Homework 5

Aiden Kenny
STAT GR5205: Linear Regression Models
Columbia University

Novermber 25, 2020

Question 1

Collaborators: None

(a) Let Y be the number of nurses in the hospital and let X be the available faculty and services. The left and middle panels of Figure 1 show the histograms of Y and X, respectively. We see that Y is skewed right, while X appears to be normally-distributed. In addition, the scatterplot of Y vs. X, which is in the third panel of Figure 1, shows that there is a nonlinear relationship between Y and X. All of these indicate that Y is suitable for a data transformation. Specifically, we would like to perform a power transformation on Y in order to make it closer to a normal distribution.

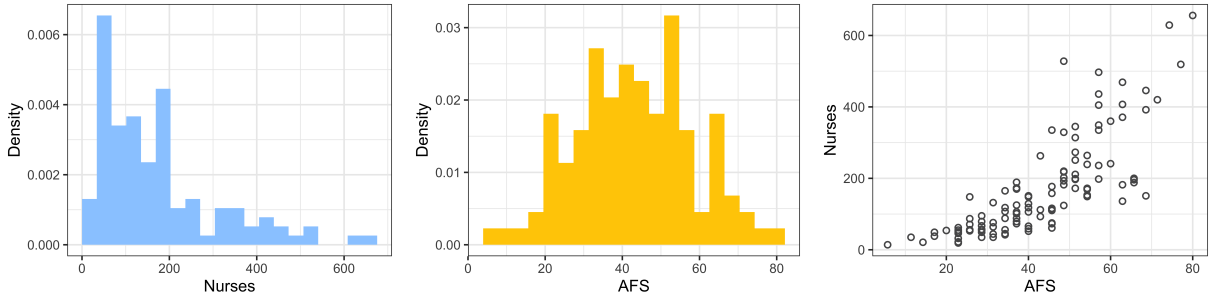


Figure 1: Histograms of Y and X and a scatterplot of Y vs. X.

(b) The power transformation function and its scaled counterpart are defined as

$$f_\lambda(Y) = \begin{cases} Y^\lambda & \text{if } \lambda \neq 0 \\ \log Y & \text{if } \lambda = 0. \end{cases} \quad \text{and} \quad g_\lambda(Y) = \begin{cases} (Y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log Y & \text{if } \lambda = 0. \end{cases}$$

When transforming Y, we first use the scaled power transform to fit the model $g_\lambda(Y) = \beta_0 + \beta_1 X + \epsilon$ in order to determine an optimal value of λ via maximum likelihood estimation. We then use f_λ to make Y closer to a normal distribution and perform any subsequent inferences. In vector form, our model is $\mathbf{g}_\lambda(\boldsymbol{Y}) = \beta_0 \mathbf{1} + \beta_1 \mathbf{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and $\mathbf{g}_\lambda(\boldsymbol{Y})$ is the transformed response (so $[\mathbf{g}_\lambda(\boldsymbol{Y})]_i = g_\lambda(Y_i)$) for some unknown λ. It can be shown (see Appendix A) that the log-likelihood function can be expressed as a function of only λ,

$$m(\lambda) \coloneqq -\frac{n}{2} \log\left(\frac{2\pi \mathrm{e}}{n}\right) - \frac{n}{2} \log\left(\mathbf{g}_\lambda^T(\mathbf{y})(\mathbf{I} - \mathbf{H})\mathbf{g}_\lambda(\mathbf{y})\right) + (\lambda - 1) \sum_{i=1}^{n} \log(y_i),$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix. $m(\lambda)$ has been plotted in the left panel of Figure 2, where we can see that the MLE is maximuzed at $\lambda = 0.085$.

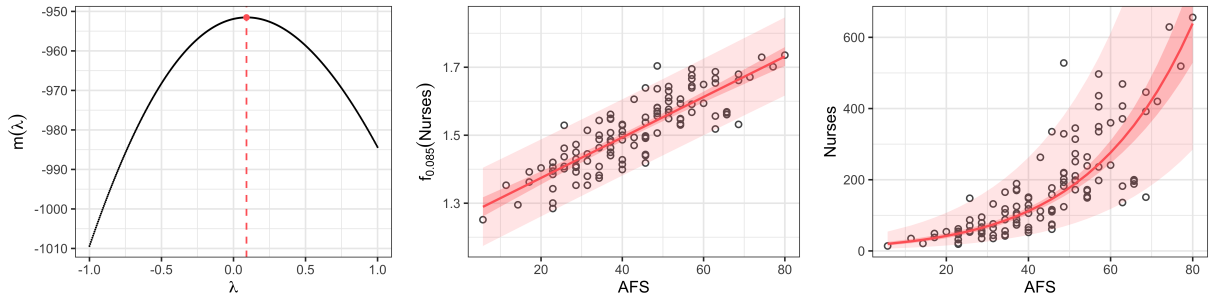


Figure 2: Relevant plots for the power transformation of Y.

(c) Now that we have our optimal λ, we will fit the model $f_{0.085}(Y) = \beta_0 + \beta_1 X + \epsilon$. We have $\hat{\beta}_0 = 1.255$ and $\hat{\beta}_1 = 0.005953$. A plot of this model, along with a 95% confidence interval, has been printed in the middle panel of Figure 2. This means that when X increases by 1 unit, $f_{0.085}(Y)$ is expected to increase by 0.005953 units. We can also make inferences about Y itself; we have $Y = f^{-1}_{0.085}(\hat{\beta}_0 + \hat{\beta}_1 X) = \big(1.255 + 0.005953X\big)^{1/0.085}$. A plot of this model has been printed in the right panel of Figure 2. We must keep in mind that this model is non-linear, so it's expected rate of change will depend on the value of X.

(d) Both diagnostic plots can be found in Figure 3, and the data seems to confirm the model assumptions very well.

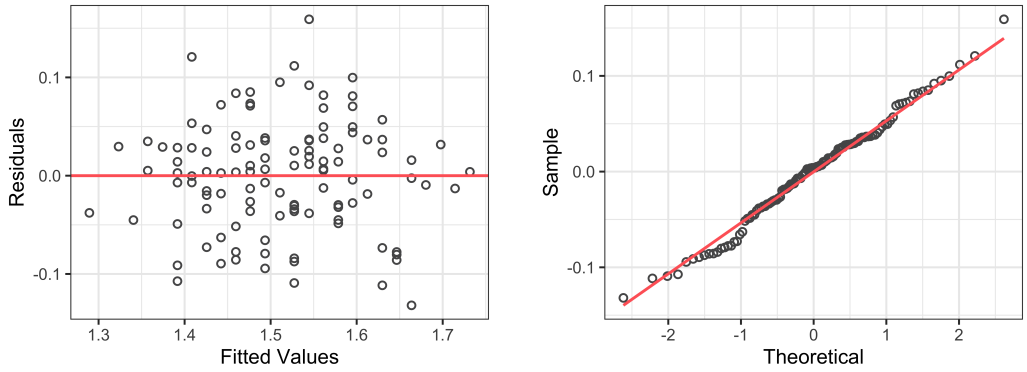


Figure 3: Diagnostic plots for the power transformation model.

(e) We cannot conduct an F-test on the two models. This is because both models have the same number of parameters that need to be estimated, meaning they both have the same number of degrees of freedom.

(f) In order to find a prediction interval for Y, we can first find a prediction interval for $f_{0.085}(Y)$ and then apply $f^{-1}_{0.085}$ to both ends of the interval. The prediction intervals for when $X = 30$ and $X = 60$ have been printed in Table 1. For reference, the prediction intervals at these observations for the linear model $Y = \beta_0 + \beta_1 + \epsilon$ have also been printed. The prediction intervals for the power model can also be seen in the middel and right plots of Figure 2. For starters, the width of the intervals for the power model drastically increases as X increases, while the increase for the linear model is negligible. When $X = 30$, the confidence interval for the power model is significantly narrower than the linear function, but when $X = 60$, the opposite is true. That is, when $X = 30$ we are confident that the next observation will lie in a smaller interval using the power model, but when $X = 60$, the linear model gives us the smaller region. Depending on the value of X, I would use the better of the two models.

	$X = 30$				$X = 60$			
Model	$\hat{Y}$	Lower	Upper	$\Delta\mathcal{I}$	$\hat{Y}$	Lower	Upper	$\Delta\mathcal{I}$
Linear	74.612	−89.781	239.004	328.785	297.769	133.036	462.503	329.467
Power	69.448	26.573	168.796	142.223	276.313	117.843	611.638	493.795

Table 1: Comparing the prediction intervals of the linear model against the power model.

Question 2

Collaborators: None

(a) We are now applying the scaled power function to both the response and each of the predictors of interest. Let Y be the length of stay and $Z_1, \ldots, Z_8$ be average age of patients, infection risk, cultering ratio, X-ray ratio, number of beds, average daily census, number of nurses, and available facilities and services, respectively. To find optimal parameters for transformation, we will fit the model

$$g_\theta(Y) = \beta_0 + \sum_{i=1}^{8} \beta_i \cdot g_{\lambda_i}(Z_i) + \epsilon, \tag{1}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$, and use MLE to find the optimal values of θ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_8)^T$. The values suggested by the initial question are $\theta_0 = -1$ and $\boldsymbol{\lambda}_0 = (1, 1, 0, 1, 0, 0, 0, 1)^T$. Using vector notation, our model can be written as $\mathbf{g}_\theta(\boldsymbol{Y}) = \mathbf{G}_{\boldsymbol{\lambda}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{G}_{\boldsymbol{\lambda}} = \begin{bmatrix}\mathbf{1} & \mathbf{g}_{\lambda_1}(\mathbf{z}_1) & \mathbf{g}_{\lambda_8}(\mathbf{z}_8)\end{bmatrix}$ is the matrix of predictors transformed by g_λ. It can be shown (see Appendix B) that the log-likelihood can be expressed as a function of only the unknown parameters θ and $\boldsymbol{\lambda}$,

$$m(\theta, \boldsymbol{\lambda}) \coloneqq -\frac{n}{2}\log\left(\frac{2\pi \mathrm{e}}{n}\right) - \frac{n}{2}\log\left(\mathbf{g}_\theta^T(\mathbf{y})(\mathbf{I} - \mathbf{H}_{\boldsymbol{\lambda}})\mathbf{g}_\theta(\mathbf{y})\right) + (\theta - 1)\sum_{i=1}^{n}\log(y_i),$$

where $\mathbf{H}_{\boldsymbol{\lambda}} = \mathbf{G}_{\boldsymbol{\lambda}}(\mathbf{G}_{\boldsymbol{\lambda}}^T\mathbf{G}_{\boldsymbol{\lambda}})^{-1}\mathbf{G}_{\boldsymbol{\lambda}}^T$ is the hat matrix. This is a non-linear equation with nine parameters that must be optimized. Using the `R` base function `optim` with initial conditions θ_0 and $\boldsymbol{\lambda}_0$, 100 itterations of a simplex method are run. The values of $\hat{\theta}_{\text{MLE}}$ and $\hat{\boldsymbol{\lambda}}_{\text{MLE}}$ can be found in Table 2, and we can see that the MLEs for θ and $\boldsymbol{\lambda}$ vary quite differently than the original guesses. We can also use `R` to show $m(\theta_0, \boldsymbol{\lambda}_0) = -172.252$ and $m(\hat{\theta}_{\text{MLE}}, \hat{\boldsymbol{\lambda}}_{\text{MLE}}) = -169.769$, so using the estimated parameters gives us a higher log-likelihood. Therefore, I am going to make the argument that the inital guess is not the best fit for the model. However, because the log-likelihood does not drastically change, I believe it is still an acceptable model, especially since it is much more interpretable. If one wants to use "cleaner" values of θ and $\boldsymbol{\lambda}$ for better interpretability, we can round the values of $\hat{\theta}_{\text{MLE}}$ and $\hat{\boldsymbol{\lambda}}_{\text{MLE}}$ as I suggest in the last row of Table 2. Using these values, we have $m(\theta_s, \boldsymbol{\lambda}_s) = -170.792$

	θ	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
Initial	−1	1	1	0	1	0	0	0	1
MLE	−0.394	0.033	2.373	0.890	0.844	−0.042	0.187	0.889	−0.848
Suggested	−1/2	0	2	1	1	0	0	1	−1

Table 2: Comparing the initial guess of the parameters against their MLEs for model (1).

(b) Let $X_i = f_{\lambda_{0,i}}(Z_i)$ and $W = f_{\theta_0}(Y)$, i.e. the predictors and response transformed by the initial values θ_0 and $\boldsymbol{\lambda}_0$. The correlation matrix, which can be found in Appendix C, shows that X_5, X_6, X_7, and X_8 are all highly-correlated with each other.

(c) Depending on the stopping criteria, we will get a differnt suggestion for the coefficients to be included in the model. If either Mallow's C_p or adjusted R^2 are used, forward stepwise regression suggests we

include X_1, X_2, X_4, X_5, X_6, and X_7, so our model will have six predictors. On the other hand, if BIC is used, we should include X_1, X_2, X_4, X_6, and X_7, so this model will only have five predictors.

(d) Regardless of the stopping criteria used, backward stepwise selection gives the same suggestions as forward stepwise regression.

(e) We will first fit our model $W = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon$. Values of the estimated coefficients can be found in the R code. The adjusted R^2 is 0.4896, so the model explains just under half of the response's variability.

Question 3

Collaborators: None

(a) A boxplot of the four test scores can be found in the left panel of Figure 4. We can see that the average score of all four tests is about the same, around 100. The second test had the hightest average score, and is negatively skewed, so more people performed well on this test. The first test has the greatest variance of the four, its range is quite significant, and there is a single outlier (the score of 150).

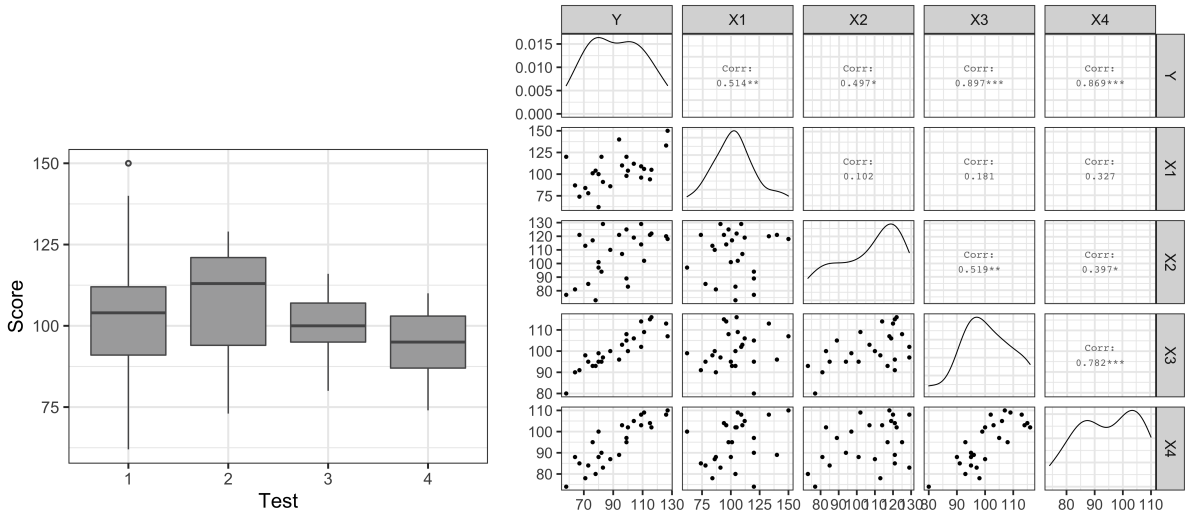


Figure 4: Information about the test scores.

(b) The scatterplot matrix can be found in the right panel of Figure 4. The only two variables that seem to have a strong correlation with each other are tests 3 and 4, while tests 2 and 3 have somewhat of a correlation. We can also see that the job proficiency is highly-correlated with tests 3 and 4.

(c)

Appendix A

In this section, we will derive the result for $m(\lambda)$, the log-likelikelihood as a function of only λ. Suppose we have our response vector $\mathbf{y}$ and our observed data $\mathbf{X} = \begin{bmatrix}\mathbf{1} & \mathbf{x}\end{bmatrix}$. We are interested in fitting the model $\mathbf{g}_\lambda(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{g}_\lambda(\mathbf{y})$ is the transformed response vector (i.e. the ith element is given by $g_\lambda(y_i)$), $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, and $\text{Var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$. For notational ease, we will denote $\mathbf{g}_\lambda(\mathbf{y})$ as $\mathbf{g}_\lambda$. If we make the further assumption that $\boldsymbol{\epsilon}$ is normally distributed, i.e. $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I})$, then our response vector is also normally

distributed, where $\mathbf{g}_\lambda \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. It's density function (and thus it's likelihood function) is given by

$$\begin{aligned} f(\mathbf{g}_\lambda \,|\, \boldsymbol{\beta}, \sigma^2, \lambda) &= \frac{1}{\sqrt{\det(2\pi\sigma^2\mathbf{I})}} \cdot \exp\left(-\frac{(\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})^T \left(\sigma^2\mathbf{I}\right)^{-1} (\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})}{2} \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{(\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right). \end{aligned}$$

Since $\mathbf{y}$ is a transformation of $\mathbf{g}_\lambda$ (via g_λ^{-1}), we can derive the density for $\mathbf{y}$ as well. Notationally, this result may be somewhat confusing; even though we are finding the density for $\mathbf{y}$, we will still express the density (partly) in terms of $\mathbf{g}_\lambda$. It is important to remember that $\mathbf{g}_\lambda$ is a function of $\mathbf{y}$. Because the ith element of $\mathbf{g}_\lambda$ only depends on the ith element of $\mathbf{y}$, the Jacobian will be a diagonal matrix, and so

$$\mathbf{J} = \frac{\partial \mathbf{g}_\lambda}{\partial \mathbf{y}} = \mathrm{diag}\left(\frac{\partial g_\lambda(y_1)}{\partial y_1}, \ldots, \frac{\partial g_\lambda(y_n)}{\partial y_n} \right) = \mathrm{diag}\big(y_1^{\lambda-1}, \ldots, y_n^{\lambda-1}\big),$$

and so the density (and thus the likelihood) of $\mathbf{y}$ is given by

$$g(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \lambda) = f\big(\mathbf{g}_\lambda(\mathbf{y})\big) \cdot \big|\det(\mathbf{J})\big| = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{(\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right) \cdot \prod_{i=1}^{n} y_i^{\lambda-1}.$$

The log-likelihood $\ell(\mathbf{y}) = \log g(\mathbf{y})$ is given by

$$\ell(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{(\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{g}_\lambda - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} + (\lambda - 1)\sum_{i=1}^{n} \log(y_i).$$

As is standard with maximum likelihood estimation, we now differentiate ℓ with respect to the unknown parameters, set the derivatives to zero, and solve to get the maximum value of ℓ. For now, we are going to leave λ fixed and differentiate with respect to $\boldsymbol{\beta}$ and σ^2. Doing this for both gives us $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{g}_\lambda$ and $\hat{\sigma}^2_{\mathrm{MLE}} = \mathbf{g}_\lambda^T(\mathbf{I} - \mathbf{H})\mathbf{g}_\lambda / n$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix. It is worth noting that both $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ and $\hat{\sigma}^2_{\mathrm{MLE}}$ are functions of λ. Plugging these values back into ℓ will maximize it with respect to $\boldsymbol{\beta}$ and σ^2, which means we will only have to maximize it with respect to λ. With some simplification, our new loss function is

$$m(\lambda) \coloneqq \ell(\mathbf{y} \,|\, \hat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \hat{\sigma}^2_{\mathrm{MLE}}, \lambda) = -\frac{n}{2}\log\left(\frac{2\pi e}{n}\right) - \frac{n}{2}\log\left(\mathbf{g}_\lambda^T(\mathbf{I} - \mathbf{H})\mathbf{g}_\lambda\right) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i).$$

Ideally, we would differentiate m with respect to λ, set $\partial m / \partial \lambda = 0$, and solve for λ. I was unable to derive a closed form solution for the result, but it is still possible to use graphical techniques or numerical methods to find the optimal value of λ.

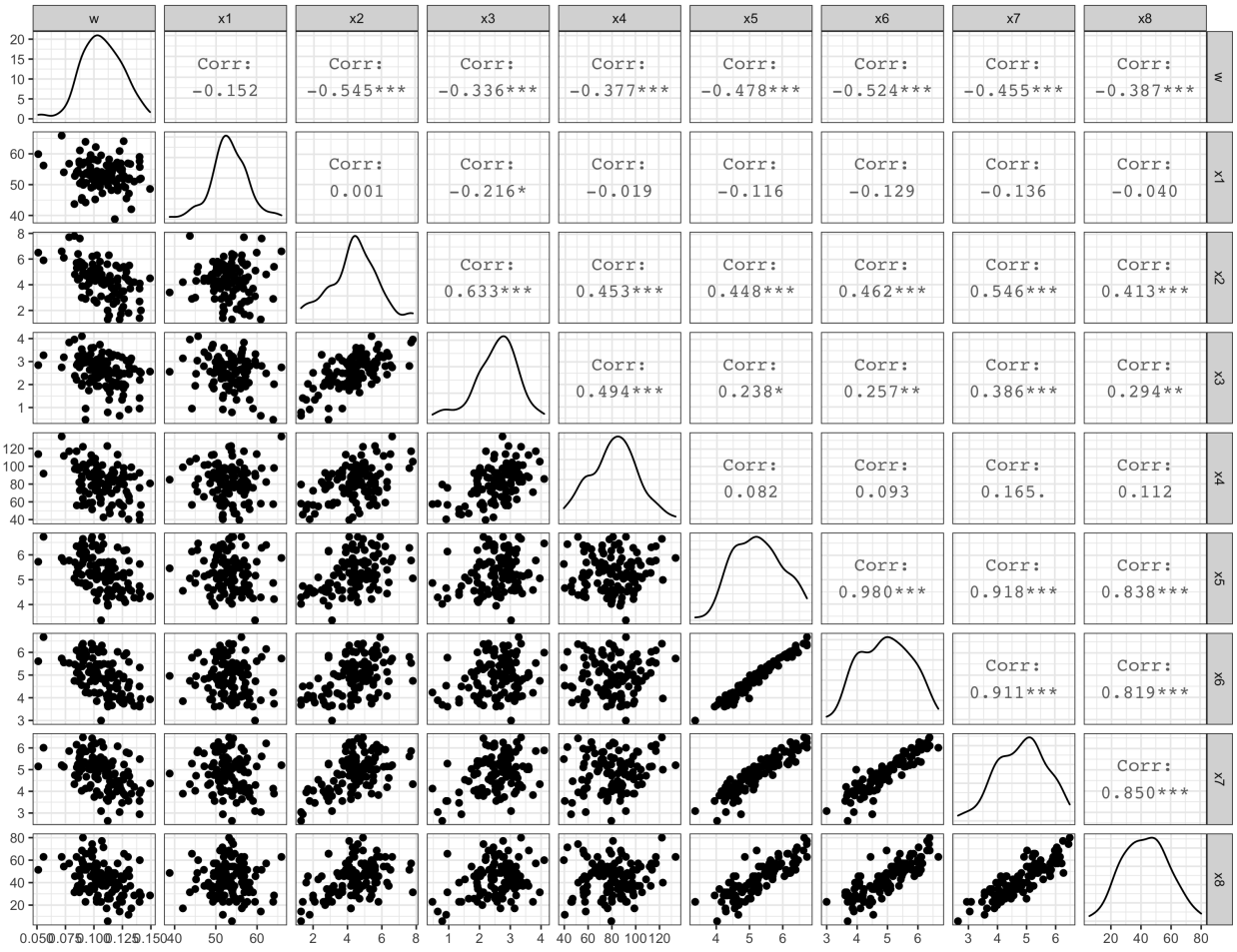
We recall that both $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ and $\hat{\sigma}^2_{\mathrm{MLE}}$ are functions of λ, so we cannot know their value until $\hat{\lambda}_{\mathrm{MLE}}$ has been determined. Because of this, as we just showed, the likelihood function can be expressed as a function $m(\lambda)$ that only depends on λ, which can be maximized to find $\hat{\lambda}_{\mathrm{MLE}}$. Once we find $\hat{\lambda}_{\mathrm{MLE}}$, we can use this value to determine $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ and $\hat{\sigma}^2_{\mathrm{MLE}}$.

Appendix B

Here we will derive the log-likelihood as a function of θ and $\boldsymbol{\lambda}$. The methodology is identical to that of Appendix A, only now our response vector is $\mathbf{g}_\theta(\mathbf{y})$ and the data matrix is $\mathbf{G}_{\boldsymbol{\lambda}}$. Using the exact same steps, we get our result $m(\theta, \boldsymbol{\lambda})$.

Appendix C

Here is the correlation matrix for part (b) of question 2.


w
x1
x2
x3
x4
x5
x6
x7
x8
Corr: -0.152
Corr: -0.545***
Corr: -0.336***
Corr: -0.377***
Corr: -0.478***
Corr: -0.524***
Corr: -0.455***
Corr: -0.387***
Corr: 0.001
Corr: -0.216*
Corr: -0.019
Corr: -0.116
Corr: -0.129
Corr: -0.136
Corr: -0.040
Corr: 0.633***
Corr: 0.453***
Corr: 0.448***
Corr: 0.462***
Corr: 0.546***
Corr: 0.413***
Corr: 0.494***
Corr: 0.238*
Corr: 0.257**
Corr: 0.386***
Corr: 0.294**
Corr: 0.082
Corr: 0.093
Corr: 0.165.
Corr: 0.112
Corr: 0.980***
Corr: 0.918***
Corr: 0.838***
Corr: 0.911***
Corr: 0.819***
Corr: 0.850***