# Statistics 206

Homework 1 Solution

*Due : October 7, 2015, In Class*

1. Please tell true or false of the following statements.

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

**ANS.** True. since $y = \overline{Y} + \hat{\beta}_1(x - \overline{X})$.

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

**ANS.** False. LS line fits the data better.

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

**ANS.** True since $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$

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

**ANS.** True. since $s\{\hat{\beta}_0\}$ and $s\{\hat{\beta}_1\}$ has $\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}$ in the denominator.

(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.

**ANS.** True. since the squared standard error of the prediction of a new observation is the sum of the squared standard error of the estimation of the mean response plus MSE; and the width of a confidence or prediction interval is proportional to the respective standard error.

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

$$P(0.3 \leq \beta_0 \leq 0.5) = 0.95.$$

**ANS.** False. $\beta_0$ is a fixed number, so it is either in between 0.3 and 0.5 or not, so the probably is either 1 or 0.

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

**ANS.** True. Decision rule depends on whether the alternative is left-sided or right-sided or two-sided.

(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.

**ANS.** True. The width of the confidence interval is proportional to $s\{\widehat{Y}_h\}$ $= \sqrt{MSE[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}]}$, so it is bigger when $|X_h - \bar{X}|$ is bigger.

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

**ANS.** TRUE. When $|r_{XY}| < 1$, then one standard deviation change in $X$ would amount to less than one standard deviation change in $Y$ (on average) and such a phenomena is called the regression effect.

2. **Exercise.** Derive the following.

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

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

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + (\overline{X})^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}(\overline{X})^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n(\overline{X})^2 + n(\overline{X})^2 \\
&= \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2.
\end{aligned}$$

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

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\overline{Y} - \sum_{i=1}^{n} \overline{X}Y_i + \sum_{i=1}^{n} \overline{X}\,\overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - n\overline{X}\,\overline{Y} - n\overline{X}\,\overline{Y} + n\overline{X}\,\overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - n\overline{X}\,\overline{Y}.
\end{aligned}$$

(c) Derive the LS estimators for simple linear regression model.
From the lecture notes, the LS estimators can be derived by finding $b_0$ and $b_1$ which satisfy the normal equations.

$$nb_0 + b_1\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i....(1)$$

$$b_0\sum_{i=1}^{n} X_i + b_1\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_iY_i....(2)$$

From equation (1), $b_0 = \overline{Y} - b_1\overline{X}$.
Using this in equation (2) we have

$$\begin{aligned} b_0 n\overline{X} + b_1 \textstyle\sum_{i=1}^n X_i^2 &= \textstyle\sum_{i=1}^n X_iY_i \\ \Rightarrow n\overline{X}\overline{Y} + b_1[\textstyle\sum_{i=1}^n X_i^2 - n\overline{X}^2] &= \textstyle\sum_{i=1}^n X_iY_i \\ \Rightarrow b_1 &= \tfrac{\sum_{i=1}^n X_iY_i - n\overline{X}\overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} \end{aligned}$$

Now from 2(a) and (b), $\hat{\beta}_1 = \frac{\sum_{i=1}^n X_iY_i - n\overline{X}\overline{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^n (X_i-\overline{X})^2}$.
From equation (1), $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$.

3. Properties of the residuals. 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) $\sum_{i=1}^n e_i = 0$.

$$\begin{aligned} \sum_{i=1}^n e_i &= \sum_{i=1}^n (Y_i - \overline{Y}) - \sum_{i=1}^n \hat{\beta}_1(X_i - \overline{X}) \\ &= (\sum_{i=1}^n Y_i - \sum_{i=1}^n \overline{Y}) - \hat{\beta}_1(\sum_{i=1}^n X_i - \sum_{i=1}^n \overline{X}) \\ &= (n\overline{Y} - n\overline{Y}) - \hat{\beta}_1(n\overline{X} - n\overline{X}) \\ &= 0. \end{aligned}$$

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

$$\begin{aligned} \sum_{i=1}^n X_ie_i &= \sum_{i=1}^n X_i((Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})) \\ &= (\sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_i\overline{Y}) - \hat{\beta}_1(\sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_i\overline{X}) \\ &= (\sum_{i=1}^n X_iY_i - n\overline{X}\ \overline{Y}) - \hat{\beta}_1(\sum_{i=1}^n X_i^2 - n(\overline{X})^2) \\ &= \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X})^2 \\ &= \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) - \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \sum_{i=1}^n (X_i - \overline{X})^2 \\ &= 0 \end{aligned}$$

(c) $\sum_{i=1}^{n} \widehat{Y}_i e_i = 0$.
By parts (a) and (b), and $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

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

(a) The LS estimator $\hat{\beta}_0$ is an unbiased estimator of $\beta_0$ and derive the formula for its variance.

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y}) - E(\hat{\beta}_1 \bar{X}) = \frac{1}{n} \sum (\beta_0 + \beta_1 X_i) - \bar{X} E(\hat{\beta}_1) \\ &= \beta_0 + \beta_1 \bar{X} - \bar{X}\beta_1 \quad \text{use the fact that } E(\hat{\beta}_1) = \beta_1 \\ &= \beta_0. \end{aligned}$$

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

$$\begin{aligned} \bar{Y} &= \frac{1}{n} \sum_{i=1}^{n} Y_i, \quad \hat{\beta}_1 = \sum \frac{X_i - \bar{X}}{\sum_{j=1}^{n} (X_j - \bar{X})^2} Y_i \\ Cov(\bar{Y}, \hat{\beta}_1) &= \sum_{i=1}^{n} \frac{1}{n} \frac{X_i - \bar{X}}{\sum_{j=1}^{n} (X_j - \bar{X})^2} Var(Y_i) \quad \text{(by } Y_i\text{s are uncorrelated)} \\ &= \frac{1}{n} \frac{\sum_{i=1}^{n} X_i - n\bar{X}}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \sigma^2 = 0 \quad \text{since } \sum_{i=1}^{n} X_i - n\bar{X} = 0. \end{aligned}$$

5. Under the simple linear regression model:

(a) Show that

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

$$\begin{aligned} \hat{Y}_i &= \hat{\beta_0} + \hat{\beta_1} X_i = \overline{Y} + \hat{\beta_1}(X_i - \overline{X}) \\ SSE &= \textstyle\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y} - \hat{\beta_1}(X_i - \overline{X}))^2 \\ &= \textstyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2 + \hat{\beta_1}^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 - 2\hat{\beta_1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \textstyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2 + \hat{\beta_1}^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 - 2\hat{\beta_1}.\hat{\beta_1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \\ \text{since} &\qquad \textstyle\hat{\beta_1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \textstyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2 - \hat{\beta_1}^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 \end{aligned}$$

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

$$\begin{aligned}
E(\hat{\beta}_1^2) &= \mathrm{Var}(\hat{\beta}_1) + E(\hat{\beta}_1)^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_1^2 \\
\text{since }\ \mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
\text{and }\ E(\hat{\beta}_1) &= \beta_1 \ \text{ from lecture notes.}
\end{aligned}$$

(c) Show that

$$E((Y_i - \bar{Y})^2) = \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2),$$

where $\bar{\epsilon} = \frac{1}{n}\sum_{i=1}^n \epsilon_i$.

$$\begin{aligned}
E((Y_i - \bar{Y})^2) &= E((\beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 - \beta_1 \bar{X} - \bar{\epsilon}))^2 \\
&= E(\beta_1 (X_i - \bar{X}) + \epsilon_i - \bar{\epsilon})^2 \\
&= \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2) + 2\beta_1 (X_i - \bar{X}) E(\epsilon_i - \bar{\epsilon}) \\
&= \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2) \ \text{ as } \ E(\epsilon_i - \bar{\epsilon}) = E(\epsilon_i) - E(\bar{\epsilon}) = 0
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{Var}(\epsilon_i - \bar{\epsilon}) &= \mathrm{Var}(\epsilon_i) + \mathrm{Var}(\bar{\epsilon}) - 2\mathrm{Cov}(\epsilon_i, \bar{\epsilon}) \\
&= \sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} \\
\text{as }\ \mathrm{Cov}(\epsilon_i, \bar{\epsilon}) = \mathrm{Cov}(\epsilon_i, \frac{\sum_{i=1}^n \epsilon_i}{n}) &= \frac{\mathrm{Var}(\epsilon_i)}{n} \\
&= \frac{(n-1)\sigma^2}{n}
\end{aligned}$$

$$\therefore E(\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})^2) = \sum_{i=1}^n E(\epsilon_i - \bar{\epsilon})^2 = \sum_{i=1}^n \mathrm{Var}(\epsilon_i - \bar{\epsilon}) = \frac{n(n-1)\sigma^2}{n} = (n-1)\sigma^2$$

From 5(a),

$$\begin{aligned}
E(SSE) &= \sum_{i=1}^n E(Y_i - \overline{Y})^2 - E(\hat{\beta}_1^2) \sum_{i=1}^n (X_i - \overline{X})^2, \ \text{ then from 5(b) and (c)} \\
&= \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n E((\epsilon_i - \bar{\epsilon})^2) - (\frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_1^2) \sum_{i=1}^n (X_i - \overline{X})^2 \\
&= \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 + (n-1)\sigma^2 - \sigma^2 - \beta_1^2 \sum_{i=1}^n (X_i - \overline{X})^2 \\
&= (n-2)\sigma^2
\end{aligned}$$

6. Under the simple linear regression model:

(a) Show that the regression sum of squares

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2$$

$$= \sum_{i=1}^{n} (\hat{\beta}_1 (X_i - \bar{X}))^2$$

$$= \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

(b) Derive $E(SSR)$.

$$E(SSR) = E(\hat{\beta}_1^2) \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$= (\frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} + \beta_1^2) \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$= \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

(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.

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 =$$

$$= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y})$$

Now,

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = \sum_{i=1}^{n} e_i(\widehat{Y}_i - \overline{Y})$$

$$= \sum_{i=1}^{n} \widehat{Y}_i e_i - \overline{Y}\sum_{i=1}^{n} e_i = 0 - \overline{Y}.0 = 0$$

$$\therefore SSTO = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

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:

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

Perform analysis under the simple linear regression model.

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

$$\overline{X} = \sum_{i=1}^{84} X_i/84 = 6602/84 = 78.6, \ \overline{Y} = \sum_{i=1}^{84} Y_i/84 = 597341/84 = 7111.2,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{84} X_iY_i - 84 \times \overline{X}\ \overline{Y}}{\sum_{i=1}^{84} X_i^2 - 84 \times (\overline{X})^2} = \frac{46400230 - 84 \times 78.6 \times 7111.2}{522098 - 84 \times 78.6^2} = -174.88,$$

$$\hat{\beta}_0 = \overline{Y} - 84 \times \overline{X} = 7111.2 - (-174.88) \times 78.6 = 20856.77.$$

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

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2 - \hat{\beta}_1^2 (\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2) \\ &= 4796548849 - 84 \times (7111.2)^2 - (-174.88)^2 \times (522098 - 84 \times 78.6^2) \\ &= 452422030, \end{aligned}$$

$$MSE = SSE/(n-2) = 452422030/(84-2) = 5517342.$$

The degrees of freedom of SSE is 82.

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

$$\begin{aligned} s\{\hat{\beta}_0\} &= \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]} \\ &= \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}\right]} \\ &= \sqrt{5517342(\frac{1}{84} + \frac{78.6^2}{522098 - 84 \times 78.6^2})} \\ &= 3299.819, \end{aligned}$$

$$\begin{aligned} s\{\hat{\beta}_1\} &= \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \\ &= \sqrt{\frac{MSE}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}} \\ &= \sqrt{\frac{5517342}{522098 - 84 \times 78.6^2}} \\ &= 41.8556 \end{aligned}$$

(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.

$$H_0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$$

$$T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}} = -174.88/41.8556 = -4.178,$$

Under null hypothesis, $T^* \sim t_{(82)}$. The critical value for two sided test at $\alpha = 0.01$ is $t(0.995, 82) \approx 2.66$. Since the observed $|T^*| = 4.178 > 2.66$, we reject the null hypothesis at 0.01 level. We conclude that there is a significant linear association between crime rate and percentage of high school graduates.

(e) What is an unbiased estimator for $\beta_0$? Construct a 99% confidence interval for $\beta_0$. Interpret your confidence interval.
The LS estimator $\hat{\beta}_0$ is an unbiased estimator for $\beta_0$.
99% confidence interval for $\beta_0$:

$$\begin{aligned} \hat{\beta}_0 \pm t(0.995; 82) s\{\hat{\beta}_0\} &= 20856.77 \pm 2.66 \times 3299.819 \\ &= [12079.25, 29634.29] \end{aligned}$$

We are 99% confident that the regression interception is in between 12079.25 and 29634.29.

(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.
95% confidence interval for $E(Y_h)$:

$$\begin{aligned} & \widehat{Y}_h \pm t(1-\alpha/2; n-2) s(\widehat{Y}_h) \\ = \ & \widehat{Y}_h \pm t(0.975; 82) \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]} \\ = \ & (\hat{\beta}_0 + \hat{\beta}_1 X_h) \pm t(0.975; 82) \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n X_i^2 - n(\overline{X})^2}\right]} \\ = \ & [20856.77 + (-174.88) \times 85] \pm 2 \times \sqrt{5517342\left[\frac{1}{84} + \frac{(85 - 78.6)^2}{522098 - 84 \times 78.6^2}\right]} \\ = \ & 5991.97 \pm (2 \times 370.73) \\ = \ & [5250.51, 6733.43] \end{aligned}$$

Note $t(0.975; 82) \approx 2$. We are 95% confident that the mean crime rate is in between 5250.51 and 6733.43 for counties with percentage of high school graduates being 85.

(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?
Predicted crime rate of county A:

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = 20856.77 + (-174.88) \times 85 = 5991.97$$

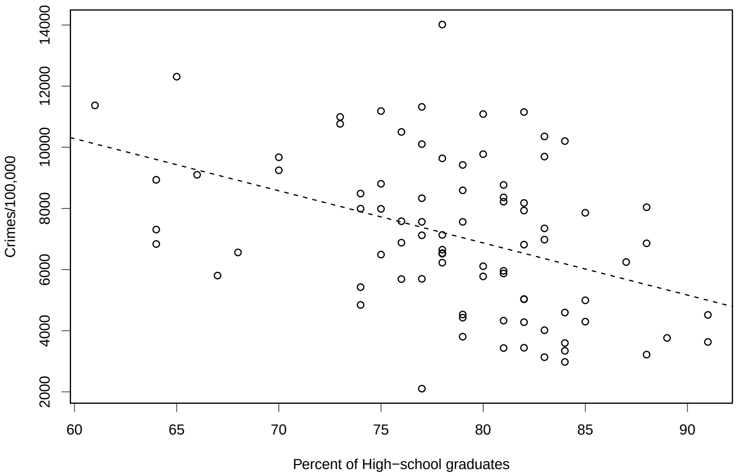
95% prediction interval for $Y_{h(new)}$:

$$\begin{aligned}
& \widehat{Y}_h \pm t(1-\alpha/2; n-2)s(pred) \\
= \;& \widehat{Y}_h \pm t(0.975; 82)\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]} \\
= \;& \widehat{Y}_h \pm t(0.975; 82)\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}\right]} \\
= \;& 5991.97 \pm (2 \times 2377.98) \\
= \;& [1236.01, 10747.93].
\end{aligned}$$

We are 95% confident that the predicted crime rate is in between 1236.01 and 10747.93. This prediction interval is much wider than the corresponding confidence interval of the mean response from part (f) because the $s\{pred\}$ is much larger.

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

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



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

$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2 = 4796548849 - 84 * (7111.2)^2 = 548738952$$

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \overline{X})^2 = \hat{\beta}_1^2[\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2] \\ &= (-174.88)^2 \times (522098 - 84 \times 78.6^2) = 96316922 \end{aligned}$$

$$SSE = SSTO - SSR = 452422030$$

The respective degrees of freedom are 83, 1, 82.

(b) Calculate the mean squares.

$$MSR = \frac{SSR}{1} = 96316922$$

$$MSE = \frac{SSE}{n-2} = \frac{452422030}{82} = 5517342$$

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

| Source<br>of Variation | SS | d.f. | MS | $F^*$ |
|---|---|---|---|---|
| Regression | $SSR = 96316922$ | d.f.$(SSR) = 1$ | $MSR = 96316922$ | $F^* = MSR/MSE$ |
| Error | $SSE = 452422030$ | d.f.$(SSE) = 82$ | $MSE = 5517342$ | =17.46 |
| Total | $SSTO = 548738952$ | d.f.$(SSTO) = 83$ | | |

(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.

$$H_0 : \beta_1 = 0 \ vs. \ H_a : \beta_1 \neq 0$$

Test statistic $F^* = MSR/MSE = 17.46$. Under null hypothesis, $F^* \sim F_{1,82}$. Since $F(0.99; 1, 82) = 6.95 < F^* = 17.46$, reject $H_0$ and conclude that there is a significant linear association between crime rate and percentage of high school graduates.

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

$$F^* = 17.46 = (-4.178)^2 = (T^*)^2, \quad F(0.99; 1, 82) = 6.95 = 2.637^2 = (T(0.995, 82))^2$$

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 $\beta_0, \beta_1$, respectively.

(b) The MLE of $\sigma^2$ is $SSE/n$.

The likelihood function is $\prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\Big\{ -\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\Big\}$. Taking the negative log of the likelihood and we get:

$$L(\beta_0, \beta_1, \sigma^2) = \frac{n}{2}(\log 2\pi + \log\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

Now take derivative w.r.t. $\beta_0, \beta_1$ and solve the minimizers:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

Which are the same as the LS estimators.
To find the MLE of $\sigma^2$ we minimze the negative log likelihood over $\sigma^2$, (use also the derivative approach) and get $\hat{\sigma}^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/n = SSE/n$.