Introductory Econometrics
Multiple Hypothesis Testing
Suggested Solution

by Hieu Nguyen

Fall 2024

1.

File `wage.csv` contains a cross-sectional dataset on 526 working individuals for the year 1976 in the US. Using this labor market data, estimate a simple model describing the impact of years of education and work experience on hourly wage in USD per hour:

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \epsilon.$$

(a) Import data into Gretl from the .csv file. Carry out a basic inspection of data (display values, visually, descriptive statistics).

(b) Comment on the expected signs of coefficients β_1 and β_2 first and then estimate the model.

(c) Evaluate the statistical significance of β_1 and β_2 based on the Gretl output.

(d) How much of the variation in wage for these 526 individuals is explained by `educ` and `exper`? Explain.

(e) Estimate also the model without `exper`, compare R^2 and R^2_{adj}. Which is a better model? Why?

(f) Test formally the following hypotheses at the 5% significance level:

(i) Education has a significant impact on wages.

(ii) Workforce experience has a significantly positive impact on wages.

(iii) The regression is overall significant.

(g) Set up a 90% confidence interval for β_2 (and a 99% confidence interval for β_1).

(h) How would the estimated coefficients, standard errors, and t-statistics have differed if we transformed the wage variable into monthly income and `exper` into decades? Explain.

Solution:

(a) To open the data (in Mac): File—Open data—User file—select *.csv as a type of files—find `wage.csv` in your directory and select it—No—Close. Path in the Gretl menu to open the data using lab computers (Window): File—Open data—Import—text/CSV—comma (,)—find `wage.csv` in your directory and select it—No—close the Gretl info window. Or you can drag and drop data directly into the Gretl window. To conduct a basic data inspection, use the right mouse click on a specific variable or the View option from the Gretl menu.

(b) Before estimation, we state our expectations about signs of coefficients (intuition behind the 'wage equation'). Then follow the path in the Gretl menu: Model—Ordinary Least Squares—select `wage` as the dependent var—select independent variables—OK:

	Coefficient	Std. Error	t-ratio	p-value
const	-3.39054	0.766566	-4.4230	0.0000
educ	0.644272	0.0538061	11.9740	0.0000
exper	0.0700954	0.0109776	6.3853	0.0000

(c) Both estimated regression coefficients have expected signs. Moreover, using the Gretl-default two-sided t-test with $H_0 : \beta_i = 0$ vs $H_A : \beta_i \neq 0$ at the 5% significance level (critical value $t_{523,0.975} = 1.96$ or you can use the rule of thumb with 3) and t-statistics (t-ratios) from the Gretl output, we strongly reject H_0 for both coefficients that are thus statistically significant. Considering p-values, both coefficients would have been statistically significant even at the 1% significance level because you can see the *** in the Gretl output.

(d) $R^2 = 0.225$, i.e., 22.5% variation in wage is explained by the variation in `educ` and `exper`, leaving 77.5% variation in wage can be explained by other variables, not included in the model.

(e)
- Model with `exper`: $R^2 = 0.225$, $R^2_{\text{adj}} = 0.222$.
- Model without `exper`: $R^2 = 0.165$, $R^2_{\text{adj}} = 0.163$.

The model with `exper` is better and will be used in further analysis because:

(a) Both RHS variables follow a good theoretical economic motivation to be included.

(b) Both estimated regression coefficients have expected signs and are statistically significant at usual significant levels (individually based on t-tests as well as jointly based on the F-test).

(c) It explains more variation in the dependent variable based on R^2_{adj} (as well as based on R^2, in fact).

(f) (i) This is an example of a two-sided t-test (because the focus is only on significance, not the direction of the impact):

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_A : \beta_1 \neq 0 \implies t_{\beta_1} = \frac{b_{\beta_1}}{s.e.(b_{\beta_1})} \sim t_{n-k-1}.$$

We simply compute from the regression output: $t_{\beta_1} = \frac{0.644}{0.054} \approx 11.9$. The critical value for a two-sided t-test is $t_{n-k-1,0.975} = t_{523,0.975} = 1.96$. We reject H_0 if $|t| > 1.96$, otherwise we do not reject H_0. Hence we reject H_0 for the coefficient β_1, which is thus statistically significant at the given significance level.

(ii) This is an example of a one-sided t-test (because the focus is also on the direction of the impact):

$$H_0 : \beta_2 \leq 0 \quad \text{vs} \quad H_A : \beta_2 > 0 \implies t_{\beta_1} = \frac{b_{\beta_1}}{s.e.(b_{\beta_1})} \sim t_{n-k-1}.$$

We simply compute from the regression output: $t_{\beta_2} = \frac{0.070}{0.011} \approx 6.4$. The critical value for a one-sided t-test is $t_{n-k-1,0.95} = t_{523,0.95} = 1.645$. We reject H_0 if $|t| > 1.645$, and t also has the sign implied by H_A, otherwise we do not reject H_0. Hence we reject H_0 for the coefficient β_2, which is thus statistically significantly positive at the given significance level.

(iii) Here we test the overall significance of the regression, i.e., we test for this complete set of two joint hypotheses using an F-test:

$$H_0 : \begin{cases} \beta_1 = 0 \\ \beta_2 = 0 \end{cases} \quad \text{vs} \quad H_A : \begin{cases} \beta_1 \neq 0 \\ \text{or } \beta_2 \neq 0 \end{cases}$$

We need to estimate also the restricted model:

$$\text{wage} = \beta_0 + \epsilon$$

and compute from the regression outputs using the F-statistic formula from lecture #5 slides:

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n-k-1)} = \frac{(7160.4 - 5548.2)/2}{5548.2/523} = \frac{806.1}{10.6} \approx 76.05 \sim F_{J,n-k-1}.$$

The critical value for an F-test is $F_{J,n-k-1,0.95} = F_{2,523,0.95} = 3$. We reject H_0 if $F > 3$, otherwise we do not reject H_0. Hence, we reject the joint H_0 in favor of the H_A at the given significance level, and the regression is overall statistically significant.

(g) Since

$$\frac{b_\beta}{s.e.(b_\beta)} \sim t_{n-k-1}, \quad \text{we derive the 90\% confidence interval for } \beta_2 \text{ as:}$$

$$b_{\beta_2} \pm t_{n-k-1,1-\frac{\alpha}{2}=0.95} \cdot s.e.(b_{\beta_2}) = 0.070 \pm 1.645 \cdot 0.011 = [0.052, 0.088]$$

Hence, $\beta_2 \in [0.052, 0.088]$ with 90% probability.

Similarly,

$$b_{\beta_1} \pm t_{n-k-1,0.995} \cdot s.e.(b_{\beta_1}) = 0.644 \pm 2.576 \cdot 0.054 = [0.505, 0.783]$$

Hence, $\beta_1 \in [0.505, 0.783]$ with 99% probability.

(h) This is, in fact, just a linear transformation (multiplication/scaling) of data by a constant; see seminar #4, exercise 2. Assuming 20 workdays per month and 8 work hours per day, the impact of data transformation can be summarized as follows:

- b_{β_0}, after transformation $= b_{\beta_0} \cdot 20 \cdot 8$;
- b_{β_1}, after transformation $= b_{\beta_1} \cdot 20 \cdot 8$;
- b_{β_2}, after transformation $= b_{\beta_2} \cdot 20 \cdot 8 \cdot 10$;
- Respective standard errors change accordingly;
- t-statistics not affected.

2.

Answer the following questions about data on the sales prices of houses in the UK. The variables in this study are:

- $PRICE_i$: sales price for house i;
- $ASSESS_i$: assessed price of house i;
- $LOTSIZE_i$: size of lot (in square feet) for house i;
- $BDRMS_i$: number of bedrooms for house i;
- $BATH_i$: number of bathrooms for house i;
- $OCEAN_i$: a variable equal to 1 if house i is located within 10 miles of the ocean, 0 otherwise;
- $URBAN_i$: a variable equal to 1 if house i is located in an area classified as urban, 0 otherwise;
- $LAKE_i$: a variable equal to 1 if house i is located within 10 miles of a lake, 0 otherwise;
- $INTERCEPT$: intercept in the model.

Table 1 lists estimated coefficients with standard errors in parentheses below.

(a) Using the reported regressions, could you test whether the value of the house near water was different from the value of the house away from water at the 5% significance level, controlling for assessed value, lot size, and the number of bedrooms? If so, perform the test. If not, explain what results you would need to do the test.

(b) Could you test whether bathrooms change the house value, controlling for assessed value, lot size, and the number of bedrooms at the 5% significance level? If so, perform the test. If not, explain what results you would need to do the test.

(c) Can you test whether the assessed value and number of bedrooms are jointly significant, controlling for lot size? If yes, perform the test at the 5% significance level. If not, explain what you would need to perform this test.

(d) Could you test whether all 7 of the listed variables (excluding the intercept) are jointly significant at the 5% significance level? Be sure to state any assumptions you are making.

Table 1: Results of regressions

Dependent variable $\texttt{PRICE}_i$, $n = 238$							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\texttt{ASSESS}_i$	0.90	0.90	0.91	0.90	0.89	0.90	0.90
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
$\texttt{LOTSIZE}_i$	0.0035	0.00059	0.00059	0.00057	0.00058	0.00059	0.00060
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
$\texttt{BDRMS}_i$	11.5	9.74	7.65	8.74	10.43		
	(2.32)	(3.11)	(3.29)	(3.54)	(3.77)		
$\texttt{BATH}_i$			3.57	3.78			
			(2.24)	(1.11)			
$\texttt{OCEAN}_i$	15.6	14.32	16.76	15.32	14.56		
	(11.43)	(5.21)	(4.32)	(4.98)	(7.01)		
$\texttt{URBAN}_i$		9.54	10.29	12.32			
		(8.99)	(5.43)	(5.22)			
$\texttt{LAKE}_i$				11.36	12.87	11.98	
				(4.28)	(8.32)	(6.43)	
INTERCEPT	261.9	-38.91	-40.30	-43.21	-36.54	-42.37	-38.44
	(11.98)	(6.78)	(7.32)	(6.99)	(5.87)	(7.22)	(9.43)
RSS	145.69	142.99	136.66	134.54	135.38	135.22	136.54
R^2	0.143	0.159	0.196	0.209	0.204	0.205	0.197

Solution:

(a) Here we test the joint significance of two coefficients because we test for this (incomplete) set of two joint hypotheses using an F-test:

$$H_0 : \begin{cases} \beta_{\text{OCEAN}} = 0 \\ \beta_{\text{LAKE}} = 0 \end{cases} \quad \text{vs} \quad H_A : \begin{cases} \beta_{\text{OCEAN}} \neq 0 \\ \text{or } \beta_{\text{LAKE}} \neq 0 \end{cases}$$

We have $J = 2$ (the number of restrictions), $n = 238$ (sample size), $k = 5$ (the number of independent variables).

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n-k-1)} \sim F_{J,n-k-1}.$$

Unfortunately, while we have the unrestricted model of (6), we don't have the restricted model. Therefore, we cannot find the value of F-test and hence, cannot make decision of reject or accept the H0 and H1.

To perform the test, we need to have the regression output of restricted model to get SSR_R to find F-test, then find F-critical value and compare between F-test and F-critical value to make decision of reject or accept the H0 and H1.

(b)

$$H_0 : \beta_{\text{BATH}} = 0 \quad \text{vs} \quad H_A : \beta_{\text{BATH}} \neq 0 \implies t_{\beta_{\text{BATH}}} = \frac{b_{\beta_{\text{BATH}}}}{s.e.(b_{\beta_{\text{BATH}}})} \sim t_{n-k-1}.$$

This is a standard two-sided t-test; however, we cannot conduct it because we do not have the model with only 4 mentioned explanatory variables (`ASSESS`, `LOTSIZE`, `BDRMS`, `BATH`).

(c) Again, we test the joint significance of two coefficients, i.e., we test for this (incomplete) set of two joint hypotheses:

$$H_0 : \begin{cases} \beta_{\text{ASSESS}} = 0 \\ \beta_{\text{BDRMS}} = 0 \end{cases} \quad \text{vs} \quad H_A : \begin{cases} \beta_{\text{ASSESS}} \neq 0 \\ \text{or } \beta_{\text{BDRMS}} \neq 0 \end{cases}$$

Unrestricted model: none, restricted model: none.

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n-k-1)} =\sim F_{J,n-k-1} = F_{2,234,0.95} = 3.$$

Unfortunately, we don't have unrestricted model nor restricted model to get the values of SSE_U and SSE_R to perform the F-test. Hence, we cannot make decision of accept or reject null hypothesis.

To perform the test, we need to have the regression output of both unrestricted and restricted models to get SSR_R, SSR_U to find F-test, then find F-critical value and compare between F-test and F-critical value to make decision of reject or accept the H0 and H1.

(d) This is another example of testing the overall significance of the regression because we consider the complete set of all 7 variables. Although we do not have the restricted model $\text{PRICE}_i = \beta_0 + \epsilon_i$ in Table 1, we utilize the fact that if we regress on a constant only, $R^2 = 0$. Unrestricted model: (4), Restricted model will be $\hat{price} = \beta_0 + u$ and hypotheses are
H_0 : all $\beta = 0$ and H_A : at least one $\beta \neq 0$.

$$F = \frac{R^2/J}{(1-R^2)/(n-k-1)} = \frac{(0.209-0)/7}{(1-0.209)/(238-8)} = 8.7 \sim F_{7,230,0.95} = 2.01$$

Hence we can reject the joint H_0 in favor of the H_A at the given significance level, and we can conclude that all 7 of the listed variables are jointly significant.

The assumption under which we can compute F-test statistics based on R^2s instead of RSSs is that $TSS_U = TSS_R$, i.e., that Total Sums of Squares are the same for our unrestricted and restricted model. Since $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ we can safely suppose in our case that the mentioned assumption is fulfilled, as we use in both models the same dependent variable (`PRICE`), thus we have the same observations y_i and also the same $\bar{y}$.