STAT120C Homework 5

Assigned Monday May 6th, 2019

Due Tuesday May 14th, 2019 by 5pm in the Dropbox in DBH

1. For a sample of responses Y_i with associated predictors X_i, $i = 1, \ldots, n$, consider the linear regression model

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

where $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

   (a) Show that $\hat{\beta}_0$ derived in class is unbiased for β_0.

   (b) Compute the MLE for the variance $\hat{\sigma}^2$. Write this estimator in terms of the sum of squared error, defined as

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

   (c) What is the bias for $\hat{\sigma}^2$ as an estimator of σ^2?

   (d) What is an unbiased estimator of σ^2? Write this estimator in terms of the SSE.

   (e) Define the *Regression sum of squares* as

$$SSR = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2.$$

   What is $\mathbb{E}[SSR]$?

   (f) Show that the point $(\bar{X}, \bar{Y})$ is on the best fit line determined by $\hat{\beta}_0, \hat{\beta}_1$.

   (g) Use the properties of the normal distribution and the formulas for $\hat{\beta}_0, \hat{\beta}_1$ to explain why these estimators are normally distributed.

2. Now consider the alternative linear regression model with no β_0:

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

with $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

   (a) Find the MLE for β_1.

   (b) Omitting β_0 from the model imposes a particular constraint on the types of lines that can be estimated. Describe the differences between the model that does not include β_0 and the model that does include β_0.

3. A study of lung function in adolescents measured the forced expository volume (FEV) in 1 second from a random sample of 654 children. Recorded covariates included age in years, gender, height in inches, and smoking status (yes/no). In order to understand how lung size is related to FEV, we will consider height as a proxy for lung size, and estimate the relationship of FEV and height using the linear regression model.

   (a) Use the R code for this problem to download the FEV data set and produce the scatterplot of FEV against height.

   (b) Let $Y_i = \text{FEV}$ and $X_i = height$ for the ith subject. Use the R code to compute $\bar{X}$, $\bar{Y}$, $\sum(X_i - \bar{X})^2$, $\sum X_i(Y_i - \bar{Y})$.

   (c) Use the values you just computed to calculate the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (d) Use the R code to plot the best fit line using *ggplot*. (Note that the slope and intercept of this line should match $\hat{\beta}_0$ and $\hat{\beta}_1$ that you calculated.)

   (e) Use the R code to fit the linear regression model using the built-in *lm* function. Include a nicely formatted table of the point estimates and standard errors of the model coefficients.