MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

SGTA Week 10: Polynomial and Multiple Regression 2023

Part 1

Question 1

Ashton, Burke, and Layne (2007) measured the carapace length (in mm) of 18 female gopher tortoises (Gopherus polyphemus) in Okeeheelee County Park, Florida. They were then x-rayed to determine the number of eggs in each. The data are contained in the file **turtles.csv**.

a. Fit a simple regression model for eggs versus carapace and validate the model through the usual diagnostic checks and comment.

b. Apply polynomial fits of order 2 and order 3 (quadratic and cubic)

c. Assess the linear, quadratic and cubic based fits using a Sequential Sum of Squares (**Hint**: A sequential ANOVA is necessary).

d. Identify the correct model and fit this, and check assumptions.

e. Produce a scatterplot of the data with the fitted line/curve of the best model chosen in d.

Question 2

Data from a sex discrimination case (see Ramsey and Schafer 2012) are contained in the file **sexdiscrim.txt**. These are data for male and female clerical workers at a bank hired between 1965 and 1975. Variables are

BSAL	Beginning salary
SEX	Male (0) or Female (1)
AGE	Age in months
EDUC	Education (in years)
EXPER	Previous experience (in months)

The research question is whether females and males had different starting salaries after correcting for age, education and previous experience.

a. Produce a matrix scatterplot with smoother, colour code the observations by `SEX`. What effect does sex have on Beginning Salary? What other relationships are obvious? **Hint:** Notice that you may have to convert the `SEX` variable into a factor vector for some code to work.

b. Fit a regression model for Beginning salary with predictors sex, age, education and experience, and validate the model. Use `BSAL` as Response, `AGE EDUC EXPER` as continuous predictor and `SEX` as Categorical Predictor (intercept adjustment).

c. Log transform Beginning Salary i.e. create a column `lnbsal` inside your dataframe. Repeat the analysis with `lnbsal` as the response.

d. Comment on any problems with the assumptions, and possible solutions, but don't fit further models.

e. Which of the two models is the better model? Explain.

f. Remove any predictors which are not significant at the 0.2 level for both models, starting with the least significant predictor.

g. Interpret the effect of sex for each of the models together with a 95% confidence interval.

Part 2

Question 1

`RMarkdown` files can include `R` language as inline `R` code or `R` code chunks.

- inline code means a single output to appear (within a sentence).
- code chunk means larger computations that are one or more R code input statements.

For example, the codes below utilise both to calculate the area of a circle of radius 5:

````
```{r}
x = 5
```
For a circle with the radius `r x` , its area is `r pi * x^2`.
````

R inline code uses a single pair of back ticks and r, where back ticks are the key in the top left corner of a standard US qwerty keyboard. R code chunks are fenced by ```.

You can use a shortcut for inserting an `R` code chunk:

- Within an .Rmd, press
  - Mac: `Cmd + Opt + i`
  - PC: `Ctrl + Alt + i`

Now let's create a `RMarkdown` file with default output being `PDF` and put it in the same directory of our `sexdiscrim.txt` file. Then:

a. Check the working directory of the `RMarkdown` session that just popped up.
   - Hint: you could create a `R` code chunk in the `RMarkdown` document and then include `getwd()` in this chunk.

b. Import the `sexdiscrim.txt` file to this `RMarkdown` session and print out the first few rows with `head()`.
   - Hint: you could create a `R` code chunk in the `RMarkdown` document and then apply `read.table()` in this chunk.

Question 2

`RMarkdown` files can also include and print out mathematical equations. For example, the codes below gives the roots of quadratic polynomials:

```
$$
x  = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}
$$
```

- Inline formulas are surrounded by dollar signs: `$1 + 1 = 2$`.
- To display equations in their own line

Recall that we have imported `sexdiscrim.txt` to our previous `RMarkdown` session. Now:

a. Using the `R` code chunk, fit a regression model using `BSAL` as Response, `AGE`, `EDUC` and `EXPER` as continuous predictor and `SEX` as Categorical Predictor. Print out the summary of this regression to the `PDF` file.

b. Using inline `R` codes, generate the regression equation where the estimated regression coefficients are automatically extracted from the summary in the previous sub-question. Then print this regression equation to the `PDF` file.

To learn more about how to typeset mathematics in RMarkdown, here are some resources to get you started.

- Mathematics in R Markdown: https://rmd4sci.njtierney.com/math.html.

References

Ashton, Kyle G, Russell L Burke, and James N Layne. 2007. "Geographic Variation in Body and Clutch Size of Gopher Tortoises." *Copeia* 2007 (2): 355–63.

Ramsey, Fred, and Daniel Schafer. 2012. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Cengage Learning.