THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

ECON2300 - Introductory Econometrics

Tutorial 2: Linear Regression with a Single Regressor

Tutor: Francisco Tavares Garcia

Do you have the latest versions of R and RStudio?

Install R – 4.5.1

https://cran.r-project.org/

Install RStudio – 2025.05.1+513

https://posit.co/download/rstudio-desktop/

Update all packages –

In RStudio >>

Tools >>

Check for Package Updates >>

Select All >>

Install Updates

Quiz 1 is available!

Course Announcements

Quiz 1 Now Available

Hi everyone,

This is a reminder that Quiz 1 is now available under the Assessment folder.

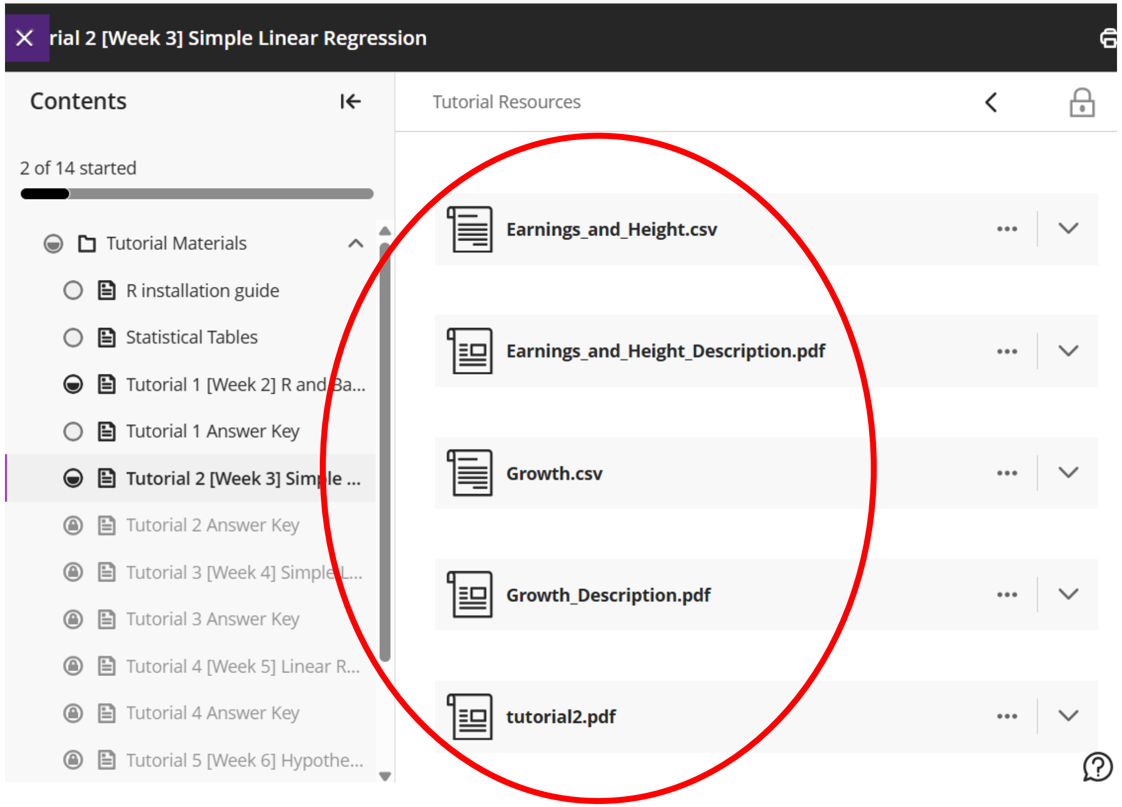
The due date is Thursday, 14th August, 16:00.

Please save and submit your quiz before the due date.

ECON2300 Team

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

- Download the files for tutorial 02 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.

E4.1 The file **Growth.csv** contains data on average growth rates from 1960 through 1995 for 65 countries, along with variables that are potentially related to growth. A detailed description is given in **Growth_Description.pdf**. You will investigate the relationship between growth and trade.

	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

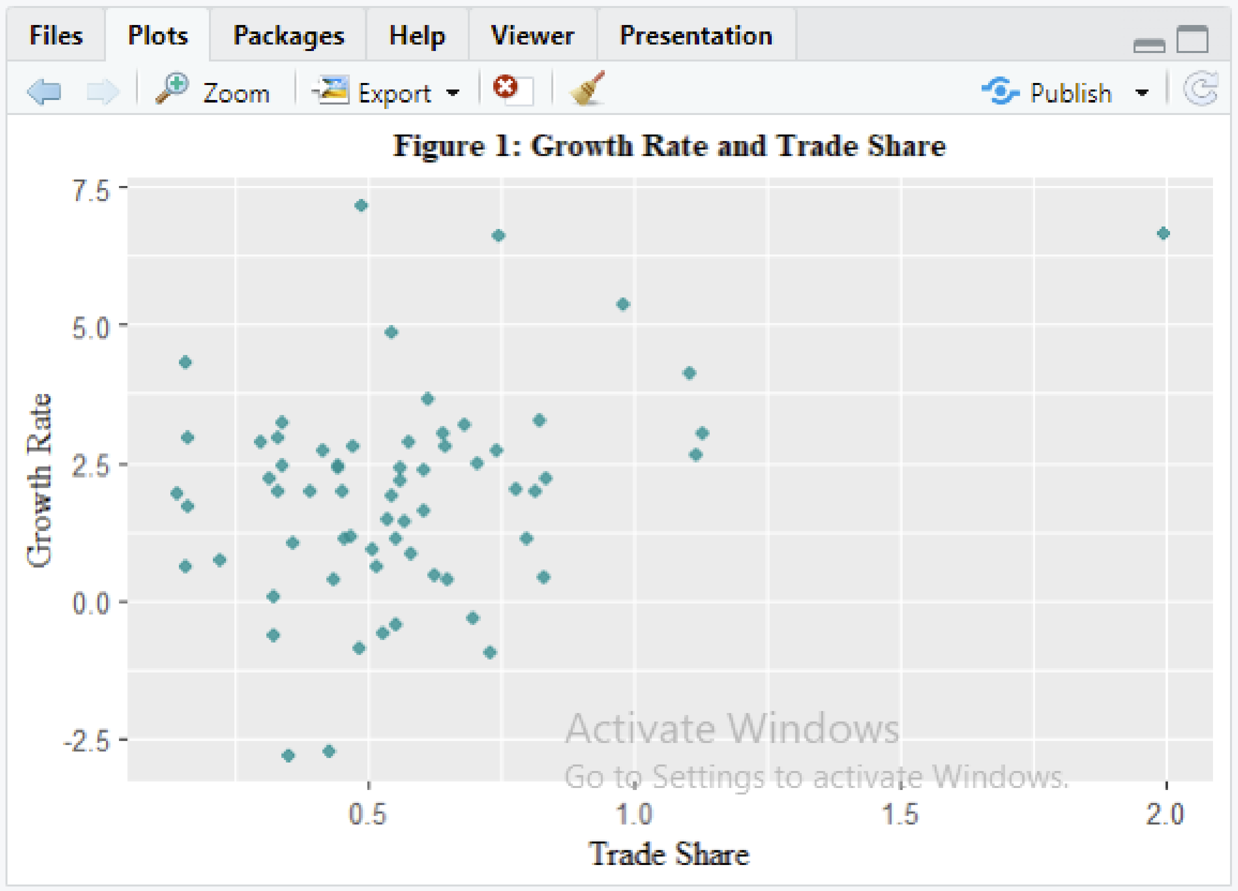
E4.1 The file **Growth.csv** contains data on average growth rates from 1960 through 1995 for 65 countries, along with variables that are potentially related to growth. A detailed description is given in **Growth_Description.pdf**. You will investigate the relationship between growth and trade.

```r
# load non-built-in R packages (you should install them first!)

install.packages("readr")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("estimatr")
install.packages("Hmisc")

library(readr)    # package for fast read rectangular data
library(dplyr)    # package for data manipulation
library(ggplot2)  # package for elegant data visualisation
library(estimatr) # package for commonly used estimators with robust SE
library(Hmisc)    # package for statistics functions


### SW E4.1

# Set working directory (make sure you edit to your own WD)
# Ex Win: setwd("G:/My Drive/BEcon/TUTOR/ECON2300/02")
# Ex Mac: setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial02")


# To use the following line:
# save this file in the same directory as the data files
# install.packages("rstudioapi")
setwd(dirname(rstudioapi::getSourceEditorContext()$path))

# Clean Working Environment
rm(list = ls())

# load csv data
Growth <- read_csv("Growth.csv")
```

(a) Construct a scatterplot of average annual growth rate, **growth**, on the average trade share, **tradeshare**. Does there appear to be a relationship between the variables?

See Figure 1. Yes, there appears to be a weak positive relationship.

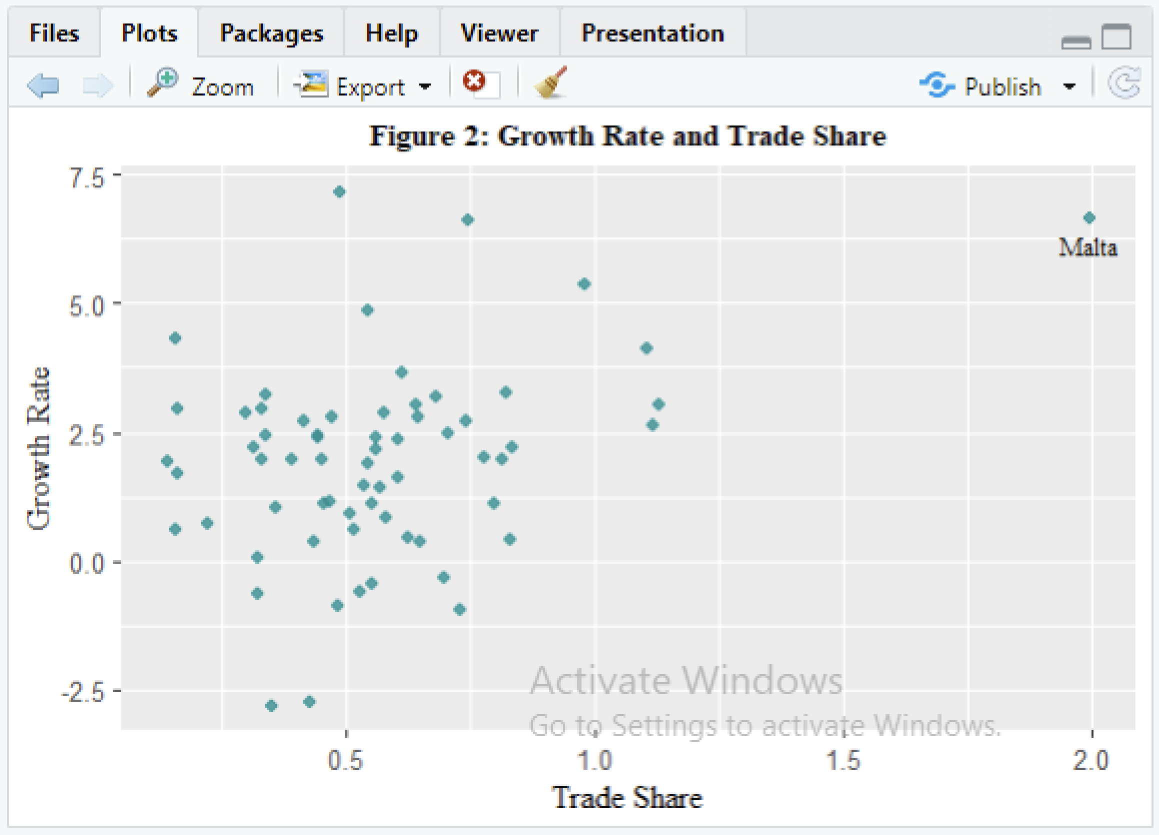
```
# set up figure with data and aes
fig1 <- ggplot(Growth, aes(tradeshare, growth)) +
  # add "point" geom and modify it
  geom_point(alpha = .75, size = 1.5, color = "cyan4") +
  # add title and axis labels
  labs(title = "Figure 1: Growth Rate and Trade Share",
       x = "Trade Share", y = "Growth Rate") +
  # modify theme characteristics
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, family = "serif",
                                   face = "bold", size = 10))
print(fig1)
```



(b) One country, Malta, has a trade share much larger than the other countries. Find Malta on the scatterplot. Does Malta look like an outlier?

See Figure 2. Malta is the "outlying" observation with a trade share of 2.

```
fig2 <- ggplot(Growth, aes(tradeshare, growth)) +
  geom_point(alpha = .75, size = 1.5, color = "cyan4") +
  labs(title = "Figure 2: Growth Rate and Trade Share",
       x = "Trade Share", y = "Growth Rate") +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, size = 10, family = "serif",
                                  face = "bold")) +
  # annotate Malta
  annotate("text", family = "serif", size = 3,
           x = Growth$tradeshare[Growth$country_name == "Malta"],
           y = Growth$growth[Growth$country_name == "Malta"] -.5,
           label = "Malta")
print(fig2)
```



(c) Using all observations, run a regression of **growth** on **tradeshare**. What is the estimated slope? What is the estimated intercept? Use the regression to predict the growth rate for a country with a trade share of 0.5 and with a trade equal to 1.0

```
# regression with robust standard errors
reg1 = lm_robust(growth ~ tradeshare, data = Growth, se_type = "stata")
# present output table
summary(reg1)
```

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

Standard error type:  HC1

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
(Intercept)   0.6403     0.4591   1.394 0.1680736  -0.2773    1.558 63
tradeshare    2.3064     0.6633   3.477 0.0009235   0.9810    3.632 63

Multiple R-squared:  0.1237 ,   Adjusted R-squared:  0.1098
F-statistic: 12.09 on 1 and 63 DF,  p-value: 0.0009235
```

```
# prediction based on regression model
predict(reg1, newdata = data.frame(tradeshare = c(0.5, 1)))
```

```
> predict(reg1, newdata = data.frame(tradeshare = c(0.5, 1)))
       1        2
1.793482 2.946699
```

The fitted regression line is

$$\widehat{\texttt{growth}} = 0.64 + 2.31 \times \texttt{tradeshare}$$

where the estimated slope and intercept are 2.31 and 0.64, respectively. Moreover, the predicted growth, $\widehat{\texttt{growth}}$, at $\texttt{tradeshare} = 1$ is $0.64+2.31\times 1 = 2.95$. Similarly, the predicted growth, $\widehat{\texttt{growth}}$, at $\texttt{tradeshare} = 0.5$ is $0.64 + 2.31 \times 0.5 = 1.80$.

(d) Estimate the same regression, excluding the data from Malta. Answer the same questions in (c).

```r
reg2 = lm_robust(growth ~ tradeshare, data = subset(Growth, country_name != "Malta"),
                 se_type = "stata")
summary(reg2)
```

```
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
```

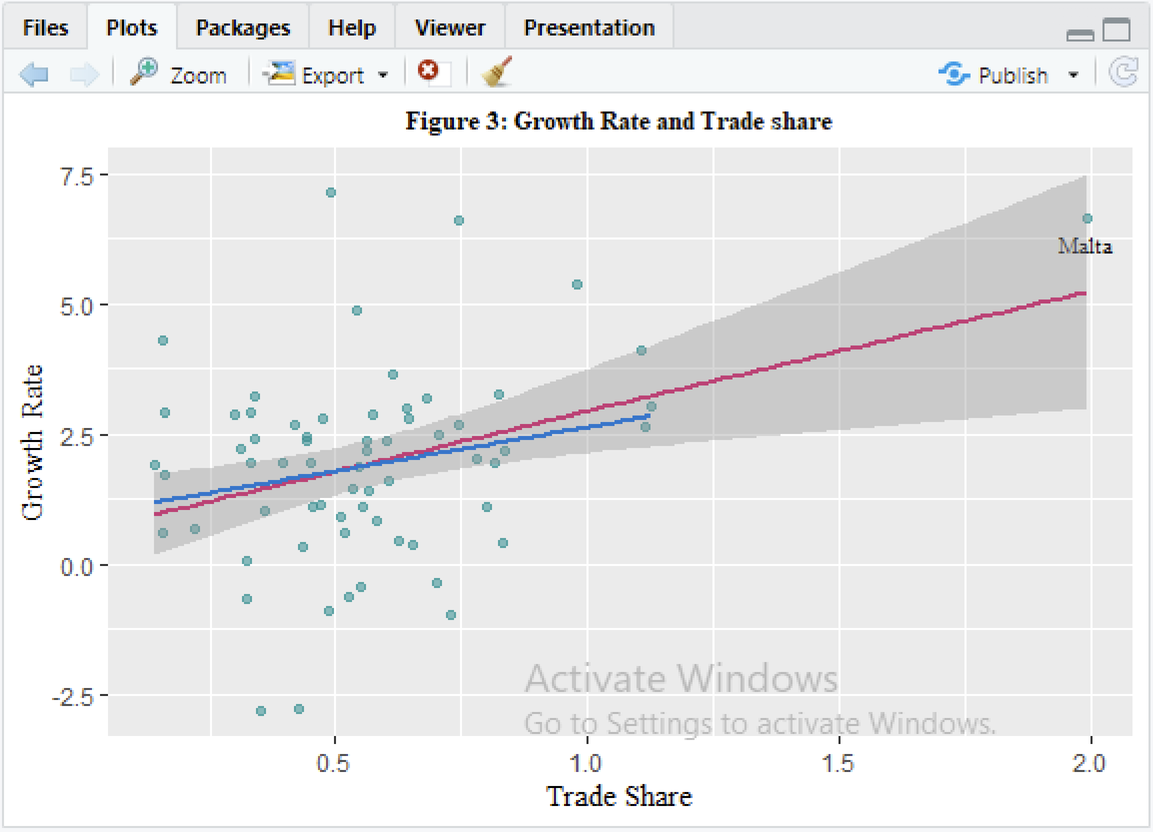
```r
predict(reg2, newdata = data.frame(tradeshare = c(0.5, 1)))
```

```
> predict(reg2, newdata = data.frame(tradeshare = c(0.5, 1)))
       1        2
1.797863 2.638315
```

The estimated slope and intercept are 1.68 and 0.96, respectively. Moreover, the predicted growth, $\widehat{\texttt{growth}}$, at $\texttt{tradeshare} = 1$ is $0.96 + 1.68 \times 1 = 2.64$. Similarly, the predicted growth, $\widehat{\texttt{growth}}$, at $\texttt{tradeshare} = 0.5$ is $0.96 + 1.68 \times 0.50 = 1.80$.

(e) Plot the estimated regression functions from (c) and (d). Using the scatterplot in (a), explain why the regression function that includes Malta is steeper than the regression function that excludes Malta.

```
fig3 <- ggplot(Growth, aes(tradeshare, growth)) +
  geom_point(alpha = .5, size = 1.5, color = "cyan4") +
  labs(title = "Figure 3: Growth Rate and Trade Share",
       x = "Trade Share", y = "Growth Rate") +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, size = 10, family = "serif",
                                  face = "bold")) +
  # fitted line using all data
  geom_smooth(method = "lm", se = FALSE, size = 0.75, color = "violetred3") +
  # fitted line using all but Malta data
  geom_smooth(data = subset(Growth, country_name != "Malta"),
              method = "lm", se = FALSE, size = 0.75, color = "dodgerblue3")
print(fig3)
```



The data point of Malta is far away from the cloud of other points with very high trade share and growth rate. As the sample size is small, such an outlier (influential point) can greatly affect the slope of the regression line. In this case, as the growth rate of Malta is higher than most other countries, including Malta makes the regression line steeper.

(f) Where is Malta? Why is the Malta trade share so large? Should Malta be included or excluded from the analysis?

Malta is an island nation in the Mediterranean Sea, south of Sicily. Malta is a freight transport site, which explains its large "trade share." Many goods coming into Malta (imports into Malta) are immediately transported to other countries (as exports from Malta). Thus, Malta's imports and exports are unlike the imports and exports of most other countries. Malta should not be included in the analysis.

E4.2 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. You will investigate the relationship between earnings and height.

	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)

(a) What is the median value of height in the sample?

```
rm(list = ls())
setwd("/Users/uqdkim7/Dropbox/Teaching/R tutorials/Tutorial02")
EH <- read_csv("Earnings_and_Height.csv")
```

(a)

The median height in the sample is 67 inches.

```
# descriptive/summary statistics
describe(EH$height, descript = "height")
```

```
> describe(EH$height, descript = "height")
height
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   17870        0       34    0.995    66.96    4.518       61       62       64       67       70       72       74

lowest : 48 49 51 52 53, highest: 79 80 81 83 84
```

(b) i. Estimate average earnings for workers whose height is at most 67 inches.

ii. Estimate average earnings for workers whose height is greater than 67 inches.

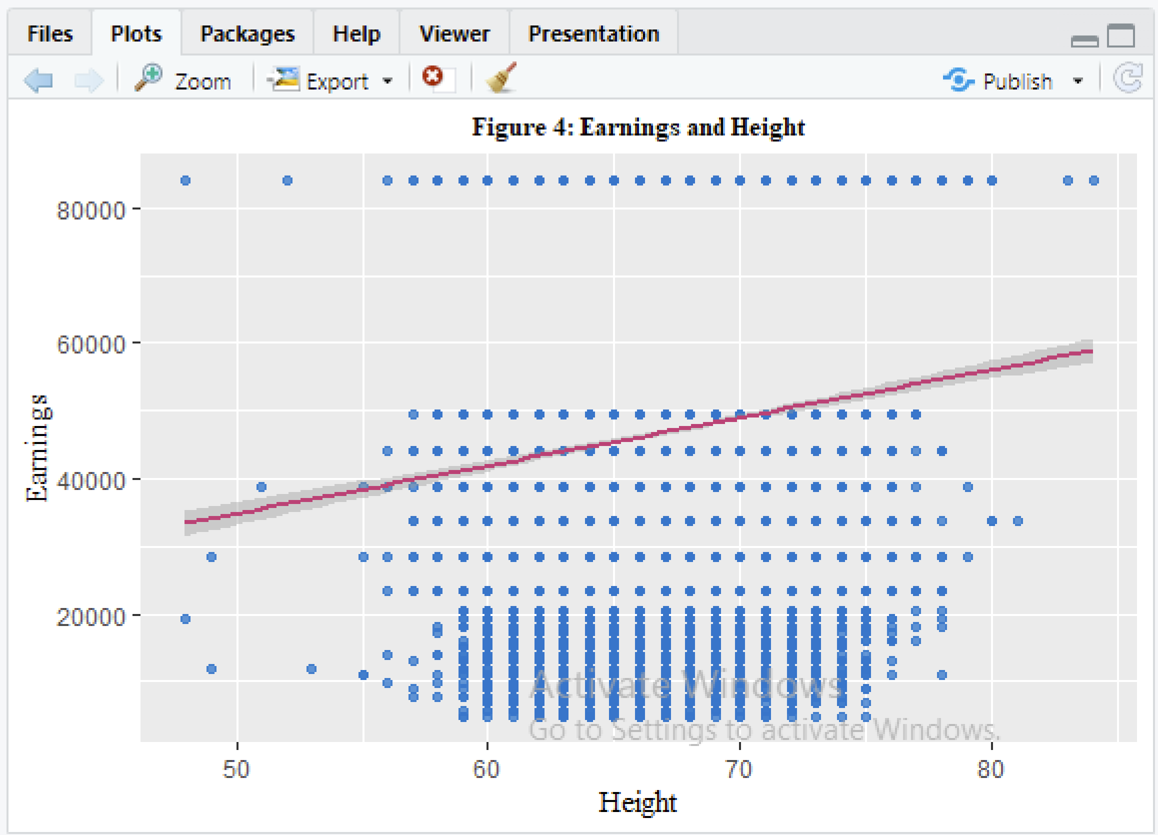
iii. On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

```
> describe(EH$earnings[EH$height <= 67], descript = "earnings for height <= 67")
earnings for height <= 67
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   10114        0       23    0.974    44488    29337    10865    15082    23363    38925    84055    84055    84055

lowest :  4726.391  5675.895  6711.288  7782.854  8826.651, highest: 33712.969 38925.336 44152.160 49430.109 84054.750
> describe(EH$earnings[EH$height > 67], descript = "earnings for height > 67")
earnings for height > 67
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
    7756        0       23    0.952    49988    29682    14021    18169    28560    44152    84055    84055    84055

lowest :  4726.391  5675.895  6711.288  7782.854  8826.651, highest: 33712.969 38925.336 44152.160 49430.109 84054.750
> # unpair two-Samples t-test
> t.test(EH$earnings[EH$height <= 67], EH$earnings[EH$height > 67])

        Welch Two Sample t-test

data:  EH$earnings[EH$height <= 67] and EH$earnings[EH$height > 67]
t = -13.59, df = 16624, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6292.643 -4706.237
sample estimates:
mean of x mean of y
 44488.44  49987.88
```

The estimated average annual earnings for shorter workers is \$44,488, is \$49,988 for taller workers, for a difference of \$5,499. The 95% confidence interval is \$4,706 to \$6,293. The difference is large (more than 10% of average earnings), precisely estimated and statistically significantly different from zero (p-value is essentially zero).

(c) Construct a scatterplot of annual earnings, **earnings**, on height, **height**. Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of **earnings**). Why? (*Hint: Carefully read the detailed data description.*)

```
fig4 <- ggplot(EH, aes(height, earnings)) +
  geom_point(alpha = .75, size = 1.5, color = "dodgerblue3") +
  labs(title = "Figure 4: Earnings and Height",
       x = "Height", y = "Earnings") +
  theme(axis.title = element_text(family = "serif"),
        plot.title = element_text(hjust = 0.5, family = "serif",
                                  face = "bold", size = 10)) +
  geom_smooth(method = "lm", se = FALSE, size = 0.75, color = "violetred3")
print(fig4)
```



The data documentation reports that individual earnings were reported in 23 brackets, and a single average value is reported for earnings in the same bracket. Thus, the dataset contains 23 distinct values of earnings.

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

i. What is the estimated slope?

ii. Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

```
> reg3 <- lm_robust(earnings ~ height, data = EH, se_type = "stata")
> summary(reg3)

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
> predict(reg3, newdata = data.frame(height = c(65, 67, 70)))
       1        2        3
45485.92 46901.26 49024.28
```

$$\widehat{\texttt{earnings}} = -512.7 + 707.7 \times \texttt{height}, \qquad R^2 = 0.011$$

The estimated slope is 707.7 (\$ per year). The estimated earnings are

Height in inches	$\widehat{\texttt{earnings}}$ in \$ per year
65	45,486
67	46,901
70	49,024

(e) Suppose height were measured in centimeters instead of inches. Answer the following questions about the earnings on height (in cm) regression.

i. What is the estimated slope of the regression?

ii. What is he estimated intercept?

iii. What is the R^2?

iv. What is the standard error of the regression?

Compute SER:

```
sqrt(reg3$res_var)
```

```
## [1] 26777.24
```

Recall that 1 cm = 0.394 inches. The estimated regression in (d), with units shown, is

$$\widehat{\texttt{earnings}}(\$) = -512.7(\$) + 707.7(\$/inch) \times \texttt{height}(inch),$$

and we have $R^2 = 0.011$ which is unit free and $SER = 26,777(\$)$, the same unit as the LHS variable. Note that

$$\begin{aligned}\widehat{\texttt{earnings}}(\$) &= -512.7(\$) + 707.7(\$/inch) \times \texttt{height}(inch)\\ &= -512.7(\$) + 707.7(\$/inch) \times (0.394 inch/cm) \times \texttt{height}(cm)\\ &= -512.7(\$) + 278.8(\$/cm) \times \texttt{height}(cm)\end{aligned}$$

So the regression is

$$\widehat{\texttt{earnings}}(\$) = -512.7(\$) + 278.8(\$/cm) \times \texttt{height}(cm)$$

with $R^2 = 0.011$ and $SER = 26,777(\$)$.

(f) Run a regression of **earnings** on **height** using data for female workers only.

i. What is the estimated slope?

ii. A randomly selected woman is 1 inch taller than he average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

```
reg4 = lm_robust(earnings ~ height, data = subset(EH, sex == "0:female"), se_type = "stata")
summary(reg4)
```

```
Call:
lm_robust(formula = earnings ~ height, data = subset(EH, sex ==
    "0:female"), 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
```

$$\widehat{\texttt{earnings}} = 12650 + 511.2 \times \texttt{height}, \qquad R^2 = 0.003$$

A women who is one inch taller than average is predicted to have earnings that are \$511.2 per year higher than average.

(f) Run a regression of **earnings** on **height** using data for female workers only.

i. What is the estimated slope?

ii. A randomly selected woman is 1 inch taller than he average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

(g) Repeat (f) for male workers.

```r
reg5 = lm_robust(earnings ~ height, data = subset(EH, sex == "1:male"), se_type = "stata")
summary(reg5)
```

```
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
```

$$\widehat{\texttt{earnings}} = -43130 + 1307 \times \texttt{height}, \qquad R^2 = 0.021$$

A man who is one inch taller than average is predicted to have earnings that are \$1307 per year higher than average.

(h) Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, say u_i, has a conditional mean of zero, given `height` (X_i)?

Height may be correlated with other factors that cause earnings. For example, height may be correlated with "strength," and in some occupations, stronger workers may by more productive. There are many other potential factors that may be correlated with height and cause earnings and you will investigate of these in future exercises.

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.