STAT3032 Homework 3

Problem I:

This problem is based on the data file and context of an exercise question in Chapter 3.

The data file `airfares.txt` on the book website gives the one-way airfare (in US dollars) and distance (in miles) from city A to 17 other cities in the US. Interest centers on modeling airfare as a function of distance:

$$Fare \;=\; \beta_0 + \beta_1\, Distance \;+\; e$$

(a). Fit a simple linear regression model and interpret the estimated intercept in the context of the question.

(b). Draw two residual plots. The first one has residual as the vertical axis and the predictor variable (Distance) as the horizontal axis. The second one has residual as the vertical axis and fitted value ($\hat{Fare}$) as the horizontal axis. Compare and contrast these two residual plots. What is similar? What is different?

(c). Which observation has the largest leverage? Is this a leverage point according to the rule on P56 of the textbook? Hint: Use R function `hatvalues(YourModel)` to obtain the leverages.

(d). Which observation has the largest standardized residual? Is this an outlier according to the rule on P60 of the textbook? Hint: Use R function `rstandard(YourModel)` to obtain the standardized residuals.

(e). Which observation has the largest Cook's Distance? Using the rough cutoff suggested by Fox (See P68 of textbook), is the largest value of the Cook's Distance noteworthy?
Hint: Use R function `cooks.distance(YourModel)` to obtain the Cook's Distance.

Problem II:

This problem focuses on calculating the quantities (leverages, standardized residual, Cook's Distance) **by hand**. You may use R as a calculator.

Consider a simple linear regression model with one continuous predictor variable. We have sample size n = 100, RSS = 24.5, SXX = 400, $\bar{X} = 20$. The following table includes the statistics for two of the observations.

	x_i	y_i	$\hat{y}_i$
Observation 11	30.5	5	5.3
Observation 69	26.5	4.5	4.3

(a). Which observation has a larger leverage (h_{ii})? Don't do any calculation yet!

(b). Calculate the leverages (h_{ii}) for the two observations. Are the results in (b) consistent with your answer in (a)?

(c). Calculate the residuals ($\hat{e}_i$) for the two observations.

(d). Calculate the standardized residuals (r_i) for the two observations.

(e). Calculate the Cook's Distance (D_i) for the two observations.

Problem III:

Download the dataset brainsPartial.csv from Moodle. The data gives the average body weight in kilograms (X) and the average brain weight in grams (Y) for 59 species of mammals. We want to build a model to predict a mammal's brain weight based on the body weight.

(a). Draw a scatter plot to visualize the relationship between the average body weight and average brain weight of the 59 mammals in the dataset. Is the linearity assumption violated?

(b). Fit a regression model and print out the summary output of the refitted model. Can you remove the slope from the model?

(c). Use the **normal QQ plot** to determine if the normality assumption has been violated.

(d). Use the **standardized residual plot** to determine if the constant variance assumption has been violated. Hint: The vertical axis is the standardized residuals. The horizontal axis is the fitted value.

(e). Refit the model using the transformed response variable (log(Y)) and predictor variable (log(X)). "Log" represents natural logarithm. Comment on whether the model based on the transformed data violates the assumptions for linear regression.