Statistics for Data Science / Exercise 03

Pierpaolo Brutti

Simple linear regression / Recap

Assume y_i represents the value of what is generally known as the *response variable* on the $i^{\texttt{th}}$ individual and that x_i represents the individual's values on what is most often called an *explanatory variable*. The simple[1] linear regression model is then

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i \in \{1, \dots, n\},$$

where β_0 is the intercept and β_1 is the slope of the linear relationship assumed between the response and explanatory variables and ε_i is an error term. Typically the error terms are assumed to be independent random variables having a Normal distribution with mean zero and constant variance σ_ε^2 but this is not needed for *prediction*.

The regression coefficients, β_0 and β_1, may be estimated as $\widehat{\beta}_0$ and $\widehat{\beta}_1$ using **least squares method**, in which the sum of squared differences between the observed values of the response variable y_i and the values *predicted* by the regression equation $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1\, x_i$ is minimised. By taking derivatives w.r.t. β_0 and β_1 and setting to zero, this leads to the least squares estimates:

$$\widehat{\beta}_1 = \frac{\widehat{\sigma}_{X,Y}}{\widehat{\sigma}_X^2}$$
$$\widehat{\beta}_0 = \bar{y} + \widehat{\beta}_1 \cdot \bar{x},$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(covariate empirical average)}$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{(response empirical average)}$$
$$\widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{(covariate empirical variance)}$$
$$\widehat{\sigma}_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{x}) \quad \text{(empirical covariance)}$$

Remember that the *predicted values* of the response variable y from the model are

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i,$$

The variance σ_ε^2 of the error terms is estimated as

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

whereas the estimated variance of the estimate of the slope parameter is

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

[1] Only one explanatory variable X.

Exercise: How Old is the Universe?

Description

The dataset `hubble` contains the relative velocity and the distance of 24 galaxies, according to measurements made using the Hubble Space Telescope.

More specifically, *velocities* are assessed by measuring the Doppler red shift in the spectrum of light observed from the galaxies concerned, although some correction for "local" velocity components is required. *Distances* are measured using the known relationship between the period of Cepheid variable stars and their luminosity.

In this exercise you will use *simple* linear regression to estimate the age of the universe.

Exercise

0. Place the file `hubble.RData` in a folder named `ex03`. In `RStudio`, create a new project inside this folder.
1. Load into `R` the data contained in the `hubble.RData` file.
2. Take a quick look using the appropriate `R` functions: what kind of object are you looking at? How many observations? How many variables? What are the names of the variables?
3. Evaluate basic *univariate* statistics for all the variables involved (e.g. `mean()`, `median()`, `sd()`, etc.)
4. Produce suitable graphical summaries for each variable in the dataset (e.g. `hist()`, `boxplot()`). Briefly comment the result.
5. Produce a scatterplot of velocity vs distance. Just by looking at this plot, what do you expect in terms of linear correlation? Do you think a linear relationship is appropriate to describe the dependency that link these two variables?
6. Evaluate the empirical correlation coefficient between velocity and distance. Briefly comment the value you get.
7. Although here we have only few data points, use `createDataPartition()` from `caret` to create a train and a test set with an 80%/20% split.
8. The next step is to fit a simple linear regression model to the data, but in this case the physical nature of the data requires a model without intercept because if distance is zero so is relative speed. So the model to be fitted to these data is

$$\text{velocity} = \beta_1 \cdot \text{distance} + \varepsilon,$$

   This is essentially what astronomers call Hubble's Law and β_1 is known as *Hubble's constant*; $1/\beta_1$ gives an approximate age of the universe. Use the function `lm()` on the train set to fit this model imposing a zero intercept using `-1` in the model formula.
9. Produce again the scatterplot of velocity vs distance and use `abline()` to add the regression line to this plot. Use `print()`, `summary()` and `plot()` on the `lm` object to explore the fitted model. What are the least squares estimates? And their standard errors? Is there evidence of lack of fit? Comment the R^2 value.
10. Predict on train and test and evaluate the corresponding RMSE. Briefly comment the results.
11. Now we can use the estimated value of β_1 to find an approximate value for the age of the universe. The Hubble constant itself has units of km $\times$ sec^{-1} $\times$ Mpc^{-1}. A mega-parsec (Mpc) is 3.09 $\times$ 1019 km, so we need to divide the estimated value of β_1 by this amount in order to obtain Hubble's constant with units of sec^{-1}. The approximate age of the universe in seconds will then be the inverse of this calculation, and it is equal to...
12. Fit a quadratic regression model without intercept, i.e, a model of the form

$$\text{velocity} = \beta_1 \times \text{distance} + \beta_2 \times \text{distance}^2.$$

Plot the fitted curve and the simple linear regression fit on a scatterplot of the data. Which model do you consider most sensible considering the nature of the data? (Remark: The "quadratic model" here is still regarded as a linear regression model since the term linear relates to the parameters of the model not to the powers of the explanatory variable!).
