Practice Problems for Quiz 1
Linear Regression

Directions: This is a (simulated) closed book exam. You do not need a calculator to complete the exam– your real exam will not allow use of a calculator. You must show your work for full credit. Partial credit can only be given if your thoughts are clear and can be followed. Make sure your name is written on every page.

1. *Understanding the simple linear regression model.* Say whether each of the following statements are true or false and **explain why**. (The majority of the points come from the **explanation**.)

   (a) A 95% confidence interval for the slope β_1 based on the observed data was calculated as $[-1, 0.5]$. Therefore
   $$P(-1 \leqslant \beta_1 \leqslant 0.5) = 0.95.$$

   (b) For the simple linear regression model, the larger the quantity $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is, the smaller the standard error of the least squares slope estimator tends to be.

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

a. False. There's nothing random in the statement "$-1 \leq \beta_1 \leq 0.5$" so it is either true or false depending on β_1. The probability is 1 if true, 0 if false.

b. Yes - $SE(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} = \sqrt{\sigma^2/SSX} = \sigma/\sqrt{SSX}$.

As SSX ↑, then $SE(\hat{\beta}_1)$ ↓.

c. Yes, since $\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SSX}} > \hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SSX}}$

$\widehat{SE}(\hat{y} - y_{new})$ ⇕ prediction int.

$\widehat{SE}(\hat{y} - \beta_0 - \beta_1 x)$ ⇕ confidence int

the width of the PI is always bigger

2. *Analysis, inference, and interpretation.* A biologist conducted an experiment to determine the relationship between auditory stimulation and heart rate. Eighteen subjects were placed in rooms of various sound levels (X), measured in decibels. The pulse rate (Y) of each subject, measured in beats per minute, was measured 60 seconds after exposure to the room's sound. Summary statistics are given below.

$$\bar{x} = 50, \quad \bar{y} = 150$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2500, \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 4100, \quad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 2500$$

We want to regress the pulse rate on the sound level using a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n.$$

(a) State the classical assumptions of the simple linear regression model.

(b) Calculate the least squares estimators of the intercept and the slope.

(c) The biologist wants to test whether the slope is 0 or not. Conduct the hypothesis test using a t-statistic and state your conclusion at 0.05 significance level. **Interpret the results of the test in the real life context of the problem.** You can use the facts that $t^*_{16}(0.95) = 1.75$ and $t^*_{16}(0.975) = 2.12$.

(d) Conduct the same hypothesis test using an F-test and state your conclusion at 0.05 significance level. Are the findings consistent? You can use the facts that $F^*_{1,16}(0.975) = 6.11$, $F^*_{2,16}(0.975) = 4.69$, $F^*_{1,16}(0.95) = 4.49$, and $F^*_{2,16}(0.95) = 3.63$.

(e) Construct and interpret the 95% confidence interval for the mean heart rate for a subject standing in a room of noise level $x = 50$ dB.

a. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

w/ x_i fixed & $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall\ i \neq j$ &

(classical 4th assumption) $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

b. $\hat{\beta}_1 = \frac{SSXY}{SSX} = \frac{2500}{2500} = 1$ $\Rightarrow \hat{y} = 100 + x$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 150 - 1(50) = 100$

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

$t = \frac{\hat{\beta}_1 - 0}{\hat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{SSX}}$

What is $\hat{\sigma}$? ... $\hat{\sigma}^2 = \frac{SSE}{n-2}$

Kind of tricky here:

$SSE = SST - SSR$

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

$= 4100 - \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1\bar{x})^2$

$= 4100 - \sum_{i=1}^{n}(\hat{\beta}_1(x_i-\bar{x}))^2$

$= 4100 - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2$

$= 4100 - (1)(2500) = 1600$

$\therefore \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{1600}{18-2} = \frac{1600}{16} = 100$

$\therefore \hat{\sigma} = 10$

so $t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{SS_x}} = \frac{1-0}{10/\sqrt{2500}} = \frac{1-0}{10/50} = \frac{1}{(1/5)} = 5$

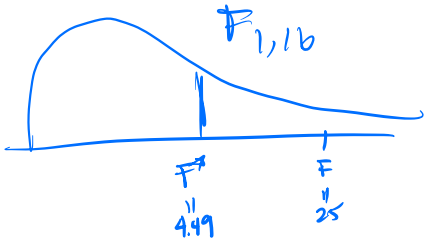
$\therefore t = 5 > t^*_{.975,16} = 2.12$

& we have evidence to reject H_0.

Interpretation in real-life context:

We have evidence that sound level is a significant predictor of pulse rate in this experimental setting.

d. $F = \frac{SSR/1}{SSE/n-2} = \frac{2500/1}{1600/16} = \frac{2500}{100} = 25 > F^*_{1,16}(.95) = 4.49$



⇒ reject H_0 & conclude that sound level is a significant predictor of pulse rate.

↑ same conclusion!

c. 95% CI for $E(Y|X=50) = \beta_0 + \beta_1(50)$

$$\hat{y}(x=50) \pm t^*_{.975,16} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(50-\bar{x})^2}{SS_X}}$$

$$(100 + (1)50) \pm 2.12(10)\sqrt{\frac{1}{16} + \frac{(50-50)^2}{2500}}$$

$\Rightarrow 150 \pm 2.12(10)(\sqrt{1/16})$ ← you can stop here on test for calculations but it is doable by hand:

$\Rightarrow 150 \pm 2.12(10)(1/4)$

$\Rightarrow 150 \pm \left(\frac{2.12}{2}\right)\left(\frac{10}{2}\right)$

$\Rightarrow 150 \pm 1.06(5)$

$\Rightarrow 150 \pm 5.3$

$\Rightarrow [144.7, 155.3]$ bpm

Interpretation:
We're 95% confident that the true avg. pulse rate of subjects when exposed to 50 dB noise levels is between 144.7 & 155.3 bpm.

3. *Rigorous proofs.* For the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n,$$

where we assume $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, consider the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$. Define the fitted values and residuals as usual: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $e_i = y_i - \hat{y}_i$, $\quad i = 1, ..., n$. Prove the following claims, citing any reasoning or results you use:

(a) $E(\bar{y}) = \beta_0 + \beta_1 \bar{x}$

(b) $\text{Var}(\hat{y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \right)$

(c) $\sum_{i=1}^n e_i (\hat{y}_i - \bar{y}) = 0$

(d) $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$

a

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_i y_i\right] = \frac{1}{n}\sum_i E(y_i)$$

$$= \frac{1}{n}\sum_i [E(\beta_0 + \beta_1 x_i + \epsilon_i)]$$

$$= \frac{1}{n}\sum_{i=1}^n [\beta_0 + \beta_1 x_i + E(\epsilon_i)] \quad (E(\epsilon_i) = 0)$$

$$= \frac{1}{n}\left(\sum_{i=1}^n \beta_0 + \beta_1 \sum_{i=1}^n x_i\right)$$

$$= \frac{n\beta_0}{n} + \beta_1 \frac{\sum_{i=1}^n x_i}{n} = \beta_0 + \beta_1 \bar{x}$$

b. $\text{Var}(\hat{y}_i) =$

$c_j = \frac{1}{n} - \bar{x} k_j$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \sum_{j=1}^n c_j y_j + x_i \sum_{j=1}^n k_j y_j$$

$k_j = \frac{x_j - \bar{x}}{SSx}$

$$= \sum_{j=1}^n (c_j + x_i k_j) y_j$$

$$= \sum_{j=1}^n \left(\frac{1}{n} - \bar{x} k_j + x_i k_j\right) y_j$$

$$= \sum_{j=1}^n \left(\frac{1}{n} + (x_i - \bar{x}) k_j\right) y_j$$

$$Var(\hat{y}_i) = Var\left(\sum_{j=1}^{n}\left(\frac{1}{n}+(x_i-\bar{x})k_j\right)y_j\right)$$

$y_i \perp y_j$ if $i \neq j$ $\Rightarrow$ (⊥)

$\therefore$ no covariance terms

$$= \sum_{j=1}^{n} Var\left(\left(\frac{1}{n}+(x_i-\bar{x})k_j\right)y_j\right)$$

$$= \sum_{j=1}^{n}\left(\frac{1}{n}+(x_i-\bar{x})k_j\right)^2 Var(y_j)$$

$$= \sum_{j=1}^{n}\left(\frac{1}{n}+(x_i-\bar{x})k_j\right)^2 \sigma^2$$

$$= \sigma^2 \sum_{j=1}^{n}\left(\frac{1}{n}+(x_i-\bar{x})k_j\right)^2$$

$$= \sigma^2 \sum_{j=1}^{n}\left(\frac{1}{n^2}+\frac{2(x_i-\bar{x})k_j}{n}+[(x_i-\bar{x})k_j]^2\right)$$

$$= \sigma^2\left[\frac{1}{n}+\frac{2(x_i-\bar{x})}{n}\sum_{j=1}^{n}k_j + (x_i-\bar{x})^2\sum_{j=1}^{n}k_j^2\right]$$

(1) under $\sum_{j=1}^{n}k_j$ (the crossed-out term $\to 0$); (2) under $\sum_{j=1}^{n}k_j^2$

(1) $$\sum_{j=1}^{n}k_j = \sum_{j=1}^{n}\frac{x_j-\bar{x}}{SSX}$$

$$= \frac{1}{SSX}\sum_{j=1}^{n}(x_j-\bar{x})$$

$$= \frac{1}{SSX}(n\bar{x}-n\bar{x}) = 0$$

(2) $$\sum_{j=1}^{n}k_j^2 = \sum_{j=1}^{n}\left(\frac{(x_j-\bar{x})}{SSX}\right)^2$$

$$= \frac{SSX}{(SSX)^2} = 1/SSX$$

$\therefore$ $$Var(\hat{y}_i) = \sigma^2\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SSX}\right)$$

c. $\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} e_i (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$= \sum_{i=1}^{n} e_i \hat{\beta}_0 + \sum_{i=1}^{n} e_i \hat{\beta}_1 x_i$

$= \hat{\beta}_0 \left[\sum_{i=1}^{n} e_i\right] + \hat{\beta}_1 \left[\sum_{i=1}^{n} e_i x_i\right] = 0$

① ②

$\sum_{i=1}^{n} e_i \bar{y} = \bar{y} \sum_{i=1}^{n} e_i$ ②

$= \bar{y}(0)$

$= 0$

① $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$= \Sigma_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \Sigma_i x$

$= n\bar{y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x}$

$= n\bar{y} - n(\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x}$

$= n\bar{y} - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 n\bar{x} = 0$

② $\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i$

$= \sum_{i=1}^{n} (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) x_i$

$= \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})) x_i$

$= \sum_{i=1}^{n} (y_i - \bar{y}) x_i - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i$

$= SSXY - \hat{\beta}_1 SSX$

can you prove these?

$= SSXy - \frac{SSXY}{SSX} \cdot SSX = 0.$ ☺

∴ $\sum_{i=1}^{n} e_i (\hat{y}_i - \bar{y}) =$

$\sum_{i=1}^{n} e_i \hat{y}_i - \sum_{i=1}^{n} e_i \bar{y} = 0 - 0 = 0.$

□

d. First) $(\hat{y}_i - \bar{y}) = (\hat{\beta}_0 + \hat{\beta}_1 x_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})$

$= \hat{\beta}_1 (x_i - \bar{x})$

Then

$\Rightarrow \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\beta}_1 (x_i - \bar{x}))^2 = \sum_{i=1}^{n} (\hat{\beta}_1^2)(x_i - \bar{x})^2$

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