Homework 1

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia Univeristy
September 21, 2020

Question 1

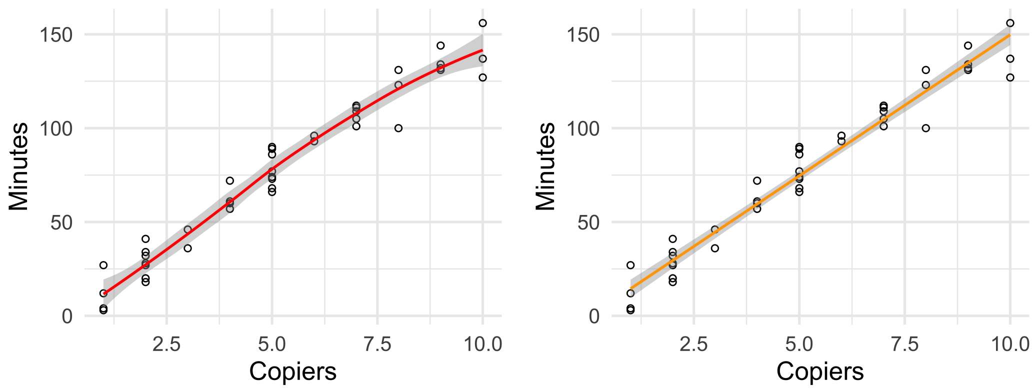

Figure 1: Left: overlaying a loess smoother to a scatterplot of the data. Right: overlaying the estimated linear regression model to the scatterplot.

(a) Not sure what she wants here, ask more about LOWESS smoothers.

(b) Using R, our estimated coefficients are given by $b_0 = -0.5801567$ and $b_1 = 15.0352480$, and so our estimated linear regression function is given by

$$\hat{Y} = -0.5801567 + 15.0352480X.$$

The estimated linear regression model has been overlayed on a scatterplot of the data in the right plot in Figure 1, and the estimated function seems to fit the data well. The general trend, where an increase in number of copiers results in an increased number of minutes on call, is captured by the model.

(c) b_1 can be interpreted as follows. If the number of copiers serviced during a call increased by one, the total number of minutes of the call is expected to *increase* by 15.0352480 minutes.

(d) b_0 can be interpreted as follows. If there are zero copiers serviced during a call, then we can expect the call to last for, on average, -0.5801567 minutes. This does *not* provide any useful or relevant information; a call cannot ever have negative time, and a customer would never call if they did not have any copiers to service (where $X = 0$).

(e)

(f)

(g) Using R, we can see that the residuals sum to 0; this is easy to do since the residuals are are included in the fitted model. We can think of Q as a *function* of β_0 and β_1, and we want to find the values of β_0 and β_1 that minimize Q. The observed residuals e_i, when plugged into Q, give the smallest value of Q that can possibly be obtained. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be the random vector containing the n residuals, and let $\mathbf{e} = (e_1, \ldots, e_n)^T$ be the n realized residuals from b_0 and b_1. Using this notation, we have $Q = \|\boldsymbol{\varepsilon}\|_2^2$, and

$$\|\mathbf{e}\|_2^2 = \min \|\boldsymbol{\varepsilon}\|_2^2 = \min Q.$$

Question 2 For this question, let `Stay` denote a patient's average stay in the hospital, `Risk` denote a patient's risk of infection, `AFS` denote the hospital's available facilities and services, and `Xray` denote a patient's routine chest X-ray ratio.

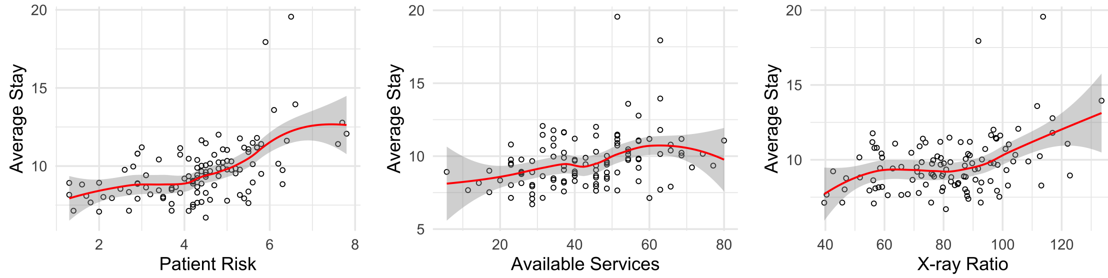

Figure 2: Applying LOESS smoothers to the scatterplots of `Stay` against the three predictor variables: `Risk`, `AFS`, and `Xray`.

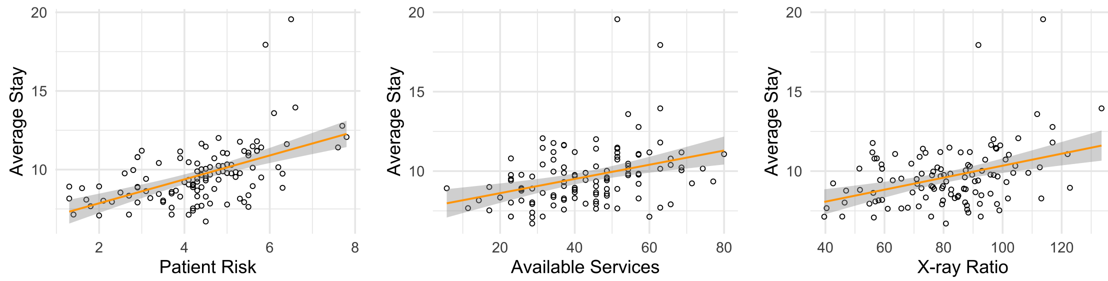

Figure 3: Overlaying the estimated linear regression function for `Stay` against each of the three predictor variables.

(a) The loess smoothers have been overlayed each of their respective scatterplots in Figure 2. We see that in each case the three curves are not too volatile, so a linear regression model would not be completely out of the question. However, for `Risk` we see that the curve begins to flatten out as we approach the end of the interval, and for `AFS` the curve begins to descend, both indications that the relationship is nonlinear. For `Xray`, while the curve is ascending at both the beginning and the end, it is completely flat in the middle, another indication that the relationship is nonlinear.

(b) For each predictor, the linear regression model was fit using `R` and overlayed on a scatterplot of `Stay` against that predictor. While it is impossible for a linear model (or any model, for that matter) to account for the variance of the residuals, we can see that in all three cases, the slope of the regression line is positive. However, the lines are not steep, which indicates that, while there may be a positive linear relationship, it is not a strong one. Numerical details about each model can be found in part (c).

(c) We use `R` to determine the estimated coefficients, mean squared error (MSE), and R^2 value for each of the three models, all of which can be found in Table 1. Recall that MSE $= \frac{1}{n}\|\mathbf{y} - b_0\mathbf{1} - b_1\mathbf{x}\|_2^2$, where $(\mathbf{x}, \mathbf{y})$ are the vectors corresponding to the realized predictor and response variables, respectively. For a given model, a lower MSE (relative to the other models) indicates that the model is a better fit than the others, since the residuals are smaller. In this case, we can see that the lowest MSE occurs when `Risk` is the predictor variable, which means that the residuals for the `Risk` model are generally lower than the other two. However, because the MSE for `Risk` is only marginally smaller than the other two, the residuals are not *that* much smaller; Figure 3 serves as a gut check for this, as the points spread out in all three plots, so it would not be obvious that `Risk` has the lowest MSE. We also see that `Risk` has

X	b_0	b_1	MSE	R^2
`Risk`	6.3368	0.7604	2.590837	0.2846
`AFS`	7.71877	0.04471	3.163568	0.1264
`Xray`	6.566373	0.037756	3.091558	0.1463

Table 1: Information about each of the three linear regression models for question 2.

the highest R^2 value; it is much higher relative to the other two models, but is still extremely low in its own right. That is, even though `Risk` does a much better job than the other models explaining the variability in `Stay` than `AFS` or `Xray`, it still does a pretty bad job overall.

Question 3 For reference, the coefficients for the estimated linear regression function $\hat{Y} = b_0 + b_1 X$ are given by

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x}.$$

The two *normal equations* are given by

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i\,(y_i - b_0 - b_1 x_i) = 0$$

We also note three useful manipulations: $\sum x_i = n\bar{x}$, $\sum y_i = n\bar{y}$, and $\bar{y} = b_0 + b_1\bar{x}$. The first two come from manipulating the definition of $\bar{x}$ and $\bar{y}$, respectively, and the third comes from manipulating the definition of b_0.

(a) This is an immediate result of the first normal equation:

$$0 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} e_i.$$

(b) This can be show directly:

$$\sum_{i=1}^{n}\hat{y}_i = \sum_{i=1}^{n}(b_0 + b_1 x_i) = b_0\sum_{i=1}^{n}1 + b_1\sum_{i=1}^{n}x_i = nb_0 + nb_1\bar{x} = n(b_0 + b_1\bar{x}) = n\bar{y} = \sum_{i=1}^{n}y_i.$$

(c) This is an immediate result of the second normal equation:

$$0 = \sum_{i=1}^{n} x_i\,(y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} x_i e_i.$$

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ be the vector of predictor observations and $\mathbf{e} = (e_1, \ldots, e_n)^T$ be the vector of residuals. This results implies that $\langle \mathbf{x}, \mathbf{e} \rangle = 0$, meaning $\mathbf{x}$ and $\mathbf{e}$ are *orthogonal*.

(d) Combining the results of parts (a) and (c), we have

$$\sum_{i=1}^{n}\hat{y}_i e_i = \sum_{i=1}^{n}(b_0 + b_1 x_i)\,e_i = b_0\sum_{i=1}^{n}e_i + b_1\sum_{i=1}^{n}x_i e_i = b_0 \cdot 0 + b_1 \cdot 0 = 0.$$

If $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ is the vector of fitted values, then $\langle \hat{\mathbf{y}}, \mathbf{e} \rangle = 0$.

(e) We want to show that $\hat{y}_i = \bar{y}$ when $x_i = \bar{x}$. This is actually an immediate result of the definition of b_0 (the third manipulation), since $\hat{y}(\bar{x}) = b_0 + b_1\bar{x} = \bar{y}$.

Question 4 Let $(X, Y) \sim p(x, y)$ with $\mathbb{E}[X^2 + Y^2 < \infty]$.

(a) We want to find the value of c that minimizes $\mathbb{E}\left[(Y-c)^2\right]$. Expanding out the inside gives us

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

Given that this is a function of c, we will now differentiate this equation with respect to c, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}c}\Big(\mathbb{E}\left[Y^2\right] - 2c\mathbb{E}\left[Y\right] + c^2\Big) = -2\mathbb{E}[Y] + 2c \overset{\text{set}}{=} 0,$$

and solving for c gives us $c = \mathbb{E}[Y]$.

(b) Using the subtle identity $0 = \mathbb{E}[Y|X] - \mathbb{E}[Y|X]$, we have

$$\begin{aligned}
\big(Y - g(X)\big)^2 &= \big(Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - g(X)\big)^2 \\
&= \Big(\big(Y - \mathbb{E}[Y|X]\big) + \big(\mathbb{E}[Y|X] - g(X)\big)\Big)^2 \\
&= \big(Y - \mathbb{E}[Y|X]\big)^2 - 2\big(Y - \mathbb{E}[Y|X]\big)\big(\mathbb{E}[Y|X] - g(X)\big) + \big(\mathbb{E}[Y|X] - g(X)\big)^2.
\end{aligned}$$

Taking the expected value of this gives us

$$\begin{aligned}
&\mathbb{E}\left[\big(Y - g(X)\big)^2\right] \\
&= \mathbb{E}\left[\big(Y - \mathbb{E}[Y|X]\big)^2 - 2\big(Y - \mathbb{E}[Y|X]\big)\big(\mathbb{E}[Y|X] - g(X)\big) + \big(\mathbb{E}[Y|X] - g(X)\big)^2\right] \\
&= \mathbb{E}\left[\big(Y - \mathbb{E}[Y|X]\big)^2\right] + \mathbb{E}\left[-2\big(Y - \mathbb{E}[Y|X]\big)\big(\mathbb{E}[Y|X] - g(X)\big)\right] + \mathbb{E}\left[\big(\mathbb{E}[Y|X] - g(X)\big)^2\right]
\end{aligned}$$