Tree Regression of the Demand for Beef in the United States

Michael Streyle, Carly Mester, Matt Foundos

November 11, 2017

# Set global figure size  
knitr::opts\_chunk$set(fig.width=6, fig.height=3.5, echo = TRUE, message=FALSE)

## Introduction

First, we import our beef data straight from the web address and attach it. The data has 9 predictor variables (Chicken price,Year, DPI, etc..) that we will use to analyze how they affect the price of beef. Some of the variables are adjusted for inflation and some intentionally leave the inflation adjustment out. There are 36 observations from 1965-2000 in the data set, none of which need to be converted to categorical data.

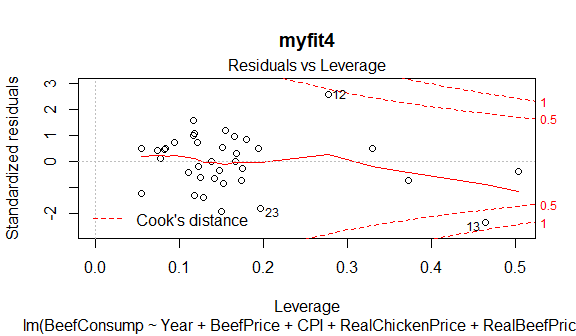
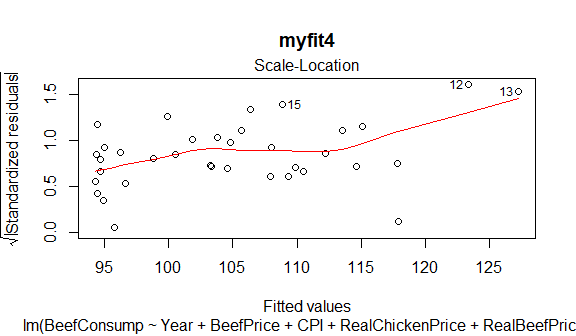
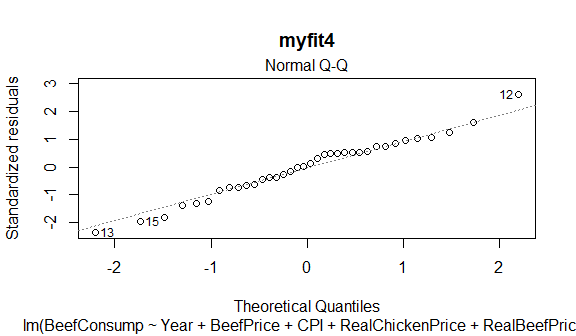
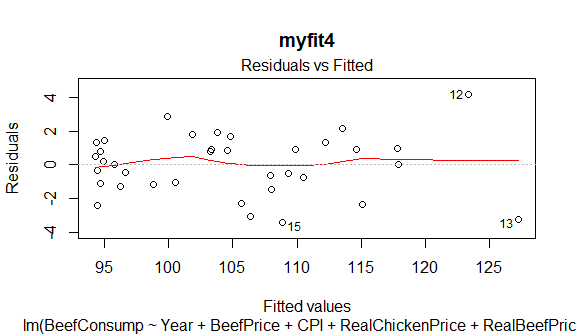
beef <- read.table('http://ww2.amstat.org/publications/jse/v22n1/kopcso/BeefDemand.txt', header = TRUE)  
  
  
attach(beef)  
#creating a new variable called myRealDPI because there is an error in X.RDPI.Mean..sq  
myRealDPIsq=(RealDPI-mean(RealDPI))^2

The final model from our Linear Regression Analysis is as follows:

fm <- lm(formula = BeefConsump ~ Year + BeefPrice + CPI + RealChickenPrice +   
 RealBeefPrice)  
summary(fm)

##   
## 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(fm, main='myfit4')



## Analysis Method - Recursive Partitioning

Start with modeling BeefConsump vs all predictors.

#This code demonstrates recursive partition analysis  
library (rpart)  
tree1 = rpart (BeefConsump ~ Year + ChickPrice + BeefPrice + CPI + DPI + RealChickenPrice +  
 RealBeefPrice + myRealDPIsq, maxdepth=10)  
print (tree1$cptable)

## CP nsplit rel error xerror xstd  
## 1 0.5854851 0 1.0000000 1.0615016 0.2408794  
## 2 0.1431249 1 0.4145149 0.5175737 0.1506503  
## 3 0.0100000 2 0.2713900 0.4097589 0.1153329

Find the tree with the smallest xerror:

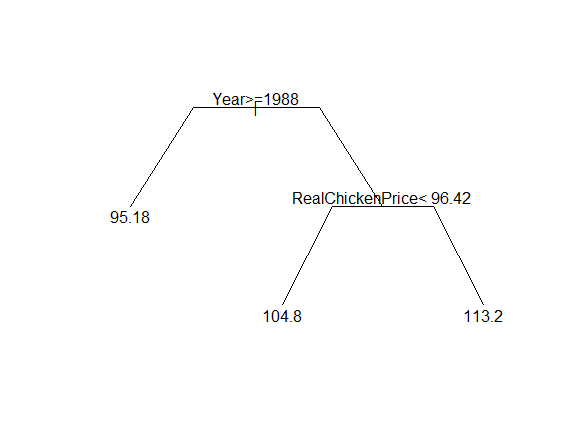
opt1 = which.min (tree1$cptable [,"xerror"])  
opt1

## 3   
## 3

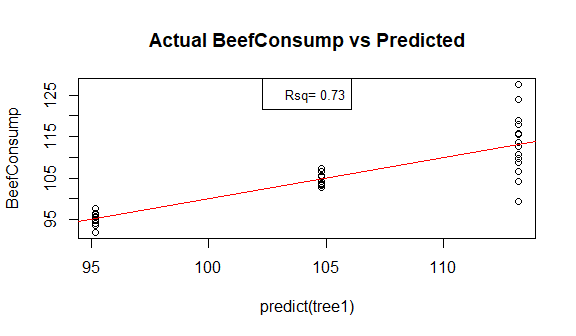
Plot the tree:

This tree tells us that Year is the single most important predictor. If the Year is greater or equal to 1988, then the average BeefConsumption is 95.18. If the Year is less than 1988 then there is an interaction between Year and RealChickenPrice. For years less than 1988 and RealChickenPrice less than 96.42, then the average BeefConsumption is 104.8. Likewise, if a year is less than 1988, and ReaLChickenPrice is greater than 96.42, then the average BeefConsumption is 113.2. In general older years have higher BeefComsumption.

par (mfrow=c(1,1))  
plot(tree1, uniform = TRUE, margin = 0.1, branch = 0.5,   
 compress = TRUE)  
text(tree1)

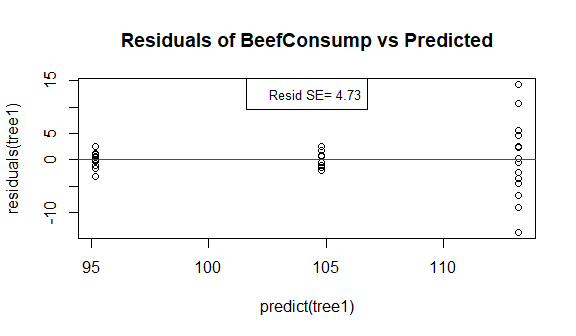
 Plot BeefConsumption vs predicted.

plot (predict(tree1), BeefConsump, main="Actual BeefConsump vs Predicted")  
abline (0, 1, col='red')  
rsq1 = cor (predict(tree1), BeefConsump)^2  
legend ('top', c(paste("Rsq=", round (rsq1, 2))), cex=0.8)



Residual plot:

plot (predict(tree1), residuals(tree1), main="Residuals of BeefConsump vs Predicted")  
abline (0, 0, col='red')  
resid.se = sd (residuals (tree1))  
legend ('top', c(paste ("Resid SE=", round (resid.se, 2))), cex=0.8)



In our linear regression modeling, we did not need to do any transformations on BeefConsumption, so it is not neccesary here either, at least not for the purposes of comparing Tree Regression vs Linear Regression.

# Summary

The for the BeefConsumption tree is 0.73, which is much lower than the for the final linear regression model, (adjusted R-squared) 0.9563.

The residual SE for the tree is 4.73, which is significantly higher than the residual SE for the final linear regression model, 1.899, both are in the original BeefConsump units.

For this dataset, multiple linear regression seems to be a much better way to model BeefConsumption, however, Tree Regression wasn't particularily awful either and seemed much more straightforward and faster. Although Tree Regression wasn't great, it still might be useful with this type of data if the only variables that can be collected are Year, and RealChickenPrice. It is somewhat astonishing to us that Tree Regression can model BeefConsumption somewhat accurately with only two variables.