Recitation 1: Practice Problems

Matthew Davis (UNI: wad2113)

May 30, 2023

The preamble chunk is where we load all the packages we'll be using. Notice the "include = FALSE" below when viewing this Notebook's .Rmd file; this prevents the preamble chunk from appearing in the knitted pdf document.

The preamble will only run successfully if you've installed of them so make sure you've done so before running them.

You'll notice two things in the chunk output: first, some packages "mask" one another's functions. This means you've loaded two or more packages which each have a function with the same name and so they conflict. By default, R assigns the conflicting name to the function from the package that was most recently loaded. Thus, using lag() will use dplyr's version of lag instead of stats' version of lag. If you wanted to make sure you're using a particular package's version, you can preface the function with the name of the packages like "stats::lag()". Masked functions are one of the most frustrating sources of coding problems for both new and experienced R users.

If in future R Notebooks, you might not want the preamble output to appear in your final pdf document since they can be quite long and take up a lot of space. To omit the output, simply start the chunk with "r, include=FALSE" instead of just "r"

Question 1: Stock-Watson, non-empirical Exercise 4.1

1a)

The predicted average test score is given by

$$\widehat{\text{Score}} = 520.4 - 5.82 \times 22$$

```
520.4-5.82*22
```

```
## [1] 392.36
```

1b)

The predicted decrease in the classroom average test score is

$$\Delta\widehat{\text{Score}} = (-5.82 \times 19) - (-5.82 \times 23)$$

```
(-5.82*19)-(-5.82*23)
```

```
## [1] 23.28
```

Alternatively, the predicted change is given by

$$\widehat{\Delta \text{Score}} = (-5.82 \times 23) - (-5.82 \times 19)$$

```
(-5.82*23)-(-5)
```

```
## [1] -128.86
```

1c)

Using the formula for $\widehat{\beta_0}$, we know the sample average of the test scores across the 100 classrooms is

$$\overline{Score} = \widehat{\beta}_0 + \widehat{\beta}_1 \times \overline{CS} = 520.4 - 5.82 \times 21.4$$

```
520.4-5.82*21.4
```

```
## [1] 395.852
```

1d)

$$SSR = (n-2)SER^2 = (100-2) \times 11.5^2 = 12961$$

The sum of squared residuals (SSR) is given by the following formula

$$SSR = (n-2)SER^2$$

Plugging in the given values:

```
(100-2)*11.5^2
```

```
## [1] 12960.5
```

Then use the following formula for R^2 to get the total sum of square (TSS):

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

```
12961/(1-0.08)
```

```
## [1] 14088.04
```

Finally, plug this into the sample variance formula

$$s_Y^2 = \frac{TSS}{n-1}$$

This works out to be:

```
14088/99
```

```
## [1] 142.303
```

We can also compute the standard deviation s_Y as the square root of the sample variance:

```
sqrt(1408)
```

```
## [1] 37.52333
```

Question 2: Non-textbook, non-empirical

2a: What kinds of factors are contained in u? Are these likely to be correlated with level of education?

Income, age, and family background (such as number of siblings) are just a few possibilities. It seems that each of these could be correlated with years of education.(Income and education are probably positively correlated; age and education may be negatively correlated because women in more recent cohorts have, on average, more education; and number of siblings and education are probably negatively correlated.)

2b: Will simple regression of kids on EDUC uncover the ceteris paribus ('all else equal') effect of education on fertility? Explain.

Not if the factors we listed in part (i) are correlated with EDUC. Because we would like to hold these factors fixed, they are part of the error term. But if u is correlated with EDUC, then E(u|EDUC) is not zero, and thus OLS Assumption (A2) fails.

Question 3: Stock-Watson Exercise 5.1

3a)

```
-5.82-1.96*2.21
```

```
## [1] -10.1516
```

```
-5.82+1.96*2.21
```

```
## [1] -1.4884
```

Thus, the 95% confidence interval is given by $-10.2 \leq \beta_1 \leq -1.48$

3b)

The t-statistic is given by

$$t^{act} = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)}$$

Plugging in numbers, the t-statistic is:

```
(-5.82-0)/2.21
```

```
## [1] -2.633484
```

The p-value for the test of the hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is

$$p = 2\Phi(-|t^{act}|)$$

We can compute this in R using the inverse normal function:

```
2*pnorm(-2.6335)
```

```
## [1] 0.008450984
```

This p-value is less than 0.01 so we can reject the null hypothesis at the 5% significance level

3c)

The t-statistic is -0.10. The p-value for the test is

```
2*pnorm(-0.10)
```

```
## [1] 0.9203443
```

The p-value is larger than the t-statistic so we cannot reject the null hypothesis at the 5% significance level. This also means that the value of -5.6 is contained in the 95% confidence interval centered at 0.

3d)

The 99% confidence interval for β_0 is

```
520.4-2.58*20.4
```

```
## [1] 467.768
```

```
520.4+2.58*20.4
```

```
## [1] 573.032
```

That is, $467.8 \leq \beta_0 \leq 573.0$

Question 4: Stock-Watson Empirical Exercise 4.1

4a) Scatterplot of growth vs. tradeshare

Load the growth dataset

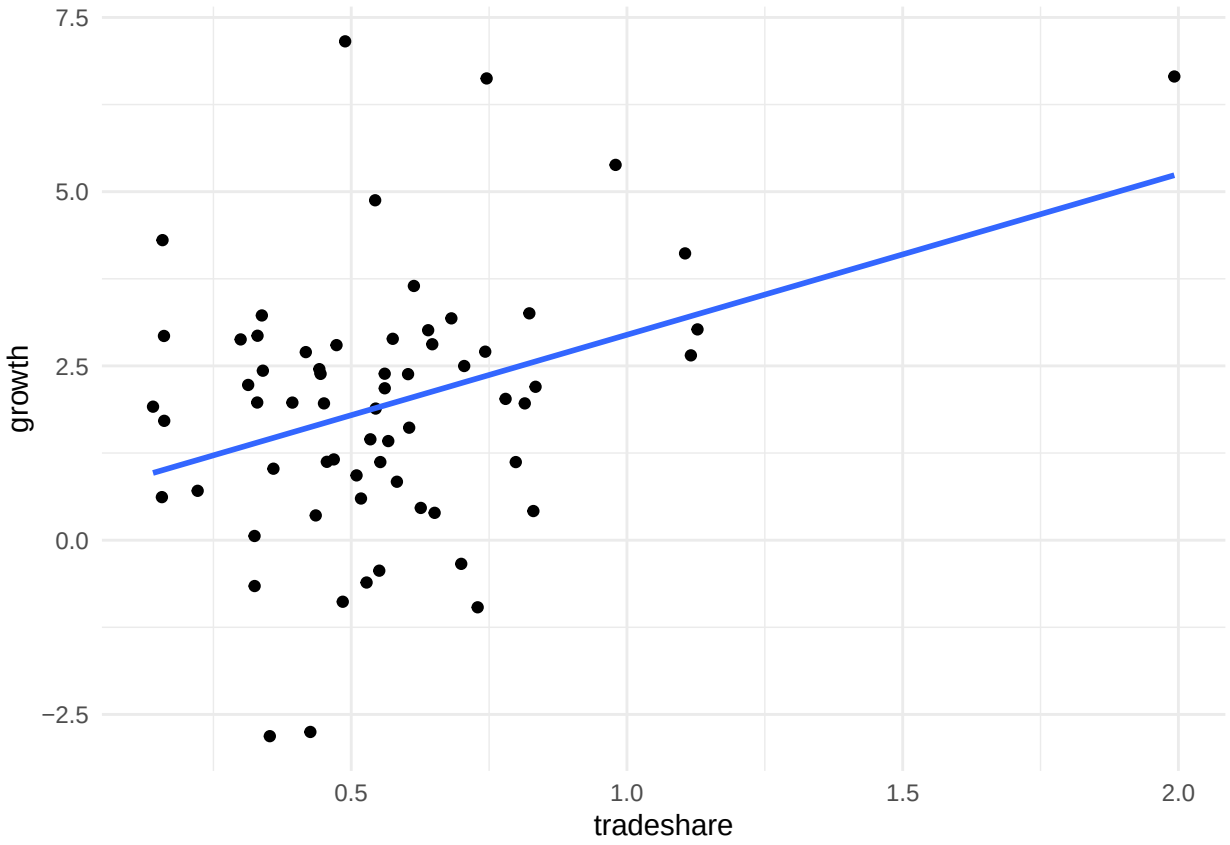
```
growth <- read.dta13('data/Growth.dta') # File names are case sensitive
head(growth)
```

```
##    country_name    growth oil    rgdp60 tradeshare yearsschool rev_coups
## 1         India 1.9151679   0  765.9998  0.1405020        1.45 0.1333333
## 2     Argentina 0.6176451   0 4462.0015  0.1566230        4.99 0.9333333
## 3         Japan 4.3047590   0 2953.9995  0.1577032        6.71 0.0000000
## 4        Brazil 2.9300966   0 1783.9999  0.1604051        2.89 0.1000000
## 5 United States 1.7122649   0 9895.0039  0.1608150        8.66 0.0000000
## 6    Bangladesh 0.7082631   0  951.9998  0.2214584        0.79 0.3064815
##   assasinations
## 1     0.8666667
## 2     1.9333333
## 3     0.2000000
## 4     0.1000000
## 5     0.4333333
## 6     0.1750000
```

We want to construct a scatterplot of growth on average tradeshare:

```
ggplot(growth, aes(x = tradeshare, y = growth)) +
  theme_minimal() +
  geom_point() +
  geom_smooth(method = 'lm', se = F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



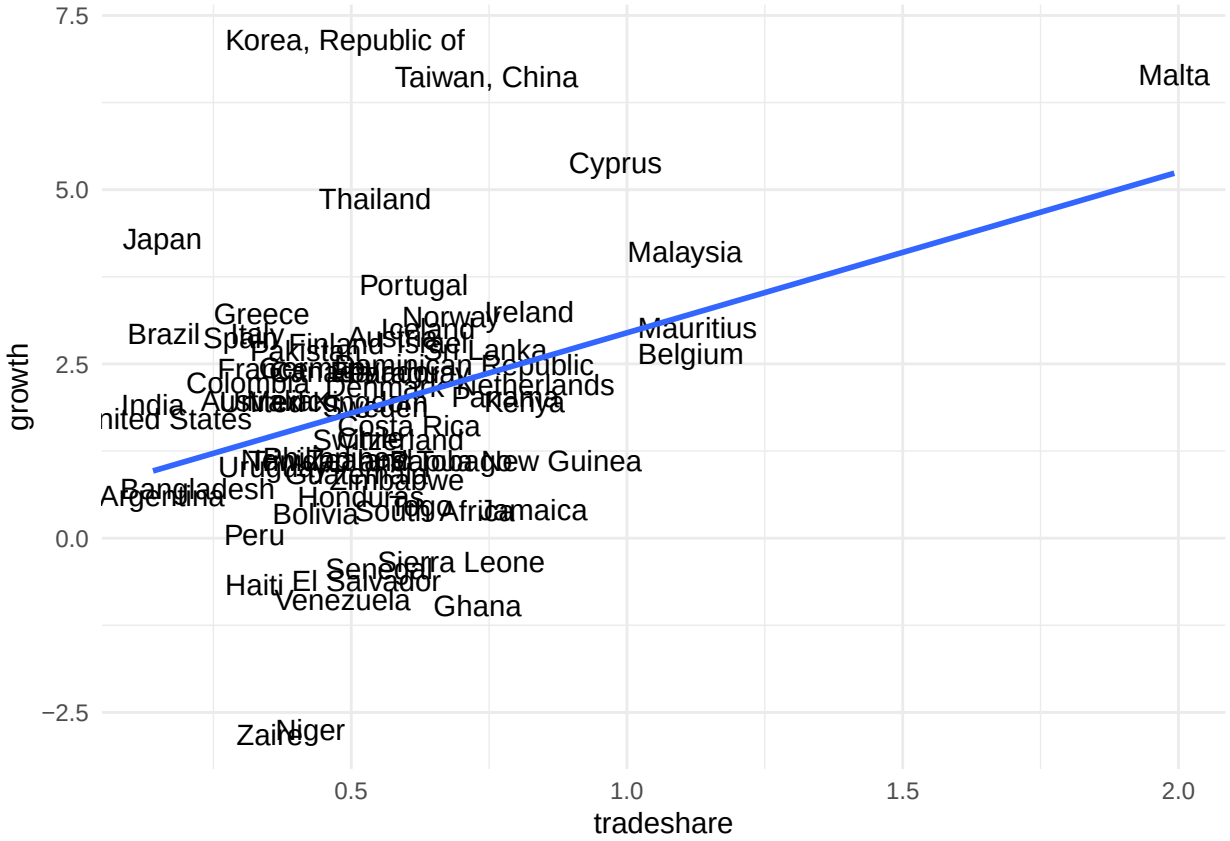
4b) Malta as an outlier

Notice that there is a very anomalous value here whose tradeshare is far greater than any other country. We can identify it by including labels in our graph:

```
ggplot(growth, aes(x = tradeshare, y = growth, label = country_name)) + # new argument: label
  theme_minimal() +
  geom_text() + # new command: replace geom_point() with geom_text()
  geom_smooth(method = 'lm', se = F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: label
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



It's Malta. As it's an outlier, we might want to drop it.

4c) Full regression: estimates and predictions

```
growth.model <- lm(growth ~ tradeshare,
                   data = growth)
extract_eq(growth.model,
           use_coefs = TRUE)
```

$$\widehat{\text{growth}} = 0.64 + 2.31(\text{tradeshare}) \tag{1}$$

```
summary(growth.model)
```

```
##
## Call:
## lm(formula = growth ~ tradeshare, data = growth)
##
## 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
```

$$\widehat{\text{growth}} = 0.64 + 2.31(\text{tradeshare})$$

Countries with tradeshares of 0.5 and 1.0 will have predicted growth rates

```
0.6403 + 2.31*0.5
```

```
## [1] 1.7953
```

```
0.6403 + 2.31*1
```

```
## [1] 2.9503
```

4d) Regression without Malta

Let's use the *dplyr* package (install if you haven't) to modify the dataset. dplyr is one of the most popular packages in R so we'll probably be using it a lot.

```
library(dplyr)
growth2 <- filter(growth, tradeshare <= 1.5)
growth.model2 <- lm(growth ~ tradeshare, data = growth2)
```

This creates a new dataset that is the same as the original dataset but filters out any observations whose tradeshare is less than or equal to 1.5.

Alternatively we could do this in one line without defining a new dataset:

```
growth.model2 <- lm(growth ~ tradeshare,
                    data = filter(growth, tradeshare <= 1.5))
```

Then we have:

```
extract_eq(growth.model2,
           use_coefs = TRUE)
```

$$\widehat{\text{growth}} = 0.96 + 1.68(\text{tradeshare}) \tag{2}$$

```
summary(growth.model2)
```

```
##
## Call:
## lm(formula = growth ~ tradeshare, data = filter(growth, tradeshare <=
##     1.5))
##
## 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
```

$$\widehat{\text{growth}} = 0.96 + 1.68(\text{tradeshare})$$

```
0.9574 + 1.6809*0.5
```

```
## [1] 1.79785
```

```
0.9574 + 1.6809*1
```

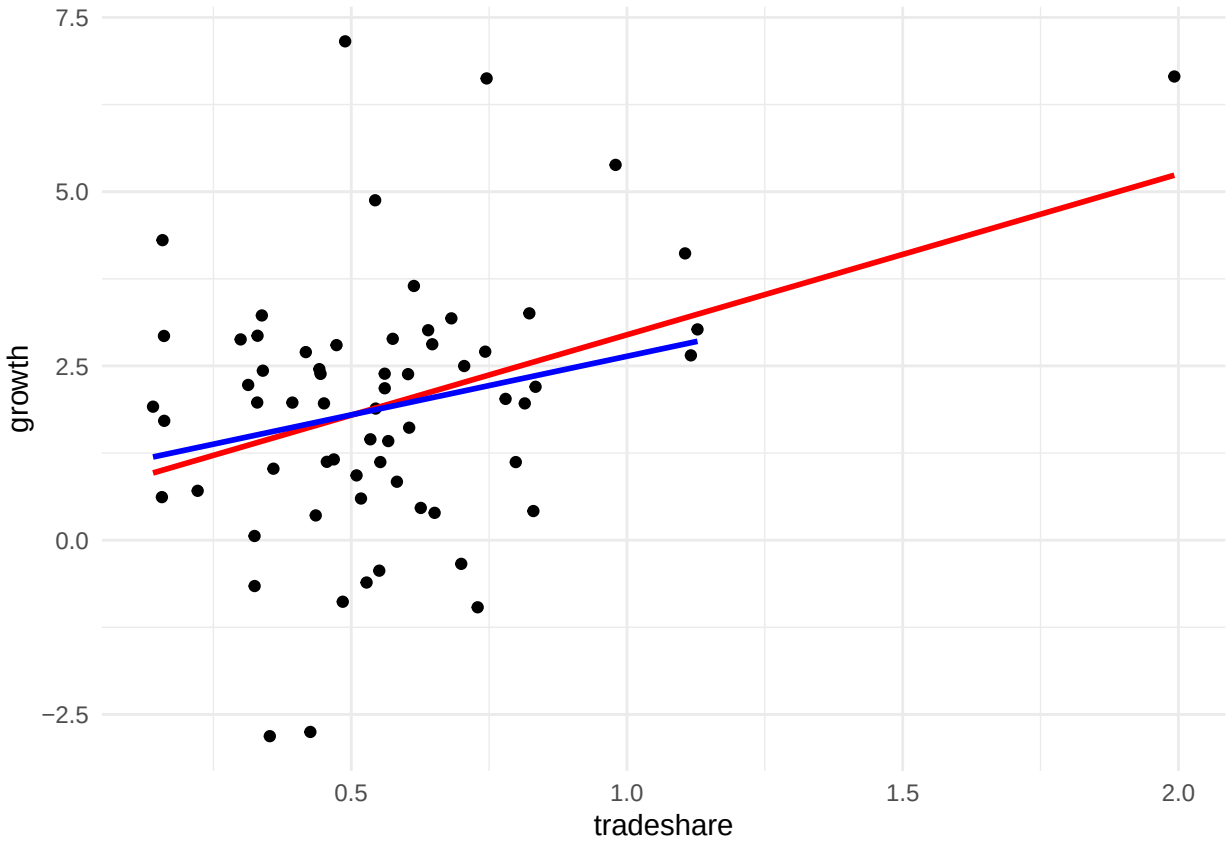
```
## [1] 2.6383
```

4e) Plot regression lines from both

We'll plot both lines of best fit on the original scatterplot

```
ggplot(growth, aes(x = tradeshare, y = growth)) +
  theme_minimal() +
  geom_point() +
  geom_smooth(method = 'lm', se = F, color = 'red') +
  geom_smooth(method = 'lm', se = F, color = 'blue', data = growth2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Here, we specified the filtered growth dataset as an argument for the second line of best fit. The blue line is shorter because the Malta point is not in that dataset. In comparison, the base R graph just has the lines extending indefinitely in both directions.

4f)

Something about Malta being a small island nation, etc.

5 (Bonus): Stock-Watson Empirical Exercise 5.2

```
growth <- read.dta13('data/Growth.dta')
growth.model <- lm_robust(growth ~ tradeshare,
                          data = filter(growth, tradeshare < 1.5),
                          se_type = 'HC1')
summary(growth.model)
```

```
##
## Call:
## lm_robust(formula = growth ~ tradeshare, data = filter(growth,
##     tradeshare < 1.5), se_type = "HC1")
##
## 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
```

5a)

The p-value on tradeshare is 0.0567. This means we can reject the null hypothesis $H_0 : \beta_1 = 0$ vs. a two-sided alternative hypothesis at the 10% level, but not at either a 5% or 1% significance level, at least using robust standard errors.

5b)

The p-value associated with the coefficient's t-statistic is 0.0567, as mentioned above.

5c)

```
confint(growth.model, level = 0.9)
```

```
##                    5 %     95 %
## (Intercept) 0.06229901 1.852522
## tradeshare  0.23549365 3.126316
```

The 90% confidence interval is reported in the regression output is (0.235, 3.13)