Homework 3

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia University

Octover 23, 2020

Throughout this assingment, we will be using a variety of base R functions to easily obtain the desired measurements. In addition, all measurements will be rounded to two decimal places or less.

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

$$\hat{y} = -0.58 + 15.04x \tag{1}$$

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

$$\mathrm{E}[y] \in \big(86.81, 92.45\big).$$

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

$$\widehat{y} \in \big(71.44, 107.83\big).$$

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

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

That is, about 95.75% 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).

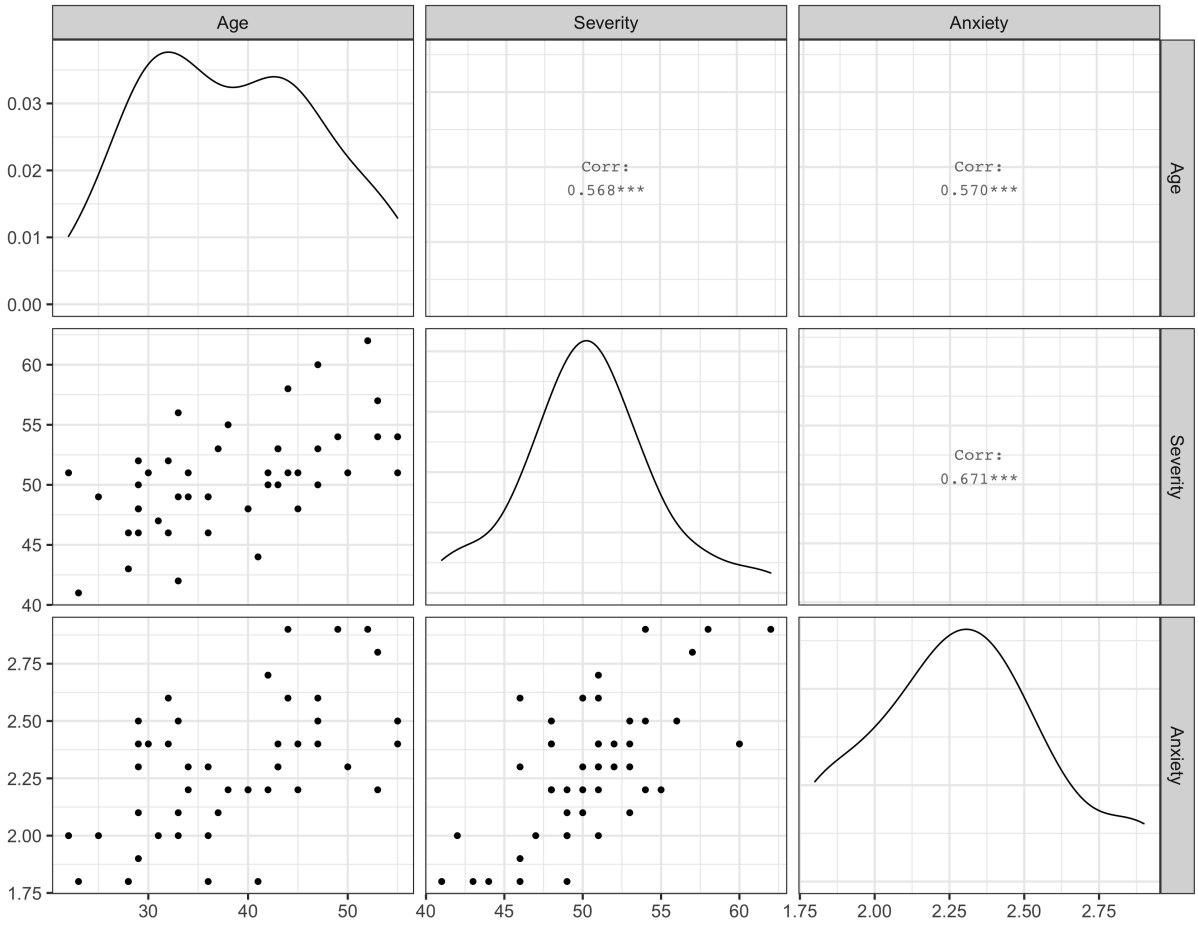

Figure 1: The scatterplot matrix for Age, Severity, and Anxiety.

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 $\boldsymbol{\beta}$ 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) A plot of the residuals against the response and each of the predictors can be found in Figure 2. For each, we see that the residuals appear to be centered around 0. We also see that there is no underlying pattern in the residuals as either of the four variables increases. This indicates that none of the model assumptions are violated.

(d) A Q-Q plot of the residuals can be found in Figure 3. The upper tail of the Q-Q plot greatly deviates from the theoretical values, which could mean that the residuals may not be normally distributed.

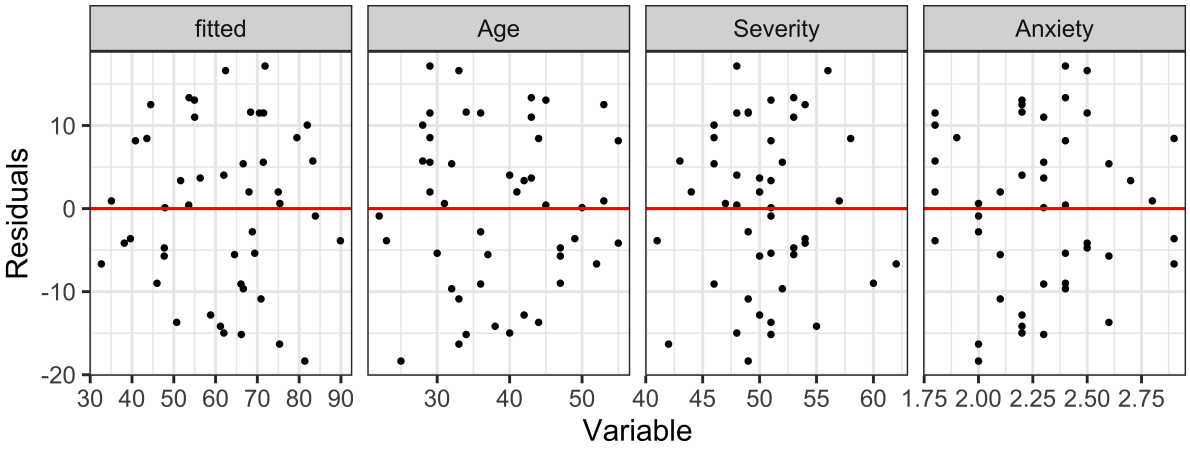

Figure 2: The residuals plotted against $\hat{y}$ and each of the predictors.

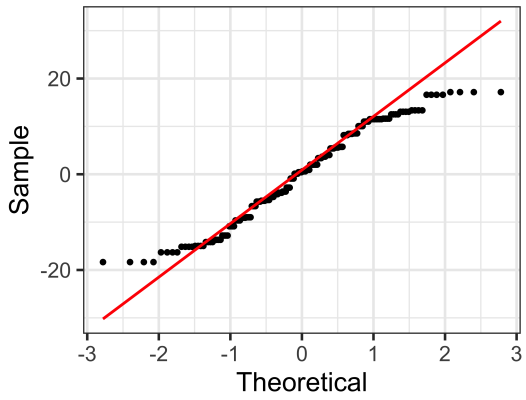

Figure 3: A Q-Q plot of the residuals for model (2).

Question 3

We are assuming that model (2) is appropriate.

(a) To see if there is *any* linear relationship between Satisfaction and Age, Severity and Anxiety, we will conduct an F-test, where $H_0 : \boldsymbol{\beta} = \mathbf{0}$ against $H_a : \boldsymbol{\beta} \neq \mathbf{0}$. Our p-value is given by $p \approx 1.542 \times 10^{-10}$, which is significantly lower than $\alpha = 0.05$. Therefore, we reject H_0; the data indicates that there is a linear relationship.

(b) A joint 90% (not 95%) confidence interval for each of the predictor coefficients is given by

$$\begin{aligned}\beta_1 &\in \big(-1.50, -0.78\big),\\ \beta_2 &\in \big(-1.27, 0.39\big),\\ \beta_3 &\in \big(-25.41, -1.53\big).\end{aligned}$$

Of the three intervals, we see that the interval for β_2 contains zero at this significance level, while the other two intervals do not. That is, we are 90% sure that $\beta_1, \beta_3 \neq 0$ and $\beta_2 = 0$ (so overall, $\boldsymbol{\beta} \neq \mathbf{0}$). We believe that Age and Anxiety have a significant effect on Satisfaction, while Severity does not.

(c) The R^2 value here is $R^2 = 0.6822$, which means about 68.22% of the response's variability is explained by the model, meaning the model is decent.

(d) When a patient is 35 years old, has a severity rating of 45, and an anxiety rating of 2.2, a 95% confidence interval for the average satisfaction rating is $\big(63.63, 74.39\big)$. This means we are 95% sure that for all patients who meet this criteria, the average of all of their satisfaction ratings will fall in this interval.

(e) Given the same conditions, a 95% prediction interval is given by $\big(48.012, 90.01\big)$. We are 95% sure that the satisfaction rating of the next patient we observe will fall in this interval.

Question 4

Let

$$\begin{bmatrix} X_1 \\ X_2 \\ \epsilon \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 & 0 \\ \rho\sigma^2 & \sigma^2 & 0 \\ 0 & 0 & \tilde{\sigma}^2 \end{bmatrix} \right).$$

That is, $X_1 \sim \mathrm{N}(\mu_1, \sigma^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma^2)$ (i.e. X_1 and X_2 have different means but the same variance), $\epsilon \sim \mathrm{N}(0, \tilde{\sigma}^2)$, $\mathrm{Cor}[X_1, X_2] = \rho$ (which makes $\rho\sigma^2 = \mathrm{Cov}[X_1, X_2]$), and $\mathrm{Cov}[X_1, \epsilon] = \mathrm{Cov}[X_2, \epsilon] = 0$. We now consider $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, a linear combination.

(a) Since Y is the sum of normal random variables, Y must also be normally distributed. It's mean and variance are given by

$$\begin{aligned}\mathrm{E}[Y] &= \mathrm{E}[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon] = \beta_0 + \beta_1 \mathrm{E}[X_1] + \beta_2 \mathrm{E}[X_2] + \mathrm{E}[\epsilon] \\ &= \beta_0 + \beta_1 \mu_1 + \beta_2 \mu_2, \\ \mathrm{Var}[Y] &= \mathrm{Var}[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon] = \beta_1^2 \mathrm{Var}[X_1] + 2\beta_1\beta_2 \mathrm{Cov}[X_1, X_2] + \beta_2^2 \mathrm{Var}[X_2] + \mathrm{Var}[\epsilon] \\ &= \sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big) + \tilde{\sigma}^2.\end{aligned}$$

(b) We first expand the numerator to get

$$\begin{aligned}\mathrm{Cov}[X_1, Y] &= \mathrm{Cov}[X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon] = \beta_1 \mathrm{Var}[X_1] + \beta_2 \mathrm{Cov}[X_1, X_2] = \beta_1\sigma^2 + \rho\beta_2\sigma^2 \\ &= \sigma^2\big(\beta_1 + \rho\beta_2\big),\end{aligned}$$

and so

$$r_1 \coloneqq \frac{\mathrm{Cov}[X_1, Y]}{\sqrt{\mathrm{Var}[X_1] \cdot \mathrm{Var}[Y]}} = \frac{\sigma^2\big(\beta_1 + \rho\beta_2\big)}{\sqrt{\sigma^2\Big(\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big) + \tilde{\sigma}^2\Big)}} = \frac{\sigma\big(\beta_1 + \rho\beta_2\big)}{\sqrt{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big) + \tilde{\sigma}^2}}.$$

(c) We have

$$\begin{aligned}\text{Cov}[\beta_1 X_1 + \beta_2 X_2, Y] &= \beta_1 \text{Cov}[X_1, Y] + \beta_2 \text{Cov}[X_2, Y] = \sigma^2\Big(\beta_1(\beta_1 + \rho\beta_2) + \beta_2(\rho\beta_1 + \beta_2)\Big) \\ &= \sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)\end{aligned}$$

and

$$\text{Var}[\beta_1 X_1 + \beta_2 X_2]\beta_1^2 \text{Var}[X_1] + 2\beta_1\beta_2 \text{Cov}[X_1, X_2] + \beta_2^2 \text{Var}[X_2] = \sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)$$

(which means $\text{Cov}[\beta_1 X_1 + \beta_2 X_2, Y] = \text{Var}[\beta_1 X_1 + \beta_2 X_2]$). Plugging this in gives us

$$r_2 \coloneqq \frac{\text{Cov}[\beta_1 X_1 + \beta_2 X_2, Y]}{\sqrt{\text{Var}[\beta_1 X_1 + \beta_2 X_2] \cdot \text{Var}[Y]}} = \frac{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)}{\sqrt{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big) \cdot \Big(\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big) + \tilde{\sigma}^2\Big)}}$$

$$= \sqrt{\frac{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)}{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big) + \tilde{\sigma}^2}}$$

(d) For notational ease, we are going to substitute $\text{Var}[Y]$ back into r_1 and r_2. We first note that

$$r_1^2 = \frac{\sigma^2(\beta_1 + \rho\beta_2)^2}{\text{Var}[Y]} \quad \text{and} \quad r_2^2 = \frac{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)}{\text{Var}[Y]}.$$

We will start with the assumption $r_1^2 \leq r_2^2$, and then manipulate the inequality until we get a result that will always be true:

$r_1^2 \leq r_2^2$	(iniital assumption)
$\frac{\sigma^2(\beta_1 + \rho\beta_2)^2}{\text{Var}[Y]} \leq \frac{\sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)}{\text{Var}[Y]}$	(substituting in)
$\sigma^2(\beta_1 + \rho\beta_2)^2 \leq \sigma^2\big(\beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2\big)$	(multiply both sides by $\text{Var}[Y]$)
$(\beta_1 + \rho\beta_2)^2 \leq \beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2$	(divide both sides by σ^2)
$\beta_1^2 + 2\rho\beta_1\beta_2 + \rho^2\beta_2^2 \leq \beta_1^2 + 2\rho\beta_1\beta_2 + \beta_2^2$	(expand left-hand side)
$\rho^2\beta_2^2 \leq \beta_2^2$	(subtract $\beta_1^2 + 2\rho\beta_1\beta_2$ from both sides)
$\rho^2 \leq 1$	(dibide both sides by β_2^2)
$\rho \leq 1$	(take the square root of both sides)

Since ρ is a correlation, it must be bounded between -1 and 1, so the final condition is always true. As a result, we will always have $r_1^2 \leq r_2^2$.