Week 8 - Homework

STAT 420, Summer 2017, Dalpiaz

Directions

- Be sure to remove this section if you use this .Rmd file as a template.
- You may leave the questions in your final document.

Exercise 1 (Writing Functions)

(a) Write a function named diagnostics that takes as input the arguments:

- model, an object of class lm(), that is a model fit vialm()‘
- pcol, for controlling point colors in plots, with a default value of black
- lcol, for controlling line colors in plots, with a default value of white
- alpha, the significance level of any test that will be performed inside the function, with a default value of 0.05
- plotit, a logical value for controlling display of plots with default value TRUE
- testit, a logical value for controlling outputting the results of tests with default value TRUE

The function should output:

- A list with two elements when testit is TRUE:
  - p_val, the p-value for the Shapiro-Wilk test for assesing normality
  - decision, the decision made when performing the Shapiro-Wilk test using the alpha value input to the function. “Reject” if the null hypothesis is rejected, otherwise “Fail to Reject”.
- Two plots, side-by-side, when plotit is TRUE:
  - A fitted versus residuals plot that adds a horizontal line at $y = 0$, and labels the x-axis “Fitted” and the y-axis “Residuals”. The points and line should be colored according to the input arguments. Give the plot a title.
  - A Normal Q-Q plot of the residuals that adds the appropriate line using qqline(). The points and line should be colored according to the input arguments. Be sure the plot has a title.

(b) Run the following code.

```r
set.seed(42)
data1 = data.frame(x = runif(n = 20, min = 0, max = 10),
                   y = rep(x = 0, times = 20))
data1$y = with(data1, 5 + 2 * x + rnorm(n = 20))
fit1 = lm(y ~ x, data = data1)

data2 = data.frame(x = runif(n = 30, min = 0, max = 10),
                   y = rep(x = 0, times = 30))
data2$y = with(data2, 2 + 1 * x + rexp(n = 30))
fit2 = lm(y ~ x, data = data2)

data3 = data.frame(x = runif(n = 40, min = 0, max = 10),
                   y = rep(x = 0, times = 40))
data3$y = with(data3, 2 + 1 * x + rnorm(n = 40, sd = x))
fit3 = lm(y ~ x, data = data3)

diagnostics(fit1, plotit = FALSE)$p_val
diagnostics(fit1, testit = FALSE, pcol = "darkorange", lcol = "dodgerblue")

diagnostics(fit2, plotit = FALSE)$decision
diagnostics(fit2, testit = FALSE, pcol = "grey", lcol = "green")

diagnostics(fit3)
```

Exercise 2 (Swiss Fertility Data)

For this exercise, we will use the `swiss` data, which can be found in the `faraway` package. After loading the `faraway` package, use `?swiss` to learn about this dataset.

```r
library(faraway)
```

(a) Fit an additive multiple regression model with `Fertility` as the response and the remaining variables in the `swiss` dataset as predictors. Report the R^2 for this model.

(b) Check the constant variance assumption for this model. Do you feel it has been violated? Justify your answer.

(c) Check the normality assumption for this model. Do you feel it has been violated? Justify your answer.

(d) Check for any high leverage observations. Report any observations you determine to have high leverage.

(e) Check for any influential observations. Report any observations you determine to be influential.

(f) Refit the additive multiple regression model without any points you identified as influential. Compare the coefficients of this fitted model to the previously fitted model.

(g) Create a data frame that stores the observations that were "removed" because they were influential. Use the two models you have fit to make predictions with these observations. Comment on the difference between these two sets of predictions.

Exercise 3 (Why Bother?)

Why do we care about violations of assumptions? One key reason is that the distributions of the parameters that we have used are all reliant on these assumptions. When the assumptions are violated, the distributional results are not correct, so our tests are garbage. **Garbage In, Garbage Out!**

Consider the following setup that we will use for the remainder of the exercise. We choose a sample size of 50.

```r
n = 100
set.seed(42)
x_1 = runif(n, -2, 2)
x_2 = runif(n, 0, 5)
```

Consider the model,

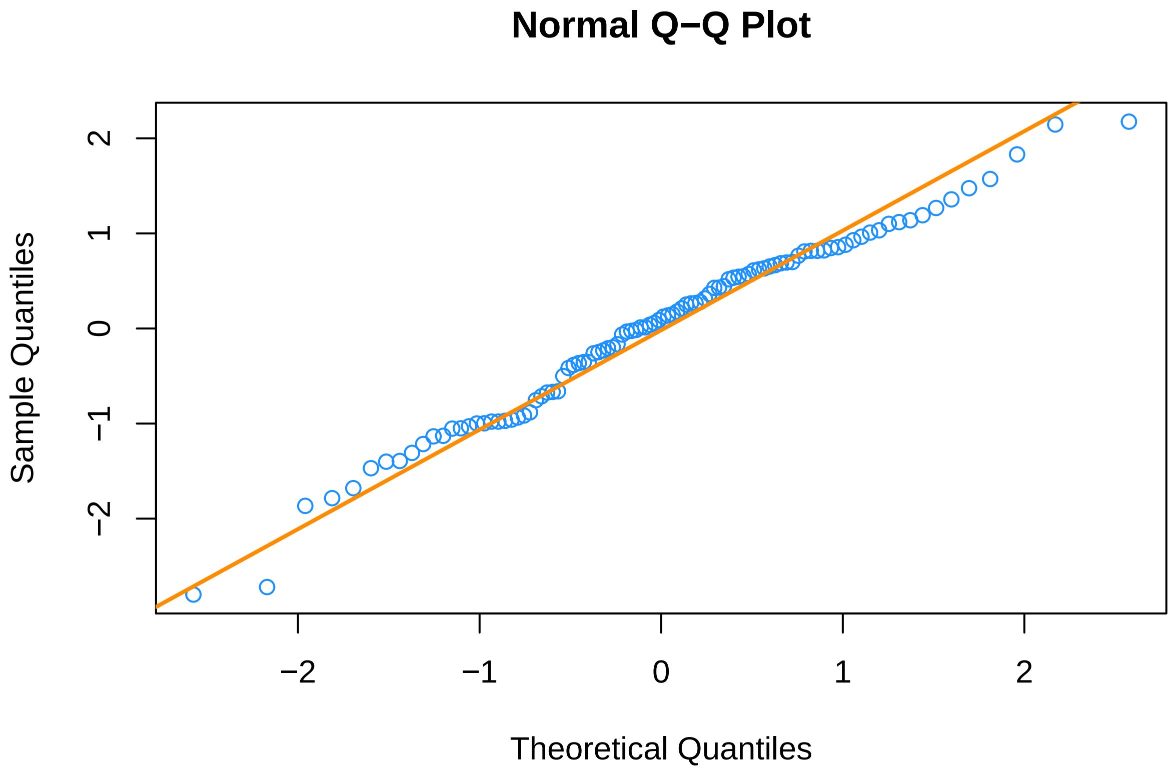
$$Y = 5 + 0x_1 + 1x_2 + \epsilon.$$

That is,

- $\beta_0 = 5$
- $\beta_1 = 0$
- $\beta_2 = 1$

We now simulate y_1 in a manner that does not violate any assumptions, which we will verify. In this case $\epsilon \sim N(0, 1)$.

```
set.seed(420)
y_1 = 5 + 0 * x_1 + 1 * x_2 + rnorm(n = n, mean = 0, sd = 1)
fit_1 = lm(y_1 ~ x_1 + x_2)
qqnorm(resid(fit_1), col = "dodgerblue")
qqline(resid(fit_1), col = "darkorange", lwd = 2)
```

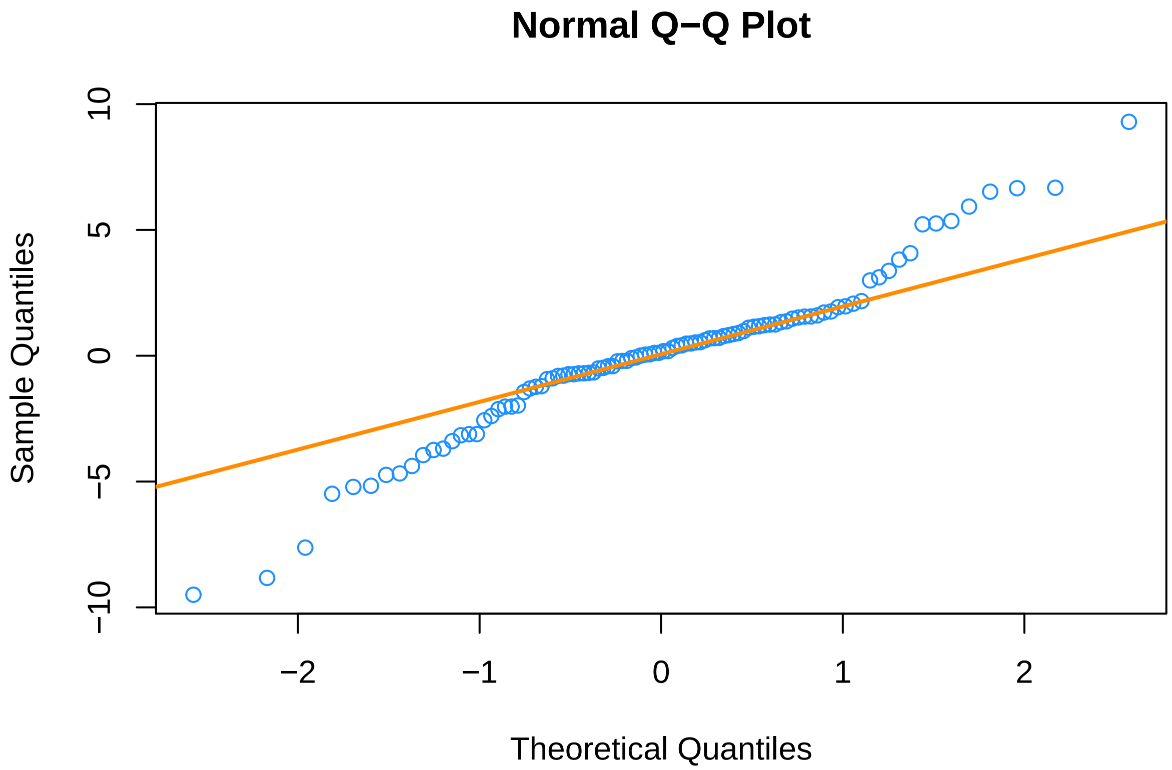


```
shapiro.test(resid(fit_1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_1)
## W = 0.98, p-value = 0.2
```

Then, we simulate y_2 in a manner that **does** violate assumptions, which we again verify. In this case $\epsilon \sim N(0, \sigma = |x_2|)$.

```
set.seed(42)
y_2 = 5 + 0 * x_1 + 1 * x_2  + rnorm(n = n, mean = 0, sd = abs(x_2))
fit_2 = lm(y_2 ~ x_1 + x_2)
qqnorm(resid(fit_2), col = "dodgerblue")
qqline(resid(fit_2), col = "darkorange", lwd = 2)
```



```
shapiro.test(resid(fit_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_2)
## W = 0.96, p-value = 0.002
```

(a) Use the following code after changing `birthday` to your birthday.

```
num_sims = 2500
p_val_1 = rep(0, num_sims)
p_val_2 = rep(0, num_sims)
birthday = 19081014
set.seed(birthday)
```

Repeat the above process of generating `y_1` and `y_2` as defined above, and fit models with each as the response `2500` times. Each time, store the p-value for testing,

$$\beta_1 = 0,$$

using both models, in the appropriate variables defined above. (You do not need to use a data frame as we have in the past. Although, feel free to modify the code to instead use a data frame.)

(b) What proportion of the `p_val_1` values are less than 0.01? Less than 0.05? Less than 0.10? What proportion of the `p_val_2` values are less than 0.01? Less than 0.05? Less than 0.10? Arrange your results in a table. Briefly explain these results.

Exercise 4 (TV Is Healthy?)

For this exercise, we will use the `tvdoctor` data, which can be found in the `faraway` package. After loading the `faraway` package, use `?tvdoctor` to learn about this dataset.

```
library(faraway)
```

(a) Fit a simple linear regression with `life` as the response and `tv` as the predictor. Plot a scatterplot and add the fitted line. Check the assumptions of this model.

(b) Fit higher order polynomial models of degree 3, 5, and 7. For each, plot a fitted versus residuals plot and comment on the constant variance assumption. Based on those plots, which of these three models do you think are acceptable? Use a statistical test(s) to compare the models you just chose. Based on the test, which is preferred? Check the normality assumption of this model. Identify any influential observations of this model.

Exercise 5 (Brains)

The data set `mammals` from the `MASS` package contains the average body weight in kilograms (x) and the average brain weight in grams (y) for 62 species of land mammals. Use `?mammals` to learn more.

```
library(MASS)
```

(a) Plot average brain weight (y) versus average body weight (x).

(b) Fit a linear model with `brain` as the response and `body` as the predictor. Test for significance of regression. Do you think this is an appropriate model?

(c) Since the body weights do range over more than one order of magnitude and are strictly positive, we will use log(body weight) as our *predictor*, with no further justification. (Recall, *the log rule*: if the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.) Use the Box-Cox method to verify that log(brain weight) is then a "recommended" transformation of the *response* variable. That is, verify that $\lambda = 0$ is among the "recommended" values of λ when considering,

$$g_\lambda(y) = \beta_0 + \beta_1 \log(\text{body weight}) + \epsilon$$

Include the relevant plot in your results, using an appropriate zoom onto the relevant values.

(d) Fit the model justified in part **(c)**. That is, fit a model with log(brain weight) as the response and log(body weight) as a predictor. Plot log(brain weight) versus log(body weight) and add the regression line to the plot. Does a linear relationship seem to be appropriate here?

(e) Use a Q-Q plot to check the normality of the errors for the model fit in part **(d)**.

(f) Use the model from part **(d)** to predict the brain weight of a male Snorlax which has a body weight of 1014.1 pounds. (A Snorlax would be a mammal, right?) Construct a 90% prediction interval.