SDA Group Submission Assignment 7

CD

For your group submission, upload one .pdf in Canvas ☺

If you follow all installation instructions given in Canvas and still have big issues with R Markdown installation, you can exceptionally submit a .pdf generated in a different way (e.g., by copying code 'manually' into a text editor), which should still include all relevant explanations, code snippets and graphics.

Always remember that there are several ways to code something, and that the useful `R` functions provided for each task are just some suggestions. Finally, the symbol ϟ indicates something generally worthy of attention. Have fun!

Exercise 1

a. Using the data set `mammals` in the `R` package `openintro`, regress `brain_wt` (brain weight of the mammal in g) on `body_wt` (total body weight of the mammal in kg) and `gestation` (gestation time in days). Display the model summary and plot `R`'s default four diagnostic plots in a 2×2 grid.

   Hint: using `plot` with an object of type `lm` as first argument will plot four default diagnostic plots. You can check the help page `?"plot.lm"` for more information about the available diagnostic plots and which ones are plotted by default.

b. Comment on the diagnostic plots: is the linearity assumption satisfied? Are the residuals normally distributed? Does the variance of error terms seem constant? Justify your answers. Two points have Cook's distance larger than 1 and can thus be identified as influence points: to which mammals do they correspond (note that the first column of `mammals` is `species`[1])?

   Hint: the `Residual vs Leverage` plot has dashed lines (contours) corresponding by default to values of Cook's distance 0.5 and 1 (if no such lines are seen, the Cook's distance of all points in the model is well below 0.5). ϟ `R` by default displays the indices of the three 'most extreme' observations, no matter how extreme (or not) they really are, thus the highlighted points are not influence points by default but need further investigation (e.g., one can check their position with respect to Cook's contour lines). If still in doubt, you can use `cooks.distance(...)` on an object of type `lm` to see the values of Cook's distance for all observations, while a graphical summary is given by using `plot(..., which = 4)` (a `plot.lm` diagnostic plot which is *not* plotted by default) on an object of type `lm`.

c. Now, perform simple linear regressions: regress `brain_wt` on `body_wt` only and then on `gestation` only using `R` (note: 'only' refers to the amount of predictors, both models should be fitted *with* intercept like all models in this assignment). Plot two scatterplots side-by-side describing the relationship between `brain_wt` and `body_wt` resp. between `brain_wt` and `gestation`, and add to each the corresponding regression line you just estimated estimated. Comment on the plots.

   Hint: `brain_wt` should be on the y-axis in each plot, so that you can use `abline` with an appropriate `lm` object as first argument to plot the regression line.

d. Regress `log(brain_wt)` on `log(body_wt)` and `log(gestation)`. Then, redo part a. and b. of this exercise for this model. What changed?

[1] Column `species` is of class `factor`, thus it has so-called `levels`. Subsetting/indexing a factor will typically display, below the extracted values, a summary of the levels: you can ignore that and focus on the values extracted.

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The code, summary and graphical output required by a.
- Your answers for b.
- Your code, graphical output and comment for c.
- The code, summary, graphical output and answers required by d.

Useful `R` functions

`lm`, `plot`, `abline`, `log`, …

Exercise 2

a. Create synthetic data by running the following lines in `R`:

```
set.seed(123)
x1 <- runif(100, min = 0, max = 10)
x2 <- runif(100, min = 0, max = 5)
y <- 7 + 0.8 * x1 + 0.4 * x2 + rnorm(100, mean = 0, sd = 1.5)
```

- As you can see, `y` is a function of `x1` and `x2`. Write down the corresponding linear model: what are the values β_0, β_1 and β_2?
- Estimate the model in `R` and compare coefficient estimates to their true values.
- Do you reject the null hypothesis $H_0 : \beta_1 = 0$? And what about $H_0 : \beta_2 = 0$ (significance level $\alpha = 5\%$)? Justify your answers.

b. Now, create other synthetic data by running:

```
set.seed(123)
x1 <- runif(100, min = 0, max = 10)
x2 <- 0.5 * x1 + rnorm(100, sd = 0.1)
y <- 7 + 0.8 * x1 + 0.4 * x2 + rnorm(100, mean = 0, sd = 1.5)
```

- As you can see, `y` again is a function of `x1` and `x2`. Write down the corresponding linear model: what are the values β_0, β_1 and β_2?
- Estimate the model in `R` and compare coefficient estimates to their true values.
- Do you reject the null hypothesis $H_0 : \beta_1 = 0$? And what about $H_0 : \beta_2 = 0$ (significance level $\alpha = 5\%$)? Justify your answers.
- What has changed compared to part a. and which issue does it create?

c. The VIF for a predictor X_j is given by $\text{VIF}_j = \frac{1}{1-\mathcal{R}_j^2}$, where $\mathcal{R}_j^2$ denotes the determination coefficient resulting from a regression of X_j on all of the other predictors in the model. Using the data you created in b., compute VIF_1 and VIF_2 and interpret the results. Finally, give a thorough explanation of why VIF_1 and VIF_2 are equal.

Hint: how can $\mathcal{R}^2$ be computed for a simple linear regression, without having to fit the linear model first?

Deliverables

- For all subtasks, to aid in correction: all relevant code
- The answers, code and comments required by a. and b.
- The required calculation, interpretation and explanation for c.

Useful `R` functions

`set.seed`, `runif`, `rnorm`, `lm`, `summary`, …