Queen's University
School of Graduate Studies and Research
Department of Economics

Economics 850 **Econometrics I** **Fall, 2022**

Professor James MacKinnon

Final Examination

December 12, 2022. Time: 3 hours

Notes: The examination is in two parts. Please answer the only question in Part I and three (3) questions from Part II. Tables with some critical values of the χ^2 and Student's t distributions appear at the end of the examination.

Part I. Please answer the following question, which is worth 28% of the final mark.

1. Consider the linear regression model with N observations and k regressors,

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \tag{1}$$

where $\boldsymbol{X}_1$ is an $N \times k_1$ matrix of observations on k_1 exogenous regressors, $\boldsymbol{x}_2$ is an N-vector of observations on a single exogenous regressor, and $\boldsymbol{y}$ and $\boldsymbol{u}$ are N-vectors of observations on a dependent variable and disturbances, respectively.

a) Suppose the elements of $\boldsymbol{u}$, say u_i, are normally and independently distributed with unknown variance σ^2. How would you estimate β_2? How would you then test the hypothesis that $\beta_2 = 0.75$? Write down your test statistic explicitly as a function of $\boldsymbol{y}$, $\boldsymbol{X}_1$, and $\boldsymbol{x}_2$ (or quantities that depend on them). How is this test statistic distributed when $N = 37$ and $k_1 = 3$?

b) Suppose the u_i are independently distributed with unknown variances σ_i^2 that may be related to the regressors. How would you estimate β_2? Write down the test statistic you would use to test the hypothesis that $\beta_2 = 0.75$ as a function of $\boldsymbol{y}$, $\boldsymbol{X}_1$, and $\boldsymbol{x}_2$ (or quantities that depend on them). What can you say about the distribution of this test statistic when $N = 37$ and $k_1 = 3$? What can you say about it when $N = 4{,}758$ and $k_1 = 44$?

c) Suppose the data used to estimate (1) fall into 13 clusters, indexed by g, for $g = 1, \ldots, 13$. Let $\boldsymbol{u}_g$ denote the disturbance vector for the g^{th} cluster. The $\boldsymbol{u}_g$ are assumed to be independent across clusters but to have unknown variances and covariances. There are 4,758 observations, with cluster sizes ranging from 43 to 984, and $k_1 = 44$. How could you test the hypothesis that $\beta_2 = 0.75$ at the .05 level without estimating the model more than once? You do not need to write down your test statistic explicitly, but it should look similar to the test statistic for part b). What distribution will you pretend that it follows?

d) Suppose that $\hat\beta_2 = 0.934$ and that the test statistic of part c) is 2.194. Would you feel confident in rejecting the null hypothesis at the .05 level? Briefly explain how you could test the hypothesis that $\beta_2 = 0.75$ using an alternative procedure that involves 14 OLS regressions. Would you expect the resulting test statistic to be larger or smaller than 2.194? Why?

ANSWER [7 marks for each part]

a) Using the FWL Theorem, we can write

$$\hat\beta_2 = \frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2},$$

where $\boldsymbol{M}_1$ projects orthogonally off $\boldsymbol{X}_1$. The estimate of σ^2 is

$$s^2 = \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}/(N-k),$$

and the standard error of $\hat\beta_2$ is

$$s(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1/2}.$$

Thus the test statistic for $\beta_2 = 0.75$ is

$$\frac{\hat\beta_2 - 0.75}{s(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1/2}}.$$

When $N = 37$ and $k_1 = 3$, this follows the $t(33)$ distribution under the specified assumptions.

b) The estimate of β_2 is still the OLS estimate $\hat\beta_2$. However, instead of the standard error above, we need to use one that is heteroskedasticity-robust. One way would be to use the square root of the k^{th} diagonal element of the matrix

$$\frac{N}{N-k}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{i=1}^{N} \hat u_i^2 \boldsymbol{X}_i^\top \boldsymbol{X}_i\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.$$

This element can also be written as

$$\frac{N}{N-k}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1}\left(\sum_{i=1}^{N} \hat u_i^2 (\boldsymbol{M}_1\boldsymbol{x}_2)_i^\top (\boldsymbol{M}_1\boldsymbol{x}_2)_i\right)(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1},$$

but there is no need to use this more complicated expression. The test statistic is just

$$\frac{\hat\beta_2 - .075}{\text{se}(\hat\beta_2)},$$

where $\operatorname{se}(\hat\beta_2)$ is the HR standard error.

This test statistic is asymptotically $N(0,1)$. When $N = 4{,}758$ and $k_1 = 44$, this is probably a good approximation. But when $N = 37$ and $k_1 = 3$, it is probably a poor one. In all likelihood, $t(33)$ would provide a better approximation, but probably still not a very good one.

c) In this case, we need to use a CRVE. Since only one regression is to be run, it is presumably CV$_1$. Students don't need to write it down, but it is

$$\frac{N}{N-k}\frac{G}{G-1}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g^\top\boldsymbol{X}_g\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1},$$

where $\boldsymbol{X}_g$ contains the rows of $\boldsymbol{X}$ for the g^{th} cluster. The test statistic is just

$$\frac{\hat\beta_2 - .075}{\operatorname{se}(\hat\beta_2)},$$

where now $\operatorname{se}(\hat\beta_2)$ is the square root of the k^{th} diagonal element of the CV$_1$ matrix. We pretend that it follows the $t(12)$ distribution.

d) Because there are only 13 clusters and they vary greatly in size, the $t(12)$ approximation is likely to be poor. The test statistic of 2.194 is only a little larger than the $t(12)$ critical value of 2.179, so even if it were good, we could just barely reject at the .05 level.

An alternative procedure would use the cluster jackknife or CV$_3$ variance matrix estimator. We run 13 additional regressions, each of them omitting one of the 13 clusters. These yield estimates $\hat{\boldsymbol{\beta}}^{(g)}$ for the whole parameter vector and $\hat\beta_2^{(g)}$ for β_2. The estimated standard error of $\hat\beta_2$ is then the square root of

$$\frac{13}{12}\sum_{g=1}^{13}(\hat\beta_2^{(g)} - \hat\beta_2)^2.$$

If we use this standard error to construct a test statistic, the test statistic is very likely (although not certain) to be smaller because the standard error is very likely to be larger.

Part II. Please answer three (3) of the following five (5) questions. Each question has four (4) parts and is worth 24% of the final mark.

2. Consider the nonlinear regression model

$$y_i = \beta_1 + \beta_2 x_{2i}^{\beta_3} x_{3i}^{1-\beta_3} + u_i, \quad u_i \sim \text{IID}(0,\sigma^2), \tag{2}$$

where the regressors x_{2i} and x_{3i} are assumed to be exogenous, and there are 76 observations.

a) When you estimate regression (2) by nonlinear least squares, the SSR is 142.85. When you impose the restriction that $\beta_3 = 0.5$, the SSR is 151.26. If you assume that $u_i \sim \text{IID}(0, \sigma^2)$, can you reject the null hypothesis that $\beta_3 = 0.5$ at the .05 level using an asymptotic test?

b) Explain how to estimate (2) subject to the restriction that $\beta_3 = 0.5$. Then explain how you would test the hypothesis that $\beta_3 = 0.5$ using a Gauss-Newton regression, or GNR, without doing any nonlinear estimation. If you are not sure what the needed derivatives are, do not waste time on them.

c) Suppose you relaxed the assumption that $u_i \sim \text{IID}(0, \sigma^2)$ and instead assumed that $\text{E}(u_i^2 \,|\, \boldsymbol{X}_i) = \sigma_i^2$, with the σ_i^2 unknown. Here $\boldsymbol{X}_i$ denotes the row vector containing 1, x_{2i}, and x_{3i}. Explain how you could use the GNR of part b) to test the hypothesis that $\beta_3 = 0.5$ under this weaker assumption.

d) Suppose you estimate all three parameters by NLS. Discuss how you would obtain a 99% asymptotic confidence interval for β_3 under the assumptions of part c). Explain how you would compute the needed standard error for $\hat\beta_3$.

ANSWER [6 marks a) and b), 4 for c), 8 for d)]

a) The F statistic, which is already in χ^2 form because it has only one degree of freedom, is

$$\frac{151.26 - 142.85}{(142.35/73)} = 4.3128.$$

Since this exceeds the .05 critical value of 3.841, we can reject the null hypothesis that $\beta_3 = 0.5$.

b) Estimating subject to the restriction is easy. We just regress y_i on a constant and $x_{2i}^{1/2} x_{3i}^{1/2}$. This yields restricted estimates of β_1 and β_2.

The GNR in this case is

$$y_i - \beta_1 - \beta_2 x_{2i}^{\beta_3} x_{3i}^{1-\beta_3} = b_1 + b_2 x_{2i}^{\beta_3} x_{3i}^{1-\beta_3} + b_3\big(\log(x_{2i}) - \log(x_{3i})\big) x_{2i}^{\beta_3} x_{3i}^{1-\beta_3} + \text{res.}$$

When evaluated at the restricted estimates, this becomes

$$y_i - \tilde\beta_1 - \tilde\beta_2 x_{2i}^{0.5} x_{3i}^{0.5} = b_1 + b_2 x_{2i}^{0.5} x_{3i}^{0.5} + b_3\big(\log(x_{2i}) - \log(x_{3i})\big) x_{2i}^{0.5} x_{3i}^{0.5} + \text{res.}$$

The ordinary t-statistic for $b_3 = 0$ can be used to test the hypothesis that $\beta_3 = 0.5$.

c) You can still use the same GNR, but now you have to employ a heteroskedasticity-robust standard error to obtain the t-statistic.

d) This time, we have to run the GNR for the unrestricted model. The estimates of b_1 through b_3 should now be zero, but their hetero-robust standard errors are what we want. A 99% confidence interval for β_3 is

$$[\hat\beta_3 - c(.995)\text{se}(\hat\beta_3), \;\; \hat\beta_3 + c(.995)\text{se}(\hat\beta_3)],$$

where $\operatorname{se}(\hat{\beta}_3)$ is the hetero-robust standard error from the GNR and $c(.995)$ denotes the .995 quantile of the standard normal distribution, which is 2.5758.

3. Consider the linear regression model

$$y_{gi} = \beta_1 + \beta_2 x_{gi} + u_{gi}, \quad g = 1, \ldots, G, \quad i = 1, \ldots, N_g, \tag{3}$$

which is to be estimated using a sample of $N = 18{,}674$ observations divided into $G = 16$ clusters. The regressor x_{gi} is assumed to be exogenous. You are interested in the parameter $\gamma \equiv \exp(\beta_2)$.

a) Explain how you would obtain an estimate $\hat{\gamma}$ and an asymptotically valid standard error $s(\hat{\gamma})$ analytically under the assumption that the disturbances in (3) are independent across clusters but may be correlated and/or heteroskedastic within each cluster. Then show how to construct two 95% asymptotic confidence intervals for γ, one symmetric and one asymmetric.

b) There are two natural ways to generate bootstrap samples for (3) under the assumptions of part a) without imposing any restrictions. Briefly explain how each of them would work in this case. Will both of them generate bootstrap samples with 18,674 observations? Explain.

c) Using whichever of the methods from part b) you prefer, explain how you could obtain a bootstrap standard error $s^*(\hat{\gamma})$ and how you would you use that standard error to construct a 95% confidence interval for γ. Would this interval be symmetric around $\hat{\gamma}$? Explain.

d) Explain how you would construct a 95% studentized bootstrap confidence interval for γ using the standard error $s(\hat{\gamma})$ from part a) and the bootstrap DGP from part c). Would this interval be symmetric around $\hat{\gamma}$? Explain.

ANSWER [7 marks for part a), 5 for part b), 6 for c) and d)]

a) Run an OLS regression of y on a constant and x. Obtain a CRVE for the two parameters. Then set $\hat{\gamma} = \exp(\hat{\beta}_2)$. Use the delta method to obtain a standard error for $\hat{\gamma}$. Because the derivative of $\exp(x)$ is just $\exp(x)$, the standard error is

$$\operatorname{se}(\hat{\gamma}) = \exp(\hat{\beta}_2)\operatorname{se}(\hat{\beta}_2),$$

where $\operatorname{se}(\hat{\beta}_2)$ is a cluster-robust standard error.

Two 95% asymptotic confidence intervals are

$$[\hat{\gamma} - 2.131\operatorname{se}(\hat{\gamma}), \;\; \hat{\gamma} + 2.131\operatorname{se}(\hat{\gamma})]$$

and

$$\big[\exp\big(\hat{\beta}_2 - 2.131\operatorname{se}(\hat{\beta}_2)\big), \;\; \exp\big(\hat{\beta}_2 + 2.131\operatorname{se}(\hat{\beta}_2)\big)\big].$$

Note that 2.131 is the 0.975 quantile of $t(15)$.

b) The natural choices are the pairs cluster bootstrap and the wild cluster bootstrap. The former resamples $[\boldsymbol{y}_g, \boldsymbol{X}_g]$ pairs from the 16 clusters. The latter generates bootstrap data as

$$\boldsymbol{y}_g^* = \hat{\beta}_1 + \hat{\beta}_2 \boldsymbol{x}_g + v_g \hat{\boldsymbol{u}}_g,$$

where v_g is Rademacher.

All the wild cluster bootstrap samples will have the same number of observations, but the pairs cluster bootstrap samples will have different numbers.

c) They should prefer the wild cluster bootstrap. For each bootstrap sample, estimate the model to obtain $\hat{\gamma}^*$. Then estimate the bootstrap standard error

$$\text{se}^*(\hat{\gamma}) = \left(\frac{1}{B} \sum_{b=1}^{B} (\hat{\gamma}_b^* - \bar{\gamma}^*)^2 \right)^{1/2}.$$

Use this to construct a symmetric confidence interval in the usual way, as $\hat{\gamma}$ plus or minus 1.96 (or 2.131?) standard errors.

d) For each bootstrap sample, construct the bootstrap test statistic

$$t_b^* = \frac{\hat{\gamma}^* - \hat{\gamma}}{\text{se}(\hat{\gamma}^*)}.$$

Find the .025 and .975 quantiles of the t_b^*. Call these $c_{.025}^*$ and $c_{.975}^*$. Then the studentized bootstrap interval is

$$\big[\hat{\gamma} - c_{.975}^* \text{se}(\hat{\gamma}), \;\; \hat{\gamma} - c_{.025}^* \text{se}(\hat{\gamma})\big].$$

This interval will not be symmetric, because $c_{.975}^* \neq -c_{.025}^*$.

4. This question deals with the linear regression model

$$y_{i1} = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 y_{i2} + u_i, \tag{4}$$

which is to be estimated using a dataset with 516 observations. The matrix $\boldsymbol{X}$ has typical row $[1 \quad z_{i1} \quad z_{i2} \quad y_{i2}]$, where the z_{ij} are predetermined and y_{i2} may be endogenous. It is assumed that both $\boldsymbol{X}^\top \boldsymbol{X}$ and the matrix that $1/N$ times it tends to asymptotically have full rank, and that the u_i are homoskedastic and independent. Three predetermined variables, w_{i1}, w_{i2}, and w_{i3}, are also observed. They are believed to be uncorrelated with u_i but correlated with y_{i2}. They are also assumed to satisfy standard regularity conditions.

a) Explain how you would test the null hypothesis that the OLS estimates of the coefficients in (4) are consistent. What would you conclude if the null hypothesis were rejected?

b) How would you obtain consistent estimates of all the coefficients in (4) using an IV estimator? Write down the covariance matrix of the IV estimates that you would report. Would you expect the standard error of $\hat{\beta}_{3\mathrm{IV}}$ to be larger or smaller than the standard error of $\hat{\beta}_{3\mathrm{OLS}}$? Explain.

c) The IV estimator of part b) involves a first-stage regression. Just what is this regression? Investigators often report the value of a certain test statistic associated with this regression. What is this test statistic, how many restrictions is it testing, and how is it distributed asymptotically? What would you conclude if the P value associated with this test statistic were 0.00013? Explain.

d) Suppose the u_i in (4) are assumed to be heteroskedastic. What is the covariance matrix of the IV estimates that you would report now? If the t-statistic for $\beta_3 = 0$ based on the appropriate diagonal element of this matrix were 2.027, would you be comfortable rejecting the hypothesis that $\beta_3 = 0$ at the .05 level? Explain why or why not.

ANSWER [6 marks for each part]

a) First, run the first-stage regression and retrieve either the residuals or the fitted values. This regression is

$$y_{i2} = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \pi_3 w_{i1} + \pi_4 w_{i2} + \pi_5 w_{i3} + v_i.$$

Call the residuals $\hat{v}_i$. Next, run the regression

$$y_{i1} = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 y_{i2} + \delta \hat{v}_i + u_i.$$

Perform an ordinary t-test or F test for $\delta = 0$. If the null is rejected, then either y_2 should be treated as endogenous or at least one of the w_j should have been included as a regressor in (4).

b) The IV estimator is

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\boldsymbol{X}^\top \boldsymbol{P_W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P_W} \boldsymbol{y}_1,$$

where $\boldsymbol{X}$ contains N observations on the constant, the two z_{ji} regressors, and the three w_{ji} instruments. The IV covariance matrix that you would report is

$$\hat{\sigma}^2_{\mathrm{IV}} (\boldsymbol{X}^\top \boldsymbol{P_W} \boldsymbol{X})^{-1},$$

where $\hat{\sigma}^2_{\mathrm{IV}}$ is $1/N$ times the sum of squared IV residuals. Note that the vector of IV residuals is

$$\boldsymbol{y}_1 - \boldsymbol{X} \hat{\boldsymbol{\beta}}_{\mathrm{IV}}.$$

It is not the vector of residuals from the second-stage regression.

You would expect the standard error of $\hat{\beta}_{3\,\mathrm{IV}}$ to be larger than the standard error of $\hat{\beta}_{3\,\mathrm{OLS}}$. This follows from the fact that the length of $\boldsymbol{P_W X}$ is less than the length of $\boldsymbol{X}$. Hence the difference between $(\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1}$ and $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is a positive semidefinite matrix.

c) We already wrote down the first-stage regression. It is

$$y_{i2} = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \pi_3 w_{i1} + \pi_4 w_{i2} + \pi_5 w_{i3} + v_i.$$

The test statistic of interest here is the F statistic for $\pi_3 = \pi_4 = \pi_5 = 0$. It is testing 3 restrictions, and it is approximately distributed as $F(3, 510)$. Since normality was not assumed, this distribution is not exact. Asymptotically, 3 times it is distributed as $\chi^2(3)$. A P value of 0.00013 suggests that the test statistic is quite large, so that IV inference should not be too unreliable. However, it is actually not large enough to gain the approval of the Stock-Yogo tables.

d) The covariance matrix is now a sandwich estimator:

$$(\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1} \left(\sum_{i=1}^{516} \hat{u}^2_{\mathrm{iv}\,i} (\boldsymbol{P_W X})_i^\top (\boldsymbol{P_W X})_i \right) (\boldsymbol{X}^\top \boldsymbol{P_W X})^{-1}.$$

With a t-statistic of 2.027, I would definitely not feel comfortable rejecting the null hypothesis. Even when the Stock-Yogo conditions are satisfied (and they are not here), IV t-statistics often do not follow their asymptotic distribution very well. We know nothing about the correlation between the reduced-form and structural errors (the v_i and the u_i). If we knew that correlation was small, we might be more comfortable. But there are 2 over-identifying restrictions, and since 2.027 is only a little larger than 1.96, the evidence seems pretty weak.

5. Suppose you are given a sample of 2,365 observations on the incomes of corporate lawyers in 2019. The largest income is \$14,688,455 and the second-largest is \$2,371,346. The mean is \$725,234, and the median is about 2/3 of the mean. The rest of the distribution is more or less as you would expect it to be given these values.

a) If you were to plot the empirical distribution function for this sample, what would it look like? Would it have any interesting features? For example, how would the distance between the α quantile and the median be related to the distance between the median and the $1 - \alpha$ quantile for $\alpha = 0.10$?

b) Explain how you would estimate the first quartile, the median, and the third quartile of the population distribution using this sample. Then explain how you would construct standard errors for these estimates using the bootstrap. Which of the three bootstrap standard errors would you expect to be the largest? Explain.

c) For the sample mean, you could easily construct a sample standard error. Then you could use the bootstrap to form a studentized bootstrap confidence interval at the 0.95 level. Explain how you would do this. Would this interval be symmetric around \$725,234? What would it look like?

d) Suppose that, in addition to the original sample of 2,365 incomes for corporate lawyers, you are given a sample of 1,465 incomes for tax lawyers. Discuss how you could use bootstrap methods to test the hypothesis that the income distributions for corporate lawyers and tax lawyers are the same.

ANSWER [4 marks for a), 7 for b) and d), 6 for c)]

a) The EDF will be very asymmetrical. Half the observations are below the median, which is about $2/3 \times 725{,}234 = 483{,}490$, and half are above it. But the former run from an unspecified minimum above zero to 483,490, and the latter run from 483,490 to 14,688,455. Evidently, the distance between the 0.10 quantile and the median is going to be much smaller than the distance between the median and the 0.90 quantile.

b) The estimates are approximately numbers 0.25×2365, 0.5×2365, and 0.75×2365 in the sorted list. These indices are 591.25, 1182.5, and 1772.75, respectively. For the median, a good estimate is the average of numbers 1182 and 1183. For the two quartiles, you could just use numbers 591 and 1773, or you could take weighted averages with weights 1/4 and 3/4. For example, the estimate of the third quartile would be $0.25 \times \#1772 + 0.75 \times \#1773$.

This is a case where the classic resampling bootstrap makes sense. Generate B bootstrap samples by resampling without replacement. Pick, say, $B = 10{,}000$. In this case, there is no reason for B to end in 99. For each bootstrap sample, estimate the three quartiles in the same way as you did with the real data. Then calculate the variances of each set of estimates. For example, if the estimated medians are m_1^* through m_B^*, compute

$$\bar{m}^* = \frac{1}{B}\sum_{b=1}^{B} m_b^* \quad \text{and} \quad \mathrm{Var}(m^*) = \frac{1}{B-1}\sum_{b=1}^{B}(m_b^* - \bar{m}^*)^2.$$

Then the bootstrap standard error is the square root of $\mathrm{Var}(m^*)$.

The bootstrap standard error for the third quartile is sure to be much larger than the other two bootstrap standard errors, because the EDF is much flatter in the neighborhood of that quartile. It is not obvious which of the other two will be largest. For a symmetric distribution, the median will be estimated more accurately than either of the other quartiles. But, in this case, the EDF might possibly be steeper near the 0.25 quartile than near the median.

c) For the sample mean $\bar{y}$, we can compute $s(\bar{y})$ in the usual way. The square of it is just the sum of squared deviations between y_i and $\bar{y}$, divided by $N(N-1)$.

Equivalently, we could regress y_i on a constant and use the standard error for the constant terms from the regression. Then, for B bootstrap samples (where B now should end in 99, so perhaps $B = 9999$), we can compute

$$t_i^* = \frac{y_i^* - \bar{y}}{s(\bar{y}^*)}.$$

Find the .025 and .975 quantiles of the t_i^*. When $B = 9999$, these are numbers 250 and 9750. Call them $c^*_{.025}$ and $c^*_{.975}$. Then the studentized bootstrap interval is

$$\left[\bar{y} - c^*_{.975}s(\bar{y}),\ \ \bar{y} - c^*_{.025}s(\bar{y})\right].$$

This is evidently not symmetric. It will be skewed to the left (!), because $c^*_{.975}$ is almost certainly greater than $|c^*_{.025}|$.

d) The first step is to combine the two samples and resample from them jointly. If the null hypothesis is true, then any quantity that can be calculated for each of the two samples should follow the same distribution. Since distributions can differ in many respects, there is potentially a large number of bootstrap test statistics that can be computed.

A simple one is the difference between the sample medians. For each of the two subsamples, we can compute the medians m_c and m_t and their difference $\varDelta = m_c - m_t$. Then we can generate B pairs of bootstrap samples (ending in 99), by resampling from the joint empirical distribution, and use each of them to compute $\varDelta_b^*$. If $\varDelta$ is extreme relative to the distribution of the $\varDelta_b^*$, we can reject the null hypothesis. Here an equal-tail test makes sense. We would compute

$$\frac{2}{B}\min\left(\sum_{b=1}^{B}\mathbb{I}(\varDelta_b^* - \varDelta),\ \ \sum_{b=1}^{B}\mathbb{I}(\varDelta - \varDelta_b^*)\right).$$

If this is less than α, we can reject the null at level α.

A test based on a difference of medians is not very interesting. We could base it on anything that can be computed separately for the two subsamples. For example, we could use the Kolmogorov-Smirnov statistic, which is the largest vertical distance between the EDFs for the two samples.

It is tempting to use several tests, such as differences for several quantiles in addition to the median. However, the probability that at least one test out of, say, J, will reject is greater than the probability that just one test rejects. So care will have to be taken to avoid over-rejection.

If you want to use multiple tests, one approach would be to define $\varDelta$ and its bootstrap analog as a function of all the test statistics. For example, it could be the (signed) maximum difference between nine estimated quantiles (from .10 to .90) for the two subsamples. Then we could calculate a bootstrap P value just as we did above.

6. Consider the linear regression model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i, \tag{5}$$

where the regressors are assumed to be exogenous and the error terms are assumed to be independent and identically distributed with mean zero and variance σ^2.

a) Suppose you have 270 observations, which naturally divide into two subsamples, the first with 160 observations and the second with 110. The sums of squared residuals from OLS estimation of (5) over the whole sample and each of the two subsamples are 23.61, 13.33, and 9.45, respectively. Can you reject the null hypothesis that all the parameters are the same for both subsamples using an asymptotic test at the .01 level? Explain.

b) Explain precisely how you would perform a bootstrap test of the hypothesis that all the parameters are the same for both subsamples using no more than 10^4 bootstrap samples. Be sure to specify the bootstrap DGP and explain how you would decide whether or not to reject the null hypothesis at the .01 level.

c) A more restrictive alternative hypothesis is that

$$y_i = \gamma_1 d_{1i} + \gamma_2(1 - d_{1i}) + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i, \tag{6}$$

where d_{1i} is a dummy variable that is equal to 1 if observation i belongs to the first subsample and equal to 0 otherwise. Suppose the SSR from OLS estimation of (6) were 23.18. Using an asymptotic test, would you reject (5) against (6) at the .01 level?

d) Because of the IID assumption, the tests you have done so far must have assumed that the variance of the error terms is the same for both subsamples. Suppose you want to relax this assumption by allowing each of the u_i to have its own variance σ_i^2. Explain how you would generate 9999 bootstrap samples assuming that (5) holds under this weaker assumption. Then explain how you would use these bootstrap samples to test (5) against (6). Be sure to explain why you would, or would not, use the same test statistic as in part b).

ANSWER [5 marks a), 6 for b) and c), 7 for d)]

a) The F statistic is

$$\frac{(23.61 - 13.33 - 9.45)/3}{(13.33 + 9.45)/(270 - 6)} = 3.2063.$$

In χ^2 form, it is 9.619. Since the .01 critical value for $\chi^2(3)$ is 11.345, we cannot reject at the .01 level.

b) Because of the IID assumption, we can use the residual bootstrap. Estimate the model for the entire sample to obtain $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}$. Then resample from the $\tilde{u}_i$ and generate bootstrap samples as

$$y_i^* = \boldsymbol{X}_i\tilde{\boldsymbol{\beta}} + u_i^*, \quad u_i^* \sim \text{EDF}(\tilde{u}_i).$$

The best choice of B given the restriction that it not exceed 10^4 is 9999. For each bootstrap sample, compute the F statistic in the same way as in part a). Then the bootstrap P value is

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(F_b^* > F).$$

Reject the null if $p^* < 0.01$.

c) This hypothesis involves 4 parameters, so there is just one restriction. The F statistic is now

$$\frac{23.61 - 23.18}{23.18/(270-4)} = 4.9344.$$

This is less than the .01 critical value of 6.635, so we do not reject the null.

d) Use the restricted wild bootstrap to generate the bootstrap samples:

$$y_i^* = \boldsymbol{X}_i\tilde{\boldsymbol{\beta}} + u_i^*, \quad u_i^* = v_i^*\tilde{u}_i,$$

where v_i^* is Rademacher.

You cannot use the same test statistic as before, because it assumes that the disturbances are homoskedastic. Instead, you need one that is hetero-robust. A natural one is the Wald statistic, in t form,

$$t_\gamma = \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\text{se}(\hat{\gamma}_1 - \hat{\gamma}_2)},$$

where $\text{se}(\hat{\gamma}_1 - \hat{\gamma}_2)$ is a hetero-robust standard error. This just requires OLS estimation of (6), rewritten so that $\gamma_1 - \gamma_2$ is a coefficient, using an HR variance matrix. You can calculate t_γ^* for each of the bootstrap samples and then compute the bootstrap P value

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(t_\gamma^{*2} - t_\gamma^2).$$

This yields a symmetric bootstrap P value, which is comparable to what we used in the homoskedastic case.