Statistics 206

Homework 2 (Solution)

Due : Oct. 13, 2022, 11:59PM

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.rmd".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

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.

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.

   (b) 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.

(c) If all observations Y_i fall on one straight line, then the coefficient of determination $R^2 = 1$.

ANS. True. If all Y_i are on one straight line then $SSE = 0$ and $R^2 = 1$.

(d) A large R^2 always means that the fitted regression line is a good fit of the data, while a small R^2 always means that the predictor and the response are not related.

ANS. False. The predictor and response may be nonlinearly related and then R^2 would be misleading for such cases.

(e) The regression sum of squares SSR tends to be large if the estimated regression slope is large in magnitude or the dispersion of the predictor values is large.

ANS. True. $SSR = \hat{\beta}_1^2 \sum (X_i - \overline{X})^2$.

2. Under the simple linear regression model:

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

ANS.

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

(b) Show that the regression sum of squares

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

Proof.

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2 \\ &= \sum_{i=1}^n (\hat{\beta}_1 (X_i - \bar{X}))^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2 \end{aligned}$$

□

(c) Derive $E(SSR)$.

ANS.

$$\begin{aligned} E(SSR) &= E(\hat{\beta}_1^2)\sum_{i=1}^{n}(X_i-\overline{X})^2 \\ &= \left(\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}+\beta_1^2\right)\sum_{i=1}^{n}(X_i-\overline{X})^2 \\ &= \sigma^2+\beta_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2 \end{aligned}$$

3. 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.,

$$\mathrm{Cov}(e_i,\hat{\beta}_0)=0, \quad \mathrm{Cov}(e_i,\hat{\beta}_1)=0$$

for $i=1,\cdots,n$.

Proof. Note that

$$e_i = Y_i-\hat{Y}_i = Y_i-\bar{Y}-\hat{\beta}_1(X_i-\bar{X})$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(X_i-\overline{X})Y_i}{\sum_{i=1}^{n}(X_i-\overline{X})^2}=\sum_{i=1}^{n}K_iY_i, \quad K_i=\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$\begin{aligned} \mathrm{Cov}(e_i,\hat{\beta}_1) &= \mathrm{Cov}(Y_i-\bar{Y}-\hat{\beta}_1(X_i-\bar{X}),\hat{\beta}_1) \\ &= \mathrm{Cov}(Y_i,\hat{\beta}_1)-\mathrm{Cov}(\bar{Y},\hat{\beta}_1)-\mathrm{Var}(\hat{\beta}_1)(X_i-\bar{X}) \\ &= K_i\sigma^2-0-K_i\sigma^2=0 \end{aligned}$$

where

$$\mathrm{Cov}(Y_i,\hat{\beta}_1)=\mathrm{Cov}(Y_i,\sum_{i=1}^{n}K_iY_i)=K_i\mathrm{Var}(Y_i)=K_i\sigma^2$$

$$\mathrm{Cov}(\bar{Y},\hat{\beta}_1)=\mathrm{Cov}\left(\frac{\sum_{i=1}^{n}Y_i}{n},\sum_{i=1}^{n}K_iY_i\right)=\frac{\sum_{i=1}^{n}K_i\mathrm{Var}(Y_i)}{n}=\frac{\sum_{i=1}^{n}K_i}{n}\sigma^2=0 \ \text{ as } \ \sum_{i=1}^{n}K_i=0$$

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

$$\mathrm{Cov}(e_i,\hat{\beta}_0)=\mathrm{Cov}(e_i,\bar{Y}-\hat{\beta}_1\bar{X})=\mathrm{Cov}(e_i,\bar{Y})-\mathrm{Cov}(e_i,\hat{\beta}_1)\bar{X}=\mathrm{Cov}(e_i,\bar{Y})$$

where

$$\begin{aligned} \mathrm{Cov}(e_i,\bar{Y}) &= \mathrm{Cov}(Y_i-\bar{Y}-\hat{\beta}_1(X_i-\bar{X}),\bar{Y}) \\ &= \mathrm{Cov}(Y_i,\frac{\sum_{i=1}^{n}Y_i}{n})-Var(\bar{Y})-(X_i-\bar{X})\mathrm{Cov}(\hat{\beta}_1,\bar{Y}) \\ &= \frac{\sigma^2}{n}-\frac{\sigma^2}{n}-0=0 \ \text{ as } \ \mathrm{Cov}(\hat{\beta}_1,\bar{Y})=0. \end{aligned}$$

□

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

   Proof. Let $e = (e_1, e_2, .., e_n)$. From problem 2, $Cov(e, \hat{\beta}_0) = 0$ and $Cov(e, \hat{\beta}_1) = 0$. Since $e, \hat{\beta}_0, \hat{\beta}_1$ are jointly normally distributed, hence $Cov(e, \hat{\beta}_0) = 0$ and $Cov(e, \hat{\beta}_1) = 0$ imply e and $\hat{\beta}_0$ are independent and e and $\hat{\beta}_1$ are independent.

   Since $SSE = e'e$ is a function of e, it is independent of the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. □

5. Under the simple linear regression model, derive $\mathrm{Var}(\widehat{Y}_h)$, where

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

   is the estimator of the mean response $\beta_0 + \beta_1 X_h$.

ANS.

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_h = \bar{Y} + \hat{\beta}_1 (X_h - \bar{X})$$

$$\begin{aligned} \mathrm{Var}(\widehat{Y}_h) &= \mathrm{Var}(\bar{Y}) + \mathrm{Var}(\hat{\beta}_1)(X_h - \bar{X})^2 + 2(X_h - \bar{X})\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}(X_h - \bar{X})^2 + 0 \quad \text{as} \quad \mathrm{Cov}(\bar{Y}, \hat{\beta}_1) = 0 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]. \end{aligned}$$

6. **Crime Rate and Education.** 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:

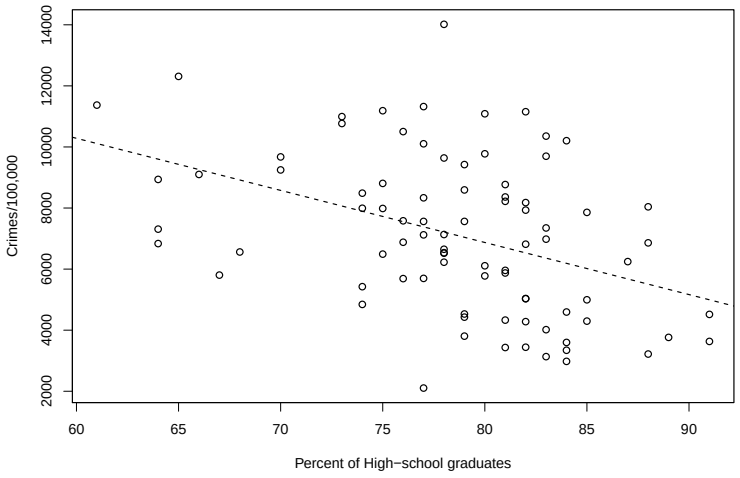
County	Crimes/100,000	Percent-of-High-school-graduates
1	8487	74
2	8179	82
3	8362	81
4	8220	81
5	6246	87
6	9100	66
..., ...		

   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.

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



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

ANS. The relationship between percentage of high school graduates and crime rate looks linear. The crime rate seems to decrease with higher percentage of high school graduates.

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

ANS.

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

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

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

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

ANS.

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

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

The degrees of freedom of SSE is 82.

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

ANS.

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

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

(e) 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.

ANS.

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

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

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.

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

ANS. The LS estimator $\hat{\beta}_0$ is an unbiased estimator for β_0.
99% confidence interval for β_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 intercept is in between 12079.25 and 29634.29.

(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.

ANS. 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.

(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 observe?

ANS. 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 (g) because the $s\{pred\}$ is much larger.

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

ANS. Normal errors

7. 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?

ANS.

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

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

$$SSE = SSTO - SSR = 452422030$$

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

(b) Calculate the mean squares.

ANS.

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

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

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

Source 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.

ANS.

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

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 part (e) from Problem 4. What do you observe?

ANS.

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

We observe that F-test for the regression effect is equivalant to t-test for the regression coefficient. Since $F^* = (T^*)^2$ and $F(1-\alpha; 1, n-1) = (t(1-\frac{\alpha}{2}; n-1))^2$, rejecting null for large value of F^* is equivalant to rejecting null for large value of $|T^*|$.

(f) Calculate the coefficient of determination R^2.

ANS. $R^2 = SSR/SSTO = 0.1755$

(g) Calculate the correlation coefficient between crime rate and percentage of high school graduates. Compare r^2 and R^2. What do you observe?

ANS.

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^n (X_i - \overline{X})^2\right)\left(\sum_{i=1}^n (Y_i - \overline{Y})^2\right)}} \\ &= \frac{\sum_{i=1}^n X_i Y_i - n\overline{XY}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right)\left(\sum_{i=1}^n Y_i^2 - n\overline{Y}^2\right)}} \\ &= -0.41896 \end{aligned}$$

We observe that $r^2 = R^2$. Note that

$$\begin{aligned} r^2 &= \frac{\left(\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})\right)^2}{\left(\sum_{i=1}^n (X_i - \overline{X})^2\right)\left(\sum_{i=1}^n (Y_i - \overline{Y})^2\right)} \\ &= \left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)^2 \cdot \sum_{i=1}^n (X_i - \overline{X})^2 \cdot \frac{1}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \\ &= \frac{\hat{\beta}_1^2 \left(\sum_{i=1}^n (X_i - \overline{X})^2\right)}{SSTO} = \frac{SSR}{SSTO} = R^2. \end{aligned}$$

8. **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?

Proof. 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. β_0, β_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 σ^2 we minimze the negative log likelihood over σ^2, (use also the derivative approach) and get $\hat{\sigma}^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/n = SSE/n$.

$$E(\hat{\sigma}^2) = E(SSE/n) = (n-2)\sigma^2/n,$$

so MLE of σ^2 is biased. □