THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

ECON2300 - Introductory Econometrics

Tutorial 4: Linear Regression with Multiple Regressors

Tutor: Francisco Tavares Garcia

Quiz 3 is available!

Quiz 3 Available

Item is not available.

Posted on: Monday, 24 March 2025 06:00:00 o'clock AEST

The following test has been made available in Quizzes: Problem Solving, Data Analysis and Short Report: ECON2300 Quiz 3 (Semester 1, 2025).

The due date for Quiz 3 is 4pm, Thursday, 27 March 2025.

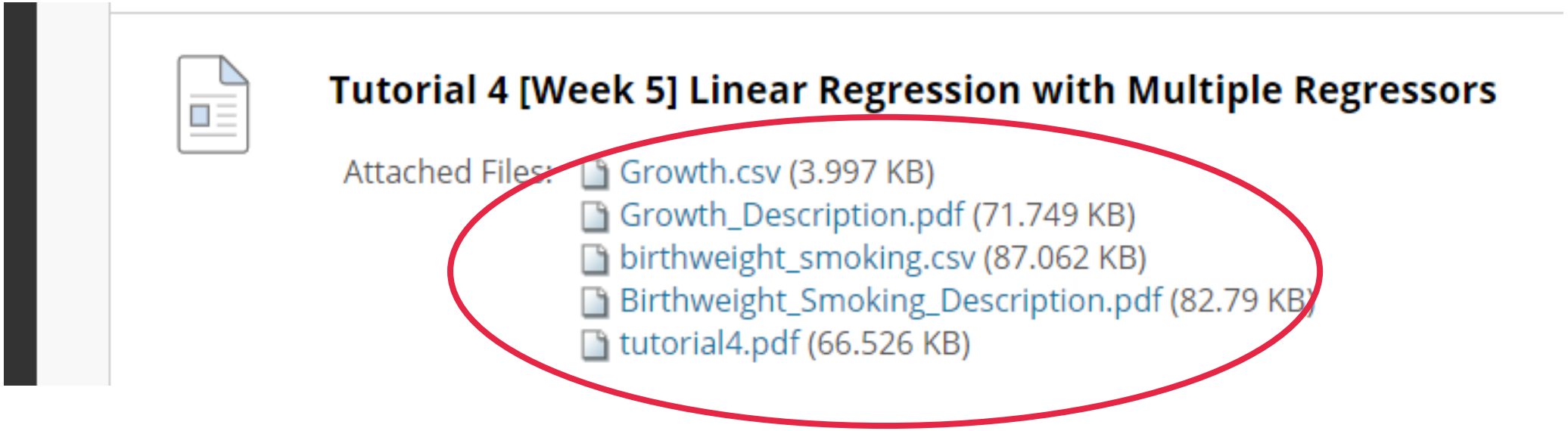
Important information about variables used in the Questions:

Questions 1-4 will study the effect of the student-computer ratio on test scores. We use two variables: (1) students per computer ratio (STC) and (2) 8th grade score (TestScore), which is the sum of math, English and science in the ataset "MASchool".

Questions 5-7 will seek to learn the effect of student-teacher ratio on test scores, we use two variables: (1) student-teacher ratio (STR) and (2) 8th grade score (TestScore), which is the sum of math, English and science in the dataset "MASchool".

Course Link /Assessment/Quizzes: Problem Solving, Data Analysis and Short Report/ECON2300 Quiz 3 (Semester 1, 2025)

- Download the files for tutorial 04 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.

E6.1 Use the `Birthweight_Smoking.csv` introduced in E5.3 to answer the following questions.

	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

E6.1 Use the `Birthweight_Smoking.csv` introduced in E5.3 to answer the following questions.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nprevist	alcohol	tripre1	tripre2	tripre3	tripre0	birthweig	smoker	unmarriec	educ	age	drinks
2	12	0	1	0	0	0	4253	1	1	12	27	0
3	5	0	0	1	0	0	3459	0	0	16	24	0
4	12	0	1	0	0	0	2920	1	0	11	23	0
5	13	0	1	0	0	0	2600	0	0	17	28	0
6	9	0	1	0	0	0	3742	0	0	13	27	0
7	11	0	1	0	0	0	3420	0	0	16	33	0
8	12	0	1	0	0	0	2325	1	0	14	24	0
9	10	0	1	0	0	0	4536	0	0	13	38	0
10	13	0	1	0	0	0	2850	0	0	17	29	0

E6.1 Use the `Birthweight_Smoking.csv` introduced in E5.3 to answer the following questions.

```
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(psych)      # package containing many functions useful for data analysis
```

SW E6.1

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial04")
BW <- read_csv("birthweight_smoking.csv")
attach(BW)
```

E6.1 Use the `Birthweight_Smoking.csv` introduced in E5.3 to answer the following questions.

(a) Regress `birthweight` on `smoker`.

```
reg1 = lm_robust(birthweight ~ smoker, data = BW, se_type = "stata")
summary(reg1)
```

```
Call:
lm_robust(formula = birthweight ~ smoker, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   3432.1      11.89 288.638 0.000e+00   3408.7   3455.4 2998
smoker        -253.2      26.81  -9.445 6.903e-21   -305.8   -200.7 2998

Multiple R-squared:  0.0286 ,   Adjusted R-squared:  0.02828
F-statistic: 89.21 on 1 and 2998 DF,  p-value: < 2.2e-16
```

The estimated regression is

$$\widehat{\texttt{birthweight}} = \underset{(11.89)}{3432.1} - \underset{(26.81)}{253.2} \times \texttt{smoker}$$

The estimated effect of smoking on birthweight is -253.2 grams.

(b) Regress `birthweight` on `smoker`, `alcohol`, and `nprevist`.

i. Using the two conditions for omitted variable bias, explain why the exclusion of `alcohol` and `nprevist` could lead to omitted variable bias in the regression estimated in (a)

```
reg2 = lm_robust(birthweight ~ smoker + alcohol + nprevist, data = BW, se_type = "stata")
summary(reg2)
```

```
Call:
lm_robust(formula = birthweight ~ smoker + alcohol + nprevist,
    se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)  3051.25     43.714  69.800 0.000e+00  2965.54  3136.96 2996
smoker       -217.58     26.108  -8.334 1.175e-16  -268.77  -166.39 2996
alcohol       -30.49     72.597  -0.420 6.745e-01  -172.84   111.85 2996
nprevist       34.07      3.608   9.442 7.109e-21    26.99    41.14 2996

Multiple R-squared:  0.07285 ,  Adjusted R-squared:  0.07192
F-statistic: 59.48 on 3 and 2996 DF,  p-value: < 2.2e-16
```

i Smoking may be correlated with both alcohol and the number of prenatal doctor visits, thus satisfying (1) in Key Concept 6.1. Moreover, both alcohol consumption and the number of doctor visits may have their own independent affects on birth weight, thus satisfying (2) in Key Concept 6.1.

ii. Is the estimated effect of smoking on birth weight substantially different from the regression that excludes `alcohol` and `nprevist`? Does the regression in (a) seem to suffer from omitted variable bias?

```
Call:
lm_robust(formula = birthweight ~ smoker + alcohol + nprevist,
    se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)  3051.25     43.714  69.800 0.000e+00  2965.54  3136.96 2996
smoker       -217.58     26.108  -8.334 1.175e-16  -268.77  -166.39 2996
alcohol       -30.49     72.597  -0.420 6.745e-01  -172.84   111.85 2996
nprevist       34.07      3.608   9.442 7.109e-21    26.99    41.14 2996

Multiple R-squared:  0.07285 ,  Adjusted R-squared:  0.07192
F-statistic: 59.48 on 3 and 2996 DF,  p-value: < 2.2e-16
```

ii The estimate is somewhat smaller: it has fallen to 217.6 grams from 253.2 grams, so the regression in (a) may suffer from omitted variable bias.

iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.

```
predict(reg2, newdata = data.frame(smoker = 1, alcohol = 0, nprevist = 8))
```

```
##        1
## 3106.228
```

iii

$$\widehat{\texttt{birthweight}} = 3051.25 - 217.58 \times 1 - 30.49 \times 0 + 34.07 \times 8 = 3106.23$$

iv. Compute R^2 and $\overline{R}^2$. Why are they so similar?

iv They are nearly identical because the sample size is very large ($n = 3000$).

R^2 and adjusted R^2

- The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$, $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$

- The R^2 always increases when you add another regressor – a bit of a problem for a measure of "fit"
- The $\overline{R}^2$ (the "adjusted R^2") corrects this problem by "penalizing" you for including another regressor – the $\overline{R}^2$ does not necessarily increase when you add another regressor.

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS}$$

Note that $\overline{R}^2 \leq R^2$, however if n is large the two will be very close.

(c) An alternative way to control for prenatal visits is to use the binary variables **tripre0** through **tripre3**. Regress **birthweight** on **smoker**, **alcohol**, **tripre0**, **tripre2**, and **tripre3**.

i. Why is **tripre1** excluded from the regression? What would happen if you included it in the regression?

```
reg3 = lm_robust(birthweight ~ smoker + alcohol + tripre0 + tripre2 + tripre3,
                 data = BW, se_type = "stata")
summary(reg3)
```

```
Call:
lm_robust(formula = birthweight ~ smoker + alcohol + tripre0 +
    tripre2 + tripre3, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error  t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   3454.5      12.48 276.7697 0.000e+00   3430.1  3479.02 2994
smoker        -228.8      26.55  -8.6199 1.068e-17   -280.9  -176.79 2994
alcohol        -15.1      69.70  -0.2166 8.285e-01   -151.8   121.57 2994
tripre0       -698.0     146.58  -4.7617 2.011e-06   -985.4  -410.56 2994
tripre2       -100.8      31.55  -3.1958 1.409e-03   -162.7   -38.97 2994
tripre3       -137.0      67.70  -2.0231 4.315e-02   -269.7    -4.22 2994

Multiple R-squared:  0.04647 ,  Adjusted R-squared:  0.04487
F-statistic: 23.22 on 5 and 2994 DF,  p-value: < 2.2e-16
```

(c) An alternative way to control for prenatal visits is to use the binary variables `tripre0` through `tripre3`. Regress `birthweight` on `smoker`, `alcohol`, `tripre0`, `tripre2`, and `tripre3`.

i. Why is `tripre1` excluded from the regression? What would happen if you included it in the regression?

i `tripre1` is omitted to avoid perfect multicollinearity. (`tripre0` + `tripre1` + `tripre2` + `tripre3` = 1, the value of the "constant" regressor that determines the intercept). The regression would not run, or the software will report results from an arbitrary normalization if `tripre0`, `tripre1`, `tripre2`, `tripre3`, and the constant term all included in the regression.

tripre0	tripre1	tripre2	tripre3	Sum
1	0	0	0	1
0	1	0	0	1
0	0	1	0	1
0	0	0	1	1

ii. The estimated coefficient on `tripre0` is large and negative. What does this coefficient measure? Interpret its value.

ii Babies born to women who had no prenatal doctor visits (`tripre0` $= 1$) had birth weights that on average were 698.0 grams ($\approx$ 1.5 lbs) lower than babies from others who saw a doctor during the first trimester (`tripre1` $= 1$).

iii. Interpret the value of the estimated coefficients on `tripre2` and `tripre3`

iii Babies born to women whose first doctor visit was during the second trimester (`tripre2` $= 1$) had birth weights that on average were 100.8 grams ($\approx$ 0.2 lbs) lower than babies from others who saw a doctor during the first trimester (`tripre1` $= 1$). Babies born to women whose first doctor visit was during the third trimester (`tripre3` $= 1$) had birth weights that on average were 137 grams ($\approx$ 0.3 lbs) lower than babies from others who saw a doctor during the first trimester (`tripre1` $= 1$).

iv. Does the regression in (c) explain a larger fraction of the variance in birth weight than the regression in (b)?

iv No. The R^2 for the regression in (c) is 0.046 (4.6% variance in birth weight is explained by the regression model), while the R^2 for the regression in (b) is 0.073.

E6.2 Using the dataset `Growth.csv`, but excluding the data for Malta, run a regression of `growth` on `tradeshare`.

	A	B	C	D	E	F	G	H
1	country_n	growth	oil	rgdp60	tradeshar	yearsscho	rev_coups	assasinatic
2	India	1.915168	0	765.9998	0.140502	1.45	0.133333	0.866667
3	Argentina	0.617645	0	4462.002	0.156623	4.99	0.933333	1.933333
4	Japan	4.304759	0	2954	0.157703	6.71	0	0.2
5	Brazil	2.930097	0	1784	0.160405	2.89	0.1	0.1
6	United Sta	1.712265	0	9895.004	0.160815	8.66	0	0.433333
7	Banglades	0.708263	0	951.9998	0.221458	0.79	0.306481	0.175
8	Spain	2.880327	0	3123.002	0.299406	3.8	0.066667	1.433333
9	Colombia	2.227014	0	1684	0.313073	2.97	0.1	0.766667
10	Peru	0.060206	0	2019	0.324613	3.02	0.266667	0.566667
11	Haiti	-0.65793	0	923.9999	0.324746	0.7	0.374074	0.2
12	Australia	1.975147	0	7782.002	0.329479	9.03	0	0.066667
13	Italy	2.932982	0	4564.001	0.330022	4.56	0.033333	1.2
14	Greece	3.22405	0	2093	0.337879	4.37	0.166667	0.166667
15	France	2.431281	0	5823.001	0.339706	4.65	0	0.3
16	Zaire	-2.81194	0	488.9999	0.352318	0.54	0.148148	0.055556
17	Uruguay	1.025309	0	3968	0.358857	5.07	0	0.166667

Variable Definitions

Variable	Definition
Country_name	Name of country
growth	Average annual percentage growth of real Gross Domestic Product (GDP)* from 1960 to 1995.
rgdp60	The value of GDP* per capita in 1960, converted to 1960 US dollars
tradehare	The average share of trade in the economy from 1960 to 1995, measured as the sum of exports plus imports, divided by GDP; that is, the average value of $(X + M)/GDP$ from 1960 to 1995, where X = exports and M = imports (both X and M are positive).
yearsshcool	Average number of years of schooling of adult residents in that country in 1960
rev_coups	Average annual number of revolutions, insurrections (successful or not) and coup d'etats in that country from 1960 to 1995
assasinations	Average annual number of political assassinations in that country from 1960 to 1995 (per million population)
oil	= 1 if oil accounted for at least half of exports in 1960 = 0 otherwise

E6.2 Using the dataset **Growth.csv**, but excluding the data for Malta, run a regression of **growth** on **tradeshare**.

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial04")
Growth <- read_csv("Growth.csv") %>%
  filter(country_name != "Malta")
attach(Growth)
```

E6.2 Using the dataset `Growth.csv`, but excluding the data for Malta, run a regression of `growth` on `tradeshare`.

(a) Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series `growth`, `tradeshare`, `yearsschool`, `oil`, `rev_coups`, `assassinations`, and `rgdp60`.

```
describe(data.frame(growth, tradeshare, yearsschool, oil, rev_coups, assasinations,
                    rgdp60), fast = T)
```

```
##                vars  n    mean      sd    min     max   range     se
## growth            1 64    1.87    1.82  -2.81    7.16    9.97   0.23
## tradeshare        2 64    0.54    0.23   0.14    1.13    0.99   0.03
## yearsschool       3 64    3.96    2.55   0.20   10.07    9.87   0.32
## oil               4 64    0.00    0.00   0.00    0.00    0.00   0.00
## rev_coups         5 64    0.17    0.23   0.00    0.97    0.97   0.03
## assasinations     6 64    0.28    0.49   0.00    2.47    2.47   0.06
## rgdp60            7 64 3130.81 2522.98 367.00 9895.00 9528.00 315.37
```

(b) Run a regression of `growth` on `tradeshare`, `yearsschool`, `oil`, `rev_coups`, `assassinations`, and `rgdp60`. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

```
reg4 = lm_robust(growth ~ tradeshare + yearsschool + oil + rev_coups + assasinations +
                 rgdp60, data = Growth, se_type = "stata")
summary(reg4)
```

```
Call:
lm_robust(formula = growth ~ tradeshare + yearsschool + oil +
    rev_coups + assasinations + rgdp60, se_type = "stata")

Standard error type:  HC1

Coefficients: (1 not defined because the design matrix is rank deficient)
               Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper DF
(Intercept)   0.6268915  0.8690927  0.7213 4.736e-01 -1.1127866  2.3665695 58
tradeshare    1.3408193  0.8819886  1.5202 1.339e-01 -0.4246727  3.1063114 58
yearsschool   0.5642445  0.1294907  4.3574 5.442e-05  0.3050408  0.8234482 58
oil                  NA         NA      NA        NA         NA         NA NA
rev_coups    -2.1504256  0.8746010 -2.4588 1.695e-02 -3.9011297 -0.3997215 58
assasinations 0.3225844  0.3803478  0.8481 3.999e-01 -0.4387644  1.0839333 58
rgdp60       -0.0004613  0.0001215 -3.7968 3.529e-04 -0.0007045 -0.0002181 58

Multiple R-squared:  0.2911 ,   Adjusted R-squared:   0.23
F-statistic:    NA on 5 and 58 DF,  p-value: NA
```

(b) Run a regression of `growth` on `tradeshare`, `yearsschool`, `oil`, `rev_coups`, `assassinations`, and `rgdp60`. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

```
predict(reg4, newdata = data.frame(tradeshare = mean(tradeshare),
                                   yearsschool = mean(yearsschool),
                                   oil = mean(oil),
                                   rev_coups = mean(rev_coups),
                                   assasinations = mean(assasinations),
                                   rgdp60 = mean(rgdp60)))
```

```
##       1
## 1.86912
```

Use the sample averages for the regressors in (a) above. The predicted growth rate at the mean values for all regressors is 1.87.

(c) Repeat (b) but now assume that the country's value for **tradeshare** is one standard deviation above the mean.

```
predict(reg4, newdata = data.frame(tradeshare = mean(tradeshare) + sd(tradeshare),
                                   yearsschool = mean(yearsschool),
                                   oil = mean(oil),
                                   rev_coups = mean(rev_coups),
                                   assasinations = mean(assasinations),
                                   rgdp60 = mean(rgdp60)))
```

```
##        1
## 2.175273
```

Use the standard deviation in (a) above. The resulting predicted value is 2.18.

(d) Why is `oil` omitted from the regression? What would happen if it were included?

The variable "oil" takes on the value of 0 for all 64 countries in the sample. This would generate perfect multicollinearity, since $\texttt{oil}_i + 1 = 1$, and hence the variable is a linear combination of one of the regressors, namely the constant (interccept).

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

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.