Rachel Prokopius

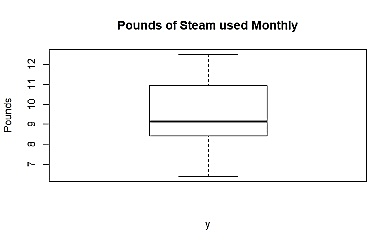
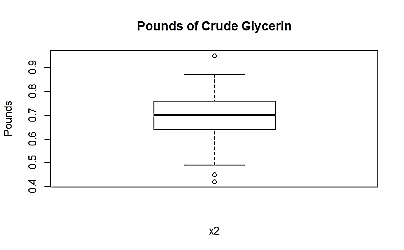
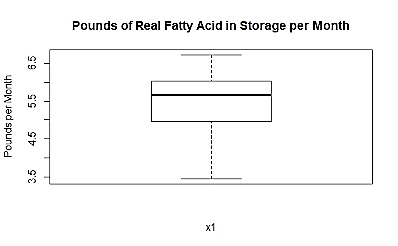
Regression Analysis

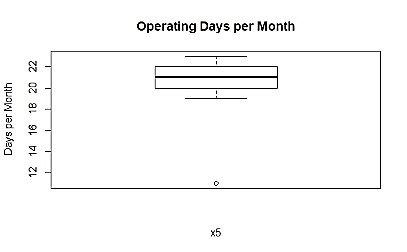
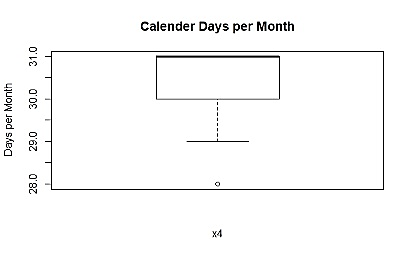
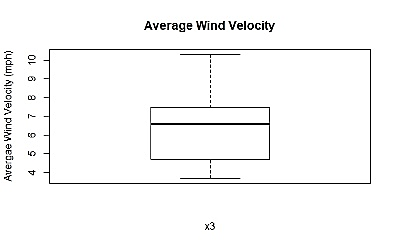
Final Exam

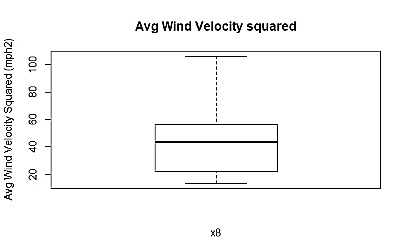
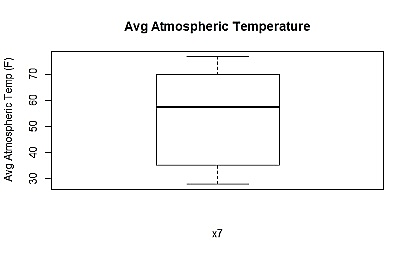
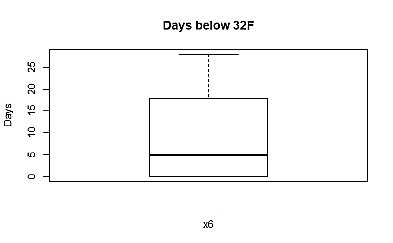
21 April 2020

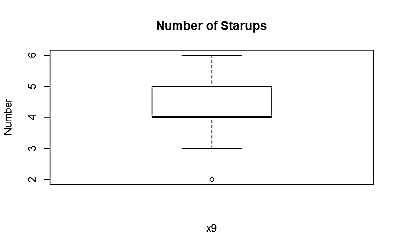
1. The following are box plots of the y variable and x variables for observations at a steam plant. Each boxplot is separate because each variable has different units.

**Figure 1**: Boxplots depicting the spread of the y variable and the regressor variables of observations at a steam plant.

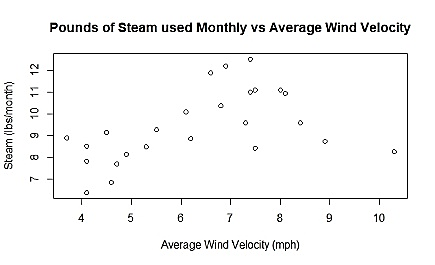
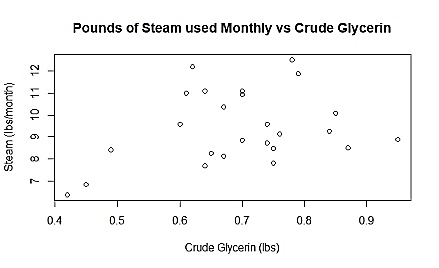
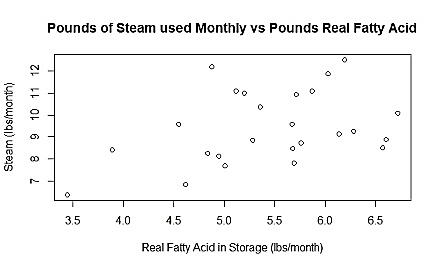


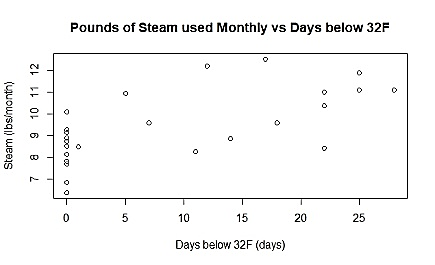
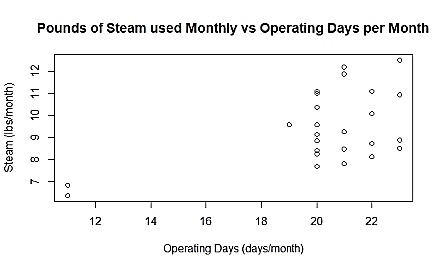
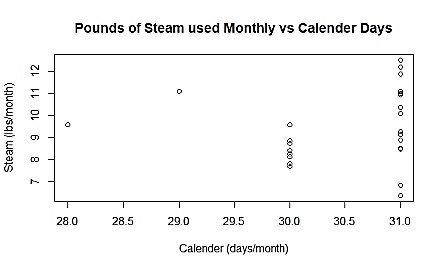


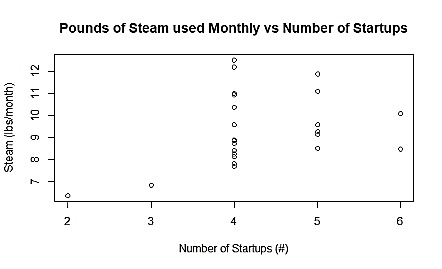
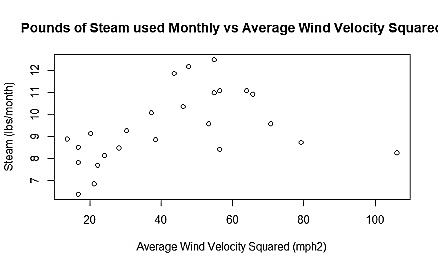
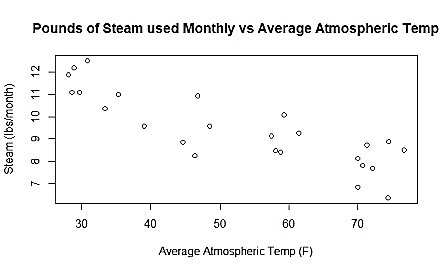


One of the requirements for a data set so that statistical analyses are valid is normal distribution, which means equal distribution of data points around the mean that have roughly 95% of the data points within two standard deviations of the mean and no outliers. Based on the box plots from this data set, certain variables appear to fulfill this requirement and others appear not to. The pounds of steam used monthly (y), pounds of real fatty acid in storage per month (x1), average wind velocity (x3), average atmospheric temperature (x7) and average wind velocity squared (x8) appear to have a normal distribution and no outliers. On the other hand, pounds of crude glycerin (x2) has outliers at both extremes, calendar days per month (x4), operating days per month (x5) and number of startups (x9) have high extreme outliers, and calendar days per month (x4) and days below 32F (x6) are missing extremes for the boxplot in the data set. Therefore, certain variables appear to fulfill one of the requirements for statistical analysis while others do not.

1. **Figure 2**: The following are scatterplots relating steam (lbs/month) to each regressor variable accounted for in observations from a steam plant.







Based on the scatter plots, certain regressor variables seem to have a linear relationship with steam (lbs/month) while others do not. Real fatty acid (lbs/month) and average atmospheric temperature (F) seem to have a fairly strong linear relationship with steam (lbs/month). Crude glycerin (lbs), average wind velocity (mph) and average wind velocity squared (mph2) seem to have more of a parabolic relationship with steam (lbs/month). Days below 32F may have a linear relationship with steam (lbs/month) but there are quite a few data points where there is a decent spread of steam (lbs/month) with zero days below 32F, so this regressor does not seem particularly linear. The same could be said for operating days per month for slightly different reasons; though fewer operating days seem to correlate with lower steam in lbs/month, more operating days correlates with a wide range of steam in lbs. Finally, calendar days and number of startups do not really seem to have a linear relationship with steam in lbs/month, but it is difficult to tell because, like operating days and number of days below 32F, it is a discrete numerical variable compared to a continuous numerical variable. Further analysis for there variables are definitely warranted.

1. The full model for relating the y variable and regressors for the steam model is as follows (**note: coefficients that are measured in the same units as the y variable (lbs/month) remain unitless in order to correctly relate the regressor to the y variable**):

**Steam (lbs/month) = 0.700(x1) -1.868[month-1](x2) +1.140[lbs\*hr\*month-1\*miles-1](x3) + 0.123[lbs\*days-1](x4) +0.180[lbs\*days-1](x5) -0.018[lbs\*days-1\*month-1](x6) -0.077[lbs\*ºF-1\*month-1](x7) -0.086[lbs\*hr2\*miles-2\*month-1](x8) -0.346[lbs\*number of startups-1\*month-1](x9) +1.761 (lbs/month)**

The summary statistics for the full model area as follows:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.76116 6.96637 0.253 0.803847

x1 0.70084 0.56248 1.246 0.231880

x2 -1.86794 4.12852 -0.452 0.657421

x3 1.14038 0.74289 1.535 0.145591

x4 0.12253 0.20374 0.601 0.556546

x5 0.17957 0.08060 2.228 0.041619 \*

x6 -0.01831 0.02440 -0.751 0.464557

x7 -0.07734 0.01652 -4.681 0.000295 \*\*\*

x8 -0.08626 0.05178 -1.666 0.116445

x9 -0.34610 0.20979 -1.650 0.119777

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

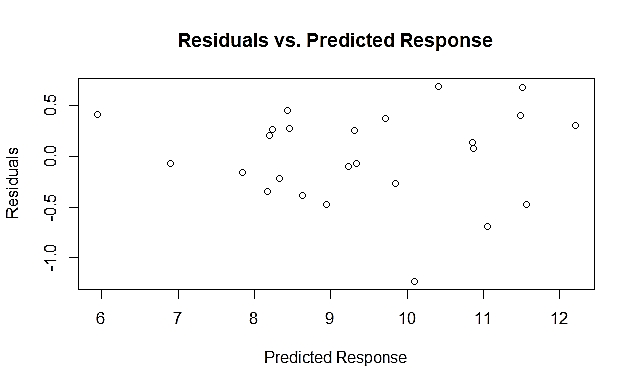
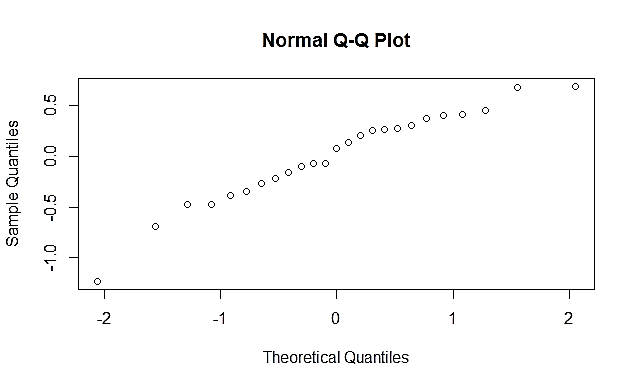
Residual standard error: 0.5673 on 15 degrees of freedom

Multiple R-squared: 0.9242, Adjusted R-squared: 0.8788

F-statistic: 20.33 on 9 and 15 DF, p-value: 7.576e-07

Though the adjusted R2 value is fairly high (0.8788), only two of the regressors (operating days per month and average atmospheric temperature) are significant, so the model likely could use some alteration.

**Figure 3.1:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model and a plot of the residuals from the full model versus the y-values the constructed model predicts.



**Table 3.1**: The following is a table of the standardized residuals for the full model relating steam (lbs/month) to all of the provided regressor variables.

**1 2 3 4 5 6**

0.2542561 1.4748225 0.6363896 0.6141057 -0.1488749 0.6723499

**7 8 9 10 11 12**

1.1986003 0.5678766 -0.7316112 -0.2035156 -1.7552012 1.6696907

**13 14 15 16 17 18**

0.8503303 0.6887397 0.1596926 -0.5628649 0.8309716 -0.4543152

**19 20 21 22 23 24**

-0.2101913 1.1567576 -0.3102859 -1.1720115 -2.4195457 -1.3487168

**25**

-1.0444930

**Table 3.2:** The following is a table of the studentized residuals for the full model relating steam (lbs/month) to all of the provided regressor variables.

**1 2 3 4 5 6**

0.2461658 1.5409084 0.6232825 0.6008842 -0.1439332 0.6595664

**7 8 9 10 11 12**

1.2177407 0.5546151 -0.7197621 -0.1968867 -1.9022436 1.7877397

**13 14 15 16 17 18**

0.8420427 0.6761632 0.1544091 -0.5496144 0.8219363 -0.4419615

**19 20 21 22 23 24**

-0.2033638 1.1709832 -0.3007313 -1.1879701 -2.9935535 -1.3899885

**25**

-1.0479030

The standardized and studentized residuals are scaled residuals that can indicate if the data set is normally distributed and has a constant variance. If the absolute value of any of these residuals is greater than 3, it could indicate an outlier in the data set. An absolute value of a studentized residual or a R-student residual that is greater than 3 indicates the associated data point’s y-coordinate is likely an outlier. None of the standardized and studentized residuals from the full model have an absolute value greater than 3 (though the studentized residual for the 23rd y value is very close, at 2.994) (Table 3.2). It is likely that there are no glaring outliers created by this model.

**Table 3.3:** The following is a table of the VIF values for the full model relating steam (lbs/month) to all of the provided regressor variables.

**x1 x2 x3 x4 x5 x6**

15.746595 20.137114 126.625618 1.836626 4.411920 4.695013

**x7 x8 x9**

6.067426 107.590891 2.385046

The standardized and studentized residuals are scaled residuals that can indicate if the data set is normally distributed and has a constant variance. If the absolute value of any of these residuals is greater than 3, it could indicate an outlier in the data set. An absolute value of a studentized residual or a R-student residual that is greater than 3 indicates the associated data point’s y-coordinate is likely an outlier. None of the standardized and studentized residuals from the full model have an absolute value greater than 3 (though the studentized residual for the 23rd y value is very close, at 2.994) (Table 3.2). It is likely that there are no glaring outliers created by this model.

The VIF values indicate that there is very little collinearity between variables. VIF values less than 1 indicate no collinearity, VIF values between 1 and 5 indicate moderate collinearity and VIF values of greater than 5 indicate high collinearity. Because the VIF values for all of the predictor variables are close to 1, it indicates that there is very little collinearity between the variables and it should not be a problem in the model.

1. Figure 12.1e: The following is a normal probability plot of residuals for the following multiple regression model that relates creatinine concentration, age and weight to creatinine clearance (in $1000’s):

**Creatinine clearance (in $1000’s)= -39.029(creatinine concentration) -0.760(age) +0.766(weight) +$120.594**



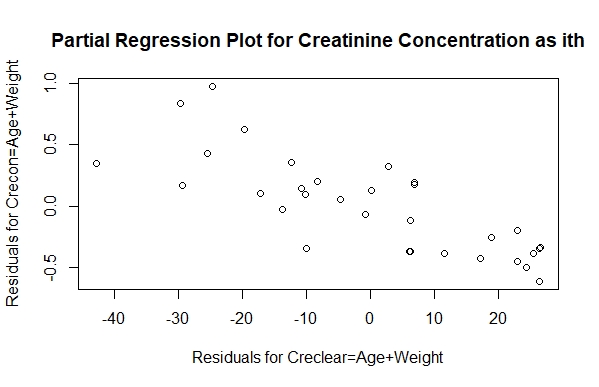
The normal probability plot of residuals, when the model follows the rules of normal distribution, independent distribution and equal variance, should have all the residuals fall on a straight line. As can be seen in the above plot, the normal QQ plot follows more of a curvilinear pattern, especially towards the higher theoretical quantile levels. This indicates that the dataset may not be normally distributed and may require transformation in order to meet the general requirements a model needs for statistical analysis to be valid (normal distribution, independent distribution and equal variance).

Figure 12.1f: The following is a plot of the residuals from the multiple linear regression from part c above versus the y-values the constructed model predicts.



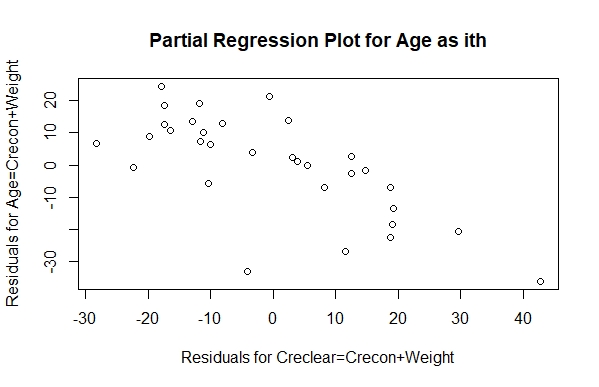
If the data set is normally distributed, independently distributed and has a constant variance, there will be no pattern in the data points in the constructed residuals vs. predicted response plot. The plot points seem to be fairly randomly distributed past the predicted response of 60, but there appears to be a weak trend for predicted responses less than 60, with a lower predicted response correlating with a higher residual value. This, similar to the suggestion of 12.1e, indicates that transformation of the dataset may be needed in order to meet the general requirements a model needs for analysis to be valid (normal distribution, independent distribution and equal variance).

Figure 12.1g: The following is a partial regression plot of residuals for creatinine concentration (Crecon) explained by the age and weight versus the residuals for the creatinine clearance (Creclear) (in $1000’s) explained by the age and weight:



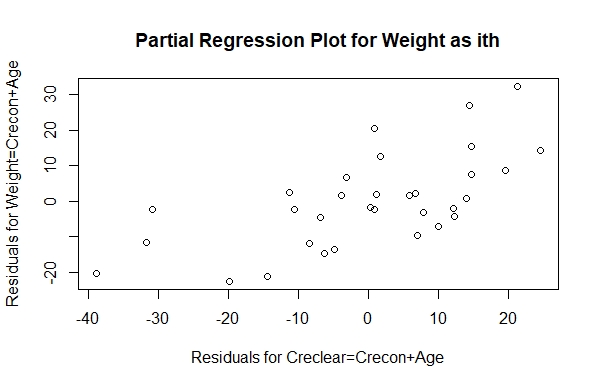
This partial regression plot treats the creatinine concentration (Crecon) as the ith term that is removed from the model, and then the model is re-run and its residuals are plotted against the residuals of the model with the ith term as the response variable. The plot helps to determine if the removal of the ith term (here, the creatinine concentration) greatly affects the outcome of the model. The fairly distinct pattern of the residuals plotted here suggests that the ith term explains much of the model’s variance after accounting for the tested regressors (here, age and weight) and the ith term being removed does have a substantial effect on the model and should remain in the model.

Figure 12.1h: The following is a partial regression plot of residuals for age explained by the creatinine concentration (Crecon) and weight versus the residuals for the creatinine clearance (Creclear) (in $1000’s) explained by the creatinine concentration (Crecon) and weight:



This partial regression plot treats the age as the ith term that is removed from the model, and then the model is re-run and its residuals are plotted against the residuals of the model with the ith term as the response variable. The plot helps to determine if the removal of the ith term (here, the age) greatly affects the outcome of the model. The fairly distinct pattern of the residuals plotted here suggests that the ith term explains much of the model’s variance after accounting for the tested regressors (here, creatinine concentration (Crecon) and weight) and the ith term being removed does have a substantial effect on the model and should remain in the model.

Figure 12.1i: The following is a partial regression plot of residuals for weight explained by the creatinine concentration (Crecon) and age versus the residuals for the creatinine clearance (Creclear) (in $1000’s) explained by the creatinine concentration (Crecon) and age:



This partial regression plot treats the weight as the ith term that is removed from the model, and then the model is re-run and its residuals are plotted against the residuals of the model with the ith term as the response variable. The plot helps to determine if the removal of the ith term (here, the weight) greatly affects the outcome of the model. The fairly distinct pattern of the residuals plotted here suggests that the ith term explains much of the model’s variance after accounting for the tested regressors (here, creatinine concentration (Crecon) and age) and the ith term being removed does have a substantial effect on the model and should remain in the model.

In conclusion, according to the partial regression plots created for the effect of creatinine concentration, age and weight on creatinine clearance (in $1000’s) indicates that all three regressors explain a decent amount of the variance of the model, and therefore should all remain in the model. Because the model, as seen by Figures 12.1e and 12.1f, does not fulfill the requirements of normal distribution, independent distribution and equal variance, and it appears that all three regressor variables are needed in the model, transformation of the model is likely needed to reconfigure the dataset in order to ready it for statistical analysis.

1. The theoretical transformed linear model relating creatinine concentration, age and weight to creatinine clearance (in $1000’s) is as follows:

**Ln(creatinine clearance (in $1000’s))= [-0.6974\*ln(creatinine concentration)] + [0.7738\*ln(140-age)] + [0.7425\*ln(weight)] -$2.1036**

**Scatter Plots:**

Figure 12.1j: The following is a scatter plot depicting the natural log of creatinine clearance (in $1000’s) versus the natural log of creatinine concentration.



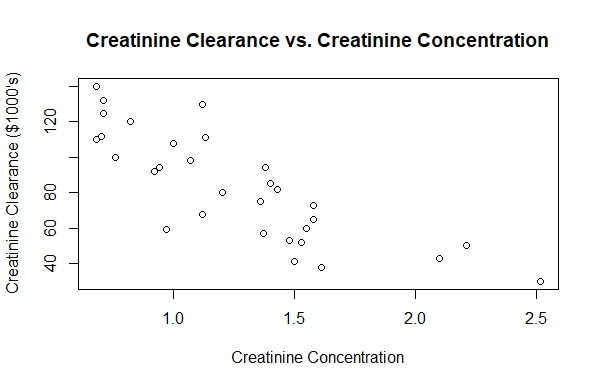
Figure 12.1k: The following is a scatter plot depicting the natural log of creatinine clearance (in $1000’s) versus the natural log of (140-age).



Figure 12.1l: The following is a scatter plot depicting the natural log of creatinine clearance (in $1000’s) versus the natural log of weight.

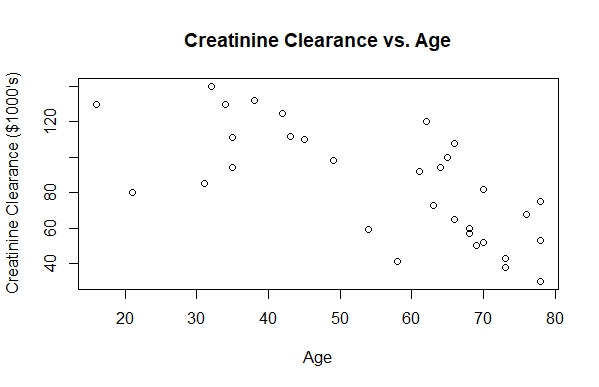


The following is a side-by-side comparison of left: Figure 12.1a (a scatter plot depicting the creatinine clearance (in $1000’s) versus the creatinine concentration) and right: Figure 12.1j (a scatter plot depicting the natural log of creatinine clearance (in $1000’s) versus the natural log of creatinine concentration)



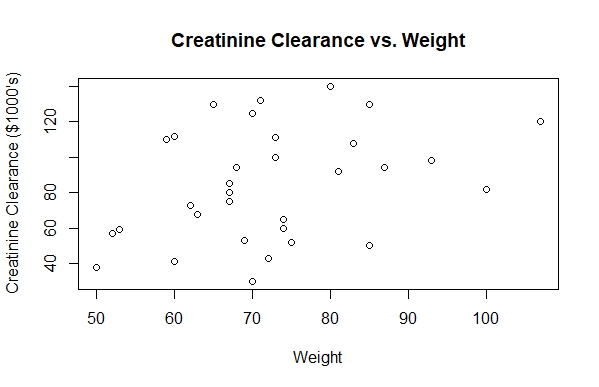
After taking the natural log of creatinine clearance (in $1000’s) and the natural log of creatinine concentration, the to variables seem to have a more linear relationship than before the transformation (Figure 12.1j). Therefore, transformation may have been necessary for these two variables in order to generate a better linear model.

The following is a side-by-side comparison of left: Figure 12.1b (a scatter plot depicting the creatinine clearance (in $1000’s) versus Age) and right: Figure 12.1k (a scatter plot depicting the natural log of creatinine clearance (in $1000’s) versus the natural log of (140-Age)):

Because of the addition of the negative sign in the transformed age variable (Figure 12.1k), the relationship between age and creatinine clearance is flipped. The transformation actually appears to make the relationship between the two variables more curvilinear than before transformation, so it may not be too favorable to transform the age regressor in this way.

The following is a side-by-side comparison of left: Figure 12.1c (a scatter plot depicting the creatinine clearance (in $1000’s) versus Weight) and right: Figure 12.1l (a scatter plot depicting the natural log of creatinine clearance (in $1000’s) versus the natural log of weight):



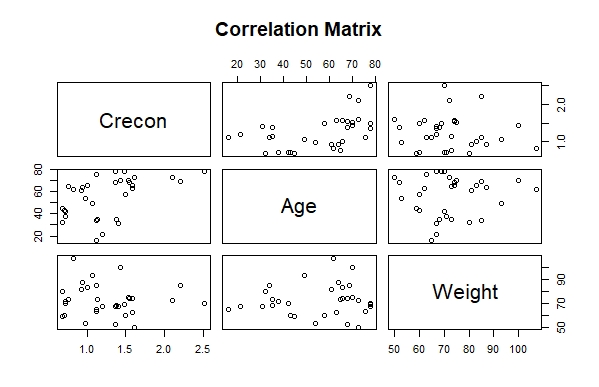
The alteration of the weight regressor does not appear to really alter the spread of the dataset in any particular way, besides changing the axes scales. There still does not appear to be a general relationship between weight and creatinine clearance after transformation with the natural log, so transforming the weight regressor may not be necessary.

**Correlation Matrices:**

Figure 12.1m: The following is a correlation matrix between the natural log of creatinine concentration, the natural log of (140-Age) and the natural log of weight.



The following is a side-by-side comparison of left: Figure 12.1d (a correlation matrix between creatinine concentration, age and weight) and right: Figure 12.1m (a correlation matrix between the natural log of creatinine concentration, the natural log of (140-Age) and the natural log of weight):

When comparing the two correlation matrix plots, there appears to be less correlation between the creatinine concentration and weight after the transformation than before (see the bottom left panel of Figure 12.1m), which suggests the transformation of the two variables may be justified to decrease the correlation between regressors. There might also be less correlation between the creatinine concentration and age after transformation (see the middle left panel of Figure 12.1m), which could suggest that the transformation of these two variables is justified to decrease the correlation between regressors.

**VIF Values:**

The following are the VIF values for the original regressors and the transformed regressors:

1. **VIF for creatinine concentration: 1.325**

**VIF for age: 1.320**

**VIF for weight: 1.020**

**VIF for ln(creatinine concentration): 1.356**

**VIF for ln(140-age): 1.348**

**VIF for ln(weight): 1.012**

The VIF values really haven’t changed very much, so the collinearity between regressor variables still seems to be fairly low, which is good for the model.

**Normality Plots:**

Figure 12.1n: The following is a normal probability plot of residuals for the following multiple regression model that relates the natural log of creatinine concentration, the natural log of (140-age) and the natural log of weight to the natural log of creatinine clearance (in $1000’s):

**Ln(creatinine clearance (in $1000’s))= [-0.6974\*ln(creatinine concentration)] + [0.7738\*ln(140-age)] + [0.7425\*ln(weight)] -$2.1036**



The following is a side-by-side comparison of left: Figure 12.1e (a normal probability plot of residuals for the following multiple regression model that relates creatinine concentration, age and weight to creatinine clearance (in $1000’s)) and right: Figure 12.1n (a normal probability plot of residuals for the following multiple regression model that relates the natural log of creatinine concentration, the natural log of (140-age) and the natural log of weight to the natural log of creatinine clearance (in $1000’s)):



The normal probability plot for the transformed model (12.1n) is much straighter than the original model, indicating that the transformed model fits the dataset better than the original model and that the dataset follows a normal distribution, independent distribution and equal variance better than the original dataset.

**Partial Regression Plots**

Figure 12.1o: The following is a plot of the residuals from the multiple linear regression from the transformed dataset versus the y-values the constructed model predicts.

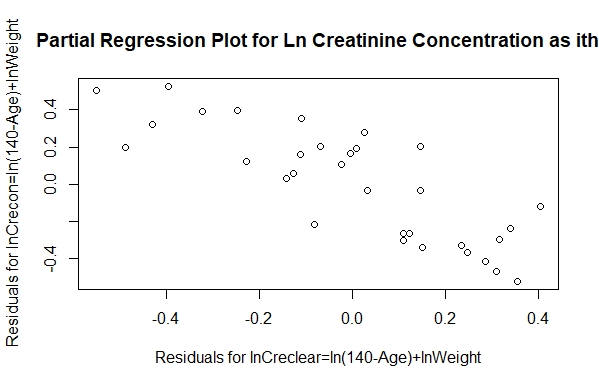


The following is a side-by-side comparison of left: Figure 12.1f (a plot of the residuals from the original multiple linear regression versus the y-values the constructed model predicts) and right: Figure 12.1o (a plot of the residuals from the multiple linear regression from the transformed dataset versus the y-values the constructed model predicts):



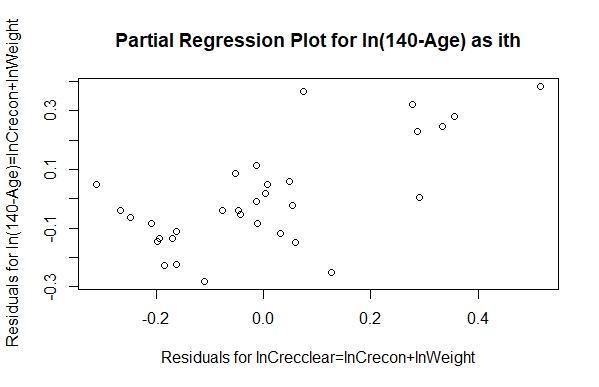
Though the data points are reconfigured, there does not appear to be a big difference in spread between the plots. However, the “middle” of the plot for the transformed data is randomly distributed and only the extreme high and low predicted responses have a slight pattern. This, perhaps, is better than the original plot, which is randomly distributed for higher predicted responses and not as much for lower predicted responses.

Figure 12.1p: The following is a partial regression plot of residuals for the natural log of creatinine concentration (lnCrecon) explained by the natural log of (140-Age) and the natural log of weight versus the residuals for the creatinine clearance (lnCrecclear) (in $1000’s) explained by the natural log of (140-Age) and the natural log of weight:



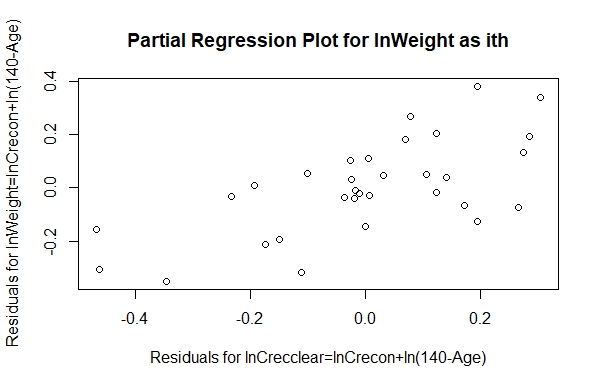
This partial regression plot treats the natural log of the creatinine concentration (lnCrecon) as the ith term that is removed from the model, and then the model is re-run and its residuals are plotted against the residuals of the model with the ith term as the response variable. The plot helps to determine if the removal of the ith term (here, the natural log of the creatinine concentration) greatly affects the outcome of the model. The fairly distinct pattern of the residuals plotted here suggests that the ith term explains much of the model’s variance after accounting for the tested regressors (here, the natural log of (140-Age) and the natural log of weight) and the ith term being removed does have a substantial effect on the model and should remain in the model.

Figure 12.1q: The following is a partial regression plot of residuals for the natural log of age (ln(140-Age)) explained by the natural log of creatinine concentration and the natural log of weight versus the residuals for the creatinine clearance (lnCrecclear) (in $1000’s) explained by the natural log of creatinine concentration and the natural log of weight:



This partial regression plot treats the natural log of age (ln(140-Age)) as the ith term that is removed from the model, and then the model is re-run and its residuals are plotted against the residuals of the model with the ith term as the response variable. The plot helps to determine if the removal of the ith term (here, the natural log of age (ln(140-Age))) greatly affects the outcome of the model. The fairly distinct pattern of the residuals plotted here suggests that the ith term explains a bit of the model’s variance after accounting for the tested regressors (here, the natural log of creatinine concentration and the natural log of weight) and the ith term being removed does have a substantial effect on the model and should remain in the model. It does not have as distinct a pattern as Figure 12.1h, but is still enough of an explanatory variable that I believe it warrants remaining in the model.

Figure 12.1r: The following is a partial regression plot of residuals for the natural log of weight explained by the natural log of age (ln(140-Age)) and the natural log of creatinine concentration versus the residuals for the creatinine clearance (lnCrecclear) (in $1000’s) explained by the natural log of creatinine concentration and the natural log of age (ln(140-Age)):



This partial regression plot treats the natural log of weight as the ith term that is removed from the model, and then the model is re-run and its residuals are plotted against the residuals of the model with the ith term as the response variable. The plot helps to determine if the removal of the ith term (here, the natural log of weight) greatly affects the outcome of the model. The fairly distinct pattern of the residuals plotted here suggests that the ith term explains much of the model’s variance after accounting for the tested regressors (here, the natural log of creatinine concentration and the natural log of (140-Age)) and the ith term being removed does have a substantial effect on the model and should remain in the model.

**Summary:**

The transformed model linearizes the relationship between creatinine concentration and creatinine clearance (in $1000’s) (Figure 12.1j), gives a slightly curvilinear to the relationship between age and creatinine clearance (in $1000’s) (Figure 12.1k) and does not appear to alter the relationship between weight and creatinine clearance (in $1000’s) (Figure 12.1l). However, according to the correlation matrix (Figure 12.1m) the regressor variables seem to be less-correlated, though the VIF values do not change very much at all. The most telling analysis that suggests the transformed model is better than the original model is the normal probability plot: it is much straighter for the transformed model, suggesting that the transformed model fits the data better than the original model (Figure 12.1n). All of the regressor variables for the transformed model appear to explain a decent amount of the variance of the model when analyzed with partial regression plots (Figures 12.1p, 12.1q and 12.1r), which is the same for the original model (Figures 12.1f, 12.1g and 12.1h). With all of this information, I would choose to use the transformed model over the original model for future analysis of the dataset.