Rachel Prokopius

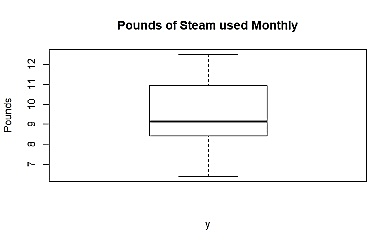
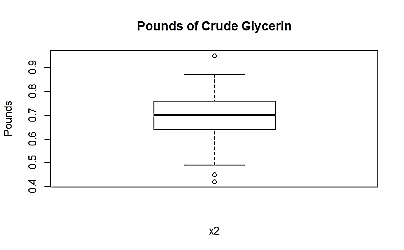
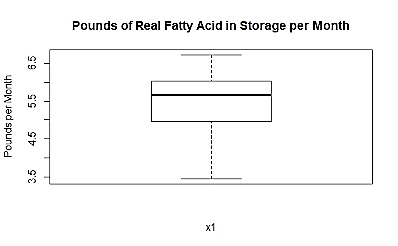
Regression Analysis

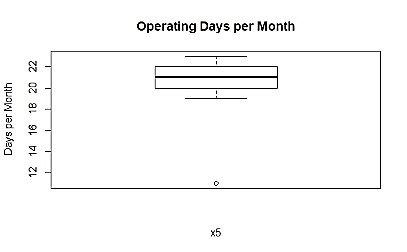
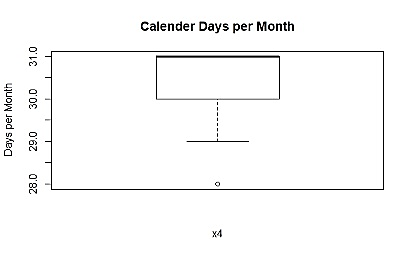
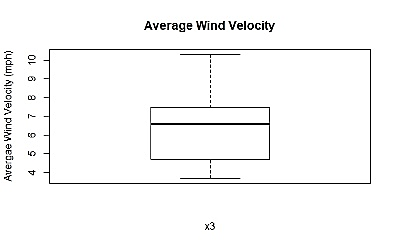
Final Exam

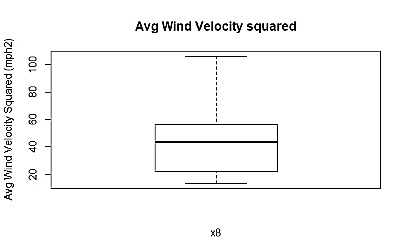
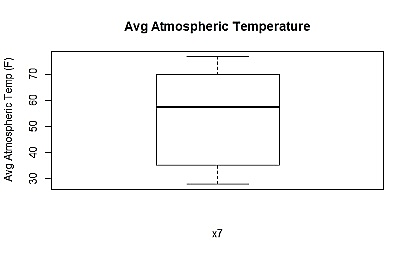
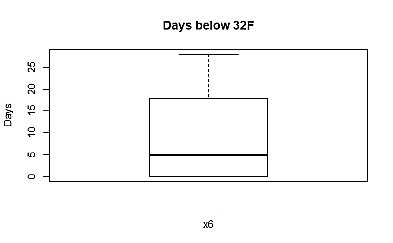
21 April 2020

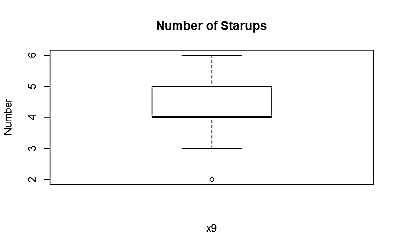
1. The following are box plots of the y variable and x variables for observations at a steam plant. Each boxplot is separate because each variable has different units.

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

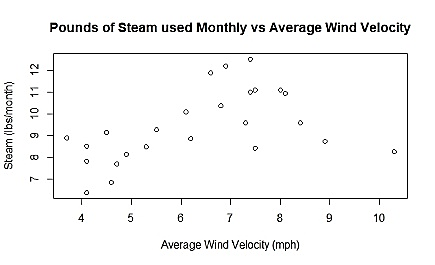
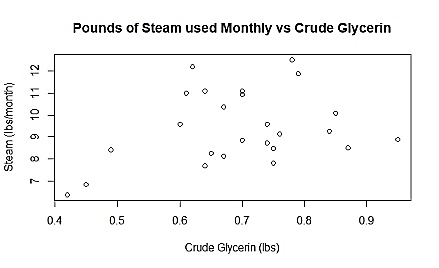
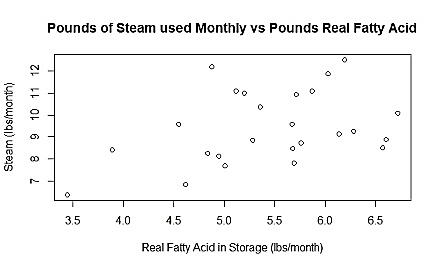


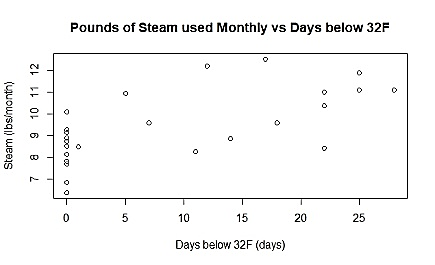
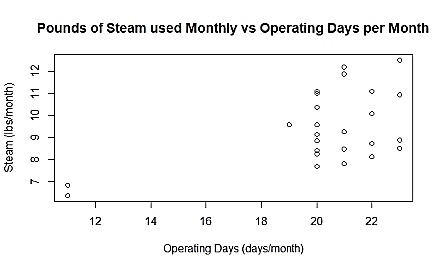
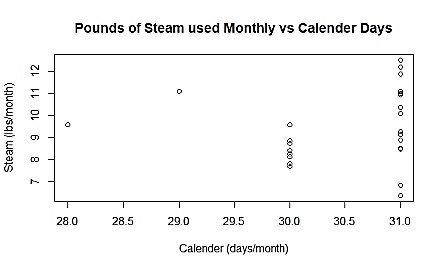


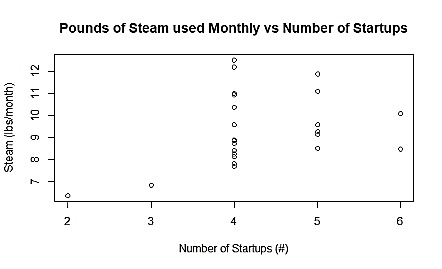
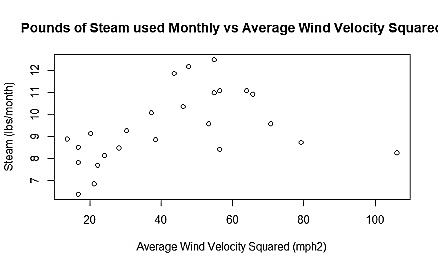
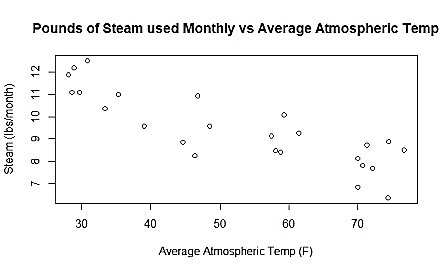


One of the requirements for a data set so that statistical analyses are valid is normal distribution, which means equal distribution of data points around the mean that have roughly 95% of the data points within two standard deviations of the mean and no outliers. Based on the box plots from this data 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

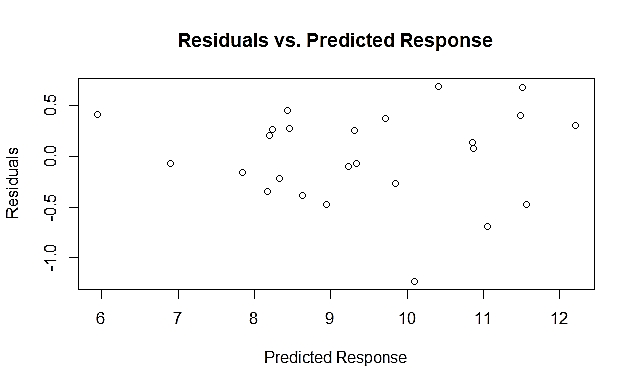
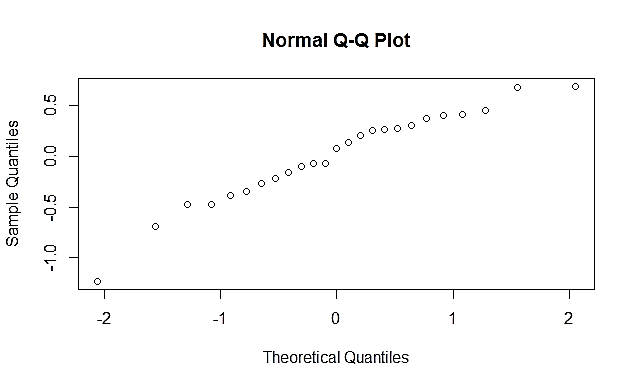
Residual standard error: 0.5673 on 15 degrees of freedom

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

**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.

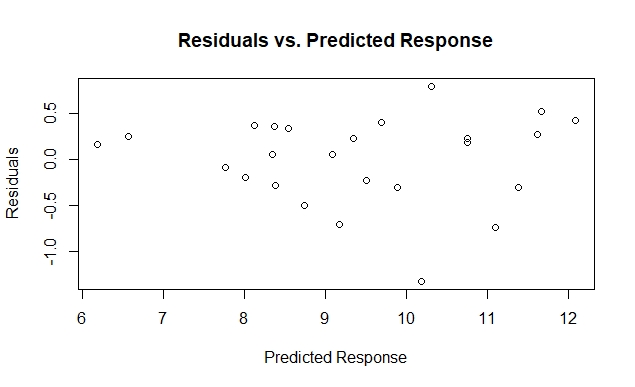
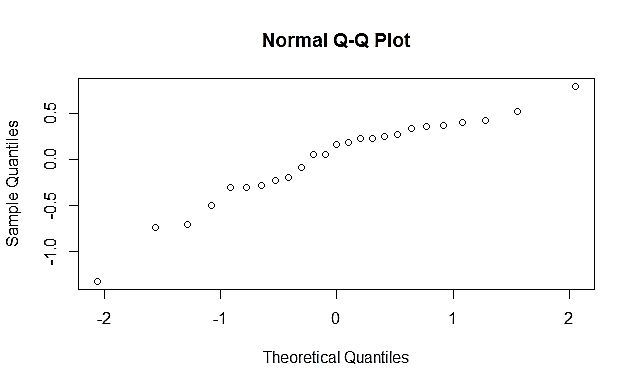
**x1 x2 x3 x4 x5 x6**

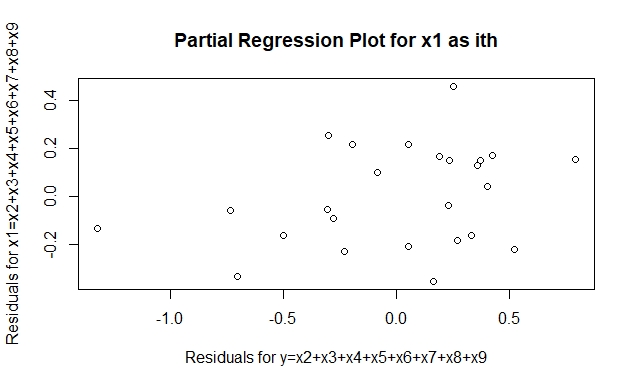
15.746595 20.137114 126.625618 1.836626 4.411920 4.695013

**x7 x8 x9**

6.067426 107.590891 2.385046

**Figure 3.2:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except pounds of real fatty acid per month (x1), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x1 versus the residuals for the x1 explained by the other regressor variables.





**Table 3.4**: The following is a table of the standardized residuals for the partial model without x1 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.4441977 1.6510694 0.8543183 0.1531863 -0.4543402 0.8720314

**7 8 9 10 11 12**

0.4103772 0.7668158 -0.3919954 0.1078524 -2.0623675 1.2142951

**13 14 15 16 17 18**

0.5545728 0.6053296 0.4001498 -0.6318953 0.8814951 -0.5757276

**19 20 21 22 23 24**

0.6058804 0.8247769 -0.1650782 -1.5598131 -2.5312539 -1.4006357

**25**

-0.6138719

**Table 3.5**: The following is a table of the studentized residuals for the partial model without x1 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.4327693 1.7551351 0.8467278 0.1484309 -0.4427785 0.8651502

**7 8 9 10 11 12**

0.3994538 0.7564971 -0.3813837 0.1044656 -2.3305308 1.2339697

**13 14 15 16 17 18**

0.5421991 0.5929368 0.3893967 -0.6196102 0.8750158 -0.5633113

**19 20 21 22 23 24**

0.5934890 0.8161244 -0.1599726 -1.6401235 -3.1652623 -1.4478207

**25**

-0.6015045

**Table 3.6**: The following is a table of the VIF values for the partial model without x1 relating steam (lbs/month) to all of the remaining regressor variables.

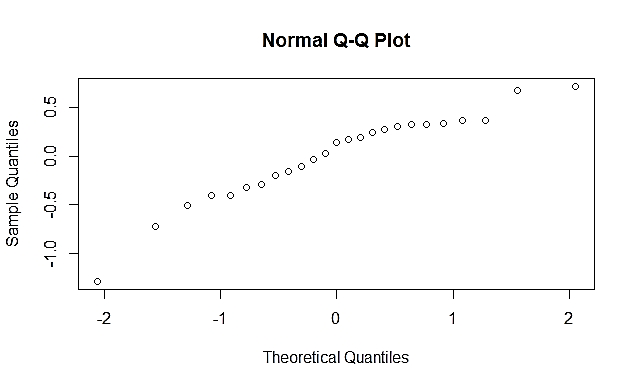
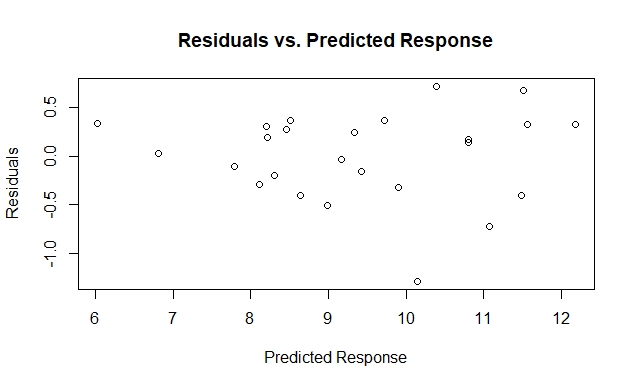
**x2 x3 x4 x5 x6 x7**

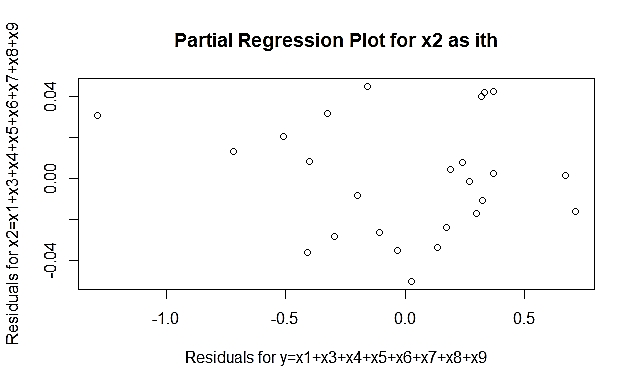
5.180428 117.669441 1.370684 4.024388 4.667488 6.064969

**x8 x9**

100.986695 2.232906

**Figure 3.3:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except pounds of crude glycerin (x2), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x2 versus the residuals for the x2 explained by the other regressor variables.



**Table 3.7**: The following is a table of the standardized residuals for the partial model without x2 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.34458364 1.56370197 0.69349876 0.60394268 -0.30824780 0.69710608

**7 8 9 10 11 12**

0.89015998 0.64587856 -0.61682186 -0.06514238 -1.85042861 1.70507243

**13 14 15 16 17 18**

0.66716441 0.66500086 0.29151391 -0.67997466 0.84178824 -0.43011399

**19 20 21 22 23 24**

0.07162692 0.89155551 -0.21381134 -1.27219570 -2.52041442 -1.42484640

**25**

-0.87446700

**Table 3.8**: The following is a table of the studentized residuals for the partial model without x2 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.33488660 1.64494915 0.68180255 0.59154650 -0.29934982 0.68545982

**7 8 9 10 11 12**

0.88406284 0.63368463 -0.60446533 -0.06308221 -2.02091565 1.82504421

**13 14 15 16 17 18**

0.65515642 0.65297132 0.28300970 -0.66810683 0.83372911 -0.41888477

**19 20 21 22 23 24**

0.06936359 0.88552121 -0.20731833 -1.29926334 -3.14275077 -1.47644878

**25**

-0.86768771

**Table 3.9**: The following is a table of the VIF for the partial model without x2 relating steam (lbs/month) to all of the remaining regressor variables.

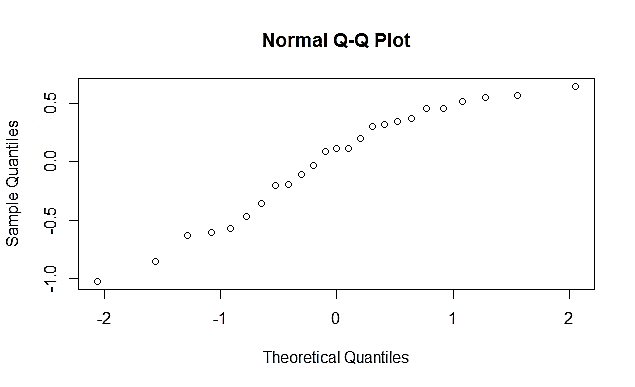
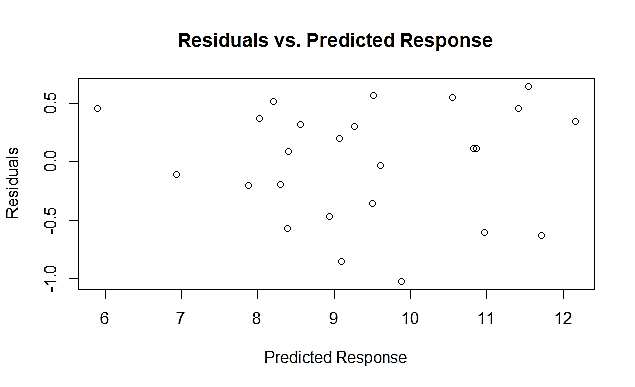
**x1 x3 x4 x5 x6 x7**

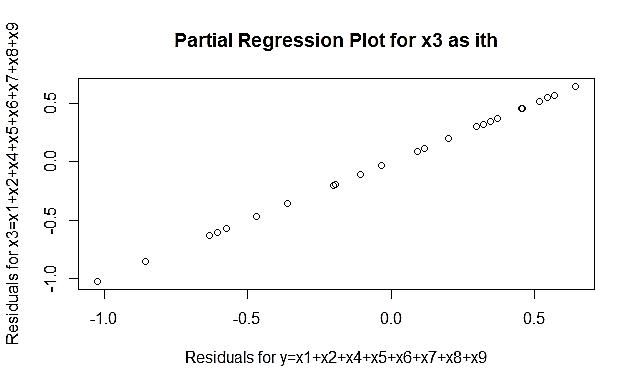
4.050933 103.260596 1.650794 2.704032 4.665612 5.999323

**x8 x9**

90.287806 2.384648

**Figure 3.4:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except average wind velocity in miles per hour (x3), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x3 versus the residuals for the x3 explained by the other regressor variables.



**Table 3.10**: The following is a table of the standardized residuals for the partial model without x3 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.21978832 1.11306319 0.69462526 1.04461250 0.37127291 1.15551286

**7 8 9 10 11 12**

1.26741150 0.18096145 -1.10497637 -0.67945367 -2.17941076 1.53195823

**13 14 15 16 17 18**

0.94586367 0.77027276 0.24323331 -0.06193097 1.17340759 -0.39424347

**19 20 21 22 23 24**

-0.32270450 0.78449261 -0.38018695 -1.12298900 -1.87259191 -1.12532471

**25**

-1.29792337

**Table 3.11**: The following is a table of the studentized residuals for the partial model without x3 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.2131311 1.1220345 0.6829445 1.0478032 0.3610420 1.1686447

**7 8 9 10 11 12**

1.2938309 0.1753948 -1.1132064 -0.6675797 -2.5165479 1.6057460

**13 14 15 16 17 18**

0.9425598 0.7600386 0.2359463 -0.0599716 1.1884327 -0.3835923

**19 20 21 22 23 24**

-0.3134791 0.7746255 -0.3697885 -1.1328924 -2.0518635 -1.1354510

**25**

-1.3285962

**Table 3.12**: The following is a table of the VIF values for the partial model without x3 relating steam (lbs/month) to all of the remaining regressor variables.

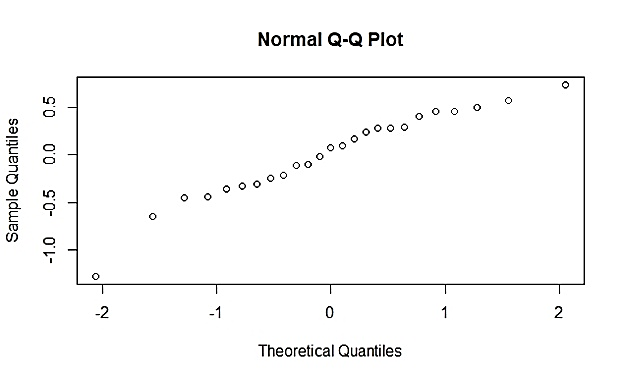
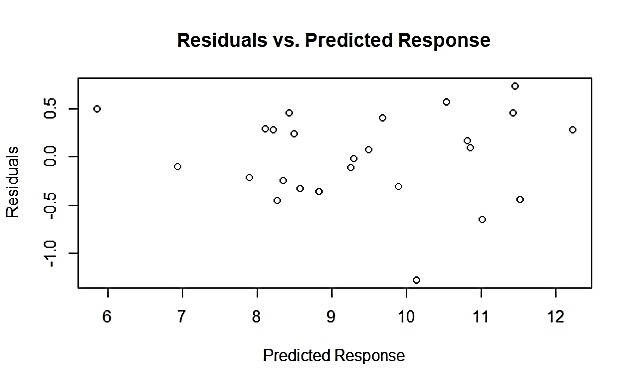
**x1 x2 x4 x5 x6 x7 x8**

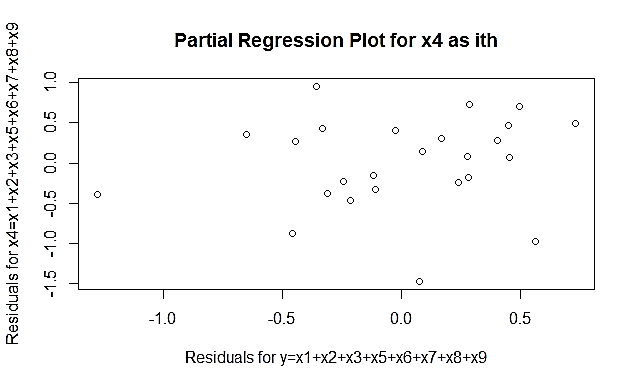
14.632845 16.421404 1.824200 3.856707 4.694497 5.056370 1.867641

**x9**

2.256168

**Figure 3.5:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except calendar days per month (x4), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x4 versus the residuals for the x4 explained by the other regressor variables.



**Table 3.13**: The following is a table of the standardized residuals for the partial model without x4 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.33290107 1.14265177 0.59932418 0.82460700 -0.04766458 0.60577340

**7 8 9 10 11 12**

1.36514236 0.60142111 -0.91375395 -0.24626395 -1.43178824 1.80107286

**13 14 15 16 17 18**

0.97282224 0.16106991 0.20167227 -0.66457396 0.91734269 -0.52126578

**19 20 21 22 23 24**

-0.33453219 1.20181466 -0.42317941 -0.81231332 -2.53402695 -1.27840162

**25**

-0.98515588

**Table 3.14**: The following is a table of the studentized residuals for the partial model without x4 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.32345220 1.15447460 0.58691850 0.81594885 -0.04615431 0.59338168

**7 8 9 10 11 12**

1.40622356 0.58901944 -0.90876778 -0.23889723 -1.48469617 1.95306785

**13 14 15 16 17 18**

0.97108823 0.15608186 0.19551699 -0.65254024 0.91253467 -0.50905441

**19 20 21 22 23 24**

-0.32504817 1.22002141 -0.41205416 -0.80325686 -3.17105230 -1.30632039

**25**

-0.98418962

**Table 3.15**: The following is a table of the VIF values for the partial model without x4 relating steam (lbs/month) to all of the remaining regressor variables.

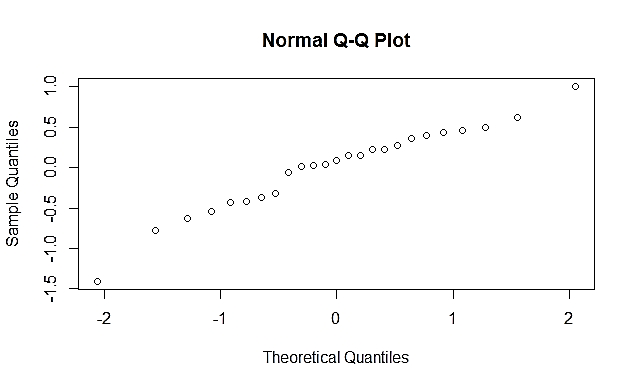
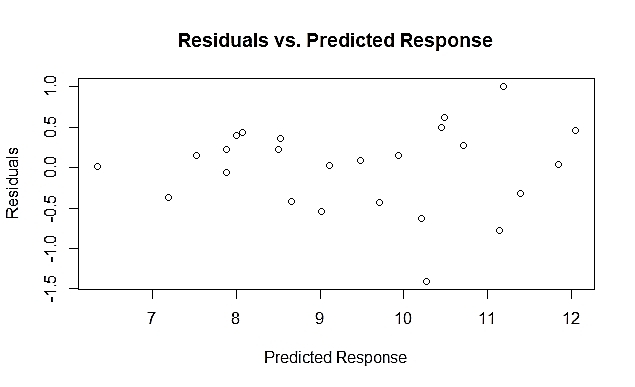
**x1 x2 x3 x5 x6 x7**

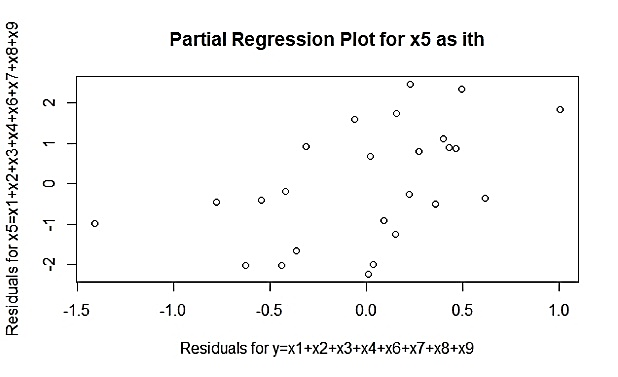
11.751771 18.099620 125.768897 4.392236 4.634851 5.862202

**x8 x9**

107.419983 2.080472

**Figure 3.6:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except operating days per month (x5), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x5 versus the residuals for the x5 explained by the other regressor variables.



**Table 3.16**: The following is a table of the standardized residuals for the partial model without x5 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.47361134 1.18993889 0.85685321 1.06250175 -0.76465028 0.49306366

**7 8 9 10 11 12**

0.02267385 0.80864384 -0.10868656 0.04194676 -1.70069400 2.09083363

**13 14 15 16 17 18**

0.06206378 0.21576074 0.89235610 -1.12637653 0.29228189 0.39875392

**19 20 21 22 23 24**

-0.94037856 0.81973694 0.26525590 -1.20661706 -2.44826380 -1.34668784

**25**

-0.60626942

**Table 3.17**: The following is a table of the studentized residuals for the partial model without x5 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.46182082 1.20678884 0.84936084 1.06709696 -0.75427958 0.48107568

**7 8 9 10 11 12**

0.02195422 0.79947330 -0.10527418 0.04061701 -1.81932195 2.37467866

**13 14 15 16 17 18**

0.06010023 0.20921402 0.88635800 -1.13660364 0.28375928 0.38802469

**19 20 21 22 23 24**

-0.93677305 0.81091801 0.25739951 -1.22538351 -2.99759874 -1.38476469

**25**

-0.59387897

**Table 3.18**: The following is a table of the VIF residuals for the partial model without x5 relating steam (lbs/month) to all of the remaining regressor variables.

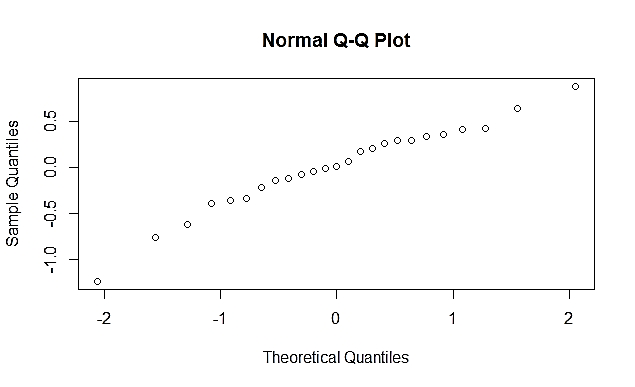
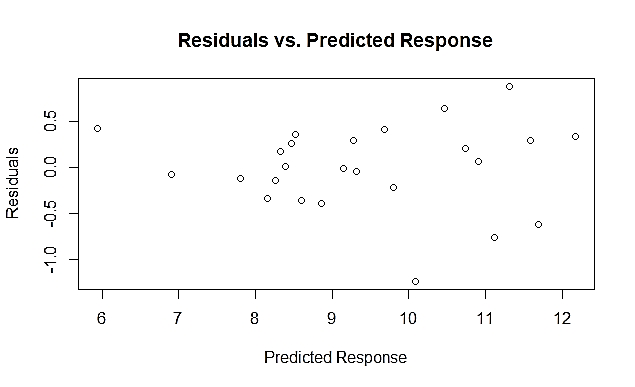
**x1 x2 x3 x4 x6 x7**

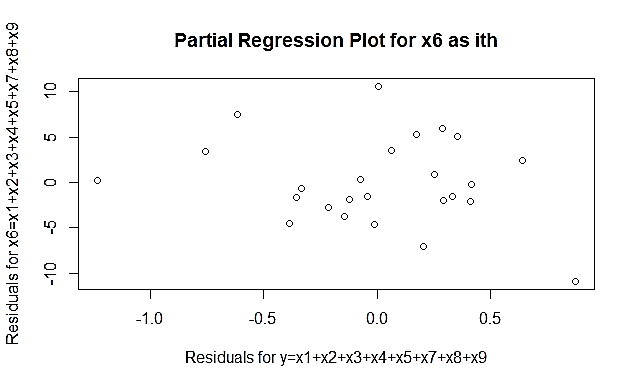
14.363452 12.341883 110.690568 1.828431 4.530146 5.840032

**x8 x9**

97.334598 2.354898

**Figure 3.7:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except days below 32F (x6), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x6 versus the residuals for the x6 explained by the other regressor variables.



**Table 3.19**: The following is a table of the standardized residuals for the partial model without x6 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.12579090 1.38574126 0.70218304 0.01215781 -0.08966443 0.63946578

**7 8 9 10 11 12**

1.22778901 0.35485753 -0.71384817 -0.02629473 -1.61570740 1.83234109

**13 14 15 16 17 18**

0.59561943 0.79049787 0.42062397 -0.45870909 0.92144628 -0.30746771

**19 20 21 22 23 24**

-0.22933923 0.88660864 -0.24655628 -0.94537103 -2.45962806 -1.47255796

**25**

-1.26552346

**Table 3.20**: The following is a table of the studentized residuals for the partial model without x6 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.12185679 1.43031190 0.69061010 0.01177180 -0.08683903 0.62722706

**7** 8 9 10 11 12

1.24909999 0.34494943 -0.70245721 -0.02546032 -1.71011900 1.99588273

**13 14 15 16 17 18**

0.58320791 0.78079527 0.40953799 -0.44709273 0.91684497 -0.29858774

**19 20 21 22 23 24**

-0.22242264 0.88035321 -0.23918189 -0.94203979 -3.01994138 -1.53349456

**25**

-1.29168898

**Table 3.21**: The following is a table of the VIF values for the partial model without x6 relating steam (lbs/month) to all of the remaining regressor variables.

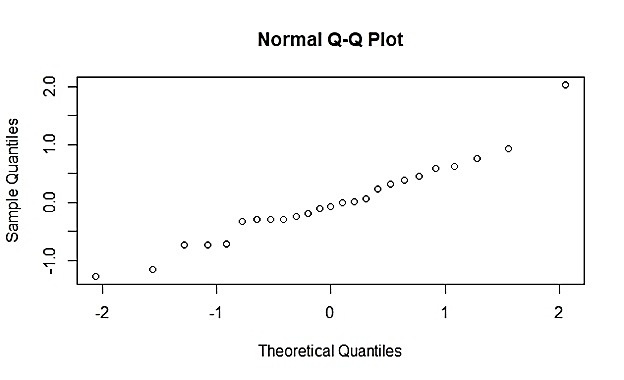
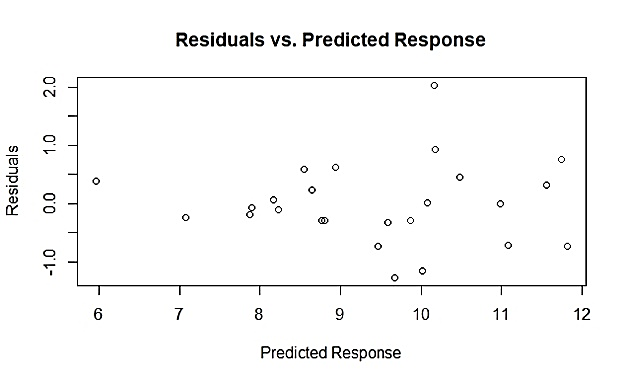
**x1 x2 x3 x4 x5 x7**

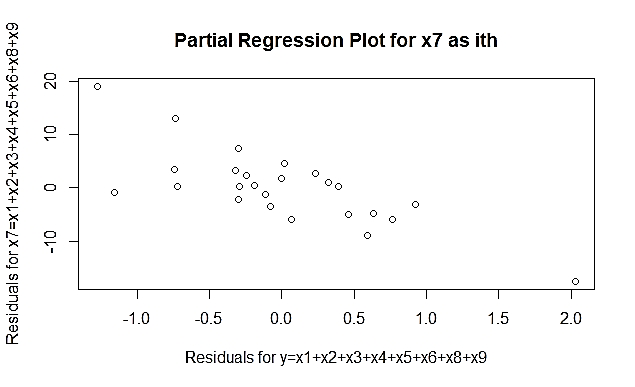
15.654279 20.011010 126.611716 1.813091 4.256994 2.707750

**x8 x9**

107.589948 2.384853

**Figure 3.8:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except average atmospheric temperature (x7), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x7 versus the residuals for the x7 explained by the other regressor variables.



**Table 3.22**: The following is a table of the standardized residuals for the partial model without x7 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5**

-0.004280291 1.308122679 1.027464977 -1.854430998 -0.435022623

**6 7 8 9 10**

-1.064714863 0.757797127 -0.404796745 -0.102563672 0.775142622

**11 12 13 14 15**

0.192660293 2.690707046 0.453713494 1.098732540 0.652335753

**16 17 18 19 20**

-0.402602580 0.024830698 -0.156057688 -0.495848853 0.398504864

**21 22 23 24 25**

-0.244553018 -0.486886978 -1.495983602 -0.917094485 -1.057002079

**Table 3.23**: The following is a table of the studentized residuals for the partial model without x7 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5**

-0.004144376 1.340281420 1.029377425 -2.026481852 -0.423722146

**6 7 8 9 10**

-1.069488942 0.747266530 -0.393965308 -0.099339509 0.765030619

**11 12 13 14 15**

0.186759282 3.520931882 0.442159823 1.106401147 0.640192143

**16 17 18 19 20**

-0.391807939 0.024042683 -0.151217336 -0.483835440 0.387779906

**21 22 23 24 25**

-0.237231229 -0.474957952 -1.561820663 -0.912274062 -1.061157694

**Table 3.24**: The following is a table of the VIF values for the partial model without x7 relating steam (lbs/month) to all of the remaining regressor variables.

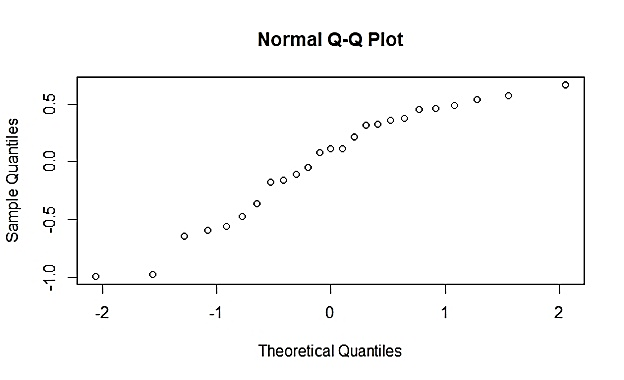
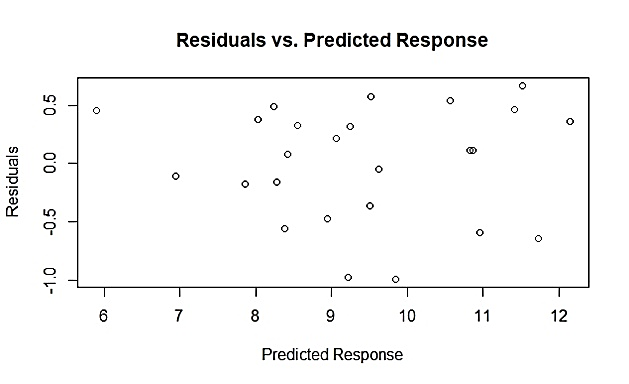
**x1 x2 x3 x4 x5 x6**

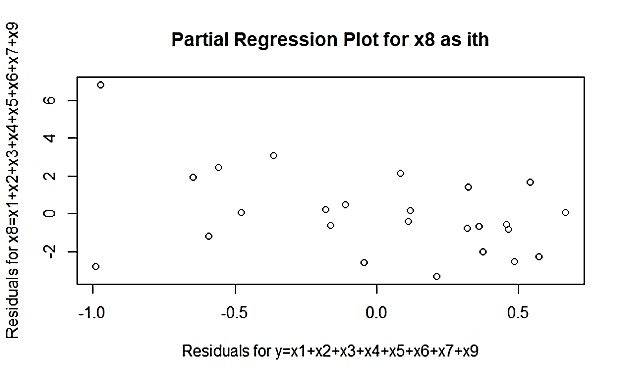
15.740218 19.911086 105.525135 1.774504 4.246571 2.095274

**x8 x9**

93.393373 2.377952

**Figure 3.9:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except average wind velocity squared (x8), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x8 versus the residuals for the x8 explained by the other regressor variables.



**Table 3.25**: The following is a table of the standardized residuals for the partial model without x8 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.21729176 1.08816727 0.71558471 1.04173314 0.39239001 1.09680110

**7 8 9 10 11 12**

1.26167965 0.16465417 -1.07432000 -0.68097205 -2.22223837 1.56615747

**13 14 15 16 17 18**

0.94789723 0.81749603 0.22764192 -0.08670867 1.16616172 -0.32307847

**19 20 21 22 23 24**

-0.32605162 0.77911432 -0.33327258 -1.12071787 -1.77808752 -1.08372175

**25**

-1.30830038

**Table 3.26**: The following is a table of the studentized residuals for the partial model without x8 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.21070296 1.09490739 0.70422246 1.04470461 0.38177135 1.10429795

**7 8 9 10 11 12**

1.28733150 0.15956096 -1.07988362 -0.66911603 -2.58777304 1.64799950

**13 14 15 16 17 18**

0.94470665 0.80860444 0.22077115 -0.08397504 1.18041000 -0.31384478

**19 20 21 22 23 24**

-0.31675218 0.76910468 -0.32381570 -1.13040583 -1.92195507 -1.09007920

**25**

-1.34048530

**Table 3.27**: The following is a table of the VIF values for the partial model without x8 relating steam (lbs/month) to all of the remaining regressor variables.

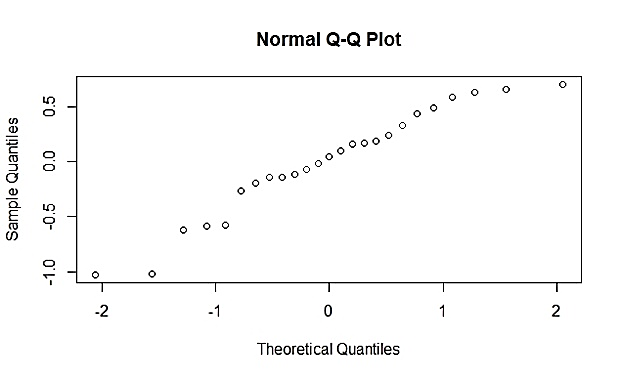
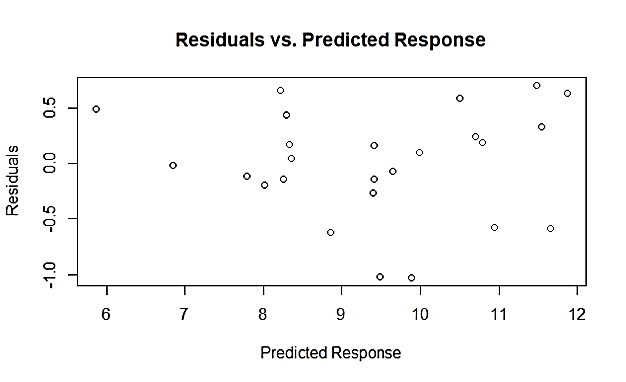
**x1 x2 x3 x4 x5 x6 x7**

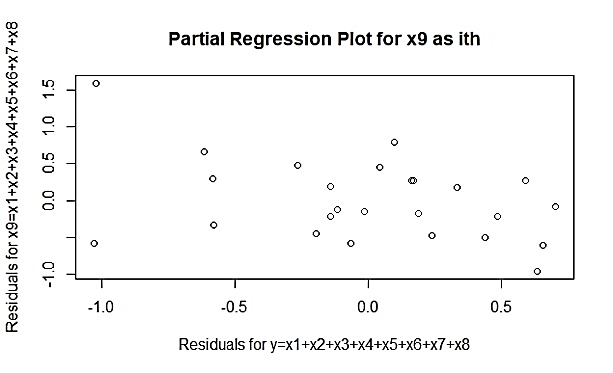
14.780030 16.898604 2.198060 1.833708 3.991346 4.694972 5.266779

**x9**

2.224572

**Figure 3.10:** The following is a normal probability plot relating steam (lbs/month) to the regressor variables in the full model except number of starups (x9), a plot of the residuals from this model versus the y-values the constructed model predicts, and a partial regression plot of residuals for steam (lbs/month) explained by the all regressors except x9 versus the residuals for the x9 explained by the other regressor variables.



**Table 3.28**: The following is a table of the standardized residuals for the partial model without x9 relating steam (lbs/month) to all of the remaining regressor variables.

**1 2 3 4 5 6**

0.35337483 1.20330596 1.16282131 0.12237207 -0.27203442 1.01269469

**7 8 9 10 11 12**

1.33820146 0.34627126 -0.37873464 -0.50954470 -2.24821617 1.65498827

**13 14 15 16 17 18**

0.67871459 0.41607185 0.48517966 -0.12808026 0.19531462 -0.28329663

**19 20 21 22 23 24**

-0.04110146 1.53347520 -0.21573584 -1.85710423 -1.86770605 -1.05948298

**25**

-1.19638859

**Table 3.29**: The following is a table of the studentized residuals for the partial model without x9 relating steam (lbs/month) to all of the remaining regressor variables.

1 2 3 4 5 6

0.34349676 1.22168583 1.17671617 0.11854173 -0.26400744 1.01355829

7 8 9 10 11 12

1.37493465 0.33653909 -0.36836314 -0.49741690 -2.63187596 1.76016016

13 14 15 16 17 18

0.66683201 0.40505710 0.47326755 -0.12407680 0.18933841 -0.27499134

19 20 21 22 23 24

-0.03979841 1.60760985 -0.20918981 -2.03020507 -2.04501474 -1.06383604

25

-1.21397061

**Table 3.30**: The following is a table of the VIF values for the partial model without x9 relating steam (lbs/month) to all of the remaining regressor variables.

**x1 x2 x3 x4 x5 x6**

14.742133 20.133751 119.783271 1.602086 4.356152 4.694632

**x7 x8**

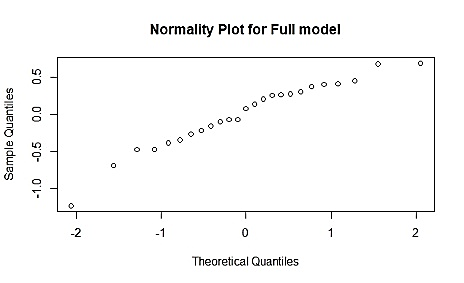
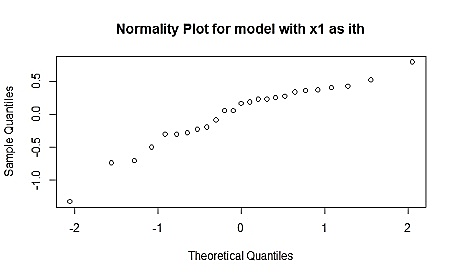
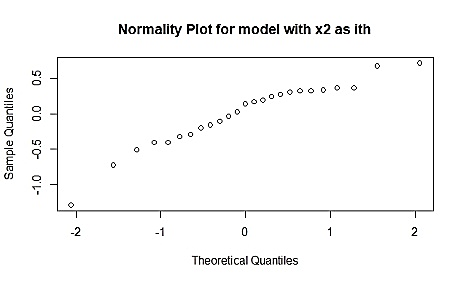
6.049379 100.351820

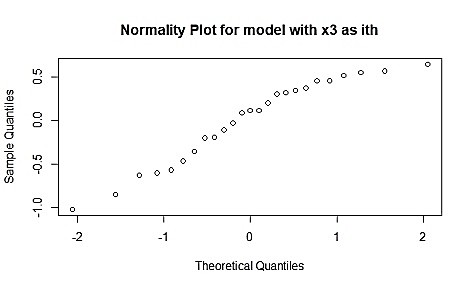
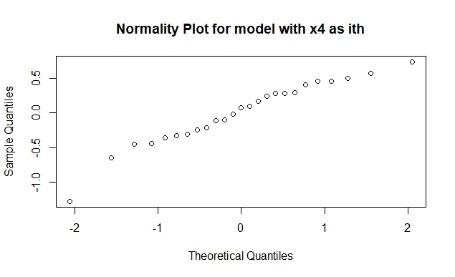
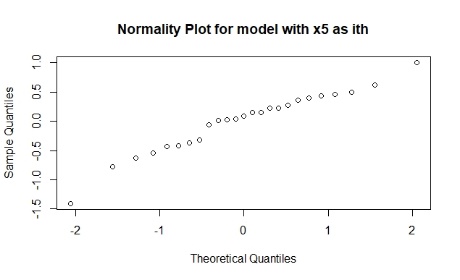
**Table 3.31**: The following is a table of the PRESS statistics for all of the models represented above.

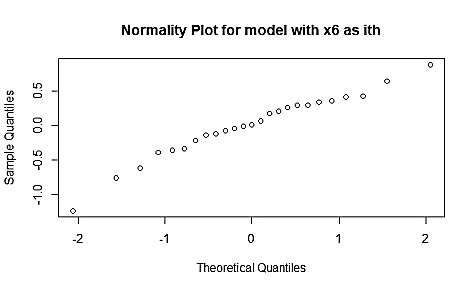
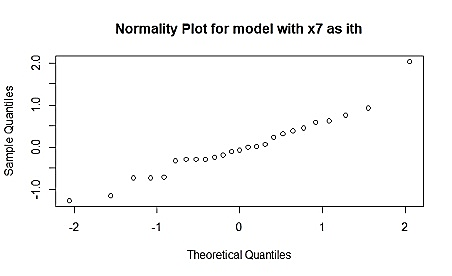
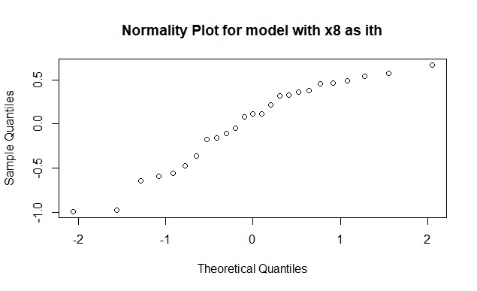
|  |  |
| --- | --- |
| **Model** | **PRESS Statistic** |
| Full model | 18.785 |
| x1 ith | 18.605 |
| x2 ith | 17.347 |
| x3 ith | 16.504 |
| x4 ith | 14.913 |
| x5 ith | 22.269 |
| x6 ith | 15.784 |
| x7 ith | 25.748 |
| x8 ith | 16.202 |
| x9 ith | 20.962 |

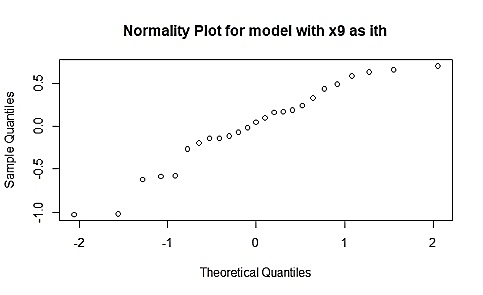
1. For analyzing model adequacy, figures have been provided with side-by-side comparison plots.

**Figure 4.1:** The following is a side-by-side comparison of all of the normality plots created in Question 3:

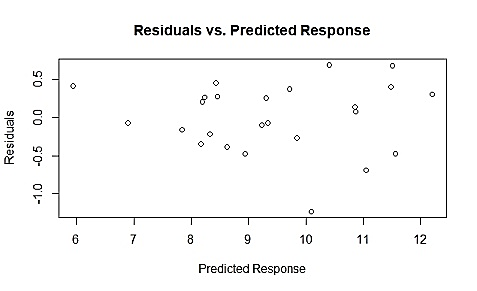
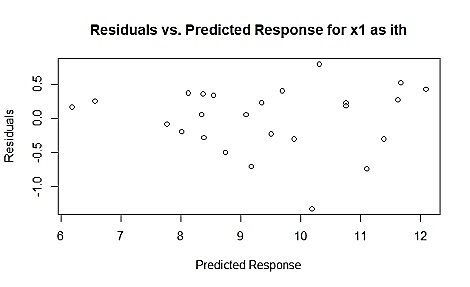
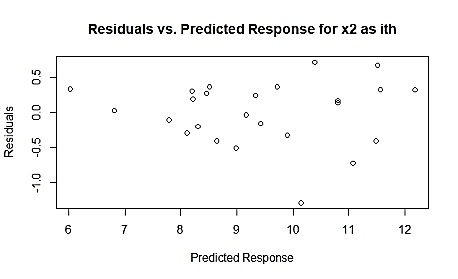
  

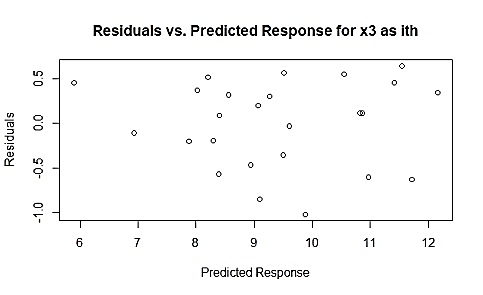
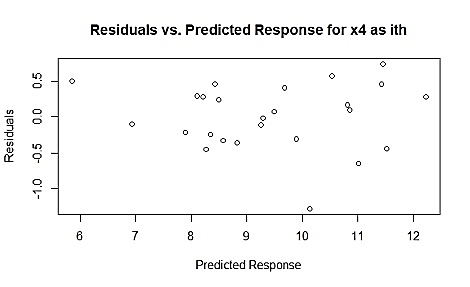
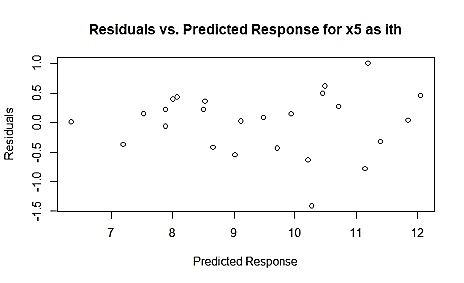
  

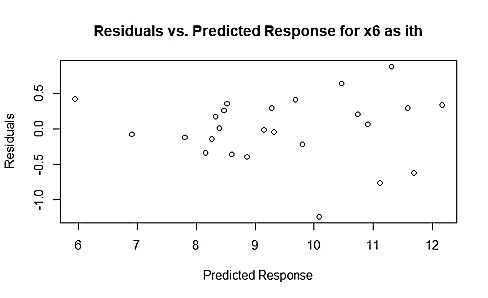
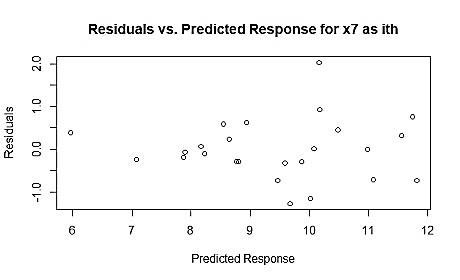
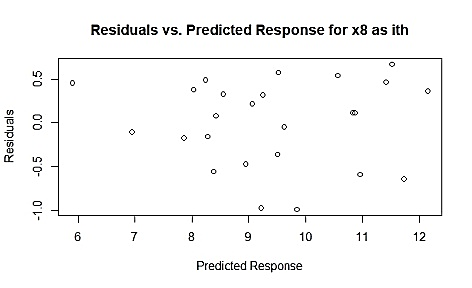
  

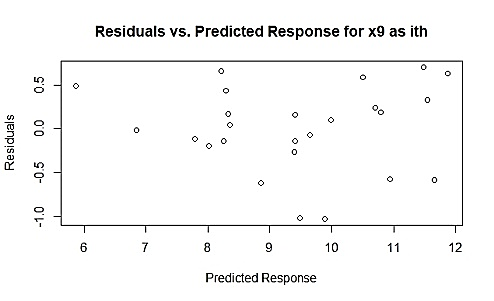


**Figure 4.2:** The following is a side-by-side comparison of all of the residual plots created in Question 3:

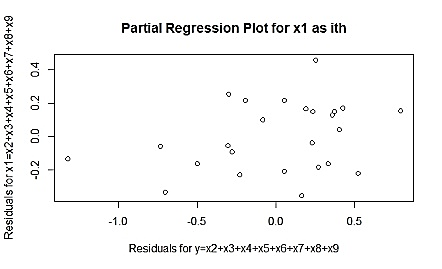
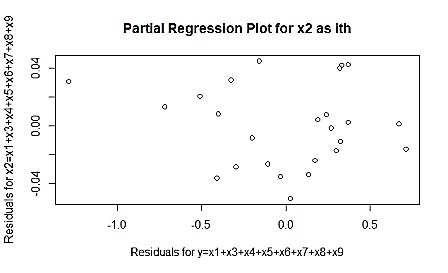
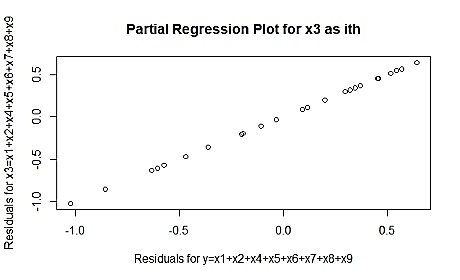
  

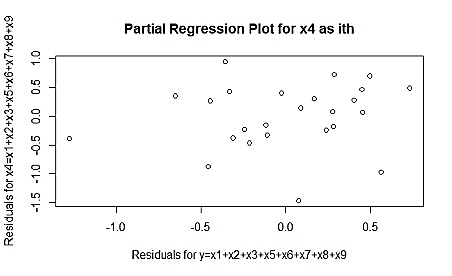
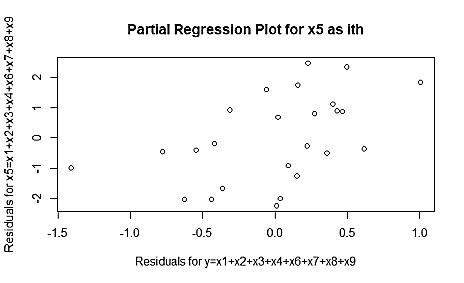
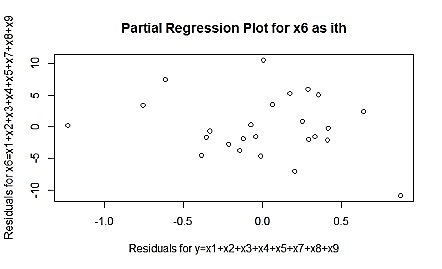
  

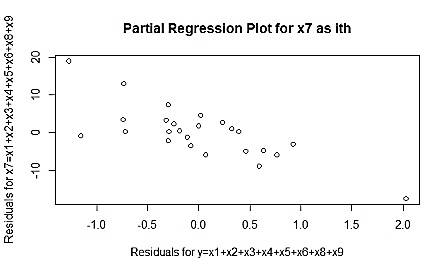
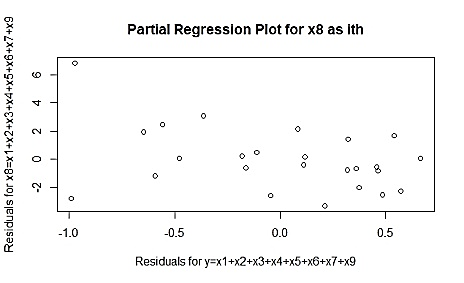
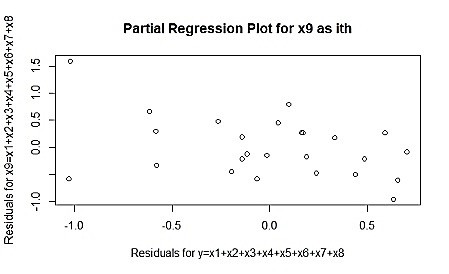
  



**Figure 4.3:** The following is a side-by-side comparison of all of the partial plots created in Question 3:

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. When comparing the normality plots for the full model to the plots with each regressor removed individually, it appears that the full model normality plot (upper left corner) deviates more from a straight line when average wind velocity in miles per hour (x3), average wind velocity squared (x8) and number of startups (x9) are removed (Figure 4.1). This suggests that the model has a more-normal distribution when these regressors are present if they are taken out individually.

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 original model has a fairly randomly-distributed residual plot, suggesting little if any transformation to the model must be done (Figure 4.2, upper left corner). Removing regressors for partial regressions also does not really alter the residual plot, suggesting removal of regressors will not yield a model with a more-constant variance.

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 partial plots do not drastically reduce the standardized and studentized residuals either, though the removal of certain regressors individually are accompanied by a spike in certain residuals over 3 when the absolute value is taken and suggest the regressor should remain in the model to control for outliers (Tables 3.5, 3.8, 3.14, 3.20). However, these spikes are seen in the studentized residuals, which are much more conservative than the standardized residuals and are therefore more likely to consider a residual an outlier than the corresponding standardized residuals.

When considering PRESS residuals, the model that has the lowest PRESS residual when one regressor is removed is when the average wind velocity in miles per hour (x3) regressor is removed. However, the partial plot for removing this regressor shows that it is the regressor that, when removed individually, explains a great amount of the variance of the model after the other regressor variables are accounted for (Figures 3.4 and 4.3). Therefore, it should remain in the model. Other regressors, when removed individually, increase the PRESS statistic compared to the original model, such as calendar days per month (x4) and average atmospheric temperature in degrees F (x7) (Table 3.31). The higher PRESS statistic means that, when the individual regressor is removed, influential points in the data set affect the model more than when the regressor was present, and suggests that the model would be benefited by the regressor in question remaining in the model.

Finally, removal of regressor variables alters the VIF values differently depending on the regressor variable that is removed. All regressors have moderate to low collinearity in the full model except pounds of real fatty acid in storage per month (x1), pounds of crude glycerin made per month (x2), average wind velocity (x4) and average wind velocity squared (x8) (Table 3.3). These collinearites make sense when considering the regressors; if more real fatty acid is stored per month, then more glycerin can be made because glycerin is comprised primarily of fatty acids. When x1 is removed from the model, x2’s VIF value drops from 20.137 to 5.180 (Tables 3.3 and 3.6). Similarly, when x2 is removed from the model, x1’s VIF value drops from 15.747 to 4.051 (Tables 3.3 and 3.9). The collinearity between average wind velocity and average wind velocity squared is obvious; average wind velocity squared is simply average wind velocity multiplied by itself. When x8 is removed, x3’s VIF value drops from 126.626 to 2.198 (Tables 3.3 and 3.27). Similarly, when x3 is removed, x8’s VIF value drops from 107.591 to 1.868 (Tables 3.3 and 3.12). However, multicollinearity is not a good enough reason to remove a regressor from a model.

Based on these analyses, there is no definitive indication that any particular regressor should be removed from the model for the model to perform better. It could be argued that certain regressors must stay in the model for it to perform well, such as average wind velocity (x3) based on its partial regression plot, but there is no evidence that any regressor should absolutely be removed from a model (Figures 3.4 and 4.3).

Based on the original scatter plots, the regressors that show a pattern with steam (lbs/month) that is nonlinear are average wind velocity (x3) and average wind velocity squared (x8). Both plots show a slight parabolic curve, suggesting that squaring these regressors may be warranted. There is a slight parabolic curve to the pounds of crude glycerin made per month (x2) as well, so this may also warrant squaring as well. Finally, because of the large number of data points at zero days for the number of days below 32F (x6), there is a slight square-root-like curve to the scatterplot and could perhaps benefit from square-rooting that regressor. The transformed model that is a better fit for the data set 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) = 0.022(x1) +1.724[lbs-1\*month](x2^2) +0.070[lbs\*hr2\*month-1\*miles-2](x3^2) + 0.206[lbs\*days-1](x4) +0.147[lbs\*days-1](x5) -0.246[lbs\*days-0.5\*month-1](sqrt[x6]) -0.091[lbs\*ºF-1\*month-1](x7) -0.00063[lbs\*hr2\*miles-2\*month-1](x8^2) -0.291[lbs\*number of startups-1\*month-1](x9) + 4.232 (lbs/month)**

The summary statistics for the new model are as follows:

Residuals:

Min 1Q Median 3Q Max

-1.07899 -0.20368 0.05293 0.26547 0.60527

Coefficients:

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

(Intercept) 4.2319065 5.5446609 0.763 0.4572

x1 0.0221917 0.4810566 0.046 0.9638

sqx2 1.7240965 2.2457111 0.768 0.4546

sqx3 0.0698773 0.0269608 2.592 0.0204 \*

x4 0.2060570 0.1731591 1.190 0.2525

x5 0.1473051 0.0631305 2.333 0.0340 \*

sqrtx6 -0.2455904 0.1387612 -1.770 0.0971 .

x7 -0.0911187 0.0170326 -5.350 8.1e-05 \*\*\*

sqx8 -0.0006255 0.0002186 -2.861 0.0119 \*

x9 -0.2908944 0.1795418 -1.620 0.1260

---

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

Residual standard error: 0.4971 on 15 degrees of freedom

Multiple R-squared: 0.9418, Adjusted R-squared: 0.9069

F-statistic: 26.98 on 9 and 15 DF, p-value: 1.106e-07

Compared to the statistics for the original model, the adjusted R2 value for the transformed model has increased, and more regressor coefficients have become significant for the model (x3 and x8 have become significant to an alpha level of 0.05, and x6 has become significant to an alpha value of 0.1).

1. Based on the results from the stepwise regression in R, the model of best-fit 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) = 0.488(x1) +0.108[lbs\*days-1](x5) -0.076[lbs\*ºF-1\*month-1](x7) + 8.556 (lbs/month)**

The output and summary statistics for this model are as follows:

Stepwise Selection Summary

------------------------------------------------------------------------------------

Added/ Adj.

Step Variable Removed R-Square R-Square C(p) AIC RMSE

------------------------------------------------------------------------------------

1 x7 addition 0.714 0.701 52.7850 69.0538 0.8903

2 x1 addition 0.860 0.847 17.1190 53.1954 0.6369

3 x5 addition 0.880 0.862 14.0670 51.4289 0.6046

Call:

lm(formula = y ~ x1 + x5 + x7)

Residuals:

Min 1Q Median 3Q Max

-1.33205 -0.30490 0.05466 0.35996 1.16505

Coefficients:

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

(Intercept) 8.55609 1.03675 8.253 4.98e-08 \*\*\*

x1 0.48842 0.21162 2.308 0.0313 \*

x5 0.10827 0.05859 1.848 0.0788 .

x7 -0.07572 0.00746 -10.150 1.49e-09 \*\*\*

---

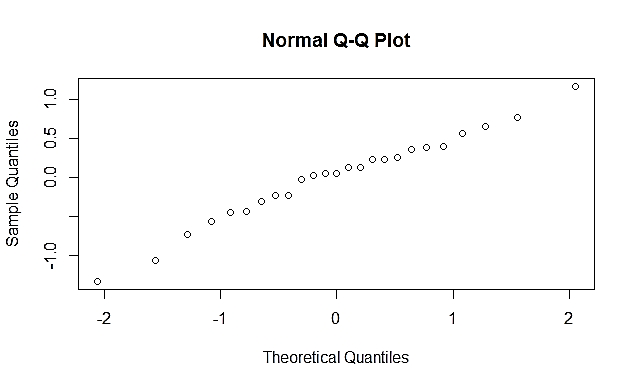
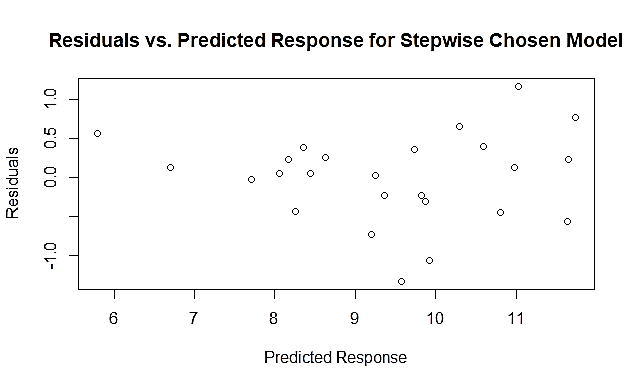
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6046 on 21 degrees of freedom

Multiple R-squared: 0.8795, Adjusted R-squared: 0.8623

F-statistic: 51.1 on 3 and 21 DF, p-value: 7.988e-10

**Figure 5.1:** The following is a normal probability plot relating steam (lbs/month) to the pounds of real fatty acid per month (x1), operating days per month (x5) and average atmospheric temperature (F) in the stepwise model and a plot of the residuals from the full model versus the y-values the constructed model predicts.

Though the adjusted R2 value is lower than the original model and the transformed model, it is still quite high and only uses three of the nine regressors of the model, which suggests it is a much more efficient model and uses the regressors that best explain the y variable. The normality plot is more linear than the original normality plot, and the residual plot is still fairly randomly distributed, suggesting a normally-distributed data set with independence and constant variance (Figure 5.1). To summarize, the computer is much better at predicting regressors to use in a model than I am.