Topics in Regression

Rob McCulloch

1. Understanding Multiple Regression
2. Regression Model Assumptions
3. Residual Plots
4. Non Linearity
    4.1. Problem: Quadratic Fit to the OJ Data
5. Variable Interaction
    5.1. Problem: Nbhd Size Interaction
6. The Log, Outliers and Standardized Residuals
    6.1. Problem: Log the OJ Data
7. Trees
    7.1. Problem: Midcity House Data Tree

4.1. Problem: Quadratic Fit to the OJ Data

Get the data OJ.csv from the webpage.

A chain of gas station convenience stores was interested in the dependency between price of and Sales for orange juice... They decided to run an experiment and change prices randomly at different locations.

(a)

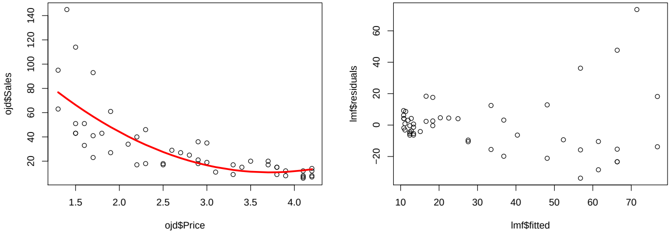
Plot Price vs. Sales.
Clearly the relationship is not linear!!

Plot the fitted values vs. residuals for the linear regression of Sales on Price and Price squared.

What does the residual plot tell us about the appropriateness of the quadratic model?

Solution

(a)



The quadratic fit would certainly be an improvement over a linear fit but the residual plot suggest that there is still some nonlinearity not captured and a non-constant variance.

R code:

```
ojd = read.csv("OJ.csv")

plot(ojd$Price,ojd$Sales)

ojd$P2 = ojd$Price^2

lmf = lm(Sales~.,ojd)

par(mfrow=c(1,2))

plot(ojd$Price,ojd$Sales)
oo = order(ojd$Price)
lines(ojd$Price[oo],lmf$fitted[oo],col="red",lwd=3)

plot(lmf$fitted,lmf$residuals)
```

5.1. Problem: Nbhd Size Interaction

Here is the R output for the fit of the model:

$$price = \beta_0 + \beta_1 size + \beta_2 n3 + \epsilon$$

where n3 is a dummy for neighborhood 3.

```
Call:
lm(formula = price ~ size + n3, data = ddf)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.396  -9.610  -1.762   8.778  38.551 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   18.153     13.574   1.337    0.184    
size          50.675      6.852   7.396 1.78e-11 ***
n3            35.699      3.137  11.379  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 15.81 on 125 degrees of freedom
Multiple R-squared:  0.659,Adjusted R-squared:  0.6536 
F-statistic: 120.8 on 2 and 125 DF,  p-value: < 2.2e-16
```

In the notes we fit the regression:

$$price = \beta_0 + \beta_1 size + \beta_2 d1 + \beta_3 d2 + \epsilon$$

where d1 and d2 are dummies for neighborhoods 1 and 2.

(a)

What is the interpretation of the model having size and n3?

Based on the regression outputs, how does the model with n3 compare to the model with d1 and d2?

(b)

Let's stick with the model having size and n3 and see if the slope should depend on the neighborhood.

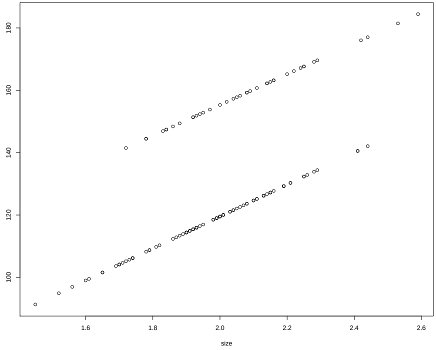
Let's fit the model:

$$price = \beta_0 + \beta_1 size + \beta_2 n3 + \beta_3 size \times n3 + \epsilon$$

Here is the regression output where n3size = n3 $\times$ size.

```
Call:
lm(formula = price ~ ., data = ddf)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.411  -9.770  -1.701   8.942  38.579 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   16.967     16.355   1.037    0.302    
size          51.278      8.275   6.197 7.81e-09 ***
n3            39.692     30.611   1.297    0.197    
n3size        -1.952     14.887  -0.131    0.896    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 15.88 on 124 degrees of freedom
Multiple R-squared:  0.6591,Adjusted R-squared:  0.6508 
F-statistic:  79.9 on 3 and 124 DF,  p-value: < 2.2e-16 
```

Here is the plot of the fit:



Do we need the interaction term in the model?

Solution

(a)

The model with with size and n3 lumps neighborhoods 1 and 2 together.

The $\hat{\sigma}$ (15.26 and 15.81) and the R^2 (.685 and .66) are not very different. Suggests we could just use the n3 dummy.
(b)

Both the ouput and the plot suggest we don't need the interaction term. The simple linear model seems ok.

```
## read in data and change compute price and size in thousands
hd = read.csv("midcity.csv")
price = hd$Price/1000
size = hd$SqFt/1000

## make dummy and interaction, but in ddf data.frame
n3 = as.numeric(hd$Nbhd==3)
ddf = data.frame(price,size,n3,n3size=n3*size)

## reg with size,n3,n3*size
lmf = lm(price~.,ddf)
print(summary(lmf))
plot(size,lmf$fitted)

## reg with size and n3
lmf1 = lm(price~size+n3,ddf)
print(summary(lmf1))
```

6.1. Problem: Log the OJ Data

Get the data OJ.csv from the webpage.

A chain of gas station convenience stores was interested in the dependency between price of and Sales for orange juice... They decided to run an experiment and change prices randomly at different locations.

(a)

Plot Price vs. Sales and log(Price) vs. log(Sales).

What does this say about using linear regression to relate Sales to Price??

(b)

Run the regression of log(Sales) on log(Price).

Plot the residuals vs. the fitted values.
What does this tell you?

Plot the standardized residuals vs. the fitted values.
Any outliers?

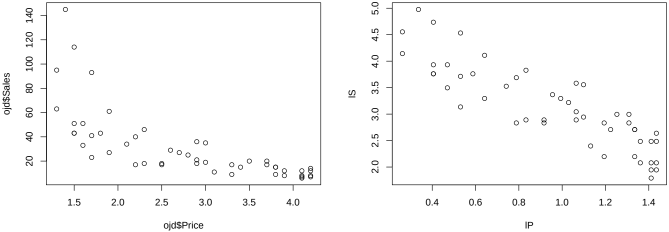
(c)

Run the regression of log(Sales) on log(Price).

What is your prediction for sales give price=3.0?

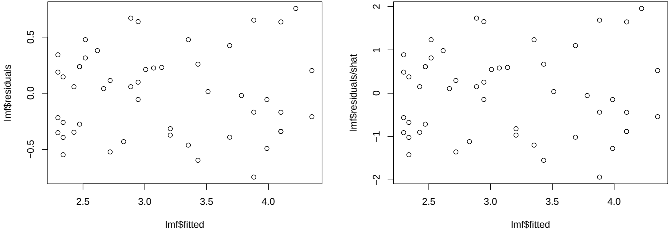
Solution

(a)



Logged data looks much better.

(b)



No obvious pattern or outliers, looks good!!!

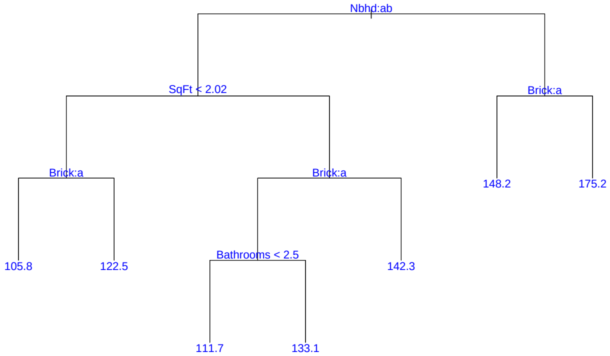
(c)

(i) log the price of 3.
(ii) plug the value from (i) into reg.
(iii) exponentiate result from (ii).

```
Call:
lm(formula = lS ~ lP, data = ddf)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7463 -0.3399  0.0279  0.2358  0.7547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.812      0.148   32.50  < 2e-16 ***
lP            -1.752      0.144  -12.17 2.77e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3858 on 48 degrees of freedom
Multiple R-squared:  0.7553,Adjusted R-squared:  0.7502
F-statistic: 148.2 on 1 and 48 DF,  p-value: 2.773e-16


> lp = log(3.0)
> lp
[1] 1.098612
> ls = 4.812 -1.752*lp
> ls
[1] 2.887231
> exp(ls)
[1] 17.94356
```

7.1. Problem: Midcity House Data Tree

Here is a tree fit to the Midcity Housing data having 7 bottom nodes (leaves).



Remember, for a categorical variable a means the first level and b means the second level and so on. So, for Nbhd, (a, b, c) corresponds to $(1, 2, 3)$ and for Brick a to No and b to Yes.

(a)

Using the tree, what price would you predict for non-brick house in Neibhborhood c=3 ?

(b)

According to the tree, what seems to be the best neighborhood?

(c)

According to the tree, what kind of house has the lowest price?

Solution

(a) 148.2

(b) c=3, the right side of three has higher prices than the left.

(c) A house in Nbhds 1 or 2 (ab), with size less than 2.02 and not made of brick.

```
hd = read.csv("midcity.csv")
hd$Nbhd = as.factor(hd$Nbhd)
hd$SqFt = hd$SqFt/1000
hd$Price = hd$Price/1000

library(tree)

#first get a big tree using a small value of mindev
temp = tree(Price~.,data=hd,mindev=.0001)
cat('first big tree size: \n')
print(length(unique(temp$where)))

#then prune it down to one with 7 leaves
hd.tree=prune.tree(temp,best=7)
cat('pruned tree size: \n')
print(length(unique(hd.tree$where)))

par(mfrow=c(1,1))

#plot the tree
plot(hd.tree,type="uniform")
text(hd.tree,col="blue",label=c("yval"),cex=1.2)
```