Statistics 206

Homework 2

Due : October 9, 2019, In Class

1. Tell true or false of the following statements and provide a brief explanation to your answer.

   (a) Under the same confidence level, the prediction interval of a new observation is always wider than the confidence interval for the corresponding mean response.

   (b) A 95% confidence interval for β_0 based on the observed data is calculated to be $[0.3, 0.5]$. Therefore
   $$P(0.3 \leq \beta_0 \leq 0.5) = 0.95.$$

   (c) In t-tests, how critical values and pvalues should be derived will depend on the form of the alternative hypothesis.

   (d) When estimating the mean response corresponding to X_h, the further X_h is from the sample mean $\overline{X}$, the wider the confidence interval for the mean response tends to be.

2. Under the simple linear regression model, show that the residuals e_i's are uncorrelated with the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, i.e.,
   $$Cov(e_i, \hat{\beta}_0) = 0, \quad Cov(e_i, \hat{\beta}_1) = 0$$
   for $i = 1, \cdots, n$.

3. Under the Normal error model: Show that SSE is independent with the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

4. Under the simple linear regression model, derive $Var(\widehat{Y}_h)$, where
   $$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$
   is the estimator of the mean response $\beta_0 + \beta_1 X_h$.

5. Simulation by R. Please attach your codes. (Hint: use the `help` function if needed)

   (a) Create a sequence of consecutive integers ranging from 1 to 100. Record these in a vector x. (Hint: use the `seq` function)

   (b) Create a new vector w by the formula: $w = 2 + 0.5 * x$.

   (c) Randomly sample 100 numbers from a Normal distribution with mean zero and standard deviation 5. Calculate the sample mean and sample variance and draw a histogram. What do you observe? (Hint: use the `rnorm` function)

(d) Add (element-wise) the numbers created in part (c) to the vector w. Record the new vector as y.

(e) Draw the scatter plot of y versus x.

(f) Estimate the regression coefficients of y on x. Add the fitted regression line to the scatter plot in part (e). What do you observe?

(g) Calculate the residuals and draw a scatter plot of residuals versus x. What do you observe? Derive MSE.

(h) Repeat parts (c) – (g) a couple of times. What do you observe?

(i) **(Optional problem).** Repeat parts (c) – (d) 1000 times. Each time, derive the fitted regression coefficients and MSE and record them. Draw histogram and calculate sample mean and sample variance for each of the three estimators. Summarize your observations.

6. A criminologist studied the relationship between level of education and crime rate. He collected data from 84 medium-sized US counties. Two variables were measured: X – the percentage of individuals having at least a high-school diploma; and Y – the crime rate (crimes reported per $100,000$ residents) in the previous year. A snapshot of the data and a scatter plot are shown here:

County	Crimes/100,000	Percent-of-High-school-graduates
1	8487	74
2	8179	82
3	8362	81
4	8220	81
5	6246	87
6	9100	66
..., ...		

Some summary statistics are also given:

$$\sum_{i=1}^{84} X_i = 6602, \quad \sum_{i=1}^{84} Y_i = 597341, \quad \sum_{i=1}^{84} X_i^2 = 522098, \quad \sum_{i=1}^{84} Y_i^2 = 4796548849, \quad \sum_{i=1}^{84} X_i Y_i = 46400230.$$

Perform analysis under the simple linear regression model.

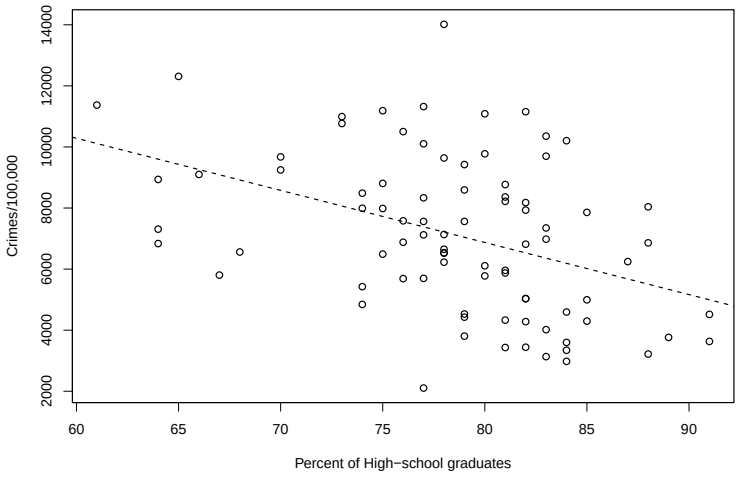
(a) Based on the scatter plot, comment on the relationship between percentage of high school graduates and crime rate.

(b) Calculate the least squares estimators: $\hat{\beta}_0, \quad \hat{\beta}_1$. Write down the fitted regression line. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$.

(c) Calculate error sum of squares (SSE) and mean squared error (MSE). What is the degrees of freedom of SSE?

(d) Calculate the standard errors for the LS estimators $\hat{\beta}_0, \quad \hat{\beta}_1$, respectively.

Figure 1: Scatter plot of Crime rate vs. Percentage of high school graduates



(e) Test whether or not there is a linear association between crime rate and percentage of high school graduates at significance level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

(f) What is an unbiased estimator for β_0? Construct a 99% confidence interval for β_0. Interpret your confidence interval.

(g) Construct a 95% confidence interval for the mean crime rate for counties with percentage of high school graduates being 85. Interpret your confidence interval.

(h) County A has a high-school graduates percentage being 85. What is the predicted crime rate of county A? Construct a 95% prediction interval for the crime rate. Compare this interval with the one from part (g), what do you find?

(i) Would additional assumption be needed in order to conduct parts (e)-(h)? If so, please state what it is.

7. **Optional Problem.** Under the Normal error model, show that

(a) LS estimators $\hat{\beta}_0, \hat{\beta}_1$ are maximum likelihood estimators (MLE) of β_0, β_1, respectively.

(b) The MLE of σ^2 is SSE/n. Is MLE of σ^2 unbiased?