Problem 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.

(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.

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

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

(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.

Solution:

(a) True. This is because

$$\sigma^2(pred_h) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) > \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) = \sigma^2(\hat{Y}_h)$$

(b) True. Observe,

$$Var(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \geq \frac{\sigma^2}{n} = Var(\hat{Y}_{\overline{X}})$$

(c) True. Because $SSR/SSTO = 1 \implies Y_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for all i.

(d) False. Pearson correlation fails to account for nonlinear relationships in data.

(e) True. This is because SSR is directly related to slope and dispersion by the formula:

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

Problem 2
Under the simple linear regression model:

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

(b) Show that the regression sum of squares

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

(c) Derive $E(SSR)$.

Solution:

(a) We remember the formula for variance as:

$$Var(Z) = E[Z^2] - E[Z]^2 \implies E[Z^2] = Var(Z) + E[Z]^2$$

On the last homework assignment, we showed that

$$E[\hat{\beta}_1] = \beta_1$$

To calculate the variance, we know that

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\underbrace{\sum_{i=1}^n (X_i - \overline{X})^2}_{S_{XX}}}\right) \\
&= \sum_{i=1}^n Var\left(\frac{(X_i - \overline{X})(Y_i - \overline{Y})}{S_{XX}}\right) \\
&= \sum_{i=1}^n \left(\frac{(X_i - \overline{X})}{S_{XX}}\right)^2 Var(Y_i - \overline{Y}) \\
&= \sum_{i=1}^n \left(\frac{(X_i - \overline{X})}{S_{XX}}\right)^2 \sigma^2 \\
&= \frac{\sigma^2}{S_{XX}^2} \sum_{i=1}^n (X_i - \overline{X})^2 \\
&= \frac{\sigma^2}{S_{XX}} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}
\end{aligned}$$

Therefore, we see that

$$E[\hat{\beta}_1^2] = \beta_1^2 + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

(b) Using the alternative form of $\hat{Y}_i = \overline{Y} + \hat{\beta}_1(X_i - \overline{X})$, we see that

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

(c) Lastly, since each X_i and $\overline{X}$ is assumed to be a fixed quantity, it follows that by linearity,

$$E[SSR] = E[\hat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2] = E[\hat{\beta}_1^2]\left(\sum_{i=1}^n (X_i - \overline{X})^2\right) = \beta_1^2\left(\sum_{i=1}^n (X_i - \overline{X})^2\right) + \sigma^2$$

Problem 3
Under the simple linear regression model, show that the residuals e_i's are uncorrelated with the LS estimators β_0 and β_1, i.e.,

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

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

Solution: Recall that for each i,

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_i + \epsilon_i$$

We know from the slides that in our simple linear regression model, we assume:

$$Cov(\epsilon_i, \epsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

$$\implies E[\epsilon_i \epsilon_j] = Cov(\epsilon_i, \epsilon_j) - E[\epsilon_i]E[\epsilon_j] = Cov(\epsilon_i, \epsilon_j)$$

$$\begin{aligned}
E[\epsilon_j \hat{\beta}_0] &= E\left[\epsilon_j \left(\frac{1}{n}\sum_{i=1}^n Y_i - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^n E[\epsilon_j(\beta_0 + \beta_1 X_i + \epsilon_i)] - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})E[\epsilon_j(Y_i - \overline{Y})]}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{1}{n}\sum_{i=1}^n E[\epsilon_i \epsilon_j] - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})E[\epsilon_j(Y_i - \overline{Y})]}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{1}{n}\sum_{i=1}^n E[\epsilon_i \epsilon_j] - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_j(\beta_0 + \beta_1 X_i + \epsilon_i)] - E[\epsilon_j \overline{Y}]\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{1}{n}\sum_{i=1}^n E[\epsilon_i \epsilon_j] - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_i \epsilon_j] - E[\epsilon_j \frac{1}{n}\sum_{k=1}^n Y_i]\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{1}{n}\sum_{i=1}^n E[\epsilon_i \epsilon_j] - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_i \epsilon_j] - \frac{1}{n}\sum_{k=1}^n E[\epsilon_j \epsilon_k]\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sigma^2}{n} - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_i \epsilon_j] - \frac{\sigma^2}{n}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sigma^2}{n} - \overline{X}\frac{\sum_{i=1}^n (X_i - \overline{X})(E[\epsilon_i \epsilon_j]) - \frac{\sigma^2}{n}\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sigma^2}{n} - \overline{X}\frac{\sigma^2 (X_j - \overline{X}) - \frac{\sigma^2}{n}\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2 \overline{X}(X_j - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}
\end{aligned}$$

Observe,

$$\begin{aligned}
Cov(e_i, \hat{\beta}_0) &= E[e_i\hat{\beta}_0] - E[e_i]E[\hat{\beta}_0] \\
&= E[\left((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_i + \epsilon_i\right)\hat{\beta}_0] - E[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_i + \epsilon_i]E[\hat{\beta}_0] \\
&= E[(\beta_0 - \hat{\beta}_0)\hat{\beta}_0] + E[(\beta_1 - \hat{\beta}_1)\hat{\beta}_0 X_i + E[\epsilon_i\hat{\beta}_0] - \left(E[(\beta_0 - \hat{\beta}_0)] + E[(\beta_1 - \hat{\beta}_1)X_i] + E[\epsilon_i]\right)E[\hat{\beta}_0] \\
&= E[\beta_0\hat{\beta}_0] - E[\hat{\beta}_0\hat{\beta}_0] + E[\beta_1\hat{\beta}_0]X_i - E[\hat{\beta}_0\hat{\beta}_1]X_i + E[\epsilon_i\hat{\beta}_0] - \left(E[(\beta_0 - \hat{\beta}_0)] + E[(\beta_1 - \hat{\beta}_1)X_i] + E[\epsilon_i]\right)E[\hat{\beta}_0] \\
&= \beta_0 E[\hat{\beta}_0] - E[\hat{\beta}_0^2] + \beta_1 X_i E[\hat{\beta}_0] - E[\hat{\beta}_0\hat{\beta}_1]X_i + E[\epsilon_i\hat{\beta}_0] - \left(\beta_0 - E[\hat{\beta}_0] + \beta_1 X_i - E[\hat{\beta}_1]X_i + E[\epsilon_i]\right)E[\hat{\beta}_0] \\
&= \beta_0 E[\hat{\beta}_0] - E[\hat{\beta}_0^2] + \beta_1 X_i E[\hat{\beta}_0] - E[\hat{\beta}_0\hat{\beta}_1]X_i + E[\epsilon_i\hat{\beta}_0] - (\beta_0 - \beta_0 + \beta_1 X_i - \beta_1 X_i + 0)\,\beta_0 \\
&= \beta_0 E[\hat{\beta}_0] - E[\hat{\beta}_0^2] + \beta_1 X_i E[\hat{\beta}_0] - E[\hat{\beta}_0\hat{\beta}_1]X_i + E[\epsilon_i\hat{\beta}_0] \\
&= \beta_0^2 - \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) - \beta_0^2 + \beta_0\beta_1 X_i - \left(\frac{-\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_0\beta_1\right)X_i + E[\epsilon_i\hat{\beta}_0] \\
&= -\sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) - \left(\frac{-\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)X_i + E[\epsilon_i\hat{\beta}_0] \\
&= 0
\end{aligned}$$

Similarly,

$$\begin{aligned}
E[\epsilon_j\hat{\beta}_1] &= E\left[\epsilon_j\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right] \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})E[\epsilon_j(Y_i - \overline{Y})]}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})E[\epsilon_j(Y_i - \overline{Y})]}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_j(\beta_0 + \beta_1 X_i + \epsilon_i)] - E[\epsilon_j\overline{Y}]\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_i\epsilon_j] - E[\epsilon_j\frac{1}{n}\sum_{k=1}^n Y_i]\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_i\epsilon_j] - \frac{1}{n}\sum_{k=1}^n E[\epsilon_j\epsilon_k]\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})\left(E[\epsilon_i\epsilon_j] - \frac{\sigma^2}{n}\right)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})(E[\epsilon_i\epsilon_j]) - \frac{\sigma^2}{n}\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sigma^2(X_j - \overline{X}) - \frac{\sigma^2}{n}\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sigma^2(X_j - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}
\end{aligned}$$

$$
\begin{aligned}
Cov(e_i, \hat{\beta}_1) &= E[e_i\hat{\beta}_1] - E[e_i]E[\hat{\beta}_1] \\
&= E[\Big((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_i + \epsilon_i\Big)\hat{\beta}_1] - E[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)X_i + \epsilon_i]E[\hat{\beta}_1] \\
&= E[(\beta_0 - \hat{\beta}_0)\hat{\beta}_1] + E[(\beta_1 - \hat{\beta}_1)\hat{\beta}_1 X_i + E[\epsilon_i\hat{\beta}_1] - \Big(E[(\beta_0 - \hat{\beta}_0)] + E[(\beta_1 - \hat{\beta}_1)X_i] + E[\epsilon_i]\Big)E[\hat{\beta}_1] \\
&= E[\beta_0\hat{\beta}_1] - E[\hat{\beta}_0\hat{\beta}_1] + E[\beta_1\hat{\beta}_1]X_i - E[\hat{\beta}_1^2]X_i + E[\epsilon_i\hat{\beta}_1] - \Big(E[(\beta_0 - \hat{\beta}_0)] + E[(\beta_1 - \hat{\beta}_1)X_i] + E[\epsilon_i]\Big)E[\hat{\beta}_1] \\
&= \beta_0 E[\hat{\beta}_1] - E[\hat{\beta}_0\hat{\beta}_1] + \beta_1 E[\hat{\beta}_1]X_i - E[\hat{\beta}_1^2]X_i + E[\epsilon_i\hat{\beta}_1] \\
&= \beta_0\beta_1 - \left(\frac{-\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_0\beta_1\right) + \beta_1^2 X_i - X_i\left(\beta_1^2 + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + E[\epsilon_i\hat{\beta}_1] \\
&= \left(\frac{\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) - X_i\left(\frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + E[\epsilon_i\hat{\beta}_1] \\
&= 0
\end{aligned}
$$

Problem 4
Under the Normal error model: Show that SSE is independent with the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. (Hint: Use the fact that, if two sets of random variables, say $(Z_1, \ldots, Z_s)$ and $(W_1, \ldots, W_t)$, are independent with each other, then their functions, say $f(Z_1, \ldots, Z_s)$ and $g(W_1, \ldots, W_t)$, are independent.)

Solution: Observe, we can define the function

$$SSE = \sum_{i=1}^n e_i^2 = f(e_1, \ldots, e_n)$$

We also know that under the normal error model,

$$cov(e_i, \hat{\beta}_0) = 0 \implies e_i \perp\!\!\!\perp \hat{\beta}_0$$

$$cov(e_i, \hat{\beta}_1) = 0 \implies e_i \perp\!\!\!\perp \hat{\beta}_1$$

for all $1 \le i \le n$, which was proven in Problem 3. Therefore, if we let $g = id : \mathbb{R} \to \mathbb{R}$, then we see that by the fact mentioned in the prompt,

$$e_i \perp\!\!\!\perp \hat{\beta}_0 \forall i \implies f(e_1, \ldots, e_n) \perp\!\!\!\perp g(\beta_0) \implies SSE \perp\!\!\!\perp \hat{\beta}_0$$
$$e_i \perp\!\!\!\perp \hat{\beta}_1 \forall i \implies f(e_1, \ldots, e_n) \perp\!\!\!\perp g(\beta_1) \implies SSE \perp\!\!\!\perp \hat{\beta}_1$$

Problem 5
Under the simple linear regression model, derive $Var(\hat{Y}_h)$, where

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

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

Solution: We need some necessary quantities:

$$E[\hat{\beta}_0^2] = Var(\hat{\beta}_0^2) + E[\hat{\beta}_0]^2 = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + \beta_0^2$$

and

$$\begin{aligned}
E[\hat{\beta}_0\hat{\beta}_1] &= Cov(\hat{\beta}_0, \hat{\beta}_1) + E[\hat{\beta}_0]E[\hat{\beta}_1] \\
&= \frac{-\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_0\beta_1
\end{aligned}$$

Now, observe,

$$\begin{aligned}
Var(\hat{Y}_h) &= Var(\hat{\beta}_0 + \hat{\beta}_1 X_h) \\
&= E[(\hat{\beta}_0 + \hat{\beta}_1 X_h)^2] - E[\hat{\beta}_0 + \hat{\beta}_1 X_h]^2 \\
&= E[\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 X_h + \hat{\beta}_1^2 X_h^2] - E[\hat{\beta}_0 + \hat{\beta}_1 X_h]^2 \\
&= E[\hat{\beta}_0^2] + 2X_h E[\hat{\beta}_0\hat{\beta}_1] + X_h^2 E[\hat{\beta}_1^2] - \Big(E[\hat{\beta}_0] + X_h E[\hat{\beta}_1]\Big)^2 \\
&= E[\hat{\beta}_0^2] + 2X_h E[\hat{\beta}_0\hat{\beta}_1] + X_h^2 E[\hat{\beta}_1^2] - \Big(E[\hat{\beta}_0]^2 + 2X_h E[\hat{\beta}_0]E[\hat{\beta}_1] + X_h^2 E[\hat{\beta}_1]^2\Big) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + \beta_0^2 + 2X_h E[\hat{\beta}_0\hat{\beta}_1] + X_h^2\beta_1^2 + X_h^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&- \left(\beta_0^2 + 2X_h\beta_0\beta_1 + X_h^2\beta_1^2\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + 2X_h E[\hat{\beta}_0\hat{\beta}_1] + X_h^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} - 2X_h\beta_0\beta_1 \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + 2X_h\left(\frac{-\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_0\beta_1\right) + X_h^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} - 2X_h\beta_0\beta_1 \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + 2X_h\left(\frac{-\sigma^2\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) + X_h^2 \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2 - 2X_h\overline{X} + X_h^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \\
&= \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)
\end{aligned}$$

Problem 6
Submitted as a Markdown file.

Problem 7
Submitted as a Markdown file.