Homework 4

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia University

November 9, 2020

Question 1

Collaborators: None

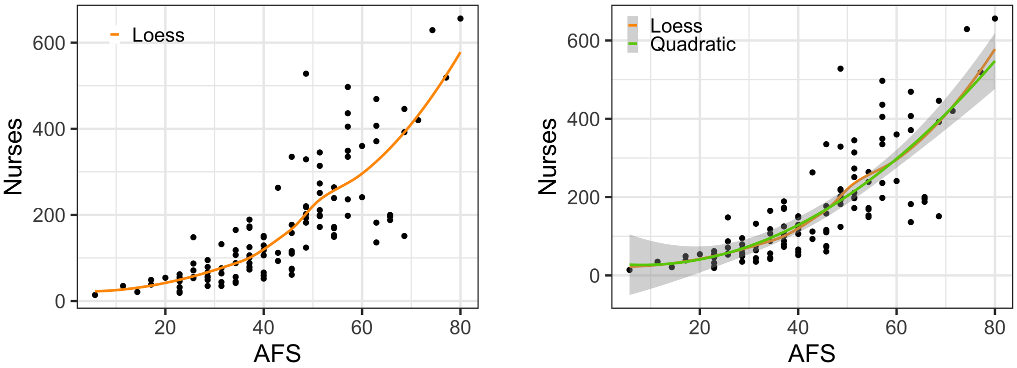

Figure 1: Overlaying a loess smoother and a quadratic polynomial to the scatterplot of AFS vs. Nurses. The left plot displays only the loess smoother, while the right plot displays both.

(a) The loess smoother is the blue line displayed in Figure 1. For clarity, the left plot displays only the loess smoother. The loess smoother indicates that a linear function does not seem plausible for this data, as the curve is too non-linear.

(b) The fitted quadratic model is given by $\hat{Y} = 33.548 - 1.666x + 0.101x^2$, and can be see in the right panel of Figure 1. As we can see, the quadratic model almost perfectly overlays the loess smoother, further indicating that a linear relationship is not plausible.

(c) Given our model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$, to determine if the quadratic term can be dropped from the model, we will test $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$. The p-value for this test is 0.00032, so for any reasonable significance level, we will reject H_0. It seems that there is a non-negligible quadratic relationship. Interestingly, if one were to re-run the same test for β_0 and β_1, the corresponding p-values would be 0.515 and 0.495, which would not give us grounds to reject either of those null hypotheses. That is, we could say the slope and linear term of this model are insignificant, but the quadratic term is not.

(d) When AFS is 30%, a 95% prediction interval is $\big(-89.781, 239.004\big)$, and when AFS is 60%, a 95% prediction interval is $\big(133.136, 462.503\big)$. The simultaneous confidence level for both of these intervals is $0.95^2 = 0.9025$, so we are 90.25% confident that both of these predictions will occur.

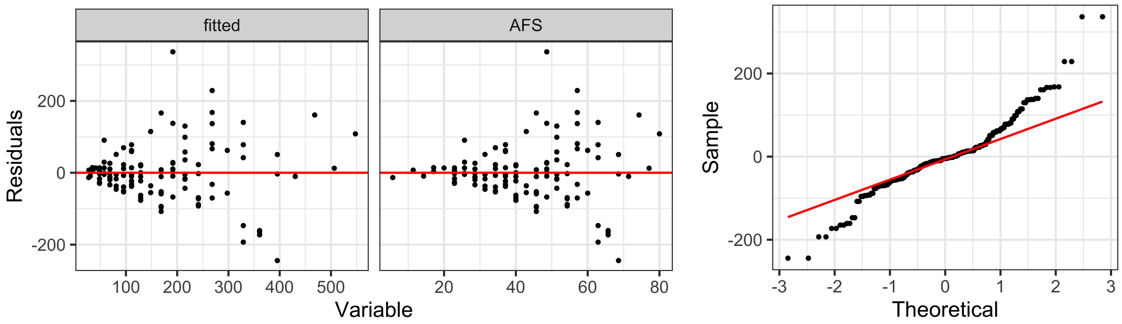

Figure 2: Diagnostic plots for the quadratic model.

(e) The diagnostic plots have been printed in Figure 2. We can see right away that the assumption of equal variance among the residuals is violated. For both residual plots, the residuals create a cone-like pattern, where the variance in the residuals starts small, and then increases. In addition, there are several significant outliers in the QQ-plot, meaning that the assumption of normally-distributed residuals is most likely being violated as well.

Question 2

Collaborators: None

(a) See the attached `R` code.

(b) The scatterplot matrix can be seen in Figure 3. It seems that there is a difference between hospitals with or without a medical school affiliation.

(c) Let X_1, X_2, x_3, and X_4 correspond to Stay, Age, Xray, and MS, respectively. We are considering two models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4, \tag{1}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_4 + \beta_6 X_2 X_4 + \beta_7 X_3 X_4. \tag{2}$$

Model (1) is a linear model with the four predictors, and model (2) has the same linear terms plus three interaction terms. For notational ease, let $\tilde{\boldsymbol{\beta}} = (\beta_5, \beta_6, \beta_7)^T$, i.e. $\tilde{\boldsymbol{\beta}}$ is a vector that contains the coefficients of the interaction terms. We want to test that the interaction terms are negligible, so we will test $H_0 : \tilde{\boldsymbol{\beta}} = \mathbf{0}$ against $H_a : \tilde{\boldsymbol{\beta}} \neq \mathbf{0}$. Using `R`, we can easily fit a linear model both with and without interactions. In addition, we can conduct an F-test to get a p-value of 0.1539, so we fail to reject the null hypothesis. The data seems to indicate that the interaction terms are negligible in the model.

(d) Using model (1), a 95% confidence interval for β_4 is $\big(-0.320, 0.896\big)$. Since zero is withing this confidence interval, we can conclude that the effect of medical school affiliation is negligible in this model.

Question 3

Collaborators: None

(a) A scatterplot matrix has been printed in Figure 4. One thing we can see is that Cen and AFS are highly-correlated.

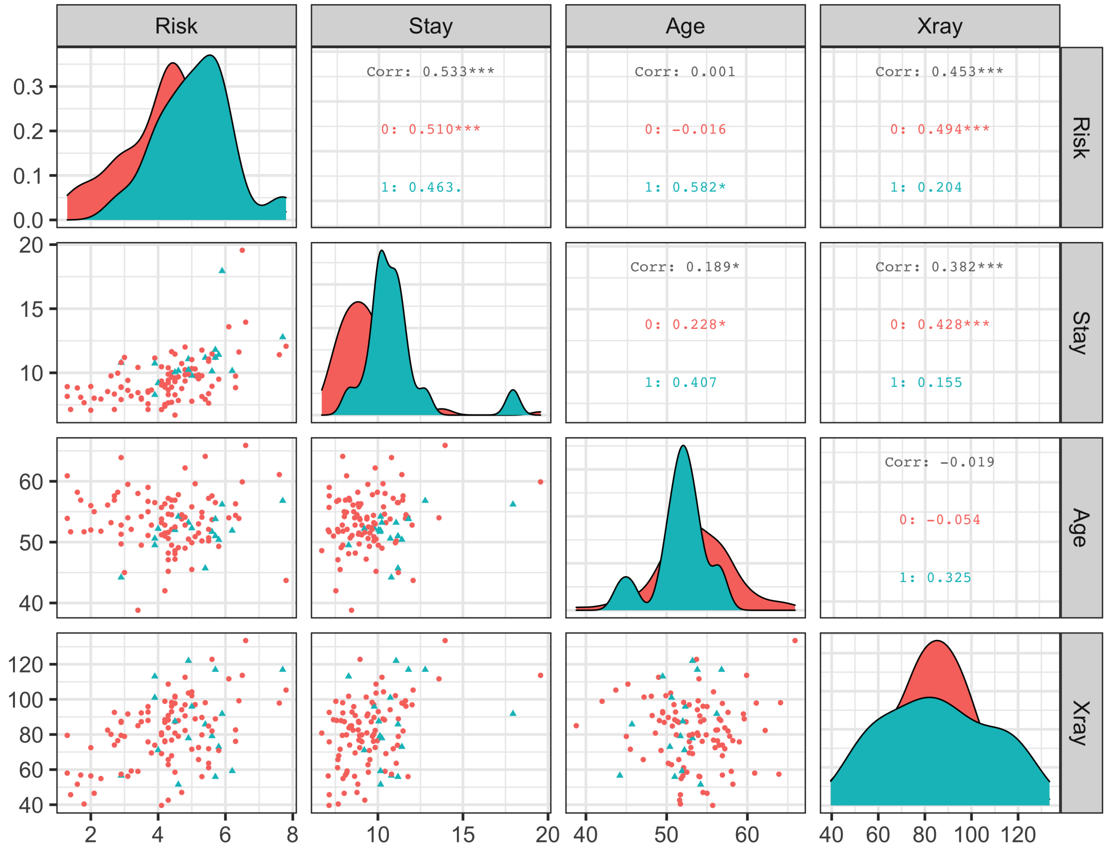

Figure 3: A scatterplot matrix of Risk against Stay, Age, and Xray. Here, hospitals without a medical school affiliation are labeled as red circles, and those with a medical school affiliation are labeled as blue triangles.

(b) Let $X_1, \ldots, X_4$ denote Age, Cult, Cen, and AFS, respectively. Our four linear models are given by

$$\hat{Y} = 4.198 + 0.104X_1 + 0.0403X_2 + 0.00660X_3 - 0.0208X_4 \tag{3a}$$
$$\hat{Y} = 3.238 + 0.104X_1 + 0.0403X_2 + 0.00660X_3 - 0.0208X_4 \tag{3b}$$
$$\hat{Y} = 2.681 + 0.104X_1 + 0.0403X_2 + 0.00660X_3 - 0.0208X_4 \tag{3c}$$
$$\hat{Y} = 2.048 + 0.104X_1 + 0.0403X_2 + 0.00660X_3 - 0.0208X_4 \tag{3d}$$

Model (3a) is for when we are in region 1, (3b) for region 2, and so on.

(c) From our four models, we have $\hat{\beta}_2 = 0.0403$. A 99% confidence interval for β_2 is $\big(0.00278, 0.0778\big)$.

(d) We can alternatively write our four separate estimated models as a single model using indicator functions. Let $X_5 = \mathbb{I}(\text{Region} = 2)$, $X_6 = \mathbb{I}(\text{Region} = 3)$, and $X_7 = \mathbb{I}(\text{Region} = 4)$. Then our linear function takes the form

$$Y = \beta_0 + \sum_{i=1}^{7} \beta_i X_i \tag{4}$$

In this form, the estimated model parameters are given by $\hat{\beta}_0 = 4.198$, $\hat{\beta}_5 = -0.960$, $\hat{\beta})_6 = -1.517$, and $\hat{\beta}_7 = -2.150$. Letting $\tilde{\boldsymbol{\beta}} = (\beta_5, \beta_6, \beta_7)^T$, if we want to test whether or not the average stay does not depend on the region, we will test $H_0 : \tilde{\boldsymbol{\beta}} = \mathbf{0}$ against $H_a : \tilde{\boldsymbol{\beta}} \neq \mathbf{0}$. One method to test this is to fit a reduced model without accounting for region, and then fitting an anova model between the two. Doing

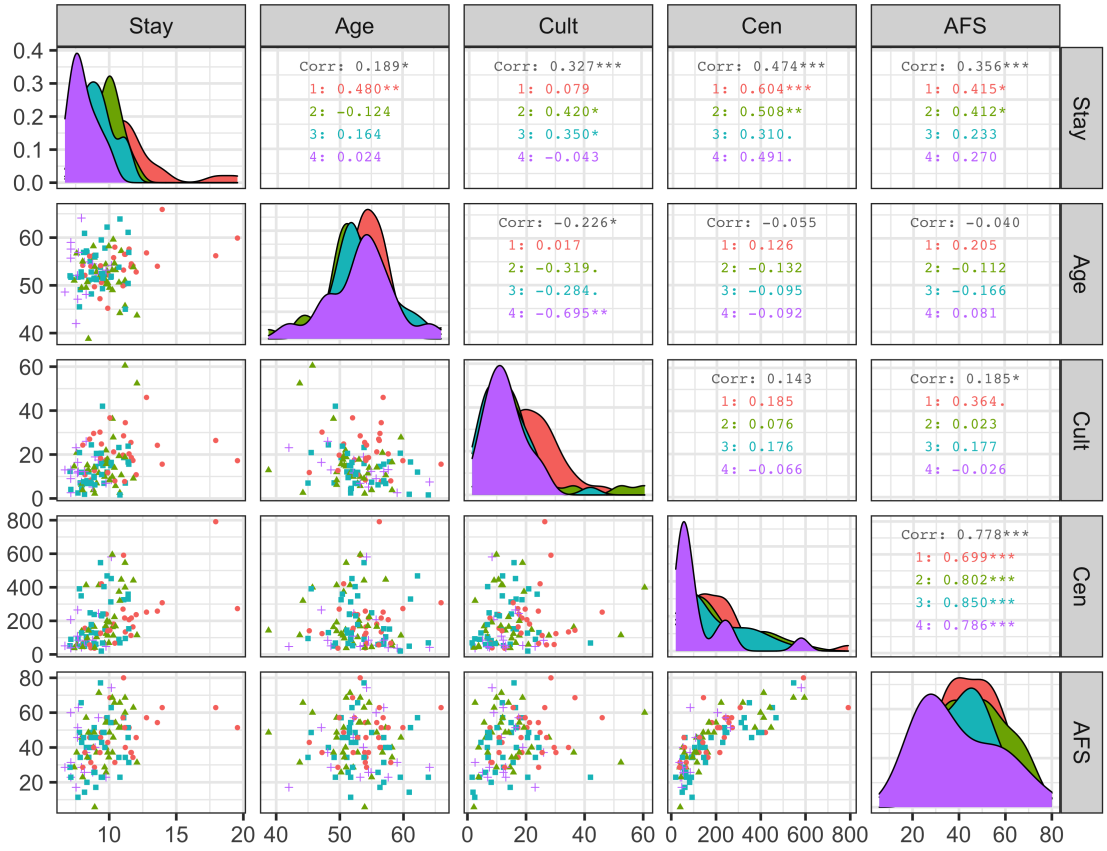

Figure 4: A scatterplot matrix of Stay against Age, Cult, Cen, and AFS.

this, we have a p-value of 3.77×10^{-5}, so for any reasonable significance level we will reject H_0. The data seems to indicate that the geographic region does play a significant role in average patient stay.

Question 4

Collaborators: None

(a) For this question we will assume the matrix $\mathbf{X}$ is *centered*, so that each variable (column) has mean zero. For the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we are no longer estimating an estimator term β_0. In n-dimensional Euclidean vector space, for any $\mathbf{v} \in \mathbb{R}^n$ we have $\|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{v}$, so the ridge regression loss function becomes

$$Q(\boldsymbol{\beta}; \lambda) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

Differentiating the loss function with respect to $\boldsymbol{\beta}$ gives us

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} = -2\mathbf{X}^T\mathbf{y} + 2\big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)\boldsymbol{\beta} \overset{\text{set}}{=} \mathbf{0},$$

and finally solving for $\boldsymbol{\beta}$ gives us $\hat{\boldsymbol{\beta}}_\lambda = \big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y}$.

(b) When $\lambda = 0$ we have $\hat{\boldsymbol{\beta}}_{\lambda=0} = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{y}$, the OLS estimator. We can re-write $\hat{\boldsymbol{\beta}}_\lambda$ as

$$\hat{\boldsymbol{\beta}}_\lambda = \big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y} = \big(\lambda\lambda^{-1}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y} = \lambda^{-1}\big(\lambda^{-1}\mathbf{X}^T\mathbf{X} + \mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y}.$$

From here, as $\lambda \to \infty$, we have $\lambda^{-1} \to 0$, and so $\hat{\boldsymbol{\beta}}_\lambda \to \mathbf{0}$.

(c) The alternative ridge estimator can be re-written as

$$\tilde{\boldsymbol{\beta}}_\lambda = \lambda\big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y} = (\lambda^{-1})^{-1}\big(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y} = \big(\lambda^{-1}\mathbf{X}^T\mathbf{X} + \mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y},$$

and so

$$\lim_{\lambda\to\infty}\tilde{\boldsymbol{\beta}}_\lambda = \lim_{\lambda\to\infty}\big(\lambda^{-1}\mathbf{X}^T\mathbf{X} + \mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{I}^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}.$$