorrelated with and unexpected movements in forward rates. This is hardly to be satisfied in practice.

8.8 Brief and exhaustive

1. We know that $\mathbb{E}[w] = \mu$ and $\mathbb{E}\left[(w-\mu)^4\right] = 3\left(\mathbb{E}\left[(w-\mu)^2\right]\right)^2$. It is trivial to take care of the former. To take care of the latter, introduce a constant $\sigma^2 = \mathbb{E}\left[(w-\mu)^2\right]$, then we have $\mathbb{E}\left[(w-\mu)^4\right] = 3\left(\sigma^2\right)^2$. Together, the system of moment conditions is

$$\mathbb{E}\left[\begin{pmatrix} w-\mu \\ (w-\mu)^2 - \sigma^2 \\ (w-\mu)^4 - 3\left(\sigma^2\right)^2 \end{pmatrix}\right] = \underset{3\times 1}{0}.$$

2. The argument would be fine if the model for the conditional mean was known to be correctly specified. Then one could blame instruments for a high value of the J-statistic. But in our time series regression of the type $\mathbb{E}_t[y_{t+1}] = g(x_t)$, if this regression was correctly specified, then the variables from time t information set *must* be valid instruments! The failure of the model may be associated with incorrect functional form of $g(\cdot)$, or with specification of conditional information. Lastly, asymptotic theory may give a poor approximation to exact distribution of the J-statistic.

3. Indeed, we are supposed to recenter, but only when there is overidentification. When the parameter is just identified, as in the case of the OLS estimator, the moment conditions hold exactly in the sample, so the "center" is zero anyway.

8.9 Efficiency of MLE in GMM class

The theorem we proved in class began with the following. The true parameter θ solves the maximization problem

$$\theta = \arg\max_{q\in\Theta} \mathbb{E}\left[h(z,q)\right]$$

with a first order condition

$$\mathbb{E}\left[\frac{\partial}{\partial q}h(z,\theta)\right] = \underset{k\times 1}{0}.$$

Consider the GMM minimization problem

$$\theta = \arg\min_{q\in\Theta} \mathbb{E}\left[m(z,q)\right]' W \mathbb{E}\left[m(z,q)\right]$$

with FOC

$$2\mathbb{E}\left[\frac{\partial}{\partial q'}m(z,\theta)\right]' W \mathbb{E}\left[m(z,\theta)\right] = \underset{k\times 1}{0},$$

or, equivalently,

$$\mathbb{E}\left[\mathbb{E}\left[\frac{\partial}{\partial q}m(z,\theta)'\right] W m(z,\theta)\right] = \underset{k\times 1}{0}.$$

Now treat the vector $\mathbb{E}\left[\frac{\partial}{\partial q}m(z,\theta)'\right] W m(z,q)$ as $\frac{\partial}{\partial q}h(z,q)$ in the given proof, and we are done.

9. PANEL DATA

9.1 Alternating individual effects

It is convenient to use three indices instead of two in indexing the data. Namely, let

$$t = 2(s-1) + q, \text{ where } q \in \{1, 2\},\ s \in \{1, \cdots, T\}.$$

Then $q = 1$ corresponds to odd periods, while $q = 2$ corresponds to even periods. The dummy variables will have the form of the Kronecker product of three matrices, which is defined recursively as $A \otimes B \otimes C = A \otimes (B \otimes C)$.

Part 1. (a) In this case we rearrange the data column as follows:

$$y_{isq} = y_{it};\ y_{is} = \begin{pmatrix} y_{is1} \\ y_{is2} \end{pmatrix};\ y_i = \begin{pmatrix} y_{i1} \\ \dots \\ y_{iT} \end{pmatrix};\ y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix};$$

$\mu = (\mu_1^O\ \mu_1^E\ \cdots\ \mu_n^O\ \mu_n^E)'$. The regressors and errors are rearranged in the same manner as y's. Then the regression can be rewritten as

$$y = D\mu + X\beta + v, \tag{9.1}$$

where $D = I_n \otimes i_T \otimes I_2$, and $i_T = (1\ \cdots\ 1)'$ ($T \times 1$ vector). Clearly,

$$D'D = I_n \otimes i_T' i_t \otimes I_2 = T \cdot I_{n \cdot T \cdot 2},$$

$$D(D'D)^{-1}D' = \frac{1}{T} I_n \otimes i_T i_t' \otimes I_2 = \frac{1}{T} I_n \otimes J_T \otimes I_2,$$

where $J_T = i_T i_T'$. In other words, $D(D'D)^{-1}D'$ is block-diagonal with n $2T \times 2T$-blocks of the form:

$$\begin{pmatrix} \frac{1}{T} & 0 & \dots & \frac{1}{T} & 0 \\ 0 & \frac{1}{T} & \dots & 0 & \frac{1}{T} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{1}{T} & 0 & \dots & \frac{1}{T} & 0 \\ 0 & \frac{1}{T} & \dots & 0 & \frac{1}{T} \end{pmatrix}.$$

The Q-matrix is then $Q = I_{n \cdot T \cdot 2} - \frac{1}{T} I_n \otimes J_T \otimes I_2$. Note that Q is an orthogonal projection and $QD = 0$. Thus we have from (9.1)

$$Qy = QX\beta + Qv. \tag{9.2}$$

Note that $\frac{1}{T} J_T$ is the operator of taking the mean over the s-index (i.e. over odd or even periods depending on the value of q). Therefore, the transformed regression is:

$$y_{isq} - \bar{y}_{iq} = (x_{isq} - \bar{x}_{iq})'\beta + v^*, \tag{9.3}$$

where $\bar{y}_{iq} = \sum_{s=1}^{T} y_{isq}$.

(b) This time the data are rearranged in the following manner:

$$y_{qis} = y_{it};\ y_{qi} = \begin{pmatrix} y_{i1} \\ \dots \\ y_{iT} \end{pmatrix};\ y_q = \begin{pmatrix} y_{q1} \\ \dots \\ y_{qn} \end{pmatrix};\ y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix};$$

$\mu = (\mu_1^O \; \cdots \; \mu_n^O \; \mu_1^E \; \cdots \; \mu_n^E)'$. In matrix form the regression is again (9.1) with $D = I_2 \otimes I_n \otimes i_T$, and

$$D(D'D)^{-1}D' = \frac{1}{T} I_2 \otimes I_n \otimes i_T i_t' = \frac{1}{T} I_2 \otimes I_n \otimes J_T.$$

This matrix consists of $2N$ blocks on the main diagonal, each of them being $\frac{1}{T} J_T$. The Q-matrix is $Q = I_{2n \cdot T} - \frac{1}{T} I_{2n} \otimes J_T$. The rest is as in Part 1(b) with the transformed regression

$$y_{qis} - \bar{y}_{qi} = (x_{qis} - \bar{x}_{qi})'\beta + v^*, \tag{9.4}$$

with $\bar{y}_{qi} = \sum_{s=1}^T y_{qis}$, which is essentially the same as (9.3).

Part 2. Take the Q-matrix as in Part 1(b). The Within estimator is the OLS estimator in (9.4), i.e. $\hat{\beta} = (X'QX)^{-1} X'QY$, or

$$\hat{\beta} = \left(\sum_{q,i,s} (x_{qis} - \bar{x}_{qi})(x_{qis} - \bar{x}_{qi})' \right)^{-1} \sum_{q,i,s} (x_{qis} - \bar{x}_{qi})(y_{qis} - \bar{y}_{qi}).$$

Clearly, $\mathbb{E}[\hat{\beta}] = \beta$, $\hat{\beta} \xrightarrow{p} \beta$ and $\hat{\beta}$ is asymptotically normal as $n \to \infty$, T fixed. For normally distributed errors v_{qis} the standard F-test for hypothesis

$$H_0 : \mu_1^O = \mu_2^O = ... = \mu_n^O \text{ and } \mu_1^E = \mu_2^E = ... = \mu_n^E$$

is

$$F = \frac{(RSS^R - RSS^U)/(2n-2)}{RSS^U/(2nT - 2n - k)} \overset{H_0}{\sim} F(2n-2, 2nT - 2n - k)$$

(we have $2n-2$ restrictions in the hypothesis), where $RSS^U = \sum_{isq} (y_{qis} - \bar{y}_{qi} - (x_{qis} - \bar{x}_{qi})'\beta)^2$, and RSS^R is the sum of squared residuals in the restricted regression.

Part 3. Here we start with

$$y_{qis} = x_{qis}'\beta + u_{qis}, \; u_{qis} := \mu_{qi} + v_{qis}, \tag{9.5}$$

where $\mu_{1i} = \mu_i^O$ and $\mu_{2i} = \mu_i^E$; $\mathbb{E}[\mu_{qi}] = 0$. Let $\sigma_1^2 = \sigma_O^2$, $\sigma_2^2 = \sigma_E^2$. We have

$$\mathbb{E}\left[u_{qis} u_{q'i's'}\right] = \mathbb{E}\left[(\mu_{qi} + v_{qis})(\mu_{q'i'} + v_{q'i's'})\right] = \sigma_q^2 \delta_{qq'} \delta_{ii'} 1_{ss'} + \sigma_v^2 \delta_{qq'} \delta_{ii'} \delta_{ss'},$$

where $\delta_{aa'} = \{1 \text{ if } a = a', \text{ and } 0 \text{ if } a \neq a'\}$, $1_{ss'} = 1$ for all s, s'. Consequently,

$$\begin{aligned} \Omega &= \mathbb{E}[uu'] = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \otimes I_n \otimes J_T + \sigma_v^2 I_{2nT} = (T\sigma_1^2 + \sigma_v^2) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes I_n \otimes \frac{1}{T} J_T + \\ &\quad + (T\sigma_2^2 + \sigma_v^2) \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes I_n \otimes \frac{1}{T} J_T + \sigma_v^2 I_2 \otimes I_n \otimes (I_T - \frac{1}{T} J_T). \end{aligned}$$

The last expression is the spectral decomposition of Ω since all operators in it are idempotent symmetric matrices (orthogonal projections), which are orthogonal to each other and give identity in sum. Therefore,

$$\begin{aligned} \Omega^{-1/2} &= (T\sigma_1^2 + \sigma_v^2)^{-1/2} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes I_n \otimes \frac{1}{T} J_T + (T\sigma_2^2 + \sigma_v^2)^{-1/2} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes I_n \otimes \frac{1}{T} J_T + \\ &\quad + \sigma_v^{-1} I_2 \otimes I_n \otimes (I_T - \frac{1}{T} J_T). \end{aligned}$$

The GLS estimator of β is

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}Y.$$

To put it differently, $\hat{\beta}$ is the OLS estimator in the transformed regression

$$\sigma_v \Omega^{-1/2} y = \sigma_v \Omega^{-1/2} X\beta + u^*.$$

The latter may be rewritten as

$$y_{qis} - (1 - \sqrt{\theta_q})\bar{y}_{qi} = (x_{qis} - (1 - \sqrt{\theta_q})\bar{x}_{qi})'\beta + u^*,$$

where $\theta_q = \sigma_v^2/(\sigma_v^2 + T\sigma_q^2)$.

To make $\hat{\beta}$ feasible, we should consistently estimate parameter θ_q. In the case $\sigma_1^2 = \sigma_2^2$ we may apply the result obtained in class (we have $2n$ different objects and T observations for each of them – see Part 1(b)):

$$\hat{\theta} = \frac{2n - k}{2n(T-1) - k + 1} \frac{\hat{u}' Q \hat{u}}{\hat{u}' P \hat{u}},$$

where $\hat{u}$ are OLS-residuals for (9.4), and $Q = I_{2n \cdot T} - \frac{1}{T} I_{2n} \otimes J_T$, $P = I_{2nT} - Q$. Suppose now that $\sigma_1^2 \neq \sigma_2^2$. Using equations

$$\mathbb{E}\left[u_{qis}\right] = \sigma_v^2 + \sigma_q^2; \quad \mathbb{E}\left[\bar{u}_{is}\right] = \frac{1}{T}\sigma_v^2 + \sigma_q^2,$$

and repeating what was done in class, we have

$$\hat{\theta}_q = \frac{n - k}{n(T-1) - k + 1} \frac{\hat{u}' Q_q \hat{u}}{\hat{u}' P_q \hat{u}},$$

with $Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes I_n \otimes (I_T - \frac{1}{T} J_T)$, $Q_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes I_n \otimes (I_T - \frac{1}{T} J_T)$, $P_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \otimes I_n \otimes \frac{1}{T} J_T$, $P_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes I_n \otimes \frac{1}{T} J_T$.

9.2 Time invariant regressors

1. (a) Under fixed effects, the z_i variable is collinear with the dummy for μ_i. Thus, γ is unidentifiable.. The Within transformation wipes out the term $z_i\gamma$ together with individual effects μ_i, so the transformed equation looks exactly like it looks if no term $z_i\gamma$ is present in the model. Under usual assumptions about independence of v_{it} and X, the Within estimator of β is efficient.

   (b) Under random effects and mutual independence of μ_i and v_{it}, as well as their independence of X and Z, the GLS estimator is efficient, and the feasible GLS estimator is asymptotically efficient as $n \to \infty$.

2. Recall that the first-step $\hat{\beta}$ is consistent but $\hat{\pi}_i$'s are inconsistent as T stays fixed and $n \to \infty$. However, the estimator of γ so constructed is consistent under assumptions of random effects (see Part 1(b)). Observe that $\hat{\pi}_i = \bar{y}_i - \bar{x}_i'\hat{\beta}$. If we regress $\hat{\pi}_i$ on z_i, we get the OLS coefficient

$$\hat{\gamma} = \frac{\sum_{i=1}^n z_i \hat{\pi}_i}{\sum_{i=1}^n z_i^2} = \frac{\sum_{i=1}^n z_i \left(\bar{y}_i - \bar{x}_i'\hat{\beta}\right)}{\sum_{i=1}^n z_i^2} = \frac{\sum_{i=1}^n z_i \left(\bar{x}_i'\beta + z_i\gamma + \mu_i + \bar{v}_i - \bar{x}_i'\hat{\beta}\right)}{\sum_{i=1}^n z_i^2}$$
$$= \gamma + \frac{\frac{1}{n}\sum_{i=1}^n z_i\mu_i}{\frac{1}{n}\sum_{i=1}^n z_i^2} + \frac{\frac{1}{n}\sum_{i=1}^n z_i\bar{v}_i}{\frac{1}{n}\sum_{i=1}^n z_i^2} + \frac{\frac{1}{n}\sum_{i=1}^n z_i\bar{x}_i'}{\frac{1}{n}\sum_{i=1}^n z_i^2}\left(\beta - \hat{\beta}\right).$$

Now, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} z_i^2 \xrightarrow{p} \mathbb{E}\left[z_i^2\right] \neq 0, \quad \frac{1}{n}\sum_{i=1}^{n} z_i\mu_i \xrightarrow{p} \mathbb{E}\left[z_i\mu_i\right] = \mathbb{E}\left[z_i\right]\mathbb{E}\left[\mu_i\right] = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} z_i\bar{v}_i \xrightarrow{p} \mathbb{E}\left[z_i\bar{v}_i\right] = \mathbb{E}\left[z_i\right]\mathbb{E}\left[\bar{v}_i\right] = 0, \quad \frac{1}{n}\sum_{i=1}^{n} z_i\bar{x}_i' \xrightarrow{p} \mathbb{E}\left[z_i\bar{x}_i'\right], \quad \beta - \hat{\beta} \xrightarrow{p} 0.$$

In total, $\hat{\gamma} \xrightarrow{p} \gamma$. However, so constructed estimator of γ is asymptotically inefficient. A better estimator is the feasible GLS estimator of Part 1(b).

9.3 First differencing transformation

OLS on FD-transformed equations is unbiased and consistent as $n \to \infty$ since the differenced error has mean zero conditional on the matrix of differenced regressors under the standard FE assumptions. However, OLS is inefficient as the conditional variance matrix is not diagonal. The efficient estimator of structural parameters is the LSDV estimator, which is an OLS estimator on Within-transformed equations.

10. NONPARAMETRIC ESTIMATION

10.1 Nonparametric regression with discrete regressor

Fix $a_{(j)}$, $j = 1, \dots, k$. Observe that

$$g(a_{(j)}) = \mathbb{E}[y_i|x_i = a_{(j)}] = \frac{\mathbb{E}(y_i \mathbb{I}[x_i = a_{(j)}])}{\mathbb{E}(\mathbb{I}[x_i = a_{(j)}])}$$

because of the following equalities:

$$\begin{aligned} \mathbb{E}\left[\mathbb{I}\left[x_i = a_{(j)}\right]\right] &= 1 \cdot \mathbb{P}\{x_i = a_{(j)}\} + 0 \cdot \mathbb{P}\{x_i \neq a_{(j)}\} = \mathbb{P}\{x_i = a_{(j)}\}, \\ \mathbb{E}\left[y_i \mathbb{I}\left[x_i = a_{(j)}\right]\right] &= \mathbb{E}\left[y_i \mathbb{I}\left[x_i = a_{(j)}\right]|x_i = a_{(j)}\right] \cdot \mathbb{P}\{x_i = a_{(j)}\} = \mathbb{E}\left[y_i|x_i = a_{(j)}\right] \cdot \mathbb{P}\{x_i = a_{(j)}\}. \end{aligned}$$

According to the analogy principle we can construct $\hat{g}(a_{(j)})$ as

$$\hat{g}(a_{(j)}) = \frac{\sum_{i=1}^n y_i \mathbb{I}\left[x_i = a_{(j)}\right]}{\sum_{i=1}^n \mathbb{I}\left[x_i = a_{(j)}\right]}.$$

Now let us find its properties. First, according to the LLN,

$$\hat{g}(a_{(j)}) = \frac{\sum_{i=1}^n y_i \mathbb{I}\left[x_i = a_{(j)}\right]}{\sum_{i=1}^n \mathbb{I}\left[x_i = a_{(j)}\right]} \xrightarrow{p} \frac{\mathbb{E}\left[y_i \mathbb{I}[x_i = a_{(j)}]\right]}{\mathbb{E}\left[\mathbb{I}[x_i = a_{(j)}]\right]} = g(a_{(j)}).$$

Second,

$$\sqrt{n}\left(\hat{g}(a_{(j)}) - g(a_{(j)})\right) = \sqrt{n}\frac{\sum_{i=1}^n \left(y_i - \mathbb{E}\left[y_i|x_i = a_{(j)}\right]\right)\mathbb{I}\left[x_i = a_{(j)}\right]}{\sum_{i=1}^n \mathbb{I}\left[x_i = a_{(j)}\right]}.$$

According to the CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(y_i - \mathbb{E}\left[y_i|x_i = a_{(j)}\right]\right)\mathbb{I}\left[x_i = a_{(j)}\right] \xrightarrow{d} \mathcal{N}(0, \omega),$$

where

$$\begin{aligned} \omega &= \mathbb{V}\left[\left(y_i - \mathbb{E}\left[y_i|x_i = a_{(j)}\right]\right)\mathbb{I}\left[x_i = a_{(j)}\right]\right] = \mathbb{E}\left[\left(y_i - \mathbb{E}\left[y_i|x_i = a_{(j)}\right]\right)^2 |x_i = a_{(j)}\right]\mathbb{P}\{x_i = a_{(j)}\} \\ &= \mathbb{V}\left[y_i|x_i = a_{(j)}\right]\mathbb{P}\{x_i = a_{(j)}\}. \end{aligned}$$

Thus

$$\sqrt{n}\left(\hat{g}(a_{(j)}) - g(a_{(j)})\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{V}\left[y_i|x_i = a_{(j)}\right]}{\mathbb{P}\{x_i = a_{(j)}\}}\right).$$

10.2 Nonparametric density estimation

(a) Use the hint that $\mathbb{E}\left[\mathbb{I}\left[x_i \leq x\right]\right] = F(x)$ to prove the unbiasedness of estimator:

$$\mathbb{E}\left[\hat{F}(x)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{I}\left[x_i \leq x\right]\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbb{I}\left[x_i \leq x\right]\right] = \frac{1}{n}\sum_{i=1}^n F(x) = F(x).$$

(b) Use the Taylor expansion $F(x+h) = F(x) + hf(x) + \frac{1}{2}h^2 f'(x) + o(h^2)$ to see that the bias of $\hat{f}_1(x)$ is

$$\mathbb{E}\left[\hat{f}_1(x)\right] - f(x) = h^{-1}\left(F(x+h) - F(x)\right) - f(x)$$
$$= \frac{1}{h}(F(x) + hf(x) + \frac{1}{2}h^2 f'(x) + o(h^2) - F(x)) - f(x)$$
$$= \frac{1}{2}hf'(x) + o(h).$$

Therefore, $a = 1$.

(c) Use the Taylor expansions $F\left(x + \frac{h}{2}\right) = F(x) + \frac{h}{2}f(x) + \frac{1}{2}\left(\frac{h}{2}\right)^2 f'(x) + \frac{1}{6}\left(\frac{h}{2}\right)^3 f''(x) + o(h^3)$ and $F\left(x - \frac{h}{2}\right) = F(x) - \frac{h}{2}f(x) + \frac{1}{2}\left(\frac{h}{2}\right)^2 f'(x) - \frac{1}{6}\left(\frac{h}{2}\right)^3 f''(x) + o(h^3)$ to see that the bias of $\hat{f}_2(x)$ is

$$\mathbb{E}\left[\hat{f}_2(x)\right] - f(x) = h^{-1}\left(F(x+h/2) - F(x-h/2)\right) - f(x) = \frac{1}{24}h^2 f''(x) + o(h^2).$$

Therefore, $b = 2$.

Let us compare the two methods. We can find the optimal rate of convergence when the bias and variance are balanced: variance $\propto$ bias^2. The "variance" is of order nh for both methods, but the "bias" is of different order (see parts (b) and (c)). For $\hat{f}_1$, the optimal rate is $n \propto h^{-1/3}$, for $\hat{f}_2$ – the optimal rate is $n \propto h^{-1/5}$. Therefore, for the same h, with the second method we need more points to estimate f with the same accuracy.

Let us compare the performance of each method at border points like $x_{(1)}$ or $x_{(n)}$, and at a median point like $\bar{x}_n$. To estimate $f(x)$ with approximately the same variance we need an approximately same number of points in the window $[x, x+h]$ for the first method and $[x - h/2, x+h/2]$ for the second. Since concentration of points in the window at a border is much lower than in the median window, we need a much bigger sample to estimate the density at border points with the same accuracy as at median points. On the other hand, when the sample size is fixed, we need greater h for border points to meet the accuracy of estimation with that for in median points. When h increases, the bias increases with the same rate in the first method and with the double rate in the second method. Consequently, $\hat{f}_1$ is preferable for estimation at border points.

10.3 First difference transformation and nonparametric regression

1. Let us consider the following average that can be decomposed into three terms:

$$\frac{1}{n-1}\sum_{i=2}^{n}(y_i - y_{i-1})^2 = \frac{1}{n-1}\sum_{i=2}^{n}(g(z_i) - g(z_{i-1}))^2 + \frac{1}{n-1}\sum_{i=2}^{n}(e_i - e_{i-1})^2$$
$$+\frac{2}{n-1}\sum_{i=2}^{n}(g(z_i) - g(z_{i-1}))(e_i - e_{i-1}).$$

Since z_i compose a uniform grid and are increasing in order, i.e. $z_i - z_{i-1} = \frac{1}{n-1}$, we can find the limit of the first term using the Lipschitz condition:

$$\left|\frac{1}{n-1}\sum_{i=2}^{n}(g(z_i) - g(z_{i-1}))^2\right| \le \frac{G^2}{n-1}\sum_{i=2}^{n}(z_i - z_{i-1})^2 = \frac{G^2}{(n-1)^2} \underset{n\to\infty}{\to} 0$$

Using the Lipschitz condition again we can find the probability limit of the third term:

$$\left|\frac{2}{n-1}\sum_{i=2}^{n}(g(z_i)-g(z_{i-1}))(e_i-e_{i-1})\right| \le \frac{2G}{(n-1)^2}\sum_{i=2}^{n}|e_i-e_{i-1}|$$
$$\le \frac{2G}{n-1}\frac{1}{n-1}\sum_{i=2}^{n}\left(|e_i|+|e_{i-1}|\right) \underset{n\to\infty}{\overset{p}{\to}} 0$$

since $\frac{2G}{n-1} \underset{n\to\infty}{\to} 0$ and $\frac{1}{n-1}\sum_{i=2}^{n}\left(|e_i|+|e_{i-1}|\right) \underset{n\to\infty}{\overset{p}{\to}} 2\mathbb{E}\,|e_i| < \infty$. The second term has the following probability limit:

$$\frac{1}{n-1}\sum_{i=2}^{n}(e_i-e_{i-1})^2 = \frac{1}{n-1}\sum_{i=2}^{n}\left(e_i^2-2e_ie_{i-1}+e_{i-1}^2\right) \underset{n\to\infty}{\overset{p}{\to}} 2\mathbb{E}\left[e_i^2\right] = 2\sigma^2.$$

Thus the estimator for σ^2 whose consistency is proved by previous manipulations is

$$\hat{\sigma}^2 = \frac{1}{2}\frac{1}{n-1}\sum_{i=2}^{n}(y_i-y_{i-1})^2.$$

2. At the first step estimate β from the FD-regression. The FD-transformed regression is

$$y_i - y_{i-1} = (x_i - x_{i-1})'\beta + g(z_i) - g(z_{i-1}) + e_i - e_{i-1},$$

which can be rewritten as

$$\Delta y_i = \Delta x_i'\beta + \Delta g(z_i) + \Delta e_i.$$

The consistency of the following estimator for β

$$\hat{\beta} = \left(\sum_{i=2}^{n}\Delta x_i \Delta x_i'\right)^{-1}\left(\sum_{i=2}^{n}\Delta x_i \Delta y_i\right)$$

can be proved in the standard way:

$$\hat{\beta}-\beta = \left(\frac{1}{n-1}\sum_{i=2}^{n}\Delta x_i \Delta x_i'\right)^{-1}\left(\frac{1}{n-1}\sum_{i=2}^{n}\Delta x_i(\Delta g(z_i)+\Delta e_i)\right)$$

Here $\frac{1}{n-1}\sum_{i=2}^{n}\Delta x_i \Delta x_i'$ has some non-zero probability limit, $\frac{1}{n-1}\sum_{i=2}^{n}\Delta x_i \Delta e_i \underset{n\to\infty}{\overset{p}{\to}} 0$ since $\mathbb{E}[e_i|x_i,z_i]=0$, and $\left|\frac{1}{n-1}\sum_{i=2}^{n}\Delta x_i \Delta g(z_i)\right| \le \frac{G}{n-1}\frac{1}{n-1}\sum_{i=2}^{n}|\Delta x_i| \underset{n\to\infty}{\overset{p}{\to}} 0$. Now we can use standard nonparametric tools for the "regression"

$$y_i - x_i'\hat{\beta} = g(z_i) + e_i^*,$$

where $e_i^* = e_i + x_i'(\beta-\hat{\beta})$. Consider the following estimator (we use the uniform kernel for algebraic simplicity):

$$\widehat{g(z)} = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})\mathbb{I}\left[|z_i-z|\le h\right]}{\sum_{i=1}^{n}\mathbb{I}\left[|z_i-z|\le h\right]}.$$

It can be decomposed into three terms:

$$\widehat{g(z)} = \frac{\sum_{i=1}^{n}\left(g(z_i)+x_i'(\beta-\hat{\beta})+e_i\right)\mathbb{I}\left[|z_i-z|\le h\right]}{\sum_{i=1}^{n}\mathbb{I}\left[|z_i-z|\le h\right]}$$

The first term gives $g(z)$ in the limit. To show this, use Lipschitz condition:

$$\left| \frac{\sum_{i=1}^n (g(z_i) - g(z))\mathbb{I}\left[|z_i - z| \le h\right]}{\sum_{i=1}^n \mathbb{I}\left[|z_i - z| \le h\right]} \right| \le Gh,$$

and introduce the asymptotics for the smoothing parameter: $h \to 0$. Then

$$\begin{aligned}\frac{\sum_{i=1}^n g(z_i)\mathbb{I}\left[|z_i - z| \le h\right]}{\sum_{i=1}^n \mathbb{I}\left[|z_i - z| \le h\right]} &= \frac{\sum_{i=1}^n (g(z) + g(z_i) - g(z))\mathbb{I}\left[|z_i - z| \le h\right]}{\sum_{i=1}^n \mathbb{I}\left[|z_i - z| \le h\right]} = \\ &= g(z) + \frac{\sum_{i=1}^n (g(z_i) - g(z))\mathbb{I}\left[|z_i - z| \le h\right]}{\sum_{i=1}^n \mathbb{I}\left[|z_i - z| \le h\right]} \underset{n\to\infty}{\to} g(z).\end{aligned}$$

The second and the third terms have zero probability limit if the condition $nh \to \infty$ is satisfied

$$\underbrace{\frac{\sum_{i=1}^n x_i'\mathbb{I}\left[|z_i - z| \le h\right]}{\sum_{i=1}^n \mathbb{I}\left[|z_i - z| \le h\right]}}_{\substack{\downarrow^p \\ \mathbb{E}\left[x_i'\right]}} \underbrace{(\beta - \hat{\beta})}_{\substack{\downarrow^p \\ 0}} \underset{n\to\infty}{\overset{p}{\to}} 0$$

and

$$\frac{\sum_{i=1}^n e_i\mathbb{I}\left[|z_i - z| \le h\right]}{\sum_{i=1}^n \mathbb{I}\left[|z_i - z| \le h\right]} \underset{n\to\infty}{\overset{p}{\to}} \mathbb{E}\left[e_i\right] = 0.$$

Therefore, $\widehat{g(z)}$ is consistent when $n \to \infty$, $nh \to \infty$, $h \to 0$.

11. CONDITIONAL MOMENT RESTRICTIONS

11.1 Usefulness of skedastic function

Denote $\theta = \binom{\beta}{\pi}$ and $e = y - x'\beta$. The moment function is

$$m(y,x,\theta) = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} y - x'\beta \\ (y-x'\beta)^2 - h(x,\beta,\pi) \end{pmatrix}$$

The general theory for the conditional moment restriction $\mathbb{E}\left[m(w,\theta)|x\right] = 0$ gives the optimal restriction $\mathbb{E}\left[D(x)'\Omega(x)^{-1}m(w,\theta)\right] = 0$, where $D(x) = \mathbb{E}\left[\frac{\partial m}{\partial \theta'}|x\right]$ and $\Omega(x) = \mathbb{E}[mm'|x]$. The variance of the optimal estimator is $V = \left(\mathbb{E}\left[D(x)'\Omega(x)^{-1}D(x)\right]\right)^{-1}$. For the problem at hand,

$$D(x) = \mathbb{E}\left[\frac{\partial m}{\partial \theta'}|x\right] = -\mathbb{E}\left[\begin{pmatrix} x' & 0 \\ 2ex' + h'_\beta & h'_\pi \end{pmatrix}|x\right] = -\begin{pmatrix} x' & 0 \\ h'_\beta & h'_\pi \end{pmatrix},$$

$$\Omega(x) = \mathbb{E}[mm'|x] = \mathbb{E}\left[\begin{pmatrix} e^2 & e(e^2-h) \\ e(e^2-h) & (e^2-h)^2 \end{pmatrix}|x\right] = \mathbb{E}\left[\begin{pmatrix} e^2 & e^3 \\ e^3 & (e^2-h)^2 \end{pmatrix}|x\right],$$

since $\mathbb{E}[ex|x] = 0$ and $\mathbb{E}[eh|x] = 0$.

Let $\Delta(x) \equiv \det \Omega(x) = \mathbb{E}[e^2|x]\mathbb{E}[(e^2-h)^2|x] - (\mathbb{E}[e^3|x])^2$. The inverse of Ω is

$$\Omega(x)^{-1} = \frac{1}{\Delta(x)}\mathbb{E}\left[\begin{pmatrix} (e^2-h)^2 & -e^3 \\ -e^3 & e^2 \end{pmatrix}|x\right],$$

and the asymptotic variance of the efficient GMM estimator is

$$V^{-1} = \mathbb{E}\left[D(x)'\Omega(x)^{-1}D(x)\right] = \begin{pmatrix} A & B' \\ B & C \end{pmatrix},$$

where

$$\begin{aligned} A &= \mathbb{E}\left[\frac{(e^2-h)^2xx' - e^3(xh'_\beta + h_\beta x') + e^2h_\beta h'_\beta}{\Delta(x)}\right], \\ B &= \mathbb{E}\left[\frac{-e^3h_\pi x' + e^2h_\pi h'_\beta}{\Delta(x)}\right], \quad C = \mathbb{E}\left[\frac{e^2h_\pi h'_\pi}{\Delta(x)}\right]. \end{aligned}$$

Using the formula for inversion of the partitioned matrices, find that

$$V = \begin{pmatrix} (A - B'C^{-1}B)^{-1} & * \\ * & * \end{pmatrix},$$

where $*$ denote submatrices which are not of interest.

To answer the problem we need to compare $V_{11} = (A - B'C^{-1}B)^{-1}$ with $V_0 = \left(\mathbb{E}\left[\frac{xx'}{h}\right]\right)^{-1}$, the variance of the optimal GMM estimator constructed with the use of m_1 only. We need to show that $V_{11} \le V_0$, or, alternatively, $V_{11}^{-1} \ge V_0^{-1}$. Note that

$$V_{11}^{-1} - V_0^{-1} = \tilde{A} - B'C^{-1}B,$$

where $\tilde{A} = A - V_0^{-1}$ can be simplified to

$$\tilde{A} = \mathbb{E}\left[\frac{1}{\Delta(x)}\left(\frac{xx'\left(\mathbb{E}\left[e^3|x\right]\right)^2}{\mathbb{E}\left[e^2|x\right]} - e^3(xh'_\beta + h_\beta x') + e^2 h_\beta h'_\beta\right)\right].$$

Next, we can use the following representation:

$$\tilde{A} - B'C^{-1}B = \mathbb{E}[ww'],$$

where

$$w = \frac{\mathbb{E}\left[e^3|x\right]x - \mathbb{E}\left[e^2|x\right]h_\beta}{\sqrt{\mathbb{E}\left[e^2|x\right]}\sqrt{\Delta(x)}} + B'C^{-1}h_\pi\sqrt{\frac{\mathbb{E}\left[e^2|x\right]}{\Delta(x)}}.$$

This representation concludes that $V_{11}^{-1} \geq V_0^{-1}$ and gives the condition under which $V_{11} = V_0$. This condition is $w(x) = 0$ almost surely. It can be written as

$$\frac{\mathbb{E}\left[e^3|x\right]}{\mathbb{E}\left[e^2|x\right]}x = h_\beta - B'C^{-1}h_\pi \text{ almost surely.}$$

Consider the special cases.

1. $h_\beta = 0$. Then the condition modifies to

$$\frac{\mathbb{E}\left[e^3|x\right]}{\mathbb{E}\left[e^2|x\right]}x = -\mathbb{E}\left[\frac{e^3 h_\pi x'}{\Delta(x)}\right]\mathbb{E}\left[\frac{e^2 h_\pi h'_\pi}{\Delta(x)}\right]^{-1} h_\pi \text{ almost surely.}$$

2. $h_\beta = 0$ and the distribution of e_i conditional on x_i is symmetric. The previous condition is satisfied automatically since $\mathbb{E}\left[e^3|x\right] = 0$.

11.2 Symmetric regression error

Part 1. The maintained hypothesis is $\mathbb{E}\left[e|x\right] = 0$. We can use the null hypothesis $H_0 : \mathbb{E}\left[e^3|x\right] = 0$ to test for the conditional symmetry. We could in addition use more conditional moment restrictions (e.g., involving higher odd powers) to increase the power of the test, but in finite samples that would probably lead to more distorted test sizes. The alternative hypothesis is $H_1 : \mathbb{E}\left[e^3|x\right] \neq 0$.

An estimator that is consistent under both H_0 and H_1 is, for example, the OLS estimator $\hat{\alpha}_{OLS}$. The estimator that is consistent and asymptotically efficient (in the same class where $\hat{\alpha}_{OLS}$ belongs) under H_0 and (hopefully) inconsistent under H_1 is the instrumental variables (GMM) estimator $\hat{\alpha}_{OIV}$ that uses the optimal instrument for the system $\mathbb{E}\left[e|x\right] = 0$, $\mathbb{E}\left[e^3|x\right] = 0$. We derived in class that the optimal unconditional moment restriction is

$$\mathbb{E}\left[a_1(x)\left(y - \alpha x\right) + a_2(x)\left(y - \alpha x\right)^3\right] = 0,$$

where

$$\begin{pmatrix} a_1(x) \\ a_2(x) \end{pmatrix} = \frac{x}{\mu_2(x)\mu_6(x) - \mu_4(x)^2}\begin{pmatrix} \mu_6(x) - 3\mu_2(x)\mu_4(x) \\ 3\mu_2(x)^2 - \mu_4(x) \end{pmatrix}$$

and $\mu_r(x) = \mathbb{E}\left[(y - \alpha x)^r|x\right]$, $r = 2, 4, 6$. To construct a feasible $\hat{\alpha}_{OIV}$, one needs to first estimate $\mu_r(x)$ at the points x_i of the sample. This may be done nonparametrically using nearest neighbor, series expansion or other approaches. Denote the resulting estimates by $\hat{\mu}_r(x_i)$, $i = 1, \cdots, n$, $r = 2, 4, 6$ and compute $\hat{a}_1(x_i)$ and $\hat{a}_2(x_i)$, $i = 1, \cdots, n$. Then $\hat{\alpha}_{OIV}$ is a solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{a}_1(x_i)\left(y_i - \hat{\alpha}_{OIV}x_i\right) + \hat{a}_2(x_i)\left(y_i - \hat{\alpha}_{OIV}x_i\right)^3\right) = 0,$$

which can be turned into an optimization problem, if convenient.

The Hausman test statistic is then

$$H = n\frac{\left(\hat{\alpha}_{OLS} - \hat{\alpha}_{OIV}\right)^2}{\hat{V}_{OLS} - \hat{V}_{OIV}} \xrightarrow{d} \chi^2(1),$$

where $\hat{V}_{OLS} = n\left(\sum_{i=1}^{n} x_i^2\right)^{-2}\sum_{i=1}^{n} x_i^2\left(y_i - \hat{\alpha}_{OLS}x_i\right)^2$ and $\hat{V}_{OIV}$ is a consistent estimate of the efficiency bound

$$V_{OIV} = \left(\mathbb{E}\left[\frac{x_i^2(\mu_6(x_i) - 6\mu_2(x_i)\mu_4(x_i)) + 9\mu_2^3(x_i))}{\mu_2(x_i)\mu_6(x_i)) - \mu_4^2(x_i)}\right]\right)^{-1}.$$

Note that the constructed Hausman test will not work if $\hat{\alpha}_{OLS}$ is also asymptotically efficient, which may happen if the third-moment restriction is redundant and the error is conditionally homoskedastic so that the optimal instrument reduces to the one implied by OLS. Also, the test may be inconsistent (i.e., asymptotically have power less than 1) if $\hat{\alpha}_{OIV}$ happens to be consistent under conditional non-symmetry too.

Part 2. Under the assumption that $e|x \sim \mathcal{N}(0, \sigma^2)$, irrespective of whether σ^2 is known or not, the QML estimator $\hat{\alpha}_{QML}$ coincides with the OLS estimator and thus has the same asymptotic distribution

$$\sqrt{n}\left(\hat{\alpha}_{QML} - \alpha\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[x^2\left(y - \alpha x\right)^2\right]}{\left(\mathbb{E}\left[x^2\right]\right)^2}\right).$$

11.3 Optimal instrument in AR-ARCH model

Let us for convenience view a typical element of $\mathcal{Z}_t$ as $\sum_{i=1}^{\infty}\omega_i\varepsilon_{t-i}$, and let the optimal instrument be $\zeta_t = \sum_{i=1}^{\infty} a_i\varepsilon_{t-i}$.The optimality condition is

$$\mathbb{E}[v_t x_{t-1}] = \mathbb{E}\left[v_t\zeta_t\varepsilon_t^2\right] \quad \text{for all } v_t \in \mathcal{Z}_t.$$

Since it should hold for any $v_t \in \mathcal{Z}_t$, let us make it hold for $v_t = \varepsilon_{t-j}$, $j = 1, 2, \cdots$. Then we get a system of equations of the type

$$\mathbb{E}[\varepsilon_{t-j}x_{t-1}] = \mathbb{E}\left[\varepsilon_{t-j}\left(\sum_{i=1}^{\infty} a_i\varepsilon_{t-i}\right)\zeta_t\varepsilon_t^2\right].$$

The left-hand side is just ρ^{j-1} because $x_{t-1} = \sum_{i=1}^{\infty}\rho^{i-1}\varepsilon_{t-i}$ and because $\mathbb{E}[\varepsilon_t^2] = 1$. In the right-hand side, all terms are zeros due to conditional symmetry of ε_t, except $a_j\mathbb{E}\left[\varepsilon_{t-j}^2\varepsilon_t^2\right]$. Therefore,

$$a_j = \frac{\rho^{j-1}}{1 + \alpha^j(\kappa - 1)},$$

where $\kappa = \mathbb{E}\left[\varepsilon_t^4\right]$. This follows from the $ARCH(1)$ structure:

$$\mathbb{E}\left[\varepsilon_{t-j}^2\varepsilon_t^2\right] = \mathbb{E}\left[\varepsilon_{t-j}^2\mathbb{E}[\varepsilon_t^2|I_{t-1}]\right] = \mathbb{E}\left[\varepsilon_{t-j}^2\left((1-\alpha)+\alpha\varepsilon_{t-1}^2\right)\right] = (1-\alpha)+\alpha\mathbb{E}\left[\varepsilon_{t-j+1}^2\varepsilon_t^2\right],$$

so that we can recursively obtain

$$\mathbb{E}\left[\varepsilon_{t-j}^2\varepsilon_t^2\right] = 1-\alpha^j+\alpha^j\kappa.$$

Thus the optimal instrument is

$$\begin{aligned}\zeta_t &= \sum_{i=1}^{\infty}\frac{\rho^{i-1}}{1+\alpha^i(\kappa-1)}\varepsilon_{t-i} = \\ &= \frac{x_{t-1}}{1+\alpha(\kappa-1)}+(\kappa-1)(1-\alpha)\sum_{i=2}^{\infty}\frac{(\alpha\rho)^{i-1}}{\left(1+\alpha^i(\kappa-1)\right)\left(1+\alpha^{i-1}(\kappa-1)\right)}x_{t-i}.\end{aligned}$$

To construct a feasible estimator, set $\hat{\rho}$ to be the OLS estimator of ρ, $\hat{\alpha}$ to be the OLS estimator of α in the model $\hat{\varepsilon}_t^2-1=\alpha(\hat{\varepsilon}_{t-1}^2-1)+\upsilon_t$, and compute $\hat{\kappa}=T^{-1}\sum_{t=2}^{T}\hat{\varepsilon}_t^4$.

The optimal instrument based on $\mathbb{E}[\varepsilon_t|I_{t-1}]=0$ uses a large set of allowable instruments, relative to which our $\mathcal{Z}_t$ is extremely thin. Therefore, we can expect big losses in efficiency in comparison with what we could get. In fact, calculations for empirically relevant sets of parameter values reveal that this intuition is correct. Weighting by the skedastic function is much more powerful than trying to capture heteroskedasticity by using an infinite history of the basic instrument in a linear fashion.

11.4 Modified Poisson regression and PML estimators

Part 1. The mean regression function is $\mathbb{E}[y|x]=\mathbb{E}[\mathbb{E}[y|x,\varepsilon]|x]=\mathbb{E}[\exp(x'\beta+\varepsilon)|x]=\exp(x'\beta)$. The skedastic function is $\mathbb{V}[y|x]=\mathbb{E}[(y-\mathbb{E}[y|x])^2|x]=\mathbb{E}[y^2|x]-\mathbb{E}[y|x]^2$. Since

$$\begin{aligned}\mathbb{E}\left[y^2|x\right] &= \mathbb{E}\left[\mathbb{E}\left[y^2|x,\varepsilon\right]|x\right]=\mathbb{E}\left[\exp(2x'\beta+2\varepsilon)+\exp(x'\beta+\varepsilon)|x\right] \\ &= \exp(2x'\beta)\mathbb{E}\left[(\exp\varepsilon)^2|x\right]+\exp(x'\beta)=(\sigma^2+1)\exp(2x'\beta)+\exp(x'\beta),\end{aligned}$$

we have $\mathbb{V}[y|x]=\sigma^2\exp(2x'\beta)+\exp(x'\beta)$.

Part 2. Use the formula for asymptotic variance of NLLS estimator:

$$V_{NLLS}=Q_{gg}^{-1}Q_{gge^2}Q_{gg}^{-1},$$

where $Q_{gg}=\mathbb{E}\left[\frac{\partial g(x,\beta)}{\partial\beta}\frac{\partial g(x,\beta)}{\partial\beta'}\right]$ and $Q_{gge^2}=\mathbb{E}\left[\frac{\partial g(x,\beta)}{\partial\beta}\frac{\partial g(x,\beta)}{\partial\beta'}(y-g(x,\beta))^2\right]$. In our problem $g(x,\beta)=\exp(x'\beta)$ and $Q_{gg}=\mathbb{E}\left[xx'\exp(2x'\beta)\right]$,

$$\begin{aligned}Q_{gge^2} &= \mathbb{E}\left[xx'\exp(2x'\beta)(y-\exp(x'\beta))^2\right]=\mathbb{E}\left[xx'\exp(2x'\beta)\mathbb{V}[y|x]\right] \\ &= \mathbb{E}\left[xx'\exp(2x'\beta)(\sigma^2\exp(2x'\beta)+\exp(x'\beta))\right]=\mathbb{E}\left[xx'\exp(3x'\beta)\right]+\sigma^2\mathbb{E}\left[xx'\exp(4x'\beta)\right].\end{aligned}$$

To find the expectations we use the formula $\mathbb{E}[xx'\exp(nx'\beta)]=\exp(\frac{n^2}{2}\beta'\beta)(I+n^2\beta\beta')$. Now, we have $Q_{gg}=\exp(2\beta'\beta)(I+4\beta\beta')$ and $Q_{gge^2}=\exp(\frac{9}{2}\beta'\beta)(I+9\beta\beta')+\sigma^2\exp(8\beta'\beta)(I+16\beta\beta')$. Finally,

$$V_{NLLS}=(I+4\beta\beta')^{-1}\left(\exp(\frac{1}{2}\beta'\beta)(I+9\beta\beta')+\sigma^2\exp(4\beta'\beta)(I+16\beta\beta')\right)(I+4\beta\beta')^{-1}.$$

The formula for asymptotic variance of WNLLS estimator is

$$V_{WNLLS} = Q_{gg/\sigma^2}^{-1},$$

where $Q_{gg/\sigma^2} = \mathbb{E}\left[\frac{\partial g(x,\beta)}{\partial\beta}\frac{\partial g(x,\beta)}{\partial\beta'}\frac{1}{\mathbb{V}[y|x]}\right]$. In this problem

$$Q_{gg/\sigma^2} = \mathbb{E}\left[xx'\exp(2x'\beta)(\sigma^2\exp(2x'\beta)+\exp(x'\beta))^{-1}\right],$$

which can be rearranged as

$$V_{WNLLS} = \sigma^2\left(I - \mathbb{E}\left[\frac{xx'}{1+\sigma^2\exp(x'\beta)}\right]\right)^{-1}.$$

Part 3. We use the formula for asymptotic variance of PML estimator:

$$V_{PML} = \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1},$$

where

$$\begin{aligned}
\mathcal{J} &= \mathbb{E}\left[\left.\frac{\partial C}{\partial m}\right|_{m(x,\beta_0)}\frac{\partial m(x,\beta_0)}{\partial\beta}\frac{\partial m(x,\beta_0)}{\partial\beta'}\right],\\
\mathcal{I} &= \mathbb{E}\left[\left(\left.\frac{\partial C}{\partial m}\right|_{m(x,\beta_0)}\right)^2\sigma^2(x,\beta_0)\frac{\partial m(x,\beta_0)}{\partial\beta}\frac{\partial m(x,\beta_0)}{\partial\beta'}\right].
\end{aligned}$$

In this problem $m(x,\beta) = \exp(x'\beta)$ and $\sigma^2(x,\beta) = \sigma^2\exp(2x'\beta)+\exp(x'\beta)$.

(a) For the normal distribution $C(m) = m$, therefore $\frac{\partial C}{\partial m} = 1$ and $V_{NPML} = V_{NLLS}$.

(b) For the Poisson distribution $C(m) = \log m$, therefore $\frac{\partial C}{\partial m} = \frac{1}{m}$,

$$\begin{aligned}
\mathcal{J} &= \mathbb{E}[\exp(-x'\beta)xx'\exp(2x'\beta)] = \exp(\tfrac{1}{2}\beta'\beta)(I+\beta\beta'),\\
\mathcal{I} &= \mathbb{E}[\exp(-2x'\beta)(\sigma^2\exp(2x'\beta)+\exp(x'\beta))xx'\exp(2x'\beta)]\\
&= \exp(\tfrac{1}{2}\beta'\beta)(I+\beta\beta') + \sigma^2\exp(2\beta'\beta)(I+4\beta\beta').
\end{aligned}$$

Finally,

$$V_{PPML} = (I+\beta\beta')^{-1}\left(\exp(-\frac{1}{2}\beta'\beta)(I+\beta\beta') + \sigma^2\exp(\beta'\beta)(I+4\beta\beta')\right)(I+\beta\beta')^{-1}.$$

(c) For the Gamma distribution $C(m) = -\frac{\alpha}{m}$, therefore $\frac{\partial C}{\partial m} = \frac{\alpha}{m^2}$,

$$\begin{aligned}
\mathcal{J} &= \mathbb{E}[\alpha\exp(-2x'\beta)xx'\exp(2x'\beta)] = \alpha I,\\
\mathcal{I} &= \alpha^2\mathbb{E}[\exp(-4x'\beta)(\sigma^2\exp(2x'\beta)+\exp(x'\beta))xx'\exp(2x'\beta)]\\
&= \alpha^2\sigma^2 I + \alpha^2\exp(\tfrac{1}{2}\beta'\beta)(I+\beta\beta').
\end{aligned}$$

Finally,

$$V_{GPML} = \sigma^2 I + \exp(\frac{1}{2}\beta'\beta)(I+\beta\beta').$$

Part 4. We have the following variances:

$$
\begin{aligned}
V_{NLLS} &= (I+4\beta\beta')^{-1}\left(\exp(\frac{1}{2}\beta'\beta)(I+9\beta\beta')+\sigma^2\exp(4\beta'\beta)(I+16\beta\beta')\right)(I+4\beta\beta')^{-1},\\
V_{WNLLS} &= \sigma^2\left(I-\mathbb{E}\frac{xx'}{1+\sigma^2\exp(x'\beta)}\right)^{-1},\\
V_{NPML} &= V_{NLLS},\\
V_{PPML} &= (I+\beta\beta')^{-1}\left(\exp(-\frac{1}{2}\beta'\beta)(I+\beta\beta')+\sigma^2\exp(\beta'\beta)(I+4\beta\beta')\right)(I+\beta\beta')^{-1},\\
V_{GPML} &= \sigma^2 I+\exp(\frac{1}{2}\beta'\beta)(I+\beta\beta').
\end{aligned}
$$

From the theory we know that $V_{WNLLS}\le V_{NLLS}$. Next, we know that in the class of PML estimators the efficiency bound is achieved when $\left.\frac{\partial C}{\partial m}\right|_{m(x,\beta_0)}$ is proportional to $\frac{1}{\sigma^2(x,\beta_0)}$, then the bound is

$$
\mathbb{E}\left[\frac{\partial m(x,\beta)}{\partial\beta}\frac{\partial m(x,\beta)}{\partial\beta'}\frac{1}{\mathbb{V}[y|x]}\right]
$$

which is equal to V_{WNLLS} in our case. So, we have $V_{WNLLS}\le V_{PPML}$ and $V_{WNLLS}\le V_{GPML}$. The comparison of other variances is not straightforward. Consider the one-dimensional case. Then we have

$$
\begin{aligned}
V_{NLLS} &= \frac{e^{\beta^2/2}(1+9\beta^2)+\sigma^2e^{4\beta^2}(1+16\beta^2)}{(1+4\beta^2)^2},\\
V_{WNLLS} &= \sigma^2\left(1-\mathbb{E}\frac{x^2}{1+\sigma^2\exp(x\beta)}\right)^{-1},\\
V_{NPML} &= V_{NLLS},\\
V_{PPML} &= \frac{e^{\beta^2/2}(1+\beta^2)+\sigma^2e^{\beta^2}(1+4\beta^2)}{(1+\beta^2)^2},\\
V_{GPML} &= \sigma^2+e^{\beta^2/2}(1+\beta^2).
\end{aligned}
$$

We can calculate these (except V_{WNLLS}) for various parameter sets. For example, for $\sigma^2=0.01$ and $\beta^2=0.4$ $V_{NLLS}<V_{PPML}<V_{GPML}$, for $\sigma^2=0.01$ and $\beta^2=0.1$ $V_{PPML}<V_{NLLS}<V_{GPML}$, for $\sigma^2=1$ and $\beta^2=0.4$ $V_{GPML}<V_{PPML}<V_{NLLS}$, for $\sigma^2=0.5$ and $\beta^2=0.4$ $V_{PPML}<V_{GPML}<V_{NLLS}$. However, it appears impossible to make $V_{NLLS}<V_{GPML}<V_{PPML}$ or $V_{GPML}<V_{NLLS}<V_{PPML}$.

11.5 Optimal instrument and regression on constant

Part 1. We have the following moment function: $m(x,y,\theta)=(y-\alpha,\quad (y-\alpha)^2-\sigma^2x_i^2)'$ with $\theta=\binom{\alpha}{\sigma^2}$. The optimal unconditional moment restriction is $\mathbb{E}[A^*(x_i)m(x,y,\theta)]=0$, where $A^*(x_i)=D'(x_i)\Omega(x_i)^{-1}$, $D(x_i)=\mathbb{E}\left[\partial m(x,y,\theta)/\partial\theta'|x_i\right]$, $\Omega(x_i)=\mathbb{E}[m(x,y,\theta)m(x,y,\theta)'|x_i]$.

(a) For the first moment restriction $m_1(x,y,\theta)=y-\alpha$ we have $D(x_i)=-1$ and $\Omega(x_i)=\mathbb{E}\left[(y-\alpha)^2|x_i\right]=\sigma^2x_i^2$, therefore the optimal moment restriction is

$$
\mathbb{E}\left[\frac{y_i-\alpha}{x_i^2}\right]=0.
$$

(b) For the moment function $m(x,y,\theta)$ we have

$$D(x_i)=\begin{pmatrix}-1 & 0\\ 0 & -x_i^2\end{pmatrix},\quad \Omega(x_i)=\begin{pmatrix}\sigma^2x_i^2 & 0\\ 0 & \mu_4(x_i)-x_i^4\sigma^4\end{pmatrix},$$

where $\mu_4(x)=\mathbb{E}\left[(y-\alpha)^4|x\right]$. The optimal weighting matrix is

$$A^*(x_i)=\begin{pmatrix}\dfrac{1}{\sigma^2x_i^2} & 0\\ 0 & \dfrac{x_i^2}{\mu_4(x_i)-x_i^4\sigma^4}\end{pmatrix}.$$

The optimal moment restriction is

$$\mathbb{E}\left[\begin{pmatrix}\dfrac{y_i-\alpha}{x_i^2}\\ \dfrac{(y_i-\alpha)^2-\sigma^2x_i^2}{\mu_4(x_i)-x_i^4\sigma^4}x_i^2\end{pmatrix}\right]=0.$$

Part 2. **(a)** The GMM estimator is the solution of

$$\frac{1}{n}\sum_i\frac{y_i-\hat{\alpha}}{x_i^2}=0\quad\Rightarrow\quad\hat{\alpha}=\sum_i\frac{y_i}{x_i^2}\Bigg/\sum_i\frac{1}{x_i^2}.$$

The estimator for σ^2 can be drawn from the sample analog of the condition $\mathbb{E}\left[(y-\alpha)^2\right]=\sigma^2\mathbb{E}\left[x^2\right]$:

$$\tilde{\sigma}^2=\sum_i(y_i-\hat{\alpha})^2\Bigg/\sum_i x_i^2.$$

(b) The GMM estimator is the solution of

$$\sum_i\begin{pmatrix}\dfrac{y_i-\hat{\alpha}}{x_i^2}\\ \dfrac{(y_i-\hat{\alpha})^2-\hat{\sigma}^2x_i^2}{\hat{\mu}_4(x_i)-x_i^4\hat{\sigma}^4}x_i^2\end{pmatrix}=0.$$

We have the same estimator for α:

$$\hat{\alpha}=\sum_i\frac{y_i}{x_i^2}\Bigg/\sum_i\frac{1}{x_i^2},$$

$\hat{\sigma}^2$ is the solution of

$$\sum_i\frac{(y_i-\hat{\alpha})^2-\hat{\sigma}^2x_i^2}{\hat{\mu}_4(x_i)-x_i^4\hat{\sigma}^4}x_i^2=0,$$

where $\hat{\mu}_4(x_i)$ is non-parametric estimator for $\mu_4(x_i)=\mathbb{E}\left[(y-\alpha)^4|x_i\right]$, for example, a nearest neighbor or a series estimator.

Part 3. The general formula for the variance of the optimal estimator is

$$V=\left(\mathbb{E}\left[D'(x_i)\Omega(x_i)^{-1}D(x_i)\right]\right)^{-1}.$$

(a) $V_{\hat{\alpha}} = \sigma^2 \left(\mathbb{E}\left[x_i^{-2}\right]\right)^{-1}$. Use standard asymptotic techniques to find

$$V_{\tilde{\sigma}^2} = \frac{\mathbb{E}\left[(y_i - \alpha)^4\right]}{\left(\mathbb{E}\left[x_i^2\right]\right)^2} - \sigma^4.$$

(b)

$$V_{(\hat{\alpha},\hat{\sigma}^2)} = \left(\mathbb{E}\left[\begin{pmatrix} \frac{1}{\sigma^2 x_i^2} & 0 \\ 0 & \frac{x_i^4}{\mu_4(x_i) - x_i^4\sigma^4} \end{pmatrix}\right]\right)^{-1} = \begin{pmatrix} \sigma^2 \left(\mathbb{E}\left[x_i^{-2}\right]\right)^{-1} & 0 \\ 0 & \left(\mathbb{E}\left[\frac{x_i^4}{\mu_4(x_i) - x_i^4\sigma^4}\right]\right)^{-1} \end{pmatrix}.$$

When we use the optimal instrument, our estimator is more efficient, therefore $V_{\tilde{\sigma}^2} > V_{\hat{\sigma}^2}$.
Estimators of asymptotic variance can be found through sample analogs:

$$\hat{V}_{\hat{\alpha}} = \hat{\sigma}^2 \left(\frac{1}{n}\sum_i \frac{1}{x_i^2}\right)^{-1}, \quad V_{\tilde{\sigma}^2} = n\frac{\sum_i (y_i - \hat{\alpha})^4}{\left(\sum_i x_i^2\right)^2} - \tilde{\sigma}^4, \quad V_{\hat{\sigma}^2} = n\left(\sum_i \frac{x_i^4}{\hat{\mu}_4(x_i) - x_i^4\hat{\sigma}^4}\right)^{-1}.$$

Part 4. The normal distribution PML2 estimator is the solution of the following problem:

$$\widehat{\begin{pmatrix} \alpha \\ \sigma^2 \end{pmatrix}}_{PML2} = \arg\max_{\alpha,\sigma^2} \left\{\text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{\sigma^2}\sum_i \frac{(y_i - \alpha)^2}{2x_i^2}\right\}.$$

Solving gives

$$\hat{\alpha}_{PML2} = \hat{\alpha} = \sum_i \frac{y_i}{x_i^2} \Big/ \sum_i \frac{1}{x_i^2}, \quad \hat{\sigma}^2_{PML2} = \frac{1}{n}\sum_i \frac{(y_i - \hat{\alpha})^2}{x_i^2}$$

Since we have the same estimator for α, we have the same variance $V_{\hat{\alpha}} = \sigma^2 \left(\mathbb{E}\left[x_i^{-2}\right]\right)^{-1}$. It can be shown that

$$V_{\hat{\sigma}^2} = \mathbb{E}\left[\frac{\mu_4(x_i)}{x_i^4}\right] - \sigma^4.$$

12. EMPIRICAL LIKELIHOOD

12.1 Common mean

Part 1. We have the following moment function: $m(x, y, \theta) = \binom{x-\theta}{y-\theta}$. The MEL estimator is the solution of the following optimization problem.

$$\sum_i \log p_i \to \max_{p_i, \theta}$$

subject to

$$\sum_i p_i m(x_i, y_i, \theta) = 0, \quad \sum_i p_i = 1.$$

Let λ be a Lagrange multiplier for the restriction $\sum_i p_i m(x_i, y_i, \theta) = 0$, then the solution of the problem satisfies

$$\begin{aligned} p_i &= \frac{1}{n} \frac{1}{1 + \lambda' m(x_i, y_i, \theta)}, \\ 0 &= \frac{1}{n} \sum_i \frac{1}{1 + \lambda' m(x_i, y_i, \theta)} m(x_i, y_i, \theta), \\ 0 &= \frac{1}{n} \sum_i \frac{1}{1 + \lambda' m(x_i, y_i, \theta)} \left(\frac{\partial m(x_i, y_i, \theta)}{\partial \theta'} \right)' \lambda. \end{aligned}$$

In our case, $\lambda = \binom{\lambda_1}{\lambda_2}$ and the system is

$$\begin{aligned} p_i &= \frac{1}{1 + \lambda_1 (x_i - \theta) + \lambda_2 (y_i - \theta)}, \\ 0 &= \frac{1}{n} \sum_i \frac{1}{1 + \lambda_1 (x_i - \theta) + \lambda_2 (y_i - \theta)} \begin{pmatrix} x_i - \theta \\ y_i - \theta \end{pmatrix}, \\ 0 &= \frac{1}{n} \sum_i \frac{-\lambda_1 - \lambda_2}{1 + \lambda_1 (x_i - \theta) + \lambda_2 (y_i - \theta)}. \end{aligned}$$

The asymptotic distribution of the estimators is

$$\sqrt{n}(\hat{\theta}_{el} - \theta) \xrightarrow{d} N(0, V), \quad \sqrt{n} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, U),$$

where $V = \left(Q'_{\partial m} Q_{mm}^{-1} Q_{\partial m} \right)^{-1}$, $U = Q_{mm}^{-1} - Q_{mm}^{-1} Q_{\partial m} V Q'_{\partial m} Q_{mm}^{-1}$. In our case $Q_{\partial m} = \binom{-1}{-1}$ and $Q_{mm} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$, therefore

$$V = \frac{\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2}{\sigma_y^2 + \sigma_x^2 - 2\sigma_{xy}}, \quad U = \frac{1}{\sigma_y^2 + \sigma_x^2 - 2\sigma_{xy}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Estimators for V and U based on consistent estimators for σ_x^2, σ_y^2 and σ_{xy} can be constructed from sample moments.

Part 2. The last equation of the system gives $\lambda_1 = -\lambda_2 = \lambda$, so we have

$$p_i = \frac{1}{1+\lambda(x_i - y_i)}, \quad 0 = \frac{1}{n}\sum_i \frac{1}{1+\lambda(x_i-y_i)}\begin{pmatrix} x_i - \theta \\ y_i - \theta \end{pmatrix}.$$

The MEL estimator is

$$\hat{\theta}_{MEL} = \sum_i \frac{x_i}{1+\lambda(x_i-y_i)} \Bigg/ \sum_i \frac{1}{1+\lambda(x_i-y_i)} = \sum_i \frac{y_i}{1+\lambda(x_i-y_i)} \Bigg/ \sum_i \frac{1}{1+\lambda(x_i-y_i)},$$

where λ is the solution of

$$\sum_i \frac{x_i - y_i}{1+\lambda(x_i-y_i)} = 0.$$

Linearization with respect to λ around 0 gives

$$p_i = 1 - \lambda(x_i - y_i), \quad 0 = \frac{1}{n}\sum_i (1-\lambda(x_i-y_i))\begin{pmatrix} x_i - \theta \\ y_i - \theta \end{pmatrix},$$

and helps to find an approximate but explicit solution

$$\lambda = \frac{\sum_i (x_i - y_i)}{\sum_i (x_i-y_i)^2}, \quad \tilde{\theta}_{MEL} = \frac{\sum_i (1-\lambda(x_i-y_i))x_i}{\sum_i(1-\lambda(x_i-y_i))} = \frac{\sum_i (1-\lambda(x_i-y_i))y_i}{\sum_i(1-\lambda(x_i-y_i))}.$$

Observe that λ is a normalized distance between the sample means of x's and y's, $\tilde{\theta}_{el}$ is a weighted sample mean. The weights are such that the weighted mean of x's equals the weighted mean of y's. So, the moment restriction is satisfied in the sample. Moreover, the weight of observation i depends on the distance between x_i and y_i and on how the signs of $x_i - y_i$ and $\bar{x} - \bar{y}$ relate to each other. If they have the same sign, then such observation says against the hypothesis that the means are equal, thus the weight corresponding to this observation is relatively small. If they have the opposite signs, such observation supports the hypothesis that means are equal, thus the weight corresponding to this observation is relatively large.

Part 3. The technique is the same as in the MEL problem. The Lagrangian is

$$L = -\sum_i p_i \log p_i + \mu\left(\sum_i p_i - 1\right) + \lambda' \sum_i p_i m(x_i, y_i, \theta).$$

The first-order conditions are

$$-\frac{1}{n}(\log p_i + 1) + \mu + \lambda' m(x_i, y_i, \theta) = 0, \quad \lambda' \sum_i p_i \frac{\partial m(x_i, y_i, \theta)}{\partial \theta'} = 0.$$

The first equation together with the condition $\sum_i p_i = 1$ gives

$$p_i = \frac{e^{\lambda' m(x_i, y_i, \theta)}}{\sum_i e^{\lambda' m(x_i, y_i, \theta)}}.$$

Also, we have

$$0 = \sum_i p_i m(x_i, y_i, \theta), \quad 0 = \sum_i p_i \left(\frac{\partial m(x_i, y_i, \theta)}{\partial \theta'}\right)' \lambda.$$

The system for θ and λ that gives the ET estimator is

$$0 = \sum_i e^{\lambda' m(x_i, y_i, \theta)} m(x_i, y_i, \theta), \quad 0 = \sum_i e^{\lambda' m(x_i, y_i, \theta)} \left(\frac{\partial m(x_i, y_i, \theta)}{\partial \theta'}\right)' \lambda.$$

In our simple case, this system is

$$0=\sum_i e^{\lambda_1(x_i-\theta)+\lambda_2(y_i-\theta)}\begin{pmatrix}x_i-\theta\\ y_i-\theta\end{pmatrix},\quad 0=\sum_i e^{\lambda_1(x_i-\theta)+\lambda_2(y_i-\theta)}(\lambda_1+\lambda_2).$$

Here we have $\lambda_1=-\lambda_2=\lambda$ again. The ET estimator is

$$\hat{\theta}_{et}=\frac{\sum_i x_i e^{\lambda(x_i-y_i)}}{\sum_i e^{\lambda(x_i-y_i)}}=\frac{\sum_i y_i e^{\lambda(x_i-y_i)}}{\sum_i e^{\lambda(x_i-y_i)}},$$

where λ is the solution of

$$\sum_i (x_i-y_i)e^{\lambda(x_i-y_i)}=0.$$

Note, that linearization of this system gives the same result as in MEL case.

Since ET estimators are asymptotically equivalent to MEL estimators (the proof of this fact is trivial: the first-order Taylor expansion of the ET system gives the same result as that of the MEL system), there is no need to calculate the asymptotic variances, they are the same as in part 1.

12.2 Kullback–Leibler Information Criterion

1. Minimization of

$$KLIC(e:\pi)=\mathbb{E}_e\left[\log\frac{e}{\pi}\right]=\sum_i\frac{1}{n}\log\frac{1}{n\pi_i}$$

is equivalent to maximization of $\sum_i \log\pi_i$ which gives the MEL estimator.

2. Minimization of

$$KLIC(\pi:e)=\mathbb{E}_\pi\left[\log\frac{\pi}{e}\right]=\sum_i\pi_i\log\frac{\pi_i}{1/n}$$

gives the ET estimator.

3. The knowledge of probabilities p_i gives the following modification of MEL problem:

$$\sum_i p_i\log\frac{p_i}{\pi_i}\to\min_{\pi_i,\theta}\quad\text{s.t.}\ \sum\pi_i=1,\ \sum\pi_i m(z_i,\theta)=0.$$

The solution of this problem satisfies the following system:

$$\pi_i=\frac{p_i}{1+\lambda' m(x_i,y_i,\theta)},$$
$$0=\sum_i\frac{p_i}{1+\lambda' m(x_i,y_i,\theta)}m(x_i,y_i,\theta),$$
$$0=\sum_i\frac{p_i}{1+\lambda' m(x_i,y_i,\theta)}\left(\frac{\partial m(x_i,y_i,\theta)}{\partial\theta'}\right)'\lambda.$$

The knowledge of probabilities p_i gives the following modification of ET problem

$$\sum_i\pi_i\log\frac{\pi_i}{p_i}\to\min_{\pi_i,\theta}\quad\text{s.t.}\ \sum\pi_i=1,\ \sum\pi_i m(z_i,\theta)=0.$$

The solution of this problem satisfies the following system

$$\pi_i = \frac{p_i e^{\lambda' m(x_i, y_i, \theta)}}{\sum_j p_j e^{\lambda' m(x_j, y_j, \theta)}},$$

$$0 = \sum_i p_i e^{\lambda' m(x_i, y_i, \theta)} m(x_i, y_i, \theta),$$

$$0 = \sum_i p_i e^{\lambda' m(x_i, y_i, \theta)} \left(\frac{\partial m(x_i, y_i, \theta)}{\partial \theta'} \right)' \lambda.$$

4. The problem

$$KLIC(e : f) = \mathbb{E}_e \left[\log \frac{e}{f} \right] = \sum_i \frac{1}{n} \log \frac{1/n}{f(z_i, \theta)} \to \min_\theta$$

is equivalent to

$$\sum_i \log f(z_i, \theta) \to \max_\theta,$$

which gives the Maximum Likelihood estimator.