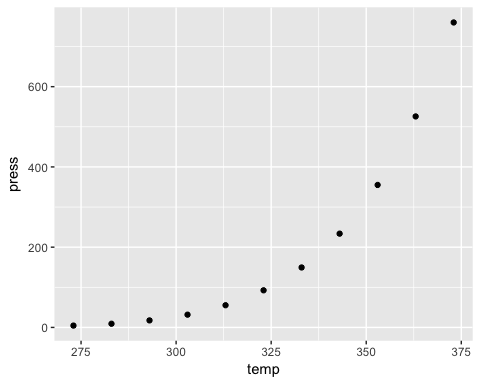
HW2

Joaquin Rodriguez

11/1/2017

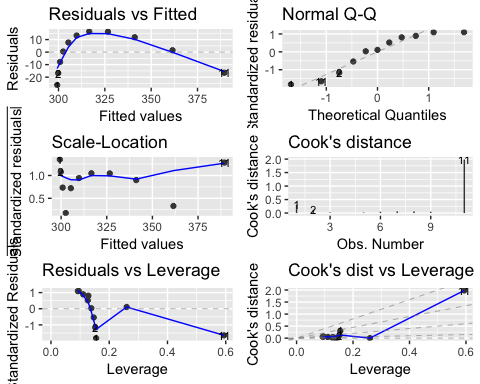
## 5.2 The following table gives the vapor pressure of water for various temperatures.

### a. Plot a scatter diagram. Does it seem likely that a straight-line model will be adequate?



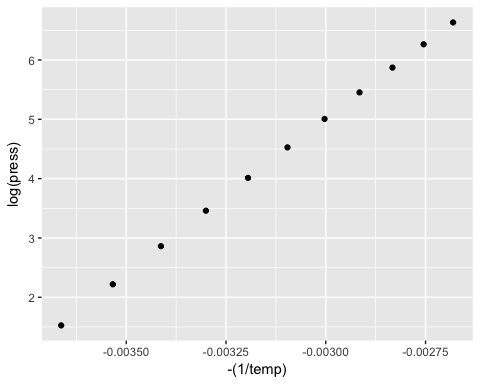
A straight line model does not seem adequate as the relationship appears to be exponential.

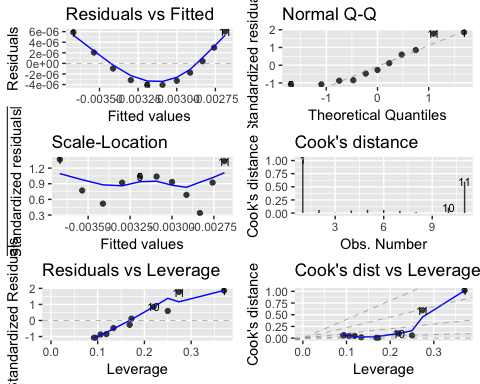
### b. Fit the straight-line model. Compute the summary statistics and the residual plots. What are your conclusions regarding model adequacy?



The summary statistics provide a good support for the goodness of the model. In fact, the Adj-Rsquared is quite high, and all the predictor and intercept are significant at 0.05.  
However, from the analysis of the residuals we observe how the model does not satisfied the model assumptions. Specifically, the homogeneity condition and the normality distribution of the errors are not satisfied.  
Therefore, the stragiht line model is not adequate to explain the relationship between pressure and temperature.

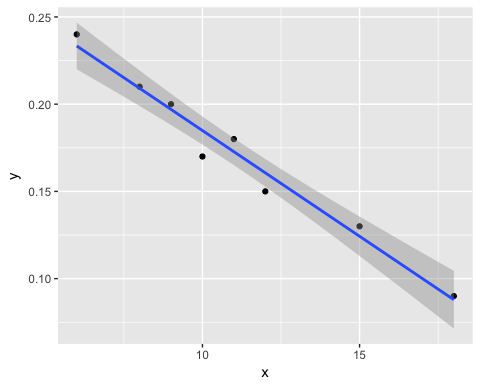
### c. From physical chemistry the Clausius-Clapeyron equation states that: Repeat part b using the appropriate transformation based on this information.





As we can observe from the summary statitics, the model with the transformed predictors and response variables perfectly explains the relationship between the variables. In fact, the Adj-Rsquared is 1, and all the are significant at 0.05.  
Furthermore, the residual plot show how the model assumptions are satisfied. Therefore, this models perdicts and explains perfectly the relationship between pressure and temperature.

## 5.4 Consider the data shown below. Construct a scatter diagram and suggest an appropriate form for the regression model. Fit this model to the data and conduct the standard tests of model adequacy.



From the scatterplot we can observe how there seems to be a strong linear relationship between the predictors and predicted variables. Therefore, I model a straight line model.

Dependent variable:

y

x

-0.012\*\*\*

p = 0.00001

Constant

0.306\*\*\*

p = 0.00000

Observations

8

R2

0.970

Adjusted R2

0.965

Residual Std. Error

0.009 (df = 6)

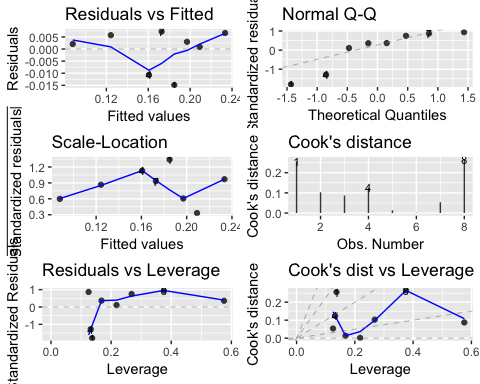
F Statistic

193.446\*\*\* (df = 1; 6)

Note:

*p<0.1;* ***p<0.05;*** p<0.01

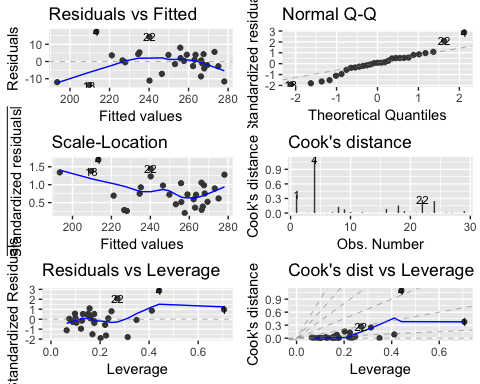
From the summary statistics we can observe how the model explains almost all variability in the data, as the R-squared is 0.97 and predictor is significant at 0.05.



As we can observe from the residual plots, the model appears to do not meet some of the assumptions. In particular, we can se how at the tails, the residuals deviate from a normal distribution. Furthermore, the rediduals do not seem to have an homogeneous variance.

## 6.1 Perform a thorough influence analysis of the solar thermal energy test data given in Table B.2. Discuss your results.

Following we will analyze the residuals in order to determine the observations that can be considered outliers, leverage, or influential.



From the Cook's D plot we can observe how there is an observation which has a value > 1. From the residuals vs leverage plot we observe how there are three observations with levarage > 0.4. From the Cook's D vs Leverage plot we observe that there are two observations that might be problematic.  
Following we identify those observations that exceed the following thresholds:  
1. hat: = 0.4137931  
2. cook's d: > 1  
3. r-student: > 2-3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| y | x1 | x2 | x3 | x4 | x5 | .fitted | .se.fit | .resid | .hat | .sigma | .cooksd | .std.resid | .squared | .r.student |
| 271.8 | 783.35 | 33.53 | 40.55 | 16.66 | 13.20 | 267.5709 | 6.752112 | 4.229148 | 0.7054608 | 8.050044 | 0.3750893 | 0.9693444 | 226.62659 | 0.5176377 |
| 230.7 | 827.80 | 33.13 | 32.52 | 17.50 | 10.53 | 213.3038 | 5.338316 | 17.396220 | 0.4409630 | 6.554166 | 1.1012177 | 2.8942192 | 96.25046 | 2.3715550 |

From the above table we can observe how there are only two observations which might be possible problematic for our model. The first one has a high hat value, meaning it is a leverage point. The cook's D and r-student values do not exceed the thresholds and therefore the observation is not influential. On the other hand, the observation in the second has a cook's D value > 1, meaning that it is an influential point. In fact, it has high hat and r-student values.

## 7.18 An article in the Journal of Pharmaceutical Sciences (80, 971–977, 1991) presents data on the observed mole fraction solubility of a solute at a constant temperature, along with x1 = dispersion partial solubility, x2 = dipolar partial solubility, and x3 = hydrogen bonding Hansen partial solubility. The response y is the negative logarithm of the mole fraction solubility.

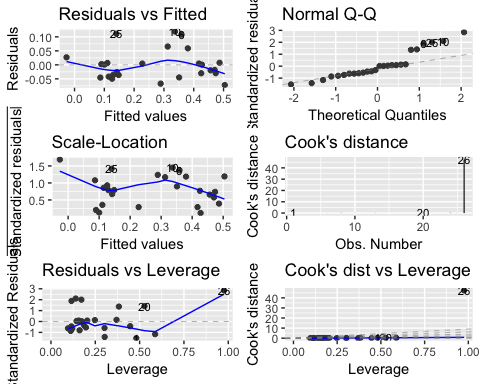
### a. Fit a complete quadratic model to the data.

### b. Test for significance of regression, and construct t statistics for each model parameter. Interpret these results.

##   
## Call:  
## lm(formula = y ~ x1 + I(x1^2) + x2 + I(x2^2) + x3 + I(x3^2),   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.072137 -0.036041 -0.008668 0.007245 0.118944   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.8446143 1.2420061 -1.485 0.15390   
## x1 0.4491296 0.2889529 1.554 0.13660   
## I(x1^2) -0.0214860 0.0166719 -1.289 0.21296   
## x2 0.0155739 0.0328394 0.474 0.64073   
## I(x2^2) 0.0011441 0.0057537 0.199 0.84450   
## x3 -0.0385807 0.0119318 -3.233 0.00437 \*\*  
## I(x3^2) 0.0005904 0.0008048 0.734 0.47211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06089 on 19 degrees of freedom  
## Multiple R-squared: 0.9015, Adjusted R-squared: 0.8704   
## F-statistic: 28.97 on 6 and 19 DF, p-value: 1.376e-08

The F-statistic for the model is significant at 0.05, therefore at least one parameter is significantly different from 0. As we can observe from the above output only one predictor parameter is significant at 0.05 (i.e. hydrogen). All other parameters are non significant. Overall, the complete quadratic model does not seem to be a good fit of the data, in fact, the high R-square is likely the result of overfitting and multicollinearity.

### c. Plot residuals and comment on model adequacy.



From the residuals plots we conform that the model does not provide a good fit for the data. In particular, we observe how the distribution of the error is significantly non-normal.

### d. Use the extra-sum-of-squares method to test the contribution of all second-order terms to the model.

Following we analyze the extra sum squares type II.

## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 0.41529 0.41529 111.9997 2.097e-09 \*\*\*  
## I(x1^2) 1 0.02751 0.02751 7.4193 0.01348 \*   
## x2 1 0.02507 0.02507 6.7619 0.01757 \*   
## I(x2^2) 1 0.00068 0.00068 0.1839 0.67283   
## x3 1 0.17405 0.17405 46.9396 1.545e-06 \*\*\*  
## I(x3^2) 1 0.00200 0.00200 0.5383 0.47211   
## Residuals 19 0.07045 0.00371   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As we can observe from the above table, the second-order term that is significant is x1^2. None of the other second-order terms seem to significantly contribute to explaining the variance of the predicted variable.

## 7.19 Consider the quadratic regression model from Problem 7.18. Find the variance inflation factors and comment on multicollinearity in this model.

## x1 I(x1^2) x2 I(x2^2) x3 I(x3^2)   
## 503.18411 494.56615 25.36125 39.63330 19.87052 31.11576

The VIF values signal strong problems of collinearity, however in this particular case it is not interpretable because by definition the second-order terms are non dependent from from the first order terms.

## 7.20 Consider the solubility data from Problem 7.18. Suppose that a point of interest is x1 = 8.0, x2 = 3.0, and x3 = 5.0.

### a. For the quadratic model from Problem 7.18, predict the response at the point of interest and find a 95% confidence interval on the mean response at that point.

## fit lwr upr  
## 1 0.2521972 0.1194738 0.3849206

## [1] 0.2654468

### b. Fit a model that includes only the main effects and two-factor interactions to the solubility data. Use this model to predict the response at the point of interest. Find a 95% confidence interval on the mean response at that point.

## fit lwr upr  
## 1 0.251212 0.1187472 0.3836769

## [1] 0.2649297

### c. Compare the lengths of the confidence intervals in parts a and b. Can you draw any conclusions about the best model from this comparison?

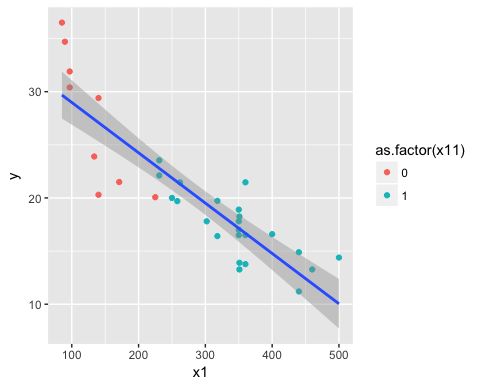
The model with only main effect and interactions provides a narrower CI, therefore it is preferred as it is able to predict with higher precition new observations.

## 8.4 Consider the automobile gasoline mileage data in Table B.3.

### a. Build a linear regression model relating gasoline mileage y to engine displacement x1 and the type of transmission x11. Does the type of transmission significantly affect the mileage performance?

No, the type of transmission does not appears to significantly influence the mileage performance of the vehicles.

### b. Modify the model developed in part a to include an interaction between engine displacement and the type of transmission. What conclusions can you draw about the effect of the type of transmission on gasoline mileage? Interpret the parameters in this model.



With the introduction of the interaction term now the for the possibility of chaging the slope of the regression line for the two types of transmission. With the introduction of the interaction effect the main trainsmission effect becomes significant. The intercept becomes relevant at explaining the data as we consider also different slopes for the two types of transmissions.

## 8.9 Suppose that a one-way analysis of variance involves four treatments but that a different number of observations (e.g., ni) has been taken under each treatment.Assuming that n1 = 3, n2 = 2, n3 = 4, and n4 = 3, write down the y vector and X matrix for analyzing these data as a multiple regression model. Are any complications introduced by the unbalanced nature of these data?

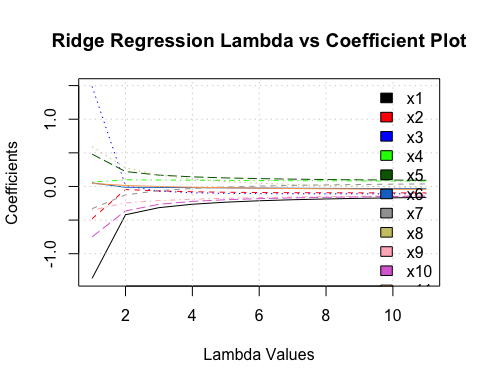
## 8.12 Two-Way Analysis of Variance.

### a. For the case a = b = n = 2, write down a regression model that corresponds to the two-way analysis of variance.

## 9.19 Estimate the parameters in a model for the gasoline mileage data in Table B.3 using ridge regression.

##   
## Call:  
## linearRidge(formula = y ~ ., data = auto)  
##   
##   
## Coefficients:  
## Estimate Scaled estimate Std. Error (scaled) t value (scaled)  
## (Intercept) 22.267878 NA NA NA  
## x1 -0.009763 -6.110955 1.602629 3.813  
## x2 -0.013665 -3.285653 2.097931 1.566  
## x3 -0.008461 -3.785128 1.480403 2.557  
## x4 1.919083 2.865588 2.728332 1.050  
## x5 1.200367 3.420072 2.666478 1.283  
## x6 -0.180319 -1.041594 2.636329 0.395  
## x7 0.283822 1.010129 2.513611 0.402  
## x8 0.001023 0.115827 2.515311 0.046  
## x9 -0.165102 -5.048734 2.616790 1.929  
## x10 -0.001058 -5.359236 1.571835 3.410  
## x11 -0.547177 -1.325329 2.902709 0.457  
## Pr(>|t|)   
## (Intercept) NA   
## x1 0.000137 \*\*\*  
## x2 0.117316   
## x3 0.010563 \*   
## x4 0.293577   
## x5 0.199626   
## x6 0.692775   
## x7 0.687784   
## x8 0.963271   
## x9 0.053686 .   
## x10 0.000651 \*\*\*  
## x11 0.647971   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Ridge parameter: 0.2665948, chosen automatically, computed using 3 PCs  
##   
## Degrees of freedom: model 4.602 , variance 3.086 , residual 6.117

### a. Use the ridge trace to select an appropriate value of k. Is the resulting model adequate?



##   
## Call:  
## lm(formula = y ~ x1 + x3 + x5 + x8 + x10, data = auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4396 -1.4704 -0.1924 1.7536 5.6469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.869116 14.962543 0.392 0.6983   
## x1 -0.026846 0.041062 -0.654 0.5195   
## x3 0.036691 0.047057 0.780 0.4432   
## x5 2.612022 1.594539 1.638 0.1144   
## x8 0.200253 0.115541 1.733 0.0959 .  
## x10 -0.009006 0.004360 -2.066 0.0498 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.016 on 24 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.8084, Adjusted R-squared: 0.7684   
## F-statistic: 20.25 on 5 and 24 DF, p-value: 6.645e-08

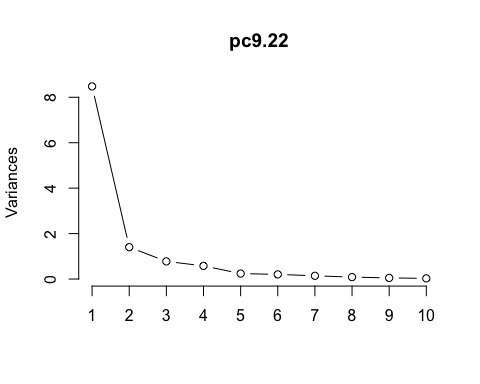
From the above plot we observe how the most important predictors are the following: x1, x3, x5, x8, x10.  
The model resulting from the ridge does not seems to be adequate as only one of the predictors is statistically significant.

### c. How much reduction in R2 has resulted from the use of ridge regression?

## [1] 3

The difference in R square from the full model and the model fitted using the k predictors suggested as an outcome of the ridge regression is 3%.

## 9.22 Estimate the model parameters for the gasoline mileage data using principal- component regression.



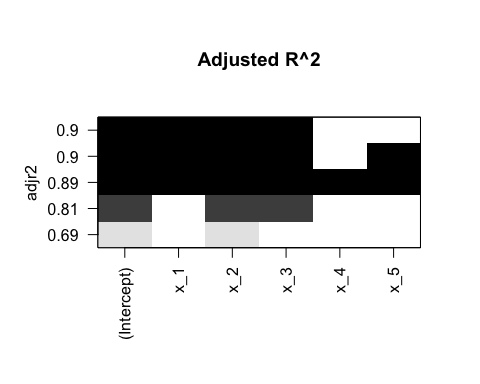
I decided to keep only the first four Principal Components to use in the regression model. In fact, after the four compoment their contribution stibilizes.

### a. How much has the residual sum of squares increased compared to least squares?

The residual sum of squares of the PC model increased compared to the full model by 34.2147847.

## 10.34 Use the all-possible-regressions selection on the methanol oxidation data in Table B.20. Perform a thorough analysis of the best candidate models. Compare your results with stepwise regression. Thoroughly discuss your recommendations.

## x\_1 x\_2 x\_3 x\_4 x\_5  
## 1 ( 1 ) \*   
## 2 ( 1 ) \* \*   
## 3 ( 1 ) \* \* \*   
## 4 ( 1 ) \* \* \* \*  
## 5 ( 1 ) \* \* \* \* \*



## Start: AIC=93.17  
## y ~ x\_1 + x\_2 + x\_3 + x\_4 + x\_5  
##   
## Df Sum of Sq RSS AIC  
## - x\_4 1 5.8 1641.6 91.235  
## - x\_5 1 6.3 1642.2 91.241  
## <none> 1635.9 93.172  
## - x\_1 1 1304.9 2940.8 101.729  
## - x\_3 1 2088.1 3724.0 105.979  
## - x\_2 1 4436.7 6072.6 114.781  
##   
## Step: AIC=91.24  
## y ~ x\_1 + x\_2 + x\_3 + x\_5  
##   
## Df Sum of Sq RSS AIC  
## - x\_5 1 27.3 1669.0 89.533  
## <none> 1641.6 91.235  
## - x\_1 1 1677.8 3319.5 101.909  
## - x\_3 1 2309.3 3951.0 105.044  
## - x\_2 1 4696.7 6338.4 113.552  
##   
## Step: AIC=89.53  
## y ~ x\_1 + x\_2 + x\_3  
##   
## Df Sum of Sq RSS AIC  
## <none> 1669.0 89.533  
## - x\_1 1 1812.9 3481.9 100.769  
## - x\_3 1 2384.1 4053.1 103.504  
## - x\_2 1 5011.9 6680.9 112.499

##   
## Call:  
## lm(formula = y ~ x\_1 + x\_2 + x\_3, data = methanol)  
##   
## Coefficients:  
## (Intercept) x\_1 x\_2 x\_3   
## -2618.418 21.973 3.818 99.624

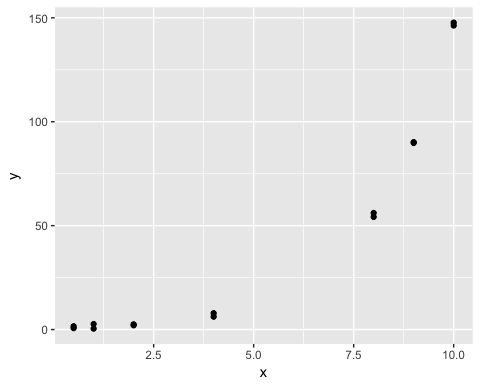
As we can observe from the leaps output plot the model with highest adj-Rsquared is the one that includes x1, x2. and x3. This result equals the one obtained by a backward step model selection.  
The model suggested by both approaches appears to be the more conservative in terms of the number of variables selected. In fact, the models with comparable results in terms of adj-Rsquared in the leaps output include one or two more variables compared to the models that includes only x1, x2, x3. Therefore, it appers that including other variables apart x1, x2, x3 do not increase significantly the explanation power of the model.

## 12.6 For the models shown below, determine whether it is a linear model, an intrinsically linear model, or a nonlinear model. If the model is intrinsically linear, show how it can be linearized by a suitable transformation.

1. Intrinsically linear
2. Non linears
3. Intrinsically linear
4. Non linear
5. Non linear

## 12.8 Consider the following observations:

### a. Fit the nonlinear regression model



##   
## Call:  
## lm(formula = log(y) ~ x, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.04644 -0.05337 -0.00899 0.11455 0.73039   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.28467 0.18455 -1.543 0.149   
## x 0.53261 0.02992 17.799 5.42e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4151 on 12 degrees of freedom  
## Multiple R-squared: 0.9635, Adjusted R-squared: 0.9605   
## F-statistic: 316.8 on 1 and 12 DF, p-value: 5.417e-10

I performed a lineariation on the non linear model in order to obtained the starting values. The linearized model is the following: $ log(u) = log(*{1}) +* {2}x$

### b. Test for significance of regression.

##   
## Formula: y ~ theta1 \* exp(theta2 \* x)  
##   
## Parameters:  
## Estimate Std. Error t value Pr(>|t|)   
## theta1 1.059107 0.048161 21.99 4.59e-11 \*\*\*  
## theta2 0.493408 0.004745 103.98 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8666 on 12 degrees of freedom  
##   
## Number of iterations to convergence: 5   
## Achieved convergence tolerance: 3.902e-06

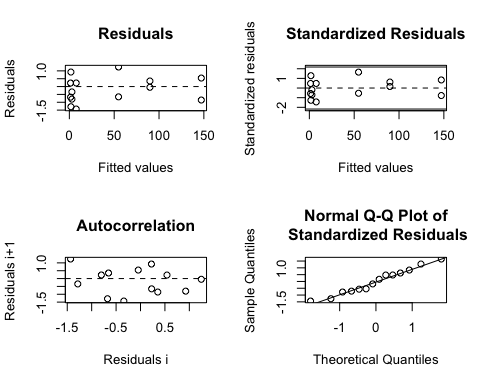
### c. Estimate the error variance σ2.

The estimate error variance equals 0.8666^2 = 0.7509956.

### d. Test the hypotheses H0: theta1 = 0 and H0: theta2 = 0. Are both model parameters different from zero? If not, refit an appropriate model.

From the model summary we observe how both model parameters are different from zero at 0.05 significance level.

### e. Analyze the residuals from this model. Discuss model adequacy.



The model appears to present a good fit of the data as the assumptions of normality and homogenesticity are soddisfied.

## 12.09 Reconsider the data in the previous problem. The response measurements in the two columns were collected on two different days. Fit a new model.

To these data, where x1 is the original regressor from Problem 12.8 and x2 is an indicator variable with x2 = 0 if the observation was made on day 1 and x2 = 1 if the observation was made on day 2. Is there any indication that there is a difference between the two days (use theta30 = 0 as the starting value).

##   
## Formula: y ~ theta3 \* x2 + theta1 \* exp(theta2 \* x)  
##   
## Parameters:  
## Estimate Std. Error t value Pr(>|t|)   
## theta1 1.065554 0.055038 19.360 7.57e-10 \*\*\*  
## theta2 0.492832 0.005315 92.732 < 2e-16 \*\*\*  
## theta3 -0.120985 0.411578 -0.294 0.774   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9016 on 11 degrees of freedom  
##   
## Number of iterations to convergence: 5   
## Achieved convergence tolerance: 3.79e-06

The parameter theta3 is not statistically diffirent from zero, therefore we have no evidence to support that the measurements collected in the two days are different.