MSAN 601 - Homework 3

Andre Guimaraes Duarte

September 19, 2016

Question 1

Table 1: Summary of results for `Transact` data.

$Y = \beta_{01} + \beta_{11}t_1 + \beta_{21}t_2$	$Y = \beta_{02} + \beta_{32}a + \beta_{42}d$	$Y = \beta_{03} + \beta_{23}t_2 + \beta_{43}d$	$Y = \beta_{04} + \beta_{14}t_1 + \beta_{24}t_2 + \beta_{34}a + \beta_{44}d$
$b_{01} = 144.36944$ $b_{11} = 5.46206$ $b_{21} = 2.03455$	$b_{02} = 144.3694$ $b_{32} = 7.4699$ $b_{42} = 1.7138$	$b_{03} = 144.3694$ $b_{23} = 7.4699$ $b_{43} = 5.4621$	$b_{04} = 144.36944$ $b_{14} = 5.46206$ $b_{24} = 2.03455$ $b_{34} = \texttt{NA}$ $b_{44} = \texttt{NA}$

a) As we can see from Table 1, the coefficient estimates b_{34} and b_{44} for β_{34} and β_{44} are labeled as `NA`. Indeed, the variables corresponding to these coefficients are a and d, which are linear combinations of t_1 and t_2, that are already in the model. Consequently, matrix X is no longer full-rank, and $X'X$ is not invertible (reminder: $X'X$ needs to be invertible to calculate $\hat{Y} = HY$, where $H = X(X'X)^{-1}X'$).

b) From Table 1, we can see that β_{0i} is the same for all four models. In addition, the F-statistic is the same (1289 on 2 and 258 degrees of freedom), as well as R^2 (0.9091) and R_a^2 (0.9083). This is true even for model (4). Indeed, since a and d are linear combinations of the other two variables, they are not included in the final model.

In addition, we can see that models (1) and (4) have the same coefficients, and are in fact the same model. In (2) and (3), we get $\beta_{32} = \beta_{23}$, but $\beta_{42} \neq \beta_{43}$.

c) We can see in Table 1 that the estimate for t_2 is different in (1) and (3). This is because the second regressor in each model is not the same (t_1 in (1) and d in (3)). The impact of t_2 when t_1 is in the model is different from the impact of t_2 when d is in the model.

Question 2

a) Here, we are modeling the variation of the log of `fertility` (the number of children per woman) as a linear function of `pctUrban` (the percent urban). With `R`, we get

$\log(\text{fertility}) = 1.500963 - 0.010163 * \text{pctUrban}.$

An increase in one unit in `pctUrban` (an additional percent in urban population) leads to an expected decrease in 0.010163 units of the log of `fertility`. Since this is hard to interpret, let's convert the equation to get rid of the log. We get

$$\begin{aligned} \text{fertility} &= \exp(1.500963 - 0.010163 * \text{pctUrban}) \\ &= \exp(1.500963) \times \exp(-0.010163 * \text{pctUrban}) \end{aligned}$$

Therefore, an increase in 25 percent in urban population is written as

$$\begin{aligned} \exp(1.500963) \times (1.25 \times \text{pctUrban})^{-0.010163} &= \exp(1.500963) \times 1.25^{-0.010163} \times \text{pctUrban}^{-0.010163} \\ &= 0.9977348 \times \exp(1.500963) \times \text{pctUrban}^{-0.010163} \\ &= 0.9977348 \times \text{fertility} \end{aligned}$$

As can be seen, this leads to a decrease in $1 - 0.9977348 = 0.0022652 = 0.22652\%$ in fertility.

b) We now have the following model

$\log(\text{fertility}) = 3.50736 - 0.06544 * \log(\text{ppgdp}) - 0.02824 * \text{lifeExpF}.$

We can re-write this equation to remove the logarithms:

$$\begin{aligned}\text{fertility} &= \exp\left(3.50736 - 0.06544 * \log(\text{ppgdp}) - 0.02824 * \text{lifeExpF}\right)\\ &= \exp\left(3.50736\right) \times \exp\left(-0.06544 * \log(\text{ppgdp})\right) \times \exp\left(-0.02824 * \text{lifeExpF}\right)\\ &= \exp\left(3.50736\right) \times \text{ppgdp}^{-0.06544} \times \exp\left(-0.02824 * \text{lifeExpF}\right)\end{aligned}$$

A 25% increase in `ppgdp` leads to the following equation:

$$\begin{aligned} & \exp\left(3.50736\right) \times (1.25 * \text{ppgdp})^{-0.06544} \times \exp\left(-0.02824 * \text{lifeExpF}\right)\\ = \; & \exp\left(3.50736\right) \times (1.25)^{-0.06544} \times \text{ppgdp}^{-0.06544} \times \exp\left(-0.02824 * \text{lifeExpF}\right)\\ = \; & 0.9855036 \times \exp\left(3.50736\right) \times \text{ppgdp}^{-0.06544} \times \exp\left(-0.02824 * \text{lifeExpF}\right)\\ = \; & 0.9855036 \times \text{fertility}\end{aligned}$$

As can be seen, this leads to a decrease in $1 - 0.9855036 = 0.0144964 = 1.44964\%$ in fertility.

Question 3

a) Using `R`, we first have a look at the scatterplot for the data, and look at the correlations using `pairs` and `cor`. The scatterplot is shown in Figure 1, and the correlation matrix is shown in Table 2. These results show us that the most likely candidate for an explanatory variable is $X2$.

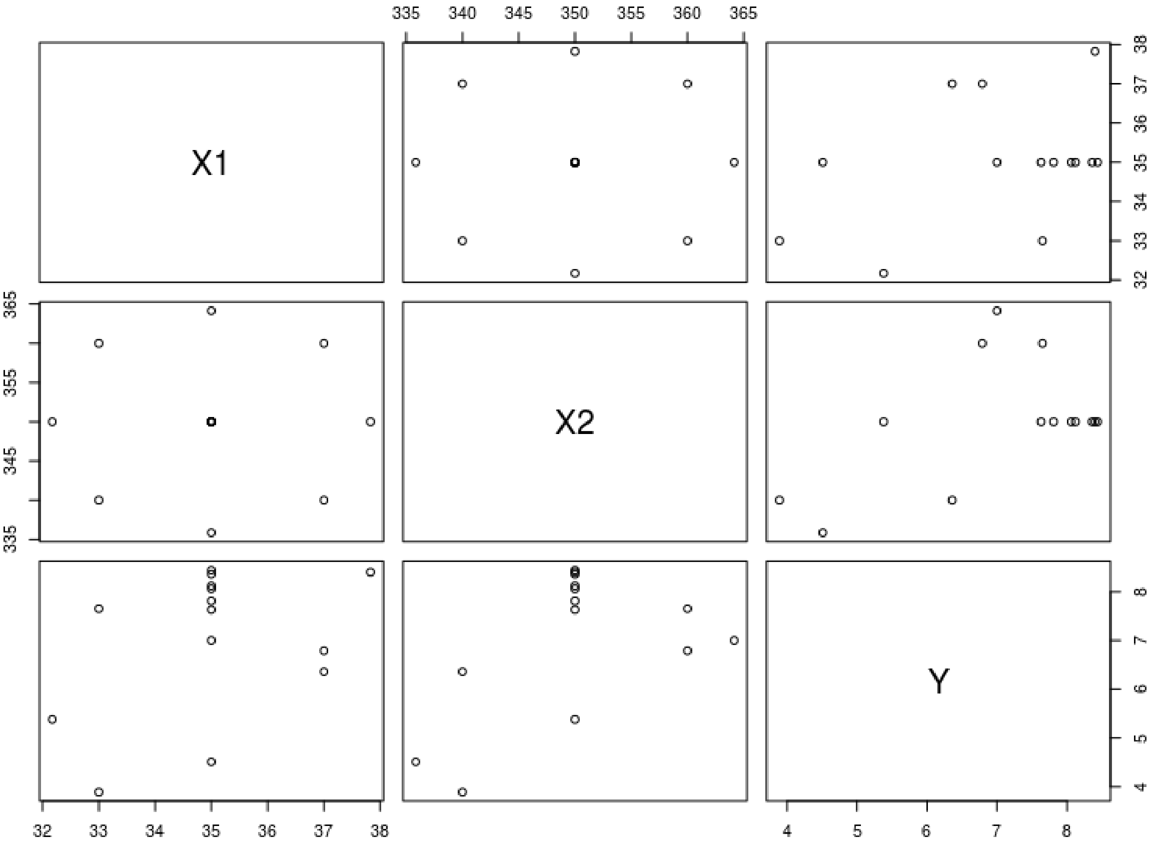


Figure 1: Pairs plot of `cakes` data.

However, if we test the model $Y = \beta_0 + \beta_1 X2$, we see that the coefficient associated to $X2$ is not significant (where $\alpha = 0.05$). In order to get a better linear fit, we transform Y using a logarithm in order to get

Table 2: Correlation matrix for `cakes` data.

	Y	X1	X2
Y	1.00000000		
X1	0.3882796	1.00000000	
X2	0.5091336	0.0000000	1.00000000

$\log(Y) = \beta_0 + \beta_1 X2$. By doing so, the resulting linear model is significant, and we get $b_1 = 0.017043$ (p = 0.0436). However, the R^2 is low, at 0.2977. Adding $X1$ does not improve the model.

b) By including the dummy variable `block`, we obtain the correlation matrix shown in Table 3. There is no clear linear relation between this variable and Y.

Table 3: Correlation matrix for `cakes` data with dummy variable `block`.

	Y	X1	X2	block
Y	1.00000000			
X1	0.3882796	1.00000000		
X2	0.5091336	0.0000000	1.00000000	
block	0.0399273	0.0000000	0.0000000	1.00000000

By including `block` in the previous model (using $\log(Y)$), we get worse results than previously. In fact, the overall F-test shows that the model is no longer statistically significant (the impact of all predictors is statistically null). `block` is not relevant for this exercise.

Question 4

The overall F-test is used to verify if there is an overall linear regression relation between the response variable `FuelC` and the collection of predictor variables `Drivers`, `Income`, `Miles`, `MPC`, `Pop`, `Tax`. The null and alternate hypotheses are

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$.
H_a: not all $\beta_k = 0 : k = 1, \ldots, 6$.

Using the `summary` function in `R`, we can easily find the value of the F statistic, or we can manually calculate it using the type I ANOVA table. We get

$F^* = 376.4$ on 6 and 44 degrees of freedom, which we compare to $F_{(6,44)} = 2.712$. Since $F^* > F_{(6,44)}$, we reject the null hypothesis H_0 that all estimator coefficients are equal to zero with $\alpha = 0.05$. At least one of them is not null and there is an overall linear regression relation between the response and the predictors.

Question 5

We use `R` to fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$. We have $p - 1 = 5 \Rightarrow p = 6$, $n = 14$. We use $\alpha = 5\%$ in this exercise.

a) **We wish to test H_0: $\beta_5 = 0$.**

We use the type II ANOVA table in order to get $SSE_f = 1.470745$ the sum of squares of the complete model, and we get $MSE_f = \frac{SSE_f}{df_f = n-p = 8} = 0.1838431$.

We also have $MSR = MSE(x_1 x_2 | x_1, x_1^2, x_2, x_2^2) = \frac{SSR(x_1 x_2 | x_1, x_1^2, x_2, x_2^2)}{df = 1} = 2.772225$.

We then perform an F-test, by calculating $F^* = \frac{MSR}{MSE_f} = \frac{2.772225}{0.1838431} = 15.0793$ and comparing it to $F_{(1,8)} = 7.570882$. Since $F^* > F_{(1,8)}$, then we reject H_0 and conclude that β_5 is not null.

b) **We wish to test H_0: $\beta_2 = 0$.**

We use the type II ANOVA table in order to get $SSE_f = 1.470745$ the sum of squares of the complete model, and we get $MSE_f = \frac{SSE_f}{df_f = n-p=8} = 0.1838431$.

We also have $MSR = MSE(x_1^2|x_1, x_2, x_2^2, x_1x_2) = \frac{SSR(x_1^2|x_1,x_2,x_2^2,x_1x_2)}{df=1} = 2.907738$.

We then perform an F-test, by calculating $F^* = \frac{MSR}{MSE_f} = \frac{2.907738}{0.1838431} = 15.81641$ and comparing it to $F_{(1,8)} = 7.570882$. Since $F^* > F_{(1,8)}$, then we reject H_0 and conclude that β_2 is not null.

c) **We wish to test** H_0**:** $\beta_1 = \beta_2 = \beta_5 = 0$.

We use the type II ANOVA table in order to get $SSE_f = 1.470745$ the sum of squares of the complete model, and we get $MSE_f = \frac{SSE_f}{df_f = n-p=8} = 0.1838431$.

We construct a reduced model assuming H_0 holds by regressing Y on x_2, x_2^2. We obtain $SSE_r = 11.47387$ and $df_r = n - (p-3) = 11$.

We get $F^* = \frac{\frac{SSE_r - SSE_f}{df_r - df_f}}{\frac{SSE_f}{df_f}} = 18.13707$. We also have $F_{(3,8)} = 5.415962$. Since $F^* > F_{(3,8)}$, then we reject H_0 and conclude that $\beta_1, \beta_2, \beta_5$ are not all null.

Question 6

The first thing we do is plot the data. Since we have four possible variables, we use the `pairs` function for this. The result is shown in Figure 2.

We can see that there seems to be a strong linear relation between Y and $X1$. In addition, there seems to be colinearity between $X2$ and $X3$. In order to verify this, we look at the correlation between these two variables (and all the others for good measure) using `cor`. The correlation matrix is shown in Table 4.

Table 4: Correlation matrix for `landrent` data.

	Y	X1	X2	X3	X4
Y	1.00000000				
X1	0.87577226	1.00000000			
X2	0.30857124	0.04871882	1.00000000		
X3	-0.32337827	-0.4998227	0.5225979	1.00000000	
X4	-0.08894927	0.08896304	-0.58343638	-0.08894927	1.00000000

We can see that the correlation between $X2$ and $X3$ is 0.5225979. In addition, this table shows a correlation between $X2$ and $X4$ as well. These results lead us to believe that $X2$ and $X4$ do not need to be included in the final model.

We decide to test the model $Y = \beta_0 + \beta_1 X1 + \beta_2 X2$, and get the following results:

- $b_0 = -6.11433$ ($p = 0.043 < 5\%$). This would be the baseline average rent per acre for alfalfa, but does not make sense in this context and can be disregarded.
- $b_1 = 0.92137$ ($p < 2 \cdot 10^{-16}$). With the density of dairy cows constant, an increase in rent for all tillable land by one unit leads to and expected increase in the mean rent per acre planted for alfalfa of 0.92137.
- $b_2 = 0.39255$ ($p = 1.59 \cdot 10^{-6}$). With the rent for all tillable land constant, an increase in the density of dairy cows by one unit leads to and expected increase in the mean rent per acre planted for alfalfa of 0.39255.
- The F-test is significant ($F^* = 165.3$ on 2 and 64 degrees of freedom, p-value: $< 2.2 \cdot 10^{-16}$), so there is an overall linear relation between Y and the two regressor variables.
- $R^2 = 0.8379$ and $R_a^2 = 0.8328$.

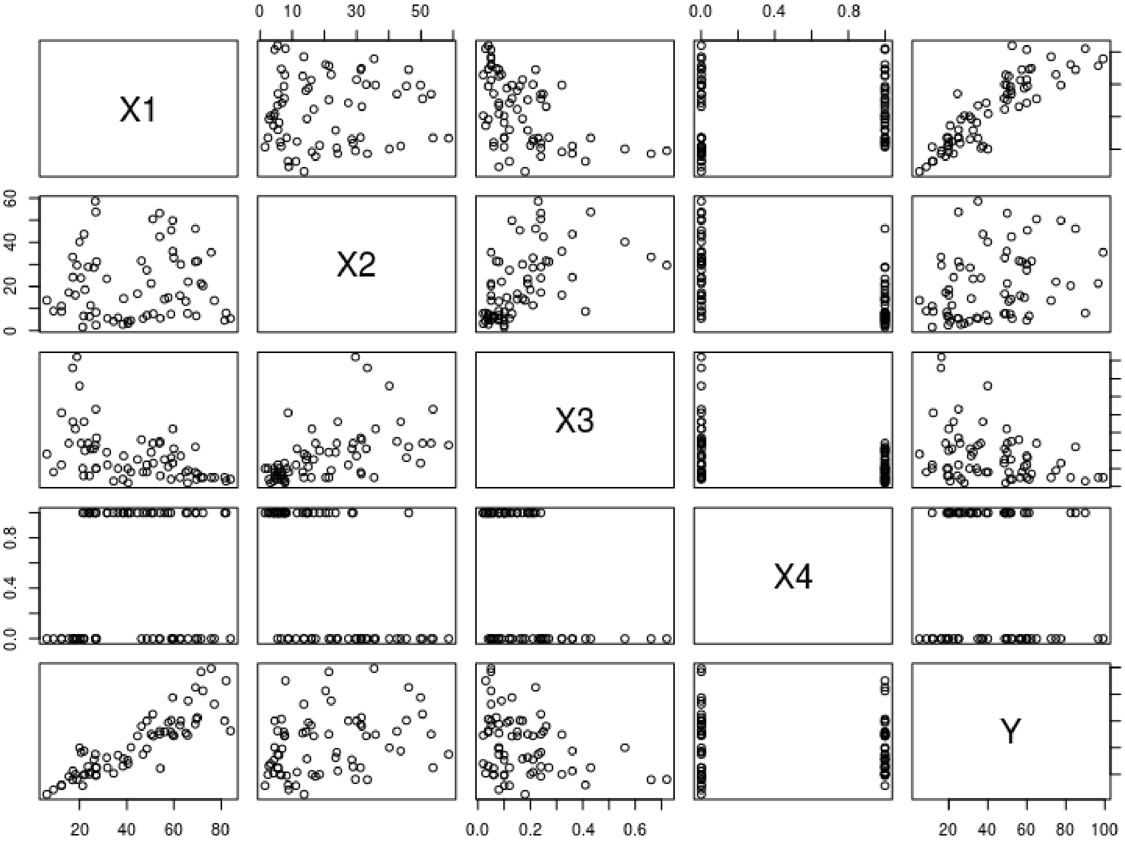


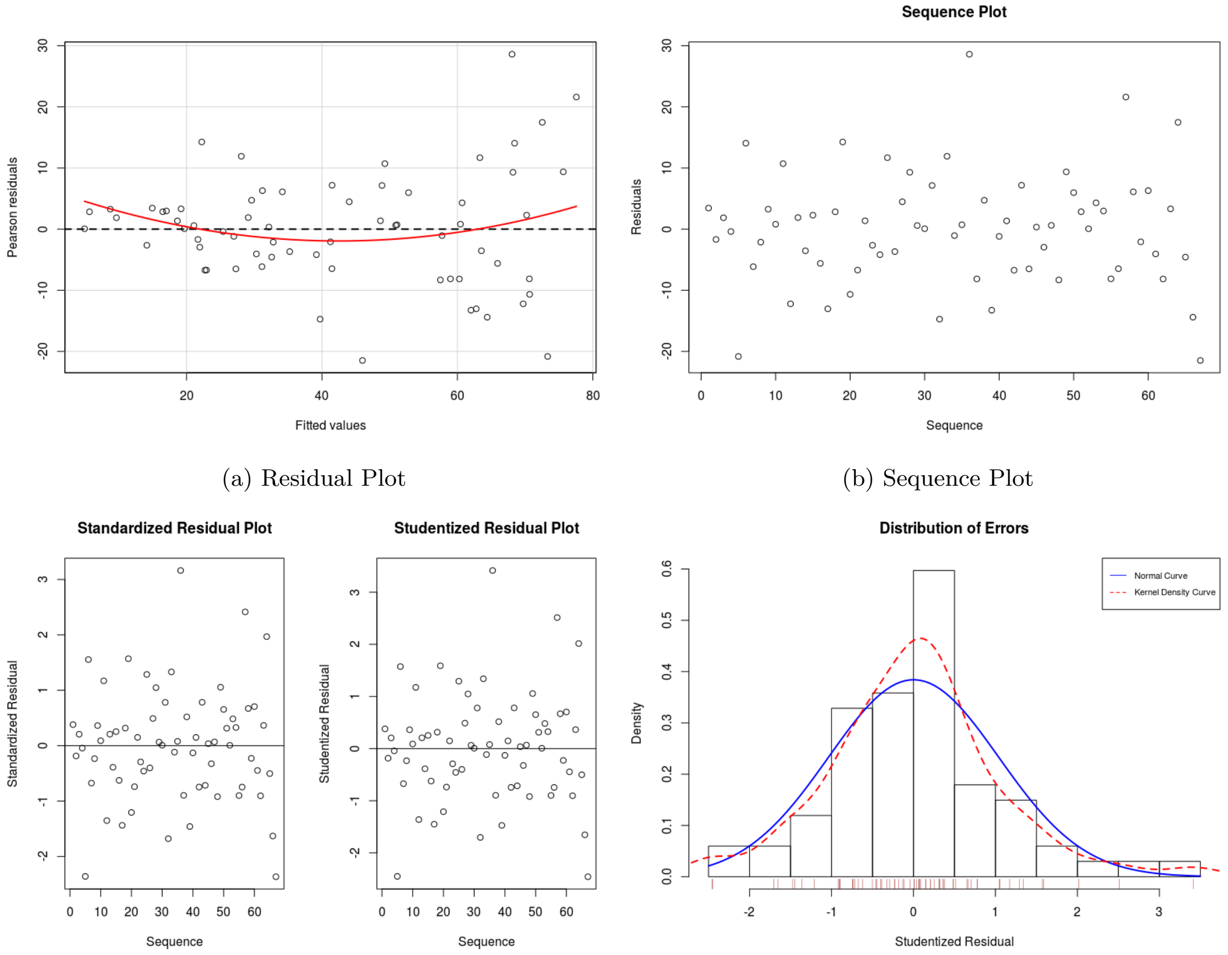
Figure 2: Pairs plot of `landrent` data.

We now make all the necessary tests on the residuals to see if the model is acceptable. All plots are shown in Figure 3.

We can see from Subfigure 3a that the residuals do not seem to be homoskedastic. We use a Brown-Forsythe test in order to determine this. We separate that data into two groups for the variable $X1$ separating along the median value. The p-value obtained this way is $0.07082 > 5\%$, so we do not reject H_0 and can say that the residuals are homoskedastic. The sequence plot in Subfigure 3b shows no sign of dependent residuals. No outliers can be discerned by the Subfigure 3c, and the distribution of the residuals seems normal from Subfigure 3d. This last statement is confirmed by a Shapiro-Wilk test (p-value = 0.3855) and a QQ-plot (not shown).

We can try to add another variable to the model, to see if it improves the model. If we add $X3$ (for the model $Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$), we get a non-significant estimator for the extra variable. If we add $X4$ (for the model $Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X4$), we also get a non-significant estimator for the extra variable. In both cases, R_a^2 decreases in relation to our first model. The effects of $X3$ and $X4$ are contained in their correlation with $X1$ and $X2$.

In conclusion, we decide to keep the model $Y = \beta_0 + \beta_1 X1 + \beta_2 X2$, for which we have $\hat{Y} = -6.11433 + 0.92137 X1 + 0.39255 X2$. The average rent per acre planted for alfalfa is positively correlated with the average rent paid for all tillable land. In addition, rent for land planted to alfalda relative to rent for other agricultural purposes is higher in areas with high density of dairy cows. However, liming does not have a significant impact of the average price per acre for alfalfa.



(a) Residual Plot

(b) Sequence Plot

(c) Outliers Plot

(d) Normality Plot

Figure 3: Regression tests for `landrent` data.

Question 7

In this question, two backward stepwise regression algorithms are implemented and tested using the `Rateprof` dataset. We start by creating the full model, using all predictor variables. Then, one variable is removed at a time, until a criterion is reached. For this exercise, two elimination criterions were implemented:

- The function `backstep_byPvalue` uses a p-value approach in order to determine when to stop removing variables. The α level is defined by the user, with 5% being the default. A variable is removed if the p-value for the significance of its predictor (t-value) is greater than this threshold. The variable with the highest p-value is chosen to be removed. The algorithm stops once all predictors are significant.
- The function `backstep_byAdjR2` uses an R_a^2 criterion to justify removing variables. We wish to get the model that gives the highest R_a^2. We compare the full model with the one with one less variable in it. If the reduced model has a higher R_a^2, then we remove that variable and repeat the process. If the reduced variable has a lower R_a^2, it means we have passed the highest return. Therefore, the full model at that step is returned.

Let's call `lm_p` and `lm_a` the two linear regression models obtained by p-value and ajusted R-squared respectively. We can see from their `summary` that the linear model with a p-value elimination criteria has only four explanatory variables in the reduced model, whereas the model using an adjusted R-squared criterion had 8 final explanatory variables. If inference is the main objective, the p-value method should be favored. However, the R_a^2 for the p-value regression model is 0.9971719, which is lower than the adjusted R-squared model, for which the R_a^2 is 0.9971979. In this case, the difference between the two does not seem important. For this dataset, the p-value method would be preferable since we have a very high R_a^2 with half the predictor variables used (it is more interpretable).

Question 8

See file `hw4q8.R` for this exercise.

Question 9

See file `hw4q9.R` for this exercise.

Question 10

In order to test whether Shapiro-Wilk's normality test is reliable, we run a little experiment.

- First, we explore the robustness of the test for non-normally distributed data when the data set is small ($n = 10$). For this, we run a random uniform distribution between 0 and 1 5000 times. We calculate the percentage of trials that are considered normal by SHapiro-Wilk. In one particular scenario, 91.56% of the trials were mistakenly labeled as normally distributed (since the nature of this test is by definition random, results may vary across experiments). So the Shapiro-Wilk normality test is not powerful for small samples.
- Then, we wish to test whether the test is accurate for samples with many data points ($n = 5000$). For this, we use a random standard normal distribution (without any added random noise). In my run, I got 4.16% of the trials that were considered non-normally distributed. The data points were picked from the same **normal** distirbution, and yet more than 4% of the trials were not considered normal by Shapiro-Wilk's normality test. So it is too powerful when the sample size is large.

This empirical example shows that Shapiro-Wilk's test by itself is not a good measure of a sample's normality. It can be used as a starting point for analysis, but other tests (either visual such as a QQ-plot) or quantitative (such as a test for skweness for example).