SS 3859A/9859A Fall 2019

Assignment 1

Due: Sep 23 by 12:00 p.m.

Total: 30 pt

Note: Please submit your assignment through OWL. Use either MS word or PDF format for your submission.

1. Suppose that we have the following 9 observations for random variable X:
$x = c(12.21, 14.37, 17.18, 11.74, 13.84, 14.26, 15.42, 13.52, 17.97)$. Conduct a t-test for the true mean of X. Specifically, test $H_0 : \mu = 16$ vs $H_1 : \mu < 16$ at significance level $\alpha = 0.05$. Give the test statistic (t value), the p-value, and your conclusion. (2 pt)

2. Consider the SLR model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. We have obtained the following data results from 10 observations (i.e. $n = 10$):

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -2022, \qquad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 102,$$
$$\bar{x} = 5, \qquad \bar{y} = -90$$

(a) Find the LS estimates of β_0 and β_1. (1 pt)

(b) Using the estimates, obtain the fitted value of y at $x = 3$. (1 pt)

(c) Suppose that $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 47.13$. Obtain an unbiased estimate of σ^2. (1 pt)

(d) Construct a 95% confidence interval for $E(Y|x = 3)$. (2 pt)

3. Consider the SLR model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

(a) Show that the fitted regression line ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) passes through the point $(\bar{x}, \bar{y})$. (2 pt)

(b) Show that SST = SSR + SSE. That is, show $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$. (2 pt)

4. For question 4, you will use an imported data. Run the following R code (in blue) and use the $hw1\_data$ data set to answer the questions.

R codes:
hw1_data = read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw1_data1.csv")

The imported data set contains 100 observations with 2 variables: x1 and x2. Include your R codes and output for the following questions.

(a) Count the number of observations whose x1 are greater than 6. (1 pt)

(b) Count the number of observations whose x1 are greater than 6 and x2 equal to H. (1 pt)

(c) Consider a subset A that contains all observations with x2 = H. Compute the mean, median and standard deviation of the x1 values in subset A. (1 pt)

(d) The sample mean of x1 is 4.435. Can we argue that the true mean of x1 differs from 4? Conduct a t-test at significance level $\alpha = 0.05$. Give the test statistic (t-value), p-value and your conclusion. (1 pt)

(e) Consider the statement: "Given that x2 equals to H, the true mean of x1 is larger than 4." Is this statement convincing? Use a t-test ($\alpha = 0.05$) to support your answer. (2 pt)

5. For question 5, you will use a subset of the *cars* data. Run the following R codes and use the *cars*2 data set to answer the questions. Include your R codes and output for the following questions.

R codes:

```
set.seed(50)
idx = sample(nrow(cars),40,replace=FALSE)
cars2 = cars[idx,]
```

(a) Make a scatterplot that shows the relationship between x and Y. From the plot, do you find any relationship between speed and dist? (1 pt)

(b) Assume that there is a linear relationship between x and Y. That is, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Obtain the LS estimates for β_0, β_1 and an unbiased estimate for σ^2. (1 pt)

(c) Using the estimates, calculate the residuals e_4, e_7 and e_{10} (i.e. residuals for the observations at the 4th, 7th and 10th rows of the cars2 data). (1 pt)

(d) Find the residuals whose absolute values are greater than 20. Indicate those residuals in the scatterplot with different a color and shape. (2 pt)

(e) Calculate the sum of the residuals (i.e. $\sum_{i=1}^{n} e_i$). (1 pt)

(f) Report the fitted model. Add the fitted regression line to the current scatterplot. Predict the distance taken to stop when the speed of the car is 17. (1 pt)

(g) State the goodness of fit for the fitted model. What percentage of the variation in the response variable is explained by the fitted model? (1 pt)

(h) Consider the statement: "If someone is driving at 100mph, according to the fitted model, the distance taken to stop will be exactly 370.9615ft." Give a brief (reasonable) criticism of the statement. (2 pt)

(i) Construct a 90% confidence interval for β_1. (2 pt)

(j) Construct a 95% confidence interval for $E(Y|x = 15)$. (1 pt)