Statistics 206

Homework 1

Due : October 7, 2015, In Class

1. Please tell true or false of the following statements with regard to simple linear regression.

   (a) The least squares line always passes the center of the data $(\overline{X}, \overline{Y})$.

   (b) The true regression line fits the data the best.

   (c) If $\overline{X} = 0, \ \ \overline{Y} = 0$, then $\hat{\beta}_0 = 0$ no matter what is $\hat{\beta}_1$.

   (d) The larger the range of X_is, the smaller the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ tend to be.

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

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

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

   (h) 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.

   (i) If the correlation coefficient r_{XY} between X_is and Y_is is such that $|r_{XY}| < 1$, then we will observe regression effect.

2. Derive the following.

   (a) $\sum_{i=1}^n (X_i - \overline{X}) = 0, \quad \sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - \overline{X}) X_i = \sum_{i=1}^n X_i^2 - n(\overline{X})^2$.

   (b) $\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^n (X_i - \overline{X}) Y_i = \sum_{i=1}^n X_i Y_i - n\overline{X}\ \overline{Y}$.

   (c) Derive the LS estimators for simple linear regression model.

3. Properties of the residuals under simple linear regression model. Recall that
   $$\begin{aligned} e_i &= Y_i - \widehat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \cdots n. \\ &= (Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X}). \end{aligned}$$

   Show that

   (a) $\sum_{i=1}^n e_i = 0$.

   (b) $\sum_{i=1}^n X_i e_i = 0$.

(c) $\sum_{i=1}^{n} \widehat{Y}_i e_i = 0$.

4. Under the simple linear regression model, show the following.

   (a) The LS estimator $\hat{\beta}_0$ is an unbiased estimator of β_0 and derive the formula for its variance.

   (b) $\overline{Y}$ and $\hat{\beta}_1$ are uncorrelated. (Hint: Write them as linear combinations of Y_is.)

5. Under the simple linear regression model:

   (a) Show that
   $$SSE = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2.$$

   (b) Derive $E(\hat{\beta}_1^2)$.

   (c) Show that
   $$E((Y_i - \bar{Y})^2) = \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2),$$
   where $\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$.

   (d) Show that $E(SSE) = (n-2)\sigma^2$. (Hint: What is $E(\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2)$?)

6. Under the simple linear regression model:

   (a) Show that the regression sum of squares
   $$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2.$$

   (b) Derive $E(SSR)$.

   (c) Derive the decomposition of total variation, i.e., show
   $$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2.$$

   Hints: Recall the partition of the total deviation:
   $$Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y}).$$

   (i) Take square of both sides and then sum over i. (ii) Recall
   $$\widehat{Y}_i - \overline{Y} = \hat{\beta}_1 (X_i - \overline{X}).$$

   (iii) Recall $Y_i - \widehat{Y}_i = e_i$ and use the properties of the residuals.

7. 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 is shown below:

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
..., ...		

Summary statistics are:

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

Perform analysis under the simple linear regression model.

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

(b) Calculate error sum of squares (SSE) and mean squared error (MSE). What is the degrees of freedom of SSE? (Hint: use the formula $SSE = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$.)

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

(d) Assume Normal error model for the rest of the problem. 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.

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

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

(g) 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 (f), what do you find?

8. Perform ANOVA on the "crime rate and education" data.

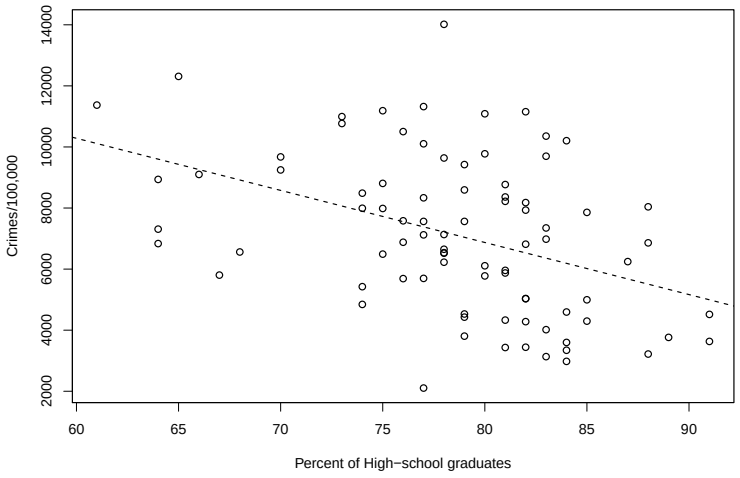
(a) Calculate sum of squares: $SSTO$, SSE and SSR. What are their respective degrees of freedom?

(b) Calculate the mean squares.

(c) Summarize results from parts (a) and (b) into an ANOVA table.

(d) Assume Normal error model, use the F test to 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.

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



(e) Compare your calculation from part (d) with those from Problem 6(d). What do you find?

9. **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.