Statistics 206

Homework 1 Solution

Due : October 2, 2019, In Class

1. Review Appendix A3 and the matrix notes on canvas under "Files".

2. $\mathbf{Z}$ is an n-dimensional random vector with expectation $\mathbf{E}(\mathbf{Z})$ and variance-covariance matrix:
$$\mathbf{Var}(\mathbf{Z}) = \mathbf{Cov}(\mathbf{Z}, \mathbf{Z}) = \Sigma.$$
A is an $s \times n$ nonrandom matrix and B is a $t \times n$ nonrandom matrix. Show the following:

   (a) $\mathbf{E}(A\mathbf{Z}) = A\mathbf{E}(\mathbf{Z})$.

   Proof.
$$(A\mathbf{Z})_j = \sum_k a_{jk}\mathbf{Z}_k, j = 1, \cdots, s$$
$$(E(A\mathbf{Z}))_j = E((A\mathbf{Z})_j) = E(\sum_k a_{jk}\mathbf{Z}_k) = \sum_k a_{jk}E(\mathbf{Z}_k) = (AE(\mathbf{Z}))_j, j = 1, \cdots, s$$
□

   (b) $\mathbf{Cov}(A\mathbf{Z}, B\mathbf{Z}) = A\Sigma B^T$. So in particular, $\mathbf{Var}(A\mathbf{Z}) = A\Sigma A^T$.

   Proof. Define
$$W = A\mathbf{Z}, U = B\mathbf{Z}, C = \mathrm{Cov}(W, U), D = A\Sigma B^T$$
$$\begin{aligned} C_{ij} &= \mathrm{Cov}(W_i, U_j) \\ &= \mathrm{Cov}(\sum_k a_{ik}\mathbf{Z}_k, \sum_l b_{jl}\mathbf{Z}_l) \\ &= \sum_k \sum_l a_{ik}b_{jl}\mathrm{Cov}(\mathbf{Z}_k, \mathbf{Z}_l) \\ &= \sum_k \sum_l a_{ik}b_{jl}\Sigma_{kl} \\ &= D_{ij}, i = 1, \cdots, s, j = 1, \cdots, t \end{aligned}$$
□

3. Derive the following.

   (a) $\sum_{i=1}^n (X_i - \overline{X}) = 0, \quad \sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - \overline{X})X_i = \sum_{i=1}^n X_i^2 - n(\overline{X})^2$.

Proof.

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \overline{X} = n\overline{X} - n\overline{X} = 0.$$

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + (\overline{X})^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}(\overline{X})^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n(\overline{X})^2 + n(\overline{X})^2 \\
&= \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2.
\end{aligned}$$

□

(b) $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i = \sum_{i=1}^{n} X_iY_i - n\overline{X}\ \overline{Y}$.

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\overline{Y} - \sum_{i=1}^{n} \overline{X}Y_i + \sum_{i=1}^{n} \overline{X}\ \overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - n\overline{X}\ \overline{Y} - n\overline{X}\ \overline{Y} + n\overline{X}\ \overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - n\overline{X}\ \overline{Y}.
\end{aligned}$$

□

4. Least-squares principle.

   (a) State the least-squares principle.

   For a given line: $y = b_0 + b_1x$, the *sum of squared vertical deviations* of the observations $\{(X_i, Y_i)\}_{i=1}^{n}$ from the corresponding points on the line is:

$$Q(b_0, b_1) = \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1X_i)\right)^2.$$

   The *least squares (LS) principle* is to fit the observed data by minimizing the sum of squared vertical deviations.

   (b) Derive the LS estimators for simple linear regression model.

Proof. From the lecture notes, the LS estimators can be derived by finding b_0 and b_1 which satisfy the normal equations.

$$nb_0 + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i ....(1)$$
$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i ....(2)$$

From equation (1), $b_0 = \overline{Y} - b_1 \overline{X}$.
Using this in equation (2) we have

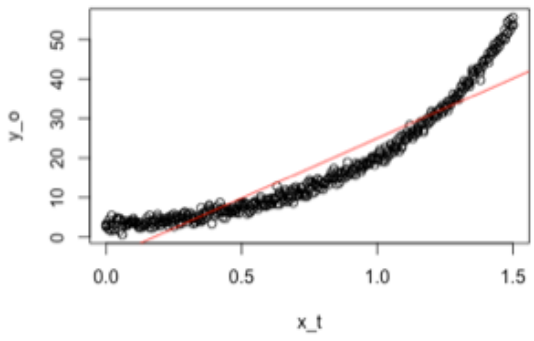
$$b_0 n\overline{X} + b_1 \textstyle\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$
$$\Rightarrow n\overline{XY} + b_1[\textstyle\sum_{i=1}^{n} X_i^2 - n\overline{X}^2] = \sum_{i=1}^{n} X_i Y_i$$
$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

Now from 2(a) and (b), $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$.
From equation (1), $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$. □

(c) Assume the observations follow:

$$Y_i = \exp(a + bX_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

where $a, b \in \mathbb{R}$ are unknown parameters and ϵ_is are uncorrelated random variables with $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$. Describe how to estimate the regression function (equivalently, a, b) by least-squares principle. (Notes: You only need to provide a description. This is an example of a nonlinear regression model.)



Plot implies linear regression model is not accurate for non-linear dataset.

$$Q(a,b) = \sum_{i=1}^{n} (Y_i - \exp(a + bX_i))^2 .$$

The *least squares (LS) principle* is to fit the observed data by minimizing the sum of squared vertical deviations.

$$(\hat{a}, \hat{b}) = argmin_{a,b} \sum_{i=1}^{n} (Y_i - \exp(a + bX_i))^2 .$$

5. Tell true or false (with a brief explanation) of the following statements with regard to simple linear regression.

   (a) The least squares line always passes the center of the data $(\overline{X}, \overline{Y})$.

ANS. True. since $y = \overline{Y} + \hat{\beta}_1(x - \overline{X})$.

   (b) The true regression line fits the data the best.

ANS. False. LS line fits the data better.

   (c) If $\overline{X} = 0$, $\overline{Y} = 0$, then $\hat{\beta}_0 = 0$ no matter what is $\hat{\beta}_1$.

ANS. True since $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

   (d) Given the sample size, the larger the range of X_is, the smaller the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ tend to be.

ANS. True. since $s\{\hat{\beta}_0\}$ and $s\{\hat{\beta}_1\}$ has $\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}$ in the denominator.

   (e) If the correlation coefficient r_{XY} between X_is and Y_is is such that $|r_{XY}| < 1$, then we will observe regression effect.

ANS. TRUE. When $|r_{XY}| < 1$, then one standard deviation change in X would amount to less than one standard deviation change in Y (on average) and such a phenomena is called the regression effect (a.k.a. regression towards the mean).

6. Properties of the residuals under simple linear regression model. Recall that

$$\begin{aligned} e_i &= Y_i - \widehat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i), \quad i = 1, \cdots n. \\ &= (Y_i - \overline{Y}) - \hat{\beta}_1 (X_i - \overline{X}). \end{aligned}$$

Show that

   (a) $\sum_{i=1}^{n} e_i = 0$.

Proof.

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(Y_i - \overline{Y}) - \sum_{i=1}^{n}\hat{\beta}_1(X_i - \overline{X}) \\ &= (\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\overline{Y}) - \hat{\beta}_1(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\overline{X}) \\ &= (n\overline{Y} - n\overline{Y}) - \hat{\beta}_1(n\overline{X} - n\overline{X}) \\ &= 0.\end{aligned}$$

□

(b) $\sum_{i=1}^{n} X_i e_i = 0$.

Proof.

$$\begin{aligned}\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i((Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})) \\ &= (\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i\overline{Y}) - \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i\overline{X}) \\ &= (\sum_{i=1}^{n} X_i Y_i - n\overline{X}\ \overline{Y}) - \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2) \\ &= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) - \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) - \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= 0\end{aligned}$$

□

(c) $\sum_{i=1}^{n} \widehat{Y}_i e_i = 0$.

Proof. By parts (a) and (b), and $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. □

7. Under the simple linear regression model, show the following.

   (a) The LS estimator $\hat{\beta}_0$ is an unbiased estimator of β_0 and derive the formula for its variance.

Proof.

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y}) - E(\hat{\beta}_1\bar{X}) = \frac{1}{n}\sum(\beta_0 + \beta_1 X_i) - \bar{X}E(\hat{\beta}_1) \\ &= \beta_0 + \beta_1\bar{X} - \bar{X}\beta_1 \quad \text{use the fact that } E(\hat{\beta}_1) = \beta_1 \\ &= \beta_0. \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2}\right) \quad \text{use 3(b)} \\ &= \frac{\sigma^2}{\sum(X_i - \overline{X})^2}, \quad \text{as } Y_i\text{s are uncorrelated} \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\overline{Y} - \hat{\beta}_1\overline{X}) \quad \text{use 4(b)} \\ &= \mathrm{Var}(\overline{Y}) + \overline{X}^2\mathrm{Var}(\hat{\beta}_1) \quad \text{use 7(b)} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right) \end{aligned}$$

□

(b) $\overline{Y}$ and $\hat{\beta}_1$ are uncorrelated. (Hint: Write them as linear combinations of Y_is.)

Proof.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \hat{\beta}_1 = \sum \frac{X_i - \bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2} Y_i$$

$$\begin{aligned} Cov(\bar{Y}, \hat{\beta}_1) &= \sum_{i=1}^{n} \frac{1}{n}\frac{X_i - \bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2} Var(Y_i) \quad \text{(by } Y_i\text{s are uncorrelated)} \\ &= \frac{1}{n}\frac{\sum_{i=1}^{n} X_i - n\bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\sigma^2 = 0 \quad \text{since } \sum_{i=1}^{n} X_i - n\bar{X} = 0. \end{aligned}$$

□

8. Under the simple linear regression model:

(a) Show that

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

Proof.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1(X_i - \overline{X})$$
$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \overline{Y} - \hat{\beta}_1(X_i - \overline{X}))^2$$
$$= \sum_{i=1}^n (Y_i - \overline{Y})^2 + \hat{\beta}_1{}^2 \sum_{i=1}^n (X_i - \overline{X})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$$
$$= \sum_{i=1}^n (Y_i - \overline{Y})^2 + \hat{\beta}_1{}^2 \sum_{i=1}^n (X_i - \overline{X})^2 - 2\hat{\beta}_1 . \hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X})^2$$
$$\text{[since} \quad \hat{\beta}_1 \sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})]$$
$$= \sum_{i=1}^n (Y_i - \overline{Y})^2 - \hat{\beta}_1{}^2 \sum_{i=1}^n (X_i - \overline{X})^2$$

□

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

Proof.

$$E(\hat{\beta}_1^2) = \text{Var}(\hat{\beta}_1) + E(\hat{\beta}_1)^2$$
$$= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_1^2$$
$$\text{since} \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$\text{and} \quad E(\hat{\beta}_1) = \beta_1 \quad \text{from lecture notes.}$$

□

(c) Show that

$$E((Y_i - \bar{Y})^2) = \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2),$$

where $\bar{\epsilon} = \frac{1}{n} \sum_{i=1}^n \epsilon_i$.

Proof.

$$E((Y_i - \bar{Y})^2) = E((\beta_0 + \beta_1 X_i + \epsilon_i - \beta_0 - \beta_1 \bar{X} - \bar{\epsilon}))^2$$
$$= E(\beta_1 (X_i - \bar{X}) + \epsilon_i - \bar{\epsilon})^2$$
$$= \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2) + 2\beta_1 (X_i - \bar{X}) E(\epsilon_i - \bar{\epsilon})$$
$$= \beta_1^2 (X_i - \bar{X})^2 + E((\epsilon_i - \bar{\epsilon})^2) \quad \text{as} \quad E(\epsilon_i - \bar{\epsilon}) = E(\epsilon_i) - E(\bar{\epsilon}) = 0$$

□

(d) Show that $E(SSE) = (n-2)\sigma^2$. (Hint: What is $E(\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})^2)$?)

Proof.

$$\begin{aligned}
\mathrm{Var}(\epsilon_i - \bar{\epsilon}) &= \mathrm{Var}(\epsilon_i) + \mathrm{Var}(\bar{\epsilon}) - 2\mathrm{Cov}(\epsilon_i, \bar{\epsilon}) \\
&= \sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n} = \frac{(n-1)\sigma^2}{n} \\
\text{as } \ \mathrm{Cov}(\epsilon_i, \bar{\epsilon}) &= \mathrm{Cov}(\epsilon_i, \frac{\sum_{i=1}^n \epsilon_i}{n}) = \frac{\mathrm{Var}(\epsilon_i)}{n} = \frac{\sigma^2}{n} \\
\therefore E(\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})^2) &= \sum_{i=1}^n E(\epsilon_i - \bar{\epsilon})^2 = \sum_{i=1}^n \mathrm{Var}(\epsilon_i - \bar{\epsilon}) = \frac{n(n-1)\sigma^2}{n} = (n-1)\sigma^2
\end{aligned}$$

From 8(a),

$$\begin{aligned}
E(SSE) &= \sum_{i=1}^n E(Y_i - \overline{Y})^2 - E(\hat{\beta}_1^2) \sum_{i=1}^n (X_i - \overline{X})^2, \ \text{ then from 8(b) and 8(c)} \\
&= \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n E((\epsilon_i - \bar{\epsilon})^2) - (\frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_1^2) \sum_{i=1}^n (X_i - \overline{X})^2 \\
&= \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 + (n-1)\sigma^2 - \sigma^2 - \beta_1^2 \sum_{i=1}^n (X_i - \overline{X})^2 \\
&= (n-2)\sigma^2
\end{aligned}$$

□