# Assignment 7-8

JARED YU

1. Use any math/stat software (e.g., www.numbergenerator.org/randomnumbergenerator) of your choice to find a random number generator to randomly select 20 rows of Table B.5 on page 558 of Textbook. Then perform a multiple regression fit to the data you generated. The multiple regression model contains the response variable $y$ ($CO_2$) and regressors $x_1$ (space time in min) and $x_6$ (solve total) and intercept.
    a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

Ans:

The rows selected are: 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 17, 18, 19, 20, 21, 22, 24, 25, and 27. After fitting a linear regression model, the resulting residuals are calculated and then the normal probability plot. This can be seen below in Figure 1. It was done using the method where the cumulative probability is used, or $P_i = \frac{\left(i-\frac{1}{2}\right)}{n}$, $i = 1, \cdots, n$. Afterward, a simple fitted line is plot through the data for comparison.

The points seem to be roughly scattered randomly along the line. However, it is not entirely ideal. For example, towards the bottom half (to the left of 0 on the $x$-axis) there is a bent under the line. It would be preferable instead for there to be a roughly random scatter along the line instead. Towards the tails of the data, there seems to be some heaviness. For example, on the top right there is one observation that is noticeable below the line and apart from the other nearby observations. On the bottom left, there are two points that seem to be flattening and straying towards some direction above the line. Another note is that these observations are relative to the fitted line. If for example some other line were drawn, a slightly different interpretation could be made. This plot in general seems to be hinting that due to the flattening occurring at the tails, the data seems to be following a heavier tailed distribution compared to the normal distribution.

It is also worth noting too that the one point at the top right and the two points at the bottom left are noticeably different from the main group of points in the center. They have a larger magnitude in comparison to the other points. This could possibly show that they're somewhat of a group of outliers. However, given that the data has so few points, and that this group is not one, but three possible outliers, it's difficult to say with certainty if they really are. The upper right point corresponds with observation 8. The furthest to the left at the bottom corresponds with observation 12 and the second furthest corresponds with observation 19.

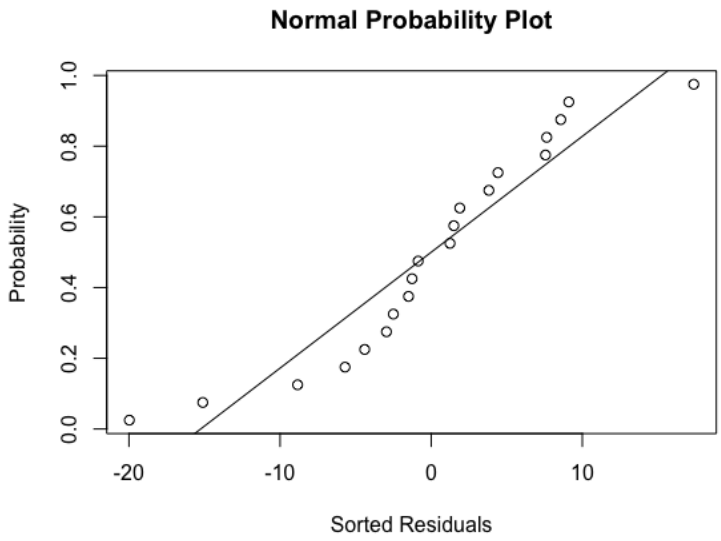


*Figure 1 A normal probability plot of the sorted residuals against the cumulative probability.*

b. Construct and interpret a plot of the residuals versus the predicted response.

Ans:

After calculating the residuals and predicted response, they were plot in a graph which can be seen below in Figure 2. In terms of the types of shapes indicated in the textbook (Figure 4.5, p.140), this one seems most similar to the outward-opening funnel pattern. It is quite noticeable that on the left, the points are in a narrower band, while moving towards the right the band is increasing. This indicates instead that the data has the problem where the variance is an increasing function of $y$.

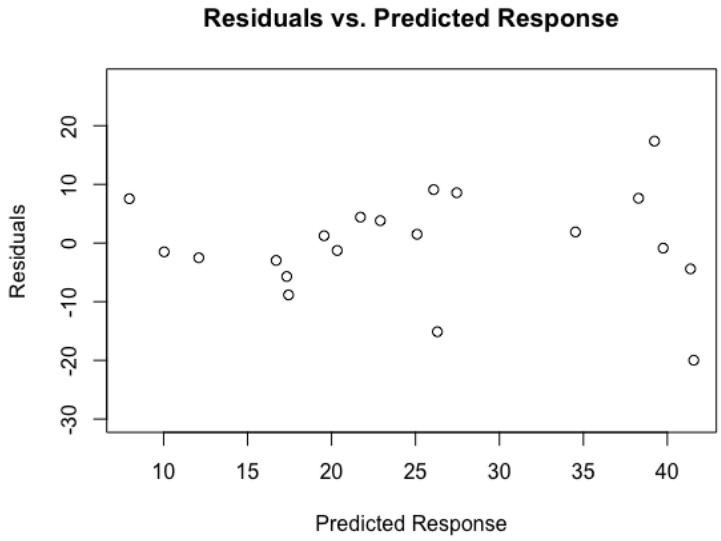


*Figure 2 A plot of the residuals versus the predicted response.*

c. Compute the studentized residuals and the $R$-student residuals for this model. What information is conveyed by these scaled residuals?

Ans:

For this part of the question, after calculating the studentized residuals and the $R$-student residuals, they were each plot into four different plots each time. These plot combinations can be seen below in Figure 3 and 4 respectively. For each of these figures, the top left shows the residuals versus predicted response, the top right shows the normal probability plot for the sorted residuals, the bottom right shows the residuals versus $x_6$, and the bottom left shows the residuals versus $x_1$.

In both figures, we can first look at the two upper plots, the type of residuals versus predicted response and the normal probability plot for the correspond type of residual. Appearance-wise, the top left plots are highly similar to Figure 2 with the ordinary residuals (i.e., $e_i$). This seems to indicate that indeed the variance is not constant. Furthermore, the variance is increasing with $y$.

In Figures 3 and 4, the bottom two plots in each figure show the plots of the residuals against each of the regressors. This was not done before so they will be analyzed separately from the upper pair of plots. Both figures show on the bottom left with the residuals versus $x_1$ (a.k.a. Space time, min) that there is clearly a funnel shape. This is indicative that there is some sort of nonconstant variance. Typically, what is desired is that instead it follows some horizontal band. The funnel direction this time is actually going the opposite direction, meaning it is an inward-opening funnel where the variance decreases as $x_1$ increases. The bottom right plots in Figure 3 and 4 also look quite similar. There is some interesting behavior here also, where on the left side of each plot, the data seems to stay roughly within a horizontal band. On the right side however, the ranges of the horizontal band are noticeably wider due to two specific sample points. There is one sample point noticeably closer to the top part of the plot, and one noticeably closer to the bottom part of the plot. It is difficult to say whether here it follows the horizontal band or a funnel shape. It is possible to stay that the behavior is non-ideal and doesn't exactly reflect what's desired for a satisfactory type of plot. The upper right point corresponds with observation 8 and the bottom right corresponds with observation 12. It is interesting to note too then that these are also the furthest extreme points pointed out in Figure 1.

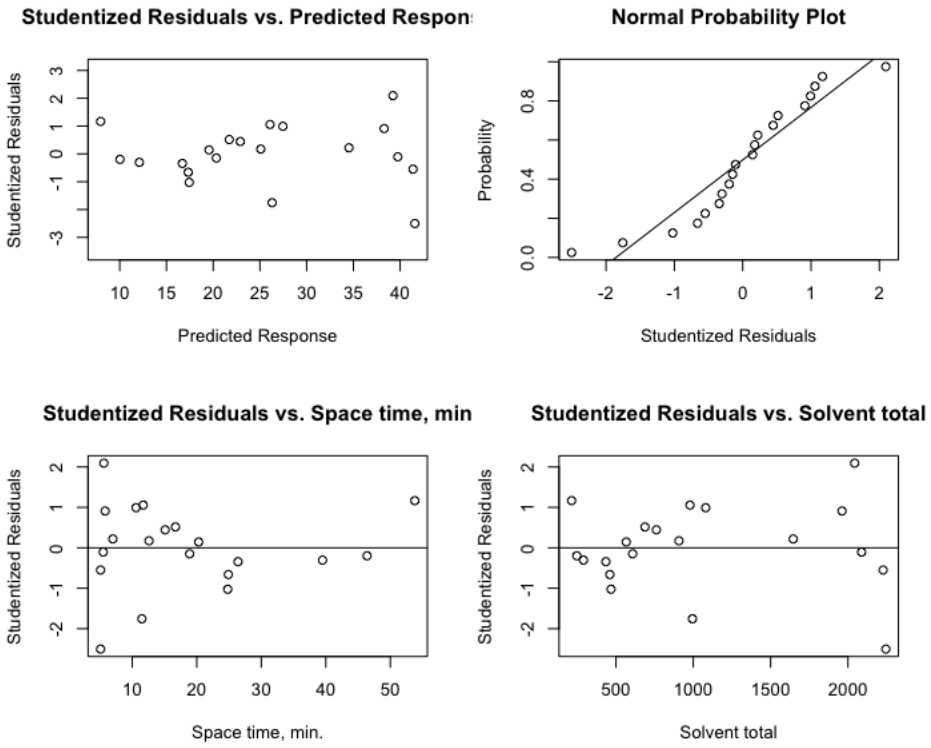


*Figure 3 A series of plots of the studentized residuals that analyzes the residuals vs. predicted response, normal probability plot, and residuals v. regressors.*

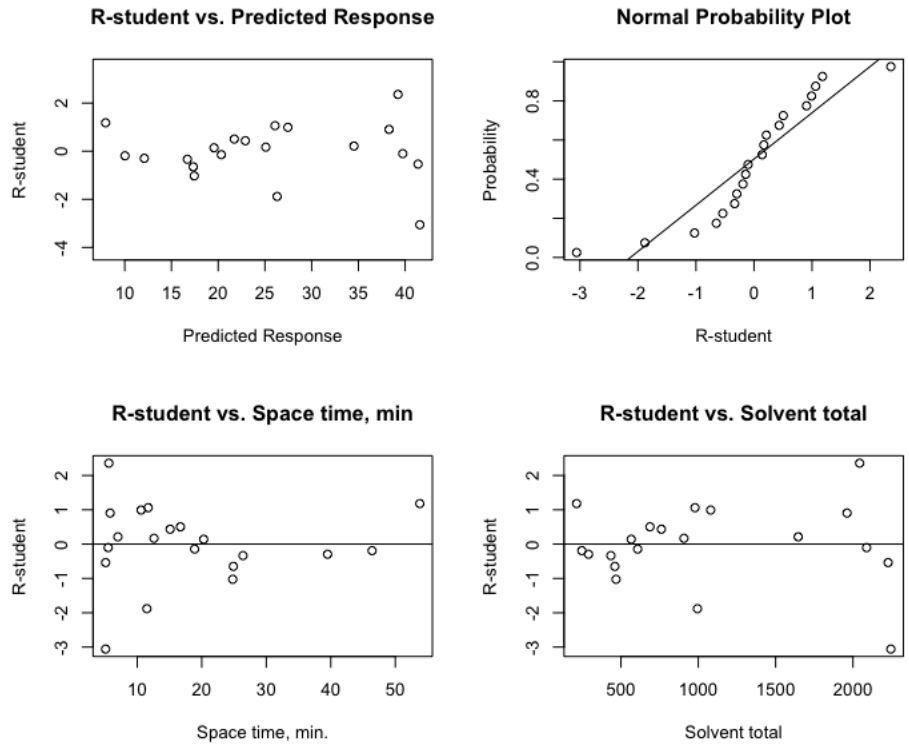


*Figure 4 A series of plots of the R-student that analyzes the residuals vs. predicted response, normal probability plot, and residuals v. regressors.*

d. Compute all other residuals (e.g., PRESS) to examine whether there are some observations that may not fit the model or potential outliers.

Ans:

The same steps that were done in part c) were done again in part d). The difference this time is that Figure 5 and 6 correspond with the standardized residuals and PRESS residuals respectively. Looking at these plots, they are quite similar to what was seen previous in Figures 1 to 4. Within the type of residuals versus predicted responses, there is the same outward-opening funnel pattern that is indicative that the variance is an increasing function of $y$. The top right plots for the normal probability plot also seem to indicate some flattening at the edges, which means that again that the residuals follow a heavier-tailed distribution compared to the normal distribution. An interesting note here however can be seen with the PRESS residuals in the upper right plot of Figure 6. It's quite obvious here that the points at the extremes are much less erratic in comparison to before. In fact, they start to follow quite closely to the line which is what's generally ideal. For example, the second furthest right point is dragged much closer to the straight line, making the furthest right point seem more normal. However, the furthest left point is still quite a bit away from the line making it seem like an outlier in comparison. This second furthest point corresponds with observation 7 and has yet to be noted as distinct from the others.

The bottom pair of plots in Figures 5 and 6 look quite similar to Figures 3 and 4. It is interesting to see that despite the different scaling of the residuals that the appearance is still highly similar. In the plot of the residuals against $x_6$ (a.k.a. solvent total), there is the same issue that there is

somewhat of a horizontal band, however observations 8 and 12 seem to again be indicating that there is not exactly the ideal appearance for the residuals relative to $x_6$. In the bottom left plot for both figures, again it shows the inward-opening funnel. This indicates instead that the residuals have decrease as $x_1$ (space time, min) increases.

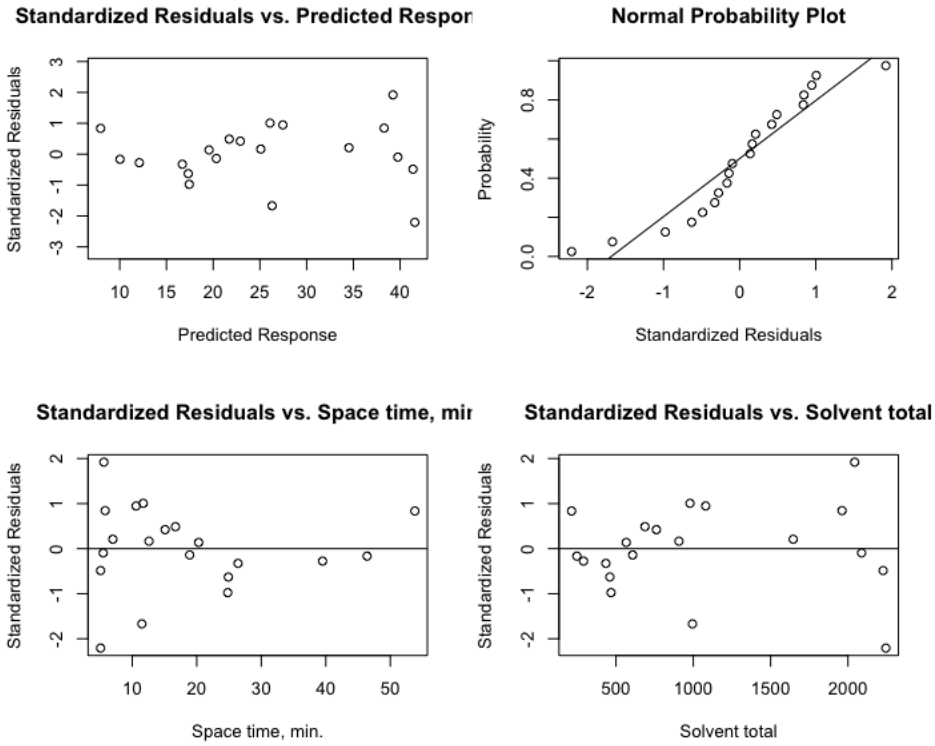


*Figure 5 A series of plots of the standardized residuals that analyzes the residuals vs. predicted response, normal probability plot, and residuals v. regressors.*

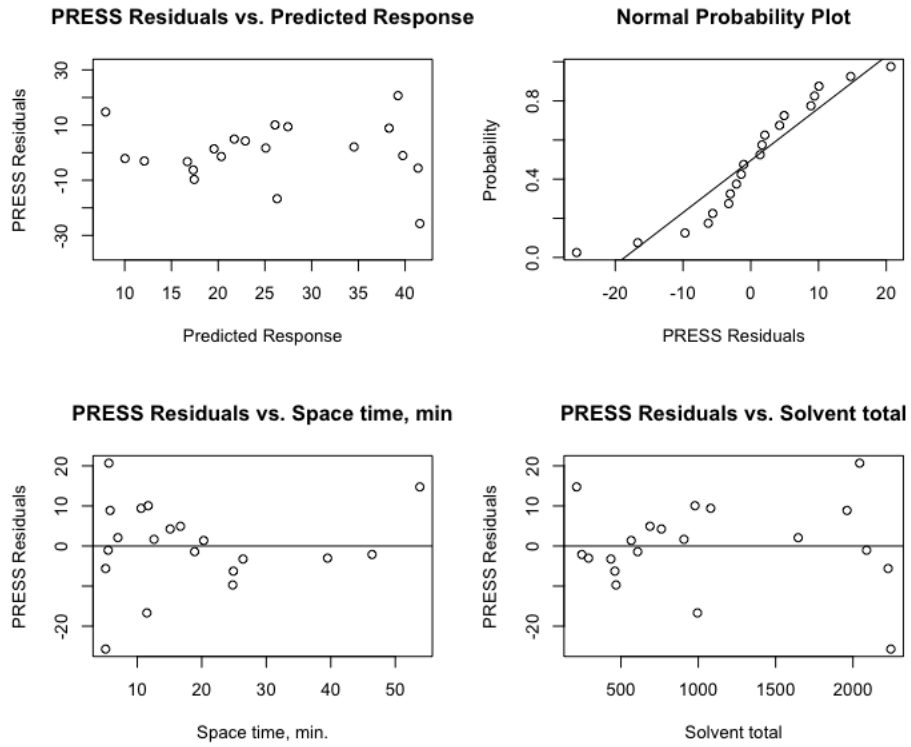


*Figure 6 A series of plots of the PRESS residuals that analyzes the residuals vs. predicted response, normal probability plot, and residuals v. regressors.*

2. Use any math/stat software (e.g., www.numbergenerator.org/randomnumbergenerator) of your choice to find a random number generator to randomly select 15 rows of Table B.4 (Property Valuation Data) on page 557 of Textbook.
    a. Perform a thorough regression analysis of $y$ on $x_4$, $x_7$, and $x_9$ including residual plots.

Ans:

The rows selected are: 1, 2, 3, 6, 8, 9, 11, 12, 13, 15, 17, 19, 21, 22, and 24. From these rows, the first-order multiple linear regression model is fit for the given regressors $x_4$, $x_7$, and $x_9$. The model for appears as follows,

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_7 + \beta_3 x_9 + \varepsilon.$$

The first goal is to test for the significance of regression of the model. The following are calculated and can be seen below in Table 1: $SS_R$, $SS_{Res}$, $SS_T$, and $F_0$. These are based on equations that can be seen on pp. 86-87 in the Textbook. Let the hypothesis be as follows,

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_1: \text{at least one } \beta_j \neq 0 \text{ for } j = 1, 2, 3.$$

After calculating the analysis-of-variance table (Table 1), the resulting critical value is $F_0 \approx 4.8249$. The critical value for $F_{3,11}$ at $\alpha = 0.05$ is $\approx 3.5874$. The F-statistic is larger than the critical value and so we reject the null hypothesis at the $\alpha = 0.05$ confidence level (*Note: This doesn't occur when* $\alpha = 0.01$.). Furthermore, the resulting p-value based on the F-statistic is $\approx 0.0222$. The conclusion then is that at least one of the $\beta_j$, for $j = 1, 2, 3$ is nonzero.

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ |
|---|---|---|---|---|
| Regression | $SS_R = 248.0808$ | $k = 3$ | $MS_R \approx 82.6936$ | $\frac{MS_R}{MS_{Res}} \approx 4.8249$ |
| Residual | $SS_{Res} = 188.5285$ | $n - k - 1 = 11$ | $MS_{Res} \approx 17.1390$ | |
| Total | $SS_T = 436.6093$ | $n - 1 = 14$ | | |

*Table 1 The above table shows the Test for Significant of Regression for the MLR model in part (a).*

Furthermore, the $R^2$ and $R^2_{Adj}$ values were calculated for the model and they can be seen below in Table 2. Ideally, we would want to see values that are closer to 1, for example around 0.9 or higher. However, it's apparent that the resulting values in Table 2 are quite low in comparison. This indicates that the model is not very strong and likely has many problems.

| $R^2$ | $R^2_{Adj}$ |
|---|---|
| $\approx 0.5682$ | $\approx 0.4504$ |

*Table 2 The table shows the resulting $R^2$ and $R^2_{Adj}$ values for the model.*

The next step will be to analyze the residuals from the model. The analysis will look at the: ordinary residuals, studentized residuals, $R$-student residuals, standardized residuals, and the PRESS residuals. The process will be similar to what was seen in the previous problem. A difference is that to examine the residuals versus the regressors, it was done only with respect to the ordinary residuals for simplicity (Figure 10). From Figures 7-9 it shows the normal probability plot (left) and residuals vs. predicted response (right). The ordinary residuals (top) and studentized residuals (bottom) are in Figure 7. The $R$-student residuals (top) and standardized residuals (bottom) are in Figure 8. The PRESS residuals are in Figure 9.

We can first look at the normal probability plots for all the different types of residuals. There are slight variations, but the overall appearance is largely the same. They seem to in fact be following the ideal appearance, where they are randomly scattered along the straight line. Towards the left side, it does slightly dip below maybe indicating a light left tail while on the right side it also slightly dips below maybe indicating a heavy right tail. These deviations however are extremely minor and never seem to stray too far from what may seem ideal. The furthest right observation is observation 15 from the original dataset. The furthest left observation is observation 3 from the original dataset.

Next we can examine the normal probability plot for the residuals for all the different types of residuals. Here there are subtle variations in the scaling of the $y$-axis, but the overall appearance remains quite similar for these plots. The overall appearance looks somewhat like a horizontal band which is ideal, but it also appears more similar to the double-bow appearance. This type of issue means that there's an indication of the errors not being constant. It is worth noting however that with only 15 observations, it's difficult to say with certainty what "shape" the residuals vs. predicted response plot shows. The textbook states that the double-bow pattern is common when the response variable is between 0 and 1, however in this case that's not what's happening.

Lastly, we can look at Figure 10 with the comparison of the residuals vs. the regressors. Looking at the top left with $x_4$ (a.k.a. Living Space ($\text{sq. ft.} \times 1000$)). In this plot, the data seems to largely be following a horizontal band which is ideal. However, again it can be seen that there's a single residual at the top which seems to be making the appearance less ideal. This is observation 15, which has been noted before on the normal probability plot. It is starting to become clearer that this observation is in fact comparatively more of an outlier than the other sample observations.

We can then look towards $x_7$ (a.k.a. Number of bedrooms) on the top right. It is worth noting that for $x_7$ and $x_9$, these are discrete variables with only 3 and 2 levels respectively. This makes it such that the analysis of these is not so good, since previously it has been such that the shapes would most likely work best with continuous data. However, the same principles will try to be applied. In this plot, there is one clear outlier with only 2 bedrooms. This coincides with observation 12 from the dataset. Also, compared to $x_7 = 4$, $x_7 = 3$ has a wider range. This wider range also includes observation 15 which has been previously noted. So even if observation 12 were for some reason removed, the plot would lack the ideal horizontal shape. However, it's difficult to say then if this would be a funnel of some sort, since there are only two levels if the third is removed. It would also be tough to say that there's some double-band shape going on, given the three levels that are available. However, that is also a possibility given the constraints.

The last plot is on the bottom left which is $x_9$ (a.k.a. Number of fireplaces). As said before, this is a discrete valued variable with only two levels. So, interpretation will be difficult in this case using the general shape principles. In this plot, it can be seen that 1/3 of the samples have one fireplace, while the other 2/3 have zero fireplaces. In this plot, it is again apparent that there's not exactly a horizontal band. Instead, with only the two levels, it's possible to guess that there's some sort of inward-opening funnel which is indicative of the variance decreasing as $x_9$ increases. The same observation 15 seems to be one of the leading points in making it appear funnel-shaped rather than some ideal horizontal band.

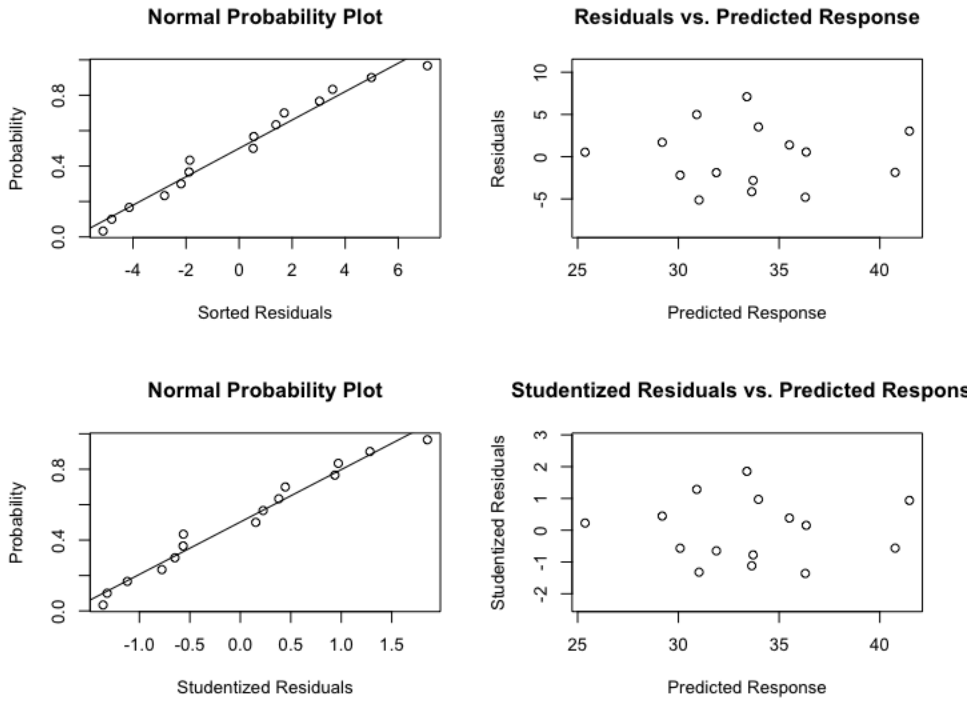


*Figure 7 The normal probability plot and residuals vs. predicted response plot for the ordinary residuals (top) and studentized residuals (bottom).*

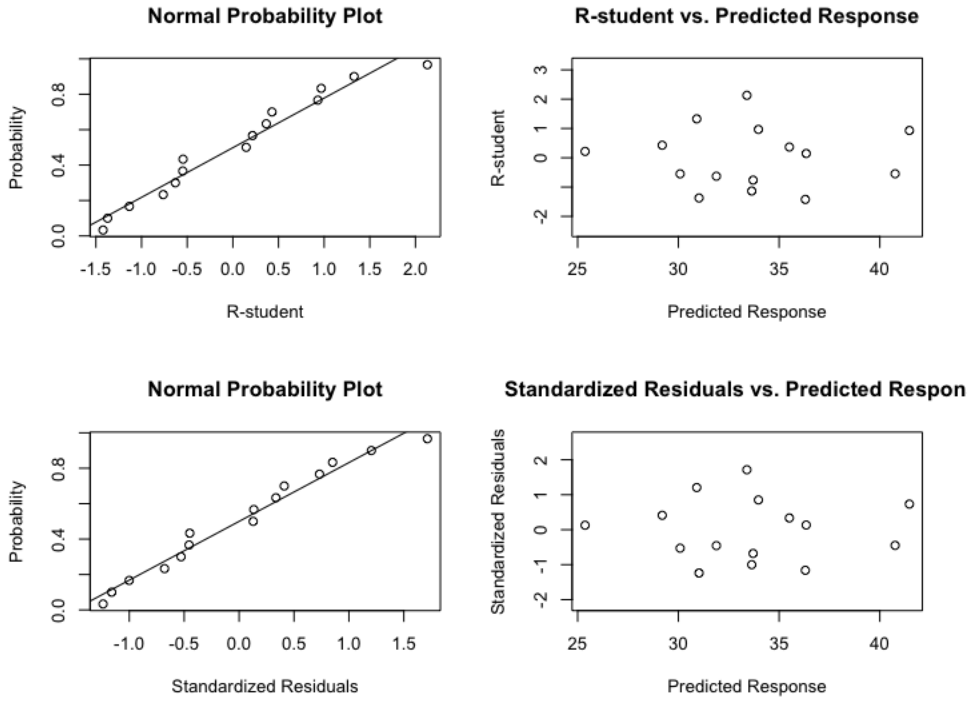


*Figure 8 The normal probability plot and residuals vs. predicted response plot for the ordinary R-student residuals (top) and standardized residuals (bottom).*

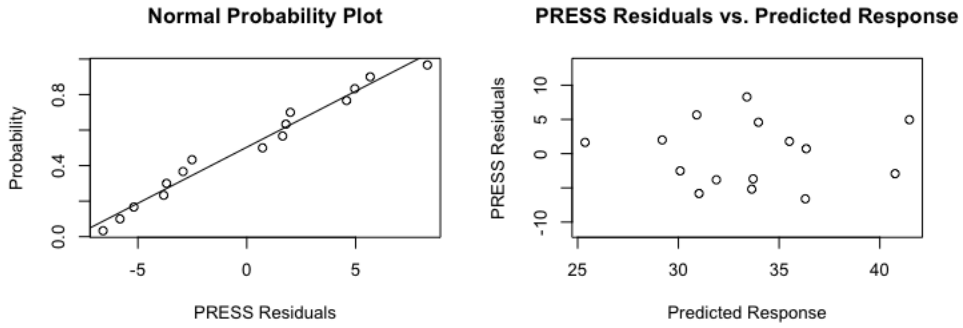


*Figure 9 The normal probability plot and residuals vs. predicted response plot for the PRESS residuals.*

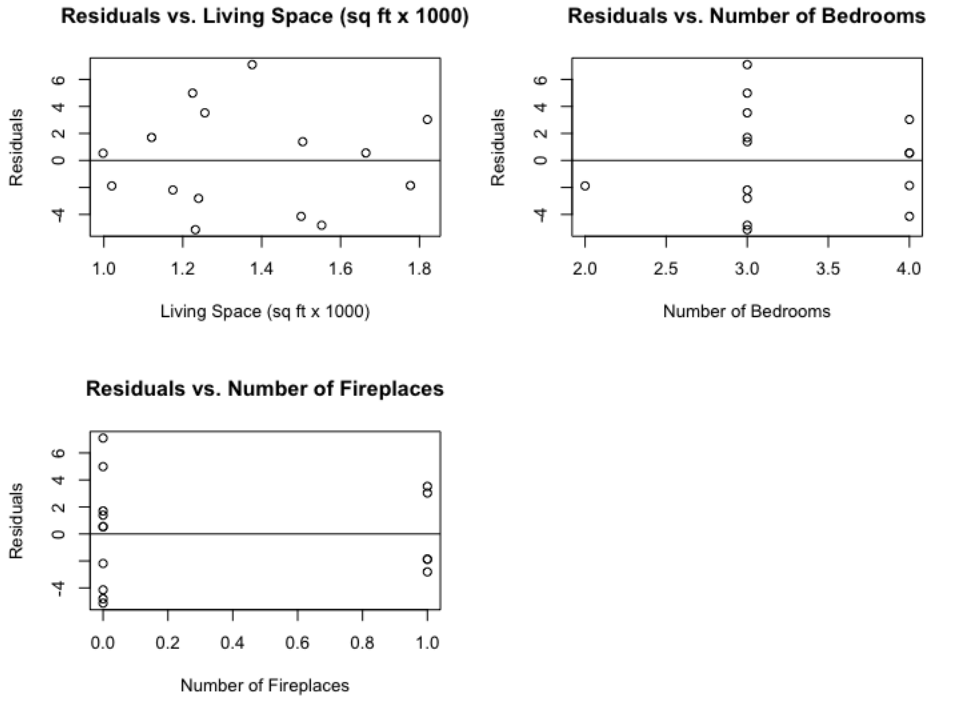


*Figure 10 The above plot shows the residuals vs. regressors for $x_4$ (top left), $x_7$ (top right), and $x_9$ (bottom left).*

b. Can an appropriate test for lack of fit be constructed? Why or why not?

Ans:

The lack of fit test requires that there are replicates of $y$ for some level of $x$. In this case, there are three duplicated values, each with a frequency of two within the randomly chosen rows. The replicated values include 25.9, 30.9, and 36.9.

Some other requirements include that there exists normality, independence, and constant-variance. Furthermore, only the first-order relationship needs to be in doubt. Given the requirements of constant-variance, it is possibly untrue given that double-bow shape of the residuals vs. fitted plots. Also, the plots of the residuals vs. regressors were difficult to interpret in the case of the discrete variables.

However, it is still possible to try it, given that the initial test for the significance of regression worked out. Furthermore, the basic requirement of there being replicates at the $y$-level for some $x$ has been met, despite there being a continuous-valued $y$ variable. In this case, it may be worth just trying it out to see what the results are and making a note of it in a report.

3. Use any math/stat software (e.g., www.numbergenerator.org/randomnumbergenerator) of your choice to find a random number generator to randomly select 7 rows of data in Problem 5.5, page 204 of Textbook. Then do (a), (b) on that page.
    a. Fit a straight-line regression model to the data and perform the standard tests for model adequacy.

Ans:

The rows selected are: 3, 4, 5, 7, 8, 10, and 12. From these rows, the first-order simple linear regression model is fit for $y = \text{Defects per } 10{,}000$ on the given regressors $x_1 = \text{Weeks}$. The simple linear regression (SLR) model appears as follows,

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

The first goal is to test for the significance of regression of the model. The following are calculated and can be seen below in Table 1: $SS_R$, $SS_{Res}$, $SS_T$, and $F_0$. These are based on equations that can be seen on pp. 86-87 in the Textbook. Let the hypothesis be as follows,

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0.$$

After calculating the analysis-of-variance table (Table 4), the resulting critical value is $F_0 \approx 23.5866$. The critical value for $F_{1,5}$ at $\alpha = 0.01$ is $\approx 16.2582$. The F-statistic is larger than the critical value and so we reject the null hypothesis at the $\alpha = 0.01$ confidence level. Furthermore, the resulting p-value based on the F-statistic is $\approx 0.0046$. The conclusion then is that $\beta_1$ is nonzero.

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ |
|---|---|---|---|---|
| Regression | $SS_R = 2{,}921.403$ | $k = 1$ | $MS_R \approx 2{,}921.403$ | $\frac{MS_R}{MS_{Res}} \approx 23.5866$ |
| Residual | $SS_{Res} = 619.2918$ | $n - k - 1 = 5$ | $MS_{Res} \approx 123.8584$ | |
| Total | $SS_T = 3{,}540.694$ | $n - 1 = 6$ | | |

*Table 4 The above table shows the Test for Significant of Regression for the SLR model in part (a).*

The resulting $R^2$ is $\approx 0.8251$, which is not ideal. This indicates that there are likely some issues with the regression model. In general, we would prefer that the $R^2$ in an SLR model is above 0.9. Furthermore, since it's and SLR model, then the $R^2_{Adj}$ won't be calculated, since that's more suited for MLR models.

The next steps will be to perform some residual analysis. The same process that was done previously in Problem 3 will be repeated here. That is, we will first plot the normal probability plot and the residuals vs. fitted plot for each of the various types of residuals. Then, we will analyze the residuals vs. regressor plot for each of the different residuals.

In figures 11 to 13, it shows the normal probability plot on the left and the residuals vs. predicted response on the right for the following types of residuals (from top to bottom): ordinary residuals, studentized residuals, $R$-student residuals, standardized residuals, and PRESS

residuals. The normal probability plots seem to indicate that the residuals do follow a normal distribution. They all seem scattered evenly in a random pattern around the straight line. There are slight variations, for example in the *R*-student plot which seems to show instead that there's some different odd pattern. However, with so few observations, it's difficult to see if what's appearing is some other pattern or just random behavior.

On the right side with the residuals vs. predicted response, the pattern seems a bit different. Looking through the different types of residuals, it seems that there's some obvious pattern where it dips down and then goes back upward. It looks rather non-linear. Generally, it indicates that some other regressor variables are required. This makes sense given that there's only one regressor in the model. However, we know that the regressor is in fact a timeseries type of variable. It is counting the number of weeks since the last furnace overhaul. Although the observations are sampled randomly from the period of 14 weeks, the sequential ordering still seems meaningful. Given the non-linear appearance in the right-hand side, it seems that since a new regressor can't be added, perhaps a transformation is suitable.

Looking to Figure 13 and 14, it shows the residuals against the regressor (Weeks). The pattern is interestingly quite similar to what is seen with the residuals vs. predicted response. Looking at the predicted response, they're a steadily increasing value, similar to how the Weeks variable is after being ordered. The textbook mentions that for timeseries data, there is a possibility of autocorrelation, which needs to be handled in a special manner. It is possible that with more weeks of data, that the up and down autocorrelation pattern would be more evident. Here, there is simply one dip, but with more time the pattern could possibly repeat itself.

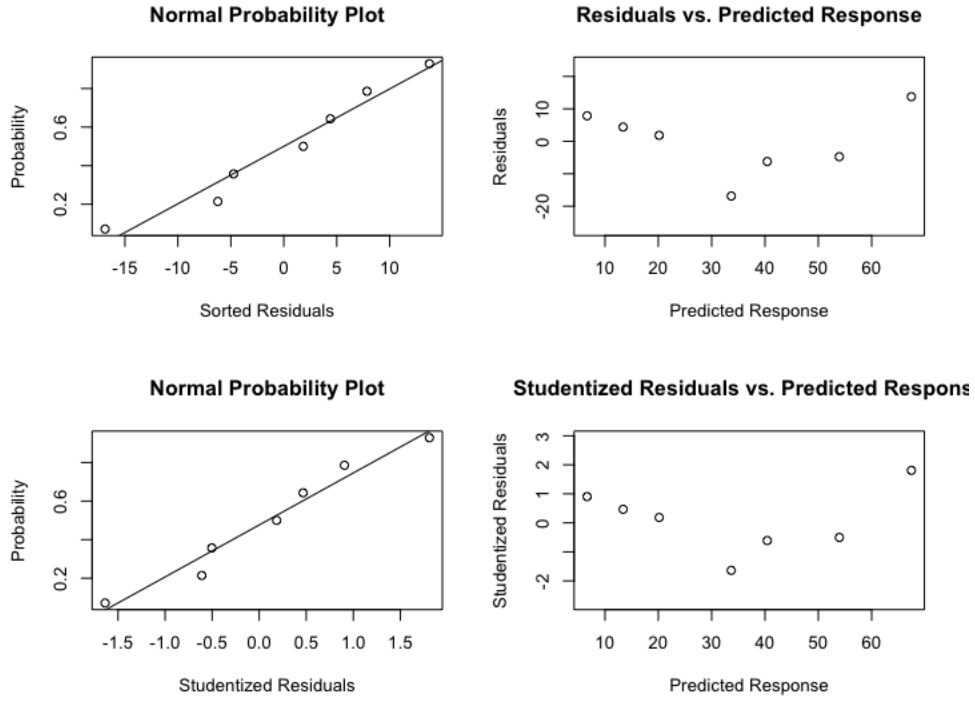


*Figure 11 The normal probability plot and residuals vs. predicted response plot for the ordinary residuals (top) and studentized residuals (bottom).*

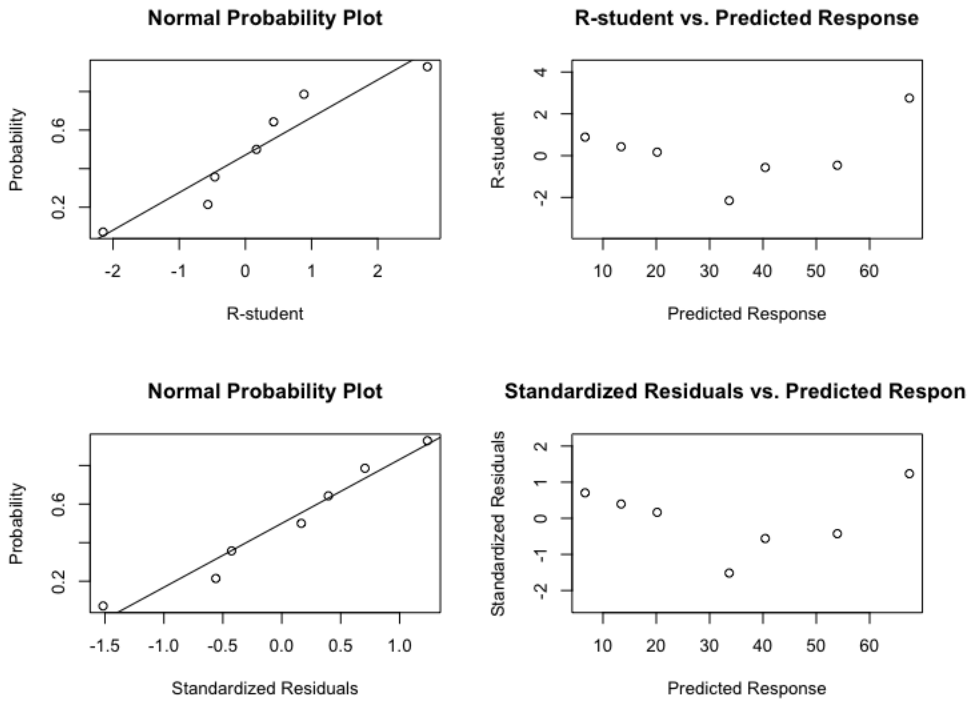


*Figure 12 The normal probability plot and residuals vs. predicted response plot for the ordinary R-student residuals (top) and standardized residuals (bottom).*

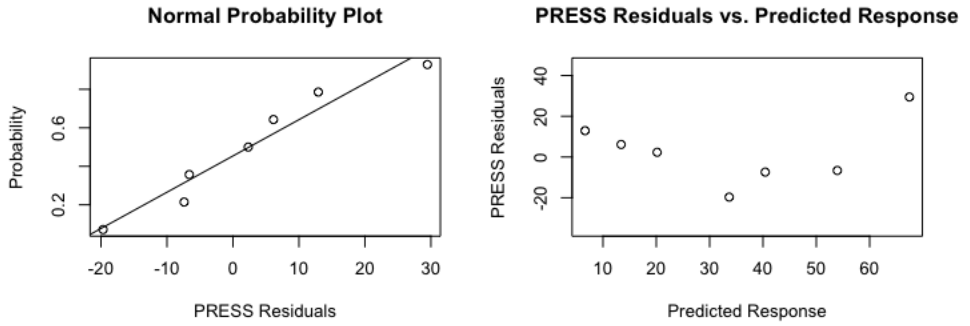


*Figure 13 The normal probability plot and residuals vs. predicted response plot for the PRESS residuals.*

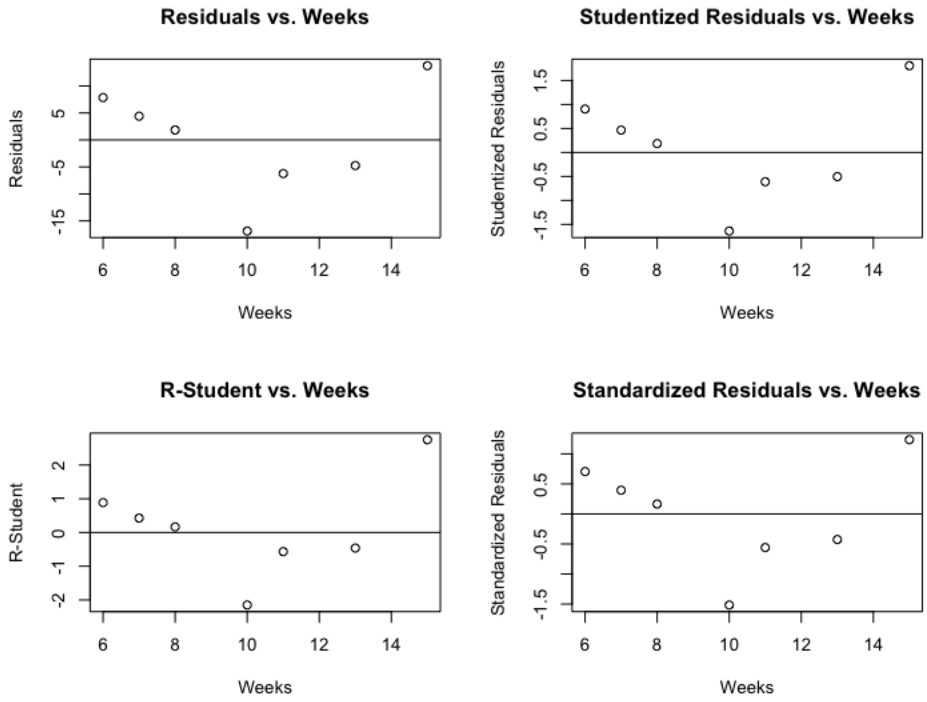


*Figure 14 The above shows the residuals vs. regressor plot for the ordinary residuals (top left), studentized residuals (top right), R-student residuals (bottom left), and standardized residuals (bottom right).*

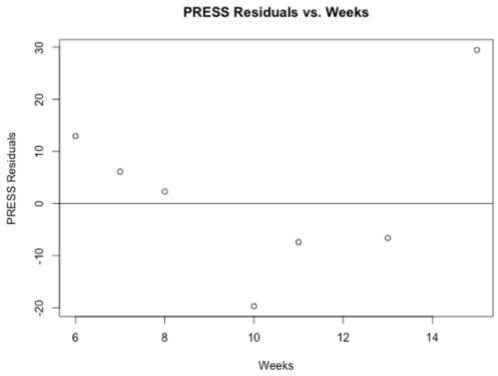


*Figure 15 The above figure shows the residuals vs. regressor for the remaining residual, PRESS residuals which was not shown in Figure 14.*

b. Suggest an appropriate transformation to eliminate the problems encountered in part a. Fit the transformed model and check for adequacy.

Ans:

Although the initial test of the SLR model in part a) showed that the coefficient $\beta_1$ is nonzero, looking at other factors made it apparent that there are some issues. The main problem is that the regressor itself is a time-related variable and so there is the major issue of autocorrelation, which can violate the assumptions of the model.

Some of the transformations mention in the textbook are in the case for when the variance is proportional to some other factor. This for example could be seen if the residual plots showed some other behavior, for instance if there was some funnel shape which could be indicative of the variance increasing with $y$. However, this type of pattern was not exactly seen, instead a non-linear behavior was seen. Furthermore, it is known to be a timeseries type of data. Therefore, the transformation is likely going to be tricky.

An initial plot simply showing $y$ vs. $x_1$, can be seen below in Figure 16 on the left. The data is clearly non-linear. It could also be misleading to think that the variance is simply proportional to something like $E(y)$, since there seems to be some other behavior going on. An alternative approach would be to try to linearize the model. Based on Figure 5.4 on p.177 in the textbook, it seems instead that the plot follows possibly a function of the form $y = \beta_0 x_1^{\beta_1}$, where the corresponding transformation is $y' = \log y$ and $x_1' = \log x_1$. The transformed variables can be seen on the right of Figure 16. The difference is not too great, but it does seem to balance out the points across the line slightly.

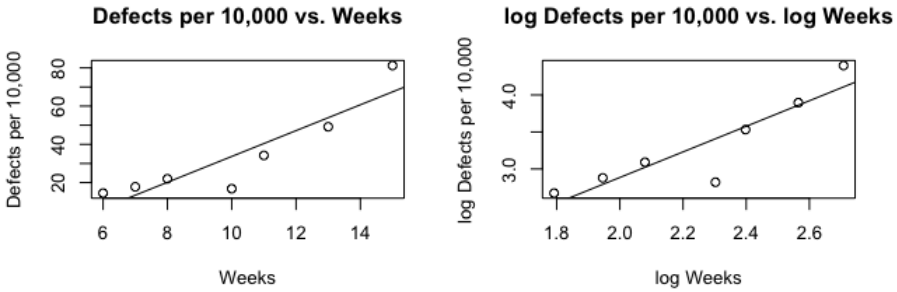


*Figure 16 A plot showing the y (Defects per 10,000) vs. x (Weeks) on the left and on the right is the same plot after taking the natural log of both variables.*

The transformed model appears as follows,

$$y' = \beta_0 + \beta_1 x_1' + \varepsilon.$$

The first goal is to test for the significance of regression of the model. The following are calculated and can be seen below in Table 1: $SS_R$, $SS_{Res}$, $SS_T$, and $F_0$. These are based on equations that can be seen on pp. 86-87 in the Textbook. Let the hypothesis be as follows,

$$H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0.$$

After calculating the analysis-of-variance table (Table 5), the resulting critical value is $F_0 \approx 21.3520$ The critical value for $F_{1,5}$ at $\alpha = 0.01$ is $\approx 16.2582$. The F-statistic is larger than the critical value and so we reject the null hypothesis at the $\alpha = 0.01$ confidence level. Furthermore, the resulting p-value based on the F-statistic is $\approx 0.0057$. The conclusion then is that $\beta_1$ is nonzero. The above results are the same as what was seen before, except the $F_0$ is slightly smaller.

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ |
|---|---|---|---|---|
| Regression | $SS_R \approx 1.9840$ | $k = 1$ | $MS_R \approx 1.9840$ | $\frac{MS_R}{MS_{Res}} \approx 21.3520$ |
| Residual | $SS_{Res} \approx 0.4646$ | $n - k - 1 = 5$ | $MS_{Res} \approx 0.0929$ | |
| Total | $SS_T \approx 2.4485$ | $n - 1 = 6$ | | |

*Table 5 The above table shows the Test for Significant of Regression for the SLR model in part (a).*

The resulting $R^2$ is $\approx 0.8251$, which is similar to what was seen before. So, it is interesting to see that despite the improved appearance of the $y$ vs. $x$ plot, the statistics so far don't seem to indicate that there's a great improvement.

In Figures 17 to 19 it shows the various types of residuals that we have seen previously, except they have been recalculated to consider the log transformations. We can again look to the left-hand side at the normal probability plots. There seems to be a significant unimprovement in the appearance compared to before. The data no longer follows the line clearly, and the observation on the far left is quite distant from the rest of the data. This pattern appears to be due to the log transformation which has greatly changed the resulting values of the data. The pattern seems to indicate that there could possibly be some sort of strong skew to the data. It is quite unclear though, since what's apparent is not really seen exactly in any of the plots on Figure 4.3 on p.136 of the textbook.

A difference however can be seen in the right-hand side with the residuals vs. predicted response. In this case, all the residuals seem to be following much closer to some horizontal movement within some fixed range. The exception however is the single point that is significantly lower. This could be an outlier that is messing up the image that the behavior is ideal. It's worth noting that this is the same observation that is an outlier in the normal probability plot. This is observation 7 from the original dataset. Previously, with the untransformed data, this observation was not so distinctly different from the others. However, in the relative position on the charts it is the same data point (i.e., the furthest left on the normal probability plot and the furthest down on the residuals vs. fitted plot).

Looking at Figures 20 and 21, it seems quite similar to before, where it mirrored quite closely the residuals vs. fitted plot. In here, it is more inclined for the viewer to think that the furthest down observation (observation 7) is an outlier. Previously, it seemed to more or less be one of the outer edge points. Here, it seems more like it is the only outlier in comparison to the others. Another interesting point though is that it's one of two points that are below 0 in terms of the residual value. There is one more that is below the line, but it is extremely close to it. The others are clearly above the line. It is possible that with more sample points that this distribution would spread out more evenly. However, since it's timeseries data, it's likely that there's an unseen pattern due to the lack of observations. So, the logic of removing a time point doesn't really make sense, since it follows a bigger picture trend. In fact, removing it to fit a linear picture would itself be misleading and likely lead to errors if such a model was pursued.

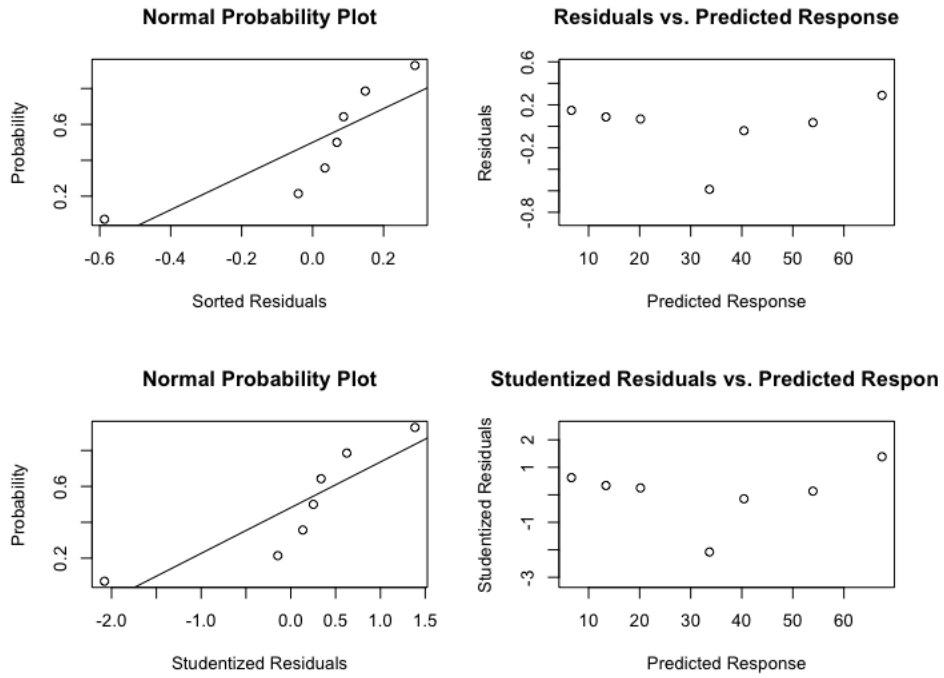


*Figure 17 The normal probability plot and residuals vs. predicted response plot for the ordinary residuals (top) and studentized residuals (bottom) (after log transformation).*

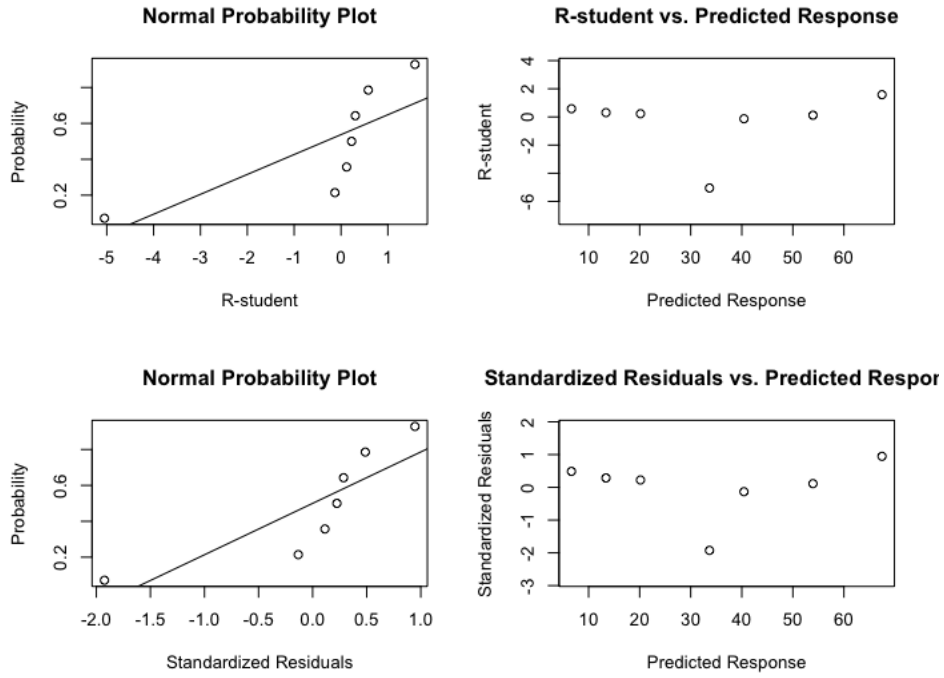


*Figure 18 The normal probability plot and residuals vs. predicted response plot for the ordinary R-student residuals (top) and standardized residuals (bottom) (after transformation).*

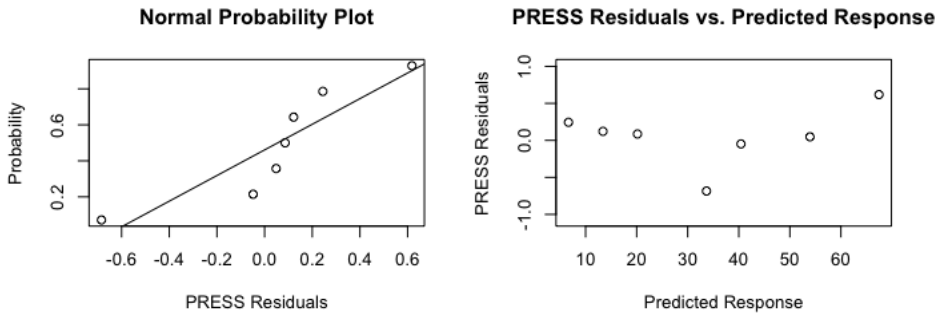


*Figure 19 The normal probability plot and residuals vs. predicted response plot for the PRESS residuals (after log transformation).*

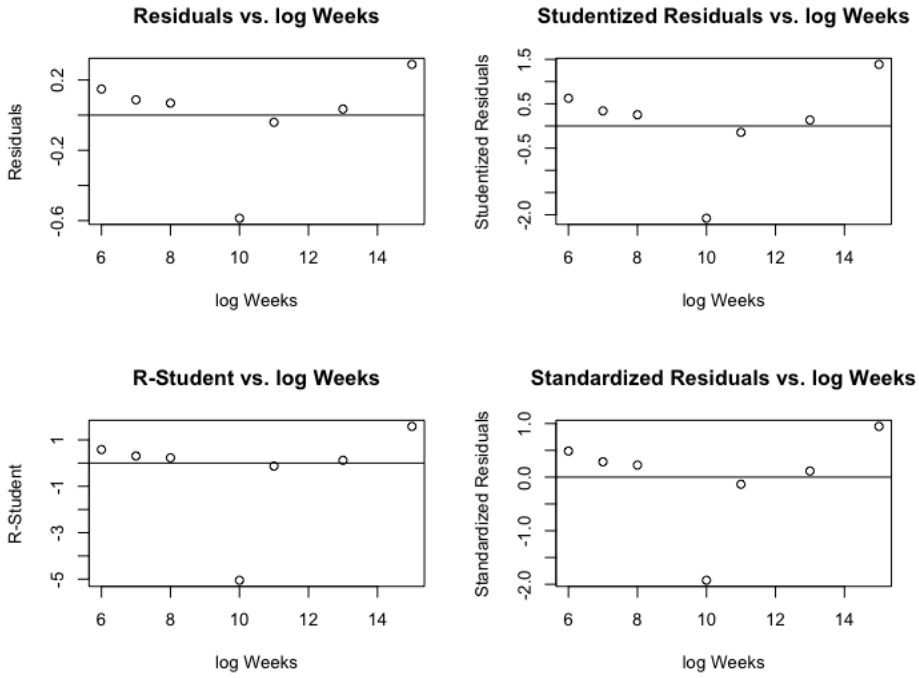


*Figure 20*

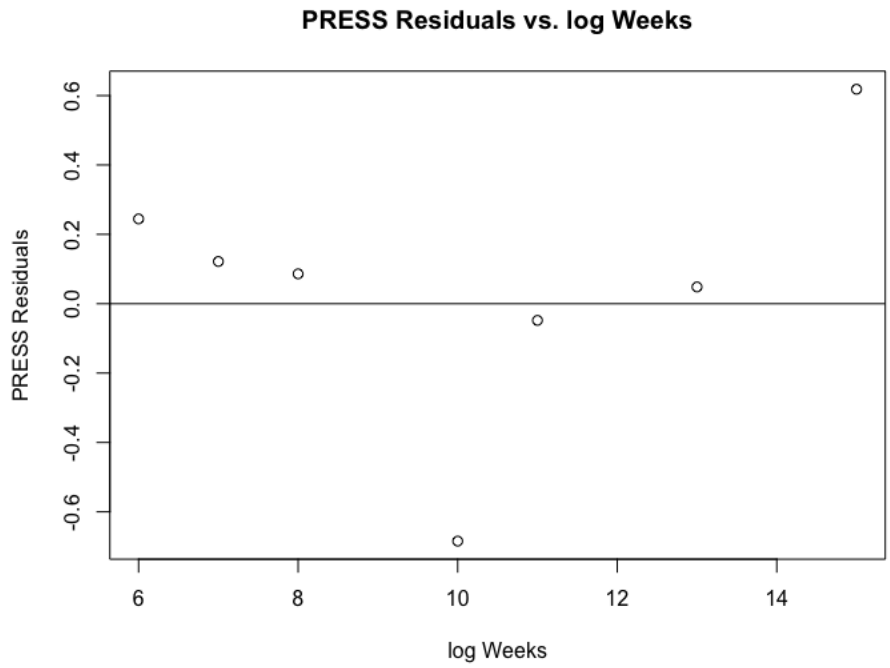


*Figure 21*

## Code Appendix

```
library(MPV)

### Problem 1
n <- 20
set.seed(1); chosen_rows <- sort(sample(seq(1,27), n))
df <- MPV::table.b5
df <- df[chosen_rows,c(1, 2, 7)]
ones <- rep(1, n)
y <- df[,1]
X <- cbind(ones, df[,c(2,3)])

beta_hat_calc <- function(X, y) {
 X <- as.matrix(X)
 beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
 return(beta_hat)
}
H_calc <- function(X) {
 X <- as.matrix(X)
 H <- X %*% solve(t(X) %*% X) %*% t(X)
 return(H)
}
y_hat_calc <- function(H, y) {
 y_hat <- H %*% y
 return(y_hat)
}
e_calc <- function(y, y_hat) {
 e <- y - y_hat
 return(e)
}

beta_hat <- beta_hat_calc(X=X, y=y)
H <- H_calc(X)
y_hat <- y_hat_calc(H=H, y=y)
```

```
e <- e_calc(y=y, y_hat=y_hat)

#### part (a)
norm_prob_plot <- function(residual_var, x_label,
             main_title = 'Normal Probability Plot',
             y_label = 'Probability', n_size=n) {
 ones <- rep(1, n)
 sorted_residuals <- sort(residual_var)
 cumulative_probability <- (1:n_size - 0.5) / n_size
 plot(sorted_residuals, cumulative_probability, main = main_title,
    xlab = x_label,
    ylab = y_label)
 X_temp <- cbind(ones, sorted_residuals)
 beta_hat_temp <- beta_hat_calc(X=X_temp,y=cumulative_probability)
 abline(beta_hat_temp)
}
norm_prob_plot(residual_var = e, x_label = 'Sorted Residuals')

order(e, decreasing = FALSE)
e[order(e, decreasing = FALSE)]

### part (b)
res_vs_fitted_plot <- function(residual_var,
                  main_title,
                  y_label,
                  x_label = 'Predicted Response',
                  pred_response = y_hat) {
 plot(pred_response, residual_var, main = main_title,
    xlab = x_label,
    ylab = y_label,
    ylim = c(min(residual_var)-sd(residual_var),
         max(residual_var)+sd(residual_var)))
}
res_vs_fitted_plot(residual_var = e,
         main_title = 'Residuals vs. Predicted Response',
         y_label = 'Residuals')

### part (c)
# studentized residuals
SS_Res_calc <- function(y, beta_hat, X) {
 SS_Res <- (t(y) %*% y) - (t(beta_hat) %*% t(X) %*% y)
 return(SS_Res)
}
SS_Res <- SS_Res_calc(y = y, beta_hat = beta_hat, X = X)
p <- ncol(X)
MS_Res <- SS_Res / (n - p)
H_diag <- diag(H)
studentized_residuals <- sapply(1:n, function(x) {
 e[x] / sqrt(MS_Res * (1 - H_diag[x]))
})

# res vs. fitted, norm prob res, res vs. x1, res vs. x6
res_vs_regressor <- function(residual_var,
                  main_title1, main_title2,
                  ylabel,
                  X_df=X) {
 x1 <- X_df[,2]; x6 <- X_df[,3]
 plot(x1, residual_var,
    main = main_title1,
    ylab = ylabel,
    xlab = 'Space time, min.')
 abline(h = 0)

 plot(x6, residual_var,
    main = main_title2,
```

```
   ylab = ylabel,
   xlab = 'Solvent total')
 abline(h = 0)
}

par(mfrow = c(2,2))
res_vs_fitted_plot(residual_var = studentized_residuals,
        main_title = 'Studentized Residuals vs. Predicted Response',
        y_label = 'Studentized Residuals')
norm_prob_plot(residual_var = studentized_residuals, x_label = 'Studentized Residuals')
res_vs_regressor(residual_var = studentized_residuals,
       main_title1 = 'Studentized Residuals vs. Space time, min',
       main_title2 = 'Studentized Residuals vs. Solvent total',
       ylabel = 'Studentized Residuals')

# Analyze the outliers
head(order(x6, decreasing = TRUE), 5)
studentized_residuals[12] # bottom right
studentized_residuals[1]
studentized_residuals[18]
studentized_residuals[8] # top right
studentized_residuals[11]

# R-student
S_squared <- sapply(1:n, function(x) {
 ((n - p) * MS_Res - ((e[x]^2) / (1 - H_diag[x]))) / (n - p - 1)
})

R_student_res <- sapply(1:n, function(x) {
 e[x] / sqrt(S_squared[x] * (1 - H_diag[x]))
})

par(mfrow = c(2,2))
res_vs_fitted_plot(residual_var = R_student_res,
        main_title = 'R-student vs. Predicted Response',
        y_label = 'R-student')
norm_prob_plot(residual_var = R_student_res, x_label = 'R-student')
res_vs_regressor(residual_var = R_student_res,
       main_title1 = 'R-student vs. Space time, min',
       main_title2 = 'R-student vs. Solvent total',
       ylabel = 'R-student')

### part (d)
# standardized residuals
standardized_res <- sapply(1:n, function(x) e[x] / sqrt(MS_Res))

par(mfrow = c(2,2))
res_vs_fitted_plot(residual_var = standardized_res,
        main_title = 'Standardized Residuals vs. Predicted Response',
        y_label = 'Standardized Residuals')
norm_prob_plot(residual_var = standardized_res, x_label = 'Standardized Residuals')
res_vs_regressor(residual_var = standardized_res,
       main_title1 = 'Standardized Residuals vs. Space time, min',
       main_title2 = 'Standardized Residuals vs. Solvent total',
       ylabel = 'Standardized Residuals')


# PRESS residuals
PRESS_res <- sapply(1:n, function(x) e[x] / (1 - H_diag[x]))

par(mfrow = c(2,2))
res_vs_fitted_plot(residual_var = PRESS_res,
        main_title = 'PRESS Residuals vs. Predicted Response',
        y_label = 'PRESS Residuals')
norm_prob_plot(residual_var = PRESS_res, x_label = 'PRESS Residuals')
```

```
res_vs_regressor(residual_var = PRESS_res,
       main_title1 = 'PRESS Residuals vs. Space time, min',
       main_title2 = 'PRESS Residuals vs. Solvent total',
       ylabel = 'PRESS Residuals')


### Problem 2
df <- MPV::table.b4
set.seed(2); chosen_rows <- sort(sample(seq(1,24), 15))
df <- df[chosen_rows,c(1, 5, 8, 10)]
n <- nrow(df)
ones <- rep(1, n)
y <- df[,1]
X <- cbind(ones, df[,c(2:4)])

beta_hat <- beta_hat_calc(X=X, y=y)
H <- H_calc(X)
y_hat <- y_hat_calc(H=H, y=y)
e <- e_calc(y=y, y_hat=y_hat)
### part (a)
SS_T_calc <- function(y) {
 n <- length(y)
 SS_T <- (t(y) %*% y) - ((sum(y)^2) / n)
 return(SS_T)
}
SS_R_calc <- function(beta_hat, X, y) {
 n <- length(y)
 SS_R <- (t(beta_hat) %*% t(X) %*% y) - ((sum(y)^2) / n)
 return(SS_R)
}

SS_Res <- SS_Res_calc(beta_hat = beta_hat, X = X, y = y)
SS_T <- SS_T_calc(y = y)
SS_R <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)

SS_Res == SS_T - SS_R
SS_Res; SS_T; SS_R
k <- 3; p <- k + 1
MS_R <- SS_R / k
MS_Res <- SS_Res / (n - k - 1)
F_0 <- MS_R / MS_Res

alpha <- 0.05
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))
pf(q = F_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

r_squared_calc <- function(SS_R, SS_T) {
 r_squared <- SS_R / SS_T
 return(r_squared)
}
adj_r_squared_calc <- function(SS_Res, SS_T, n, p) {
 adj_r_squared <- 1 - ((SS_Res / (n - p)) / (SS_T / (n - 1)))
 return(adj_r_squared)
}

r_squared <- r_squared_calc(SS_R = SS_R, SS_T = SS_T)
adj_r_squared <- adj_r_squared_calc(
 SS_Res = SS_Res, SS_T = SS_T, n = n, p = p)

# ordinary residuals
par(mfrow = c(2,2))
norm_prob_plot(residual_var = e, x_label = 'Sorted Residuals')
res_vs_fitted_plot(residual_var = e,
        main_title = 'Residuals vs. Predicted Response',
        y_label = 'Residuals')
```

```
order(e, decreasing = TRUE)

# studentized residuals
H_diag <- diag(H)
studentized_residuals <- sapply(1:n, function(x) {
 e[x] / sqrt(MS_Res * (1 - H_diag[x]))
})
norm_prob_plot(residual_var = studentized_residuals, x_label = 'Studentized Residuals')
res_vs_fitted_plot(residual_var = studentized_residuals,
        main_title = 'Studentized Residuals vs. Predicted Response',
        y_label = 'Studentized Residuals')

# R-student residuals
par(mfrow = c(2,2))
S_squared <- sapply(1:n, function(x) {
 ((n - p) * MS_Res - ((e[x]^2) / (1 - H_diag[x]))) / (n - p - 1)
})

R_student_res <- sapply(1:n, function(x) {
 e[x] / sqrt(S_squared[x] * (1 - H_diag[x]))
})

norm_prob_plot(residual_var = R_student_res, x_label = 'R-student')
res_vs_fitted_plot(residual_var = R_student_res,
        main_title = 'R-student vs. Predicted Response',
        y_label = 'R-student')

# standardized residuals
standardized_res <- sapply(1:n, function(x) e[x] / sqrt(MS_Res))
norm_prob_plot(residual_var = standardized_res, x_label = 'Standardized Residuals')
res_vs_fitted_plot(residual_var = standardized_res,
        main_title = 'Standardized Residuals vs. Predicted Response',
        y_label = 'Standardized Residuals')

# PRESS residuals
par(mfrow = c(2,2))
PRESS_res <- sapply(1:n, function(x) e[x] / (1 - H_diag[x]))
norm_prob_plot(residual_var = PRESS_res, x_label = 'PRESS Residuals')
res_vs_fitted_plot(residual_var = PRESS_res,
        main_title = 'PRESS Residuals vs. Predicted Response',
        y_label = 'PRESS Residuals')

x4 <- X[,2]; x7 <- X[,3]; x9 <- X[,4]
par(mfrow = c(2,2))
plot(x4, e,
  main = 'Residuals vs. Living Space (sq ft x 1000)',
  ylab = 'Residuals',
  xlab = 'Living Space (sq ft x 1000)')
abline(h = 0)

plot(x7, e,
  main = 'Residuals vs. Number of Bedrooms',
  ylab = 'Residuals',
  xlab = 'Number of Bedrooms')
abline(h = 0)

plot(x9, e,
  main = 'Residuals vs. Number of Fireplaces',
  ylab = 'Residuals',
  xlab = 'Number of Fireplaces')
abline(h = 0)


### part (b)
```

```
data.frame(table(y))

### Problem 3
### part (a)
df <- MPV::p5.5
set.seed(3); chosen_rows <- sort(sample(seq(1,14), 7))
df <- df[chosen_rows,]
y <- df[,1]
n <- nrow(df)
ones <- rep(1, n)
X <- cbind(ones, df[,2])

beta_hat <- beta_hat_calc(X=X, y=y)
H <- H_calc(X)
y_hat <- y_hat_calc(H=H, y=y)
e <- e_calc(y=y, y_hat=y_hat)

SS_Res <- SS_Res_calc(beta_hat = beta_hat, X = X, y = y)
SS_T <- SS_T_calc(y = y)
SS_R <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)

SS_Res == SS_T - SS_R
SS_Res; SS_T; SS_R
k <- 1; p <- k + 1
MS_R <- SS_R / k
MS_Res <- SS_Res / (n - k - 1)
F_0 <- MS_R / MS_Res

alpha <- 0.01
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))
pf(q = F_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

r_squared <- r_squared_calc(SS_R = SS_R, SS_T = SS_T)

# ordinary residuals
par(mfrow = c(2,2))
norm_prob_plot(residual_var = e, x_label = 'Sorted Residuals')
res_vs_fitted_plot(residual_var = e,
        main_title = 'Residuals vs. Predicted Response',
        y_label = 'Residuals')

order(e, decreasing = TRUE)

# studentized residuals
H_diag <- diag(H)
studentized_residuals <- sapply(1:n, function(x) {
 e[x] / sqrt(MS_Res * (1 - H_diag[x]))
})
norm_prob_plot(residual_var = studentized_residuals, x_label = 'Studentized Residuals')
res_vs_fitted_plot(residual_var = studentized_residuals,
        main_title = 'Studentized Residuals vs. Predicted Response',
        y_label = 'Studentized Residuals')

# R-student residuals
par(mfrow = c(2,2))
S_squared <- sapply(1:n, function(x) {
 ((n - p) * MS_Res - ((e[x]^2) / (1 - H_diag[x]))) / (n - p - 1)
})

R_student_res <- sapply(1:n, function(x) {
 e[x] / sqrt(S_squared[x] * (1 - H_diag[x]))
})

norm_prob_plot(residual_var = R_student_res, x_label = 'R-student')
res_vs_fitted_plot(residual_var = R_student_res,
```

```
        main_title = 'R-student vs. Predicted Response',
        y_label = 'R-student')

# standardized residuals
standardized_res <- sapply(1:n, function(x) e[x] / sqrt(MS_Res))
norm_prob_plot(residual_var = standardized_res, x_label = 'Standardized Residuals')
res_vs_fitted_plot(residual_var = standardized_res,
        main_title = 'Standardized Residuals vs. Predicted Response',
        y_label = 'Standardized Residuals')

# PRESS residuals
par(mfrow = c(2,2))
PRESS_res <- sapply(1:n, function(x) e[x] / (1 - H_diag[x]))
norm_prob_plot(residual_var = PRESS_res, x_label = 'PRESS Residuals')
res_vs_fitted_plot(residual_var = PRESS_res,
        main_title = 'PRESS Residuals vs. Predicted Response',
        y_label = 'PRESS Residuals')

par(mfrow = c(2,2))
plot(X[,2], e,
  main = 'Residuals vs. Weeks',
  ylab = 'Residuals',
  xlab = 'Weeks')
abline(h = 0)

plot(X[,2], studentized_residuals,
  main = 'Studentized Residuals vs. Weeks',
  ylab = 'Studentized Residuals',
  xlab = 'Weeks')
abline(h = 0)

plot(X[,2], R_student_res,
  main = 'R-Student vs. Weeks',
  ylab = 'R-Student',
  xlab = 'Weeks')
abline(h = 0)

plot(X[,2], standardized_res,
  main = 'Standardized Residuals vs. Weeks',
  ylab = 'Standardized Residuals',
  xlab = 'Weeks')
abline(h = 0)

par(mfrow = c(1,1))
plot(X[,2], PRESS_res,
  main = 'PRESS Residuals vs. Weeks',
  ylab = 'PRESS Residuals',
  xlab = 'Weeks')
abline(h = 0)

### part (b)
par(mfrow = c(2,2))
plot(df$weeks, df$defects,
  main = 'Defects per 10,000 vs. Weeks',
  xlab = 'Weeks', ylab = 'Defects per 10,000')
abline(beta_hat)

log_y <- log(df$defects)
log_x <- log(df$weeks)
plot(log_x, log_y,
  main = 'log Defects per 10,000 vs. log Weeks',
  xlab = 'log Weeks', ylab = 'log Defects per 10,000')
X_log <- cbind(ones, log_x)
beta_hat_log <- beta_hat_calc(X=X_log, y=log_y)
H_log <- H_calc(X_log)
```

```
y_hat_log <- y_hat_calc(H=H_log, y=log_y)
e_log <- e_calc(y=log_y, y_hat=y_hat_log)
abline(beta_hat_log)

SS_Res <- SS_Res_calc(beta_hat = beta_hat_log, X = X_log, y = log_y)
SS_T <- SS_T_calc(y = log_y)
SS_R <- SS_R_calc(beta_hat = beta_hat_log, X = X_log, y = log_y)

SS_Res == SS_T - SS_R
SS_Res; SS_T; SS_R
k <- 1; p <- k + 1
MS_R <- SS_R / k
MS_Res <- SS_Res / (n - k - 1)
F_0 <- MS_R / MS_Res

alpha <- 0.01
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))
pf(q = F_0, df1 = k, df2 = (n - k - 1), lower.tail = FALSE)

r_squared <- r_squared_calc(SS_R = SS_R, SS_T = SS_T)

# ordinary residuals
par(mfrow = c(2,2))
e <- e_log
norm_prob_plot(residual_var = e, x_label = 'Sorted Residuals')
res_vs_fitted_plot(residual_var = e,
        main_title = 'Residuals vs. Predicted Response',
        y_label = 'Residuals')

order(e, decreasing = TRUE)

# studentized residuals
H_diag <- diag(H)
studentized_residuals <- sapply(1:n, function(x) {
 e[x] / sqrt(MS_Res * (1 - H_diag[x]))
})
norm_prob_plot(residual_var = studentized_residuals, x_label = 'Studentized Residuals')
res_vs_fitted_plot(residual_var = studentized_residuals,
        main_title = 'Studentized Residuals vs. Predicted Response',
        y_label = 'Studentized Residuals')

# R-student residuals
par(mfrow = c(2,2))
S_squared <- sapply(1:n, function(x) {
 ((n - p) * MS_Res - ((e[x]^2) / (1 - H_diag[x]))) / (n - p - 1)
})

R_student_res <- sapply(1:n, function(x) {
 e[x] / sqrt(S_squared[x] * (1 - H_diag[x]))
})

norm_prob_plot(residual_var = R_student_res, x_label = 'R-student')
res_vs_fitted_plot(residual_var = R_student_res,
        main_title = 'R-student vs. Predicted Response',
        y_label = 'R-student')

# standardized residuals
standardized_res <- sapply(1:n, function(x) e[x] / sqrt(MS_Res))
norm_prob_plot(residual_var = standardized_res, x_label = 'Standardized Residuals')
res_vs_fitted_plot(residual_var = standardized_res,
        main_title = 'Standardized Residuals vs. Predicted Response',
        y_label = 'Standardized Residuals')

# PRESS residuals
par(mfrow = c(2,2))
```

```
PRESS_res <- sapply(1:n, function(x) e[x] / (1 - H_diag[x]))
norm_prob_plot(residual_var = PRESS_res, x_label = 'PRESS Residuals')
res_vs_fitted_plot(residual_var = PRESS_res,
        main_title = 'PRESS Residuals vs. Predicted Response',
        y_label = 'PRESS Residuals')

par(mfrow = c(2,2))
plot(X[,2], e,
  main = 'Residuals vs. log Weeks',
  ylab = 'Residuals',
  xlab = 'log Weeks')
abline(h = 0)

plot(X[,2], studentized_residuals,
  main = 'Studentized Residuals vs. log Weeks',
  ylab = 'Studentized Residuals',
  xlab = 'log Weeks')
abline(h = 0)

plot(X[,2], R_student_res,
  main = 'R-Student vs. log Weeks',
  ylab = 'R-Student',
  xlab = 'log Weeks')
abline(h = 0)

plot(X[,2], standardized_res,
  main = 'Standardized Residuals vs. log Weeks',
  ylab = 'Standardized Residuals',
  xlab = 'log Weeks')
abline(h = 0)

par(mfrow = c(1,1))
plot(X[,2], PRESS_res,
  main = 'PRESS Residuals vs. log Weeks',
  ylab = 'PRESS Residuals',
  xlab = 'log Weeks')
abline(h = 0)
```