hw03: Run simple linear models in R and interpret the outputs

06 July, 2017

1 Overview

Purpose

To continue working with **tidyverse** and **ggplot2** packages and run a number of linear regressions to explore the relationships in the PWT dataset in more detail. Those of you who are also taking Mr Watkins' Econometrics class will see quite a few parallels in material covered.

Instructions

In this assignment, you will

- clone the assignment repository and make a working branch (eg. `solution` branch);
- solve the problems in Section **??**;
- write the solutions in `solution.Rmd` and knit the file;
- commit `solution.Rmd` and `solution.pdf`; and
- open a Pull Request.

Set Up

For this assignement we will continue working with the pwt dataset.

```
library(tidyverse)

pwt <- haven::read_dta("~/Data/pwt90.dta")
```

Simple growth empirics - absolute convergence hypothesis

One of the predictions of the Solow model is that the income levels in countries across the world should converge to the same steady state in the long run. Specifically, we would expect the poorer countries to grow faster and as such (eventually) catch up with the more affluent countries. Let's check whether the data in the PWT supports this prediction.

First, let's create a subset of pwt that would contain all the countries in the world, with the exception of nations with a population of less than 1 million. Within this subset, let's also mutate in the variables for growth as a percentage change year on year as follows.

```
pwt_world <- pwt %>%
  filter(year >= 1960 & pop >= 1) %>%
  select(country, year, rgdpo, pop) %>%
  mutate(gdp_per_capita = rgdpo / pop) %>%
  group_by(country) %>%
  mutate(growth = (log(rgdpo) - log(lag(rgdpo)))*100)
```

Now let's group summarise the previously defined dataframe to get the growth means. The average growth rates for the first 5 countries in pwt_world are as follows:

```
growth_world <- pwt_world %>%
  group_by(country) %>%
  summarise(mean(growth, na.rm = TRUE))
head(growth_world, 5)
```

```
## # A tibble: 5 x 2
##     country `mean(growth, na.rm = TRUE)`
##       <chr>                        <dbl>
## 1   Albania                     3.417804
## 2   Algeria                     2.867912
## 3    Angola                     4.583719
## 4 Argentina                     4.798053
## 5   Armenia                     1.172207
```

We want to see if there is convergence over a long enough time scale, so let's check whether the lower income levels in 1960 correspond to a relatively higher average growth rate over 1960-2014. Eyeballing the data, we can easily tell that some countries (e.g. Aruba) don't have data for their income levels in 1960 within PWT. We filter those out and merge the resulting dataframe with the one from the previous step

```
pwtworld_1960 <- pwt_world %>%
  filter(year == 1960 & gdp_per_capita >= 0) %>%
  select(country, gdp_per_capita)

world_converge <- merge.data.frame(growth_world, pwtworld_1960,
                                   by = "country" , all = F)
colnames(world_converge) <- c("country", "growth", "income_level_1960")
```

Now, we fit a linear model where average growth is the dependent variable and gdp per worker in 1960 is the explanatory variable as follows:

```
growth_1960inc <- lm(formula = growth ~ income_level_1960, data = world_converge)
summary(growth_1960inc)
```

```
##
## Call:
## lm(formula = growth ~ income_level_1960, data = world_converge)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7705 -0.6320  0.0273  0.6628  3.9535
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.531e+00  1.911e-01  23.706  < 2e-16 ***
## income_level_1960 -1.057e-04  3.214e-05  -3.288  0.00145 **
```

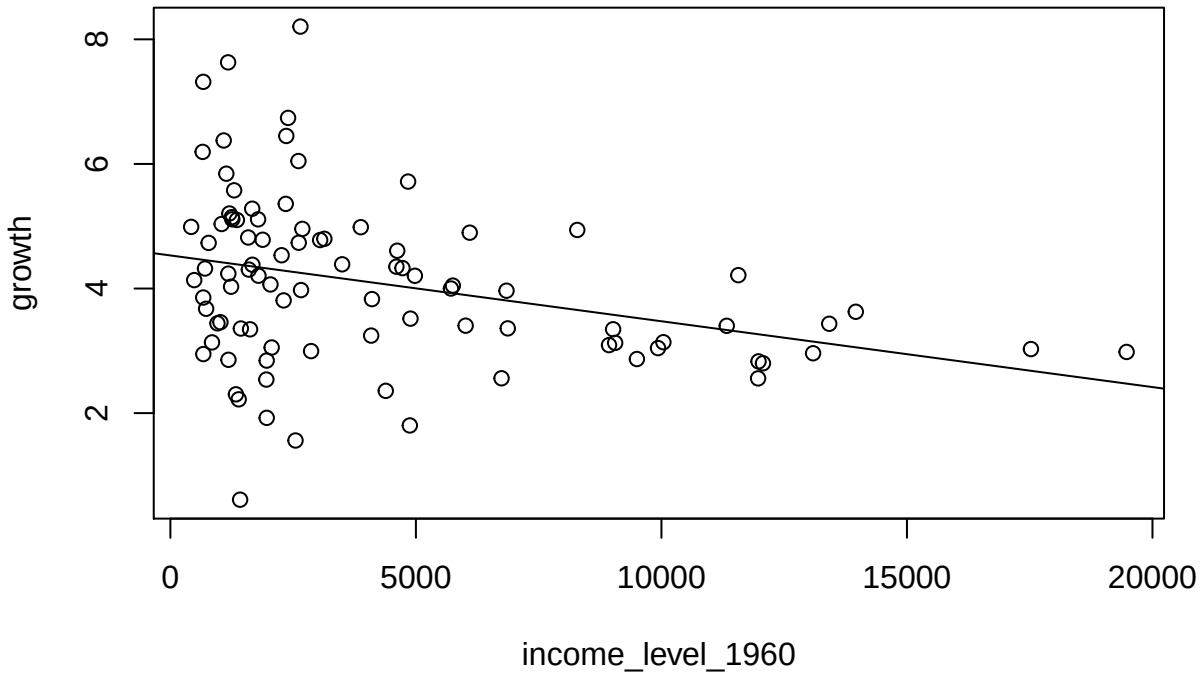
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.273 on 89 degrees of freedom
## Multiple R-squared:  0.1083, Adjusted R-squared:  0.09827
## F-statistic: 10.81 on 1 and 89 DF,  p-value: 0.001448
```

By looking at the output summary level we can note the following.[1]

The sign on income level variable is negative as we expect it to be - we predicted a negative relationship between the income level and the growth rates. The effect is also statistically significant however the estimator is very small and the correlation, given by the R-squared value, between growth and the income level is very low.

Let's plot this result, abline adds a line of best fit defined by our OLS linear model object we got previously:

```
plot(growth ~ income_level_1960, data = world_converge)
abline(growth_1960inc)
```



Alternatively, we can create a more visually meaningful plot by making use of the ggplot2 package.

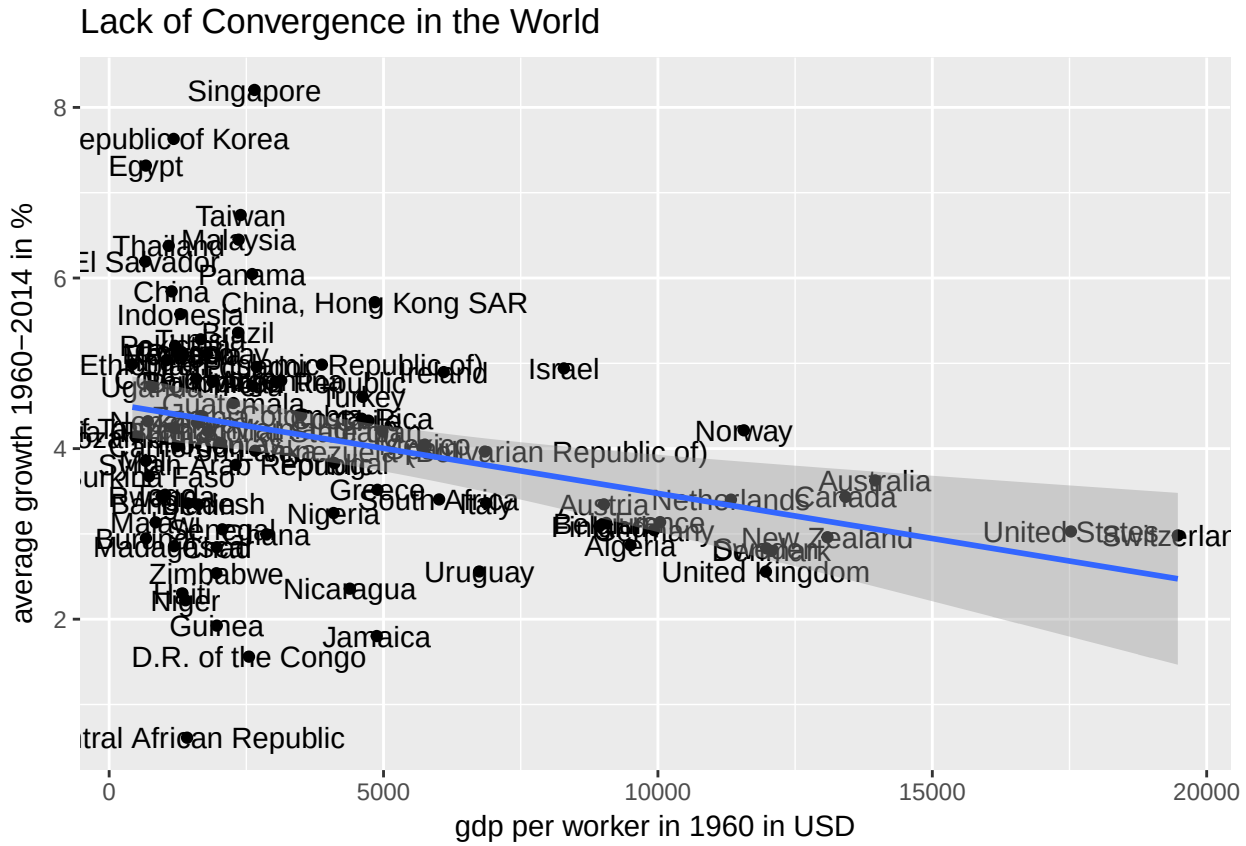
```
library(ggplot2)
p_world <- ggplot(world_converge, aes(x = world_converge[,3],
                                      y = world_converge[,2],
                                      label = country)) + geom_point(
                                      ) + geom_text(aes(label=country))
```

Using the ggplot() function we can also add an OLS type regression line to get a visual clue of whether there is correlation between income levels in 1960 and the convergence rate.

[1] We omit exploring tests on various properties and the behaviour of a specified model. To cover these important steps would require a serious study of Econometrics.

Note that in this case we didnt make an lm object, instead geom_smooth with the lm method runs an lm(y ~ x) as supplied into the ggplot function by the columns in the dataframe in the aes() argument.

```
p_world + geom_smooth(method = lm) + labs(x = "gdp per worker in 1960 in USD",
                                          y = "average growth 1960-2014 in %",
                                          title = "Lack of Convergence in the World")
```



Just as before, doesn't look like we have evidence of the convergence here. Perhaps the convergence is sensitive to the choice of countries. As an exercise, to improve your familiarity with the tools you have used in solving hw02, you will be asked to recreate these steps for the case of OECD countries.

2 Some more examples of linear regressions

Recall the following formulation from the class. Let's check relationship between capital return and real gdp output in the US over the time period.

```
usa <- pwt %>%
  filter(country == "United States") %>%
  mutate(capsh = (1 - labsh),
         caprtrn = capsh * rgdpo / rkna - delta)
```

Because here we have a time series, to correct for the time trend in both rgdpo and caprtrn we add the variable for the year. Note how you can log transform rgdpo within the lm() and also how you add additional variables to regress against gdpo.

```
rgdpo_caprtrn <- lm(formula = log(rgdpo) ~  caprtrn + year, data = usa)
summary(rgdpo_caprtrn)
```

```
##
## Call:
```

```
## lm(formula = log(rgdpo) ~ caprtrn + year, data = usa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12122 -0.03033  0.01407  0.04103  0.08216
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.867e+01  9.633e-01 -50.520   <2e-16 ***
## caprtrn     -3.508e+00  1.448e+00  -2.422   0.0184 *
## year         3.262e-02  5.206e-04  62.662   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05515 on 62 degrees of freedom
## Multiple R-squared:  0.9919, Adjusted R-squared:  0.9916
## F-statistic:  3788 on 2 and 62 DF,  p-value: < 2.2e-16
```

If the time trend is not captured by the addition of the year variable, the correlation appears much stronger.[2]

Let's have a quick look at the correlations of the variables in pwt.

Specifically, using the subset of the world data for the year 2010, we can get the **correlation matrix** of the data as follows[3]

```
pwt_world_2010 <- pwt_world %>% filter(year == 2010)
pwt_world_2010 <- pwt_world_2010[, -2]
cor(pwt_world_2010[sapply(pwt_world_2010, is.numeric)])
```

```
##                     rgdpo         pop gdp_per_capita      growth
## rgdpo          1.00000000  0.71539418     0.19185960 -0.05019895
## pop            0.71539418  1.00000000    -0.05658528  0.07375418
## gdp_per_capita 0.19185960 -0.05658528     1.00000000 -0.11047834
## growth        -0.05019895  0.07375418    -0.11047834  1.00000000
```

You can use this as a guide for the potential relationships to explore within the dataset. Other variables, which we have filtered out might also be of interest.

Problems

1) Filter out all the non-OECD countries from pwt and check for the evidence of convergence among OECD countries (reproduce the first section of this problem set). Comment on your findings.
2) Choose either a year or a specific country from the pwt dataset. Fit an lm model of gdp per capita versus capital labour ratio $\frac{K}{L}$. Comment on the output of your model.
3) Using the same subset as in (2) fit a bivariate model of both human capital and capital per labour. How does this model compare with the previous?
4) Choose a variable (or variables) in pwt and check their growth empirics. Justify your choice of variable(s).

[2] Most of this effect is due to a **spurious correlation**

[3] Note that we also dropped the now unnecessary year column and made sure that the correlation matrix ignored the character variable for country.