STAT 331: Assignment 1

Due: Wednesday May 22, 2019 by 11:59pm

Derivation Component [30 points]
Answer the questions below in the context of the following simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the ε_i's are assumed to be independent and identically distributed $\mathrm{N}(0, \sigma^2)$ random variables, $i = 1,2,\dots,n$.

1. [5] Compute the maximum likelihood estimates for β_0 and β_1 and compare them to the corresponding least squares estimates. What do you notice?

2. [5] Compute the maximum likelihood estimate for σ and compare this to the corresponding least squares estimate. What do you notice?

3. [5] Let $c_i = (x_i - \bar{x})/s_{xx}$ where $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Show
   a. $\sum_{i=1}^{n} c_i = 0$

   b. $\sum_{i=1}^{n} c_i x_i = 1$

   c. $\sum_{i=1}^{n} c_i^2 = 1/s_{xx}$

4. [5] Using the properties listed in question 3, and recognizing that $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, show

$$E[\hat{\beta}_1] = \beta_1 \text{ and } Var[\hat{\beta}_1] = \sigma^2/s_{xx}.$$

5. [5] Using the results of question 4 and the distributional assumptions associated with the linear model, show

$$E[\hat{\beta}_0] = \beta_0 \text{ and } Var[\hat{\beta}_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right).$$

6. [5] Using the results of questions 4 and 5 and the distributional assumptions associated with the linear model, show

$$E[\hat{\mu}_0] = \mu_0 \text{ and } Var[\hat{\mu}_0] = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)$$

   where $\mu_0 = \beta_0 + \beta_1 x_0$ is the expected response for a particular value x_0 of the explanatory variable.

Computation Component [58 points]

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The dataset titled `bike_share.csv` contains 10,886 hourly records of rental data spanning two years for the Capital Bikeshare program in Washington, D.C. This dataset contains observations from the following variables:

- `season`: 1 = spring, 2 = summer, 3 = fall, 4 = winter
- `weather`: 1 = nice, 2 = cloudy, 3 = rainy, 4 = stormy
- `temp`: outdoor temperature (measured in Fahrenheit)
- `humidity`: relative humidity (as a percentage)
- `windspeed`: wind speeds (measured in miles per hour)
- `count`: the number of bike rentals in a given hourly period

Interest lies in (1) understanding which factors influence bike rental demand, and (2) predicting the number of bike rentals in a given hourly period. Your job in this assignment is to investigate these questions by completing each of the tasks below. Where computation is required, you must perform the calculations using R.

7. $y =$ `count`, $x =$ `temp`

   (a) [4] Construct a scatter plot of `count` versus `temp`, being sure to add a title and appropriately label your axes. Calculate the corresponding correlation coefficient and describe the linear relationship you observe in terms of 'direction' and 'strength'. Note that you may use the `cor()` function in R.

   (b) [3] For the relationship in part (a) calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ and state the equation of the line-of-best-fit. Note that you must use the equations derived in class to perform these calculations. You may, however, use automated functions (such as `lm()` in R) to check your answers.

   (c) [2] Interpret the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (d) [5] To the scatter plot in part (a), add the least squares line-of-best-fit, the 95% confidence interval for this line and the 95% prediction interval for this line. Be sure to include a title, axis labels and a legend.

   (e) [2] Using your results from part (d) predict the number of bike rentals in hours for which the outside temperature is 70 degrees Fahrenheit. Be sure to accompany your point prediction with a 95% prediction interval.

(f) [3] Using the `lm()` function in R, fit a simple linear regression model relating `count` to `temp`. Using the output from this function, formally test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$ and draw a conclusion regarding whether bike rentals appear to be significantly influenced by the outside temperature.

8. $y =$ `count`, $x =$ `humidity`

(a) [4] Construct a scatter plot of `count` versus `humidity`, being sure to add a title and appropriately label your axes. Calculate the corresponding correlation coefficient and describe the linear relationship you observe in terms of 'direction' and 'strength'. Note that you may use the `cor()` function in R.

(b) [3] For the relationship in part (a) calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ and state the equation of the line-of-best-fit. Note that you must use the equations derived in class to perform these calculations. You may, however, use automated functions (such as `lm()` in R) to check your answers.

(c) [2] Interpret the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) [5] To the scatter plot in part (a), add the least squares line-of-best-fit, the 95% confidence interval for this line and the 95% prediction interval for this line. Be sure to include a title, axis labels and a legend.

(e) [2] Using your results from part (d) predict the number of bike rentals in hours for which the humidity is 40%. Be sure to accompany your point prediction with a 95% prediction interval.

(f) [3] Using the `lm()` function in R, fit a simple linear regression model relating `count` to `humidity`. Using the output from this function, formally test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$ and draw a conclusion regarding whether bike rentals appear to be significantly influenced by the relative humidity.

9. $y =$ `count`, $x =$ `windspeed`

(a) [4] Construct a scatter plot of `count` versus `windspeed`, being sure to add a title and appropriately label your axes. Calculate the corresponding correlation coefficient and describe the linear relationship you observe in terms of 'direction' and 'strength'. Note that you may use the `cor()` function in R.

(b) [3] For the relationship in part (a) calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ and state the equation of the line-of-best-fit. Note that you must use the equations derived in class to perform these calculations. You may, however, use automated functions (such as `lm()` in R) to check your answers.

(c) [2] Interpret the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) [5] To the scatter plot in part (a), add the least squares line-of-best-fit, the 95% confidence interval for this line and the 95% prediction interval for this line. Be sure to include a title, axis labels and a legend.

(e) [2] Using your results from part (d) predict the number of bike rentals in hours for which the wind speed is 10 miles per hour. Be sure to accompany your point prediction with a 95% prediction interval.

(f) [3] Using the `lm()` function in R, fit a simple linear regression model relating `count` to `windspeed`. Using the output from this function, formally test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$ and draw a conclusion regarding whether bike rentals appear to be significantly influenced by windspeed.

10. [1] Based on results from questions 7-9, rank the variables `temp`, `humidity` and `windspeed` in terms of the strength of their relationship with bike rentals, from most weakly associated to most strongly associated.

11. [NOT FOR SUBMISSION – JUST DO IT AND THINK ABOUT IT] Using the `lm()` function in R fit a simple linear regression model between `count` and `weather`. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$. Do these interpretations seem practically useful? Yes or No.

Submission

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically via *Crowdmark*. This means that your responses for different questions should be on separate pages.

For the derivation questions, you may upload solutions produced by LaTeX (or some other equation editing software), or handwritten solutions that have been scanned or photographed. Please note that if you scan or photograph your solutions it is your duty to ensure they are readable and of high quality. If the marking team cannot read your solutions, they simply will not be marked.

For the computational questions, I highly recommend you produce your solutions as a nicely formatting .pdf file with R Markdown. R Markdown facilitates the seamless combination of written text, R code and output, and LaTeX equations. Your submission for these questions should include the code, the corresponding output, and the interpretations where appropriate.