The University of Queensland
Australia
CREATE CHANGE

ECON2300 - Introductory Econometrics

Tutorial 3: Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

Tutor: Francisco Tavares Garcia

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Quiz 01 is available!

Posted on: Monday, 14 August 2023 06:00:00 o'clock AEST

Dear ECON2300 Students,

Quiz 1 is now available in the "Quizzes: Problem Solving Exercises" folder, which you can access via the Assessment tab.

The due date for Quiz 1 is **Friday, August 18, 2023, 4pm**

Please read all instructions carefully before commencing the quiz. For convenience, a copy of the quiz instructions has been presented below.

===

Instructions:

Please pay close attention to the number of decimal places required (if any) for each answer. The required number of decimal places may differ from question to question.

Avoid rounding during intermediate calculations where possible.

The quiz is not timed. This means that you can open the quiz and return to it as many times as you need to (provided that you do not click submit).

There is only one attempt for this quiz.

The quiz is marked out of 7, but will contribute 10% towards your final grade if it is among the highest 3 of your 5 Quiz scores across the semester.

The closing time for this quiz is **4pm on Friday, August 18, 2023**. Please make sure that you have submitted your answers by this time. Remember that **you must click submit** before the deadline for your quiz to be marked.

Please Note: If you encounter any technical issues with the quiz, please email the CML coordinator at cml.2300@uq.edu.au. Do not email quiz issues to the Course Coordinator or Course Administrator. Otherwise there may be a delay in responding to your enquiry.

Posted by: Dominic Byrne
Posted to: [ECON2300] Introductory Econometrics (St Lucia & external). Semester 2, 2023 ECON2300S_7360_61825

Adm · Tut 03 · E5.1 · E5.2 · E5.3

- Download the files for tutorial 03 from Blackboard,
- save them into a folder for this tutorial.

Tutorial 3 [Week 4] Simple Linear Regression - Inference

Attached Files:
- Earnings_and_Height_Description.pdf (111.741 KB)
- Earnings_and_Height.csv (1.651 MB)
- Growth_Description.pdf (71.749 KB)
- Growth.csv (3.997 KB)
- tutorial3.pdf (75.847 KB)
- Birthweight_Smoking_Description.pdf (82.79 KB)
- birthweight_smoking.csv (87.062 KB)

- Copy the code from Codeshare,
- https://codeshare.io/tut03
- Paste the code in a new script in RStudio,
- Save the script in the same folder as the data.

E5.1 The file `Earnings_and_Height.csv` contains data on earnings, height, and other characteristics of a random sample of U.S. workers. See `Earnings_and_Height_Description.pdf` for a detailed description of the data. Carry out the following exercises.

	A	B	C	D	E	F	G	H	I	J	K
1	sex	age	mrd	educ	cworker	region	race	earnings	height	weight	occupation
2	0:female	48	1:Married	13	1:Private (	3:South	non-hispa	84054.75	65	133	1
3	0:female	41	6:Never M	12	1:Private (	2:Midwes	non-hispa	14021.4	65	155	1
4	0:female	26	1:Married	16	1:Private (	1:Northea	non-hispa	84054.75	60	108	1
5	0:female	37	1:Married	16	1:Private (	2:Midwes	non-hispa	84054.75	67	150	1
6	0:female	35	6:Never M	16	1:Private (	1:Northea	non-hispa	28560.39	68	180	1
7	0:female	25	6:Never M	15	1:Private (	4:West	non-hispa	23362.87	63	101	1
8	0:female	29	1:Married	16	1:Private (	2:Midwes	non-hispa	38925.34	67	150	1
9	0:female	44	3:Divorced	18	3:State Gc	4:West	non-hispa	84054.75	65	125	1
10	0:female	50	6:Never M	14	2:Fed Gov	3:South	non-hispa	84054.75	67	129	1
11	0:female	38	1:Married	12	4:Local Go	3:South	non-hispa	84054.75	66	110	1
12	0:female	30	1:Married	12	1:Private (	4:West	non-hispa	84054.75	65	110	1
13	0:female	29	3:Divorced	18	2:Fed Gov	3:South	non-hispa	38925.34	68	135	1
14	0:female	26	1:Married	16	1:Private (	1:Northea	non-hispa	84054.75	65	123	1
15	0:female	50	1:Married	12	1:Private (	1:Northea	non-hispa	49430.11	63	132	1
16	0:female	65	3:Divorced	16	4:Local Go	4:West	hispanic	16081.59	65	110	1
17	0:female	45	6:Never M	17	2:Fed Gov	4:West	non-hispa	84054.75	71	202	1
18	0:female	26	6:Never M	16	1:Private (	3:South	non-hispa	23362.87	66	130	1
19	0:female	57	3:Divorced	12	2:Fed Gov	3:South	non-hispa	44152.16	68	220	1
20	0:female	40	3:Divorced	16	1:Private (	4:West	non-hispa	84054.75	66	195	1
21	0:female	36	1:Married	12	1:Private (	3:South	non-hispa	49430.11	68	135	1
22	0:female	60	1:Married	15	4:Local Go	3:South	non-hispa	84054.75	64	160	1
23	0:female	32	1:Married	12	1:Private (	2:Midwes	non-hispa	33712.97	65	115	1
24	0:female	33	1:Married	12	1:Private (	1:Northea	non-hispa	44152.16	61	125	1

Variable Name	Description
age	Age, in years
cworker	Class of Worker: 1 = Private company Employee 2 = Federal Government Employee 3 = State Government Employee 4 = Local Government Employee 5 = Incorporated Business Employee 6 = Self Employed
earnings	annual labor earnings, expressed in \$2012 (see Table notes)
educ	years of education
height	height without shoes (in inches)
mrd	Marital Status 1 = Married, Spouse in household 2 = Married, Spouse not in household 3 = Widowed 4 = Divorced 5 = Separated 6 = Never Married
occupation	Occupations in 15 categories: 1 = Exec/Manager 2 = Professionals 3 = Technicians 4 = Sales 5 = Administrat 6 = Household service 7 = Protective service 8 = Other Service 9 = Farming 10 = Mechanics 11 = Construction/Mining 12 = Precision production 13 = Machine Operator 14 = Transport 15 = Laborer
race	race/ethnicity 1 = non-Hispanic white 2 = non-Hispanic black 3 = Hispanic 4 = other
region	Region of the U.S. 1 = Northeast 2 = Midwest 3 = South 4 = West
sex	Sex, 1=Male, 0 = Female
weight	weight without shoes (in pounds)

E5.1 The file `Earnings_and_Height.csv` contains data on earnings, height, and other characteristics of a random sample of U.S. workers. See `Earnings_and_Height_Description.pdf` for a detailed description of the data. Carry out the following exercises.

```
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 E5.1

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial03")
EH <- read_csv("Earnings_and_Height.csv")
```

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(a) Run a regression of **earnings** on **height**.

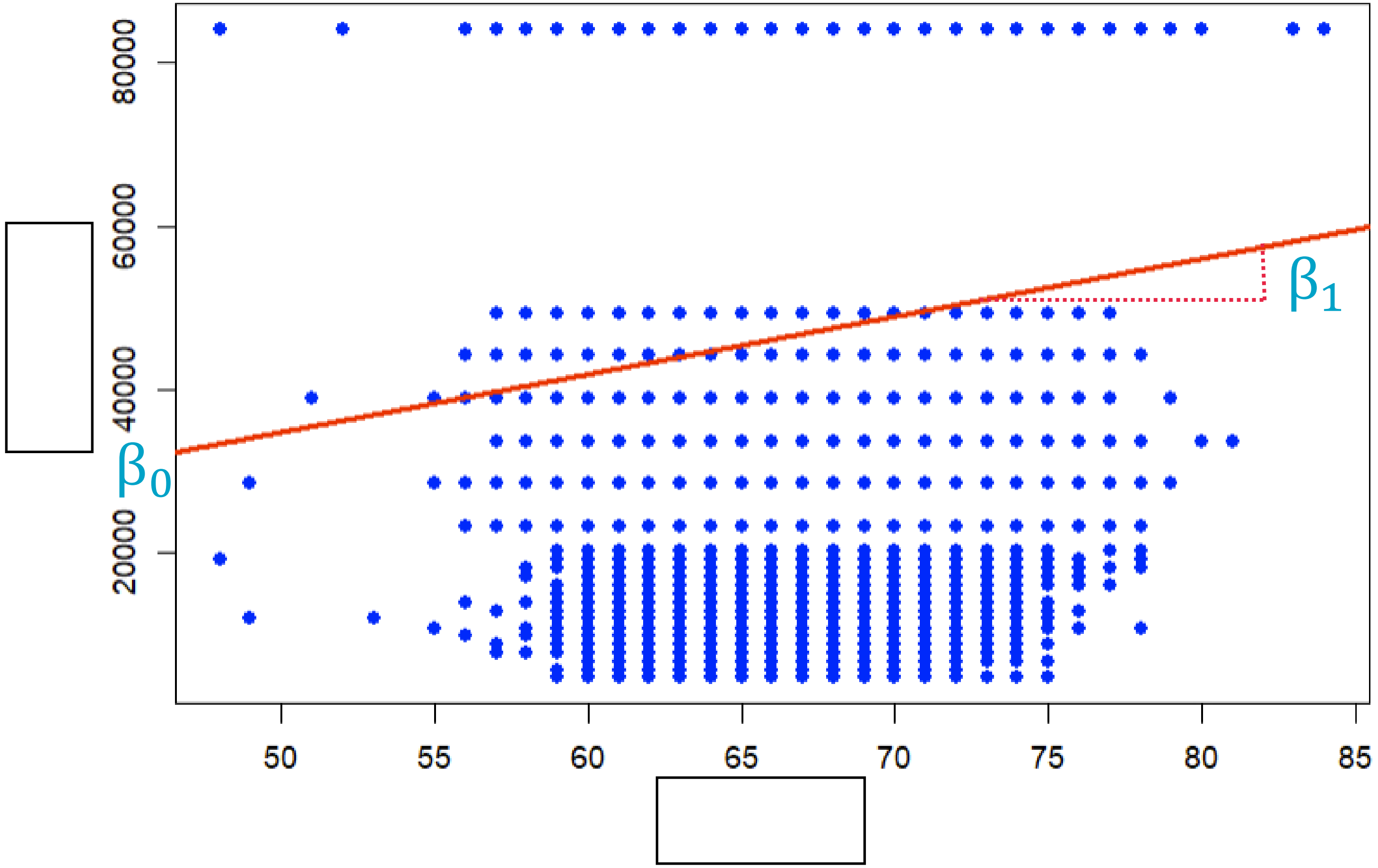
$$\square = \square + \square\square + \square$$

Which variable is the regression trying to estimate?

H
earnings
84054.75
14021.4
84054.75
84054.75
28560.39
23362.87
38925.34
84054.75
84054.75
84054.75
84054.75
38925.34
84054.75
49430.11
16081.59
84054.75
23362.87
44152.16
84054.75

I
height
65
65
60
67
68
63
67
65
67
66
65
68
65
63
65
71
66
68
66

Linear Regression Recap



(a) Run a regression of **earnings** on **height**.

i. Is the estimated slope statistically significant?

ii. Construct a 95% confidence interval for the slope coefficient.

```
reg1 = lm_robust(earnings ~ height, data = EH, se_type = "stata")
summary(reg1)
```

```
Call:
lm_robust(formula = earnings ~ height, data = EH, se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
(Intercept)   -512.7     3379.9 -0.1517 8.794e-01  -7137.6   6112.1 17868
height         707.7       50.4 14.0425 1.478e-44    608.9    806.5 17868

Multiple R-squared:  0.01088 ,  Adjusted R-squared:  0.01082
F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

The estimated regression is

$$\widehat{\texttt{earnings}} = \underset{(3379.9)}{-\,512.7} + \underset{(50.4)}{707.7} \times \texttt{height}$$

The 95% confidential interval for the slope coefficient is $707.7 \pm 1.96 \times 50.4$, or $[608.9, 806.5]$. This interval does not include $\beta_1 = 0$, so the estimated slope is significantly different from 0 at the 5% level. Alternatively, the t-statistic is $707.7/50.4 \approx 14.0$, which is greater in absolute value than the 5% critical value of 1.96. And finally, the p-value for the t-statistic is ≈ 0.000, which is smaller than 0.05.

(a) Run a regression of **earnings** on **height**.

i. Is the estimated slope statistically significant?

ii. Construct a 95% confidence interval for the slope coefficient.

(b) Repeat (a) for female observations.

```
reg2 = lm_robust(earnings ~ height, data = subset(EH, sex != "1:male"), se_type = "stata")
summary(reg2)
```

```
Call:
lm_robust(formula = earnings ~ height, data = subset(EH, sex !=
    "1:male"), se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)  12650.9    6299.15   2.008 4.463e-02    303.2  24998.5 9972
height         511.2      97.58   5.239 1.650e-07    319.9    702.5 9972

Multiple R-squared:  0.002672 , Adjusted R-squared:  0.002572
F-statistic: 27.44 on 1 and 9972 DF,  p-value: 1.65e-07
```

The estimated regression is

$$\widehat{\texttt{earnings}} = \underset{(6299.15)}{12650.9} + \underset{(97.58)}{511.2} \times \texttt{height}$$

The 95% confidential interval for the slope coefficient is $511.2 \pm 1.96 \times 97.58$. This interval does not include $\beta_{1,female} = 0$, so the estimated slope is significantly different from 0 at the 5% level.

(a) Run a regression of **earnings** on **height**.

i. Is the estimated slope statistically significant?

ii. Construct a 95% confidence interval for the slope coefficient.

(b) Repeat (a) for female observations.

(c) Repeat (a) for male observations.

```
reg3 = lm_robust(earnings ~ height, data = subset(EH, sex == "1:male"), se_type = "stata")
summary(reg3)
```

```
Call:
lm_robust(formula = earnings ~ height, data = subset(EH, sex ==
    "1:male"), se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
(Intercept)   -43130    6925.01  -6.228 4.960e-10   -56705   -29555 7894
height          1307      98.86  13.220 1.771e-39     1113     1501 7894

Multiple R-squared:  0.02086 ,  Adjusted R-squared:  0.02074
F-statistic: 174.8 on 1 and 7894 DF,  p-value: < 2.2e-16
```

The estimated regression is

$$\widehat{\texttt{earnings}} = -\underset{(6925.01)}{43130} + \underset{(98.86)}{1307} \times \texttt{height}$$

The 95% confidential interval for the slope coefficient is $1307 \pm 1.96 \times 98.86$. This interval does not include $\beta_{1,male} = 0$, so the estimated slope is significantly different from 0 at the 5% level.

(d) Test the null hypothesis that the effect of height on earnings is the same for men and women.

$\widehat{\beta}_{1,male} - \widehat{\beta}_{1,female} = 1307 - 511.2 = 795.8$, $\mathrm{SE}(\widehat{\beta}_{1,male} - \widehat{\beta}_{1,female}) = \sqrt{\mathrm{SE}(\widehat{\beta}_{1,male})^2 + \mathrm{SE}(\widehat{\beta}_{1,female})^2} = \sqrt{98.86^2 + 97.58^2} = 138.9$. The 95% confidence interval $= 795.8 \pm 1.96 \times 138.9 = [523.6, 1068]$. This interval does not include $\beta_{1,male} - \beta_{1,female} = 0$, so the estimated difference in the slopes is significantly different from 0 at the 5% level.

E5.2 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

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

E5.2 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/Tutorial03")
Growth <- read_csv("Growth.csv")
```

```
reg4 = lm_robust(growth ~ tradeshare, data = subset(Growth, country_name != "Malta"),
                 se_type = "stata")
summary(reg4)
```

```
Call:
lm_robust(formula = growth ~ tradeshare, data = subset(Growth,
    country_name != "Malta"), se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)   0.9574     0.5361   1.786  0.07899 -0.11415    2.029 62
tradeshare    1.6809     0.8656   1.942  0.05670 -0.04944    3.411 62

Multiple R-squared:  0.04466 ,  Adjusted R-squared:  0.02925
F-statistic: 3.771 on 1 and 62 DF,  p-value: 0.0567
```

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

(a) Is the estimated regression slope statistically significant? This is, can you reject the null hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ at the 5% or 1% significance level?

```
Call:
lm_robust(formula = growth ~ tradeshare, data = subset(Growth,
    country_name != "Malta"), se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)   0.9574     0.5361   1.786  0.07899 -0.11415    2.029 62
tradeshare    1.6809     0.8656   1.942  0.05670 -0.04944    3.411 62

Multiple R-squared:  0.04466 ,  Adjusted R-squared:  0.02925
F-statistic: 3.771 on 1 and 62 DF,  p-value: 0.0567
```

The fitted regression line is

$$\widehat{\texttt{growth}} = \underset{(0.54)}{0.96} + \underset{(0.87)}{1.68} \times \texttt{tradeshare}$$

The t-statistic for the slope coefficient is $t = 1.68/0.87 = 1.94$. The t-statistic is less in absolute value than the 5% and 1% critical values (1.96 and 2.58). Therefore, the null hypothesis is not rejected at the 5% or 1% levels.

(b) What is the p-value associated with the coefficient's t-statistic?

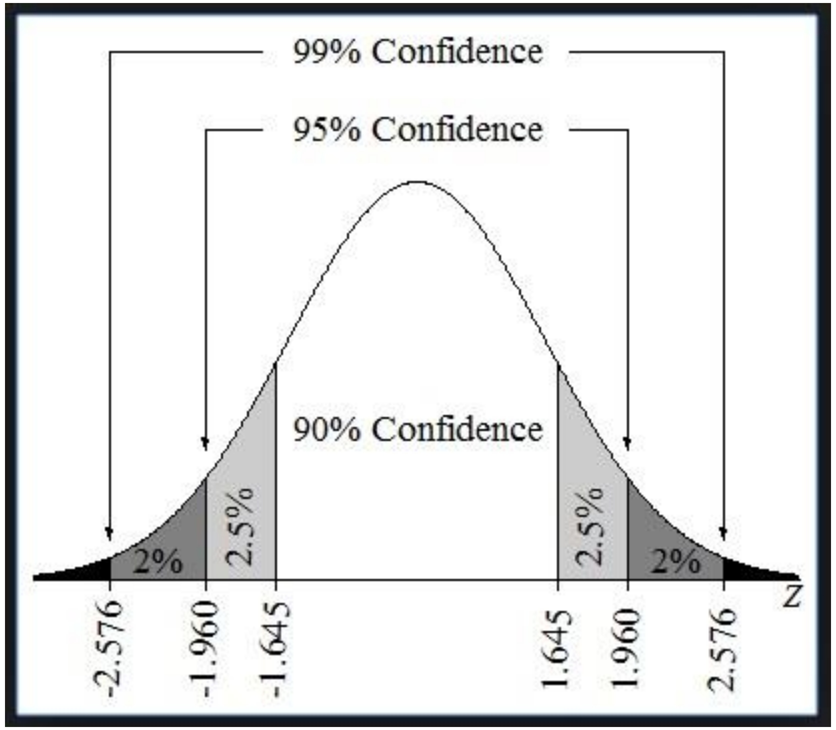
```
Call:
lm_robust(formula = growth ~ tradeshare, data = subset(Growth,
    country_name != "Malta"), se_type = "stata")

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
(Intercept)   0.9574     0.5361   1.786  0.07899 -0.11415    2.029 62
tradeshare    1.6809     0.8656   1.942  0.05670 -0.04944    3.411 62

Multiple R-squared:  0.04466 ,  Adjusted R-squared:  0.02925
F-statistic: 3.771 on 1 and 62 DF,  p-value: 0.0567
```

(c) Construct a 99% confidence interval for β_1.




```
> confint(reg4, level = 0.99)
                 0.5 %   99.5 %
(Intercept) -0.4671516 2.381973
tradeshare  -0.6194541 3.981263
```

The 99% confidence interval is $1.68 \pm 2.58 \times 0.87$.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

E5.3 The data file `Birthweight_Smoking.csv` contains data for a random sample of babies in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. See `Birthweight_Smoking_Description.pdf` for a detailed description of the data. You will investigate the relationship between birth weight and smoking during pregnancy.

	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 1^{st} prenatal care visit in 1^{st} trimester
9	tripre2	indicator=1 if 1^{st} prenatal care visit in 2^{nd} trimester
10	tripre3	indicator=1 if 1^{st} prenatal care visit in 2^{nd} trimester
11	tripre0	indicator=1 if no prenatal visits
12	nprevist	total number of prenatal visits

The University of Queensland Australia

E5.3 The data file `Birthweight_Smoking.csv` contains data for a random sample of babies in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. See `Birthweight_Smoking_Description.pdf` for a detailed description of the data. You will investigate the relationship between birth weight and smoking during pregnancy.

	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

```r
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial03")
BW <- read_csv("birthweight_smoking.csv")
# attach the data to the R search path
attach(BW)
```

(a) In the sample:

i. What is the average value of **birthweight**?
ii. What is the average value of **birthweight** for mothers who smoke?
iii. What is the average value of **birthweight** for mothers who do not smoke?

```
describe(birthweight)
```

```
   vars    n    mean     sd median trimmed    mad min  max range  skew kurtosis    se
X1    1 3000 3382.93 592.16   3420 3412.04 520.39 425 5755  5330 -0.83     2.54 10.81
```

```
tapply(birthweight, smoker, describe)
```

```
$`0`
   vars    n    mean     sd median trimmed    mad min  max range  skew kurtosis    se
X1    1 2418 3432.06 584.62   3459 3460.91 504.08 425 5755  5330 -0.82     2.54 11.89

$`1`
   vars   n    mean     sd median trimmed    mad min  max range  skew kurtosis    se
X1    1 582 3178.83 580.01   3220 3213.11 480.36 510 4763  4253 -1.02      3.1 24.04
```

The sample average of **birthweight** is 3382.93. For smoking mothers, it is 3178.83, while for nonsmoking mothers, it is 3432.06.

(b) i. Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.

ii. What is the standard error for the estimated difference in (b)i?

iii. Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

First, we can conduct the test using R-outputs in (a). The estimated difference is $\overline{X}_{\text{Smokers}} - \overline{X}_{\text{nonSmokers}} = 3178.83 - 3432.06 = -253.23$. The standard error of the difference is $SE(\overline{X}_{\text{Smokers}} - \overline{X}_{\text{nonSmokers}}) = \sqrt{SE(\overline{X}_{\text{Smokers}})^2 + SE(\overline{X}_{\text{nonSmokers}})^2} = \sqrt{11.89^2 + 24.04^2} = 26.82$. The 95% confidence for the difference is $-253.23 \pm 1.96 \times 26.82 = (-305.80, -200.66)$.

Second, we can use `t.test`:

```
t.test(birthweight[smoker == 1], birthweight[smoker == 0])
```

```
	Welch Two Sample t-test

data:  birthweight[smoker == 1] and birthweight[smoker == 0]
t = -9.4414, df = 887.15, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -305.8685 -200.5882
sample estimates:
mean of x mean of y 
 3178.832  3432.060 
```

(c) Run a regression of **birthweight** on the binary variable **smoker**.

i. How the estimated slope and intercept are related to your answers in Parts (a) and (b)?

ii. How the SE($\hat{\beta}_1$) is related to your answer in (b)ii.

iii. Construct a 95% confidence interval for the effect of smoking on birth weight.

```
reg5 = lm_robust(birthweight ~ smoker, data = BW, se_type = "stata")
summary(reg5)
```

```
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
```

```
        Welch Two Sample t-test

data:  birthweight[smoker == 1] and birthweight[smok
t = -9.4414, df = 887.15, p-value < 2.2e-16
alternative hypothesis: true difference in means is
95 percent confidence interval:
 -305.8685 -200.5882
sample estimates:
mean of x mean of y 
 3178.832  3432.060 
```

The estimated regression is

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

i The intercept is the average birthweight for non-smokers (**smoker=0**). The slope is the difference between average birthweights for smokers (**smoker=1**) and non-smokers (**smoker=0**).

ii They are the same.

iii It is the same as the 95% confidence for the difference, i.e., $-253.2 \pm 1.96 \times 26.8 = (-305.9, -200.6)$.

(c) Run a regression of **birthweight** on the binary variable **smoker**.

i. How the estimated slope and intercept are related to your answers in Parts (a) and (b)?

ii. How the SE($\hat{\beta}_1$) is related to your answer in (b)ii.

iii. Construct a 95% confidence interval for the effect of smoking on birth weight.

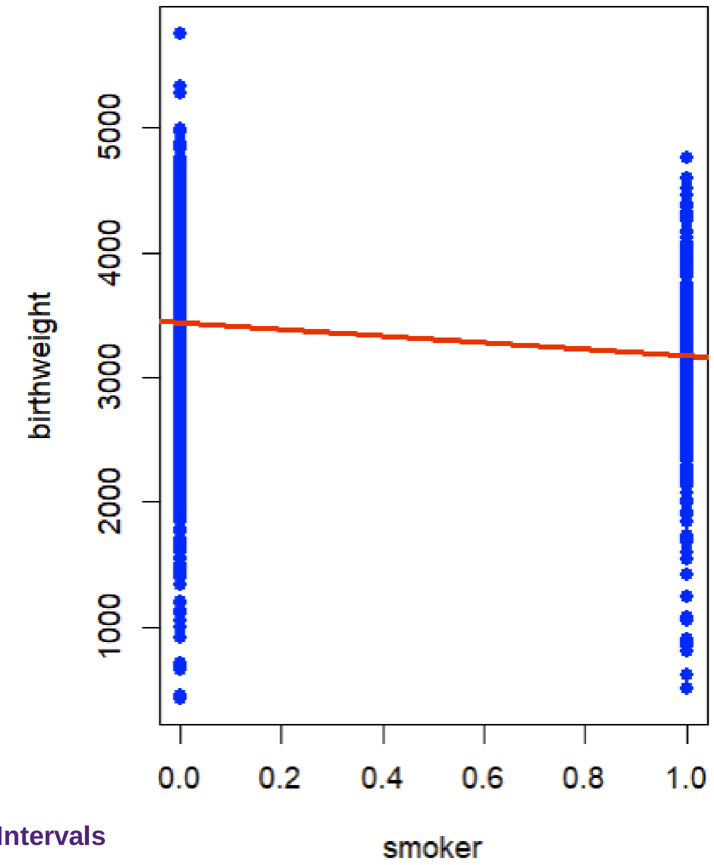
```
reg5 = lm_robust(birthweight ~ smoker, data = BW, se_type = "stata")
summary(reg5)
```

```
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
```



Thank you

Francisco Tavares Garcia | Tutor
School of Economics

Reference

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