Homework 2

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia University

Octover 23, 2020

Question 1

We are considering the linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where $\hat{y}$ is the estimated service time for a call, x is the number of copiers being serviced, and $\epsilon \sim \mathrm{N}(0, \sigma^2)$. The least-squares estimator model is given by

$$\hat{y} = -0.5802 + 15.0352x \tag{1}$$

Throughout this question, we will be using a variety of base R functions to easily obtain the desired measurements.

(a) The 95% confidence interval for the mean service time when there are six copiers is given by

$$\mathrm{E}[y] \in \Big(86.8152, 92.44746\Big).$$

Intuitively, this means that there are six copiers being serviced, we are 95% sure that the average service time for *all* service times falls within this range.

(b) The 95% prediction interval for the next service time when there are six copiers is

$$\widehat{y} \in \Big(71.43628, 107.8264\Big).$$

As expected, we notice that the prediction interval is significantly wider than the confidence interval.

(c)

(d) The ANOVA table has been printed in Table 1.

(e) To determine if there is any linear relationship between x and y, we conduct an F-test, where $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. From Table 1, we see that the associated p-value is well below the significance level $\alpha = 0.05$, and so we reject H_0. The data seems to indicate that there is in fact a linear relationship between X and Y.

(f) The total variance explained by the model is known as the R^2 value, and is given by

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{76960}{80376} \approx 0.9575.$$

That is, bout 95.7% of Y's variation is explained by model (1), quite a significant reduction.

Source of Variation	df	Sum of Squares	Mean Square	f	$\Pr(> f)$
Copiers	1	76960	76960	968.66	$< 2.2 \times 10^{-16}$
Residuals	43	3416	79	–	–
Total	44	80376	–	–	–

Table 1: The ANOVA table for model (1).

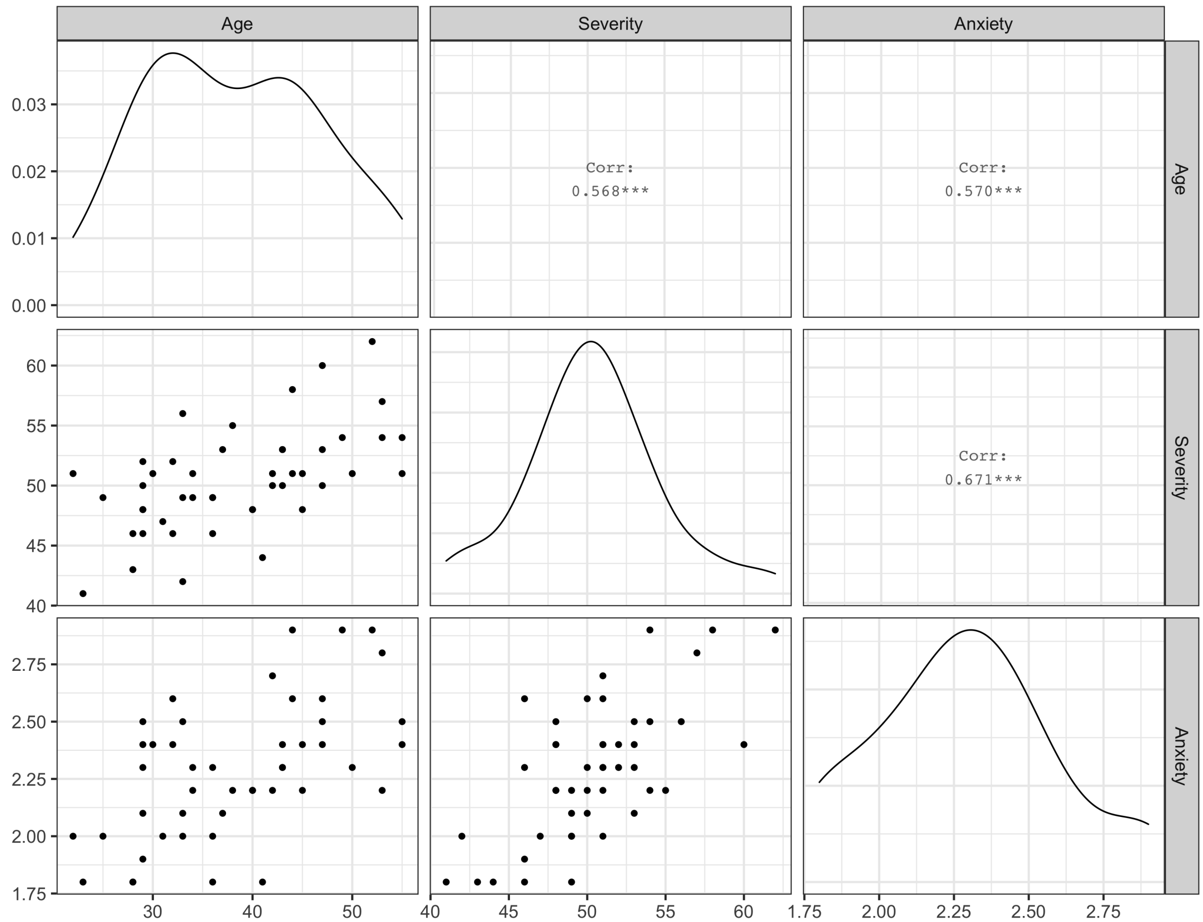

Figure 1: The scatterplot matrix for `Age`, `Severity`, and `Anxiety`.

Question 2

(a) The scatterplot matrix has been printed in Figure 1. We can see that there does seem to be a positive correlation between all three of the variables.

(b) Let $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^n$ denote `Patient Satisfaction`, `Age`, `Severity`, and `Anxiety`, respectively, and let $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix} \in \mathbb{R}^{n \times 4}$. The multiple regression model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. It is worth emphasizing that $\boldsymbol{\epsilon}$ and $\mathbf{y}$ are random vectors, while $\mathbf{X}$ and $\boldsymbol{\beta}$ are fixed. The least-squares estimate for $\boldsymbol{\beta}$ (where each entry is rounded to two decimal places) is given by

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 158.49 \\ -1.14 \\ -0.44 \\ -13.47 \end{bmatrix},$$

and our estimated model is given by $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$. What this means is that, when holding all other variables constant, increasing `Severity` by one unit will cause `Satisfaction` to *decrease* by 0.44.

(c)