Linear Regression Models
Statistics GR5205/GU4205 — Fall 2020

Homework 1

The following problems are due on Monday, September 21, 11:59pm.

1. (Problems 1.20 and 1.24 in KNN) The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data in `copier_maintainenance.txt` were collected from 45 recent calls on users to perform routine preventive maintenance service; for the ith call let x_i denote the number of copiers serviced and y_i the total number of minutes spent by the service person, for $i = 1, 2, \ldots, n = 45$.

   (a) Plot the data and overlay a *lowess* smoother. Does it seem that the simple linear regression model

   $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   is appropriate? Explain.

   (b) Obtain the least squares estimated linear regression function, and overlay *it* on a scatterplot of the data. How well does the estimated regression function fit the data?

   (c) Interpret b_1 in your estimated regression function.

   (d) Interpret b_0 in your estimated regression function. Does b_0 provide any relevant information here? Explain.

   (e) Obtain a point estimate of the mean service time for calls on which $x = 5$ copiers are serviced.

   (f) Obtain a point prediction for the service time of a single call on which $x = 5$ copiers are to be serviced.

   (g) Obtain the residuals $e_i = y_i - (b_0 + b_1 x_i)$ and confirm that they sum to zero. Explain the relation between the sum of squared residuals and the quantity

   $$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \ .$$

   (h) Obtain point estimates of $\sigma^2 = \mathrm{var}(\varepsilon_i)$ and σ. In what units is σ expressed?

2. (Problem 1.45 in KNN) The primary objective of the Study on the Efficacy of Nosocomial Infection Control (SENIC Project) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. The data set in `SENIC.txt` consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed. Each line of the data set has an identification number `ID` and provides information on 11 other variables for a single hospital.

   The average length of a stay in a hospital (`Stay`) is anticipated to be related to infection risk `Risk`, available facilities and services `AFS`, and routine chest X-ray ratio `Xray`. (See Appendix C.1 for details on these and the other variables included in the data set.)

   (a) Obtain scatterplots of average length of stay against each of the three predictor variables, and overlay *lowess* smoothers. Does a linear mean function seem plausible in each case? Explain.

   (b) Obtain the least squares estimate for the linear regression of average length of stay on each of the three predictor variables, and overlay the least squares lines on your scatterplots. Does the simple linear regression model seem plausible in each case? Explain.

   (c) Calculate *MSE* for each of the three linear regression fits. Which predictor variable leads to the smallest variability around the fitted regression line? Was this result apparent from your plots in parts (a) and (b)? Explain.

3. **Properties of the Least Squares Estimation**

Prove the following properties of the least squares estimated regression function $\hat{y} = b_0 + b_1 x$ where

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$b_0 = \bar{y} - b_1 \bar{x} \; .$$

(a) The sum of the residuals $e_i = y_i - (b_0 + b_1 x_i)$ is zero:

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

(b) The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i = b_0 + b_1 x_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

(c) The sum of the weighted residuals, weighted by the values of the predictor variable, is zero:

$$\sum_{i=1}^{n} x_i e_i = 0$$

(d) The sum of the weighted residuals, weighted by the fitted values, is zero:

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

(e) The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.

4. **Conditional Expectation as Minimum Mean Squared Error Estimator**

(from lecture notes of STAT 901 at the University of Waterloo, by Prof. Don McLeish)

Let $(X, Y) \sim p(x, y)$. Suppose $\mathbb{E}[X^2 + Y^2] < \infty$.

(a) What constant is consider to be the best fit to a random variable in the sense of smallest mean squared error? In other words, what is the value of c solving

$$\min_c \mathbb{E}\left[(Y - c)^2\right].$$

(b) Show that for any function g,

$$\mathbb{E}\left[(Y - \mathbb{E}\left[Y|X\right])^2\right] \leq \mathbb{E}\left[(Y - g(X))^2\right].$$

Hint. Consider $\mathrm{Var}[Y|X = x] = \mathbb{E}\left[(Y - \mathbb{E}\left[Y|X\right])^2 |X = x\right].$