FORE224/STAT202 Assignment 9

Non-linear transformations of variables (transformations for model fit)

Due at 12 noon, 20 October 2021

As usual, create a new project in its own folder for this assignment.

[1] Load the `tidyverse` library.

Download the file lizards.csv from Learn and save it in your project folder. The file contains data for 74 lizards. Three variables are recorded:

- the continuous variable mass_g, which is the lizard's mass in grams.
- the continuous variable SVL_mm, which is the lizard's snout-vent length in millimetres
- the coded category type, which will not be used in this analysis.

You will use this data to create some models for lizard mass.

Read the dataset from lizards.csv.

Use `set.seed` to set the random seed to your student ID number and make a subset called my_lizards of 70 cases from this lizards data using `sample_n` as in previous labs. Use this subset of the data for the rest of the assignment.

[2] Use a `ggplot` to plot mass_g (y-axis) against SVL_mm (x-axis). `ggplot2` is part of `tidyverse` so you do not need to load it separately.

[3] Create a linear model m1 for mass_g as a function of SVL_mm. Show the code used to create the model and the model summary (using `summary`) in your report.

Create the regression diagnostic plots for model m1 (at least the usual 4 plots and any others you want to add). Include the plots in your report. Comment on what the diagnostic plots show (about 80 words in total).

[4] Create a log-linear model m2 with response log(mass_g) and regressor SVL_mm. Use natural log (`log` in R) Show the code used to create the model and the model summary in your report.

Create the regression diagnostic plots for model m2 (at least the usual 4 plots and any others you want to add). Include the plots in your report. Comment on these diagnostic plots in relation to the diagnostic plots for model m1 (60-70 words in total).

[5] Use a suitable approximation to interpret the coefficient of the SVL_mm regressor in model m2 in the context of the relationship between mass and SVL (20-30 words). We did not do an example like this in class but the appendix for lecture 29 covers it :)

[6] Create a log-log model m3 for response log(mass_g) and regressor log(SVL_mm), again using natural log. Show the code used to create the model and the model summary in your report.

Create the regression diagnostic plots for model m3 (at least the usual 4 plots and any others you want to add). Include the plots in your report. Comment on these diagnostic plots in relation to the diagnostic plots for model m2 (about 60 words in total).

[7] Use a suitable approximation to interpret the coefficient of the log(SVL_mm) regressor in model m3 in the context of the relationship between mass and SVL (20-30 words).

[8] Create a quadratic model m4 for response mass_g with predictor SVL_mm (ie, model mass_g as a quadratic function of SVL_mm). Show the code used to create the model and the model summary in your report.

Create the regression diagnostic plots for model m4 (at least the usual 4 plots and any others you want to add). Include the plots in your report. Comment on these diagnostic plots in relation to the diagnostic plots for model m3 (about 70-80 words in total).

[9] Interpret the intercept and the coefficient of the SVL_mm regressor in model m4 in relation to the quadratic curve relating mass and SVL given by the model. How meaningful are these coefficients in context? (About 50-60 words in total.)

[10] Add a variable SVL_mm_c for the centered SVL_mm to the data. Show the code to create this variable in your report.

Create a quadratic model m5 for response mass_g using predictor SVL_mm_c (ie, model mass_g as a quadratic function of the centered SVL variable SVL_mm_c). Show the code used to create the model and the model summary in your report.

Interpret the intercept and the coefficient of the SVL_mm_c regressor in model m5 in relation to the quadratic curve relating mass and SVL given by the model. Explain why these coefficients more meaningful than those in model m4. (About 80-100 words in total.)

[11] Which are the only two models out of the models m1, m2, m3, m4, m5 created above that could be compared using a nested model F-test with `anova`? Briefly explain your answer in 50-70 words.

[12] Which model do you prefer? Very briefly explain your answer (about 30 words).

Optional. If you want to you can try plotting the regression models for m1, m2, m3 and m4 on a scatter plot of the data. If you have created and named the models correctly the following code should do this for you and add a suitable legend.

```
my_lizards %>% ggplot(aes(x = SVL_mm, y = mass_g)) +
  geom_point() +
  geom_function(
    fun = function(x) coef(m1)[[1]] + coef(m1)[[2]] * x,
    aes(color = "blue"), size = 1
  ) +
  # alternative for linear is geom_smooth(method = "lm", se = FALSE, aes(color =
"blue") ) +
  geom_function(
    fun = function(x) exp(coef(m2)[[1]]) * exp(x * coef(m2)[[2]]),
    aes(color = "green"), size = 1
  ) +
  geom_function(
    fun = function(x) exp(coef(m3)[[1]]) * x^coef(m3)[[2]],
    aes(color = "orange"), size = 1
  ) +
  geom_function(
    fun = function(x) coef(m4)[[1]] + coef(m4)[[2]] * x + coef(m4)[[3]] * x^2,
    aes(color = "red"), size = 1
  ) +
  labs(x = "SVL (mm)", y = "mass (g)") +
  # hack the legend!
  scale_color_identity(
    name = "Model",
```

```
    breaks = c("blue", "green", "orange", "red"),
    labels = c("Linear", "Log-linear", "Log-log", "Quadratic"),
    guide = "legend"
  ) +
  theme_bw()
```

Include all the answers and code used to obtain them in a Word document and submit it through Learn. Remember to use your own words when answering the questions.