# 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.

   **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) 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.

   (c) 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.

   (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.

   **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.

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 $i = 1, \cdots, n$.

*Proof.*

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

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

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

□

3. Under the Normal error model:

   (a) 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.
   $\therefore SSE = e'e$, a function of $e$ is independent of 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 $\beta_0 + \beta_1 X_h$.

$$\begin{aligned} \widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1X_h = \bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1X_h = \bar{Y} + \hat{\beta}_1(X_h - \bar{X}) \\ Var(\widehat{Y}_h) = Var(\bar{Y}) + Var(\hat{\beta}_1)(X_h - \bar{X})^2 + 2(X_h - \bar{X})Cov(\bar{Y}, \hat{\beta}_1) \\ = \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}(X_h - \bar{X})^2 + 0 \ \text{ as } \ Cov(\bar{Y}, \hat{\beta}_1) = 0 \\ = \sigma^2[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}]. \end{aligned}$$

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

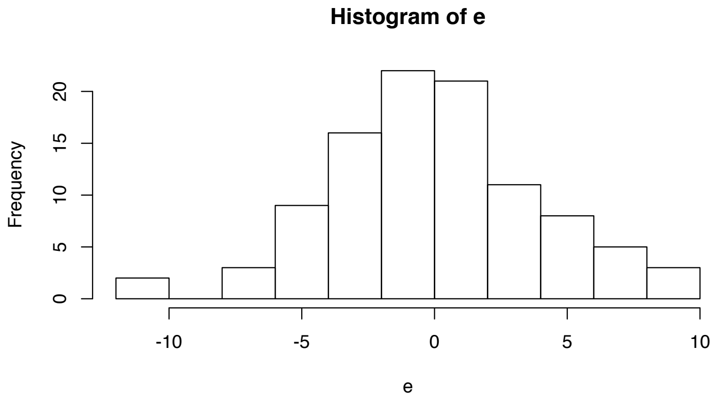
   (a) Create a sequence of consecutive integers ranging from 1 to 100. Record these in a vector x. (Hint: use the `seq` function)

   ```
   x <- seq(1, 100)
   ```

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

   ```
   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)



   ```
   e <- rnorm(n = 100, mean = 0, sd = 5)
   mean(e)
   [1] 0.5343712
   var(e)
   [1] 20.4063
   hist(e)
   ```

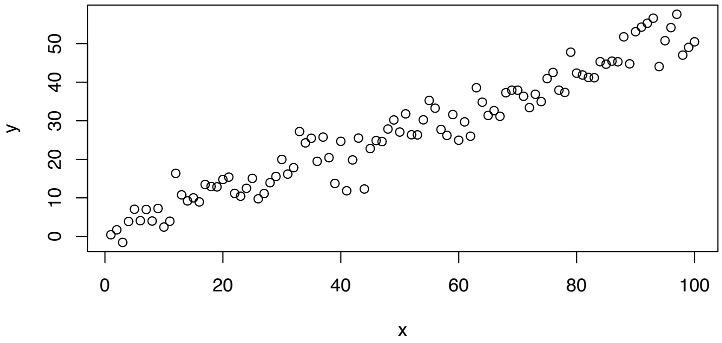
   (d) Add (element-wise) the numbers created in part (c) to the vector w. Record the new vector as y.

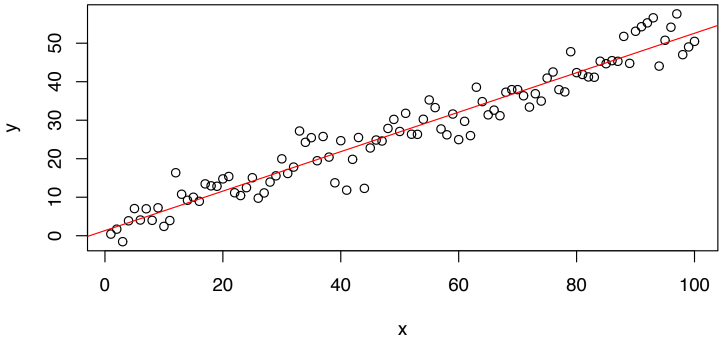
   ```
   y <- w + e
   ```

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

   ```
   plot(x, y)
   ```

   (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?



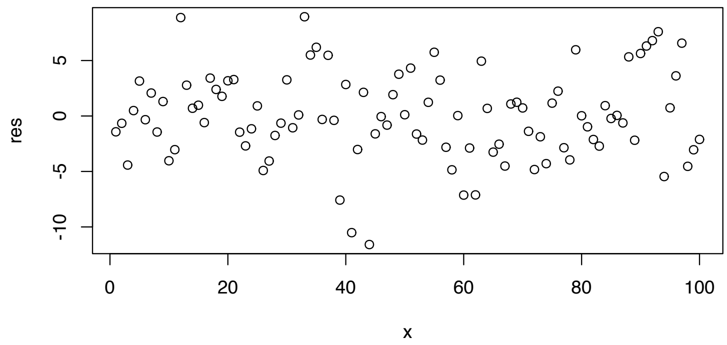


```
fit <- lm(y ~ x)
beta0 <- fit$coefficients[1]
beta1 <- fit$coefficients[2]
plot(x, y)
abline(a = beta0, b = beta1, col = ”red”)
```

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

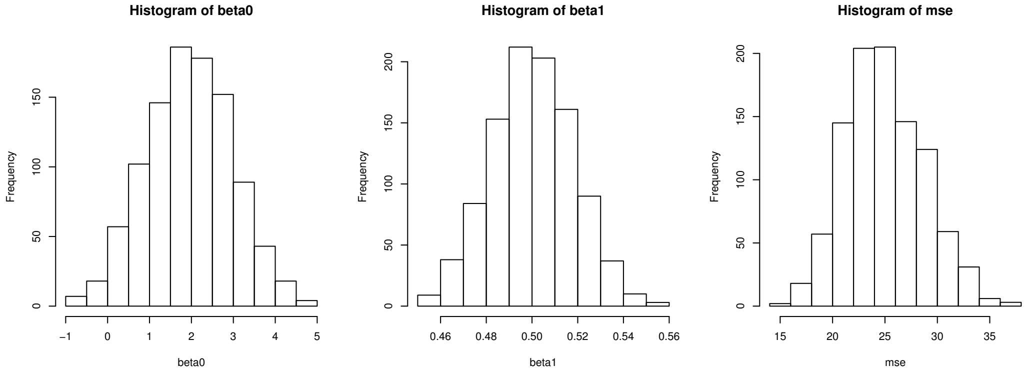
Residuals are randomly distributed around 0.

```
res <- residuals(fit)
plot(x, res)
sse <- sum(res^2)
mse <- sse/(100 - 2)
[1] 24.19991
```



(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.



```
beta0 <- c()
beta1 <- c()
mse <- c()
for(k in 1:1000){
  e <- rnorm(n = 100, mean = 0, sd = 5)
  y <- w + e
  fit <- lm(y ~ x)
  beta0[k] <- fit$coefficients[1]
  beta1[k] <- fit$coefficients[2]
  res <- residuals(fit)
  sse <- sum(res^2)
  mse[k] <- sse/(100 - 2)
}
> mean(beta0)
[1] 1.959226
> mean(beta1)
[1] 0.5005168
> mean(mse)
[1] 24.99118
> var(beta0)
[1] 1.029963
> var(beta1)
[1] 0.0003207875
> var(mse)
[1] 14.06196
```

The distribution of each of the three estimators is bell shaped, with sample mean close to true parameter.

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
..., ...
```

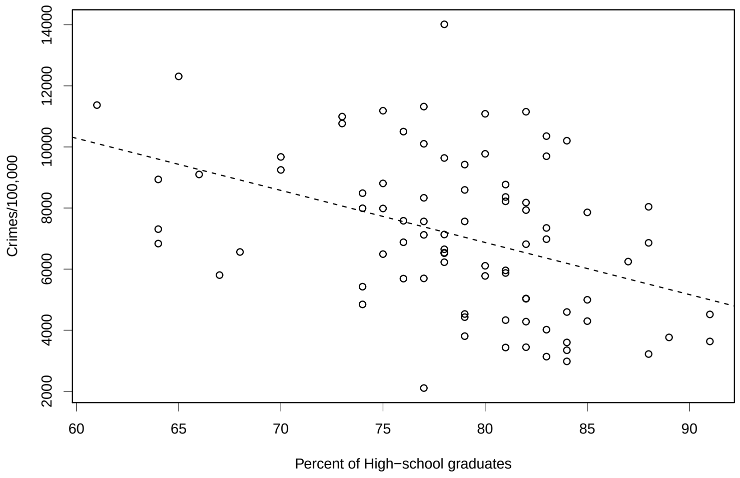
Some summary statistics are also given:

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

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.
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.

Figure 1: Scatter plot of Crime rate vs. 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 $\hat{\beta}_1$.

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

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

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

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

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

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

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

(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.

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

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

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 $\beta_0$? 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 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. 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 (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.

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

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 $\beta_0, \beta_1$, respectively.

(b) The MLE of $\sigma^2$ is $SSE/n$. Is MLE of $\sigma^2$ 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. $\beta_0, \beta_1$ 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 $\hat{\sigma}^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/n = SSE/n$.

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

so MLE of $\sigma^2$ is biased. □