1. Suppose we have $n = 10$ observations (X_i, Y_i) and fit the data with model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

with $\varepsilon_i, i = 1, ..., 10$ are IID $N(0, \sigma^2)$. We have the following calculations.

$$\bar{X} = 0.5669, \quad \bar{Y} = 0.9624, \quad \sum_{i=1}^{n} Y_i^2 = 10.2695,$$

$$\sum_{i=1}^{n} X_i^2 = 4.0169, \quad \sum_{i=1}^{n} X_i Y_i = 6.2841.$$

(a) Write down the estimated model

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = 1.030974, \qquad b_0 = \bar{Y} - b_1\bar{X} = 0.3780$$

$$SST = \sum_{i=1}^{n}(Y_i^2 - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = 1.0069$$

$$SSR = b_1^2 \sum_{i=1}^{n}(X_i^2 - \bar{X})^2 = b_1^2(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2) = 0.85427$$

$$SSE = SST - SSR = 0.15269, \quad MSE = SSE/(n-2) = 0.01908625$$

Thus

$$s(b_1) = \sqrt{MSE / \sum_{i=1}^{n}(X_i - \bar{X})^2} = 0.1541026$$

and

$$s(b_0) = \{MSE\Big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big]\}^{1/2} = 0.09767$$

$$R^2 = SSR/SST = 0.8484,$$

$$F = \frac{SSR/p}{SSE/(n-p-1)} = 44.76, \quad (\textit{with } DF = 1 \textit{ and } 8)$$

The model is

$$\begin{array}{lcccc} \hat{Y} & = & 0.3780 & + & 1.030974X \\ \text{SE} & & (0.0977) & & (0.15410) \\ \multicolumn{5}{l}{R^2 = 0.8484, \ \hat{\sigma}^2 = 0.01908625, \quad F = 44.76} \end{array}$$

(b) Test $H_0 : \beta_1 = 1$ with $\alpha = 0.05$

$$|t| = |\frac{b_1 - 1}{s(b_1)}| = 0.2010014 < t(1 - 0.025, 8) = 2.7515$$

We accept H_0.

(c) For a new $X = 1$, find the 95% CI for its expected response

$$\hat{Y} = 0.3780 + 1.030974 * 1 = 1.408974$$
$$s(\hat{Y}) = \{MSE\Big[\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\}^{1/2} = 0.07977546$$

The CI is

$$\hat{Y} \pm s(\hat{Y}) * t(1-0.025, 8) = [1.2250, 1.5929]$$

(d) For a new $X = 0.5$, find the 95% prediction interval for its possible response

$$\hat{Y} = 0.8934994,$$
$$s(pred) = \{MSE\Big[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\}^{1/2} = 0.1595317$$

The PI is

$$\hat{Y} \pm s(pred) * t(1-0.025, 8) = [0.5585214, 1.228477]$$

2. For the least square estimator b_0, b_1 of simple linear regression model, find $Cov(b_0, b_1)$

$$\begin{aligned} \mathbf{Var}\{\mathbf{b}\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Var}\{\mathbf{Y}\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i-\bar{X})^2} & \frac{-\bar{X}}{\sum_{i=1}^n (X_i-\bar{X})^2} \\ \frac{-\bar{X}}{\sum_{i=1}^n (X_i-\bar{X})^2} & \frac{1}{\sum_{i=1}^n (X_i-\bar{X})^2} \end{bmatrix} \end{aligned}$$

so, $Cov(b_0, b_1) = -\sigma^2 \frac{\bar{X}}{\sum_{i=1}^n (X_i-\bar{X})^2}$

3. Suppose $A : m \times n$ is a constant matrix and $Y : n \times 1$, is a random vector. Then

$$\mathbf{Var}(AY) = A\mathbf{Var}(Y)A'$$

Please give your proof for $m = 2, n = 3$.

By definition,

$$\begin{aligned} Var(AY) &= E[\{AY - E(AY)\}\{AY - E(AY)\}^\top] \\ &= E[\{A(Y - EY)\}\{A(Y - EY)\}^\top] \\ &= E[A\{(Y - EY)\}\{(Y - EY)\}^\top A^\top] \\ &= AE[\{(Y - EY)\}\{(Y - EY)\}^\top A^\top] \\ &= AE[\{(Y - EY)\}\{(Y - EY)\}^\top]A^\top \\ &= A\mathbf{Var}(Y)A' \end{aligned}$$

4. For multiple linear regression, the normal equations are

$$\sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} e_i X_{i1} = 0$$
$$\vdots$$
$$\sum_{i=1}^{n} e_i X_{ip} = 0$$

Prove that

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

By $1st.Eq \times b_0 + 2nd.Eq \times b_1 + ... + (p+1)th.Eq \times b_p$, we have

$$b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} e_i X_{i1} + \cdots + b_p \sum_{i=1}^{n} e_i X_{ip} = 0$$

i.e.

$$\sum_{i=1}^{n} e_i b_0 + \sum_{i=1}^{n} e_i b_1 X_{i1} + \cdots + \sum_{i=1}^{n} e_i b_p X_{ip} = 0$$

i.e.

$$\sum_{i=1}^{n} e_i (b_0 + b_1 X_{i1} + \cdots b_p X_{ip}) = 0$$

i.e.

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

For each of the following regression models, indicate whether it is a general linear regression model. If not, state whether it can be expressed in the form of a linear regression model after some suitable transformation

$$
\begin{array}{ll}
a. & Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i \\
b. & Y_i = \varepsilon_i \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2), \quad with\ \varepsilon_i > 0 \\
c. & Y_i = \beta_0 \log(\beta_1 X_{i1}) + \varepsilon_i \\
d. & Y_i = \log(\beta_1 X_{i1}) + \beta_2 \log X_{i2} + \varepsilon_i \\
e. & Y_i = [1 + \exp(\beta_0 + \beta_1 X_{i1} + \varepsilon_i]^{-1}
\end{array}
$$

No of these are linear models, but a, b, d e can be transformed to linear regression models

Consider the multiple linear regression models

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,\ i = 1, ..., n$$

where ε_i are uncorrelated with $E\varepsilon_i = 0$ and $E\varepsilon_i^2 = \sigma^2$ state the least square criterion and derive the least squares estimators for β_1 and β_2. Let

$$Q(b_1, b_2) = \sum_{i=1}^{n} \{Y_i - b_1 X_{i1} - b_2 X_{i2}\}^2$$

by calculus, we have

$$
\begin{aligned}
\frac{dQ(b_1, b_2)}{db_1} &= -2 \sum_{i=1}^{n} \{Y_i - b_1 X_{i1} - b_2 X_{i2}\} X_{i1} \\
\frac{dQ(b_1, b_2)}{db_2} &= -2 \sum_{i=1}^{n} \{Y_i - b_1 X_{i1} - b_2 X_{i2}\} X_{i2}
\end{aligned}
$$

The normal equations are

$$
\begin{aligned}
\sum_{i=1}^{n} X_{i1}^2 b_1 + \sum_{i=1}^{n} X_{i1} X_{i2} b_2 = \sum_{i=1}^{n} X_{i1} Y_i \\
\sum_{i=1}^{n} X_{i1} X_{i2} b_1 + \sum_{i=1}^{n} X_{i2}^2 b_2 = \sum_{i=1}^{n} X_{i2} Y_i
\end{aligned}
$$

Solving it we have

$$
\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1} X_{i2} \\ \sum_{i=1}^{n} X_{i1} X_{i2} & \sum_{i=1}^{n} X_{i2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} X_{i1} Y_i \\ \sum_{i=1}^{n} X_{i2} Y_i \end{pmatrix}
$$