DA3_2

Tamas Koncz

2017 november 18

Download cross-country data on life expectancy and GDP per capita. "GDP per capita, PPP (constant)" and "Life expectancy at birth (total)"

1. Delete unnecessary columns and save a csv file with three columns only: country name, life expectancy and GDP per capita. Keep countries with non-missing values for life expectancy and GDP per capita. Document what you do.

First, we'll use the WB API for getting our data set to R.

```
x <- wb(country = "countries_only", indicator = c("NY.GDP.PCAP.CD", "SP.DYN.LE00.IN", "SP.POP.TOTL") ,
countries <- data.table(x)
```

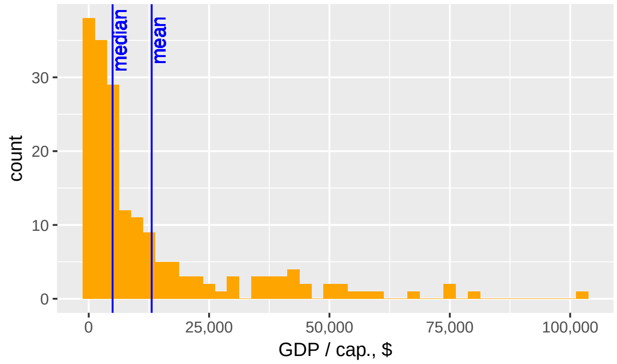
Once we have the data loaded, we format the data into a workable format. We are making four transformations:

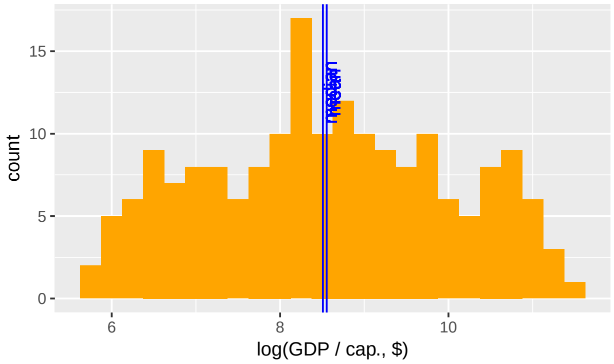
1. Separate columns created for each of our variables
2. Columns are renamed to short names for easier coding
3. Rows with NA values are removed
4. Population is redefined as population in millions for easier handling

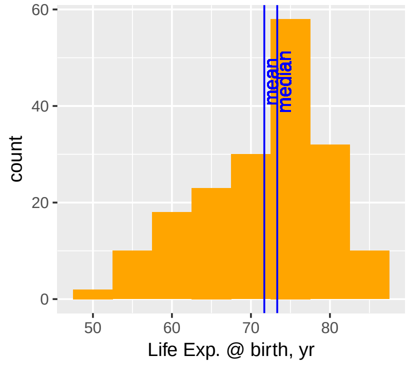
As a last step, we save our data table to a csv file as requested in the exercise.

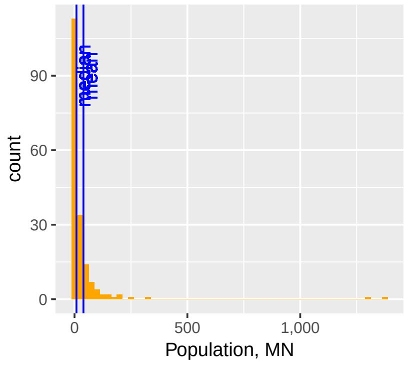
```
dt <- countries[, c('value', 'country', 'indicator')]
dt <- spread(dt, key = 'indicator', value = 'value', fill = NA)

#saving string names for later use (charts, ...)
s_country = colnames(dt)[1]
s_gdp = "GDP / cap., $" ##colnames(dt)[2]
s_life_exp = "Life Exp. @ birth, yr" ##colnames(dt)[3]
s_pop = "Population, MN" ##colnames(dt)[4]
s_lnpop = "log(Population, MN)"
s_lngdp = "log(GDP / cap., $)"

setnames(dt, old = colnames(dt), new = c('country', 'gdp', 'life_exp', 'pop'))

dt <- dt[is.na(gdp) == FALSE, ]
dt <- dt[is.na(pop) == FALSE, ]
dt <- dt[is.na(life_exp) == FALSE, ]

dt[, pop := pop / 1000000]
dt[, lngdp := log(gdp)]

write.csv(x = dt, file = 'WD_Data_Filtered.csv')
```

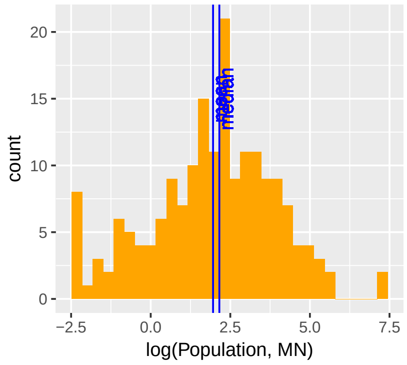
Before we create our models, let's explore our data set visually. First, using histograms to visualize single-variable distributions:



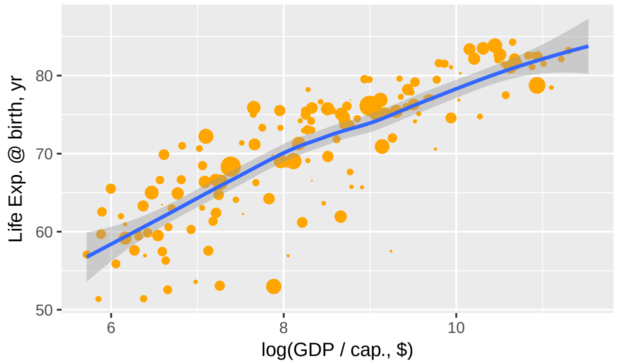


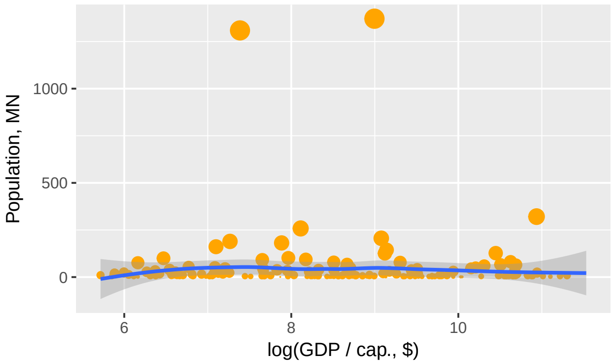


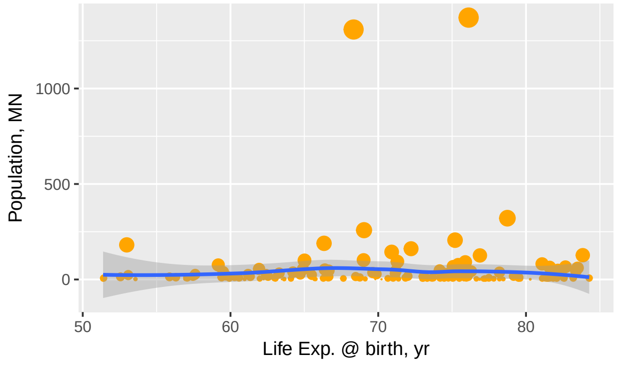


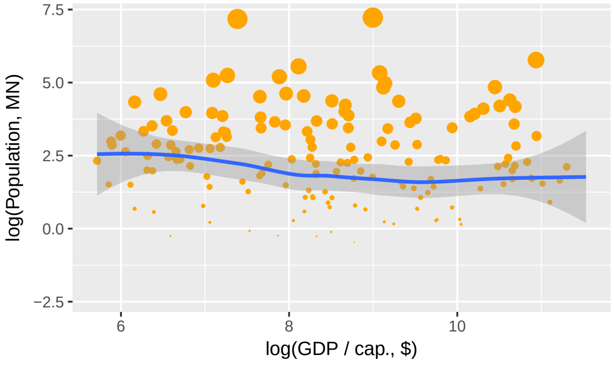


Then, we'll take a look at multi-varite distributions with scatterplots (inc. default loess lines). (Bubble sizes show the size of the population)







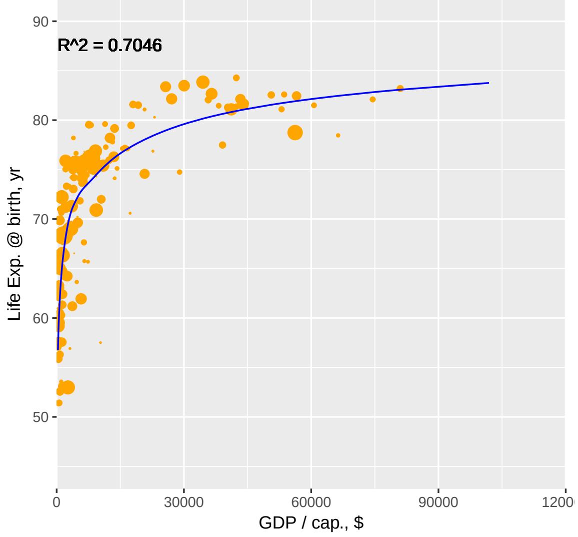


2. Estimate a lowess regression of life expectancy on ln gdp per capita. Estimate a linear regression of life expectancy on GDP per capita that best captures the nonlinearity you found (life expectancy on a piecewise linear spline or a polynomial in the explanatory variable). Argue for your choice. Report the coefficient estimates as well as their confidence interval, interpret and visualize the results.

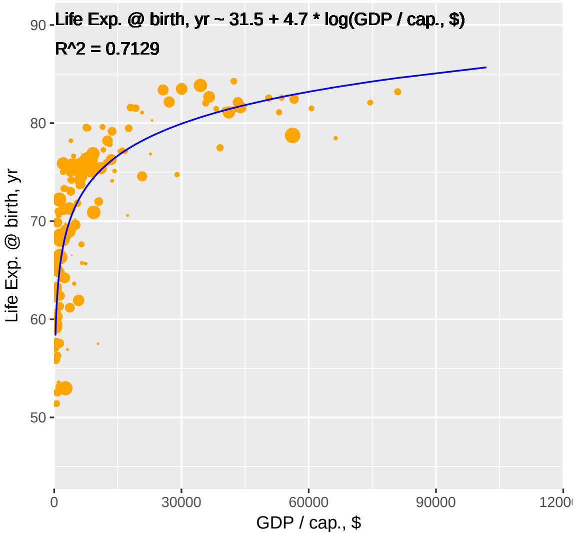
Below code was used to create our different models:

Visualizing the results:

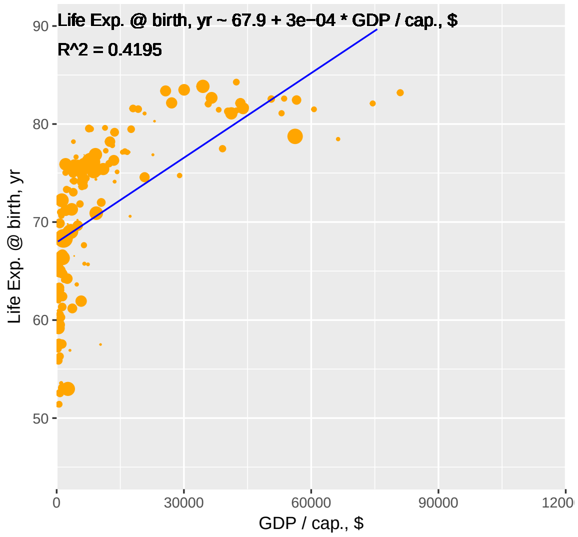
Model: Lowess



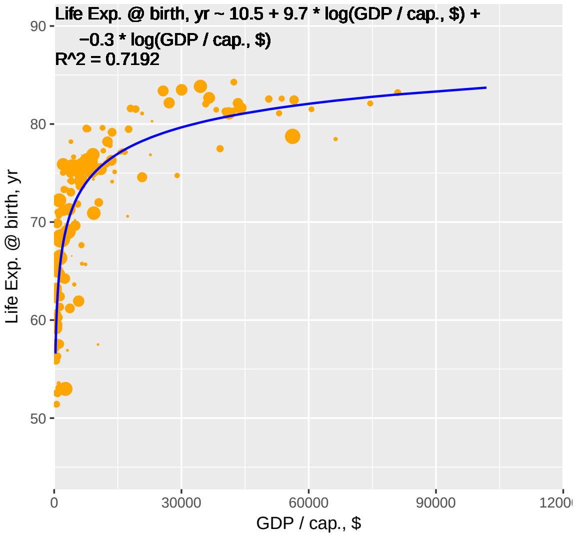
Model: Level − Log Linear Regression



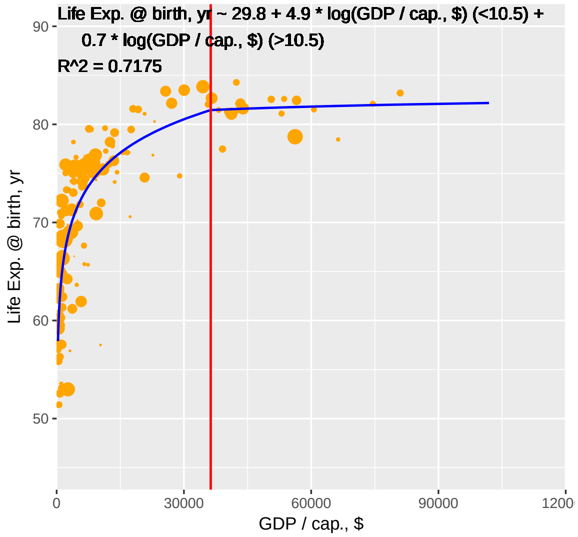
Model: Linear Regression



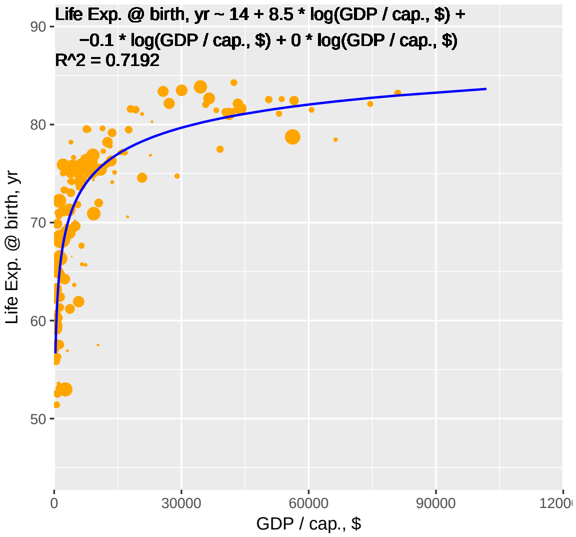
Model: Linear Regression w/ square polinom



Model: Spline w/ 1 knot



Model: Linear Regression w/ cubic polinom



Note on the visuals: The different plots has been scaled back to show the non-log GDP / capita on the X axis, even if the models were run on the log transformation.

The results follow our expactations - the more complex models tend to drive slightly better fits in the sample. Having small differences is actually key in selecting our favored model:

While a simple linear regression on non-log GDP is visibly not the best model, from models run on log(GDP per capita) we actually argue to select the simplest one, our log-linear regression.
This model's fit in the sample (measured by Rˆ2 and visually) is just insignificantly behind of the more complex models' - this fact in itself justifies the use of the simpler model, as we should only opt for more complexity if we can increase our predictive power (and even then trade-offs like possible overfit should be examined, which is not analyzed here).

Also, based on p values, only the beta of log(GDP) variable is significant on the 95% and 99% confidence levels. Supporting our decision, below is a summary for the coefficients and their confidence intervals in different models:

Dependent variable:

life_exp

(1)

(2)

(3)

(4)

(5)

lngdp

4.700***

9.746***

8.454

(0.222)

(2.517)

(22.935)

gdp

0.0003***

(0.00003)

lspline(lngdp, knot)1

4.924***

(0.256)

lspline(lngdp, knot)2

0.700

(2.345)

lngdp_sq

-0.295**

-0.141

(0.146)

(2.718)

lngdp_cub

-0.006

(0.106)

Constant

31.482***

67.925***

29.757***

10.491

14.032

(1.924)

(0.563)

(2.162)

(10.601)

(63.375)

Observations

183

183

183

183

183

R2

0.713

0.419

0.717

0.719

0.719

Adjusted R2

0.711

0.416

0.714

0.716

0.715

Residual Std. Error

4.347 (df = 181)

6.181 (df = 181)

4.324 (df = 180)

4.311 (df = 180)

4.323 (df = 179)

F Statistic

449.410*** (df = 1; 181)

130.788*** (df = 1; 181)

228.576*** (df = 2; 180)

230.519*** (df = 2; 180)

152.829*** (df = 3; 179)

Note:

p<0.1; ***p<0.05;*** p<0.01

The level-log linear model, with close to .7 R^2 actually carries a very strong predictive power. We can interpret the formula as follows: on a 95% confidence level, we can say that a country with 10% higher GDP / capita is expected to have a life expetancy higher by 0.45-0.49 years on average.

One factor that this model leaves unexplored is the difference is different countries populations. We revisit that in the next section.

3. Estimate a weighted regression (weight=population). Compare results to what we saw in class.

Modeling:

```
lm_weighted_model <- lm(life_exp ~ lngdp, weights = pop, data = dt)
dt$lm_weighted_pred <- predict(lm_weighted_model)
r2_lm_weighted <- summary(lm_weighted_model)$r.squared

formula_lm_weighted = paste(s_life_exp,"~", round(coefficients(lm_weighted_model)[1], 1),
                            "+", round(coefficients(lm_weighted_model)[2], 1), "*", s_lngdp, "\nweighted by",
```

Dependent variable:

life_exp

(1)

(2)

lngdp

4.414***

4.700***

(0.238)

(0.222)

Constant

34.688***

31.482***

(2.027)

(1.924)

Observations

183

183

R2

0.655

0.713

Adjusted R2

0.653

0.711

Residual Std. Error (df = 181)

25.359

4.347

F Statistic (df = 1; 181)

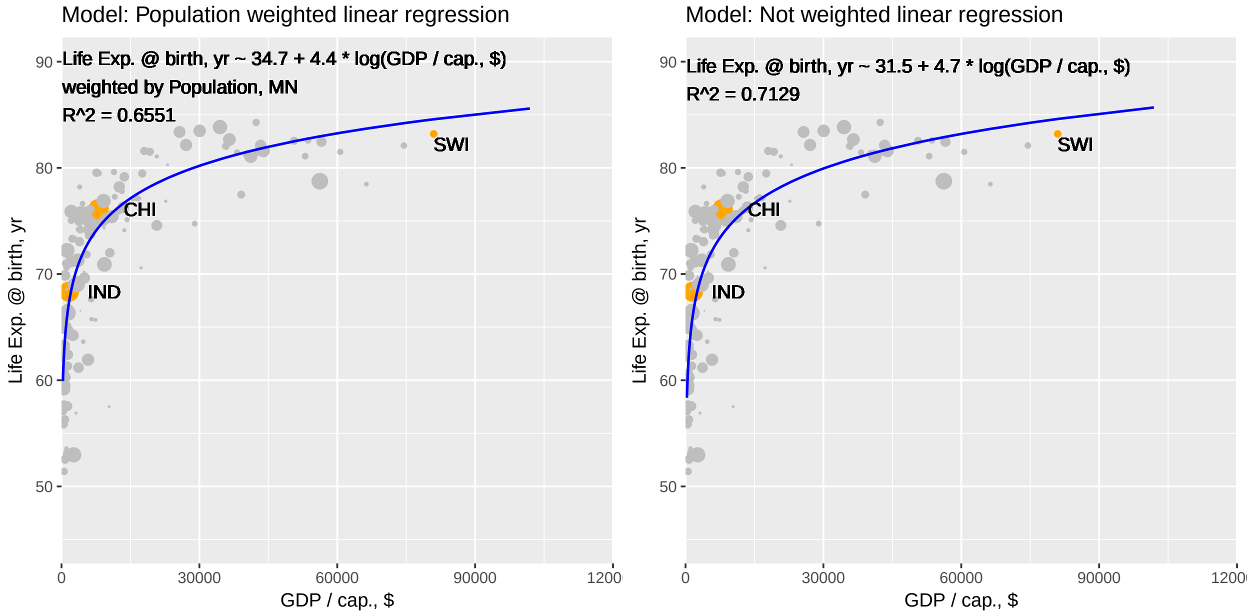
343.790***

449.410***

Note:

p<0.1; ***p<0.05;*** p<0.01

Visualizing results:



Visually, it's hard to tell the difference between the two models, however we can see it in the R^2 and regression formula.

To better understand of what is the difference between the weighted and the non-weighted models, consider the below tables showing predicted values and residuals for the two models:

Largest 5 countries:

country	life_exp	lm_weighted_pred	resid_w	lm_ln_pred	resid_nw
China	76.12	74.40	-1.72	73.77	-2.35
India	68.33	67.29	-1.04	66.20	-2.13
United States	78.74	82.96	4.22	82.89	4.15
Indonesia	69.04	70.50	1.46	69.61	0.58
Brazil	75.20	74.76	-0.44	74.15	-1.05
Pakistan	66.33	66.76	0.43	65.64	-0.70

Smallest 5 countries:

country	life_exp	lm_weighted_pred	resid_w	lm_ln_pred	resid_nw
Seychelles	73.23	77.25	4.02	76.80	3.57
Antigua and Barbuda	76.08	76.69	0.61	76.21	0.13
Virgin Islands (U.S.)	79.87	81.04	1.17	80.84	0.97
Micronesia, Fed. Sts.	69.05	70.05	1.01	69.14	0.09
Tonga	72.84	71.40	-1.44	70.58	-2.26
Grenada	73.50	74.98	1.48	74.39	0.89

What we see is that generally the fit improved for large countries, while it got worse for smaller ones (as expected).

The difference comes from the actual interpretation of the data: while the simple regression predicts that life expectancy is ~.47 years longer on average **in countries** with 10% higher GDP per capita, the weighted one says that **people** who live in countries with 10% higher GDP per capita live longer by .44 years on average.

While the actual coefficients are somewhat different from results seen in the class (even if the confidence intervals are incorporated), the dynamics are similar: most importantly, we cannot rule out confidently (on usual levels) that coefficients are different, the results are similar for both. This can be due to the fact observed in the lecture, that the largest countries are mostly around the middle of the GDP per capita distribution.