Robust regression

Contents

This assignment must be submitted before **February 23th before 10pm** on Moodle.

Data

This exercise is based on the *gapminder* dataset from the package of the same name.

> Jennifer Bryan (2017). gapminder: Data from Gapminder. R package version 0.3.0. https://CRAN.R-project.org/package=gapminder

This dataset includes the life expectancy (*lifeExp*), population (*pop*) and GDP per capita (*gdpPercap*) for 142 countries and 12 years (every 5 years between 1952 and 2007).

```
library(gapminder)
str(gapminder)
```

```
## tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
##  $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 163
##  $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

1. Effect of GDP and time on life expectancy

a) First, visualize the life expectancy as a function of GDP per capita and year. It is suggested to represent the logarithm of *gdpPercap* and to separate the different years, for example with facets in *ggplot2*: `... + facet_wrap(~year)`.

What general trends do you observe? Are there extreme values that could strongly influence a regression model? If so, try to identify these data in the table based on the position of the points in the graph.

b) Perform a linear regression (`lm`) to determine the effect of GDP per capita, year and their interaction on life expectancy. To help interpret the coefficients, perform the following transformations on the predictors:

- Take the logarithm of *gdpPercap* and standardize it with the function `scale`. *Reminder*: `scale(x)` subtracts each value of `x` from its mean and divides by its standard deviation, so the resulting variable has a mean of 0 and a standard deviation of 1; it represents the number of standard deviations above or below the mean.
- Replace *year* with the number of years since 1952.

Interpret the meaning of each of the coefficients in the model and then refer to the diagnostic graphs. Are the assumptions of the linear model met?

c) Compare the result of the model in (b) with two more robust alternatives: robust regression based on Tukey's biweight (function `lmrob` from the *robustbase* package) and median regression (function `rq` from the *quantreg* package, choosing only the median quantile). Explain how the estimates and standard errors of the coefficients differ between the three methods.

Note: Use the `showAlgo = FALSE` option when applying the `summary` function to the output of `lmrob`, to simplify the summary.

(d) Superimpose the regression lines of the three models on the graph in (a). With `ggplot` you can use the `geom_smooth` function with `method = "lm"` for linear regression and `method = "lmrob"` for robust regression. For median regression you can use `geom_quantile` as seen in the notes.

2. Variation of effects by quantile

a) Based on your observation of the data in 1(a), would it be useful to model different quantiles of life expectancy based on the predictors? Justify your answer.

b) Perform a quantile regression with the same predictors as in 1(b), with the following quantiles: `(0.1, 0.25, 0.5, 0.75, 0.9)`. Use the `plot` function on the quantile regression summary and describe how the effect of the predictors varies between quantiles.

c) Superimpose the quantile regression lines on the graph of the data. Do the trends for each quantile appear to be affected by extreme values?

Note on international comparisons

While this dataset is useful for illustrating the concepts of robust regression and quantile regression, it should be noted that this type of statistical analysis comparing variables measured at the national level has several limitations:

- It cannot be assumed that the associations detected apply at a smaller scale (e.g., the relationship between life expectancy and income when comparing national averages is not necessarily the same as the relationship between life expectancy and income at the level of individuals living in each country).
- Averages calculated in different countries are not independent observations, because environmental, social and economic conditions are correlated between nearby countries.
- There are many factors that differentiate countries, so it is difficult to interpret an association as a causal link.

Many articles, particularly in the social sciences, have been published on the methods to be used to make this type of *cross-country comparisons*.