THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

ECON2300 - Introductory Econometrics

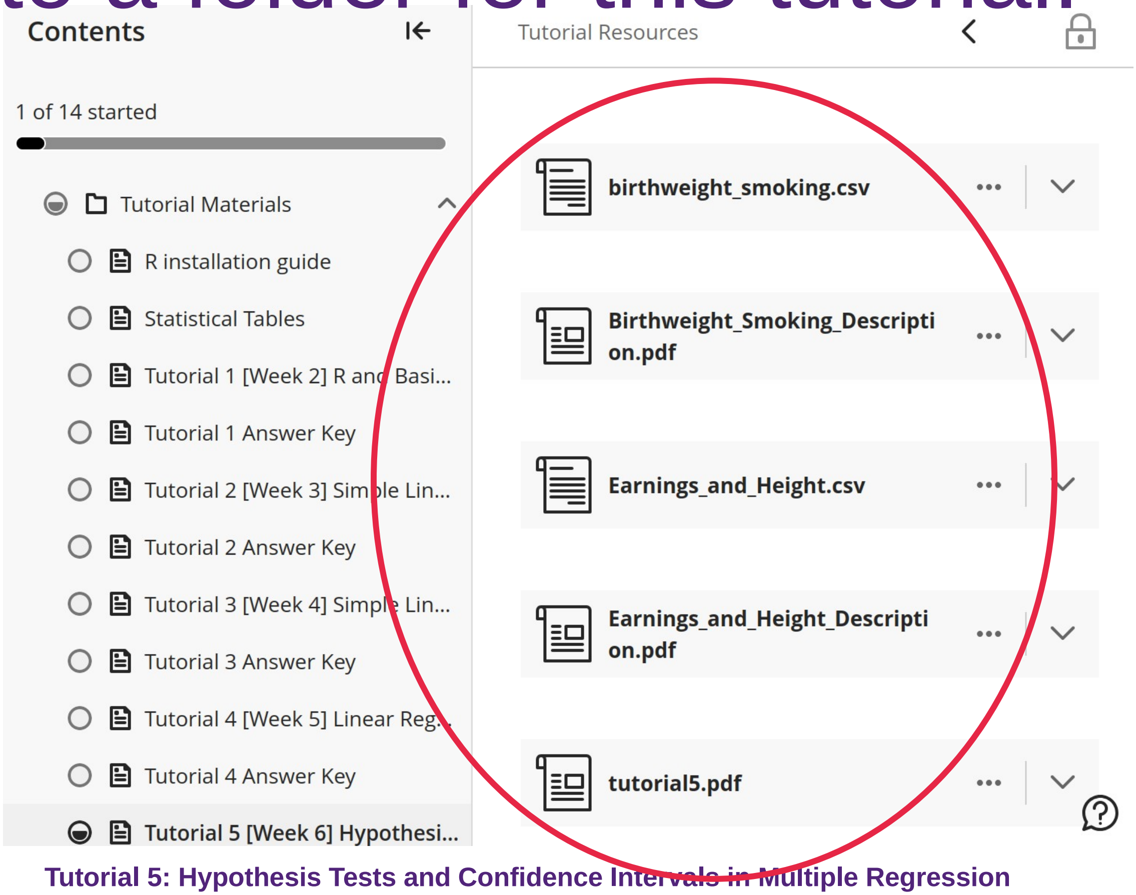
Tutorial 5: Hypothesis Tests and Confidence Intervals in Multiple Regression

Tutor: Francisco Tavares Garcia

Quiz 4 is now available under the Assessment folder.

The due date is Thursday, 4th September, 16:00.

- Download the files for tutorial 05 from Blackboard,
- save them into a folder for this tutorial.



Now, let's download the script for the tutorial.

- Copy the code from Github,
  - https://github.com/tavaresgarcia/teaching
- Save the scripts in the same folder as the data.

E7.1 Using the `Birthweight_Smoking.csv` introduced in E5.3 to answer the following questions. To begin, run three regressions:

(1) `birthweight` on `smoker`.

(2) `birthweight` on `smoker`, `alcohol`, and `nprevist`.

(3) `birthweight` on `smoker`, `alcohol`, `nprevist`, and `unmarried`.

	Variable	**Description**
		Birthweight and Smoking
1	birthweight	birth weight of infant (in grams)
2	smoker	indicator equal to one if the mother smoked during pregnancy and zero, otherwise.
		Mother's Attributes
3	age	age
4	educ	years of educational attainment (more than 16 years coded as 17)
5	unmarried	indicator =1 if mother is unmarried
		This Pregnancy
6	alcohol	indicator=1 if mother drank alcohol during pregnancy
7	drinks	number of drinks per week
8	tripre1	indicator=1 if 1st prenatal care visit in 1st trimester
9	tripre2	indicator=1 if 1st prenatal care visit in 2nd trimester
10	tripre3	indicator=1 if 1st prenatal care visit in 2nd trimester
11	tripre0	indicator=1 if no prenatal visits
12	nprevist	total number of prenatal visits

E7.1 Using the Birthweight_Smoking.csv introduced in E5.3 to answer the following questions. To begin, run three regressions:

```
library(readr)      # package for fast read rectangular data
library(dplyr)      # package for data manipulation
library(estimatr)   # package for commonly used estimators with robust SE
library(texreg)     # package converting R regression output to LaTeX/HTML tables
library(car)        # package for functions used in "An R Companion to Applied Regression"
```

SW E7.1

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial05")
BW <- read_csv("birthweight_smoking.csv")
reg1 = lm_robust(birthweight ~ smoker, data = BW, se_type = "stata")
reg2 = lm_robust(birthweight ~ smoker + alcohol + nprevist, data = BW, se_type = "stata")
reg3 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried, data = BW,
                 se_type = "stata")
reg4 = lm_robust(birthweight ~ smoker + alcohol + nprevist + unmarried + age + educ,
                 data = BW, se_type = "stata")
```

E7.1 Using the `Birthweight_Smoking.csv` introduced in E5.3 to answer the following questions. To begin, run three regressions:

Table 1: Birth Weight and Smoking

	Model 1	Model 2	Model 3	Model 4
(Intercept)	3432.06***	3051.25***	3134.40***	3199.43***
	(11.89)	(43.71)	(44.15)	(90.64)
smoker	−253.23***	−217.58***	−175.38***	−176.96***
	(26.81)	(26.11)	(26.83)	(27.33)
alcohol		−30.49	−21.08	−14.76
		(72.60)	(72.99)	(72.91)
nprevist		34.07***	29.60***	29.78***
		(3.61)	(3.58)	(3.60)
unmarried			−187.13***	−199.32***
			(27.68)	(30.99)
age				−2.49
				(2.45)
educ				0.24
				(5.53)
R^2	0.03	0.07	0.09	0.09
Adj. R^2	0.03	0.07	0.09	0.09
Num. obs.	3000	3000	3000	3000
RMSE	583.73	570.47	565.70	565.76

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

(a) What is the value of the estimated effect of smoking on birth weight in each of the regressions?

Table 1: Birth Weight and Smoking

	Model 1	Model 2	Model 3	Model 4
(Intercept)	3432.06***	3051.25***	3134.40***	3199.43***
	(11.89)	(43.71)	(44.15)	(90.64)
smoker	−253.23***	−217.58***	−175.38***	−176.96***
	(26.81)	(26.11)	(26.83)	(27.33)
alcohol		−30.49	−21.08	−14.76
		(72.60)	(72.99)	(72.91)
nprevist		34.07***	29.60***	29.78***
		(3.61)	(3.58)	(3.60)
unmarried			−187.13***	−199.32***
			(27.68)	(30.99)
age				−2.49
				(2.45)
educ				0.24
				(5.53)
R^2	0.03	0.07	0.09	0.09
Adj. R^2	0.03	0.07	0.09	0.09
Num. obs.	3000	3000	3000	3000
RMSE	583.73	570.47	565.70	565.76

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(b) Construct a 95% confidence interval for the effect of smoking on birth weight, using each of the regressions.

```
> # From Model 1 the 95% CI is -305.8 to -200.7.
> confint(reg1)
                2.5 %     97.5 %
(Intercept) 3408.746 3455.3744
smoker      -305.797 -200.6597
> # From Model 2 the 95% CI is -268.8 to -166.4.
> confint(reg2)
                 2.5 %      97.5 %
(Intercept) 2965.53519 3136.96195
smoker      -268.77078 -166.38937
alcohol     -172.83573  111.85314
nprevist      26.99487   41.14496
> # From Model 3 the 95% CI is -228.0 to -122.8.
> confint(reg3)
                 2.5 %      97.5 %
(Intercept) 3047.83544 3220.96461
smoker      -227.97772 -122.77609
alcohol     -164.20321  122.03628
nprevist      22.57766   36.62742
unmarried   -241.40139 -132.86508
```

(c) Does the coefficient on `smoker` in regression (1) suffer from omitted variable bias? Explain.

Table 1: Birth Weight and Smoking

	Model 1	Model 2	Model 3	Model 4
(Intercept)	3432.06***	3051.25***	3134.40***	3199.43***
	(11.89)	(43.71)	(44.15)	(90.64)
smoker	−253.23***	−217.58***	−175.38***	−176.96***
	(26.81)	(26.11)	(26.83)	(27.33)
alcohol		−30.49	−21.08	−14.76
		(72.60)	(72.99)	(72.91)
nprevist		34.07***	29.60***	29.78***
		(3.61)	(3.58)	(3.60)
unmarried			−187.13***	−199.32***
			(27.68)	(30.99)
age				−2.49
				(2.45)
educ				0.24
				(5.53)
R^2	0.03	0.07	0.09	0.09
Adj. R^2	0.03	0.07	0.09	0.09
Num. obs.	3000	3000	3000	3000
RMSE	583.73	570.47	565.70	565.76

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Yes, it seems so. The coefficient falls by roughly 30% in magnitude when additional regressors are added to Model 1. This change is substantively large and large relative to the standard error in Model 1.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(d) Does the coefficient on `smoker` in regression (2) suffer from omitted variable bias? Explain.

Table 1: Birth Weight and Smoking

	Model 1	Model 2	Model 3	Model 4
(Intercept)	3432.06***	3051.25***	3134.40***	3199.43***
	(11.89)	(43.71)	(44.15)	(90.64)
smoker	−253.23***	−217.58***	−175.38***	−176.96***
	(26.81)	(26.11)	(26.83)	(27.33)
alcohol		−30.49	−21.08	−14.76
		(72.60)	(72.99)	(72.91)
nprevist		34.07***	29.60***	29.78***
		(3.61)	(3.58)	(3.60)
unmarried			−187.13***	−199.32***
			(27.68)	(30.99)
age				−2.49
				(2.45)
educ				0.24
				(5.53)
R^2	0.03	0.07	0.09	0.09
Adj. R^2	0.03	0.07	0.09	0.09
Num. obs.	3000	3000	3000	3000
RMSE	583.73	570.47	565.70	565.76

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Yes, it seems so. The coefficient falls by roughly 20% in magnitude when unmarried is added as an additional regression. This change is substantively large and large relative to the standard error in Model 2.

(e) Consider the coefficient on `unmarried` in regression (3).

i. Construct a 95% confidence interval for the coefficient.

ii. Is the coefficient statistically significant? Explain.

iii. Is the magnitude of the coefficient large? Explain.

iv. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree? [*Hint:* Review the discussion of control variables in Section 7.5. Discuss some of the various factors that `unmarried` may be controlling for and how this affects the interpretation of its coefficient.]

i The 95% CI is -241.4 to -132.9 .

ii Yes. The 95% confidence interval does not include zero. Alternatively, the t-statistic is -6.76 which is large in absolute value than the 5% critical value of 1.96.

iii Yes. On average, birth weight is 187 grams lower for unmarried mothers.

iv As the question suggests, unmarried is a control variable that captures the effects of several factors that differ between married and unmarried mothers such as age, education, income, diet and other health factors, and so forth.

(f) Consider the various other control variables in the data set. Which do you think should be included in the regression? Using a table like Table 7.1, examine the robustness of the confidence interval you constructed in (b). What is a reasonable 95% confidence interval for the effect of smoking on birth weight?

I have added on additional regression in the table that includes `age` and `educ` (years of education). The coefficient on `smoker` is very similar to its value in regression Model 3. See Model 4 in Table 1.

E7.2 In the empirical exercises on earnings and height until last week, you estimated a relatively large and statistically significant effect of a worker's height on his or her earnings. One explanation for this result is omitted variable bias: Height is correlated with an omitted factor that affects earnings. For example, Case and Paxson (2008) suggest that cognitive ability (or intelligence) is the omitted factor. The mechanism they describe is straightforward: Poor nutrition and other harmful environmental factors in utero and in early childhood have, on average, deleterious effects on both cognitive and physical development. Cognitive ability affects earnings later in life and thus is an omitted variable in the regression.

(a) Suppose that the mechanism described above is correct. Explain how this leads to omitted variable bias in the OLS regression of `earnings` on `height`. Does the bias led the estimated slope to be too large or too small? [*Hint:* Review Equation (6.1) in SW.]

From Key Concept 6.1, omitted variable bias arises if X is correlated with the omitted variable and the omitted variable is a determinant of the dependent variable, Y. The mechanism described in the problem explains why height (X) and cognitive ability (the omitted variable) are correlated and why cognitive ability is a determinant of earning (Y). The mechanism suggests that height and cognitive ability are positively correlated and that cognitive ability has a positive effect on earnings. Thus, X will be positively correlated with the error leading to a positive bias in the estimated coefficient.

If the mechanism described above is correct, the estimated effect of height on earnings should disappear if a variable measuring cognitive ability is included in the regression. Unfortunately, there is not a direct measure of cognitive ability in the dataset, but the dataset does include "years of education" for each individual. Because students with higher cognitive ability are more likely to attend school longer, years of education might serve as a control variable for cognitive ability. In this case, including education in the regression will eliminate, or at least attenuate, the omitted variable bias problem.

Use the years of education variable, `educ`, to construct four indicator (dummy) variables for whether a worker has less than a high school diploma (`lt_hs` $= 1$ if `educ` < 12, and 0, otherwise), a high school diploma (`hs` $= 1$ if `educ` $= 12$, and 0, otherwise), some college, (`some_col` $= 1$ if $12 <$ `educ` < 16, and 0, otherwise), or a bachelor's degree or higher (`college` $= 1$ if `educ` ≥ 16, and 0, otherwise).

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial05")
EH <- read_csv("Earnings_and_Height.csv") %>%
  mutate(lt_hs = as.numeric(educ < 12), hs = as.numeric(educ == 12),
         col = as.numeric(educ >= 16), some_col = 1 - lt_hs - hs - col)
attach(EH)
```

(b) Focusing first on women only, run two regressions: (1) **earnings** on **height**, and (2) **earnings** on **height**, including **lt_hs**, **hs**, and **some_col** as control variables.

```
reg1 = lm_robust(earnings ~ height, data = subset(EH, sex != "1:male"), se_type = "stata")
reg2 = lm_robust(earnings ~ height + lt_hs + hs + some_col,
                 data = subset(EH, sex != "1:male"), se_type = "stata")
reg3 = lm_robust(earnings ~ height, data = subset(EH, sex == "1:male"), se_type = "stata")
reg4 = lm_robust(earnings ~ height + lt_hs + hs + some_col,
                 data = subset(EH, sex == "1:male"), se_type = "stata")
```

Table 2: Earnings and Height

	Model 1 (Women)	Model 2 (Women)	Model 3 (Men)	Model 4 (Men)
(Intercept)	12650.86*	50749.52***	−43130.34***	9862.74
	(6299.15)	(6003.82)	(6925.01)	(6541.32)
height	511.22***	135.14	1306.86***	744.68***
	(97.58)	(92.32)	(98.86)	(92.26)
lt_hs		−31857.81***		−31400.49***
		(834.96)		(869.70)
hs		−20417.89***		−20345.85***
		(637.81)		(701.64)
some_col		−12649.07***		−12610.92***
		(716.59)		(797.80)
R^2	0.00	0.14	0.02	0.17
Adj. R^2	0.00	0.14	0.02	0.17
Num. obs.	9974	9974	7896	7896
RMSE	26800.90	24917.38	26671.29	24623.22

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

i. Compare the estimated coefficients on `height` in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.

Table 2: Earnings and Height

	Model 1 (Women)	Model 2 (Women)	Model 3 (Men)	Model 4 (Men)
(Intercept)	12650.86*	50749.52***	−43130.34***	9862.74
	(6299.15)	(6003.82)	(6925.01)	(6541.32)
height	511.22***	135.14	1306.86***	744.68***
	(97.58)	(92.32)	(98.86)	(92.26)
lt_hs		−31857.81***		−31400.49***
		(834.96)		(869.70)
hs		−20417.89***		−20345.85***
		(637.81)		(701.64)
some_col		−12649.07***		−12610.92***
		(716.59)		(797.80)
R^2	0.00	0.14	0.02	0.17
Adj. R^2	0.00	0.14	0.02	0.17
Num. obs.	9974	9974	7896	7896
RMSE	26800.90	24917.38	26671.29	24623.22

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

i The estimated coefficient on height falls by approximately 75%, from 511 to 135 when the education variables are added as control variables in the regression. This is consistent with positive omitted bias in Model 1.

ii. Regression (2) omits the control variable `college`. Why?

ii The variable `college` is perfectly collinear with other education regressors and the constant regressor.

iii. Test the joint null hypothesis that the coefficients on the education variables are equal to zero.

```
> linearHypothesis(reg2, c("lt_hs = 0", "hs = 0", "some_col = 0"), test = c("F"))
Linear hypothesis test

Hypothesis:
lt_hs = 0
hs = 0
some_col = 0

Model 1: restricted model
Model 2: earnings ~ height + lt_hs + hs + some_col

  Res.Df Df      F    Pr(>F)
1   9972
2   9969  3 577.93 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iii The F-statistic is 578, and the corresponding p-value is ≈ 0.00. Therefore, the null hypothesis that the coefficients on the education variables are jointly equal to zero is rejected at the 1% significance level.

iv. Discuss the values of the estimated coefficients on `lt_hs`, `hs`, and `some_col`. (Each of the estimated coefficients is negative, and the coefficient on `lt_hs` is more negative than the coefficient on `hs`, which in turn is more negative than the coefficient on `some_col`. Why? What do the coefficients measure?)

iv The coefficients measure the effect of education on earnings relative to the omitted category, which is `college`. Thus, the estimated coefficient on the "Less than High School" regressor implies that workers with less than a high school education on average earn $31,858 less per year than a college graduate; a worker with a high school education on average earns $20,418 less per year than a college graduate; a worker with a some college on average earns $12,649 less per year than a college graduate.

(c) Repeat (b), using data for men.

i. Compare the estimated coefficients on `height` in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.

i. Compare the estimated coefficients on `height` in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.

Table 2: Earnings and Height

	Model 1 (Women)	Model 2 (Women)	Model 3 (Men)	Model 4 (Men)
(Intercept)	12650.86*	50749.52***	−43130.34***	9862.74
	(6299.15)	(6003.82)	(6925.01)	(6541.32)
height	511.22***	135.14	1306.86***	744.68***
	(97.58)	(92.32)	(98.86)	(92.26)
lt_hs		−31857.81***		−31400.49***
		(834.96)		(869.70)
hs		−20417.89***		−20345.85***
		(637.81)		(701.64)
some_col		−12649.07***		−12610.92***
		(716.59)		(797.80)
R^2	0.00	0.14	0.02	0.17
Adj. R^2	0.00	0.14	0.02	0.17
Num. obs.	9974	9974	7896	7896
RMSE	26800.90	24917.38	26671.29	24623.22

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

i The estimated coefficient on `height` falls by approximately 50%, from 1307 to 745. This is consistent with positive omitted bias in the simple regression, Model 1.

ii. Regression (2) omits the control variable `college`. Why?

ii The variable `college` is perfectly collinear with other education regressors and the constant regressor.

iii. Test the joint null hypothesis that the coefficients on the education variables are equal to zero.

```
> linearHypothesis(reg4, c("lt_hs = 0", "hs = 0", "some_col = 0"), test = c("F"))
Linear hypothesis test

Hypothesis:
lt_hs = 0
hs = 0
some_col = 0

Model 1: restricted model
Model 2: earnings ~ height + lt_hs + hs + some_col

  Res.Df Df      F    Pr(>F)
1   7894
2   7891  3 500.92 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iii The F-statistic is 500.9, and the corresponding p-value is ≈ 0.00. Therefore, the null hypothesis that the coefficients on the education variables are jointly equal to zero is rejected at the 1% significance level.

iv. Discuss the values of the estimated coefficients on `lt_hs`, `hs`, and `some_col`. (Each of the estimated coefficients is negative, and the coefficient on `lt_hs` is more negative than the coefficient on `hs`, which in turn is more negative than the coefficient on `some_col`. Why? What do the coefficients measure?)

iv The coefficients measure the effect of education on earnings relative to the omitted category, which is `college`. Thus, the estimated coefficient on the "Less than High School" regressor implies that workers with less than a high school education on average earn $31,400 less per year than a college graduate; a worker with a high school education on average earns $20,346 less per year than a college graduate; a worker with a some college on average earns $12,611 less per year than a college graduate.

Thank you

Francisco Tavares Garcia
Academic Tutor | School of Economics
tavaresgarcia.github.io

Reference

Stock, J. H., & Watson, M. W. (2019). Introduction to Econometrics, Global Edition, 4th edition. Pearson Education Limited.