ECOM20001 Econometrics 1

Tutorial 5 Semester 2, 2021

Chin Quek

Department of Economics

Introduction

- Hypothesis Testing of Sample Means
- Single Linear Regression

File downloaded from Canvas

- `tute5.R`
- `tute5_height.csv`
- `tute5_growth.csv`

Dataset 1

The first (micro) dataset, `tute5_height.csv`, has the following 5 variables:

- `id`: worker identifier
- `earnings`: annual labour earnings in $10,000's (in real terms, 2012=100)
- `height`: height without shoes in centimetres
- `weight`: weight without shoes in kilograms
- `male`: binary variable that equals 1 if worker is male and 0 otherwise
- `age`: age of the worker at time of survey

In total, the dataset contains this information for n = 17,870 U.S. workers.

```
# Loading the dataset
data1=read.csv(file="tute5_height.csv")

head(data1)
```

```
##   ï..id earnings height weight male age
## 1     1 8.405475    165     60    0  48
## 2     2 1.402139    165     70    0  41
## 3     3 8.405475    152     49    0  26
## 4     4 8.405475    170     68    0  37
## 5     5 2.856039    173     82    0  35
## 6     6 2.336287    160     46    0  25
```

```r
library(stargazer)
stargazer(data1, type = "text", digits = 2,
          summary.stat = c("n", "mean", "sd", "median", "min", "max"),
          title = "Summary statistics"
          )
```

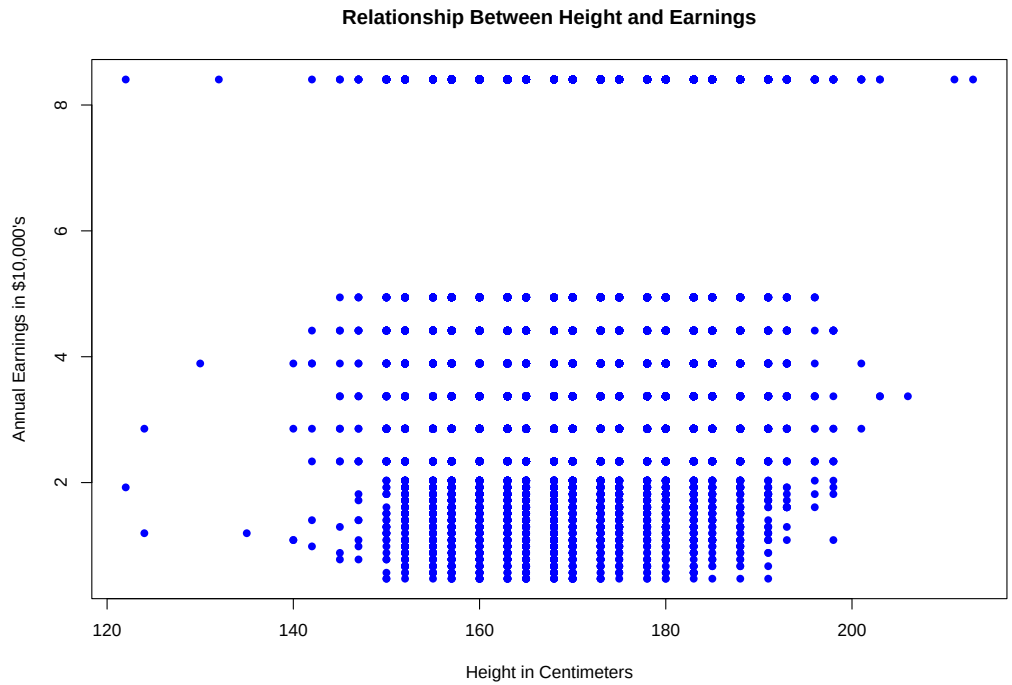
```
##
## Summary statistics
## ===================================================
## Statistic    N       Mean    St. Dev. Median  Min    Max
## ---------------------------------------------------
## ï..id      17,870 8,935.50 5,158.77 8,935.5   1    17,870
## earnings   17,870    4.69     2.69     3.89    0.47   8.41
## height     17,870   170.10   10.10      170    122    213
## weight     17,870   77.36    23.32      74      36    227
## male       17,870    0.44     0.50       0      0      1
## age        17,870   40.92    10.04      40      25     65
## ---------------------------------------------------
```

Height and Earnings

```
cor(data1$height,data1$earnings)
```

```
## [1] 0.1044771
```

```
plot(data1$height,data1$earnings,
     main="Relationship Between Height and Earnings",
     xlab="Height in Centimeters", ylab="Annual Earnings in $10,000's",
     col="blue", pch=16)
```



Comparison of two sample means

Compare sample means for people with height $\geq$ 170cm and $<$ 170cm.

```
## Mean earnings for heights above and below 170cm
mean(data1$earnings[data1$height >= 170])
```

```
## [1] 4.909318
```

```
mean(data1$earnings[data1$height < 170])
```

```
## [1] 4.44879
```

```
## Difference in means for people taller and shorter than 170cm
mean(data1$earnings[data1$height >= 170])-
  mean(data1$earnings[data1$height < 170])
```

```
## [1] 0.460528
```

Two-sample t-test

Conduct a two sample t-test to determine whether there is a significant difference in earnings between the two height groups.

```
t.test(data1$earnings[data1$height >= 170],
       data1$earnings[data1$height < 170])
```

```
##
##  Welch Two Sample t-test
##
## data:  data1$earnings[data1$height >= 170] and data1$earnings[data1$heig
## t = 11.469, df = 17783, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3818203 0.5392357
## sample estimates:
## mean of x mean of y
##  4.909318  4.448790
```

In-tutorial Question 1

Look at the results of the t-test you conducted to determine whether there is a difference between people's earnings who are under 170 cm tall and those who are 170 cm or taller.

- Write out the hypotheses.
- Using the results of your test what is your decision and conclusion?

In-tutorial Question 1

Look at the results of the t-test you conducted to determine whether there is a difference between people's earnings who are under 170 cm tall and those who are 170 cm or taller.

- Write out the hypotheses.
- Using the results of your test what is your decision and conclusion?

Let $\mu_1 =$ average earnings of people 170 cm or taller, and
$\mu_2 =$ average earnings of people under 170 cm

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 \neq 0$$

In-tutorial Question 1

Look at the results of the t-test you conducted to determine whether there is a difference between people's earnings who are under 170 cm tall and those who are 170 cm or taller.

- Write out the hypotheses.
- Using the results of your test what is your decision and conclusion?

Let μ_1 = average earnings of people 170 cm or taller, and
μ_2 = average earnings of people under 170 cm

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 \neq 0$$

```
##
##  Welch Two Sample t-test
##
## data:  data1$earnings[data1$height >= 170] and data1$earnings[data1$heig
## t = 11.469, df = 17783, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3818203 0.5392357
## sample estimates:
## mean of x mean of y
##  4.909318  4.448790
```

Is there significant difference between earnings of those under 170cm and those 170cm or taller?

- Rejects the null of equal means with a t-statistic of 11.470 and p-value less than 2.2e-16
- 95% CI is [$3818, $5392].

Is there significant difference between earnings of those under 170cm and those 170cm or taller?

- Rejects the null of equal means with a t-statistic of 11.470 and p-value less than 2.2e-16
- 95% CI is [$3818, $5392].

Note that this interval does not contain zero which corresponds to the result in the hypothesis test above.

The results provide initial evidence that taller people in the top half of the sample above the median height of 170cm have higher average income than people in the bottom half of the sample with height below 170cm.

In-tutorial Question 2

Compare the value of the t-statistic and p-value from this hypothesis test to the t-statistic and p-value you obtained for regressing earnings on the dummy variable you created.

What do you notice?

In-tutorial Question 2

Compare the value of the t-statistic and p-value from this hypothesis test to the t-statistic and p-value you obtained for regressing earnings on the dummy variable you created.

What do you notice?

Setting up the regression model

Define a dummy variable, `height.dv` taking the values - 1 = height greater than or equal to 170cm - 0 = height less than 170 cm

```
data1$height.dv <- 1*(data1$height >= 170)

head(data1)
```

```
##   ï..id earnings height weight male age height.dv
## 1     1 8.405475    165     60    0  48         0
## 2     2 1.402139    165     70    0  41         0
## 3     3 8.405475    152     49    0  26         0
## 4     4 8.405475    170     68    0  37         1
## 5     5 2.856039    173     82    0  35         1
## 6     6 2.336287    160     46    0  25         0
```

Regression of earnings on the dummy variable height.dv

```
reg.dv <- lm(earnings ~ data1$height.dv, data = data1)
summary(reg.dv)
```

```
##
## Call:
## lm(formula = earnings ~ data1$height.dv, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.437 -2.112 -1.017  3.496  3.957
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.44879    0.02892  153.85   <2e-16 ***
## data1$height.dv   0.46053    0.04016   11.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.683 on 17868 degrees of freedom
## Multiple R-squared:  0.007305,   Adjusted R-squared:  0.00725
## F-statistic: 131.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

Two-samples t-test

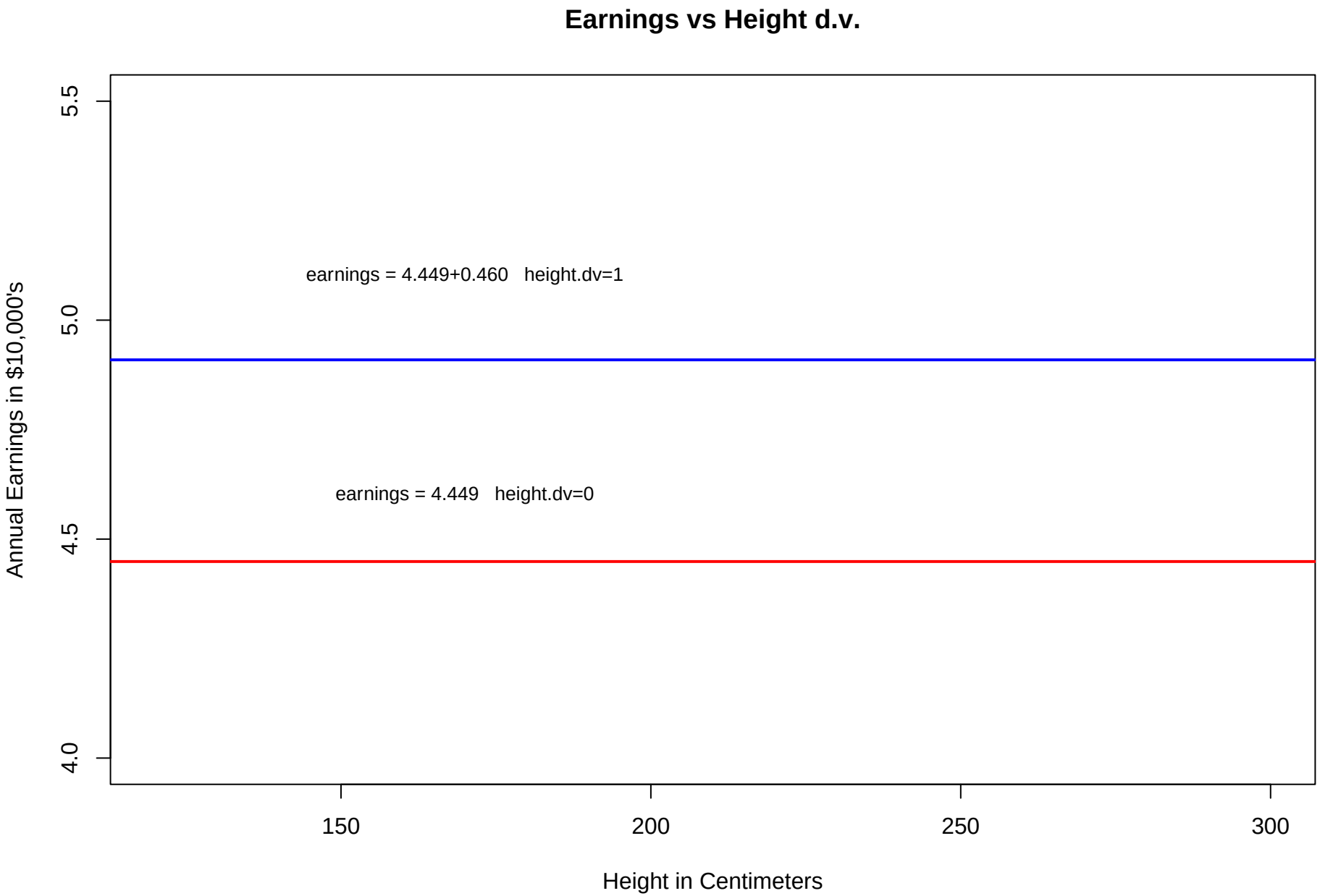
```
##
##  Welch Two Sample t-test
##
## data:  data1$earnings[data1$height >= 170] and data1$earnings[data1$heig
## t = 11.469, df = 17783, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3818203 0.5392357
## sample estimates:
## mean of x mean of y
##  4.909318  4.448790
```

Two-samples t-test

```
##
##  Welch Two Sample t-test
##
## data:  data1$earnings[data1$height >= 170] and data1$earnings[data1$heig
## t = 11.469, df = 17783, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3818203 0.5392357
## sample estimates:
## mean of x mean of y
##  4.909318  4.448790
```

- What do you notice between the two t-statistics and p-values?

Plot of sample regression lines



$$\text{earnings}_i = \beta_0 + \beta_1 \, \text{height.dv} + u_i \quad (1)$$

Dataset 2

The second (macro) dataset, `tute5_growth.csv`, has the following 5 variables:

- `country`: country name
- `growth`: average annual percentage growth rate of real GDP (1960=100) from 1960-1995
- `rgdp60`: the value of GDP per capita in 1960 (in real terms, 1960=100)
- `tradeshare`: the average share of annual trade in the economy from 1960 to 1995, measured as the sum of gross exports plus gross imports divided by nominal GDP; that is the average of (X+M)/GDP from 1960 to 1995.

In total, the dataset contains this information for n = 65 countries.

```
# Loading the dataset
data2=read.csv(file="tute5_growth.csv")
```

```
head(data2)
```

```
##   ï..country    growth    rgdp60 tradeshare
## 1   Argentina 0.6176451 4462.0015  0.1566230
## 2   Australia 1.9751474 7782.0024  0.3294792
## 3     Austria 2.8891852 5143.0010  0.5752748
## 4  Bangladesh 0.7082631  951.9998  0.2214584
## 5     Belgium 2.6513345 5495.0020  1.1159170
## 6     Bolivia 0.3550578 1147.9998  0.4355793
```

```
names(data2)
```

```
## [1] "ï..country" "growth"     "rgdp60"     "tradeshare"
```

```
dim(data2)
```

```
## [1] 65  4
```

```r
stargazer(data2, type = "text", digits = 2,
          summary.stat = c("n", "mean", "sd", "median", "min", "max"),
          title = "Summary statistics")
```

```
##
## Summary statistics
## ===================================================
## Statistic  N    Mean    St. Dev.  Median   Min     Max
## ---------------------------------------------------
## growth     65   1.94      1.90     1.98   -2.81    7.16
## rgdp60     65 3,103.78 2,512.66 2,019.00 367.00 9,895.00
## tradeshare 65   0.56      0.29     0.54    0.14    1.99
## ---------------------------------------------------
```

```
stargazer(data2, type = "text", digits = 2,
          summary.stat = c("n", "mean", "sd", "median", "min", "max"),
          title = "Summary statistics")
```

```
##
## Summary statistics
## ===================================================
## Statistic  N    Mean   St. Dev.  Median   Min     Max
## ---------------------------------------------------
## growth     65   1.94     1.90     1.98   -2.81    7.16
## rgdp60     65 3,103.78 2,512.66 2,019.00 367.00 9,895.00
## tradeshare 65   0.56     0.29     0.54    0.14    1.99
## ---------------------------------------------------
```

- Is there anything interesting about the statistics?
- Interpretation of the statistics?

```
stargazer(data2, type = "text", digits = 2,
          summary.stat = c("n", "mean", "sd", "median", "min", "max"),
          title = "Summary statistics")
```

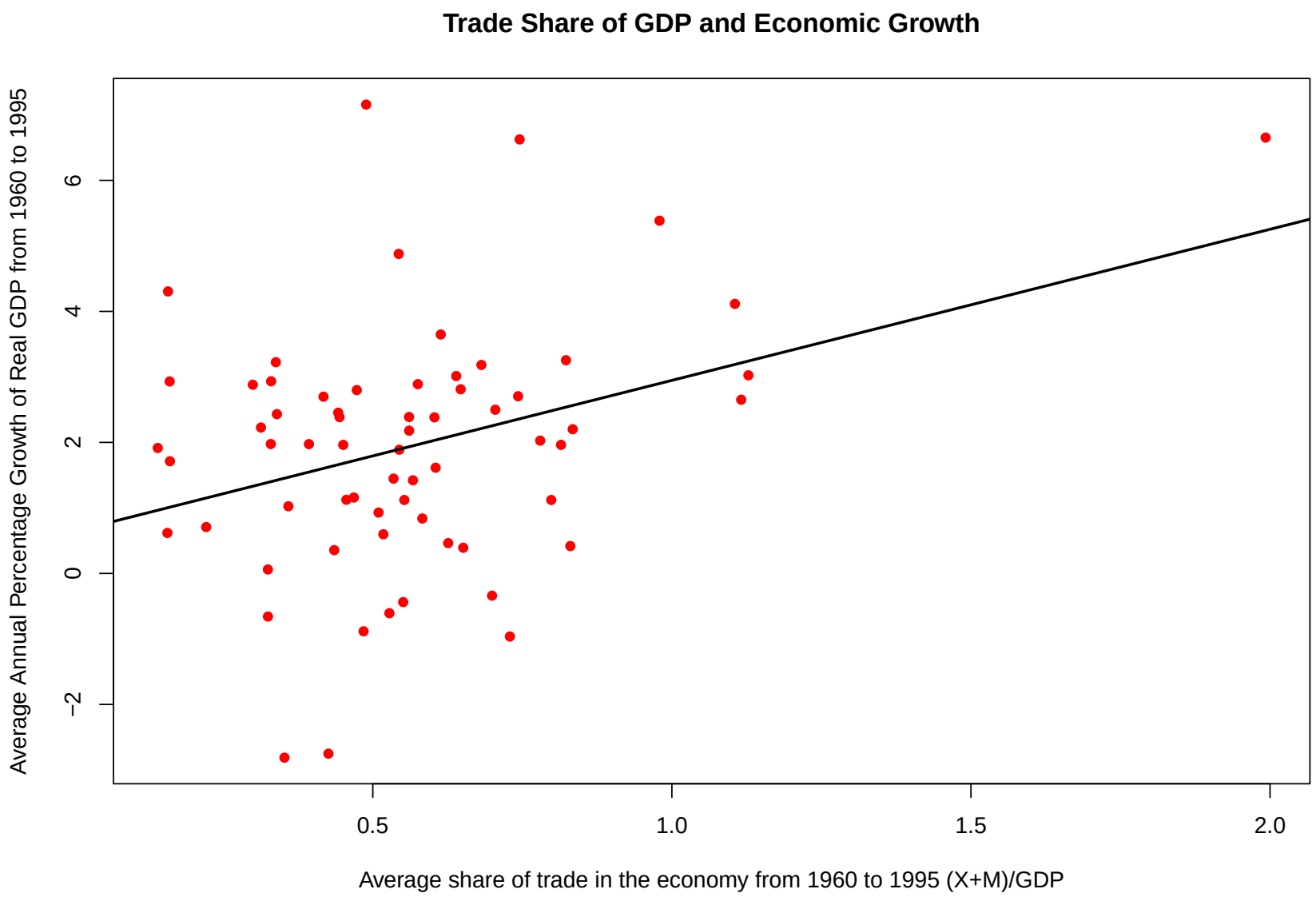
```
##
## Summary statistics
## ===================================================
## Statistic  N    Mean   St. Dev.  Median   Min     Max
## ---------------------------------------------------
## growth     65   1.94     1.90     1.98   -2.81    7.16
## rgdp60     65 3,103.78 2,512.66 2,019.00 367.00 9,895.00
## tradeshare 65   0.56     0.29     0.54    0.14    1.99
## ---------------------------------------------------
```

- Is there anything interesting about the statistics?
- Interpretation of the statistics?

A typical country has an average annual growth rate of 1.94%, real (1960=100) GDP of \$3,104 per person, and a trade share of 56.47%.

For the latter, this means the average country gross exports and imports together more than **half** of its annual GDP.

Scatter plot of tradeshare and growth



Discussion

Looking at the scatter plot between **tradeshare** and **growth** below, there does appear visually to be a positive relationship between growth and trade.

The Pearson correlation coefficient that measures the strength of the linear relationship between **tradeshare** and **growth**

```
cor(data2$tradeshare,data2$growth)
```

```
## [1] 0.351682
```

Discussion

Looking at the scatter plot between **tradeshare** and **growth** below, there does appear visually to be a positive relationship between growth and trade.

The Pearson correlation coefficient that measures the strength of the linear relationship between **tradeshare** and **growth**

```
cor(data2$tradeshare,data2$growth)
```

```
## [1] 0.351682
```

However, it appears that there are a number of unusual observations. These points in the scatter plot could be potential **outliers**.

- Can also identify these potential outliers by noting there are a few observations with a **growth** rate above 6% and below -2%.

```
data2[which(data2$growth > 6),]
```

```
##                  ctry   growth    rgdp60 tradeshare
## 33 Korea, Republic of 7.156855  904.0001  0.4889496
## 35               Malta 6.652838 1374.0000  1.9926157
## 56      Taiwan, China 6.624734 1256.0000  0.7454978
```

```
data2[which(data2$growth < -2),]
```

```
##     ctry    growth   rgdp60 tradeshare
## 40 Niger -2.751478 531.9999  0.4258372
## 64 Zaire -2.811944 488.9999  0.3523176
```

Regression on entire sample

$$\widehat{\text{growth}} = 0.64 + 2.306\,\text{tradeshare}$$

```
##
## Call:
## lm(formula = growth ~ tradeshare, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3739 -0.8864  0.2329  0.9248  5.3889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6403     0.4900   1.307  0.19606
## tradeshare    2.3064     0.7735   2.982  0.00407 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.79 on 63 degrees of freedom
## Multiple R-squared:  0.1237, Adjusted R-squared:  0.1098
## F-statistic: 8.892 on 1 and 63 DF,  p-value: 0.00407
```

Excluding potential outliers

```
exK2data <- subset(data2, ctry !="Korea, Republic of")

reg2.2<-lm(growth~tradeshare,data=exK2data)
summary(reg2.2)
```

```
exM2data  <- subset(data2, ctry !="Malta")

reg2.3 <-lm(growth~tradeshare,data=exM2data)
summary(reg2.3)
```

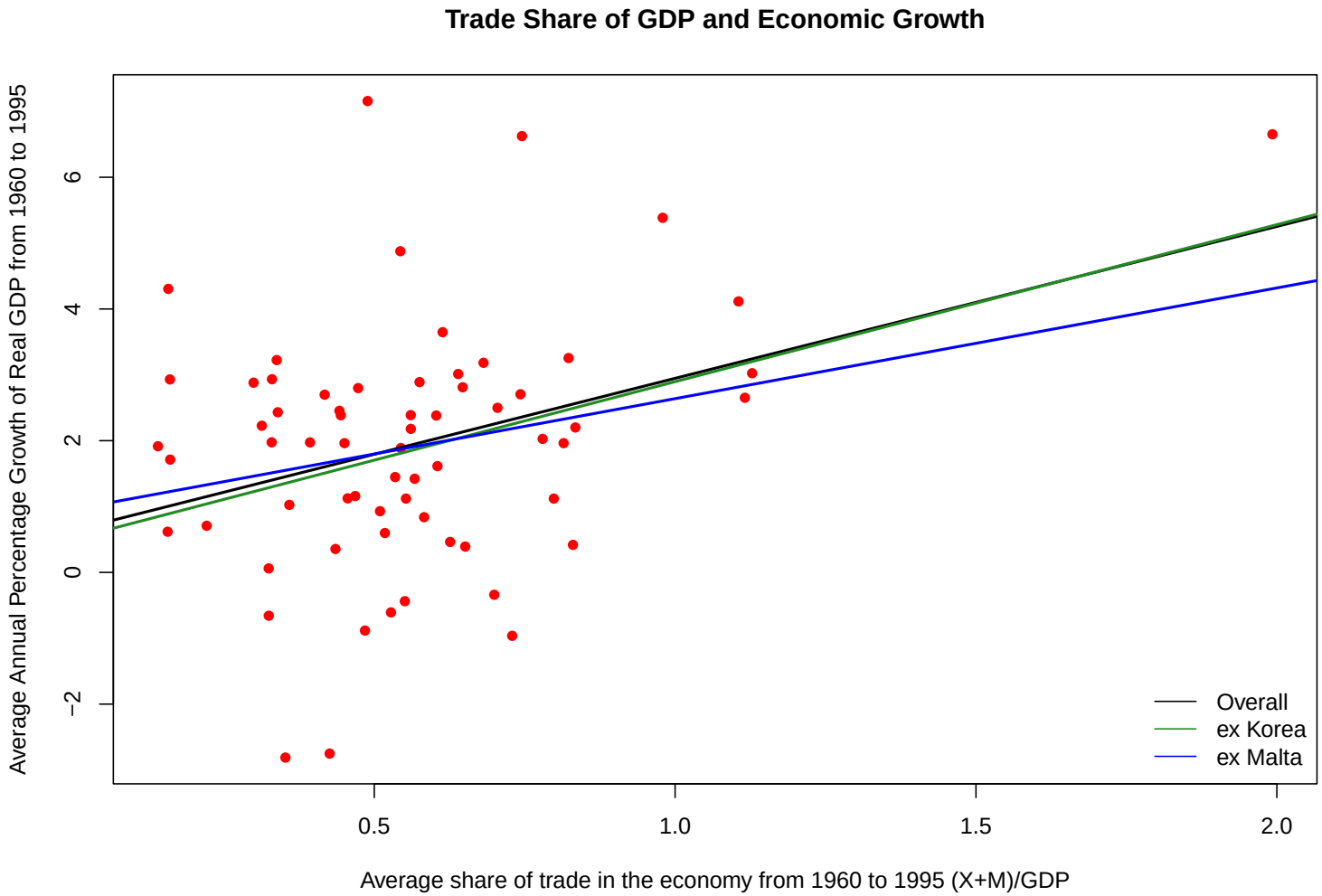
Sample regression excluding Korea

```
##
## Call:
## lm(formula = growth ~ tradeshare, data = exK2data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2789 -0.8642  0.1924  0.9832  4.3353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5122     0.4580   1.118  0.26778
## tradeshare    2.3839     0.7208   3.307  0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.667 on 62 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1363
## F-statistic: 10.94 on 1 and 62 DF,  p-value: 0.001571
```

Sample regression excluding Malta

```
##
## Call:
## lm(formula = growth ~ tradeshare, data = exM2data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4247 -0.9383  0.2091  0.9265  5.3776
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9574     0.5804   1.650   0.1041
## tradeshare    1.6809     0.9874   1.702   0.0937 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.789 on 62 degrees of freedom
## Multiple R-squared:  0.04466,    Adjusted R-squared:  0.02925
## F-statistic: 2.898 on 1 and 62 DF,  p-value: 0.09369
```

Comparison of all three sample regressions



```r
plot(data2$tradeshare,data2$growth,
     main = "Trade Share of GDP and Economic Growth",
     xlab = "Average share of trade in the economy from 1960 to 1995
     (X+M)/GDP",
     ylab = "Average Annual Percentage Growth of Real GDP from 1960 to 1995
     col = "red",
     pch = 16)

abline(reg2.1, col = "black", lwd = 2)
abline(reg2.2, col = "forestgreen", lwd = 2)
abline(reg2.3, col = "blue", lwd = 2)
legend("bottomright", legend = c("Overall","ex Korea", "ex Malta"),
       col = c("black","forestgreen",  "blue"), lty = 1, cex = 1,
       box.lty = 0)
```

- From the graph before, what are some reasons for these "unusual" observations?

```
reg2.3$coefficients
```

```
## (Intercept)  tradeshare
##   0.9574107   1.6809047
```

- Using the regression result excluding Malta, how would you interpret the slope coefficient for a one-unit increase in the trade share of a country?
- What is the effect of excluding the outlier, Republic of Korea?
- Does dropping Malta appear to have a large impact on the results?

```
stargazer(reg2.1, reg2.2, reg2.3,
          type = "text",
          intercept.bottom = FALSE)
```

```
##
## ================================================================================
##                                        Dependent variable:
##                    ------------------------------------------------------------
##                                              growth
##                            (1)                 (2)                  (3)
## --------------------------------------------------------------------------------
## Constant                  0.640               0.512                0.957
##                          (0.490)             (0.458)              (0.580)
##
## tradeshare               2.306***            2.384***             1.681*
##                          (0.773)             (0.721)              (0.987)
##
## --------------------------------------------------------------------------------
## Observations                65                  64                   64
## R2                         0.124               0.150                0.045
## Adjusted R2                0.110               0.136                0.029
## Residual Std. Error   1.790 (df = 63)     1.667 (df = 62)      1.789 (df = 62)
## F Statistic         8.892*** (df = 1; 63) 10.938*** (df = 1; 62) 2.898* (df = 1; 62)
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

Correlation and scatterplot without Malta

```
cor(data2$tradeshare[data2$tradeshare<max(data2$tradeshare)],
    data2$growth[data2$tradeshare<max(data2$tradeshare)])
```

```
## [1] 0.2113246
```