ECON-GA 2100 Econometrics I (Second Half)
Fall 2025
Problem Set 3
Due: Tuesday, November 25.

Problem 1 (Testing Linear Restrictions) This exercise asks you to compare the LM, Wald, and LR test for the homoskedastic linear model with linear restrictions of the form $R\beta = r$, where we denote the residual variance with $\sigma_0^2 := \mathbb{E}[e_i^2|x_i]$. Also, denote the residuals from unrestricted and restricted least squares estimation with $\hat{e}_U := y - X\hat{\beta}_{LS}$ and $\hat{e}_R := y - X\hat{\beta}_{CLS}$, respectively, and suppose the statistics are computed using the estimators $\hat{\sigma}_U^2 := \frac{1}{n}\hat{e}_U'\hat{e}_U$ and $\hat{\sigma}_R^2 := \frac{1}{n}\hat{e}_R'\hat{e}_R$ for the residual variance, i.e. without the degree of freedom correction.

(a) Denoting $\hat{d} := R\hat{\beta}_{LS} - r$, show that $\hat{e}_R = \hat{e}_U - X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}\hat{d}$.

(b) Show that
$$\hat{e}_R'\hat{e}_R = \hat{e}_U'\hat{e}_U + \hat{d}'\left[R(X'X)^{-1}R'\right]^{-1}\hat{d}$$

(c) Use the result from (b) to express the Wald statistic W_n, the LM statistic LM_n, and the LR statistic LR_n exclusively in terms of $\hat{e}_R'\hat{e}_R$ and $\hat{e}_U'\hat{e}_U$. *Hint:* you may use any results from class for the LR and LM statistic.

(d) Prove that for any given sample, $W_n \geq LR_n \geq LM_n$. *Hint:* use a mean-value expansion of $\log(x)$ about $x = 1$.

Problem 2 (Simulation Exercise: Bootstrap) This exercise asks you to implement and assess the performance of the bootstrap for the linear regression model. Suppose you have the linear regression model (LRM)
$$y_i = \beta_0 + \beta_1 x_i + e_i$$
where $x_i \sim U[0, 2]$, and $e_i|x_i \sim U[-1, 1]$ is uniformly distributed. For your simulations set $\beta_0 = \beta_1 = 1$.

(a) Write a code that generates i.i.d. samples of sizes $n = 10, 50, 200$ from that distribution, computes (1) the LS estimator for β, (2) the t-ratio for the LS coefficient β_1, $t_n = (\hat{\beta}_{1,LS} - 1)/(\widehat{s.e.}(\hat{\beta}_{1,LS}))$, and (3) the LS residuals $\hat{e}_i = y_i - \hat{\beta}_{0,LS} - \hat{\beta}_{1,LS}x_i$.

(b) Write a code for drawing n times at random from the discrete uniform distribution over the estimated residuals $\hat{e}_1, \ldots, \hat{e}_n$ (i.e. with replacement).

(c) Use your code from parts (a) and (b) to implement the residual bootstrap - assuming that e_i and x_i are independent - to estimate the 95th percentiles of the respective distributions of $\hat{\beta}_{1,LS}$ and t_n. Repeat each design with 200 simulation replications, drawing a new sample from the DGP for each replication.

(d) Repeat part (a) for sample sizes $n = 10, 50, 200$ with 200 replications each, where you keep the initial draws of $x_1, \ldots, x_n$ from part (a) and only generate new residuals from their conditional distribution. Compute $\hat{\beta}_{1,LS}$ and the statistic t_n using 200 independent samples of size n. Use your results to compute a simulated estimate for the 95th percentiles of the respective sampling distributions for $\hat{\beta}_{1,LS}$ and t_n.

(e) Compare your results from (c) and (d). What do you conclude about the performance of the bootstrap? How does it compare to the 95th percentile of the asymptotic distribution of t_n?

Problem 3 (2SLS and IV) Consider the linear IV model

$$y_i = x_i'\beta + e_i, \qquad \mathbb{E}[z_i e_i] = 0$$

with m regressors, $x_i \in \mathbb{R}^m$, and k instrumental variables, $z_i \in \mathbb{R}^k$. For the first part of the exercise, consider the case $m = k = 2$ with a constant, $x_{i1} = z_{i1} = 1$.

(a) Show that for this case, the rank condition is equivalent to $\mathrm{Cov}(x_{i2}, z_{i2}) \neq 0$.

(b) Show that the 2SLS estimator of the slope coefficient is given by

$$\hat{\beta}_{2,2SLS} = \frac{\mathrm{Cov}(z_{i2}, y_{i2})}{\mathrm{Cov}(z_{i2}, x_{i2})}$$

which we referred to as the instrumental variables estimator in class.

Now return to the case of general m and $k \geq m$. Then for an $k \times m$ matrix A, the *instrumental variables* estimator for β is defined as the solution of the estimating equations

$$A'Z'(y - X\hat{\beta}_{IV}) = 0$$

(c) Solve for $\hat{\beta}_{IV}$ in terms of y, X, Z and A, assuming that a solution exists.

(d) Now choose $A = \hat{\Pi}$, the least squares estimator for Π in the linear first stage

$$x_i' = z_i'\Pi + v_i', \qquad \mathbb{E}[z_i v_i'] = 0$$

Show that the resulting IV estimator is numerically equivalent to 2SLS.

Problem 4 (Nonlinear Transformations)
Suppose you want to estimate the regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

using k instrumental variables z_i satisfying the exclusion restriction $\mathbb{E}[e_i|z_i] = 0$.

(a) Suppose that the true relationship between x_i and z_i is given by a nonlinear relationship, $x_i = h(z_i, v_i)$ where $h(\cdot)$ is a nonlinear function of z_i, v_i, and v_i is an unobserved random vector. Nevertheless as you compute the 2SLS estimator, you fit a linear model for the first stage. Does this lead to an inconsistent estimator of $\beta = (\beta_0, \beta_1)$?

(b) Now assume that the model you'd like to estimate also includes a quadratic term in x_i,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

Suppose you fit a linear first stage

$$x_i = z_i'\pi + v_i$$

via LS, and in a second stage you run a linear regression of y_i on $\hat{x}_i = z_i'\hat{\pi}_{LS}$ and $(\hat{x}_i)^2$. Is the resulting estimator for $\beta = (\beta_0, \beta_1, \beta_2)$ consistent?

(c) Explain how to estimate β consistently.