Problem Set 6 (two pages)

Statistics 24510-30040 (W19)

Due Tuesday, Feb. 26, at the beginning of class.

Requirements Provide detailed derivations. Select only the relevant part of the output to be inserted. Attach your code or output as an appendix if necessary. Discussions allowed, the assignment should be devised and written by yourself completely.

Problem assignments (Relevant reading in the textbook: Chapter 14.)

1. (*MSE comparison*)
   The Mean Square Error $MSE(\hat{\theta})$ of an estimator $\hat{\theta}$ for a parameter θ is defined as $E(\hat{\theta}-\theta)^2$.
   Let $X_1, \cdots, X_n$ be independent $N(\mu, \sigma^2)$ random variables, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean.
   Consider two estimators of σ^2:
   Sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and the MLE $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.
   Compute and compare $MSE(s^2)$ and $MSE(\hat{\sigma}^2)$.

2. (*Simple linear regression*) (Base on Problem 44, chapter 14 of the text.)
   The file `asthema.txt` at https://www.stat.uchicago.edu/ meiwang/courses/w19stat2/asthma.txt lists respiratory `resistance` and `height` (cm) for children with asthma.
   (a) Find the least squares line.
   (b) Test whether the slope is significantly different from zero.
   (c) Give 95% confidence intervals for the slope and the intercept.
   (d) How much of the variation in the data is explained by the simple linear regression model?
   (e) Show a scatter-plot, a residual plot, and a normal probability plot for the residuals.
   (f) Briefly discuss your results.

3. (*Small derivations*)
   (a) Suppose the model is $Y_i = \mu + \epsilon_i, i = 1, \cdots, n$, ϵ_i are independent with mean zero and variance σ^2. Derive the least square estimate of μ.
   (b) The linear model $Y_i = \beta x_i + \epsilon_i$ with ϵ_i i.i.d. mean zero is fitted to points $(x_i, y_i), i = 1, \cdots, n$. Derive the lease squares estimate of β.
   (c) Suppose that n points $x_1, \cdots, x_n$ are to be placed in the interval $[-1, 1]$ for fitting the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i are independent with mean zero and variance σ^2. How should the x_i be chosen in order to minimize $Var(\hat{\beta}_1)$?
   (d) Fitting a linear model $Y_i = \beta_0 + \beta_1 x_{1i} + +\beta_2 x_{2i} + \epsilon_i$ on n data points yields estimates $\hat{Y}$, $\hat{\beta}_1$, and $\hat{\beta}_2$. Now suppose each x_{1i} is replaced by $x_{1i}^* = k x_{1i}, i = 1, \cdots, n$.
      i. Does $\hat{Y}$ change? If so, what's the new estimate?
      ii. Does $\hat{\beta}_1$ change? If so, what's the new estimate?
      iii. Does $\hat{\beta}_2$ change? If so, what's the new estimate?

(e) Let $X = (X_1, \cdots, X_n)$ be a random n-vector.
Let Y be a random n-vector with components $Y_1 = X_1, Y_i = X_i - X_{i-1}$, $i = 1, \cdots, n$.

i. If the X_i are independent random variables with variance σ^2, find the covariance matrix of Y.

ii. If the Y_i are independent random variables with variance σ^2, find the covariance matrix of X.

4. (*Linear model design matrix and estimators*) (Base on Problem 4, chapter 14 of the text.)

Consider a standard linear regression model in which the freshman GPA(denoted by Y) is modeled to depend linearly on high school GPA (denoted by x). Consider the model

$$Y_i = I_F(i)\beta_F + I_M(i)\beta_M + \beta_1 x_i + e_i, \quad i = 1, \cdots, n$$

where

$$I_F(i) = \begin{cases} 1, & \text{if student } i \text{ is female,} \\ 0, & \text{if otherwise.} \end{cases}, \qquad I_M(i) = \begin{cases} 1, & \text{if student } i \text{ is male,} \\ 0, & \text{if otherwise.} \end{cases}$$

(a) Give the form of the design matrix X for the model.

(b) Find $X^T X$.

(c) Find $(X^T X)^{-1}$.

(d) Find explicit expressions for the least squares estimates of the following.

i. β_F

ii. β_M

iii. β_1

5. (*Model misspecification*) Continuation of previous problem.

Suppose at a particular college that $\beta_1 = 0$, so that the true data-generating mechanism is

$$Y_i = I_F(i)\beta_F + I_M(i)\beta_M + e_i, \qquad e_i \ \ i.i.d. \ \sim N(0, \sigma^2), \qquad i = 1, \cdots, n. \tag{1}$$

In addition, suppose that girls at this college tend to have higher GPAs than boys in both high school and freshman year of college. Now suppose we fit the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad e_i \ \ i.i.d. \ \sim N(0, \sigma^2), \qquad i = 1, \cdots, n. \tag{2}$$

to a representative sample of students from this college (with roughly equal number of male and female students in the sample).

(a) Find an exact expression for $E(\hat{\beta}_1)$ obtained by fitting model (2) by least squares to the data when in fact model (1) is true.

(b) Will this expected value generally be positive, negative or (nearly) 0? Explain.

(c) Comment on what your finding says about how using model (2) rather than the correct (1) could result in misleading conclusions about the effect of high school GPA on college GPA.