Homework 2

Instructions

Due: 1:35pm on Wednesday, September 23rd

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

Based on ISLR Exercise 3.1

Write out the null hypotheses to which the p-values given in Table 3.4 (p. 75 ISLR) correspond. Explain what conclusions you can draw based on these particular p-values.

Problem 2

Based on ISLR Exercise 3.4

I collect a set of data (n = 100 observations)containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0+\beta_1 X+\beta_2 X^2+\beta_3 X^3+\epsilon$.

(a) Suppose the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. COnsider the training residual sum of squares RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that teh true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

Problem 3

Based on ISLR Exercises 3.5 and 3.6

(a) Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i\hat{\beta}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{j=1}^{n} x_j^2\right)$$

Show that we can write

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j$$

for some constants a_j. Give the explicit formula for a_j.

(b) Use equation 3.4 in the text to show that for SLR, the least squares line always passes through the point $(\bar{x}, \bar{y})$

Applied

Problem 4

1000 large seismic events around Fiji have been collected in a data set called `quakes` that is built into R. You can learn more about it with the following commands:

Earthquake detection Included in the data set is a column recording the number of stations that detected each earthquake. This refers to a global network of seismographs and it stands to reason that the larger the quake, the more widely it will be detected.

(a) Create a plot of the relationship between `stations` and `magnitude`. How would you characterize the relationship? (If you see overplotting, you may want to add `jitter` to your points or make them transparent by playing with the `alpha` value.)

(b) If here was actually *no relationship* between the two variables, what would you expect the slope of a linear model to be? What about the intercept?

(c) Fit a linear model called `m1` to the trend and add that line to the plot from exercise 1. Interpret your slope and intercept in the context of the problem. *Hint: the `geom_abline` layer adds a straight line with given slope and intercept to plot*

(d) Using formulas 3.8 and 3.9 on page 66 of ISLR , calculate a 95% confidence interval for the slope of the model that you fit in exercise 3 (you can use R as a calculator assist with arithmetic). Confirm the calculation by applying the `confint` function to your linear model.

(e) How many stations do you predict would be able to detect an earthquake of magnitude 7.0?

(f) Parts (a) - (e) in this problem involve elements of *data description*, *inference*, and/or *prediction*. Which was the dominant goal in each question?

Problem 5

One good way to assess whether your fitted model seems appropriate is to simulate data from it and see if it looks like the data that you observed. We'll do this for the `mag` and `station` data in the `quakes` data set.

(a) To begin, let's assume (perhaps unreasonably) that `mag` is normally distributed. Compute the mean and standard deviation of `mag` and store them as variables `mean_mag`, `sd_mag`.

(b) The `rnorm(n, mean, sd)` function generates `n` observations from a Normal distribution with mean of `mean` and standard deviation of `sd`. Use the values calculated in part (a) to generate 1000 data points for `mag` and store the output as the vectors `sim_mag`

(c) We now need to theorize the functional relationship between `station` and `mag`. Since we fit a linear model previously in Problem (1), we can use that as a starting point. To generate the $\hat{y}$ predicted values from your linear function based on your simulated data in (a), we can define an R function. Replace the line beginning with # in the code chunk below with the formula you found in Problem 1 (i.e. something like `10 - 2x`)

```
f_hat <- function(x){
  # your formula for the linear function goes here
}
```

(d) Generate your predicted values by applying the `f_hat` function you just made to the vector of predictors `sim_mag` and store the result as the vector `pred_stations`.

(e) Now, simulate observed y's by adding random error to each predicted value. Estimate the standard deviation of this error using the observed RSE from your model in Problem 1 and store this value as the variable `obs_rse`. Then generate 1000 independent and Normally distributed errors using `rnorm` (note that errors should have mean of 0).

(f) Create a vector of simulated observed values by adding together your vector of predicted values and the vector of errors, and save the new vector as `sim_stations`.

(g) Create a data frame of simulated data called `quakes_sim` by applying the `data.frame` function to the vectors `sim_mag` and `sim_stations`.

(h) Perform exploratory data analysis on this simulated data set. How does your simulated data compare to the actual observed data? How might you change your simulation to make the data more consistent with the observed data?

Problem 6

Based on ISLR Exercise 3.9

This question uses the `Auto` data set, loaded from the `ISLR` library, as well as the `ggpairs` function from the `GGally` library. Both libraries are loaded by running the code chunk below.

```
library(ISLR)
library(GGally)
```

You can learn more about tbe data set with the following commands:

(a) When the `ggpairs` function is applied to a data frame, it creates a matrix of pairwise scatterplots and correlations for all variables in the data frame (using `ggplot` styling conventions). Use this function to create pairwise scatterplots for all **quantitative** variables in the `Auto` data set. *You may want to adjust the displayed figure dimensions using chunk options (the gear in upper right of chunk)*

(b) Use the `lm` function to fit a MLR model with `mpg` as the response and all other quantitative variables as predictors. Then use the `summary` function to print the results.

(c) Based on your model, does there appear to be a relationship between the predictors and response? Which predictors have statistically significant relationship with the response? What does the coefficient for the `year` variable suggest? Justify your answers.

(d) Create diagnostic plots for the linear regression fit. Comment on any problems you observe. Do the residual plots suggest any unusually large outliers? Do leverage plots suggest any observations with unusually large leverage?

(e) Fit a linear regression model with at least 3 interaction terms of your choice. Do any of these interactions terms appear significant?

(f) Try two different transformations of two different variables. Comment on the effect.