STATISTICAL MODELLING

Multiple Linear Regression

20/May/2022

Content prepared by Dr. Sridhar Pappu

Multiple Linear Regression

THE OUTPUT

Multiple Linear Regression (MLR)

- Simple Linear Regression models the effect of one independent variable,, on one dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_n x_n + \varepsilon$$

- The β parameters reflect the independent contribution of each independent variable, x, to the value of the dependent variable, y.

Interpreting Regression Coefficients

A coefficient is the slope of the linear relationship between the dependent variable (DV) and the independent contribution of the independent variable (IV), i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.89666084					
R Square	0.804000661					
Adjusted R Square	0.750546296					
Standard Error	2.90902388					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	381.8467141	127.282238	15.04087945	0.00033002	
Residual	11	93.08661926	8.462419933			
Total	14	474.9333333				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	12.04617703	9.312399791	1.29356313	0.222319528	-8.450276718	32.54263077
Stock 2 ($)	0.878777607	0.26187309	3.355738482	0.006412092	0.302398821	1.455156393
Stock 3 ($)	0.220492727	0.143521894	1.536300286	0.152714573	-0.095396832	0.536382286

Assumptions of MLR

- Same as simple linear regression
  - Linearity
  - Independence of errors
  - Homoscedasticity (constant variance)
  - Normality of errors

- Methods of checking assumptions are also the same

Determining the MLR Equation

- $k + 1$ equations to solve for k independent variables and the intercept.
- In solving for intercept and slope in a simple linear regression model, we needed $\sum x$, $\sum y$, $\sum xy$, and $\sum x^2$.
- For multiple regression model with 2 independent variables, we need $\sum x_1$, $\sum x_2$, $\sum y$, $\sum {x_1}^2$, $\sum {x_2}^2$, $\sum x_1 x_2$, $\sum x_1 y$, and $\sum x_2 y$.

Determining the MLR Equation - Excel

In a real estate study, multiple variables were explored to determine the price of a house.

- # of bedrooms
- # of bathrooms
- Age of the house
- # of square feet of living space
- Total # of square feet of space
- # of garages

Find the equation if you want to predict the price of the house by total square feet and age of the house.

Determining the MLR Equation – Interpreting the Output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.860872681					
R Square	0.741101773					
Adjusted R Square	0.715211951					
Standard Error	11.96038667					
Observations	23					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	8189.723012	4094.861506	28.62521631	1.35298E-06	
Residual	20	2861.016988	143.0508494			
Total	22	11050.74				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.35074586	10.00715186	5.73097587	1.31298E-05	36.47619286	78.22529885
Area (sq ft) (x1)	0.017718036	0.00314562	5.632605205	1.63535E-05	0.011156388	0.024279685
Age of House (years) (x2)	-0.666347946	0.227996703	-2.922620973	0.008417613	-1.141940734	-0.190755157

What is the equation?

$$\hat{y} = 57.35 + 0.0177Area - 0.666Age$$

Are the coefficients and the model significant?

Yes

DSC 7402

Interpreting the Output (Residuals Analysis) - Homework Assignment

Residuals are determined the same way as in simple linear regression. The predicted value is calculated by substituting the predictor values of interest. The residual is again the difference between the observed and the predicted values, $y - \hat{y}$.

The assumptions are also tested through the same plots the same way as in Simple Linear Regression.

SSE and Standard Error of the Estimate, SE – Homework Assignment

$$SSE = \sum (y - \hat{y})^2$$

$$SE = \sqrt{\frac{SSE}{n - k - 1}}$$

Coefficient of Multiple Determination, R^2 – Homework Assignment

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Interpreting the Output (Adjusted R^2 (not R^2))

As additional independent variables are added to the regression model, R^2 increases.

$$\boldsymbol{R^2 = 1 - \frac{SSE}{SST}}$$

However, sometimes these variables are insignificant and add no real value but inflate R^2.

Adjusted R^2 takes into consideration both the additional information and the changed degrees of freedom (# of data points contributing to the variance).

$$\boldsymbol{Adjusted\ R^2} = 1 - \frac{\frac{SSE}{(n-k-1)}}{\frac{SST}{n-1}} = \boldsymbol{1 - \frac{MSE}{MST}}$$

k is the number of independent variables whose coefficients are to be computed.
Interpretation is the same as for R^2. We should always look at Adjusted R^2 in MLR.

Sample R Output

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$Sunshine + ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
      1       2       3       4       5       6       7       8
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813
      9      10
-1.1552  0.8429

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        31.6084     7.1051   4.449  0.00671 **
ToxinConc$Rain      7.0676     1.0031   7.046  0.00089 ***
ToxinConc$NoonTemp -0.4201     0.2413  -1.741  0.14215
ToxinConc$Sunshine -0.2375     0.5086  -0.467  0.66018
ToxinConc$WindSpeed -0.7936    0.2977  -2.666  0.04458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

Always remember, "significant" means $p < \alpha$ (generally, 0.05), which means null hypothesis (generally, $\beta = 0$) is rejected, i.e.,

$$\boldsymbol{significant \equiv H_0\ rejected \equiv p < 0.05}$$

Handling Special Situations

POLYNOMIAL REGRESSION

Polynomial Regression

For example, $y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \epsilon$
How is this a special case of the general linear model?
Replace ${x_1}^2$ with x_2, so that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Multiple linear regression assumes a linear fit of the regression **coefficients** and regression constant, but not necessarily a linear relationship of the independent variable values.

Polynomial Regression: Car Mileage Prediction

mtcars dataset in R

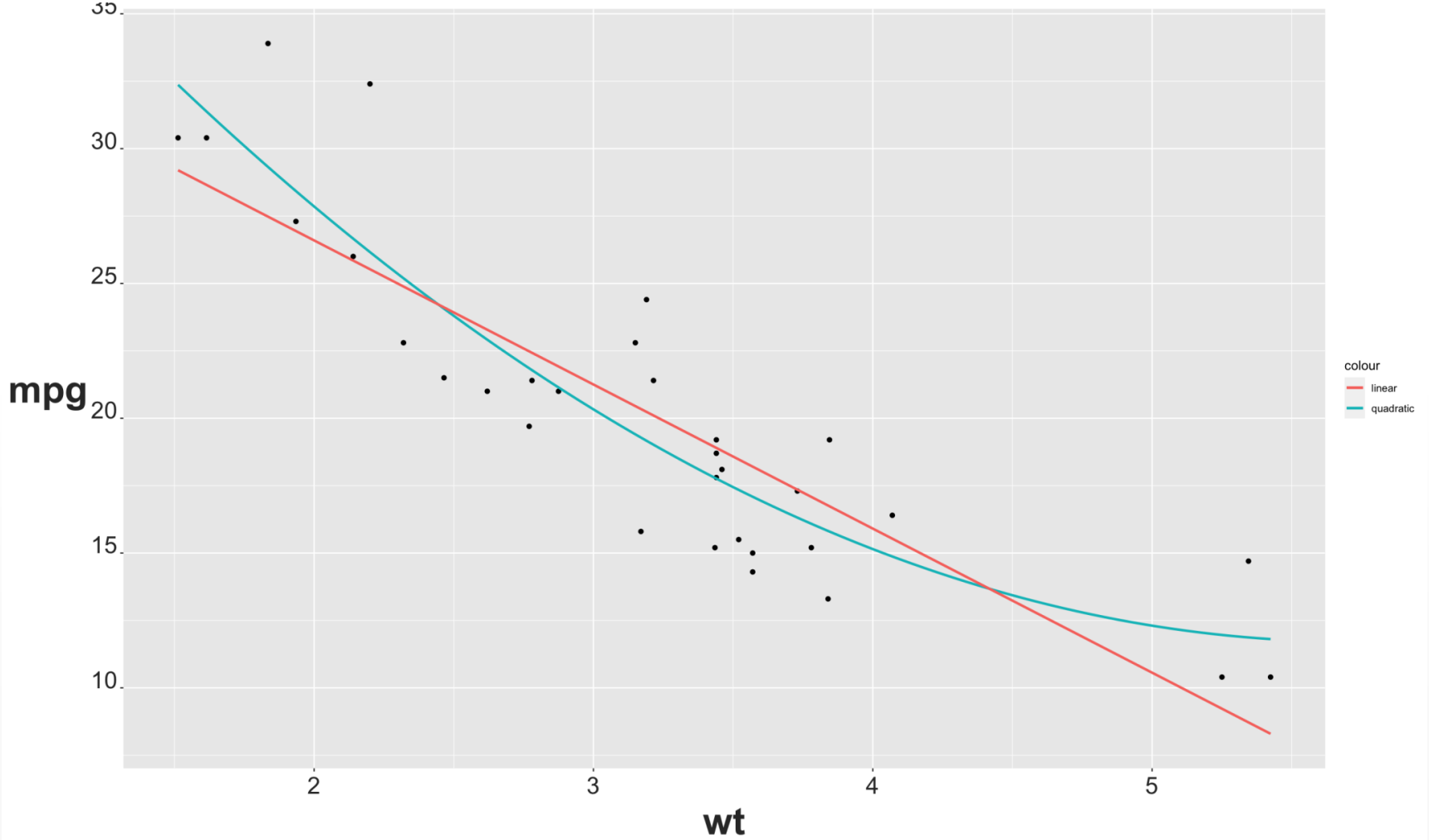
Data was extracted from the *Motor Trend* US magazine with a goal to predicting the fuel consumption (mpg) using 10 variables dealing with automobile design and performance.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Polynomial Regression – Predicting Car Mileage

Let us see the relationship between *mpg* and *wt* of the car.



Polynomial Regression – Predicting Car Mileage

Let us compare the linear and the quadratic models.

Linear Model

```
> lm1 <- lm(mpg ~ wt)
> summary(lm1)

Call:
lm(formula = mpg ~ wt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

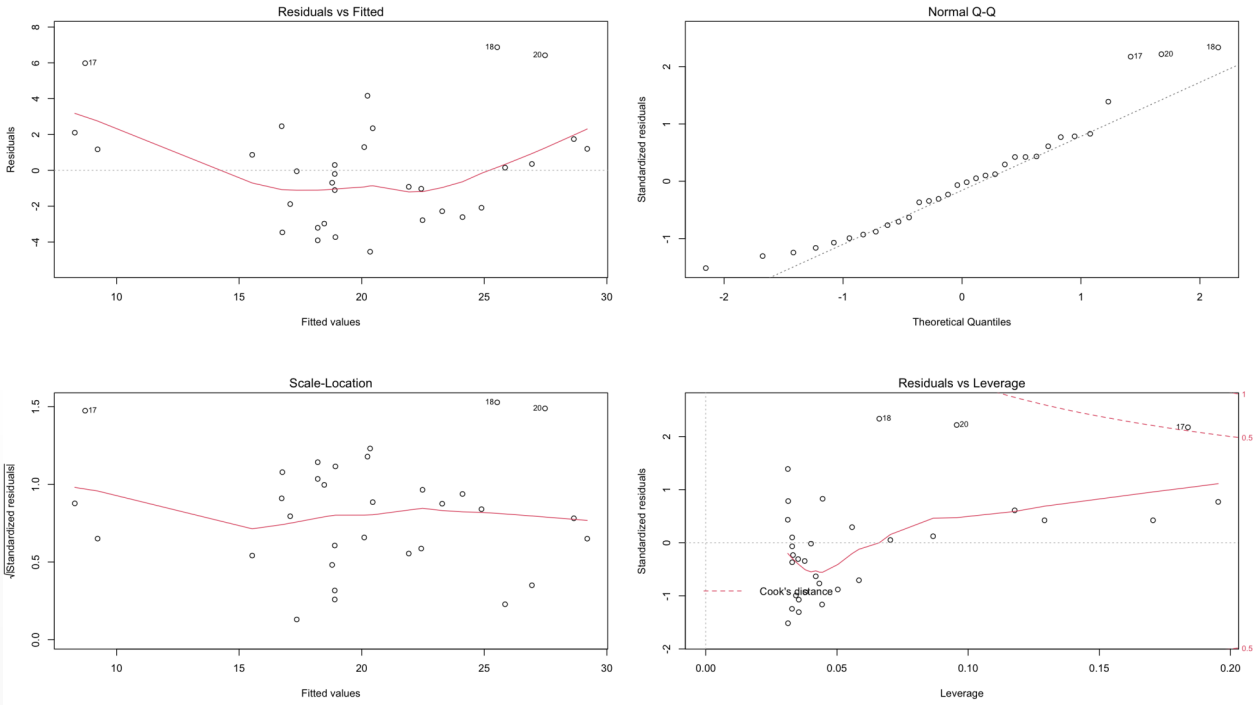
Quadratic (Polynomial of Order 2) Model

```
> lm2 <- lm(mpg ~ poly(wt,2, raw = TRUE))
> summary(lm2)

Call:
lm(formula = mpg ~ poly(wt, 2, raw = TRUE))

Residuals:
   Min     1Q Median     3Q    Max
-3.483 -1.998 -0.773  1.462  6.238

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    49.9308     4.2113  11.856 1.21e-12 ***
poly(wt, 2, raw = TRUE)1 -13.3803     2.5140  -5.322 1.04e-05 ***
poly(wt, 2, raw = TRUE)2   1.1711     0.3594   3.258  0.00286 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.651 on 29 degrees of freedom
Multiple R-squared:  0.8191,    Adjusted R-squared:  0.8066
F-statistic: 65.64 on 2 and 29 DF,  p-value: 1.715e-11
```

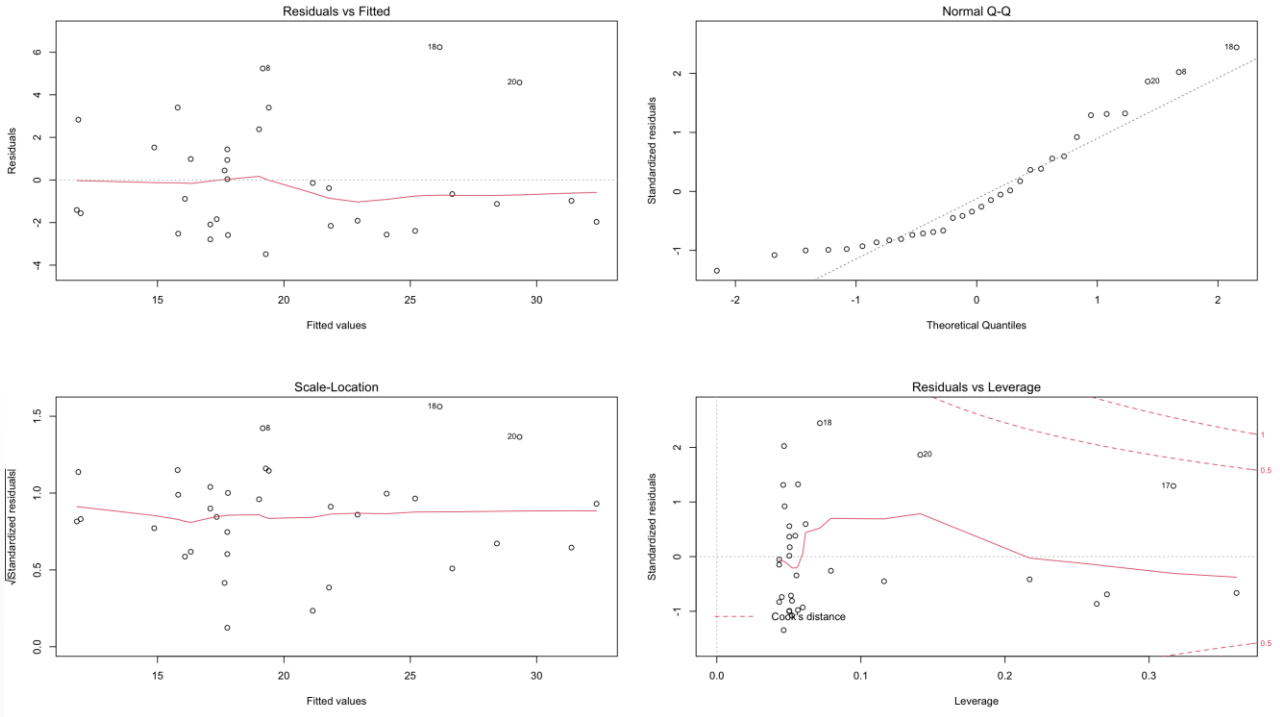

Polynomial Regression – Predicting Car Mileage

Let us compare the linear and the quadratic models.

Linear Model



Quadratic (Polynomial of Order 2) Model



SC 7402

Handling Special Situations

DATA TRANSFORMATIONS

Tukey's Ladder of Powers/Transformations

Transformations

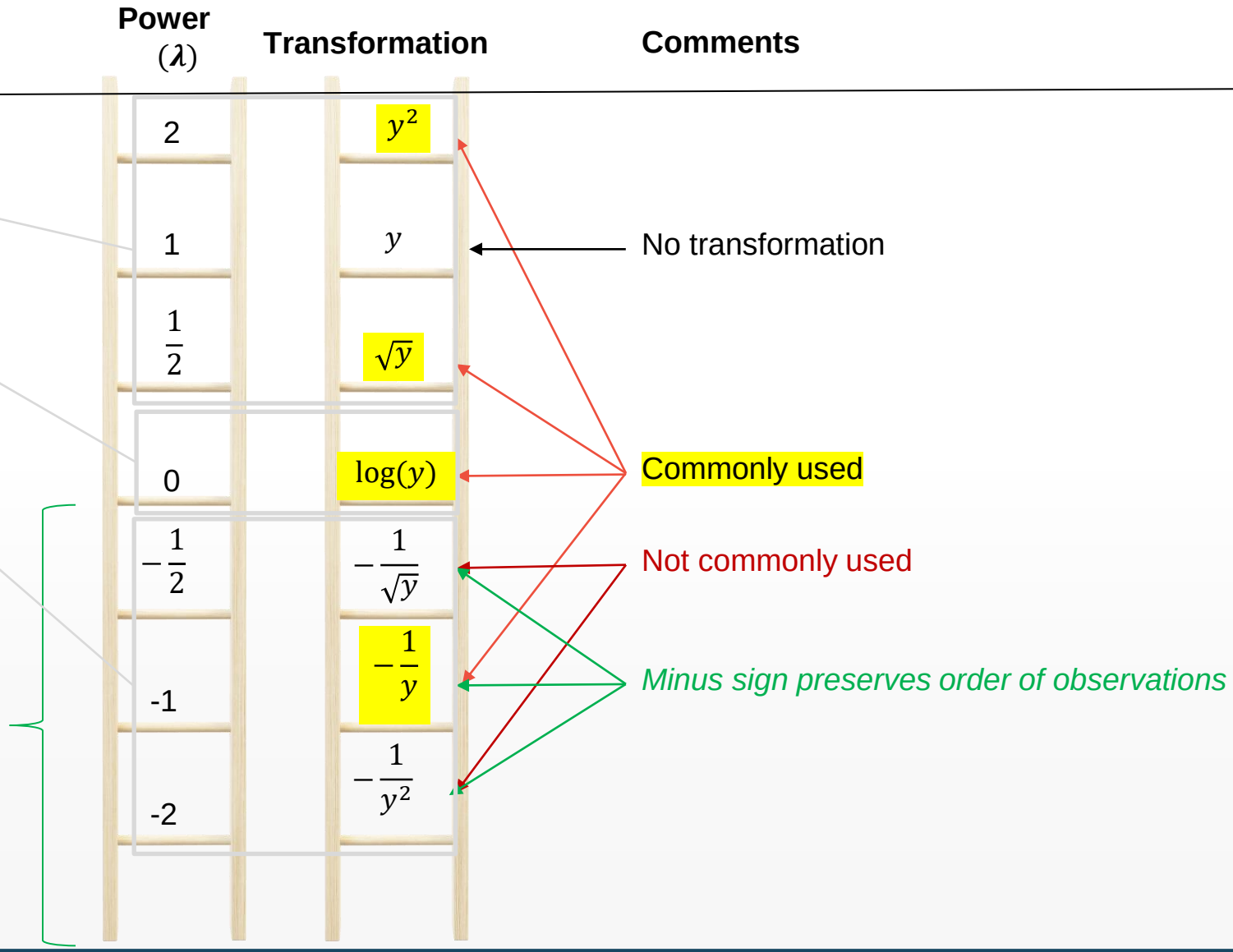
$y^{\lambda}\ if\ \lambda > 0$

$\log(y)\ if\ \lambda = 0$

$-(y^{\lambda})\ if\ \lambda < 0$

EXAMPLE

y	$\frac{1}{y}$	$-\frac{1}{y}$
2	0.5	-0.5
4	0.25	-0.25
5	0.2	-0.2
10	0.1	-0.1



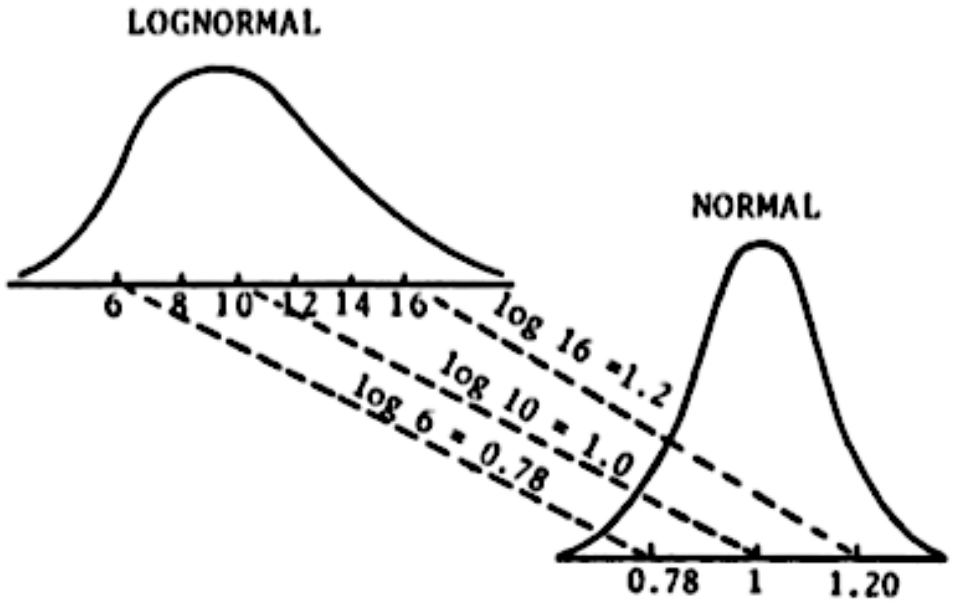
More Thoughts on Transformations

DATA TRANSFORMATION

As suggested by Tabachnick and Fidell (2007) and Howell (2007), the following guidelines (including SPSS compute commands) should be used when transforming data.

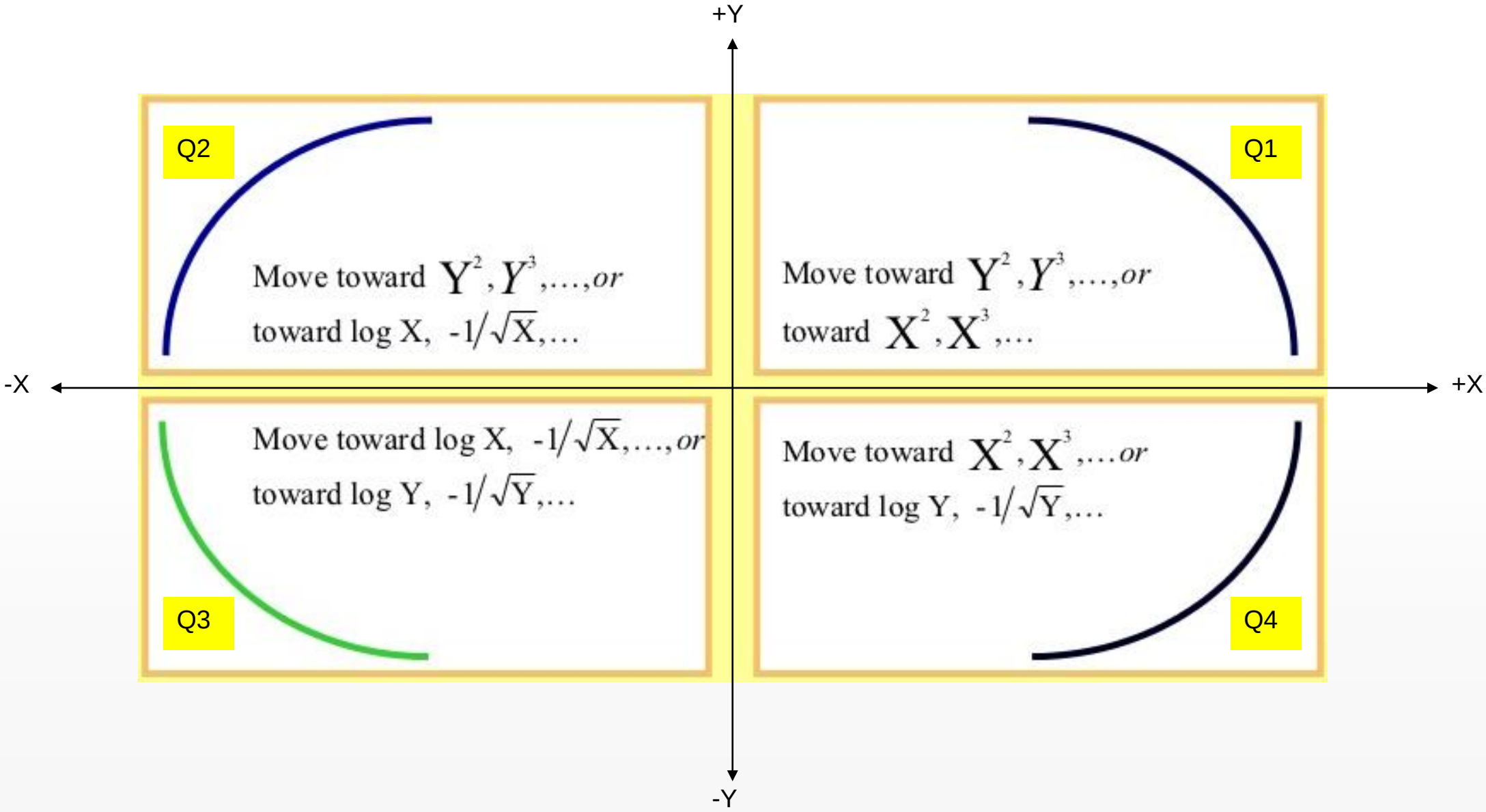
If your data distribution is…	Use this transformation method.
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

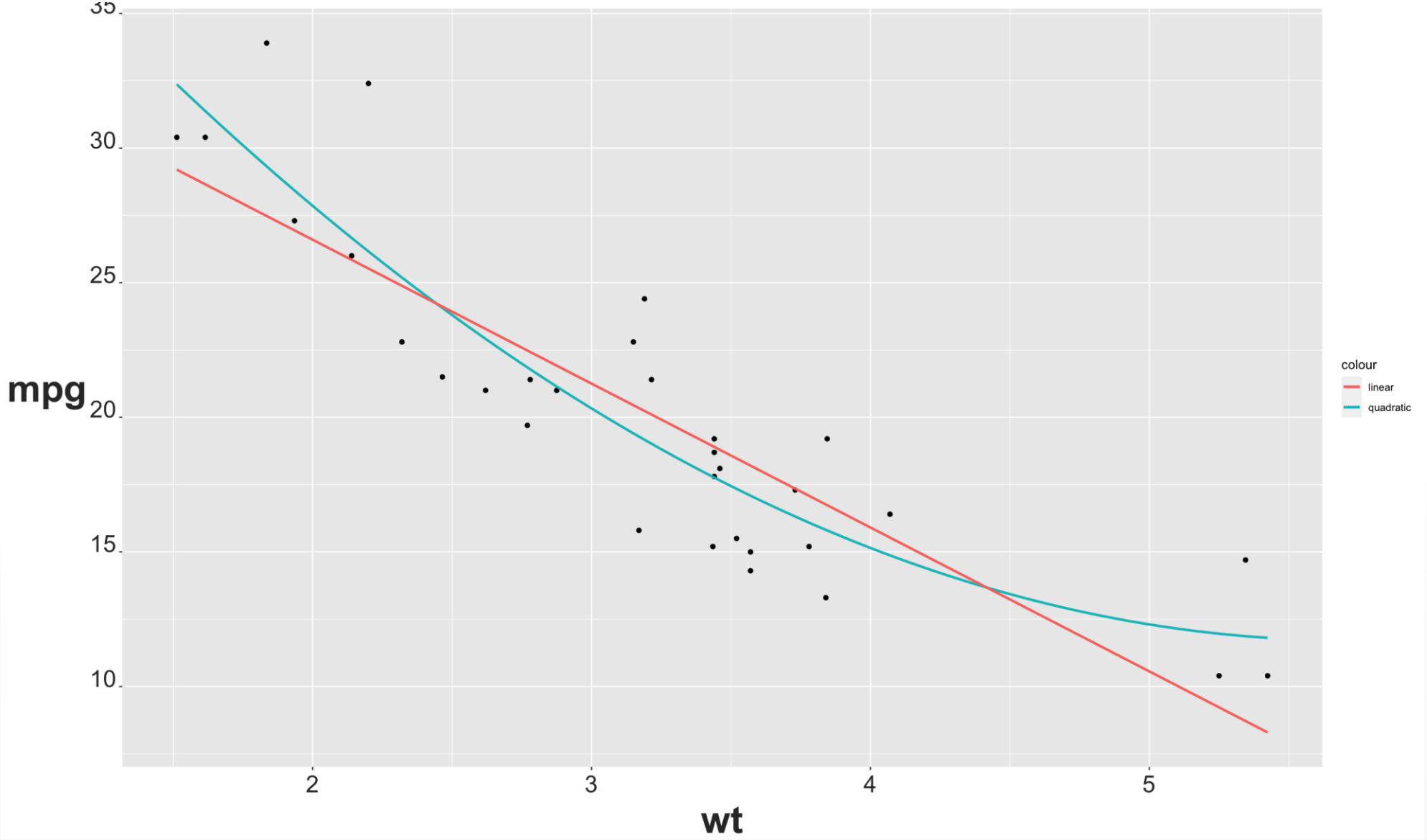


Source: http://oak.ucc.nau.edu/rh232/courses/eps625/handouts/data%20transformation%20handout.pdf
Last accessed: May 12, 2016

Tukey's Four-Quadrant Approach



Predicting Car Mileage



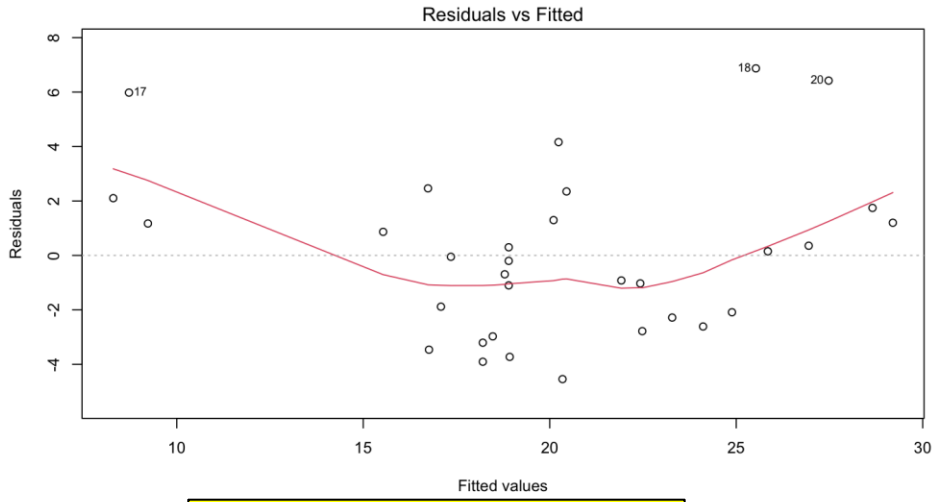
Predicting Car Mileage – R Output Analysis

```
> lm1 <- lm(mpg ~ wt)
> summary(lm1)

Call:
lm(formula = mpg ~ wt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

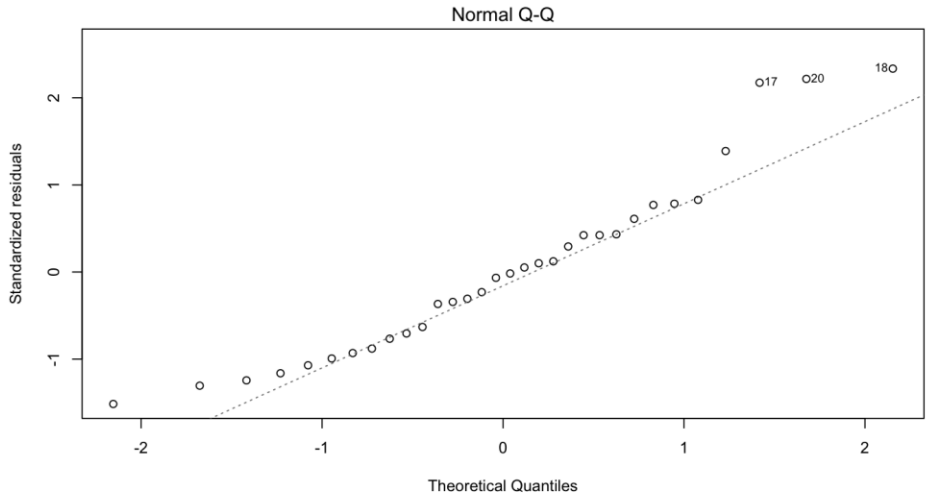
R^2 is good and the variable and the model are **significant**. Checks #1, #2 and #3 of the Linear Regression model analysis point to a good model.

Predicting Car Mileage – Residuals Analysis

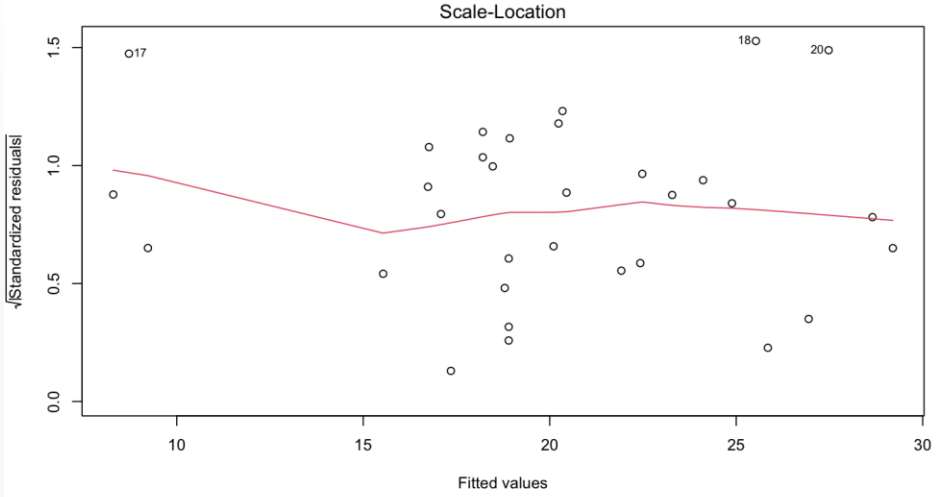
Is a wrong model fitted (linear or quadratic, etc.)?



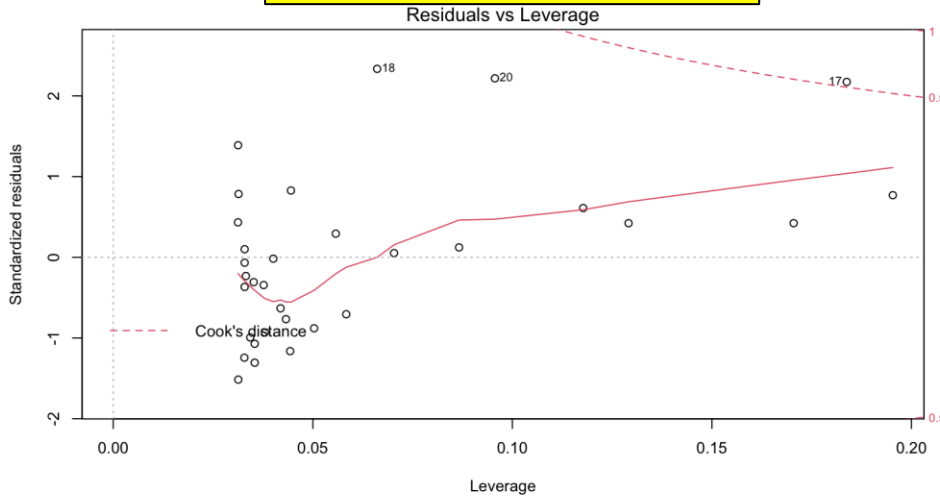
Are the residuals normally distributed?



Is the data homoscedastic?



Are there influential points?

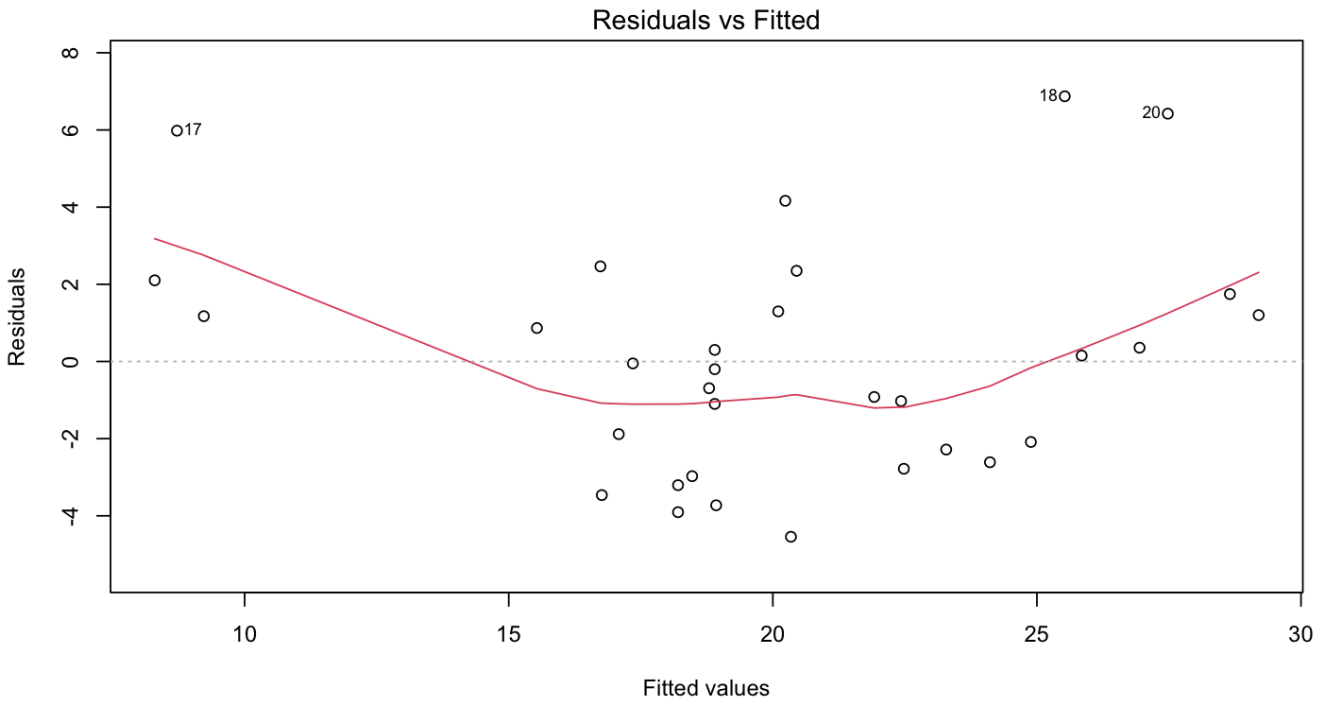


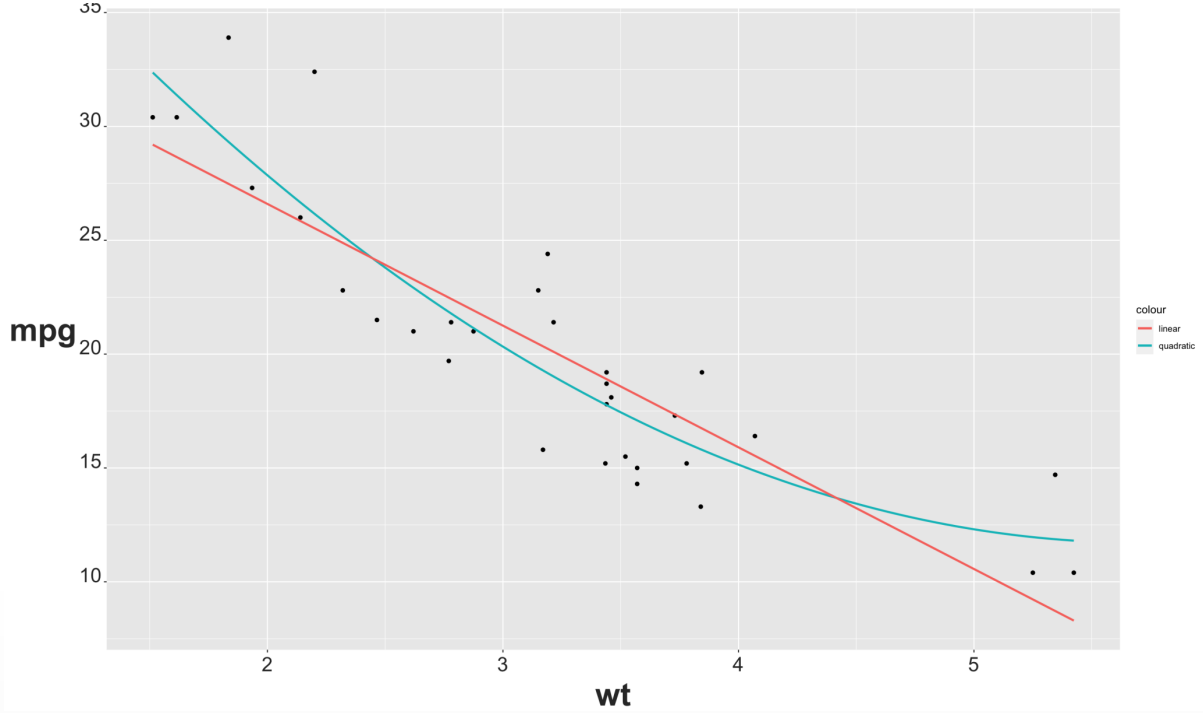
Linearity and Normality **assumptions are violated** and there is a potential **influential observation**.

Check #4 suggests a better model is possible if these issues could be fixed.

DSC 7402

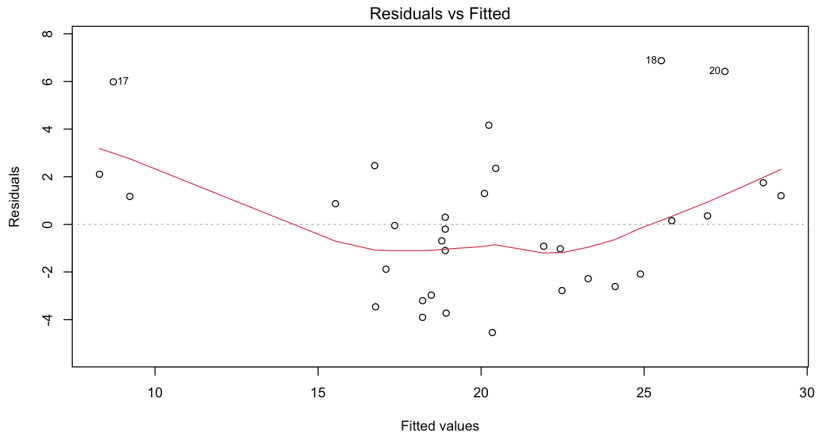
Based on Tukey's 4-Quadrant Approach, what transformation(s) do you recommend?





DSC 7402

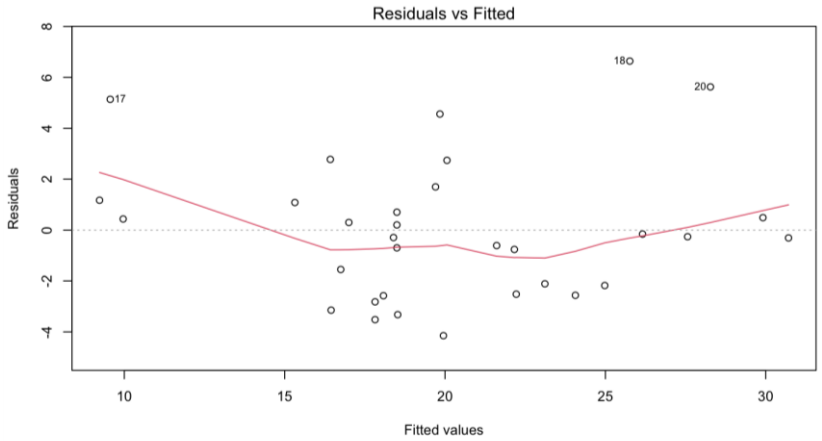
No Transformation



$\widehat{\text{mpg}} = -5.3445 * wt + 37.2851$

Adj. R^2 = 74.46%

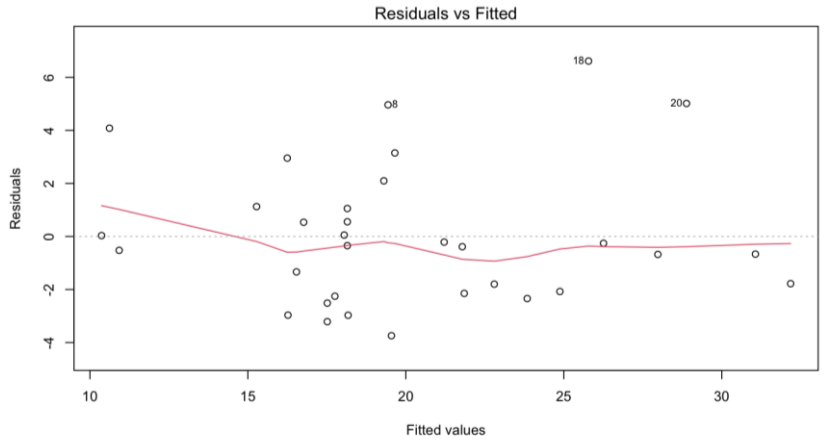
SQRT Transformation on X



$\widehat{mpg} = -19.547 * \sqrt{wt} + 54.752$

Adj. R^2 = 78.34%

Log Transformation on X



$\widehat{\text{mpg}} = -17.086 * ln(wt) + 39.257$

Adj. R^2 = 80.38%

Log transformed data appears the best, both from the residual plot and Adj. R^2. Look at the other residual plots as well. Finally, the real test would be how the model performs on new data.

Data Transformations Using R

- *transformTukey() in rcompanion package of R. Python currently does not have Tukey transformation options.*
- *boxcox() in MASS package of R*
- *boxCox() in car package of R that includes other families of transformations like Yeo-Johnson, etc.*
- *boxTidwell() in car package of R to transform predictors*

Note: Box-Cox requires strictly positive values of y whereas Yeo-Johnson can handle negative values as well.

Tukey Transformation

$$\begin{cases} y^{\lambda} \; if \; \lambda > 0 \\ log(y) \; if \; \lambda = 0 \\ -\left(y^{\lambda}\right) \; if \; \lambda < 0 \end{cases}$$

Box-Cox Transformation

$$\begin{cases} \frac{y^{\lambda}-1}{\lambda} \; if \; \lambda \neq 0 \\ log(y) \; if \; \lambda = 0 \end{cases}$$

R Code

transformTukey() conducts Ladder of Powers on a vector of values to produce a more normally distributed vector of values. For example, *transformTukey(wt)* will transform the *wt* variable in the Car Mileage example.

boxCox() transforms the dependent variable to follow a normal distribution. The argument is a formula or fitted model object of class *lm*. For example, *boxCox(lm1)* where $lm1 = lm(mpg \sim wt, data = mtcars)$

DSC 7402

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252.

Approach to determine whether to transform X or Y to achieve **linearity**, **homoscedasticity** and **normality**:

1. Often, a transformation that fixes one, fixes all.

2. In general, transforming both is not required, although sometimes it is.

3. A general rule of thumb:

    a. Transform Y first to remove heteroscedasticity and non-normality.
    b. Then transform X to remove non-linearity.

Handling Special Situations

INTERACTION TERMS

Interaction Terms

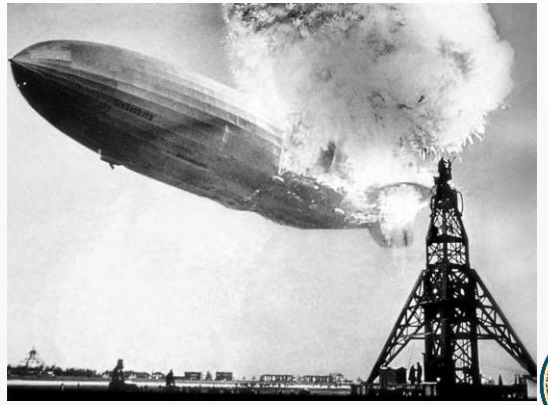
Interaction can be examined as a separate independent variable in regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

For example,

- Individually each of two drugs might improve symptoms, but when taken together, they may interact and cause a decline in health.

- Fire increases a balloon's levity (as in the hot air balloons). Hydrogen also increases levity as in the Zeppelins. But fire and hydrogen together dramatically reduce the levity (and that is an understatement).



Interaction Terms

Interaction can be examined as a separate independent variable in regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

2) Exercise

In a study of 6 type 1 diabetic patients, long-term exercise over 8 months helped improve fructosamine and HbA1c levels. However, short-term, acute exercise increased both fructosamine and HbA1c [23].

Doctors should take exercise into consideration when testing fructosamine.

2) Obesity

Obese people have lower fructosamine levels than non-obese people, even among diabetics [14, 15, 16].

If you are overweight, this test may underestimate the average glucose levels in your blood.

Source: https://labs.selfdecode.com/blog/fructosamine/
Last accessed: September 17, 2021

Interaction Terms – Interpreting Coefficients

Interaction can be examined as a separate independent variable in regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

If $x_2 = 0$, the coefficient of x_1 is β_1. If x_2 is now increased by 1 unit, i.e., $x_2 = 1$, then coefficient of x_1 is $\beta_1 + \beta_3$. β_3, thus, is the additional increase in the effect of x_1 when x_2 is increased by one unit.

The coefficient β_3 measures the *(interaction)* effect of a one-unit increase in **both** x_1 and x_2 over and above the sum of their independent effects, also called the *main* effects, which are measured by β_1 and β_2, respectively.

Interaction Terms – Predicting Car Mileage

Without Interaction Terms

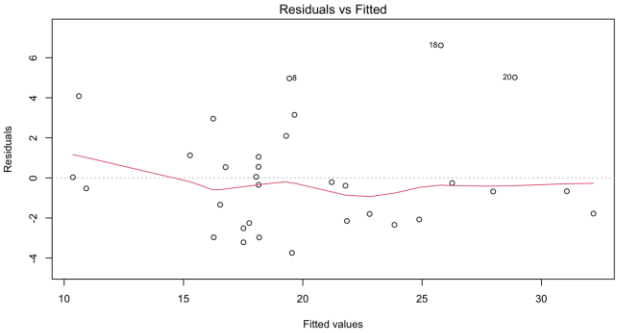
```
Call:
lm(formula = mpg ~ log(wt))

Residuals:
    Min      1Q  Median      3Q     Max
-3.7440 -2.0954 -0.3672  1.0709  6.6150

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.257      1.758   22.32  < 2e-16 ***
log(wt)      -17.086      1.510  -11.31 2.39e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.669 on 30 degrees of freedom
Multiple R-squared:  0.8101,    Adjusted R-squared:  0.8038
F-statistic:   128 on 1 and 30 DF,  p-value: 2.391e-12
```

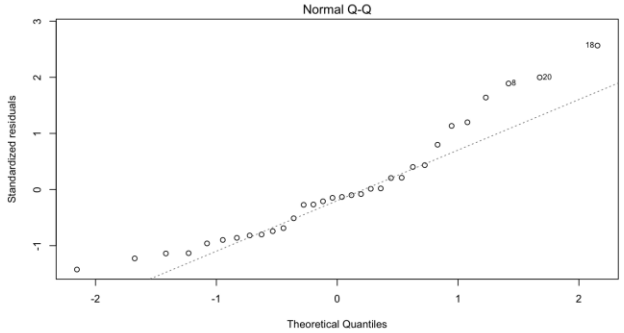
With Interaction Terms

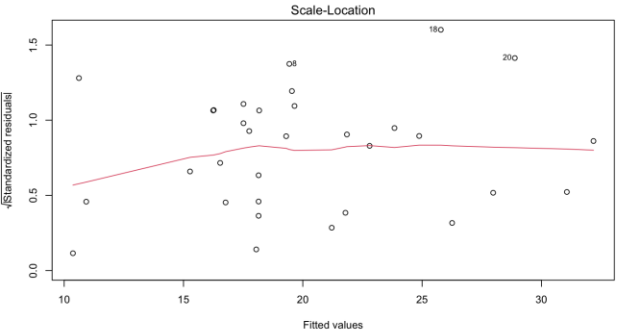
```
Call:
lm(formula = mpg ~ log(wt) * qsec)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5237 -1.5282 -0.3631  0.9451  4.9413

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.220     24.059  -0.175   0.8620
log(wt)         7.470     21.124   0.354   0.7263
qsec            2.364      1.330   1.777   0.0864 .
log(wt):qsec   -1.318      1.175  -1.121   0.2716
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.167 on 28 degrees of freedom
Multiple R-squared:  0.8832,    Adjusted R-squared:  0.8707
F-statistic:  70.6 on 3 and 28 DF,  p-value: 3.584e-13
```

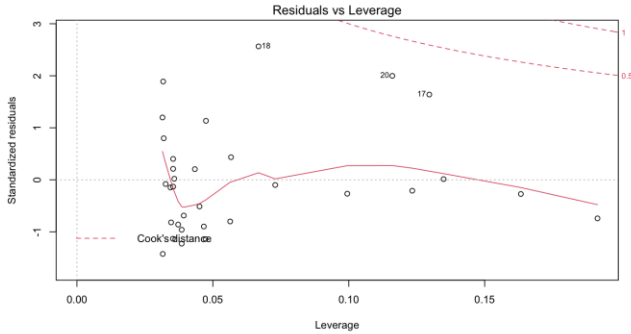

Interaction Terms – Predicting Car Mileage

Without Interaction Terms

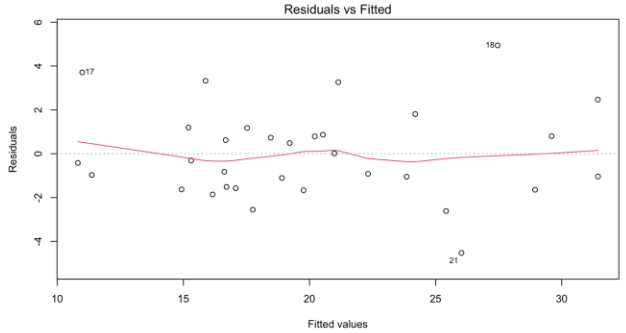


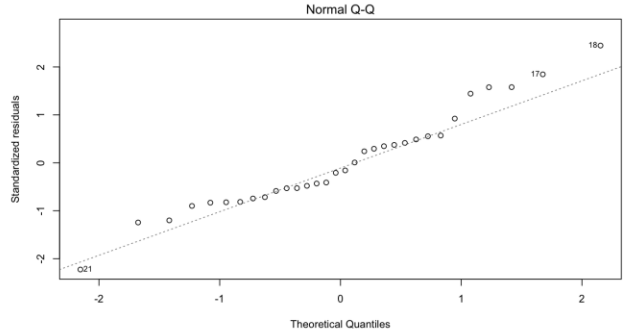


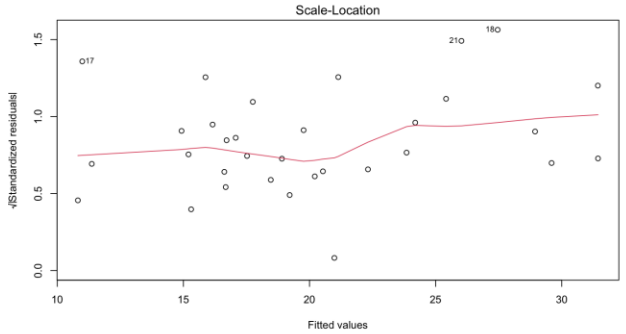


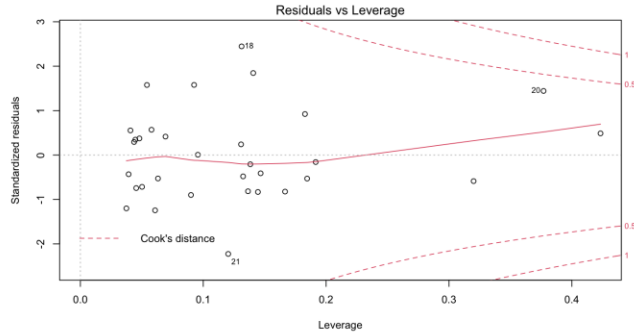


With Interaction Terms









DSC 7402

Multiple Linear Regression

HANDLING CATEGORICAL PREDICTORS

Indicator (Dummy) Variables

Categorical variables such as gender, geographic region, occupation, marital status, level of education, economic class, religion, buying/renting a home, etc. can also be used in multiple regression analysis.

If there are *n* levels in a category, *n-1* dummy variables need to be inserted into the regression analysis replacing that category.

Indicator (Dummy) Variables

If a survey question asks about the region of country your office is located in, with North, South, East and West as the options, the **recoding** can be done as follows:

Region
North
East
North
South
West
West
East

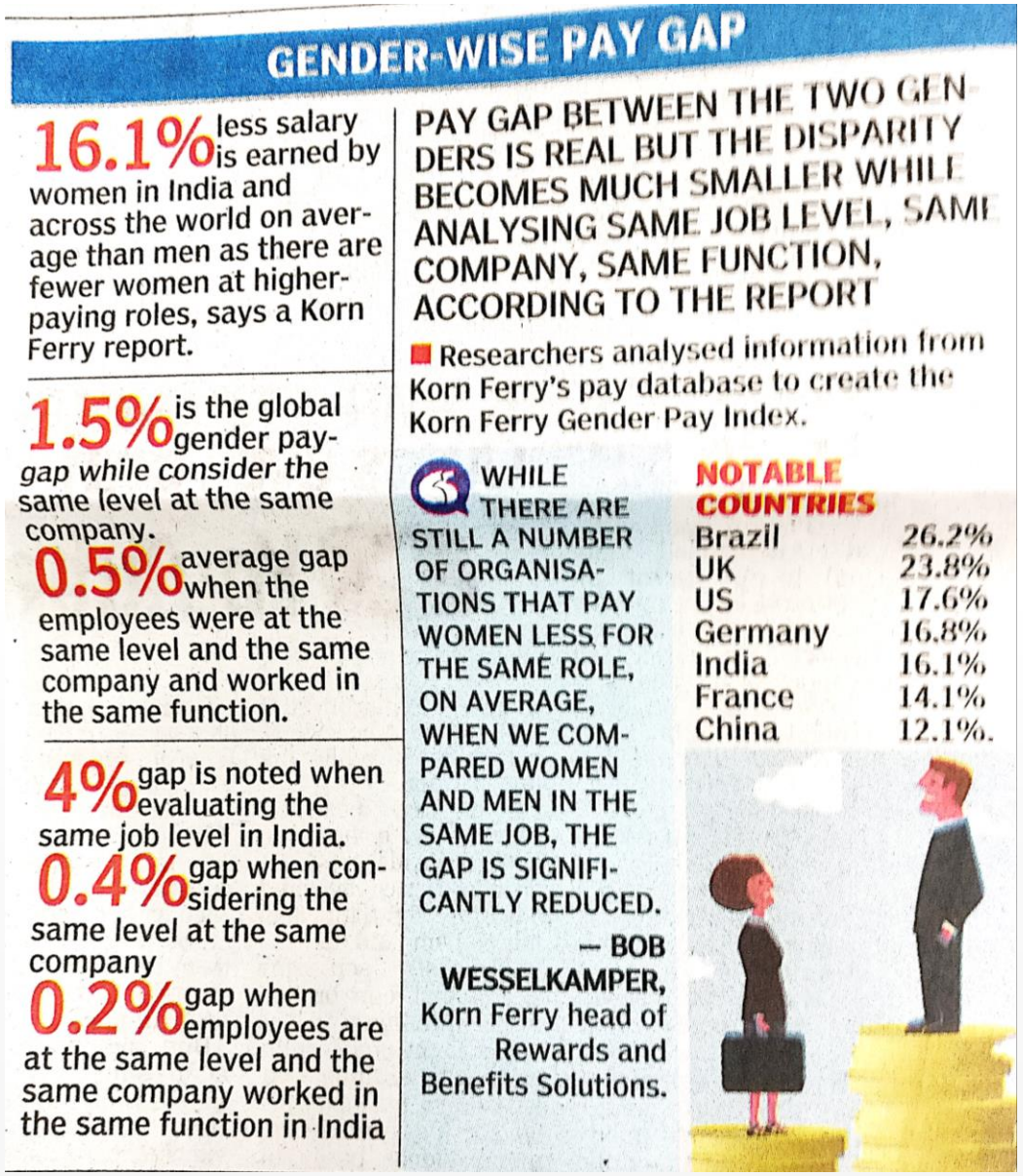
North	West	South
1	0	0
0	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	0

Original column in the dataset

New columns in the dataset REPLACING the original column (Region)

Indicator (Dummy) Variables - Excel

Consider the issue of gender discrimination in the salary earnings of workers in some industries. If there is discrimination, how much is one gender earning more than the other?



GENDER-WISE PAY GAP

16.1% less salary is earned by women in India and across the world on average than men as there are fewer women at higher-paying roles, says a Korn Ferry report.

1.5% is the global gender pay-gap while consider the same level at the same company.

0.5% average gap when the employees were at the same level and the same company and worked in the same function.

4% gap is noted when evaluating the same job level in India.

0.4% gap when considering the same level at the same company

0.2% gap when employees are at the same level and the same company worked in the same function in India

PAY GAP BETWEEN THE TWO GENDERS IS REAL BUT THE DISPARITY BECOMES MUCH SMALLER WHILE ANALYSING SAME JOB LEVEL, SAME COMPANY, SAME FUNCTION, ACCORDING TO THE REPORT

Researchers analysed information from Korn Ferry's pay database to create the Korn Ferry Gender Pay Index.

WHILE THERE ARE STILL A NUMBER OF ORGANISATIONS THAT PAY WOMEN LESS FOR THE SAME ROLE, ON AVERAGE, WHEN WE COMPARED WOMEN AND MEN IN THE SAME JOB, THE GAP IS SIGNIFICANTLY REDUCED.

— BOB WESSELKAMPER, Korn Ferry head of Rewards and Benefits Solutions.

NOTABLE COUNTRIES

Brazil	26.2%
UK	23.8%
US	17.6%
Germany	16.8%
India	16.1%
France	14.1%
China	12.1%.

Indicator (Dummy) Variables - Excel

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.933293402
R Square	0.871036574
Adjusted R Square	0.869727301
Standard Error	95.63590092
Observations	200

What is the equation?

$$Salary = 1821.9 + 8.38 * Age + 467.6 * Gender$$

Separate equation for each gender (as it can take only 2 values)

$$Salary_{Male} = 1821.9 + 8.38 * Age + 467.6 * 1$$

$$Salary_{Female} = 1821.9 + 8.38 * Age + 467.6 * 0$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	12169646.51	6084823	665.2824	2.40412E-88
Residual	197	1801806.432	9146.226		
Total	199	13971452.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1821.901302	59.56542052	30.58656	9.46E-77	1704.433585	1939.369019
Age (years)	8.375445059	1.813578901	4.618186	6.98E-06	4.798924132	11.95196599
Gender (1=Male, 0=Female)	467.628629	14.32150552	32.6522	2E-81	439.3854882	495.8717697

Indicator (Dummy) Variables – Interpreting Coefficients and Significance

Choice of reference group is not important; end results remain the same.

What will be the salary of a fresher in the two cases below where Fresher is the reference group in the 1st case and Low experience is the reference group in the 2nd?

SUMMARY OUTPUT $Salary = 2176.9 + 232.3 * ExpLow + 213.2 * ExpMed$

Regression Statistics	
Multiple R	0.376964139
R Square	0.142101962
Adjusted R Square	0.133392337
Standard Error	246.6638526
Observations	200

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	1985370.871	992685.4	16.31551	2.78E-07
Residual	197	11986082.07	60843.06		
Total	199	13971452.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2176.871087	33.2601471	65.44983	1.3E-135	2111.279	2242.463
Exp-Low	232.2707845	44.4457411	5.22594	4.4E-07	144.6203	319.9213
Exp-Med	213.1740921	43.78901842	4.868209	2.3E-06	126.8187	299.5295

SUMMARY OUTPUT $Salary = 2409.1 - 232.3 * ExpFresher - 19.1 * ExpMed$

Regression Statistics	
Multiple R	0.376964139
R Square	0.142101962
Adjusted R Square	0.133392337
Standard Error	246.6638526
Observations	200

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	2	1985370.871	992685.4	16.31551	2.78E-07
Residual	197	11986082.07	60843.06		
Total	199	13971452.94			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2409.141871	29.48196935	81.71577	6.3E-154	2351.001	2467.283
Exp-Fresher	-232.2707845	44.4457411	-5.22594	4.4E-07	-319.921	-144.62
Exp-Med	-19.09669233	40.99301484	-0.46585	0.641836	-99.9382	61.74477

p-values here indicate if the level (or group) is significantly different from the reference level (or group).

What might you do if there is no significant difference as is the case between low and medium experience?

A possible action could be to combine Low and Medium groups into a single group

Indicator (Dummy) Variables – Interpreting Coefficients

Interpret the coefficients of the numeric and categorical variables below.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.948085877					
R Square	0.898866831					
Adjusted R Square	0.896792304					
Standard Error	85.12366002					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	4	12558475.63	3139619	433.2877	8.41E-96	
Residual	195	1412977.312	7246.037			
Total	199	13971452.94				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1631.967642	59.02259438	27.64988	2.15E-69	1515.563	1748.372
Age (years)	12.25039811	1.69958175	7.20789	1.22E-11	8.898476	15.60232
Gender (1=Male, 0=Female)	430.4373177	13.72116567	31.37032	3.96E-78	403.3764	457.4983
Exp-Low	114.7447865	16.66453858	6.885566	7.7E-11	81.87892	147.6107
Exp-Med	100.5836307	16.08109074	6.254777	2.47E-09	68.86844	132.2988

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.948085877					
R Square	0.898866831					
Adjusted R Square	0.896792304					
Standard Error	85.12366002					
Observations	200					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	4	12558475.63	3139619	433.2877	8.41E-96	
Residual	195	1412977.312	7246.037			
Total	199	13971452.94				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1746.712428	54.2327351	32.20771	5.32E-80	1639.754	1853.67
Age (years)	12.25039811	1.69958175	7.20789	1.22E-11	8.898476	15.60232
Gender (1=Male, 0=Female)	430.4373177	13.72116567	31.37032	3.96E-78	403.3764	457.4983
Exp-Fresher	-114.7447865	16.66453858	-6.88557	7.7E-11	-147.611	-81.8789
Exp-Med	-14.16115576	14.18998882	-0.99797	0.319532	-42.1467	13.8244

- Numeric: For unit change in Age (numeric), Salary increases by $12.25, all other variables (Age, Gender, Exp) being the same.
- Categorical (Dummy): If a person is a fresher, (s)he makes $114.7 less than a person with low experience, all other variables (Age, Gender, Exp) being the same.

Multiple Linear Regression

MODEL BUILDING METHODS

Model Building: Search Procedures

Suppose a model to predict the world crude oil production (barrels per day) is to be developed and the predictors used are:

- US energy consumption (BTUs)
- Gross US nuclear electricity generation (kWh)
- US coal production (short-tons)
- Total US dry gas (natural gas) production (cubic feet)
- Fuel rate of US-owned automobiles (miles per gallon)

What does your intuition say about how each of these variables would affect the oil production?

CrudeOilOutput

WorldOil	USEnergy	USAutoFuelRate	USNuclear	USCoal	USDryGas
55.7	74.3	13.4	83.5	598.6	21.7
55.7	72.5	13.6	114	610	20.7
52.8	70.5	14	172.5	654.6	19.2
57.3	74.4	13.8	191.1	684.9	19.1
59.7	76.3	14.1	250.9	697.2	19.2
60.2	78.1	14.3	276.4	670.2	19.1
62.7	78.9	14.6	255.2	781.1	19.7
59.6	76	16	251.1	829.7	19.4
56.1	74	16.5	272.7	823.8	19.2
53.5	70.8	16.9	282.8	838.1	17.8
53.3	70.5	17.1	293.7	782.1	16.1
54.5	74.1	17.4	327.6	895.9	17.5
54	74	17.5	383.7	883.6	16.5
56.2	74.3	17.4	414	890.3	16.1
56.7	76.9	18	455.3	918.8	16.6
58.7	80.2	18.8	527	950.3	17.1
59.9	81.4	19	529.4	980.7	17.3
60.6	81.3	20.3	576.9	1029.1	17.8
60.2	81.1	21.2	612.6	996	17.7
60.2	82.2	21	618.8	997.5	17.8
60.2	83.9	20.6	610.3	945.4	18.1
61	85.6	20.8	640.4	1033.5	18.8
62.3	87.2	21.1	673.4	1033	18.6
64.1	90	21.2	674.7	1063.9	18.8
66.3	90.6	21.5	628.6	1089.9	18.9
67	89.7	21.6	666.8	1109.8	18.9

Model Building: Search Procedures

Two considerations in model building:

- Explaining most variation in dependent variable (high R^2 or Adjusted R^2)
- Keeping the model simple AND economical (as few variables as possible)

Quite often, the above two considerations are in conflict of each other.

If 3 variables can explain the variation nearly as well as 5 variables, the simpler model is better. Search procedures help choose the more attractive model.

Search Procedures: All Possible Regressions

All variables used in all combinations. For a dataset containing *k* independent variables, 2^k-1 models are examined. In the example of the oil production, 31 models are examined.

Tedious, Time-Consuming, Inefficient, Overwhelming.

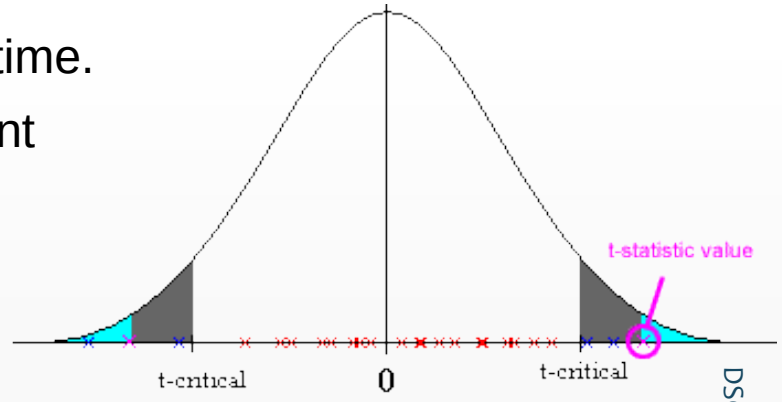
Search Procedures: Stepwise Regression

Forward, Backward and Bi-directional

Starts a model with a single predictor (or all predictors, called the full model) and then adds (or deletes) predictors one step at a time.

Also, the significance level used for adding or removing variables is higher than the typical 5% **(usually 15%)** to allow variables to enter the model easily and make it difficult for them to leave the model.

- Step 1
  - Simple regression model for each of the independent variables one at a time.
  - Model with **smallest p-value** selected and the corresponding independent variable considered the best single predictor, denoted x_1.
  - If no variable is significant, the search stops with no model.



Search Procedures: Stepwise Regression

- Step 2
  - All possible two-predictor regression models with x_1 as one variable.
  - Model with smallest p-value in conjunction with x_1 and one of the other $k - 1$ variables denoted x_2.
  - Occasionally, if x_1 becomes insignificant, it is dropped from the model and search continued with x_2.
  - If no other variables are significant, procedure stops.

- The above process continues with the 3rd variable added to the above 2 selected and so on.

Search Procedures: Stepwise Regression - Excel

Step 1

Dependent Variable, *y*	Independent Variable	*p*-value
Oil production	Energy consumption	1.86e-11
Oil production	Nuclear	0.000176
Oil production	Coal	0.000662
Oil production	Dry gas	0.292870
Oil production	Fuel rate	0.00169

$$y = 13.075 + 0.580x_1$$

Energy Consumption has the smallest p-value at 1.86 x 10^{-11}. This is the first variable entering the model.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9232519
R Square	0.85239407
Adjusted R Square	0.84624382
Standard Error	1.51547535
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	318.3066	318.3066	138.5951	1.86E-11
Residual	24	55.11997	2.296666		
Total	25	373.4265			

	Coefficients	*andard Errc*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13.0749304	3.894389	3.357377	0.002618	5.037307	21.11255
US Energy Consumption (	0.58012095	0.049277	11.77264	1.86E-11	0.478418	0.681824

Search Procedures: Stepwise Regression - Excel

Step 2

Dependent Variable, y	Independent Variable, x_1	Independent Variable, x_2	p-value
Oil production	Energy consumption	Nuclear	0.00152
Oil production	Energy consumption	Coal	0.0227
Oil production	Energy consumption	Dry gas	0.0357
Oil production	Energy consumption	Fuel rate	0.00106

$$y = 7.14 + 0.772x_1 - 0.517x_2$$

Energy Consumption is still significant with $p = 2.55 \times 10^{-11}$.

Step 3

Dependent Variable, y	Independent Variable, x_1	Independent Variable, x_2	Independent Variable, x_3	p-value
Oil production	Energy consumption	Fuel rate	Nuclear	0.672
Oil production	Energy consumption	Fuel rate	Coal	0.102
Oil production	Energy consumption	Fuel rate	Dry gas	0.650

Note significance level being used is 0.15 and not 0.05.

$$y = 8.45 + 0.754x_1 - 1.028x_2 + 0.01\ x_3$$

Energy Consumption and *Fuel rate* are still significant with $p = 4.4 \times 10^{-11}$ and 0.005, respectively.

Search Procedures: Stepwise Regression - Excel

Step 4

Dependent Variable, y	Independent Variable, x_1	Independent Variable, x_2	Independent Variable, x_3	Independent Variable, x_4	p-value
Oil production	Energy consumption	Fuel rate	Coal	Nuclear	0.528
Oil production	Energy consumption	Fuel rate	Coal	Dry gas	0.813

Neither variable is significant at $\alpha = 0.15$, i.e., neither p-value < 0.15. Process stops.

Search Procedures: Stepwise Regression - R

- R implementation:
  - stepAIC(CrudeOilOutputlm) {MASS} package
  - step(CrudeOilOutputlm) (stats} package

  where CrudeOilOutputlm is the object of class *lm*.

AIC (Akaike's Information Criterion)

- AIC is *similar to Adjusted R^2* in the sense it penalizes for adding more parameters to the model.
- $AIC = 2k + nln\left(\frac{RSS}{n}\right)$ where RSS is Residual Sum of Squares or SSE (Sum of Squared Errors).

  k is the number of parameters including intercept.
- AIC provides a means for model selection by comparing models **built on the same data**. Lower the AIC, better the model.

```
> stepAICOil <- stepAIC(CrudeOilOutputlm, direction = "both")
Start:  AIC=15.29
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal + CrudeOilOutput$USDryGas

                                Df Sum of Sq     RSS     AIC
- CrudeOilOutput$USDryGas        1     0.151  29.661  13.425
- CrudeOilOutput$USNuclear       1     0.651  30.161  13.860
<none>                                        29.510  15.293
- CrudeOilOutput$USAutoFuelRate  1     2.640  32.150  15.521
- CrudeOilOutput$USCoal          1     2.683  32.193  15.555
- CrudeOilOutput$USEnergy        1    31.720  61.231  32.270

Step:  AIC=13.42
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal

                                Df Sum of Sq     RSS     AIC
- CrudeOilOutput$USNuclear       1     0.583  30.243  11.931
<none>                                        29.661  13.425
- CrudeOilOutput$USCoal          1     4.296  33.956  14.941
- CrudeOilOutput$USAutoFuelRate  1     4.575  34.236  15.154
+ CrudeOilOutput$USDryGas        1     0.151  29.510  15.293
- CrudeOilOutput$USEnergy        1   137.158 166.818  56.329

Step:  AIC=11.93
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USCoal

                                Df Sum of Sq     RSS     AIC
<none>                                        30.243  11.931
- CrudeOilOutput$USCoal          1     3.997  34.240  13.158
+ CrudeOilOutput$USNuclear       1     0.583  29.661  13.425
+ CrudeOilOutput$USDryGas        1     0.082  30.161  13.860
- CrudeOilOutput$USAutoFuelRate  1    13.531  43.774  19.545
- CrudeOilOutput$USEnergy        1   195.845 226.088  62.234
```

Multiple Linear Regression

HANDLING MULTICOLLINEARITY

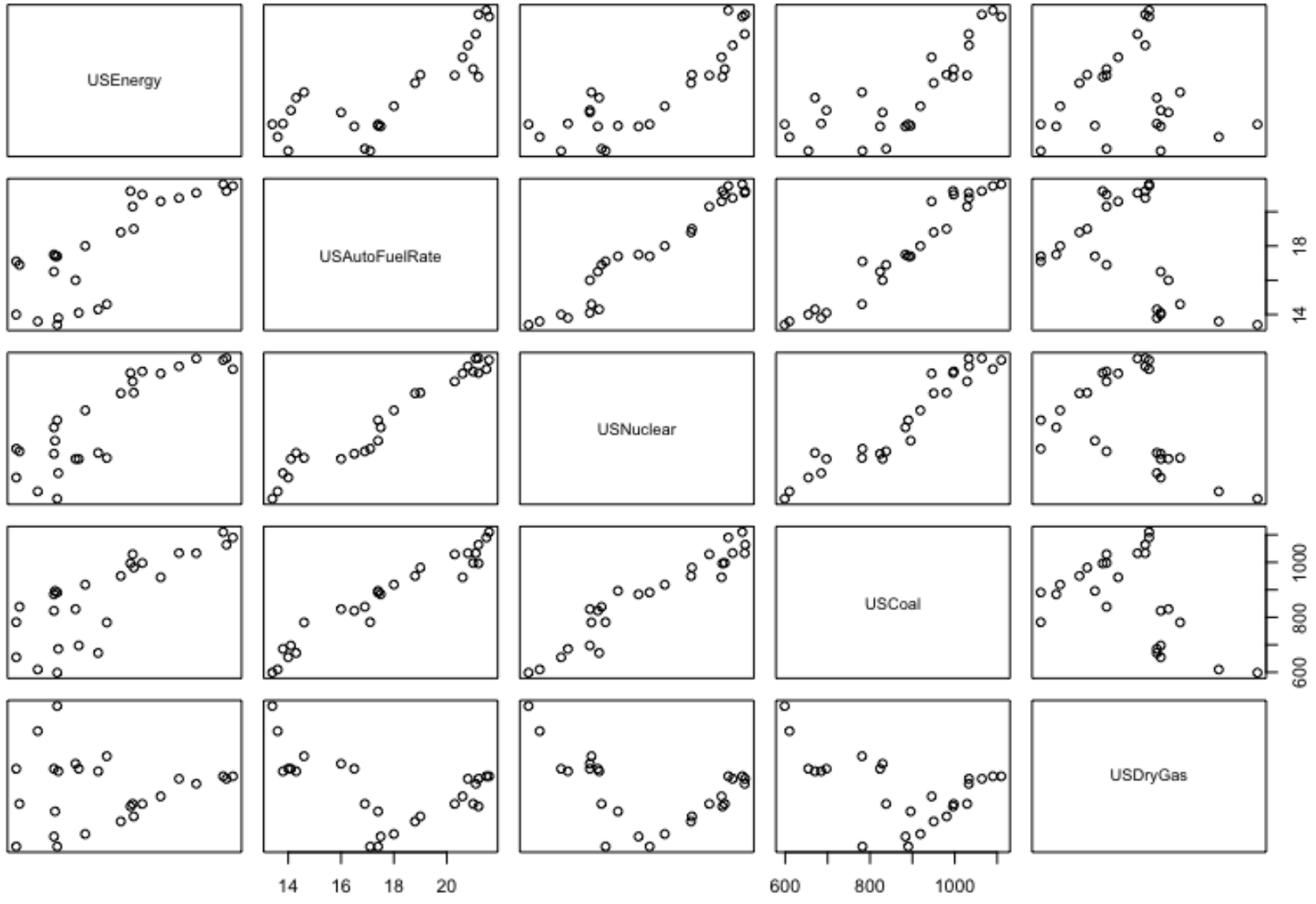
Multicollinearity - R

Two or more **independent variables** are highly correlated.

	Energy consumption	Nuclear	Coal	Dry gas	Fuel rate
Energy consumption	1				
Nuclear	0.856	1			
Coal	0.791	0.952	1		
Dry gas	0.057	-0.404	-0.448	1	
Fuel rate	0.791	0.972	0.968	-0.423	1

R Code: (correlation <- cor(CrudeOilOutput))

Multicollinearity - R



R Code

```
> CrudeOilOutput <-
read.csv("CrudeOilOutput.csv",
header = T, sep = ",")
> plot(CrudeOilOutput)
```

DSC 7402

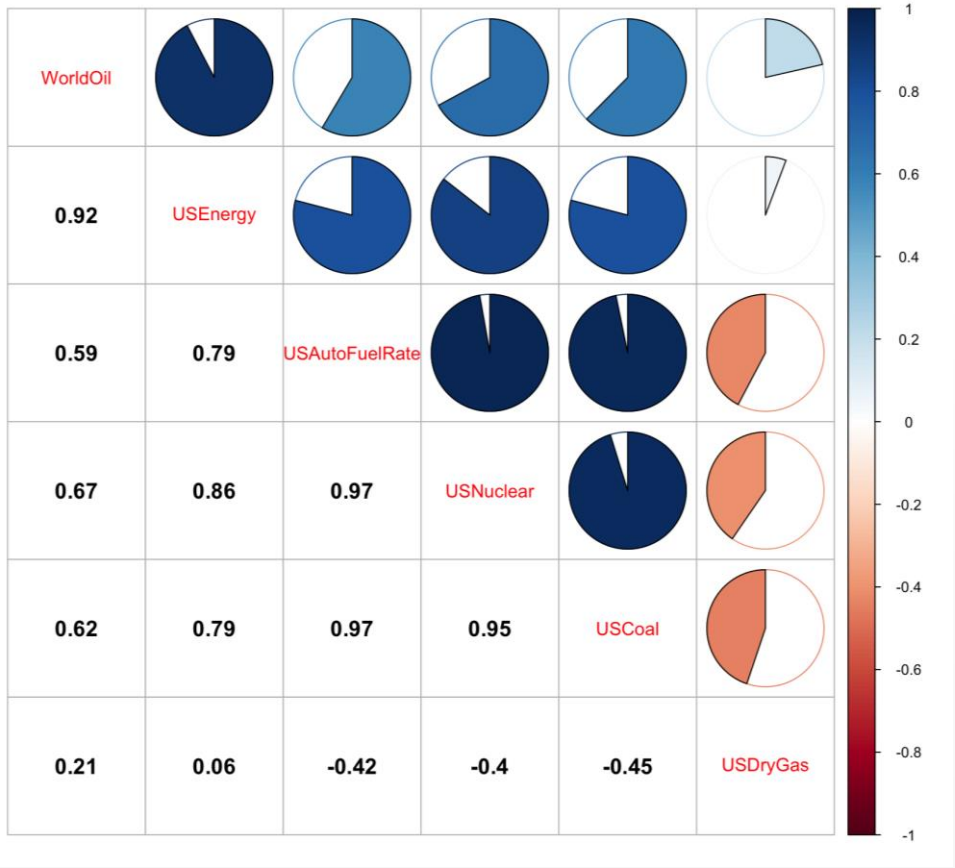
Multicollinearity - R

R Code

> correlation <- cor(CrudeOilOutput)
> corrplot.mixed(correlation,lower.col = "black",number.cex=1.2, upper = "circle")

R Code

> correlation <- cor(CrudeOilOutput)
> corrplot.mixed(correlation,lower.col = "black",number.cex=1.2, upper = "pie")



DSC 7402

Multicollinearity

Sign of estimated regression coefficient when interacting may be opposite of the signs when used as individual predictors.

For example, fuel rate and coal production are highly correlated (0.968).

$$\hat{y} = 44.869 + 0.7838(fuel\ rate)$$

$$\hat{y} = 45.072 + 0.0157(coal)$$

$$\hat{y} = 45.806 + 0.0277(coal) - 0.3934(fuel\ rate)$$

Multicollinearity

Multicollinearity can lead to a model where the model (*F* value) is significant but all individual predictors (*t* values) are insignificant.

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Multicollinearity

- **Variance Inflation Factor (VIF):** Measures how much the variance (or standard error in estimating the coefficient of a variable) gets inflated in the presence of correlated variables *compared to the situation* where they are completely uncorrelated.

- A regression analysis is conducted to predict an independent variable by the other independent variables. The independent variable being predicted becomes the dependent variable in this analysis.

$$VIF = \frac{1}{1 - {R_i}^2}$$

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Multicollinearity

CrudeOilOutput

WorldOil	USEnergy	USAutoFuelRate	USNuclear	USCoal	USDryGas
55.7	74.3	13.4	83.5	598.6	21.7
55.7	72.5	13.6	114	610	20.7
52.8	70.5	14	172.5	654.6	19.2
57.3	74.4	13.8	191.1	684.9	19.1
59.7	76.3	14.1	250.9	697.2	19.2
60.2	78.1	14.3	276.4	670.2	19.1
62.7	78.9	14.6	255.2	781.1	19.7
59.6	76	16	251.1	829.7	19.4
56.1	74	16.5	272.7	823.8	19.2
53.5	70.8	16.9	282.8	838.1	17.8
53.3	70.5	17.1	293.7	782.1	16.1
54.5	74.1	17.4	327.6	895.9	17.5
54	74	17.5	383.7	883.6	16.5
56.2	74.3	17.4	414	890.3	16.1
56.7	76.9	18	455.3	918.8	16.6
58.7	80.2	18.8	527	950.3	17.1
59.9	81.4	19	529.4	980.7	17.3
60.6	81.3	20.3	576.9	1029.1	17.8
60.2	81.1	21.2	612.6	996	17.7
60.2	82.2	21	618.8	997.5	17.8
60.2	83.9	20.6	610.3	945.4	18.1
61	85.6	20.8	640.4	1033.5	18.8
62.3	87.2	21.1	673.4	1033	18.6
64.1	90	21.2	674.7	1063.9	18.8
66.3	90.6	21.5	628.6	1089.9	18.9
67	89.7	21.6	666.8	1109.8	18.9

If we wish to understand the multicollinearity of USAutoFuelRate, we wish to understand *what % of variation in USAutoFuelRate can be explained by other independent variables*.

Recall this is what R^2 tells us, and so, we take USAutoFuelRate as the dependent variable (*for the purpose of this calculation and understanding ONLY*) and build a regression with other **independent** variables.

Multicollinearity

CrudeOilOutput

WorldOil	USEnergy	USAutoFuelRate	USNuclear	USCoal	USDryGas
55.7	74.3	13.4	83.5	598.6	21.7
55.7	72.5	13.6	114	610	20.7
52.8	70.5	14	172.5	654.6	19.2
57.3	74.4	13.8	191.1	684.9	19.1
59.7	76.3	14.1	250.9	697.2	19.2
60.2	78.1	14.3	276.4	670.2	19.1
62.7	78.9	14.6	255.2	781.1	19.7
59.6	76	16	251.1	829.7	19.4
56.1	74	16.5	272.7	823.8	19.2
53.5	70.8	16.9	282.8	838.1	17.8
53.3	70.5	17.1	293.7	782.1	16.1
54.5	74.1	17.4	327.6	895.9	17.5
54	74	17.5	383.7	883.6	16.5
56.2	74.3	17.4	414	890.3	16.1
56.7	76.9	18	455.3	918.8	16.6
58.7	80.2	18.8	527	950.3	17.1
59.9	81.4	19	529.4	980.7	17.3
60.6	81.3	20.3	576.9	1029.1	17.8
60.2	81.1	21.2	612.6	996	17.7
60.2	82.2	21	618.8	997.5	17.8
60.2	83.9	20.6	610.3	945.4	18.1
61	85.6	20.8	640.4	1033.5	18.8
62.3	87.2	21.1	673.4	1033	18.6
64.1	90	21.2	674.7	1063.9	18.8
66.3	90.6	21.5	628.6	1089.9	18.9
67	89.7	21.6	666.8	1109.8	18.9

```
Call:
lm(formula = USAutoFuelRate ~ USEnergy + USNuclear + USCoal +
    USDryGas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59522 -0.34986 -0.03581  0.16996  1.26953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.319789   2.538800   4.459 0.000217 ***
USEnergy    -0.192304   0.058091  -3.310 0.003328 **
USNuclear    0.014051   0.002439   5.760 1.02e-05 ***
USCoal       0.009076   0.002114   4.292 0.000323 ***
USDryGas     0.425035   0.152267   2.791 0.010939 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.483 on 21 degrees of freedom
Multiple R-squared:  0.9765,    Adjusted R-squared:  0.972
F-statistic: 217.8 on 4 and 21 DF,  p-value: < 2.2e-16
```


Multicollinearity

CrudeOilOutput

WorldOil	USEnergy	USAutoFuelRate	USNuclear	USCoal	USDryGas
55.7	74.3	13.4	83.5	598.6	21.7
55.7	72.5	13.6	114	610	20.7
52.8	70.5	14	172.5	654.6	19.2
57.3	74.4	13.8	191.1	684.9	19.1
59.7	76.3	14.1	250.9	697.2	19.2
60.2	78.1	14.3	276.4	670.2	19.1
62.7	78.9	14.6	255.2	781.1	19.7
59.6	76	16	251.1	829.7	19.4
56.1	74	16.5	272.7	823.8	19.2
53.5	70.8	16.9	282.8	838.1	17.8
53.3	70.5	17.1	293.7	782.1	16.1
54.5	74.1	17.4	327.6	895.9	17.5
54	74	17.5	383.7	883.6	16.5
56.2	74.3	17.4	414	890.3	16.1
56.7	76.9	18	455.3	918.8	16.6
58.7	80.2	18.8	527	950.3	17.1
59.9	81.4	19	529.4	980.7	17.3
60.6	81.3	20.3	576.9	1029.1	17.8
60.2	81.1	21.2	612.6	996	17.7
60.2	82.2	21	618.8	997.5	17.8
60.2	83.9	20.6	610.3	945.4	18.1
61	85.6	20.8	640.4	1033.5	18.8
62.3	87.2	21.1	673.4	1033	18.6
64.1	90	21.2	674.7	1063.9	18.8
66.3	90.6	21.5	628.6	1089.9	18.9
67	89.7	21.6	666.8	1109.8	18.9

```
> (rsq_AFR <- summary(AFRlm)$r.square)
[1] 0.9764673
> (vif_AFR <- 1/(1 - (rsq_AFR)))
[1] 42.49405
```

$$VIF = \frac{1}{1 - R_i^2}$$

```
> (vif(CrudeOilOutputlm))
      USEnergy USAutoFuelRate       USNuclear           USCoal
     20.822780      42.494049       61.549771       21.346305
      USDryGas
      6.188241
```

DSC 7402

Multicollinearity - VIF 

- VIF > 4 (R_i^2>0.75), 5 (R_i^2>0.80) and 10 (R_i^2>0.90) are commonly used as rules of thumb to indicate severe multicollinearity. VIF of 3 and 6 are also used.

- In practical situations, sometimes even 1.5 is considered as large VIF. ☹

- Remove such variables, rebuild models and compare with earlier model. Make decision based on whether **accuracy of prediction** is more important to the business or **interpretation of the model and the coefficients**.

- Multicollinearity is not a problem if predictions are only of interest. However, if model interpretation is of interest, coefficients can be misinterpreted as they really are not independent.

Multicollinearity

- Regression coefficients cannot be estimated precisely due to high standard errors.
- Value and sign of the coefficients may change when different samples are selected from the data.
- When standard errors are high, confidence intervals get wider and equivalently t-statistic becomes smaller. A small t-statistic leads to large p-values such that Null hypothesis doesn't get rejected.
- Thus, multicollinearity can erroneously indicate that the correlated variables are insignificant when one (or more) of them actually may be significant.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.826018723	8.015718426	0.477314511	0.638068311	-12.84358	20.4956177
US Energy Consumption (quadrillion BTUs generation per year)	0.784304704	0.079589474	9.854377252	2.50331E-09	0.61878933	0.94982008
US Fuel Rate for Automobiles (miles per gallon)	-0.825327796	0.458583006	-1.799734803	0.086287926	-1.7790034	0.12834777
US Coal Gross Production (million short-tons)	0.010932943	0.006269174	1.743920754	0.095795081	-0.0021045	0.0239704
US Nuclear Electricity (billion kilowatt-hours)	-0.00426081	0.006634459	-0.642224138	0.527677133	-0.0180579	0.0095363

Multiple Linear Regression

EVALUATING MODEL PERFORMANCE

Appropriate Error Measures for Numeric Data

- MAE (Mean Absolute Error): Mean of the absolute value of the difference between the predicted and actual values. Useful if you do not wish some outliers to skew the results.

$$\frac{\sum|y_i - \hat{y}_i|}{n}$$

- MAPE (Mean Absolute Percentage Error): Same as above but converted into percentages to allow for comparison across different scales (e.g., comparing accuracies of forecasts on BSE vs NSE) or when dealing with large numbers.

$$\frac{1}{n}\left(\sum \frac{|y_i - \hat{y}_i|}{y_i}\right) * 100$$

- MSE (Mean Square Error) or RMSE (Root Mean Square Error): Accounts for infrequent large errors, whose impact may be understated by the absolute error measures.

$$\frac{\sum(y_i - \hat{y}_i)^2}{n} \qquad \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

Putting It All Together

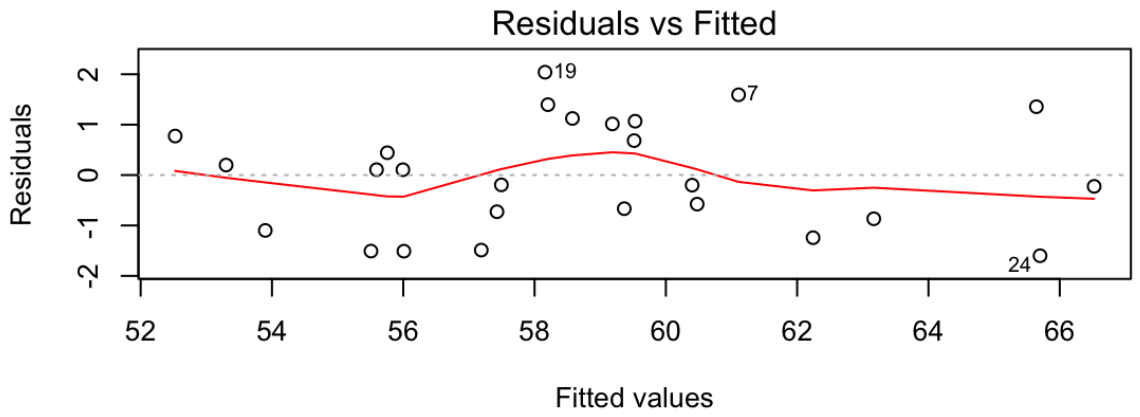
CASES

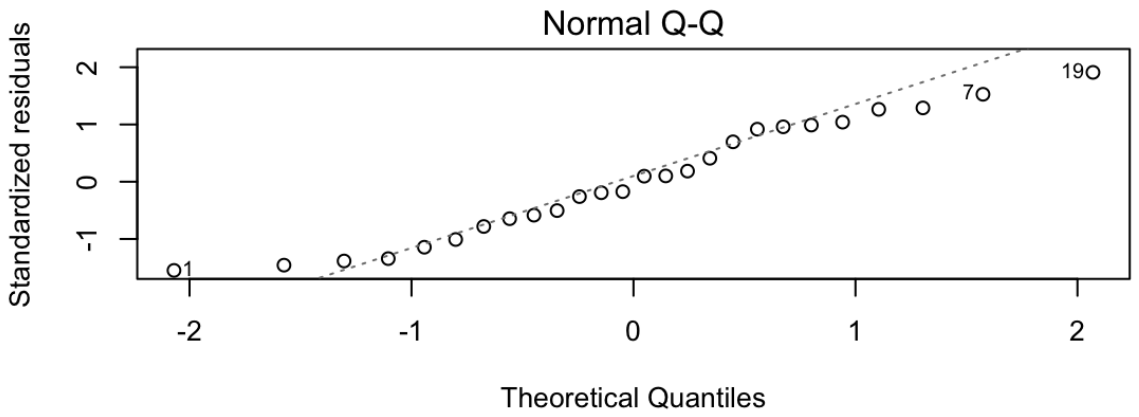
1: Predicting World Crude Oil Production

```
Call:
lm(formula = WorldOil ~ USEnergy + USAutoFuelRate + USNuclear +
    USCoal + USDryGas, data = CrudeOilOutput)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60006 -0.83219 -0.04438  0.95456  2.04060

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.708474   8.908760   0.304 0.764250
USEnergy        0.835670   0.180234   4.637 0.000159 ***
USAutoFuelRate -0.734144   0.548823  -1.338 0.196018
USNuclear      -0.006544   0.009854  -0.664 0.514197
USCoal          0.009825   0.007286   1.348 0.192596
USDryGas       -0.143211   0.448408  -0.319 0.752753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.215 on 20 degrees of freedom
Multiple R-squared:  0.921,     Adjusted R-squared:  0.9012
F-statistic: 46.62 on 5 and 20 DF,  p-value: 2.41e-10
```

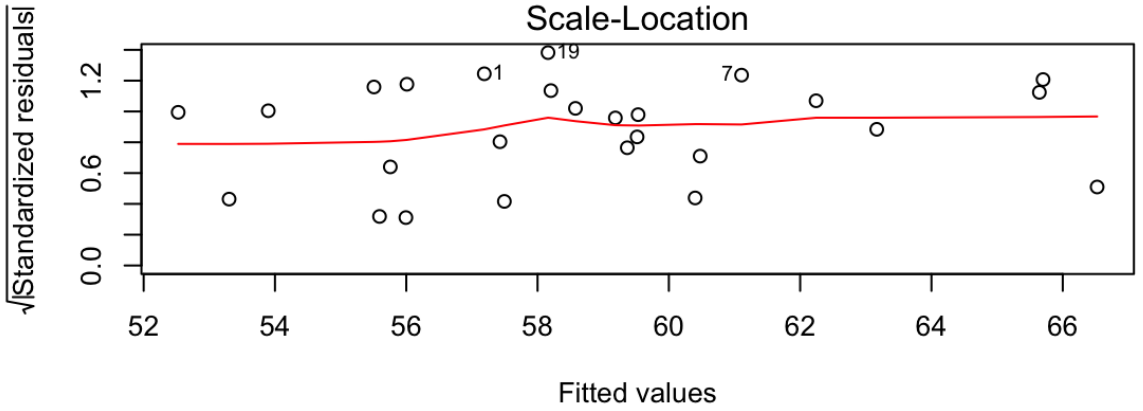
- Check #1: Strong Adj. R2.
- Check #2: Energy Consumption appears to be the only significant variable
- Check #3: Model is significant

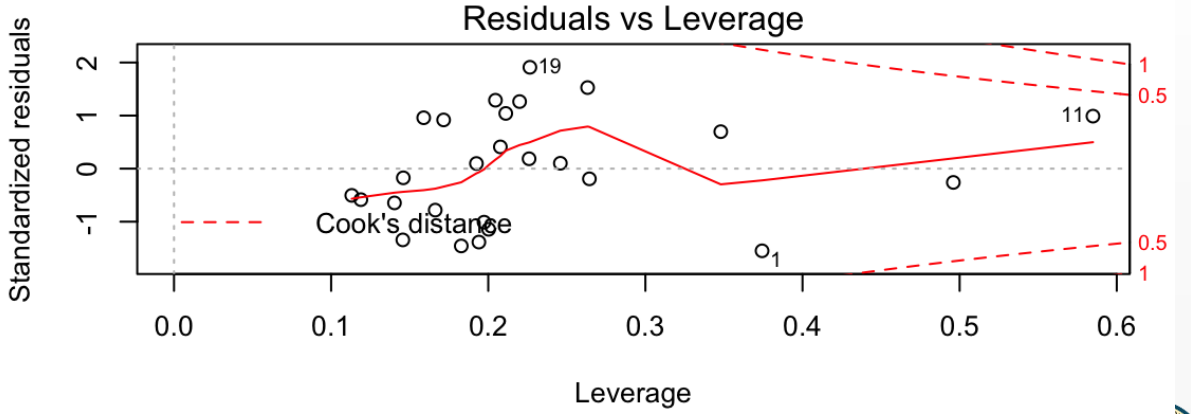
1: Predicting World Crude Oil Production





Check #4: Residuals Analysis (everything looks good)

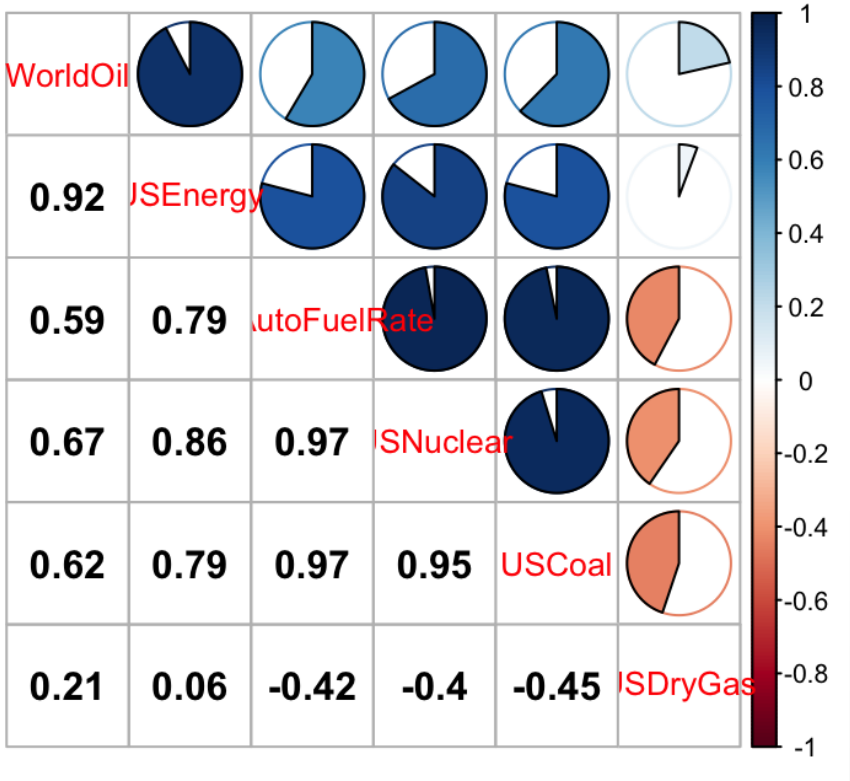




1: Predicting World Crude Oil Production

Check #5: Check for Multicollinearity

```
> (vif(CrudeOilOutputlm))
      USEnergy USAutoFuelRate       USNuclear         USCoal
     20.822780      42.494049      61.549771      21.346305
      USDryGas
      6.188241
```



- VIF values and correlation plots indicate strong multicollinearity
- Energy Consumption appears to be the only significant variable (as seen in the last slide) but multicollinearity may be causing issues
- Systematically remove highest VIF variable **<u>one at a time</u>**, or run Stepwise Regression

1: Predicting World Crude Oil Production

```
> stepAICOil <- stepAIC(CrudeOilOutputlm, direction = "both")
Start:  AIC=15.29
WorldOil ~ USEnergy + USAutoFuelRate + USNuclear + USCoal + USDryGas

                 Df Sum of Sq    RSS    AIC
- USDryGas        1     0.151 29.661 13.425
- USNuclear       1     0.651 30.161 13.860
<none>                       29.510 15.293
- USAutoFuelRate  1     2.640 32.150 15.521
- USCoal          1     2.683 32.193 15.555
- USEnergy        1    31.720 61.231 32.270

Step:  AIC=13.42
WorldOil ~ USEnergy + USAutoFuelRate + USNuclear + USCoal

                 Df Sum of Sq     RSS    AIC
- USNuclear       1     0.583  30.243 11.931
<none>                       29.661 13.425
```

- Stepwise Regression indicates USEnergy, USAutoFuelRate and USCoal are significant.
- Build the model with these 3 variables and check VIF.

```
> stepAICOil$call
lm(formula = WorldOil ~ USEnergy + USAutoFuelRate + USCoal, data = CrudeOil
Output)
```

DSC 7402

1: Predicting World Crude Oil Production

```
Call:
lm(formula = WorldOil ~ USEnergy + USAutoFuelRate + USCoal, data = CrudeOil
Output)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5570 -0.9813  0.1836  0.9138  1.9647

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.454146   3.462723   2.441  0.02313 *
USEnergy       0.753943   0.063166  11.936 4.41e-11 ***
USAutoFuelRate -1.028331   0.327772  -3.137  0.00479 **
USCoal         0.010479   0.006145   1.705  0.10225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.172 on 22 degrees of freedom
Multiple R-squared:  0.919,     Adjusted R-squared:  0.908
F-statistic: 83.21 on 3 and 22 DF,  p-value: 3.659e-12
```

```
> (vif(CrudeOilOutputlm1))
    USEnergy USAutoFuelRate          USCoal
    2.745189      16.268335       16.298435
```

- Model indicates USCoal is not significant.
- VIF also indicates strong multicollinearity.
- Remove USCoal and rebuild the model.

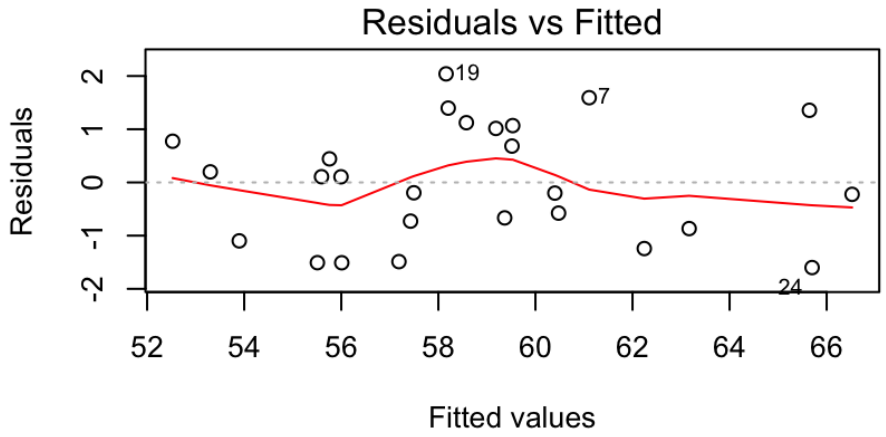
1: Predicting World Crude Oil Production

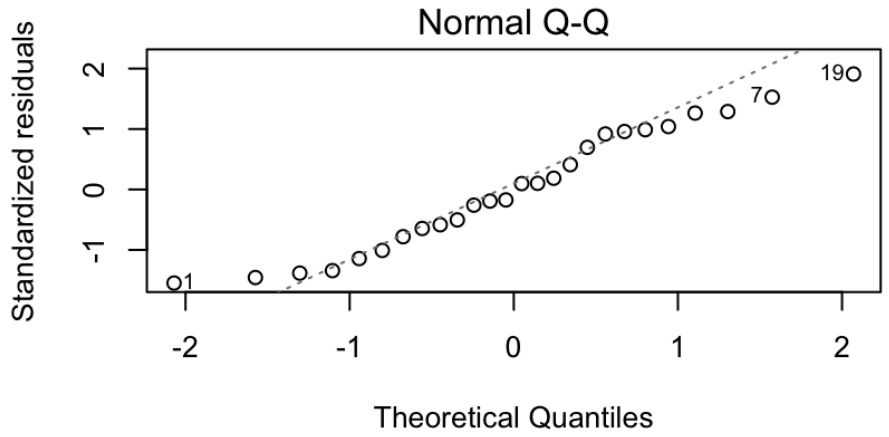
```
Call:
lm(formula = WorldOil ~ USEnergy + USAutoFuelRate, data = CrudeOilOutput)

Residuals:
     Min       1Q   Median       3Q      Max
-1.86869 -1.00279  0.01214  0.67016  2.20084

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.1403     3.5131   2.032  0.05381 .
USEnergy         0.7720     0.0648  11.913 2.55e-11 ***
USAutoFuelRate  -0.5173     0.1381  -3.745  0.00106 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.22 on 23 degrees of freedom
Multiple R-squared:  0.9083,    Adjusted R-squared:  0.9003
F-statistic: 113.9 on 2 and 23 DF,  p-value: 1.166e-12

> (vif(CrudeOilOutputlm2))
     USEnergy USAutoFuelRate
     2.667949       2.667949
```

- Multicollinearity resolved.
- Adj. R2 still good (Check #1).
- Significant variables (Check #2).
- Significant model (Check #3).

DSC 7402

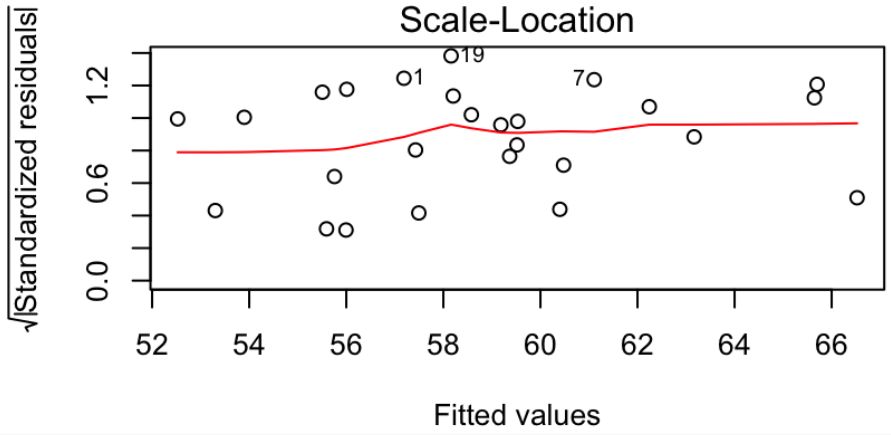
1: Predicting World Crude Oil Production

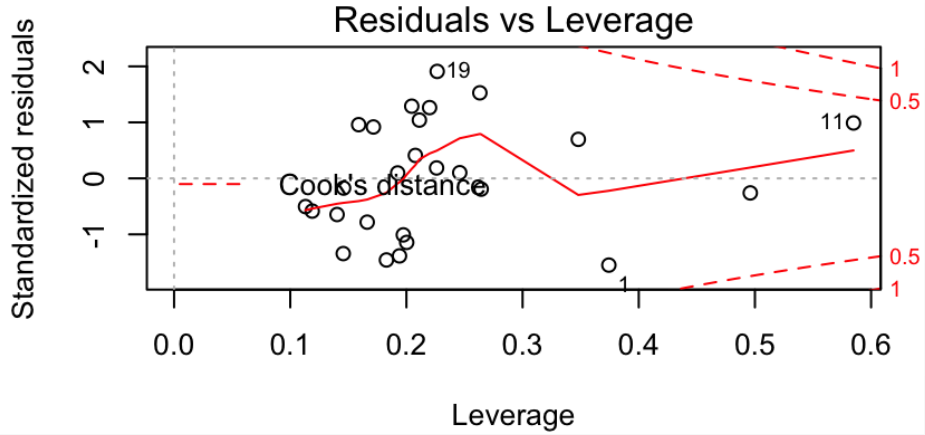




Check #4: Residuals Analysis (everything looks good)

We have a good model.





DSC 7402

2: Car Mileage Prediction in Road Tests

mtcars dataset in R

Data was extracted from the *Motor Trend* US magazine with a goal to predicting the fuel consumption (mpg) using 10 variables dealing with automobile design and performance.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

2: Car Mileage Prediction in Road Tests

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

- Very good Adjusted R2
- No significant variable at 5% significance level
- Model is significant
- Indicates multicollinearity

```
> vif(mtcarslm)
      cyl      disp        hp      drat        wt      qsec
15.373833 21.620241  9.832037  3.374620 15.164887  7.527958
       vs        am      gear      carb
 4.965873  4.648487  5.357452  7.908747
```

- Rules of thumb indicate almost everything is highly collinear
- Let's run Stepwise Regression

2: Car Mileage Prediction in Road Tests

```
> mtcarsStepAIC <- stepAIC(mtcarslm)
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- vs    1    0.2685 147.84 66.973
- carb  1    0.5201 148.09 67.028
- gear  1    1.8211 149.40 67.308
- drat  1    1.9826 149.56 67.342
- disp  1    3.9009 151.47 67.750
- hp    1    7.3632 154.94 68.473
<none>              147.57 68.915
- qsec  1   10.0933 157.67 69.032
- am    1   11.8359 159.41 69.384
- wt    1   27.0280 174.60 72.297

Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq    RSS    AIC
- carb  1    0.6855 148.53 65.121
- gear  1    2.1437 149.99 65.434
```

```
> mtcarsStepAIC

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```

- Stepwise Regression identified 3 variables as significant
- Let us build the model with these 3

2: Car Mileage Prediction in Road Tests

```
Call:
lm(formula = mpg ~ am + qsec + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
am            2.9358     1.4109   2.081 0.046716 *
qsec          1.2259     0.2887   4.247 0.000216 ***
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
> vif(mtcarslm2)
      am     qsec       wt
2.541437 1.364339 2.482952
```

- Adjusted R^2 improved
- All variables are significant
- Model is significant
- VIF values are around 2.5 or less
- Please note when we used data transformation and used interaction term, the Adj R^2 was even higher at 87.1%.

DECISION-MAKING: If accuracy of predictions is more important, use the earlier model. If explicability of the model is more important, use the current one.

3: Predicting Fungal Toxin Contamination

A drug precursor molecule is extracted from a type of nut, which is commonly contaminated by a fungal toxin that is difficult to remove during the purification process. The suspected predictors of the amount of fungus are:

- Rainfall (cm/week)
- Noon temperature (°C)
- Sunshine (h/day)
- Wind speed (km/h)

The fungal toxin concentration is measured in µg/100 g.

FungalToxinContamination

Toxin	Rain	NoonTemp	Sunshine	WindSpeed
18.1	1.3	20.9	6.23	13.3
28.6	2.28	25.4	8.13	10.8
15.9	1.11	28.2	10.21	10.9
19.2	0.74	23.7	6.96	8.2
19.3	1.32	26.5	9.04	9.8
14.8	0.51	23.9	7.84	12.3
21.7	1.56	26.7	6.69	10
16.5	1.32	30	8.3	12.2
23.8	2.05	24.9	9.22	10.7
19	1.37	22	8.37	15

3: Predicting Fungal Toxin Contamination

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$Sunshine + ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
      1       2       3       4       5       6       7       8
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813
      9      10
-1.1552  0.8429

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         31.6084     7.1051   4.449  0.00671 **
ToxinConc$Rain       7.0676     1.0031   7.046  0.00089 ***
ToxinConc$NoonTemp  -0.4201     0.2413  -1.741  0.14215
ToxinConc$Sunshine  -0.2375     0.5086  -0.467  0.66018
ToxinConc$WindSpeed -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

Multiple regression tends to remove correlated pairs of IVs, as is *possibly* the case with Noon Temperature and Sunshine here.

3: Predicting Fungal Toxin Contamination

```
> correlation
               Toxin         Rain    NoonTemp    Sunshine    WindSpeed
Toxin      1.00000000  0.868734134 -0.07319548 -0.05169949 -0.270555628
Rain       0.86873413  1.000000000  0.11691043  0.16841144 -0.002180167
NoonTemp  -0.07319548  0.116910426  1.00000000  0.50082303 -0.368972511
Sunshine  -0.05169949  0.168411437  0.50082303  1.00000000 -0.018439486
WindSpeed -0.27055563 -0.002180167 -0.36897251 -0.01843949  1.000000000
```

```
> vif(ToxinConclm)
     ToxinConc$Rain  ToxinConc$NoonTemp  ToxinConc$Sunshine ToxinConc$WindSpeed
           1.031045            1.616535            1.415269            1.209717
```

There doesn't appear to be any strongly correlated variables either using correlation values or the VIF, although in some situations, a VIF of 1.5 is considered high.

It may be worthwhile to build another model keeping one of the correlated variables in the model. The more significant can be preferred but business intuition may be cautiously used to include other statistically insignificant variable(s).

Let us run Stepwise Regression first.

3: Predicting Fungal Toxin Contamination

```
> ToxinConclm1 <- stepAIC(ToxinConclm, direction = "both")
Start:  AIC=12.14
ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp + ToxinConc$Sunshine +
    ToxinConc$WindSpeed

                       Df Sum of Sq     RSS    AIC
- ToxinConc$Sunshine    1     0.540  12.927 10.567
<none>                              12.387 12.141
- ToxinConc$NoonTemp    1     7.510  19.897 14.880
- ToxinConc$WindSpeed   1    17.603  29.990 18.983
- ToxinConc$Rain        1   122.991 135.378 34.055

Step:  AIC=10.57
ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp + ToxinConc$WindSpeed

                       Df Sum of Sq     RSS    AIC
<none>                              12.927 10.567
+ ToxinConc$Sunshine    1     0.540  12.387 12.141
- ToxinConc$NoonTemp    1    13.417  26.344 15.686
- ToxinConc$WindSpeed   1    19.688  32.615 17.822
- ToxinConc$Rain        1   122.830 135.757 32.083
```


3: Predicting Fungal Toxin Contamination

```
Call:
lm(formula = ToxinConc$Toxin ~ ToxinConc$Rain + ToxinConc$NoonTemp +
    ToxinConc$WindSpeed, data = ToxinConc)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6394 -0.9308  0.1394  0.6545  2.0909

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         31.5651     6.6253   4.764  0.00311 **
ToxinConc$Rain       7.0108     0.9285   7.551  0.00028 ***
ToxinConc$NoonTemp  -0.4790     0.1919  -2.495  0.04682 *
ToxinConc$WindSpeed -0.8218     0.2718  -3.023  0.02331 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.468 on 6 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.8726
F-statistic: 21.54 on 3 and 6 DF,  p-value: 0.001298
```

```
> vif(ToxinConclm1)
     ToxinConc$Rain  ToxinConc$NoonTemp ToxinConc$WindSpeed
           1.015857           1.175947           1.159879
```

Toxin concentrations increase with increasing rainfall and decrease in drier climates characterized by higher temperatures and wind speeds.

The business can take a decision to rent farms in drier climates if the cost benefits of saved nuts versus higher rents are high.

4: Predicting Coal Production

Goal is to predict coal production in the US based on multiple factors.

Data is from USA (2011) https://www.eia.gov/coal/data.php#production and https://www.msha.gov/mine-employment-and-coal-production

1798 rows and 16 columns

Mines with production less than 1,00,000 short tons removed leaving 691 records.

Dependent Variable: DV
Independent Variable: IV

IV IV IV DV IV IV

Historic Coal Production Data: 2011
Source: The U.S. Energy Information Administration (EIA) and the U.S. Mine Safety and Health Administration

Year	MSHA ID	Mine Name	Mine State	Mine County	Mine Basin	Mine Status	Mine Type	Company Type	Operation Type	Operating Company	Operating Company Address	Union Code	Production (short tons)	Average Employees	Labor Hours
2011	103397	Parrish Mine	Alabama	Walker	Appalachia Southern	abandoned	Surface	Indepedent Producer Operator	Mine only	Black Warrior Minerals Inc	P.O. Box 1190, Sumiton, AL 35148		6,929	8	5,062
2011	103398	Kansas Mine	Alabama	Walker	Appalachia Southern	Active	Surface	Indepedent Producer Operator	Mine only	National Coal Of Alabama, Inc	1810 Birmingham Ave, Suite 101, Jasper, AL 35501		1,42,764	30	54,137
2011	103410	Coal Valley Mine	Alabama	Walker	Appalachia Southern	Active	Surface	Indepedent Producer Operator	Mine only	Xinergy Of Alabama, Inc.	2600 Warrior Jasper Rd, Warrior, AL 35180		1,71,689	41	58,040
2011	103423	Dutton Hill Mine	Alabama	Walker	Appalachia Southern	Active	Surface	Indepedent Producer Operator	Mine only	Quality Coal Company, Inc.	17405 Hwy 69 S, Jasper, AL 35501		1,12,044	20	45,020
2011	103321	Poplar Springs	Alabama	Winston	Appalachia Southern	Active	Surface	Indepedent Producer Operator	Mine only	National Coal Of Alabama Inc	1810 Birmingham Ave, Suite 101, Jasper, AL 35501		4,49,055	73	2,07,516
2011	103358	Old Union	Alabama	Winston	Appalachia Southern	Active	Surface	Indepedent Producer Operator	Mine only	Birmingham Coal & Coke Co Inc	912 Edenton St, Birmingham, AL 35242		2,48,368	41	93,764
2011	5000030	Usibelli	Alaska	Yukon-Koyukuk	Western	Active	Surface	Indepedent Producer Operator	Mine and Preparation Plant	Usibelli Coal Mine Inc	P.O. Box 1000, Healy, AK 99743	Teamsters	21,48,926	136	3,31,784
2011	201195	Kayenta Mine	Arizona	Navajo	Western	Active	Surface	Operating Subsidiary	Mine and Preparation Plant	Peabody Western Coal Company	P.O. Box 650, Kayenta, AZ 86033	United Mine Workers of America	81,10,942	419	9,61,579
2011	301569	Penny #1	Arkansas	Sebastian	Interior	Active	Surface	Indepedent Producer Operator	Mine only	Comer Mining Company	P.O. Box 986, Greenwood, AR 72936		6,719	3	5,072
2011	301736	Sebastian Mine	Arkansas	Sebastian	Interior	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	South Central Coal Co. Inc.	P.O. Box 6, Spiro, OK 74959		1,26,628	67	1,86,355
2011	503818	Terror Creek Loadout	Colorado	Delta	Uinta Region	Active	Underground	Operating Subsidiary	Preparation Plant	Terror Creek Llc	1840 Hwy 133, Paonia, CO 81428		0	7	12,762
2011	504591	Bowie No 2 Mine	Colorado	Delta	Uinta Region	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	Bowie Resources Llc	P.O. Box 1488, Paonia, CO 81428		22,35,055	306	6,57,340
2011	503672	West Elk Mine	Colorado	Gunnison	Uinta Region	Active	Underground	Operating Subsidiary	Mine and Preparation Plant	Mountain Coal Company, L.L.C.	P.O. Box 591, Somerset, CO 81434		58,96,402	368	8,24,337
2011	504674	Elk Creek Mine	Colorado	Gunnison	Uinta Region	Active	Underground	Operating Subsidiary	Mine only	Oxbow Mining, Llc	P.O. Box 535, Somerset, CO 81434		30,07,833	349	6,79,998
2011	504864	King Ii	Colorado	La Plata	Western	Active	Underground	Indepedent Producer Operator	Mine only	Gcc Energy Llc	6473 County Road 120, Hesperus, CO 81326		6,18,132	70	1,61,320
2011	500296	New Elk Mine	Colorado	Las Animas	Western	Active	Underground	Indepedent Producer Operator	Mine only	New Elk Coal Company Llc	122 W. 1st Street, Trinidad, CO 81082		17,701	188	3,86,565
2011	504461	New Elk Prep Plant	Colorado	Las Animas	Western	not producing	Underground	Indepedent Producer Operator	Preparation Plant	New Elk Coal Company Llc	21250 Hwy 12, Weston, CO 81091		0	19	36,778
2011	502838	Trapper Mine	Colorado	Moffat	Uinta Region	Active	Surface	Indepedent Producer Operator	Mine only	Trapper Mining Inc	P.O. Box 187, Craig, CO 81626	Engineers	24,84,106	189	3,96,928
2011	502962	Colowyo Mine	Colorado	Moffat	Uinta Region	Active	Surface	Operating Subsidiary	Mine and Preparation Plant	Colowyo Coal Company L P	5731 State Highway 13, Meeker, CO 81641		25,37,904	262	4,98,819
2011	500299	New Horizon Mine	Colorado	Montrose	Western	Active	Surface	Indepedent Producer Operator	Mine only	Western Fuels-Colorado, Llc	P.O. Box 628, Nucla, CO 81424	United Mine Workers of America	3,60,009	27	50,571
2011	503505	Deserado Mine	Colorado	Rio Blanco	Uinta Region	Active	Underground	Operating Subsidiary	Mine and Preparation Plant	Blue Mountain Energy	3607 County Road #65, Rangely, CO 81648	United Mine Workers of America	19,83,581	169	3,00,545
2011	503836	Foidel Creek Mine	Colorado	Routt	Uinta Region	Active	Underground	Operating Subsidiary	Mine and Preparation Plant	Peabody Twentymile Mining Llc	29515 Routt County Rd#27, Oak Creek, CO 80467		77,48,909	451	9,05,287
2011	1103189	Mc#1 Mine	Illinois	Franklin	Illinois Basin	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	M-Class Mining Llc	11351 N. Thompsonville Rd., Macedonia, IL 62860		8,92,612	137	3,29,645
2011	1103207	Old Ben #25	Illinois	Franklin	Illinois Basin	Active	Surface	Indepedent Producer Operator	Mine and Preparation Plant	S.I. Energy Llc	1867 Antioch Road, West Frankfor, IL 62896		0	4	5,195
2011	1103207	Old Ben #25	Illinois	Franklin	Illinois Basin	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	S.I. Energy Llc	1867 Antioch Road, West Frankfor, IL 62896		0	4	5,194
2011	1102751	I-1 Mine	Illinois	Gallatin	Illinois Basin	not producing	Surface	Operating Subsidiary	Mine and Preparation Plant	Illinois Fuel Company Llc	P.O. Box 7, Herod, IL 62947		52,923	17	36,021
2011	1103017	Cottage G	Illinois	Gallatin	Illinois Basin	Active	Surface	Operating Subsidiary	Mine and Preparation Plant	Peabody Midwest Mining, Llc	7100 Eagle Crest Blvd., Suite, Evansville, IN 47715		18,97,146	193	4,63,308
2011	1103020	Creek Paum Mine	Illinois	Jackson	Illinois Basin	Active	Surface	Indepedent Producer Operator	Mine and Preparation Plant	Knight Hawk Coal Llc	500 Cuttler-Trico Road, Percy, IL 62272		3,81,479	46	98,440
2011	1103020	Creek Paum Mine	Illinois	Jackson	Illinois Basin	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	Knight Hawk Coal Llc	500 Cuttler-Trico Road, Percy, IL 62272		0	5	12,995
2011	1103140	Cora Terminal	Illinois	Jackson	Illinois Basin	Active	Surface	Indepedent Producer Operator	Preparation Plant	Kinder Morgan Operating Lp "b	262 Cora Road, Rockwood, IL 62280	Engineers	0	25	47,806
2011	1103140	Cora Terminal	Illinois	Jackson	Illinois Basin	Active	Underground	Indepedent Producer Operator	Preparation Plant	Kinder Morgan Operating Lp "b	262 Cora Road, Rockwood, IL 62280	Engineers	0	26	47,806
2011	1103209	Mine	Illinois	Mcdonough	Illinois Basin	Active	Surface	Indepedent Producer Operator	Mine and Preparation Plant	Black Nugget Llc	P.O. Box 235, Industry, IL 61440		1,65,782	26	57,828
2011	1100726	Shay #1 Mine	Illinois	Macoupin	Illinois Basin	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	Maryan Mining Llc	14300 Brushy Mound Rd, Carlinville, IL 62626		18,29,122	111	2,46,671
2011	1102632	Crown Iii Mine	Illinois	Macoupin	Illinois Basin	Active	Underground	Contractor	Mine and Preparation Plant	Tri County Coal, Llc	2 Mine Avenue/P. O. Box 139, Farmersville, IL 62533	United Mine Workers of America	9,34,682	213	5,02,981
2011	1103182	Deer Run Mine	Illinois	Montgomery	Illinois Basin	Active	Underground	Indepedent Producer Operator	Mine and Preparation Plant	Patton Mining Llc	211 N Broadway,ste 2600, St. Louis, MO 63102		4,91,227	78	1,86,086
2011	1103045	Red Hawk	Illinois	Perry	Illinois Basin	Active	Surface	Indepedent Producer Operator	Mine and Preparation Plant	Knight Hawk Coal, Llc	500 Cutler-Trico Road, Percy, IL 62272		5,32,977	36	78,126
2011	1103143	Prairie Eagle	Illinois	Perry	Illinois Basin	Active	Surface	Indepedent Producer Operator	Mine and Preparation Plant	Knight Hawk Coal, Llc	500 Cutler-Trico Road, Percy, IL 62272		3,06,871	30	63,133

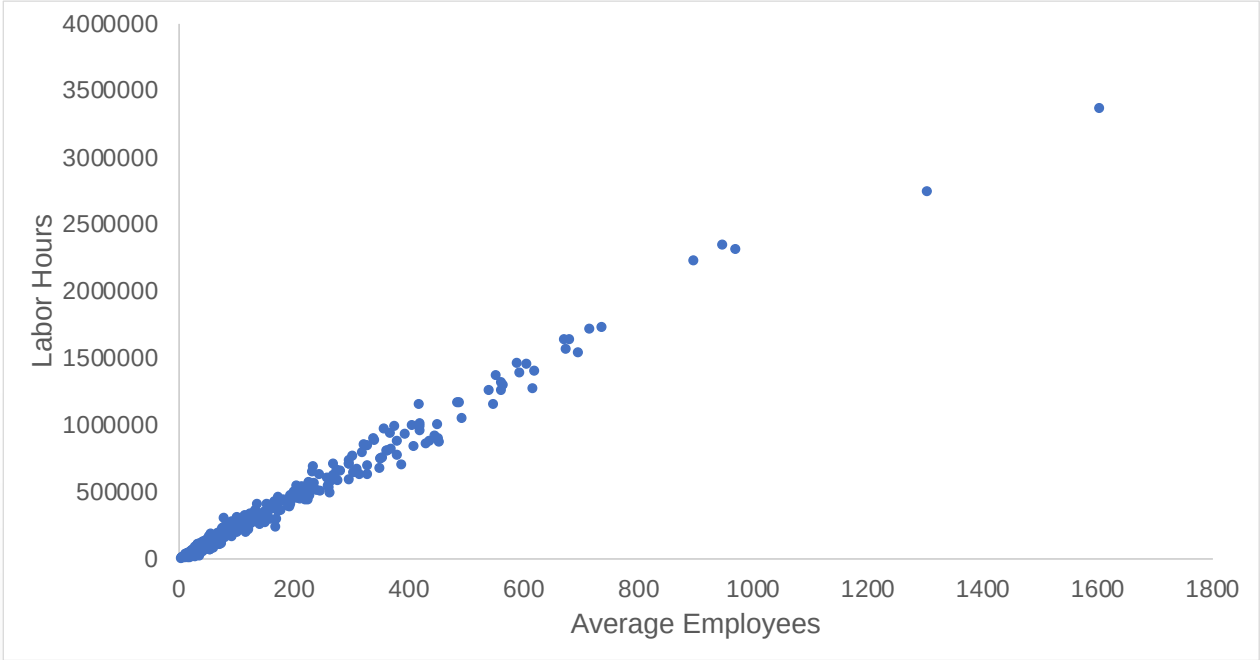
1402

4: Predicting Coal Production

DV: Production - Numeric
IV

- Average Employees – Numeric
- Labor Hours – Numeric
- Mine Basin – Nominal with 8 levels
  - Appalachia Central
  - Appalachia Northern
  - Appalachia Southern
  - Illinois Basin
  - Interior
  - Powder River Basin
  - Uinta Region
  - Western
- Mine Type – Nominal (Binary)
  - Surface
  - Underground
- Operation Type – Nominal (Binary)
  - Mine and Preparation Plant
  - Mine Only

Logical thinking and a little Exploratory Data Analysis (EDA) suggests Average Employees and Labor Hours must be correlated.



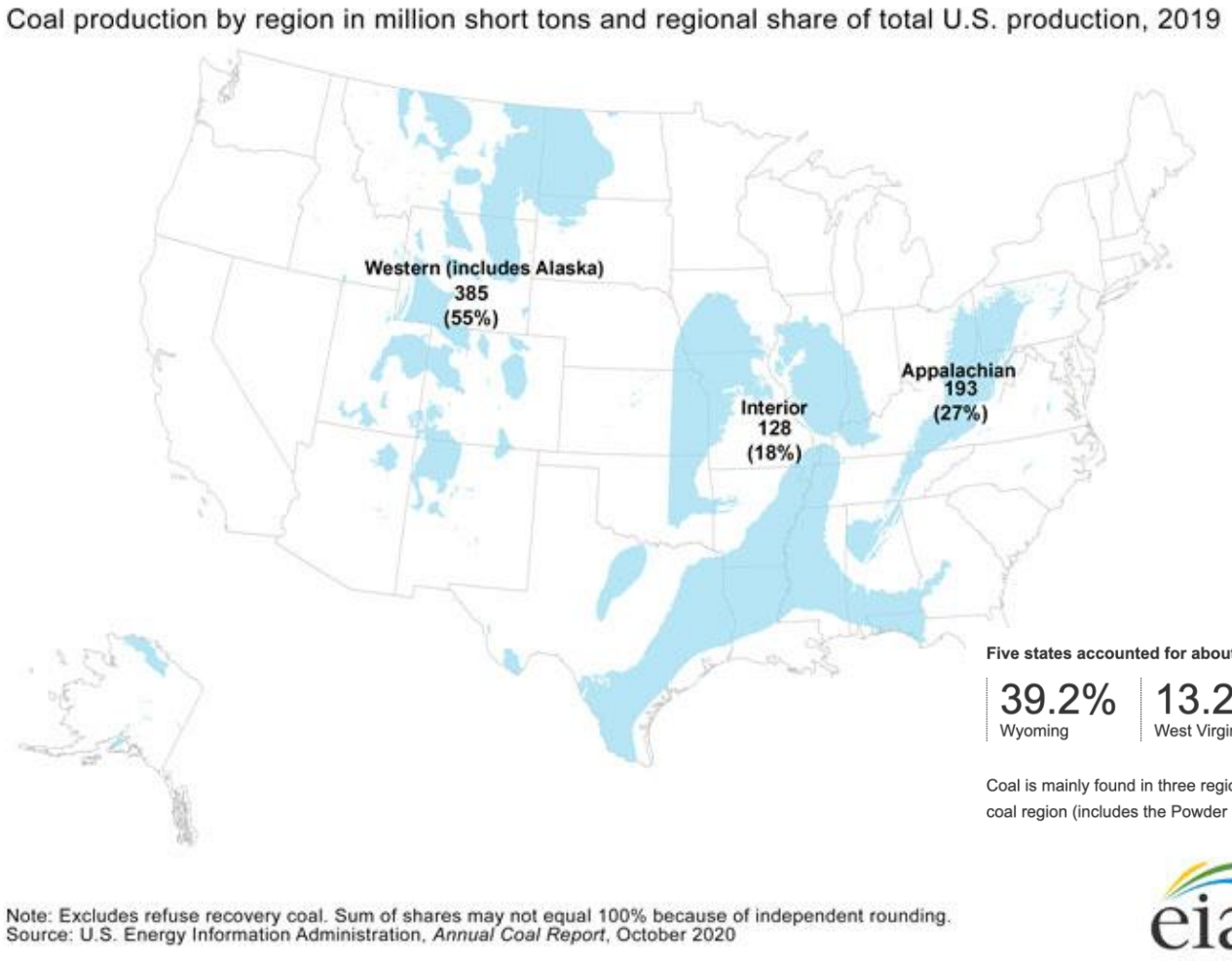

```
> cor(coalUSA2011$AverageEmployees, coalUSA2011$LaborHours)
[1] 0.9935286
```

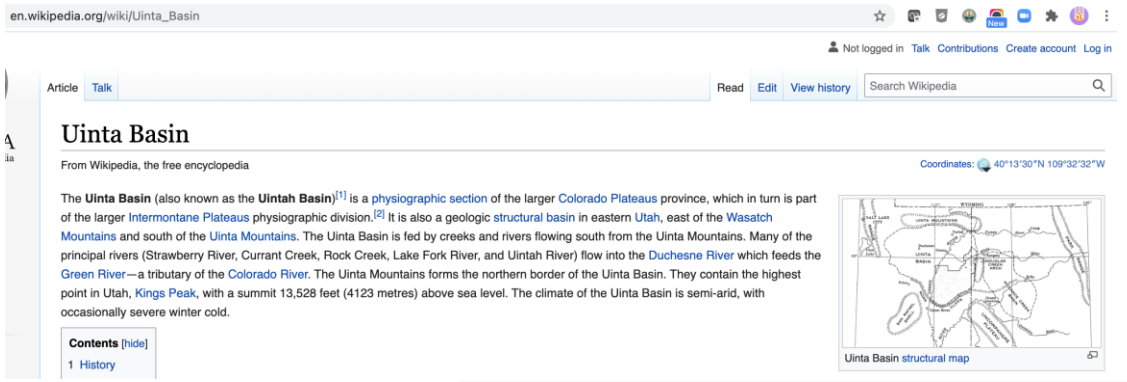
Plotting and checking correlation confirms it. Let us drop *Average Employees* from independent variables.

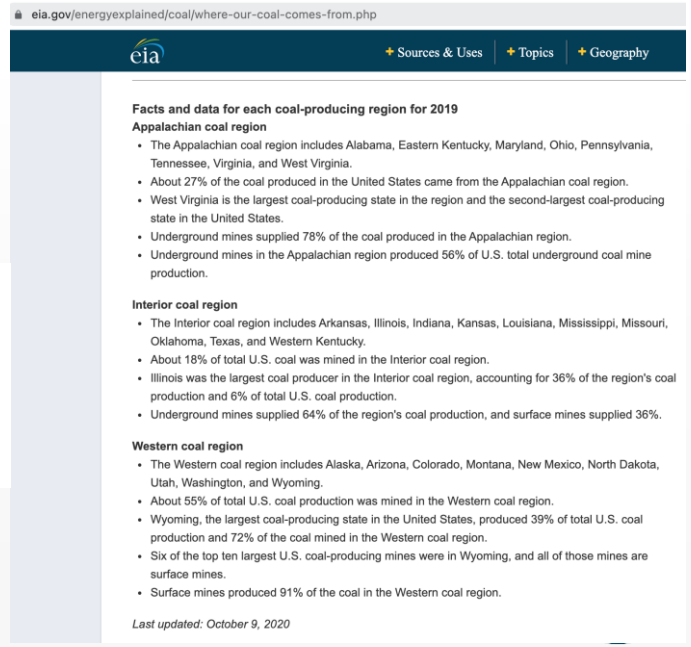
DSC 7402

4: Predicting Coal Production

A little domain study showed that there are 3 primary coal producing regions in the US: Appalachia, Interior and Western. Levels appropriately clubbed based on this information (reduced from 8 to 3).







Five states accounted for about 71% of total U.S. coal production in 2019:

39.2%	13.2%	7.1%	6.5%	5.1%
Wyoming	West Virginia	Pennsylvania	Illinois	Kentucky

Coal is mainly found in three regions: the Appalachian coal region, the Interior coal region, and the Western coal region (includes the Powder River Basin).

eia

4: Predicting Coal Production

Build the baseline model.

```
Call:
lm(formula = Production ~ MineBasin + MineType + OperationType +
    LaborHours, data = coalUSA2011)

Residuals:
      Min        1Q    Median        3Q       Max
-17850128   -697950    122378   1460278  72696609

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.015e+06  6.542e+05  -3.080  0.00215 **
MineBasinInterior       -5.245e+05  5.611e+05  -0.935  0.35027
MineBasinWestern         6.543e+06  7.642e+05   8.562  < 2e-16 ***
MineTypeUnderground     -2.390e+06  3.735e+05  -6.398 2.91e-10 ***
OperationTypeMine only   1.500e+06  6.056e+05   2.476  0.01351 *
LaborHours               1.158e+01  6.077e-01  19.063  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4706000 on 685 degrees of freedom
Multiple R-squared:  0.4896,    Adjusted R-squared:  0.4858
F-statistic: 131.4 on 5 and 685 DF,  p-value: < 2.2e-16
```

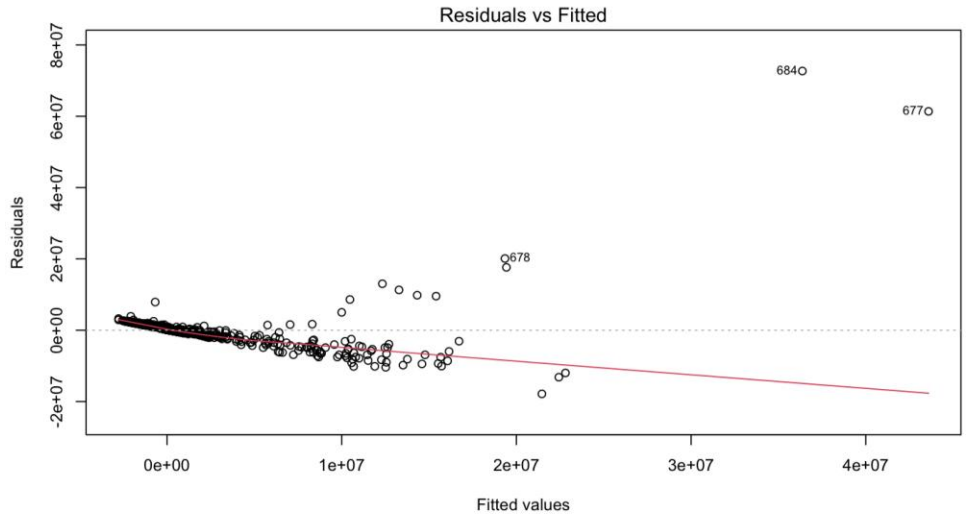
Observations and Analysis

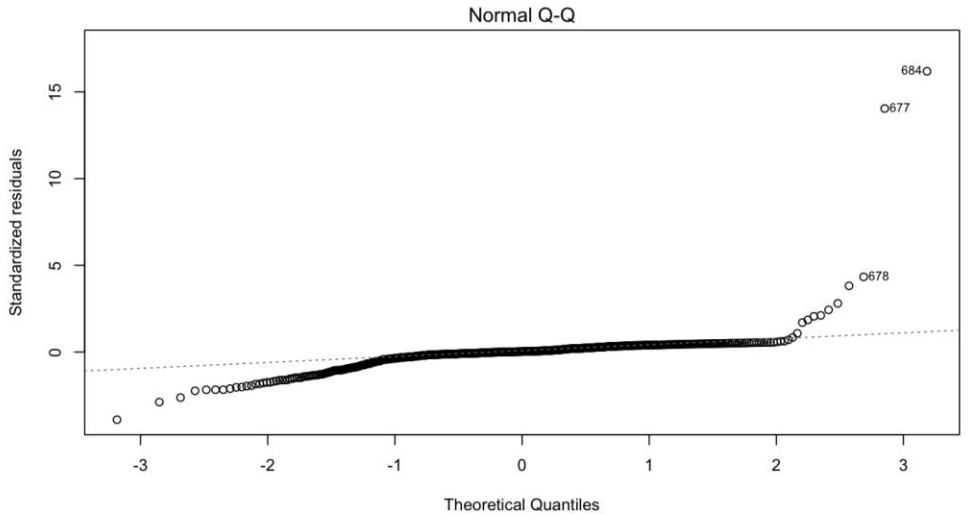
- Adj R^2 is 48.58%. Not great.
- Interior region is not significantly different from the Appalachia region (reference).
- Model is significant and so are other variables.

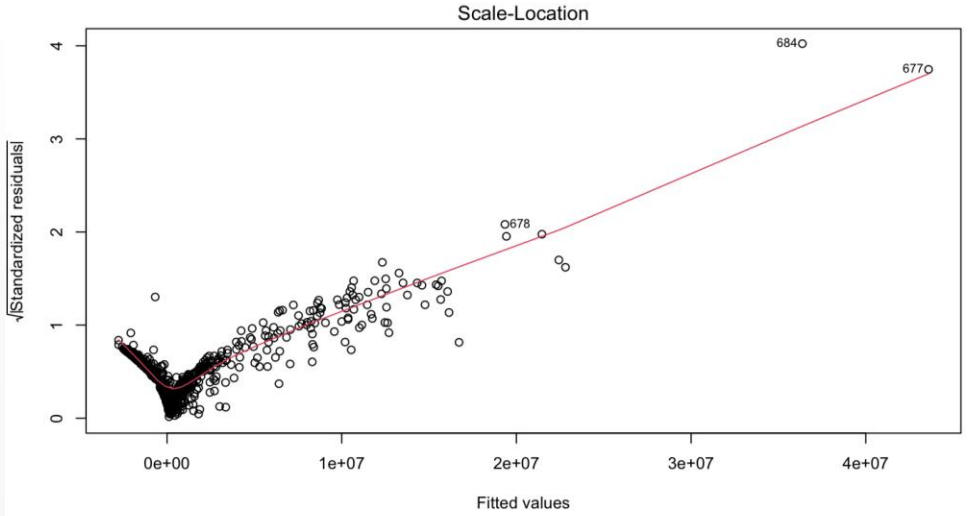
4: Predicting Coal Production

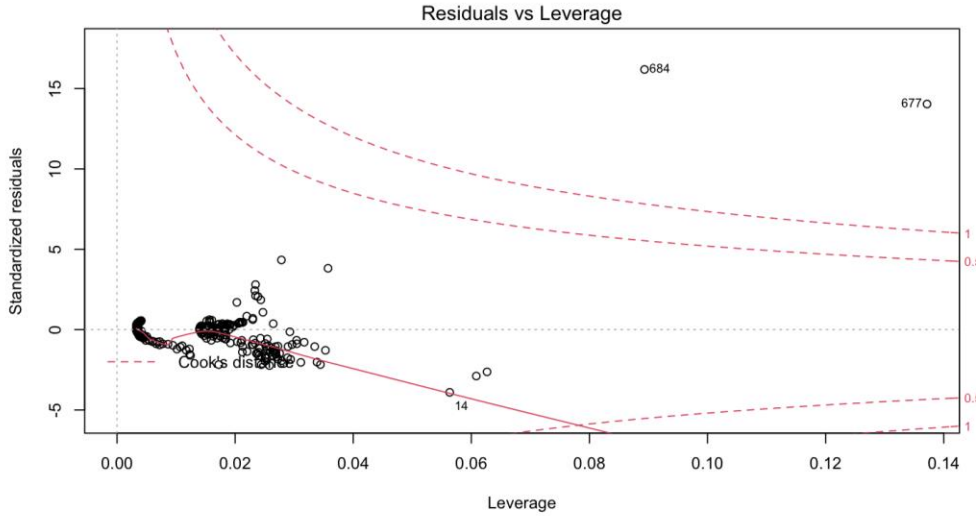
Residual analysis shows problems in all 4 plots. Points 677 and 684 appear way off.

Let us try data transformations.



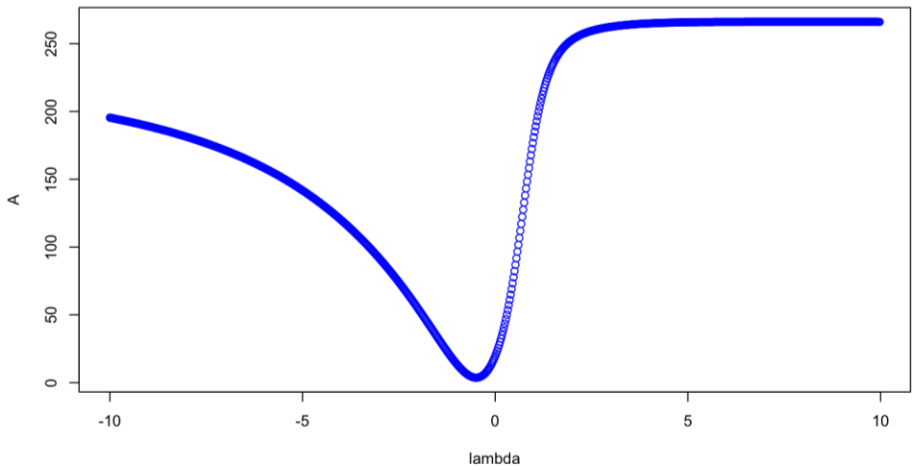


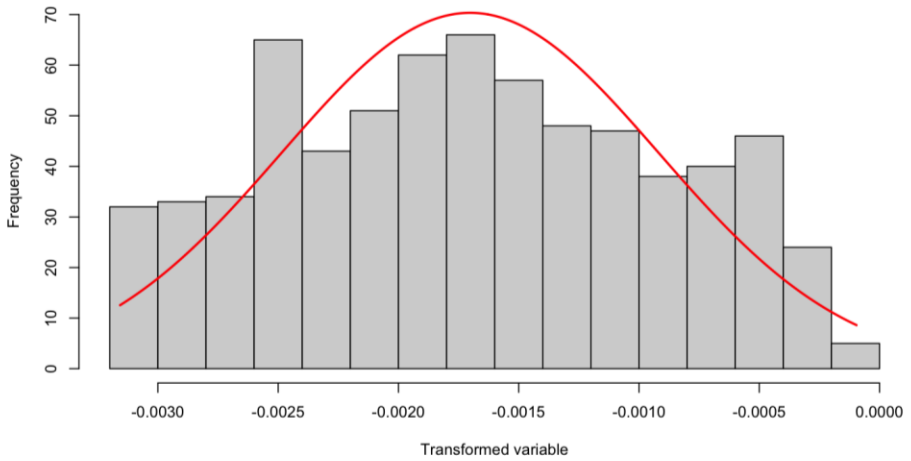


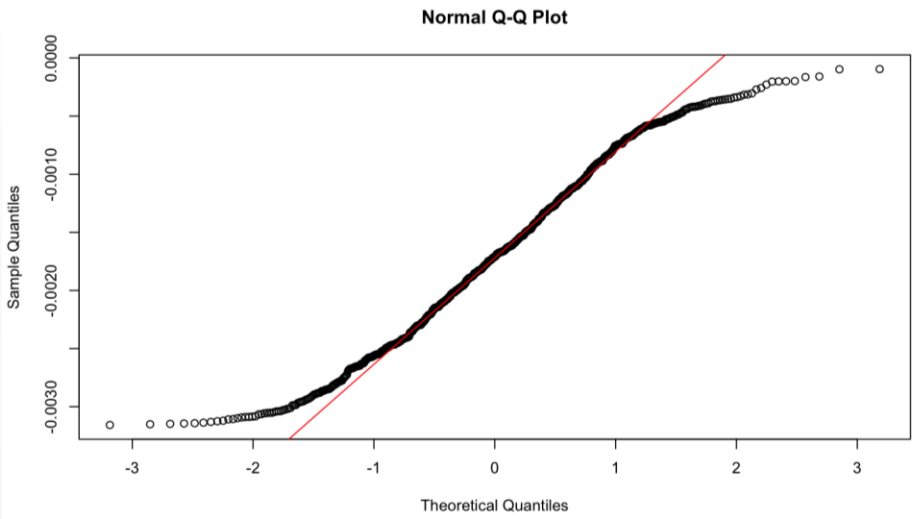


4: Predicting Coal Production

transformTukey(coalUSA2011$Production, plotit = TRUE, statistic = 2)








```
    lambda      W Shapiro.p.value     A Anderson.p.value
381   -0.5 0.9739       9.323e-10 3.619        4.841e-09
```

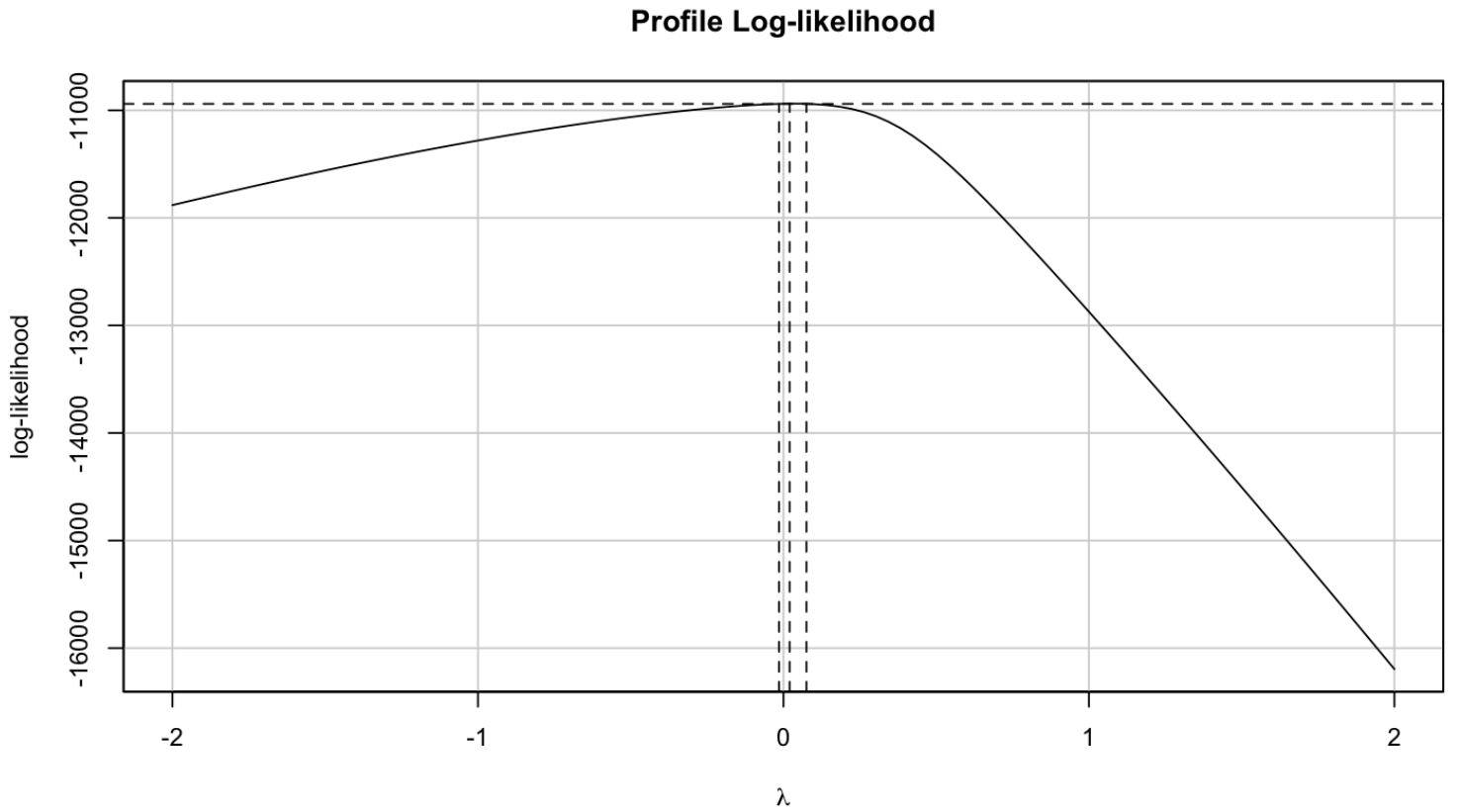
```
if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}
```

Tukey transformation suggests $-\frac{1}{\sqrt{y}}$, although the hypothesis tests still reject normality after transformation.

DSC 7402

4: Predicting Coal Production

boxCox(coalUSA2011lm, family = "bcPower")



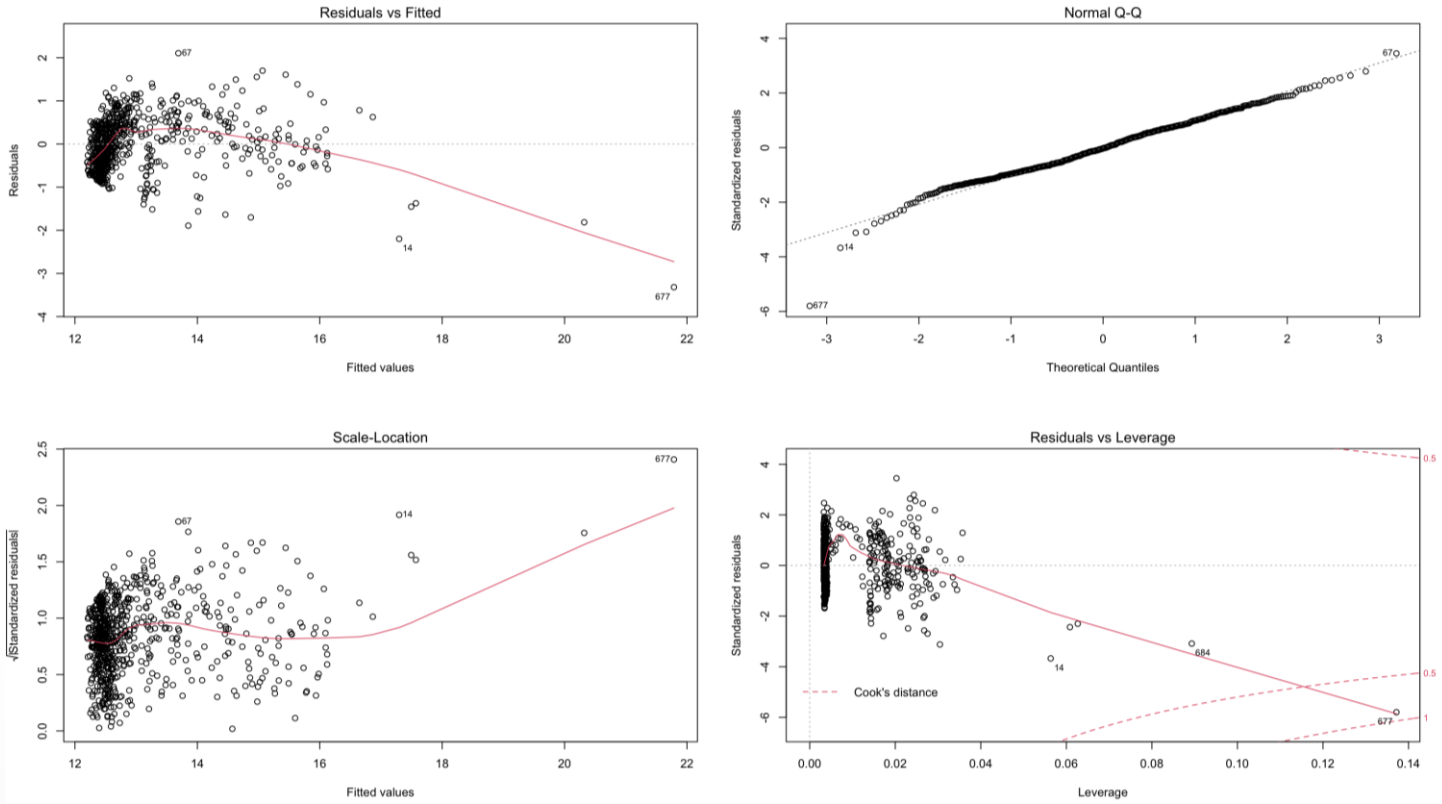
Box-Cox suggests log transformation.

Let us rebuild the model with log(Production) as the DV.

DSC 7402

4: Predicting Coal Production

```
Call:
lm(formula = log(Production) ~ MineBasin + MineType + OperationType +
    LaborHours, data = coalUSA2011)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3181 -0.4328 -0.0153  0.4263  2.1037

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.225e+01  8.567e-02 142.980  < 2e-16 ***
MineBasinInterior        7.774e-01  7.347e-02  10.581  < 2e-16 ***
MineBasinWestern         1.611e+00  1.001e-01  16.102  < 2e-16 ***
MineTypeUnderground     -1.999e-01  4.891e-02  -4.088 4.87e-05 ***
OperationTypeMine only   6.815e-02  7.930e-02   0.859     0.39
LaborHours               2.351e-06  7.957e-08  29.545  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6162 on 685 degrees of freedom
Multiple R-squared:  0.7499,    Adjusted R-squared:  0.7481
F-statistic: 410.8 on 5 and 685 DF,  p-value: < 2.2e-16
```



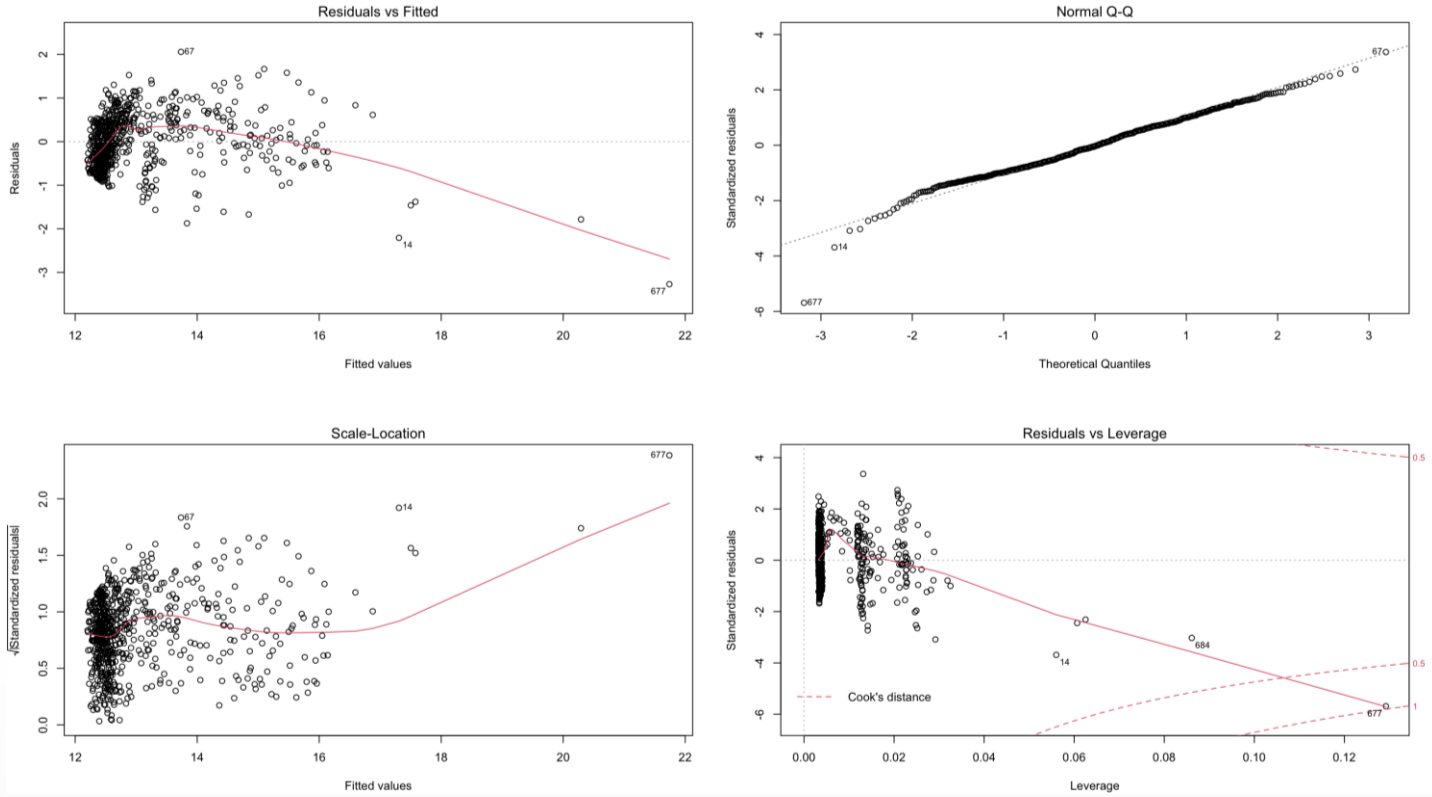
Observations and Analysis

- Adj R^2 is 74.81%. That is a significant improvement.
- Interior region and Western regions are significantly different from the Appalachia region (reference).
- Operation Type (binary) is not significant. Let us remove and build a new model.
- Residuals look better but not good enough.

DSC7402

4: Predicting Coal Production

```
Call:
lm(formula = log(Production) ~ MineBasin + MineType + LaborHours,
    data = coalUSA2011)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2708 -0.4372 -0.0200  0.4316  2.0589

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.231e+01  3.703e-02 332.584  < 2e-16 ***
MineBasinInterior    7.611e-01  7.098e-02  10.724  < 2e-16 ***
MineBasinWestern     1.591e+00  9.718e-02  16.371  < 2e-16 ***
MineTypeUnderground -1.951e-01  4.858e-02  -4.017 6.56e-05 ***
LaborHours           2.323e-06  7.277e-08  31.924  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6161 on 686 degrees of freedom
Multiple R-squared:  0.7496,    Adjusted R-squared:  0.7482
F-statistic: 513.5 on 4 and 686 DF,  p-value: < 2.2e-16
```



Observations and Analysis

- Adj R^2 is 74.82%. We have a simpler model (one variable less) with no change in Adj R^2.
- Let us check for multicollinearity and try Stepwise Regression.

DSC 7402

4: Predicting Coal Production

```
                GVIF Df GVIF^(1/(2*Df))
MineBasin   1.170567  2        1.040157
MineType    1.073257  1        1.035981
LaborHours  1.196626  1        1.093904
```

*NOTE: When categorical variables are present, G(eneralized)VIF is output instead of VIF. Rule of thumb is to take the **square** of the values shown for* $GVIF^{\frac{1}{2*df}}$. *For numeric or binary variables, this is the same as shown in GVIF column.*

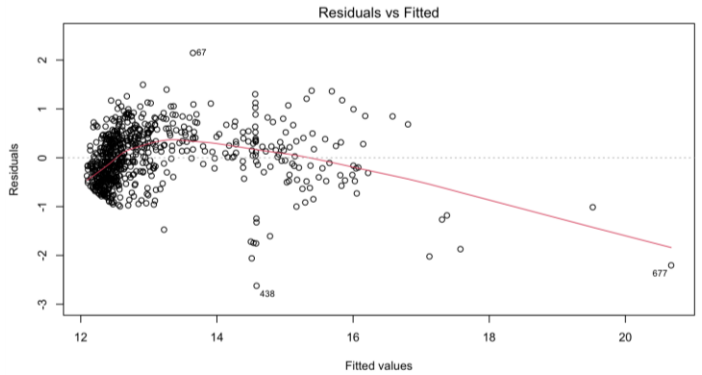
```
> stepAIC(coalUSA2011lm3, direction = "both")
Start:  AIC=-664.37
log(Production) ~ MineBasin + MineType + LaborHours

             Df Sum of Sq    RSS     AIC
<none>                    260.40 -664.37
- MineType    1      6.12 266.52 -650.31
- MineBasin   2    126.10 386.50 -395.48
- LaborHours  1    386.85 647.25  -37.20

Call:
lm(formula = log(Production) ~ MineBasin + MineType + LaborHours,
    data = coalUSA2011)

Coefficients:
        (Intercept)    MineBasinInterior     MineBasinWestern  MineTypeUnderground
          1.231e+01             7.611e-01             1.591e+00            -1.951e-01
         LaborHours
          2.323e-06
```

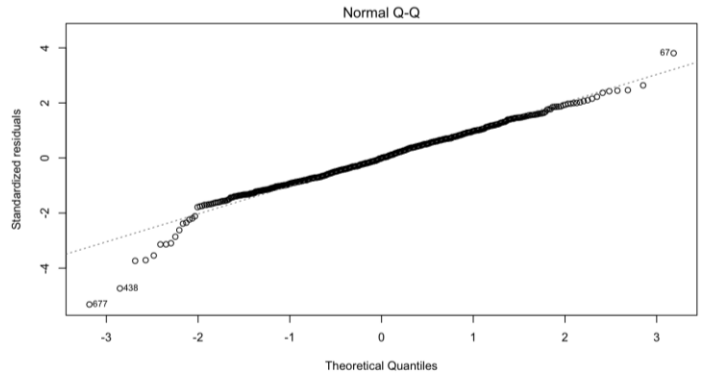
Observations and Analysis

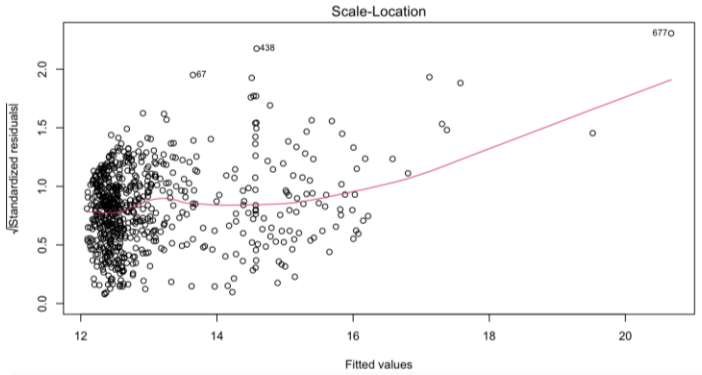
- Can we think of interaction terms?
- It is reasonable to think that there could be a dependence of the labor hours on whether the mine is on surface or underground or even in different regions.
- Let us try interaction terms as well.

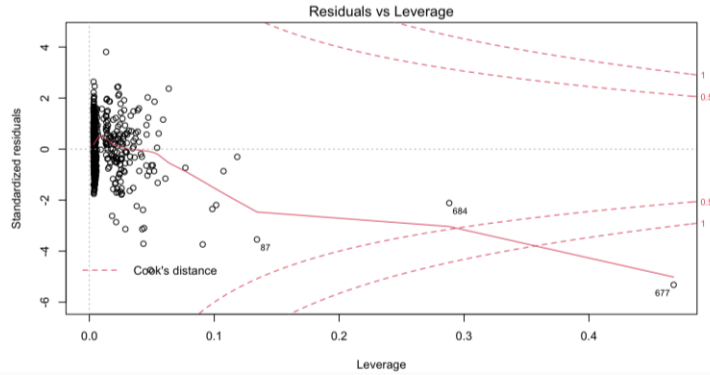
4: Predicting Coal Production

```
Call:
lm(formula = log(Production) ~ MineBasin + MineType + LaborHours +
    LaborHours:MineBasin + LaborHours:MineType, data = coalUSA2011)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62272 -0.39065 -0.00416  0.38220  2.14613

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.205e+01  4.419e-02 272.643  < 2e-16 ***
MineBasinInterior              5.978e-01  9.607e-02   6.223 8.51e-10 ***
MineBasinWestern               2.395e+00  1.303e-01  18.374  < 2e-16 ***
MineTypeUnderground            1.387e-01  5.890e-02   2.354   0.0188 *
LaborHours                     4.079e-06  1.860e-07  21.937  < 2e-16 ***
MineBasinInterior:LaborHours   1.349e-07  1.989e-07   0.678   0.4978
MineBasinWestern:LaborHours   -2.232e-06  2.267e-07  -9.848  < 2e-16 ***
MineTypeUnderground:LaborHours -1.870e-06  1.876e-07  -9.966  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5677 on 683 degrees of freedom
Multiple R-squared:  0.7884,    Adjusted R-squared:  0.7862
F-statistic: 363.5 on 7 and 683 DF,  p-value: < 2.2e-16
```







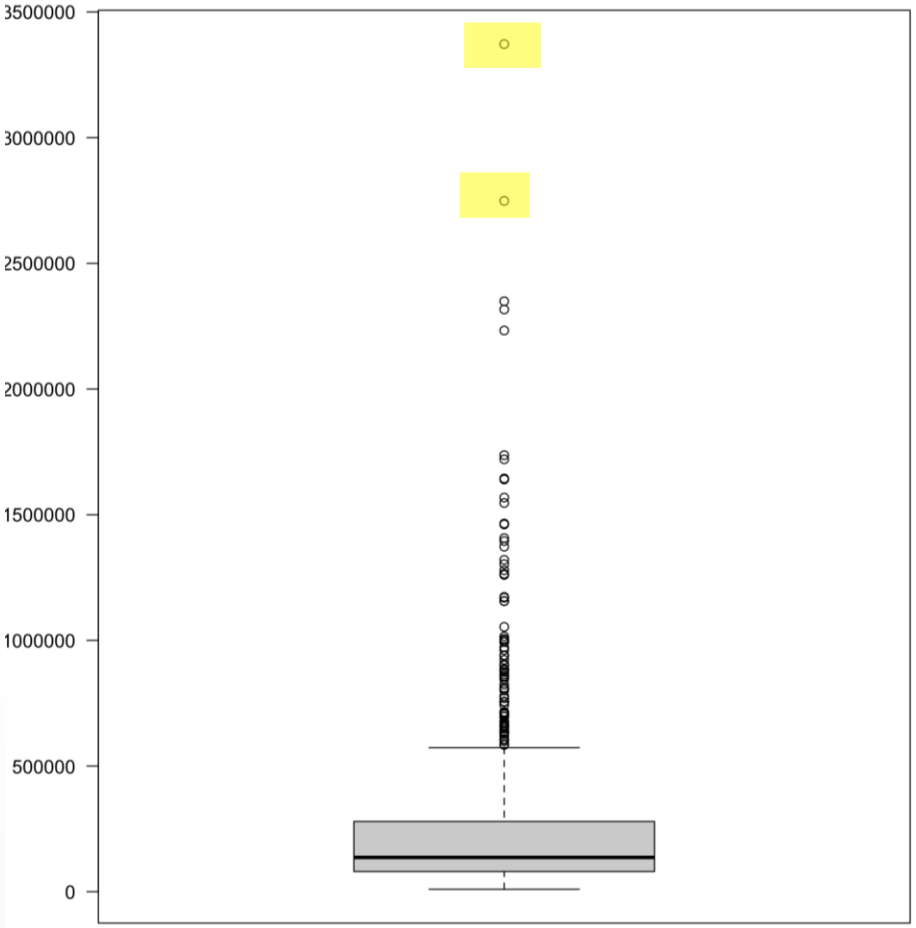


Observations and Analysis

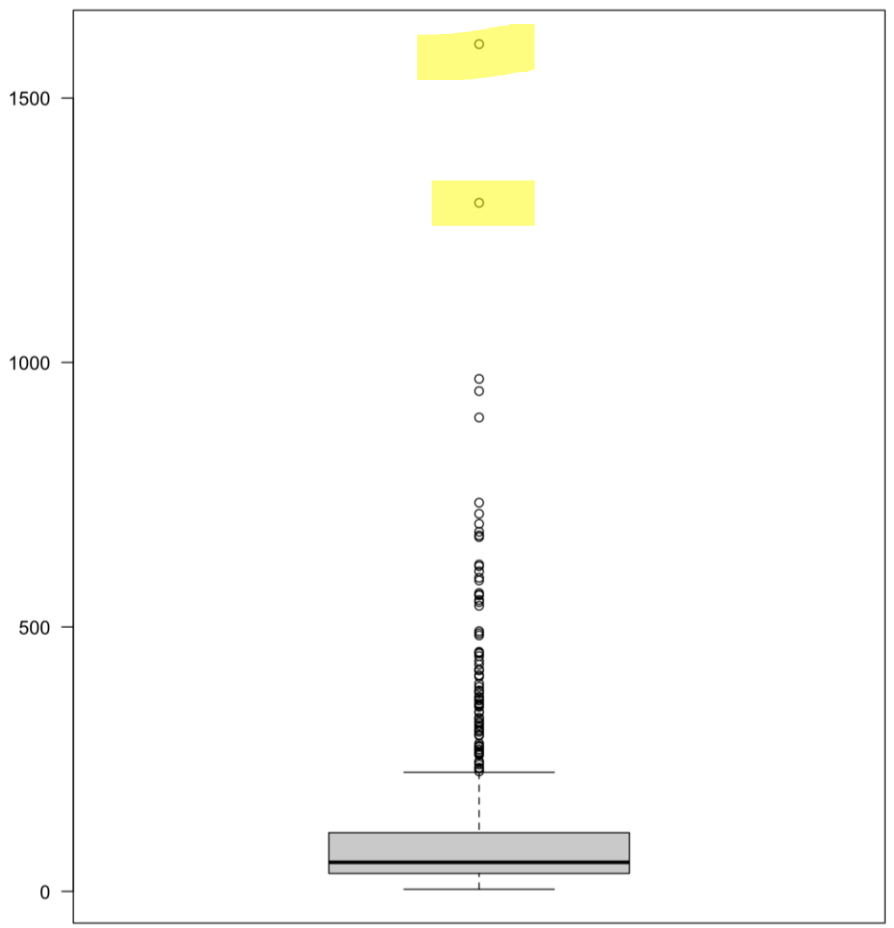
- Adj R^2 has further improved to 78.62%. Point 677 continues to be influential. We had earlier seen that point 684 too had an impact.
- Our final option is to deal with the influential points.
- Let us build a Box Plot on Average Employees and Labor Hours to understand these points more.

DSC 7402

4: Predicting Coal Production



Labor Hours



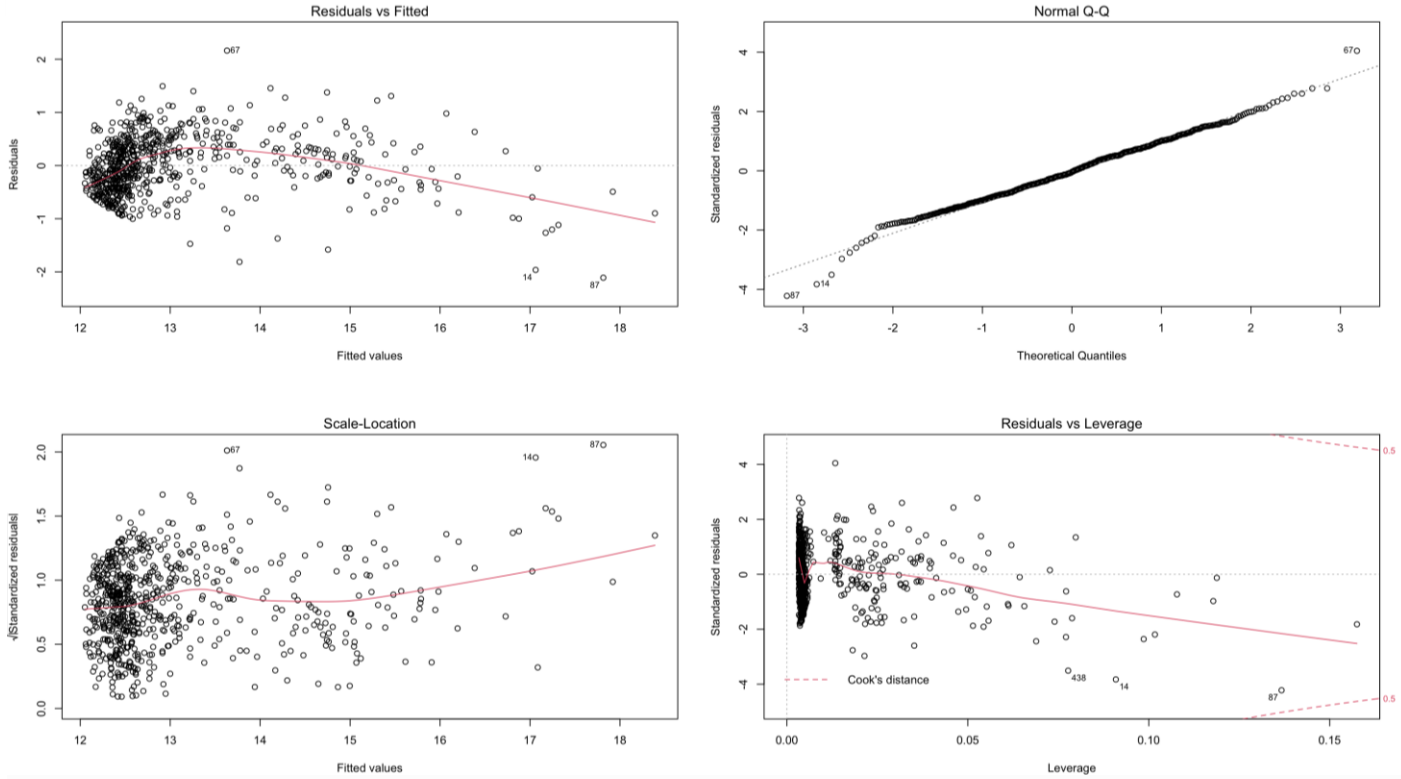
Average Employees

Observations and Analysis

Average Employees are over 1000 only in two mines represented by points 677 and 684 in the dataset. The *Labor Hours* are also high for them. Let us remove these two points and rebuild the model.

4: Predicting Coal Production

```
Call:
lm(formula = log(Production) ~ MineBasin + MineType + LaborHours +
    LaborHours:MineBasin + LaborHours:MineType, data = coalUSA2011Minus677_684)

Residuals:
     Min       1Q   Median       3Q      Max
-2.10972 -0.38744 -0.02195  0.35895  2.16180

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.201e+01  4.213e-02 285.043  < 2e-16 ***
MineBasinInterior              5.995e-01  9.103e-02   6.585 9.08e-11 ***
MineBasinWestern               1.484e+00  1.605e-01   9.248  < 2e-16 ***
MineTypeUnderground            1.919e-01  5.613e-02   3.418 0.000668 ***
LaborHours                     4.366e-06  1.792e-07  24.366  < 2e-16 ***
MineBasinInterior:LaborHours   8.561e-08  1.886e-07   0.454 0.649933
MineBasinWestern:LaborHours   -5.379e-07  2.874e-07  -1.872 0.061690 .
MineTypeUnderground:LaborHours -2.188e-06  1.814e-07 -12.062  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5379 on 681 degrees of freedom
Multiple R-squared:  0.7991,    Adjusted R-squared:  0.797
F-statistic:   387 on 7 and 681 DF,  p-value: < 2.2e-16
```



Observations and Analysis

- Adj R^2 has slightly improved to 79.7%.
- Very importantly, residual plots have improved a lot.
- We can try to transform *LaborHours* also and see if the residual plots improve further. If not, this model is still pretty good.

DSC 7402

4: Predicting Coal Production

```
> boxTidwell(log(Production) ~ LaborHours, other.x=~MineBasin + MineType,
 data = coalUSA2011Minus677_684)
 MLE of lambda Score Statistic (z)  Pr(>|z|)
       0.26647                -21.098 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  4
```

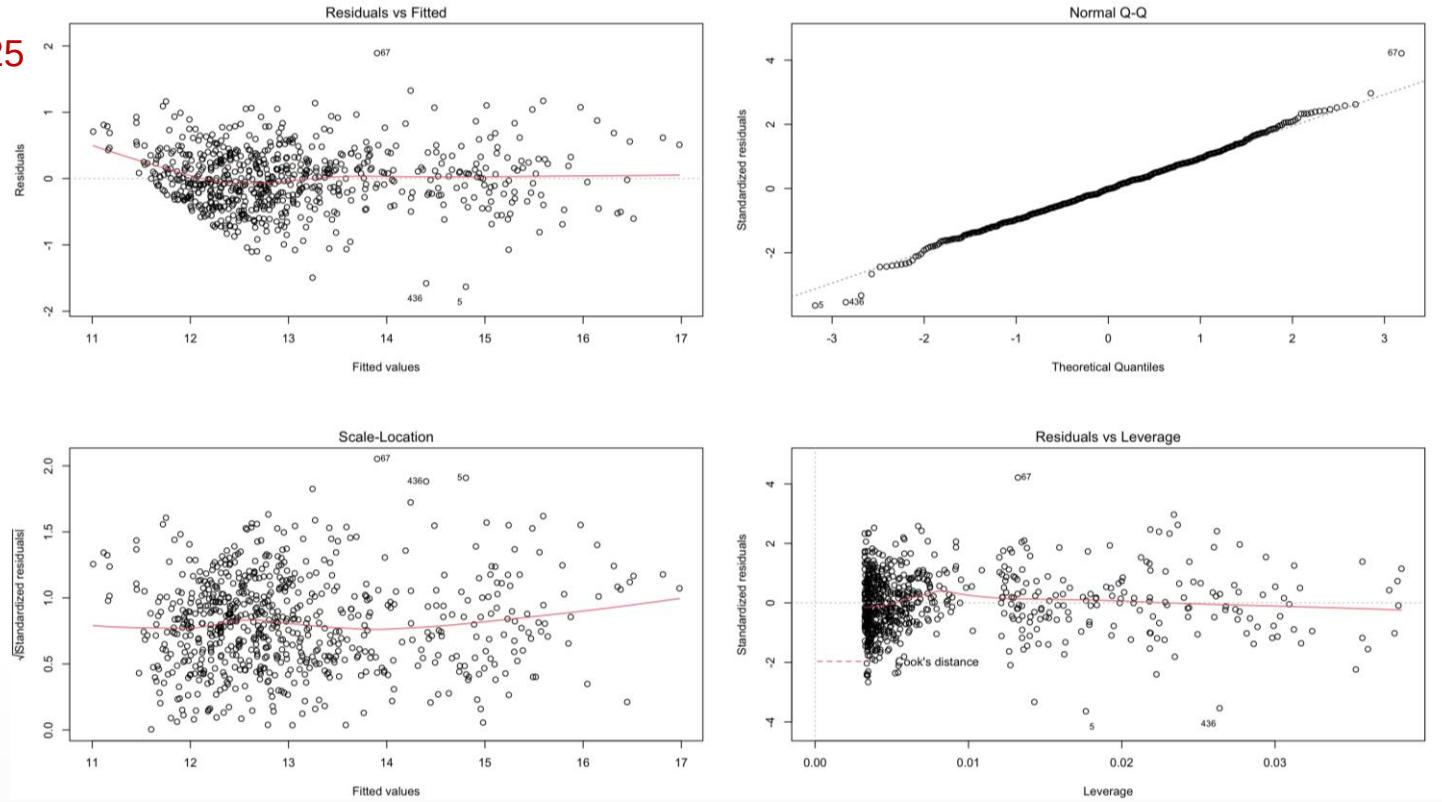
Observations and Analysis

- Box-Tidwell suggests a transformation of 0.26647 power, or approximately 1/4th power on LaborHours.
- That transformation would be hard to explain but let us see what it does.

4: Predicting Coal Production

```
> LaborHours25 <- coalUSA2011Minus677_684$LaborHours^0.25
```

```
Call:
lm(formula = log(Production) ~ MineBasin + LaborHours25 * MineType,
    data = coalUSA2011Minus677_684)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63239 -0.30166 -0.00684  0.29148  1.89168

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      8.997629   0.105797  85.047  < 2e-16 ***
MineBasinInterior                0.514340   0.053019   9.701  < 2e-16 ***
MineBasinWestern                 1.187361   0.074483  15.941  < 2e-16 ***
LaborHours25                     0.202088   0.005610  36.022  < 2e-16 ***
MineTypeUnderground              0.176664   0.147594   1.197    0.232
LaborHours25:MineTypeUnderground -0.030964   0.007091  -4.367 1.46e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4521 on 683 degrees of freedom
Multiple R-squared:  0.8577,    Adjusted R-squared:  0.8566
F-statistic: 823.2 on 5 and 683 DF,  p-value: < 2.2e-16
```



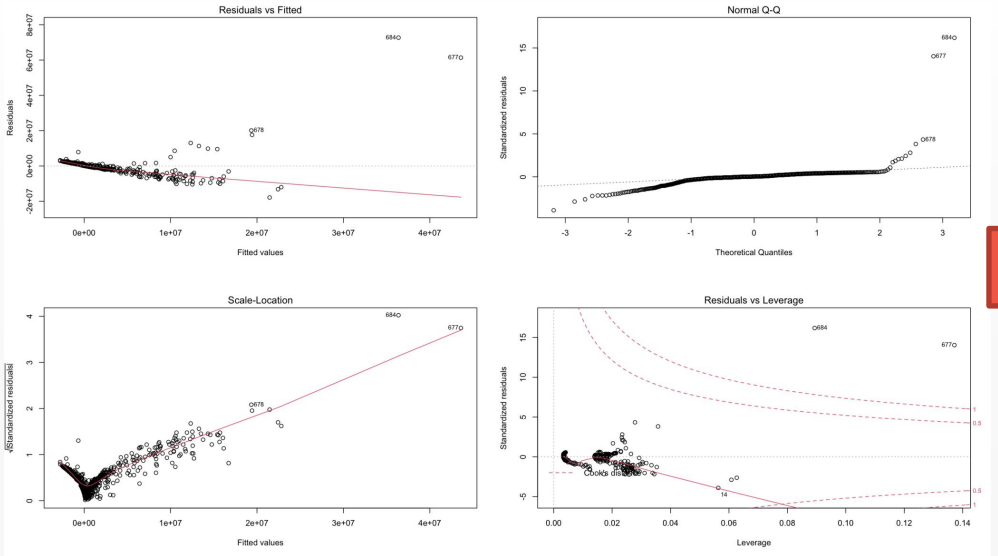
Observations and Analysis

- Adj R^2 has improved dramatically to 85.66%. Note we started with 48.58%.
- Very importantly, residual plots look perfect now.
- If **accuracy of predictions** is important, this model is best. If **explicability** is important and the transformation on *LaborHours* can be explained by a domain expert, etc., use this model; else, use the previous model.

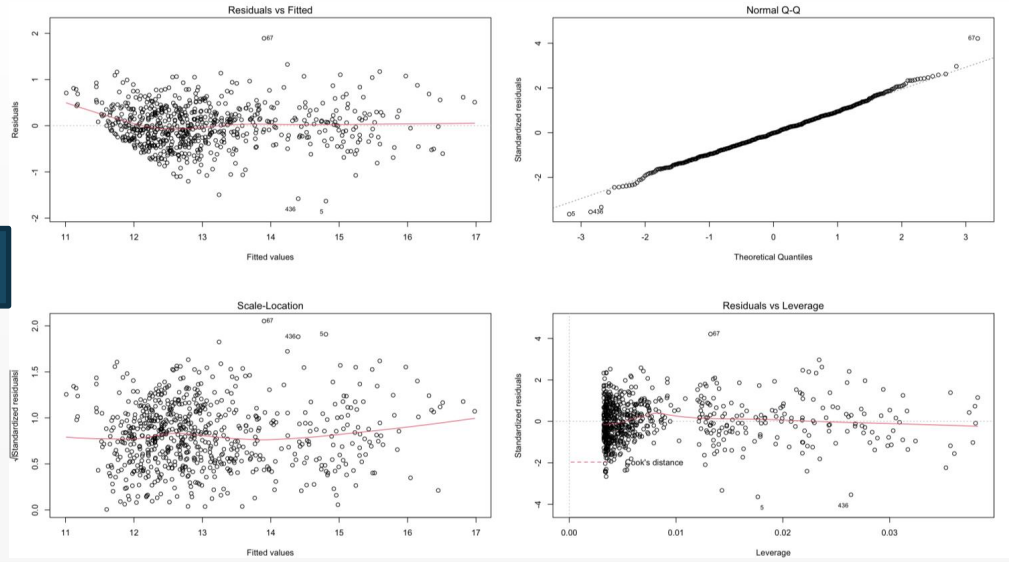
4: Predicting Coal Production

Let us check the performance of the various models built on the fitted data.

Model #	Key Idea	Adj R^2	MAPE
1	Baseline model	48.58%	3.96%
2	log transformed DV *Production*	74.81%	0.58%
3	Removed insignificant variable *OperationType*	74.82%	0.58%
4	Interaction terms	78.62%	0.52%
5	Removed 2 influential points	79.7%	0.47%
6	Transformed IV *LaborHours*	85.66%	0.38%



Model 1 → Model 6



Learning Outcomes of Today's Session

*When revising the material, keep in mind that if you can **confidently** and **fluently** answer the below, you have understood everything that needs to be understood from today's session – these are the expected outcomes from your learning today. First revise (material and videos of the class) and then ask questions.*

Be able to:

- explain adjusted R2 and how it is different from R2
- explain why polynomial regression is also linear
- explain what types of data transformations can fix what issues identified during residuals analysis
- explain the use of interaction terms in building linear regression models
- explain how to handle categorical independent variables
- explain Stepwise Regression approach in feature selection and model building
- explain AIC and its usage in model selection
- define multicollinearity and explain ways of identifying it
- explain VIF and its usage in multicollinearity
- explain use of various error metrics in performance evaluation of models

INSOFE's Vision

The BEST GLOBAL DESTINATION for individuals and organizations to learn and adopt disruptive technologies for solving business and society's challenges.

INSOFE - HYDERABAD
2nd Floor, Jyothi Imperial, Vamsiram Builders
Janardana Hills, Gachibowli
Hyderabad – 500032
+91 93199 77257

INSOFE - BENGALURU
Floors 1-3, L77, 15th Cross Road
Sector 6, HSR Layout
Bengaluru - 560102
+91 93199 77267

INSOFE - MUMBAI
4th Floor - A Wing, Spaces - Kanaki
Andheri-Kurla Road, Chakala
Andheri East, Mumbai - 400093
+91 93199 77269

Email:
info@insofe.edu.in

Website:
www.insofe.edu.in

Follow us on Social Media: /insofeglobal/ /school/insofe/ /INSOFEedu /insofe_global/ /InsofeVideos