Prof. James MacKinnon QUEEN'S UNIVERSITY Department of Economics

Economics 850 Fall, 2024

Assignment 2

Due: October 10, 2024

1. The file **house-price-data.csv** contains 546 observations, as described in the file **house-price-info.txt**.

a) Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables. Report the OLS estimates, standard errors (based on the assumption that the disturbances are IID), t-statistics, and P values.

b) One of the explanatory variables is the number of storeys, which can take on the values 1, 2, 3, and 4. A more general specification would allow the effect on log price of each number of storeys to be different. Test the original model against this more general one using an F test. Report the test statistic, the degrees of freedom parameters that you used, and the P value. Is the test you performed exact without any additional assumptions? Explain.

c) What is s, the standard error of the regression? Test the hypothesis that $\sigma = 0.20$ at the .05 level under the assumption that the disturbances are normally, identically, and independently distributed with variance σ^2. Be sure to report a P value for the test. Is the test you performed an exact test? Explain why or why not. **Hint:** How is the quantity $\boldsymbol{y}^\top \boldsymbol{M_X} \boldsymbol{y}/\sigma_0^2$ distributed under the specified assumption, if σ_0 is the true value of σ?

d) Using the same data, model, and assumptions as in part c), test the hypothesis that $\sigma \leq 0.20$ at the .05 level. Be sure to report a P value for the test.

e) Estimate the model again using data for only the first 540 observations. Use those estimates to forecast the log prices for the last 6 observations. What are the standard errors of these 6 forecasts? What are the associated 95% forecast intervals? Note that it is possible to obtain both the forecasts and the standard errors by running a single regression with 546 observations.

[36]

2. Consider the linear regression model with N observations,

$$\boldsymbol{y} = \delta_1 \boldsymbol{d}_1 + \delta_2 \boldsymbol{d}_2 + \boldsymbol{u}, \;\; \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}). \tag{1}$$

The two regressors are dummy variables, with every element of $\boldsymbol{d}_2$ equal to 1 minus the corresponding element of $\boldsymbol{d}_1$. The vector $\boldsymbol{d}_1$ has N_1 elements equal to 1, and the vector $\boldsymbol{d}_2$ has $N_2 = N - N_1$ elements equal to 1. You can think of $d_{1i} = 1$ as denoting control observations and $d_{i2} = 1$ as denoting treated observations.

a) Suppose the parameter of interest is $\gamma \equiv \delta_2 - \delta_1$. Derive the true standard error of $\hat{\gamma}$ (that is, the standard error when σ^2 is known) and write it as a function of σ, N, and N_2. **Hint:** Rewrite regression (1) so that γ is estimated directly.

b) Suppose the data for regression (3) come from a survey that you design and administer. If you can only afford to collect 900 observations, how should you choose N_1 in order to estimate γ as efficiently as possible?

c) Suppose that $\boldsymbol{X}$, a matrix of regressors that do not vary systematically with the sample size N, is added to regression (3), so that it becomes

$$\boldsymbol{y} = \delta_1 \boldsymbol{d}_1 + \delta_2 \boldsymbol{d}_2 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{2}$$

What is the true variance of $\hat{\gamma}$ in this case? Write this variance as a function of N and N_2 using same-order notation. Of what order in N is this variance if N_1 is chosen optimally? Will this variance tend to 0 as $N \to \infty$ if N_2 is held fixed? Explain.

[32]

3. Consider two linear regressions, one restricted and the other unrestricted:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}; \tag{3}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{v}. \tag{4}$$

a) Show that the SSR from OLS estimation of (3) is always equal to or greater than the SSR from OLS estimation of (4).

b) The OLS estimates from (3) and (4) are $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, respectively. Under the assumption that the disturbances in (3) and (4) are independently and identically distributed, what are $\text{Var}(\tilde{\boldsymbol{\beta}})$ and $\text{Var}(\hat{\boldsymbol{\beta}})$?

c) Show that, when $\boldsymbol{X}^\top \boldsymbol{Z} = \boldsymbol{0}$, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

d) In general, $\text{Var}(\tilde{\boldsymbol{\beta}}) \neq \text{Var}(\hat{\boldsymbol{\beta}})$. But we saw in c) that $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ when $\boldsymbol{X}^\top \boldsymbol{Z} = \boldsymbol{0}$. How can the same estimates have two different covariance matrices? Which one is correct? **Hint:** Think of $\boldsymbol{u}$ in (3) as equal to $\boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{v}$ in (4). What would happen if all the variation in $\boldsymbol{u}$ came from variation in $\boldsymbol{Z}$?

[32]