Statistics and R short course

Marc Henrion

02 December 2020

Session 5 - Practical

Exercise 1

Using the `adolescent_small.csv` data from Session 4, fit a linear model regressing weight (variable `a104wt`) on age (variable `a12age`).

Test if the regression coefficient of age $\beta_{age} = 0$.

Note:

- deviance = sum of squares
- residual = error

Exercise 2

Repeat the ANOVA test from yesterday as a linear regression model. Do you get the same results?

Note: `glm` will have the word *deviance* in its output. This is just another word for sum of squares.

Recall from yesterday:

```
aov(cd41~as.factor(hosp),data=Tbreg)
```

Exercise 3

Using the `adolescent_small.csv` data, fit the following GLM model:

Weight `a104wt` as a function of

- age `a12age`
- height `a103ht`
- hiv `hiv`
- sex `a13sex`

Produce:

- a residuals vs. fitted values graph
- histogram of the residuals
- a QQ plot.

Exercise 4

You are given the following data:

$$\mathbf{x} = (-6, -6, -4, -1, 0.5, 2, 8, 8, 11, 11.5)^T$$

$$\mathbf{y} = (-3.7, -4.3, -3.9, -4.6, 0.5, -6.9, 10.2, 16.1, 6, 19.5)^T$$

a. Fit a linear regression model to these data and show the model output.

b. Describe the resulting regression line:

   - What is the relationship between variables X and Y?
   - How much (on average) does Y change when X changes by 1?
   - What value does Y take (on average) when $X = 0$?

c. Compute the coefficient of determination R^2, the adjusted R^2, the likelihood and the AIC. Which of these tell you how good your model fits the data?

d. Compute the residuals $r_i = y_i - \hat{y}_i$ and do a normal distribution QQ plot.

e. What other diagnostic check(s) could you do? Do this and explain whether you think this is a good model.

f. Re-fit the model, but now including a term for X^2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. Check and discuss the resulting model and compare it to the previous one. Which model would you recommend for this dataset?

Exercise 5

Download (from GitHub) and load the dataset `cuse.csv`.

This is a dataset on contraceptive use. `using`, `notUsing` lists how many people in each group implied by combinations of `age`, `education`, `wantsMore` are currently using contraceptives. `age`, `education` are self-explanatory. `wantsMore` lists whether individuals want more children or not.

Model the binary variable specified by the 2 columns `using`, `notUsing` in terms of `age`, `education`, `wantsMore`.

- Discuss your results.
- What can you say about the deviance? Does it look like this is a good model?
- What happens if you include an interaction term between the age variable and the desire for more children variable?