Beef Data - First Draft

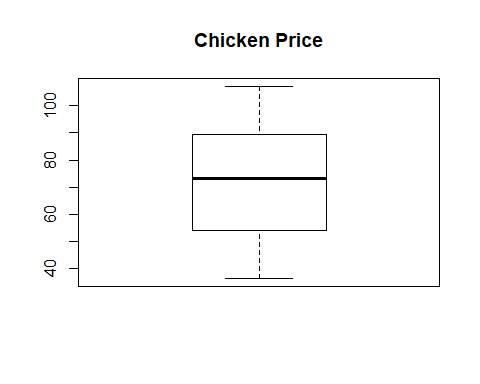
Michael Streyle, Carly Mester, Matt Foundos

October 27, 2017

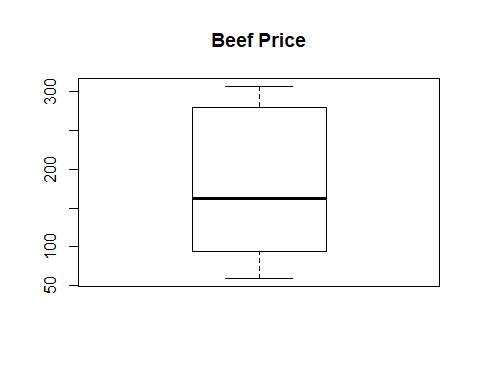
General Note: We have a lot of code that we won't end up using, but kept it included to show our exploration of the data. We will obviously be cleaning it up before the final draft.

A: Our data set examines the factors that could affect beef prices. We have 9 predictor variables(Chicken price,Year, DPI, etc..) that we will use from the start to examine how they will affect beef prices. Some of these are adjusted for inflation and some intentionally leave the adjustment out. There are 36 observations in our data set, none of which need to be converted to categorical data.

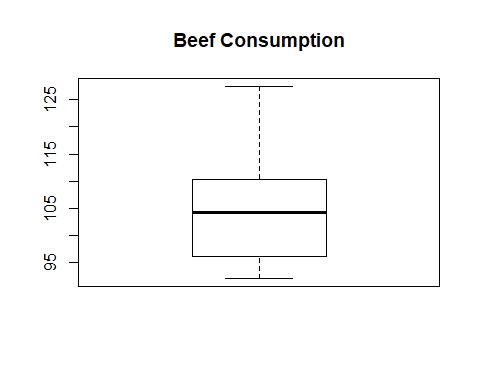
# Applied Statistics Project 1  
# Michael Streyle, Matt Foundos, Carly Mester  
  
  
beef <- read.table('http://ww2.amstat.org/publications/jse/v22n1/kopcso/BeefDemand.txt', header = TRUE)  
  
   
attach(beef)  
#creating a new variable called myRealDPI because there is an error in X.RDPI.Mean..sq  
myRealDPIsq=(RealDPI-mean(RealDPI))^2  
  
#checking distributions of each variable  
boxplot(ChickPrice, main="Chicken Price")



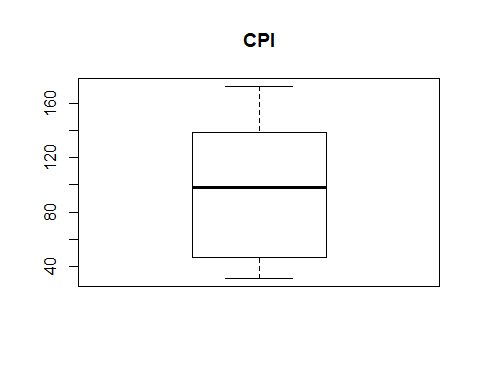
boxplot(BeefPrice, main="Beef Price")



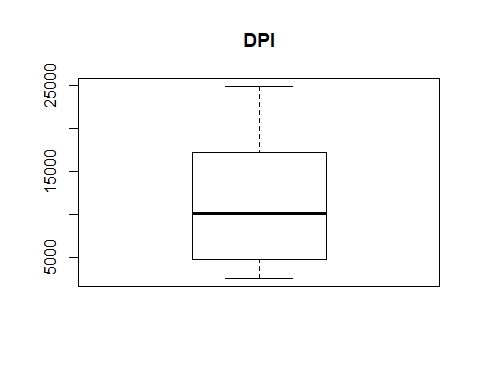
boxplot(BeefConsump, main="Beef Consumption")



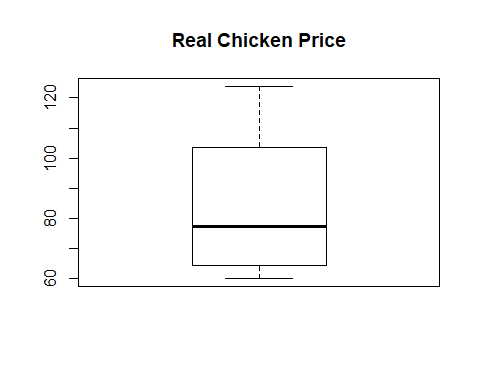
boxplot(CPI, main="CPI")



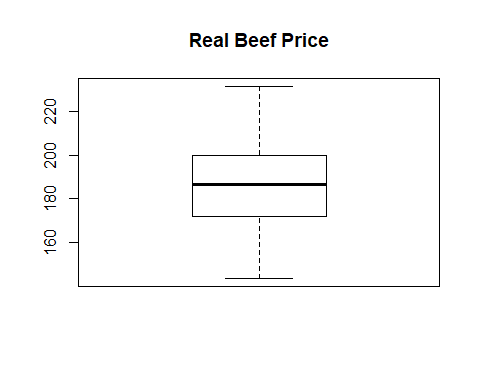
boxplot(DPI, main="DPI")



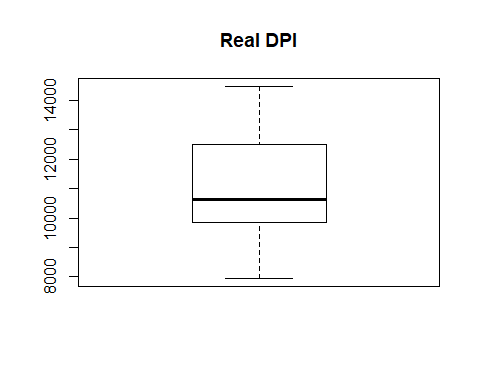
boxplot(RealChickenPrice, main="Real Chicken Price")



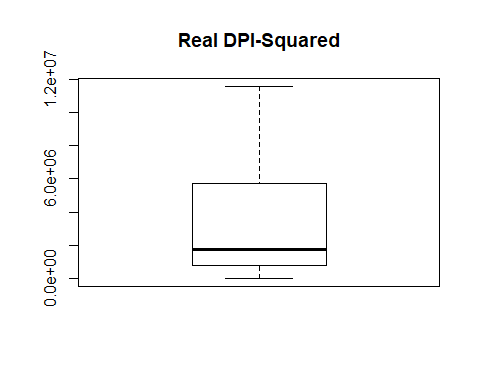
boxplot(RealBeefPrice, main="Real Beef Price")



boxplot(RealDPI, main="Real DPI")



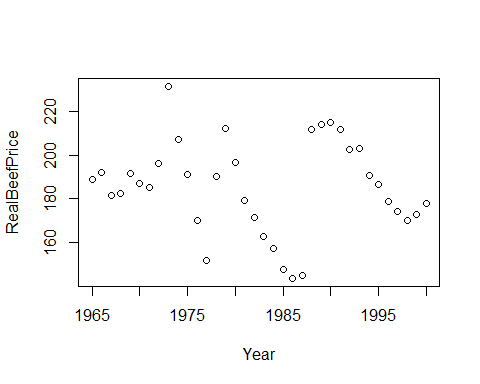
boxplot(myRealDPIsq, main="Real DPI-Squared")



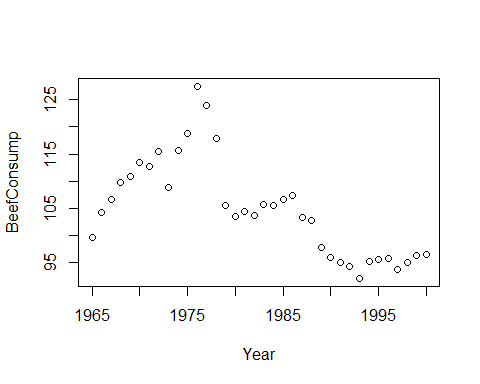
#myRealDPIsq is skewed, but since it is a derived variable, we don't know if we should mess with it

Bb: We are given a variable with an error in it (X.RDPI.Mean..sq), so we created a new corrected variable called myRealDPIsq to replace it. Next, we examined the distributions of all of our variables with seperate boxplots. We noticed that the variable we created was skewed, but since we created it we decided not to transform it any further. The other 8 variables aren't perfectly normal, however they fit well enough that we do not need to transform them. Note: A boxcox plot is conducted later, which is inconclusive.

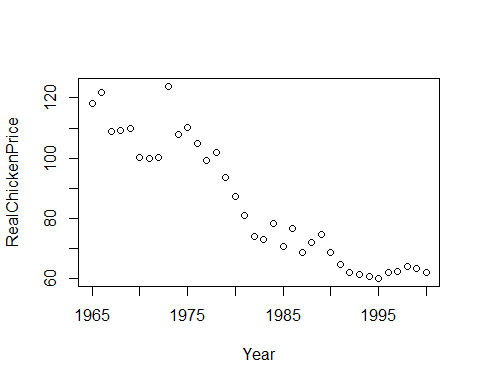
# Plot some of the data  
plot(Year,RealBeefPrice)



plot(Year,BeefConsump)



plot(Year, RealChickenPrice)



typeof(beef)

## [1] "list"

B: We plot three of our variables against the Year and see what the data looks like. The most obvious is that there seems to be a negative linear relationship between RealChickPrice and Year. The other two plots show weak nonlinear relationships and have no real pattern that can be seen throughout the entire plot.

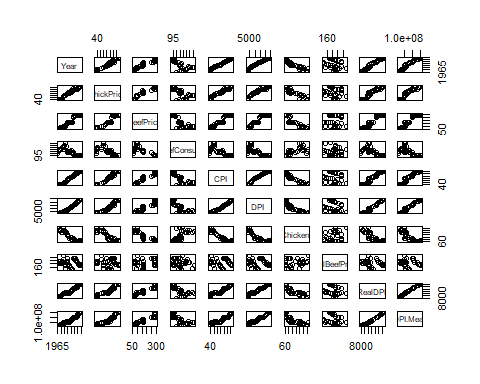
#showing CPI and DPI have a very high correlation, indicates we should not need to use them both  
cor(DPI, CPI)

## [1] 0.9907179

#shows the correlation matrix and pairs plots  
round (cor(beef), 2)

## Year ChickPrice BeefPrice BeefConsump CPI DPI  
## Year 1.00 0.98 0.96 -0.69 0.99 0.99  
## ChickPrice 0.98 1.00 0.94 -0.67 0.97 0.96  
## BeefPrice 0.96 0.94 1.00 -0.79 0.97 0.96  
## BeefConsump -0.69 -0.67 -0.79 1.00 -0.76 -0.76  
## CPI 0.99 0.97 0.97 -0.76 1.00 0.99  
## DPI 0.99 0.96 0.96 -0.76 0.99 1.00  
## RealChickenPrice -0.94 -0.90 -0.91 0.71 -0.95 -0.91  
## RealBeefPrice -0.11 -0.11 0.12 -0.17 -0.12 -0.10  
## RealDPI 0.98 0.96 0.95 -0.66 0.97 0.98  
## X.RDPI.Mean..sq 0.97 0.95 0.94 -0.67 0.96 0.98  
## RealChickenPrice RealBeefPrice RealDPI X.RDPI.Mean..sq  
## Year -0.94 -0.11 0.98 0.97  
## ChickPrice -0.90 -0.11 0.96 0.95  
## BeefPrice -0.91 0.12 0.95 0.94  
## BeefConsump 0.71 -0.17 -0.66 -0.67  
## CPI -0.95 -0.12 0.97 0.96  
## DPI -0.91 -0.10 0.98 0.98  
## RealChickenPrice 1.00 0.21 -0.87 -0.86  
## RealBeefPrice 0.21 1.00 -0.04 -0.04  
## RealDPI -0.87 -0.04 1.00 1.00  
## X.RDPI.Mean..sq -0.86 -0.04 1.00 1.00

pairs(beef)



C: First we check the correlation between CPI and DPI and we see there is a high level which indicates we probably do not need to use both variables. We then produced our correlation matrix and examine our correlations in our pair plots. Between these two tools, it is easy to see which variables correlate and which we should start examining for use from here on out.

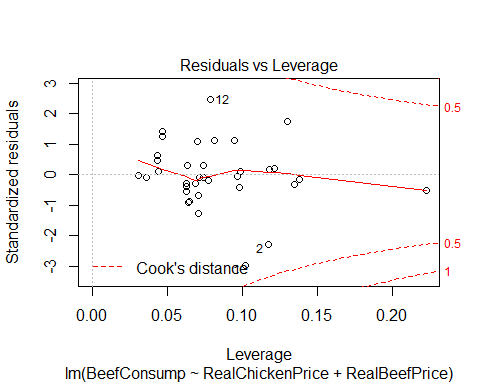
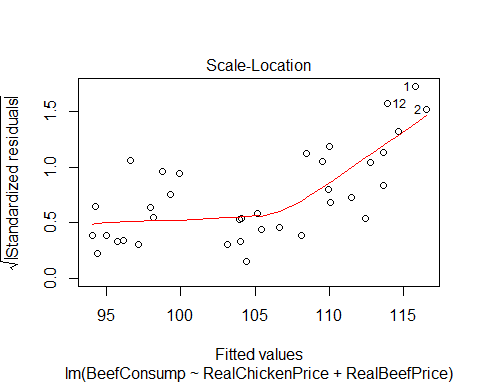
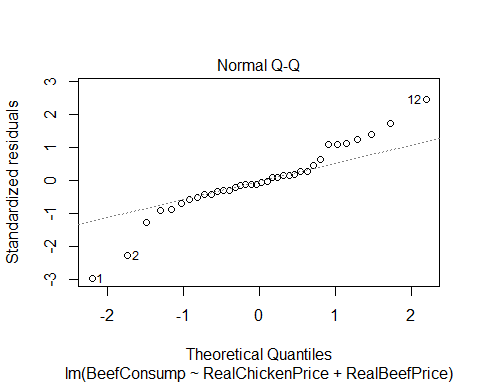
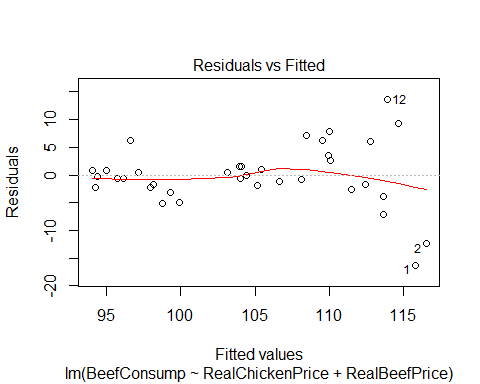
#trying to use RealChickenPrice and RealBeefPrice to model BeefConsump  
myfit <- lm (BeefConsump ~ RealChickenPrice + RealBeefPrice)  
myfit

##   
## Call:  
## lm(formula = BeefConsump ~ RealChickenPrice + RealBeefPrice)  
##   
## Coefficients:  
## (Intercept) RealChickenPrice RealBeefPrice   
## 102.7655 0.3429 -0.1459

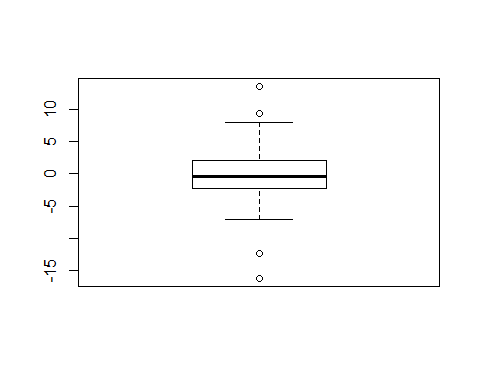
summary(myfit)

##   
## Call:  
## lm(formula = BeefConsump ~ RealChickenPrice + RealBeefPrice)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.2544 -2.2652 -0.4347 1.8816 13.6002   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 102.76552 8.85661 11.603 3.45e-13 \*\*\*  
## RealChickenPrice 0.34293 0.04802 7.142 3.50e-08 \*\*\*  
## RealBeefPrice -0.14590 0.04701 -3.104 0.00391 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.771 on 33 degrees of freedom  
## Multiple R-squared: 0.6192, Adjusted R-squared: 0.5961   
## F-statistic: 26.83 on 2 and 33 DF, p-value: 1.208e-07

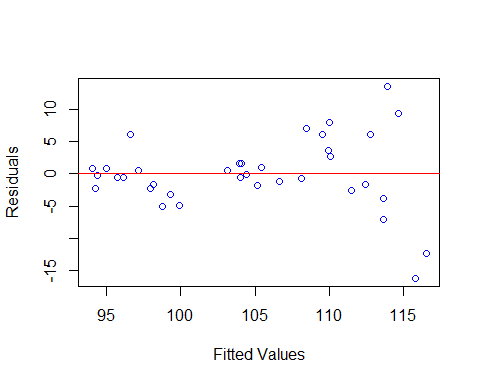
plot(myfit)



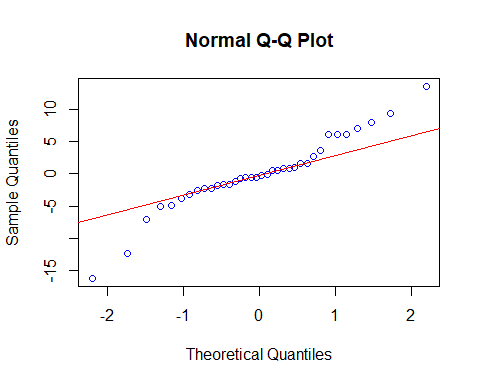
#this boxplot shows there are outliers on both ends of the data  
resid = myfit$residuals  
fitted = myfit$fitted.values  
boxplot(resid)



plot (fitted, resid, xlab="Fitted Values", ylab="Residuals", col='blue')  
abline(h=0, col='red')



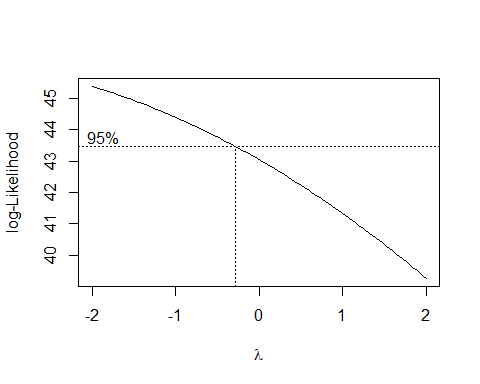
#qqplot  
  
qqnorm (resid, col='blue')  
qqline (resid, col='red')



#this indicates a transformation might be helpful.  
#need to do boxcox of myfit otherwise it is seen as a linear model  
library(MASS)

## Warning: package 'MASS' was built under R version 3.4.2

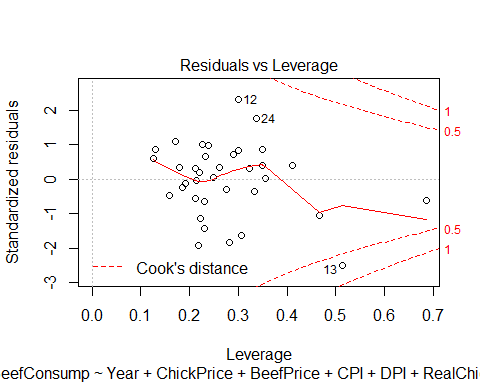
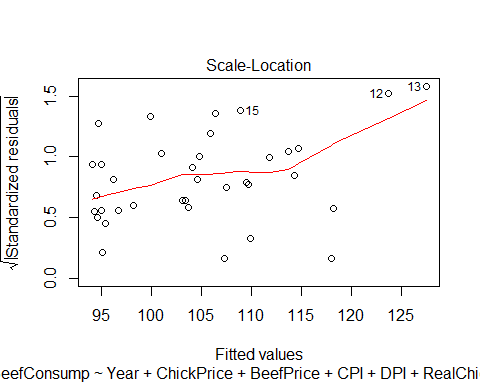
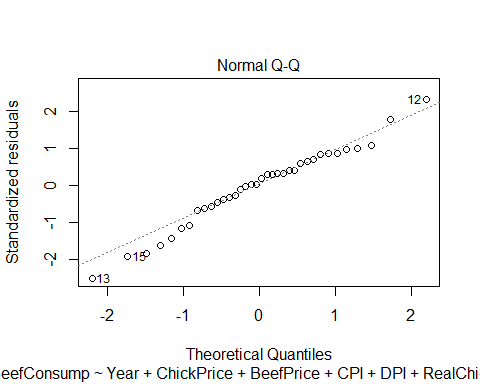
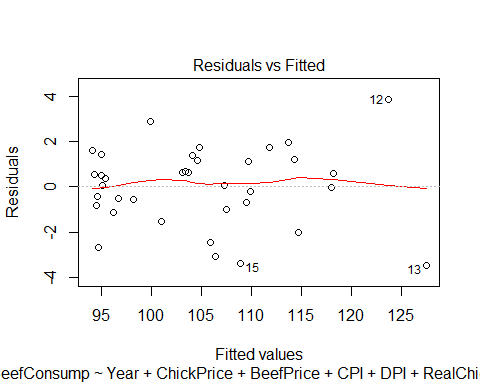
boxcox(myfit)



#boxcox plots are not conclusive

D: Here we are trying to use RealChickenPrice and RealBeefPrice to model BeefConsump. From all of our plots we have throughout this chunk of code, we can tell many things. Our data seems to be "heavily-tailed" on both ends, but none far enough to be outliers. There seems to be a non-constant variance throughout the data and it seems to be normally distributed. At the end of this chunk of code, we look to see if a transformation might be useful but our boxcox plots come back inconclusive.

#fit a model with all variables   
myfit1 <- lm(BeefConsump ~ Year + ChickPrice + BeefPrice + CPI + DPI + RealChickenPrice + RealBeefPrice + RealDPI + myRealDPIsq)  
plot(myfit1)



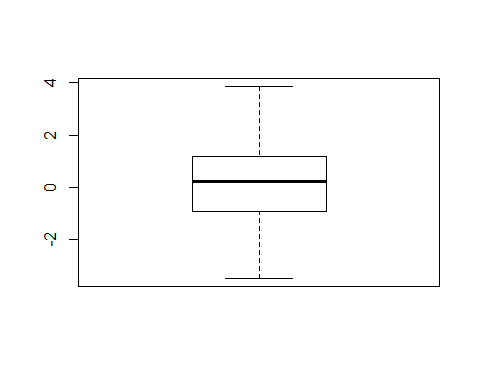
summary(myfit1)

##   
## Call:  
## lm(formula = BeefConsump ~ Year + ChickPrice + BeefPrice + CPI +   
## DPI + RealChickenPrice + RealBeefPrice + RealDPI + myRealDPIsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4656 -0.8820 0.2151 1.1653 3.8570   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.695e+03 1.863e+03 -5.205 1.96e-05 \*\*\*  
## Year 5.010e+00 9.437e-01 5.308 1.49e-05 \*\*\*  
## ChickPrice -1.783e-01 2.162e-01 -0.825 0.4171   
## BeefPrice 2.854e-01 5.810e-02 4.912 4.24e-05 \*\*\*  
## CPI -1.035e+00 6.892e-01 -1.501 0.1453   
## DPI -4.834e-03 6.016e-03 -0.804 0.4289   
## RealChickenPrice 2.390e-01 1.355e-01 1.764 0.0895 .   
## RealBeefPrice -4.116e-01 6.128e-02 -6.716 3.99e-07 \*\*\*  
## RealDPI 3.299e-03 6.629e-03 0.498 0.6229   
## myRealDPIsq 1.005e-06 1.208e-06 0.832 0.4130   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.989 on 26 degrees of freedom  
## Multiple R-squared: 0.9644, Adjusted R-squared: 0.952   
## F-statistic: 78.16 on 9 and 26 DF, p-value: < 2.2e-16

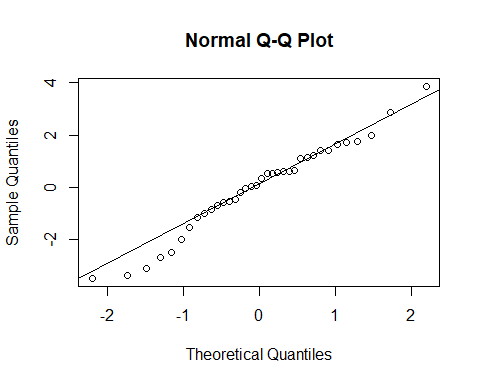
anova(myfit1)

## Analysis of Variance Table  
##   
## Response: BeefConsump  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Year 1 1386.17 1386.17 350.4102 < 2.2e-16 \*\*\*  
## ChickPrice 1 12.25 12.25 3.0957 0.09027 .   
## BeefPrice 1 615.57 615.57 155.6117 1.766e-12 \*\*\*  
## CPI 1 571.95 571.95 144.5842 4.008e-12 \*\*\*  
## DPI 1 12.24 12.24 3.0931 0.09039 .   
## RealChickenPrice 1 4.02 4.02 1.0170 0.32251   
## RealBeefPrice 1 176.66 176.66 44.6580 4.333e-07 \*\*\*  
## RealDPI 1 1.25 1.25 0.3157 0.57904   
## myRealDPIsq 1 2.74 2.74 0.6922 0.41299   
## Residuals 26 102.85 3.96   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

boxplot(myfit1$residuals)

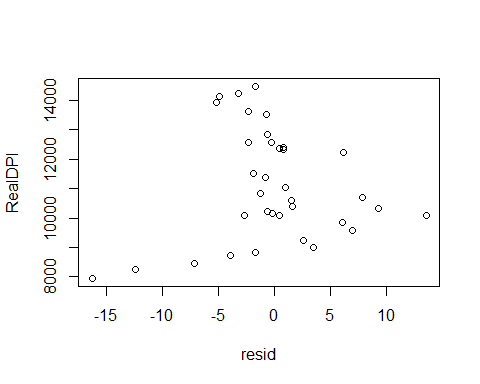


qqnorm(myfit1$residuals)  
qqline(myfit1$residuals)

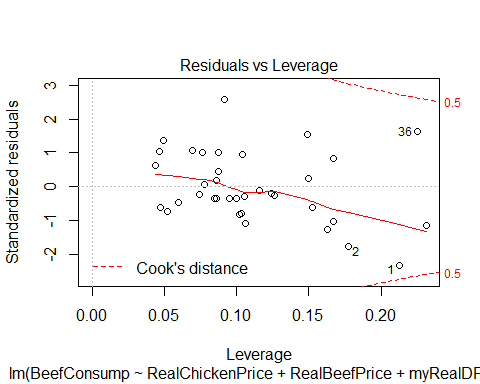
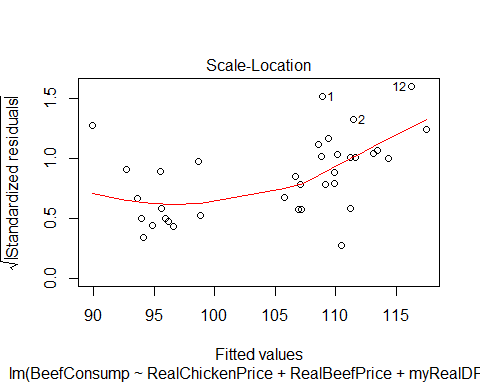
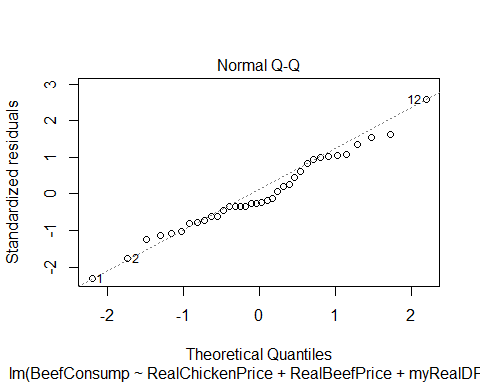
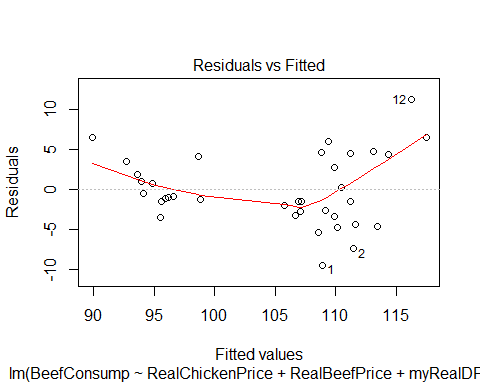


E: Here we create our myfit1, which contains all of our variables. There is a slight curvature of the residuals, however I would say it is close to linear. There also seems to be a constant variance with normally distributed residuals. We then started looking at some residual analysis and the residuals look to be normally distributed and relatively normal. With our new "myfit1" our Residual Standard Error seems to have gotten much better. The scale of our residual boxplots has decreased significantly and our Multiple R-squared has increased quite a bit as well.

#myfit3 sucks but keeping it for now  
plot(resid, RealDPI)



myfit3 <- lm(BeefConsump ~ RealChickenPrice + RealBeefPrice + myRealDPIsq)  
plot(myfit3)



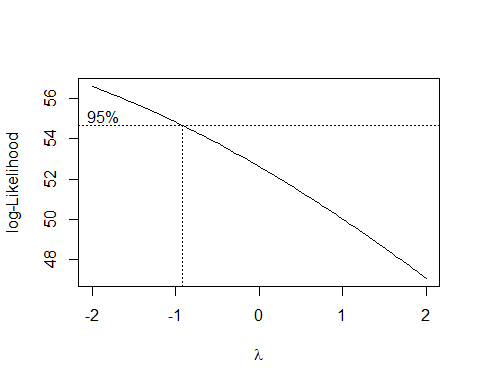
myfit3

##   
## Call:  
## lm(formula = BeefConsump ~ RealChickenPrice + RealBeefPrice +   
## myRealDPIsq)  
##   
## Coefficients:  
## (Intercept) RealChickenPrice RealBeefPrice myRealDPIsq   
## 1.075e+02 3.341e-01 -1.490e-01 -1.028e-06

summary(myfit3)

##   
## Call:  
## lm(formula = BeefConsump ~ RealChickenPrice + RealBeefPrice +   
## myRealDPIsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.424 -2.884 -1.048 3.656 11.285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.075e+02 7.144e+00 15.047 4.55e-16 \*\*\*  
## RealChickenPrice 3.341e-01 3.835e-02 8.712 5.90e-10 \*\*\*  
## RealBeefPrice -1.490e-01 3.750e-02 -3.973 0.000378 \*\*\*  
## myRealDPIsq -1.027e-06 2.305e-07 -4.457 9.55e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.603 on 32 degrees of freedom  
## Multiple R-squared: 0.765, Adjusted R-squared: 0.743   
## F-statistic: 34.73 on 3 and 32 DF, p-value: 3.519e-10

boxcox(myfit3)



F:

#what exactly is X.RDPI.Mean..sq?  
# it is The square of the difference between Inflation-adjusted Disposable Personal Income per capita and its mean  
  
  
#bp test  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(myfit1, studentize=FALSE)

##   
## Breusch-Pagan test  
##   
## data: myfit1  
## BP = 11.204, df = 9, p-value = 0.262

#the results of the bptest indicate constant variance since the p-value is 0.262 which is greater than 0.01 (my chosen level of significance)  
  
  
#lack of fit test  
  
aggregate (BeefConsump, by=list(RealBeefPrice, RealChickenPrice, RealDPI), FUN='mean')

## Group.1 Group.2 Group.3 x  
## 1 188.8889 118.23812 7952.381 99.50000  
## 2 191.9753 121.73305 8256.173 104.10000  
## 3 181.7365 108.93864 8467.066 106.50000  
## 4 182.4713 109.22129 8727.011 109.70000  
## 5 191.8256 109.81201 8825.613 110.80000  
## 6 186.8557 100.42270 8992.268 113.50000  
## 7 185.1852 100.02726 9234.568 112.70000  
## 8 196.4115 100.14691 9569.378 115.50000  
## 9 191.0781 110.33862 9834.572 118.80000  
## 10 207.0994 107.91987 9847.870 115.70000  
## 11 171.5913 74.19689 10076.684 103.70000  
## 12 231.5315 123.84456 10092.342 108.80000  
## 13 170.2988 104.92091 10094.903 127.50000  
## 14 179.1639 81.07811 10168.317 104.30000  
## 15 196.8186 87.37864 10219.660 103.40000  
## 16 151.8152 99.17492 10333.333 124.00000  
## 17 162.8298 73.09237 10381.526 105.70000  
## 18 212.1232 93.66391 10581.267 105.50000  
## 19 190.2130 101.99386 10687.117 117.90000  
## 20 157.4027 78.34456 10834.456 105.50000  
## 21 147.5571 70.63197 11023.234 106.50000  
## 22 143.6210 76.64234 11376.825 107.30000  
## 23 145.1622 68.66197 11526.408 103.30000  
## 24 211.6159 72.16399 12237.532 102.79858  
## 25 211.6924 64.63289 12309.838 94.85576  
## 26 214.2406 74.75806 12344.355 97.66680  
## 27 215.0089 68.79878 12398.623 95.79480  
## 28 203.0738 61.60554 12562.630 92.00207  
## 29 202.8570 61.95296 12570.207 94.14389  
## 30 190.8738 60.78947 12822.537 95.16543  
## 31 186.5704 60.15092 13519.685 95.54054  
## 32 178.6063 61.99490 13623.327 95.69209  
## 33 174.1641 62.42368 13901.558 93.61479  
## 34 170.0102 64.03067 14120.245 95.00751  
## 35 172.7256 63.37935 14221.489 96.16141  
## 36 177.9326 62.20674 14464.576 96.46501

# Fit the full model:  
full = lm (BeefConsump ~ 0 + as.factor(RealBeefPrice) \* as.factor(RealChickenPrice) \* as.factor(RealDPI))  
anova (myfit3, full)

## Analysis of Variance Table  
##   
## Model 1: BeefConsump ~ RealChickenPrice + RealBeefPrice + myRealDPIsq  
## Model 2: BeefConsump ~ 0 + as.factor(RealBeefPrice) \* as.factor(RealChickenPrice) \*   
## as.factor(RealDPI)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 32 678.05   
## 2 0 0.00 32 678.05

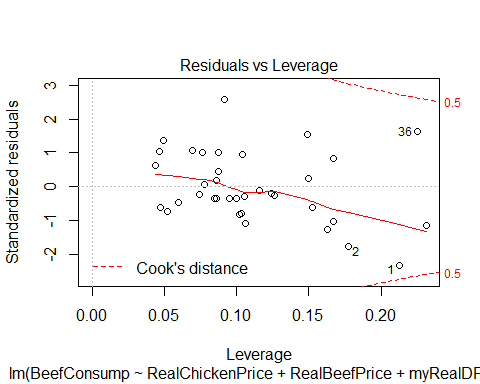
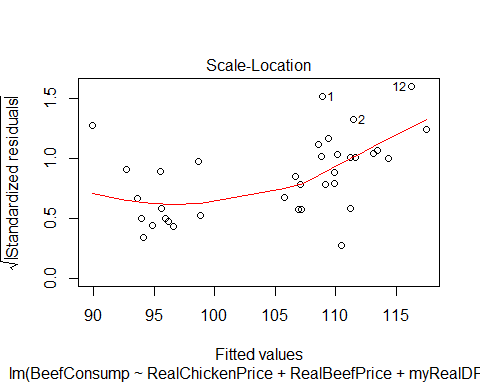
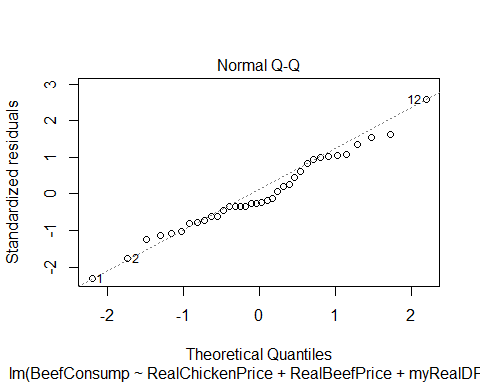
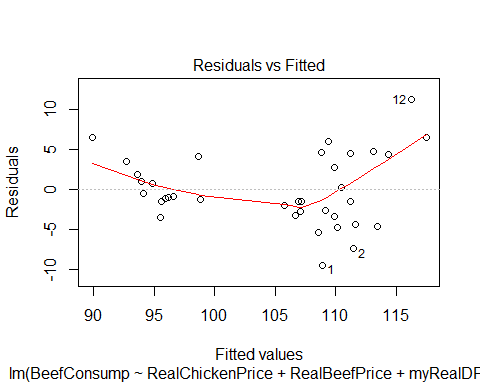
summary(myfit3)

##   
## Call:  
## lm(formula = BeefConsump ~ RealChickenPrice + RealBeefPrice +   
## myRealDPIsq)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.424 -2.884 -1.048 3.656 11.285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.075e+02 7.144e+00 15.047 4.55e-16 \*\*\*  
## RealChickenPrice 3.341e-01 3.835e-02 8.712 5.90e-10 \*\*\*  
## RealBeefPrice -1.490e-01 3.750e-02 -3.973 0.000378 \*\*\*  
## myRealDPIsq -1.027e-06 2.305e-07 -4.457 9.55e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.603 on 32 degrees of freedom  
## Multiple R-squared: 0.765, Adjusted R-squared: 0.743   
## F-statistic: 34.73 on 3 and 32 DF, p-value: 3.519e-10

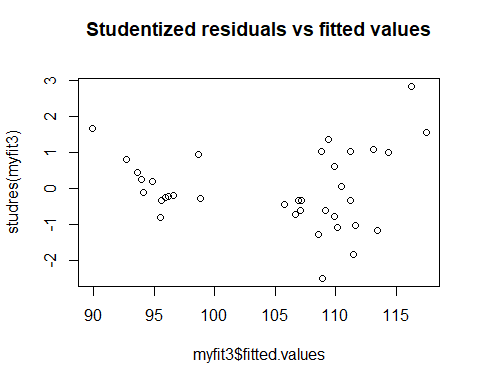
#the summary shows a p-value of 6.909e-07 which is small suggesting a good regression relation

G:

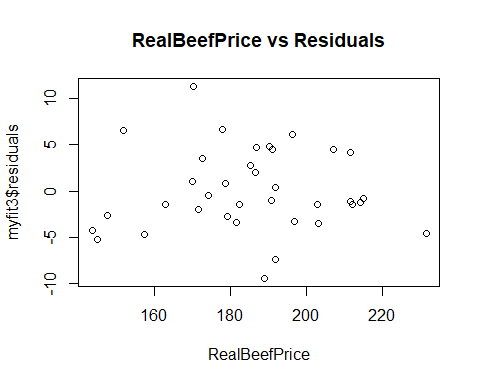
#looking at some residual diagnostics of myfit3  
  
plot(myfit3)



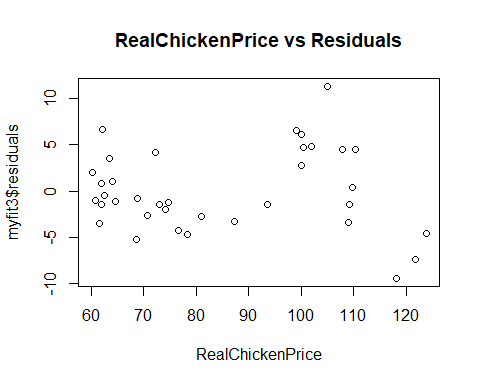
#semistudentized residuals  
plot(myfit3$fitted.values, studres(myfit3), main="Studentized residuals vs fitted values  
")



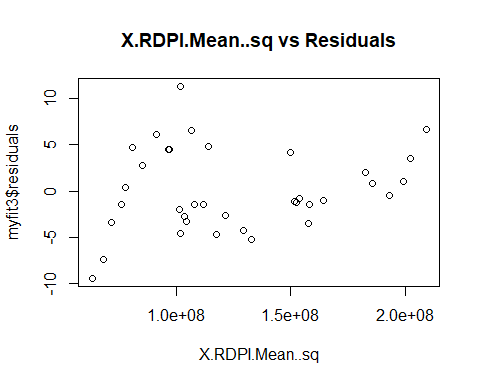
#note that none of the studentized residuals are greater than the absolute value of 4 - no outliers  
plot( RealBeefPrice, myfit3$residuals, main="RealBeefPrice vs Residuals")



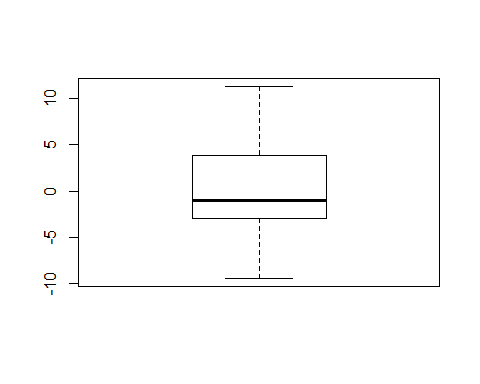
plot( RealChickenPrice, myfit3$residuals, main="RealChickenPrice vs Residuals")



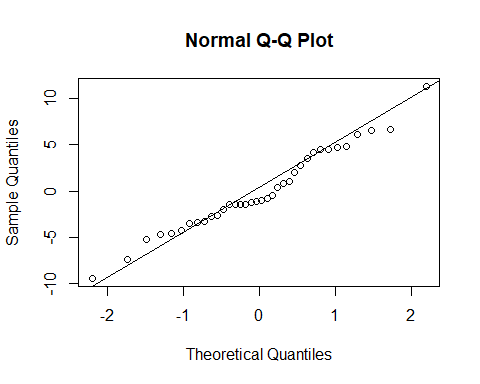
plot( X.RDPI.Mean..sq, myfit3$residuals, main="X.RDPI.Mean..sq vs Residuals")



#the X.RDPI.Mean..sq vs residuals seems interesting.   
  
boxplot(myfit3$residuals)

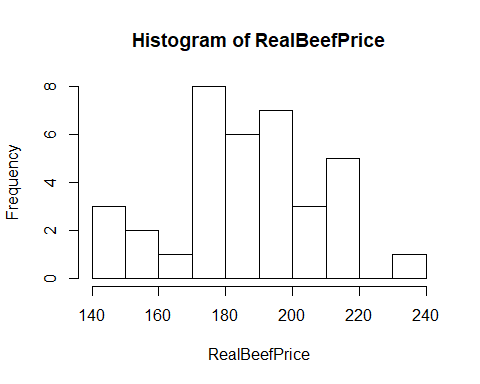


qqnorm(myfit3$residuals)  
qqline(myfit3$residuals)

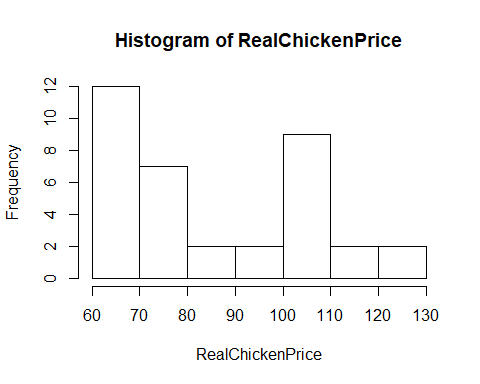


H:

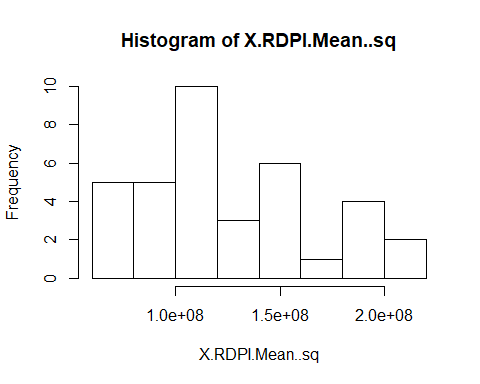
#before running the following lines, enlarge your plots window, or it might produce an error  
hist(RealBeefPrice)



hist(RealChickenPrice) #these show some interesting distributions



hist(X.RDPI.Mean..sq)



#RealChickenPrice and X.RDPI.Mean..sq show some indication of being right skewed so  
#even though other indicators don't suggest this, maybe try a log transformation (as suggested in advice doc on katie)

I:

#random lines of code  
confint(myfit1, level=0.99)

## 0.5 % 99.5 %  
## (Intercept) -1.487126e+04 -4.519363e+03  
## Year 2.387436e+00 7.632152e+00  
## ChickPrice -7.790587e-01 4.224971e-01  
## BeefPrice 1.239726e-01 4.468823e-01  
## CPI -2.949718e+00 8.804035e-01  
## DPI -2.155046e-02 1.188204e-02  
## RealChickenPrice -1.374801e-01 6.154453e-01  
## RealBeefPrice -5.818716e-01 -2.412883e-01  
## RealDPI -1.512015e-02 2.171898e-02  
## myRealDPIsq -2.351713e-06 4.361867e-06

summary(myfit1)$r.squared

## [1] 0.9643581

cor(BeefConsump, myfit1$fitted.values)^2 #same as line 112

## [1] 0.9643581

#some confidence interval stuff  
mean1 = mean(myRealDPIsq) #setting myRealDPIsq to its mean because using a random number for it seems somewhat meaningless because it is a derived variable. (Maybe I should derive its value with 50,50)  
  
#predict (myfit1, data.frame (RealBeefPrice=50, RealChickenPrice=50, X.RDPI.Mean..sq=mean1), interval='confidence', level=0.99)  
#predict (myfit3, data.frame (RealBeefPrice=50, RealChickenPrice=50, X.RDPI.Mean..sq=mean1), interval='prediction', level=0.99)

J: Backward elimination method - automatic using step function

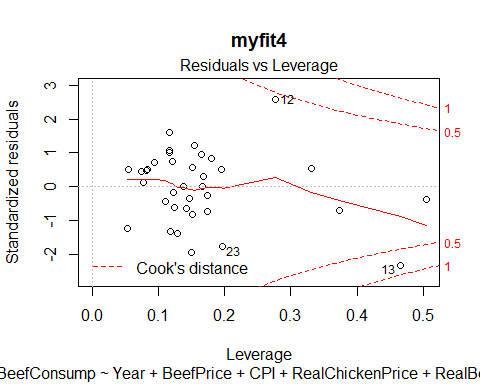
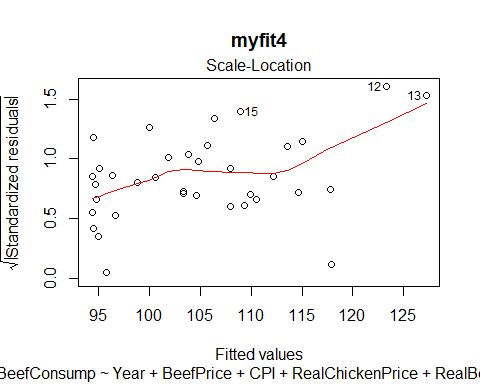
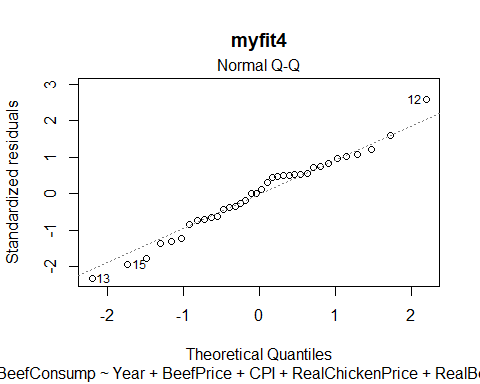
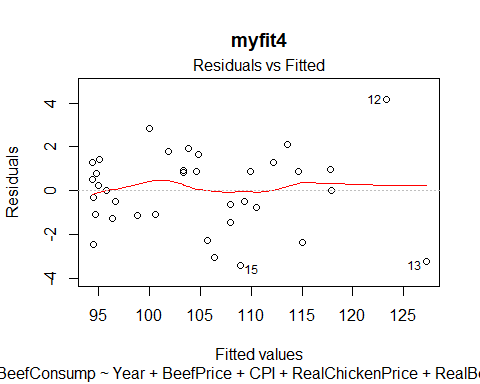
step(myfit1, direction="backward", trace=FALSE)

##   
## Call:  
## lm(formula = BeefConsump ~ Year + BeefPrice + CPI + RealChickenPrice +   
## RealBeefPrice)  
##   
## Coefficients:  
## (Intercept) Year BeefPrice CPI   
## -7955.7024 4.1470 0.2681 -1.5211   
## RealChickenPrice RealBeefPrice   
## 0.1298 -0.4026

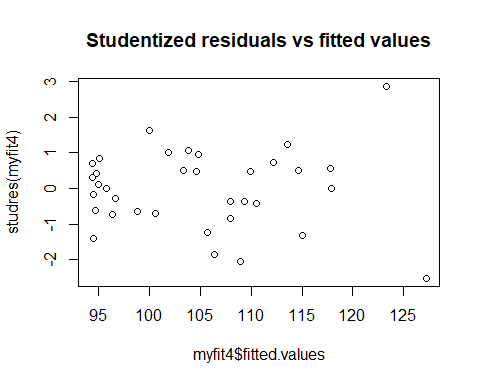
#result of backward elimination method:  
  
myfit4 <- lm(formula = BeefConsump ~ Year + BeefPrice + CPI + RealChickenPrice +   
 RealBeefPrice)  
summary(myfit4)

##   
## Call:  
## lm(formula = BeefConsump ~ Year + BeefPrice + CPI + RealChickenPrice +   
## RealBeefPrice)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4016 -1.1130 0.1231 1.0582 4.1643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.956e+03 5.221e+02 -15.236 1.16e-15 \*\*\*  
## Year 4.147e+00 2.671e-01 15.525 7.00e-16 \*\*\*  
## BeefPrice 2.681e-01 5.279e-02 5.077 1.87e-05 \*\*\*  
## CPI -1.521e+00 1.162e-01 -13.095 6.11e-14 \*\*\*  
## RealChickenPrice 1.298e-01 5.980e-02 2.170 0.0381 \*   
## RealBeefPrice -4.026e-01 5.715e-02 -7.045 7.86e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.899 on 30 degrees of freedom  
## Multiple R-squared: 0.9625, Adjusted R-squared: 0.9563   
## F-statistic: 154.1 on 5 and 30 DF, p-value: < 2.2e-16

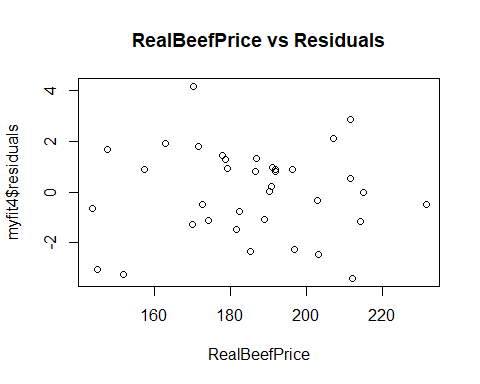
plot(myfit4, main='myfit4')



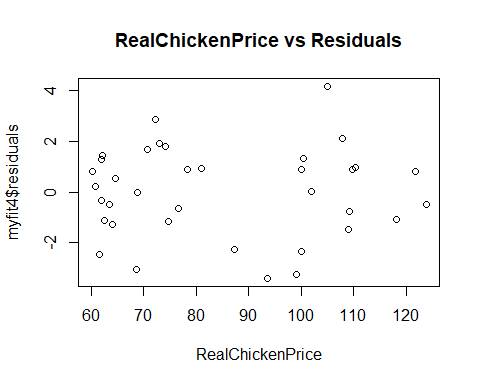
#semistudentized residuals  
plot(myfit4$fitted.values, studres(myfit4), main="Studentized residuals vs fitted values")



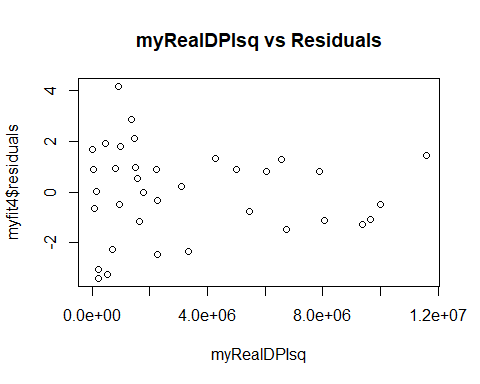
#note that none of the studentized residuals are greater than the absolute value of 4 - no outliers  
plot( RealBeefPrice, myfit4$residuals, main="RealBeefPrice vs Residuals")



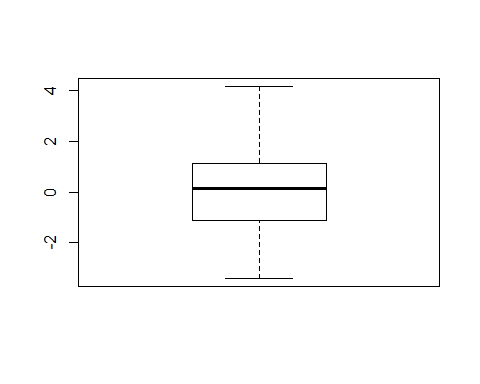
plot( RealChickenPrice, myfit4$residuals, main="RealChickenPrice vs Residuals")



plot( myRealDPIsq, myfit4$residuals, main="myRealDPIsq vs Residuals")



boxplot(myfit4$residuals)



qqnorm(myfit4$residuals)  
qqline(myfit4$residuals)

