Homework for Section 5

Rob McCulloch

August 16, 2020

1 Homework for Section 5

1.1 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.

```
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:

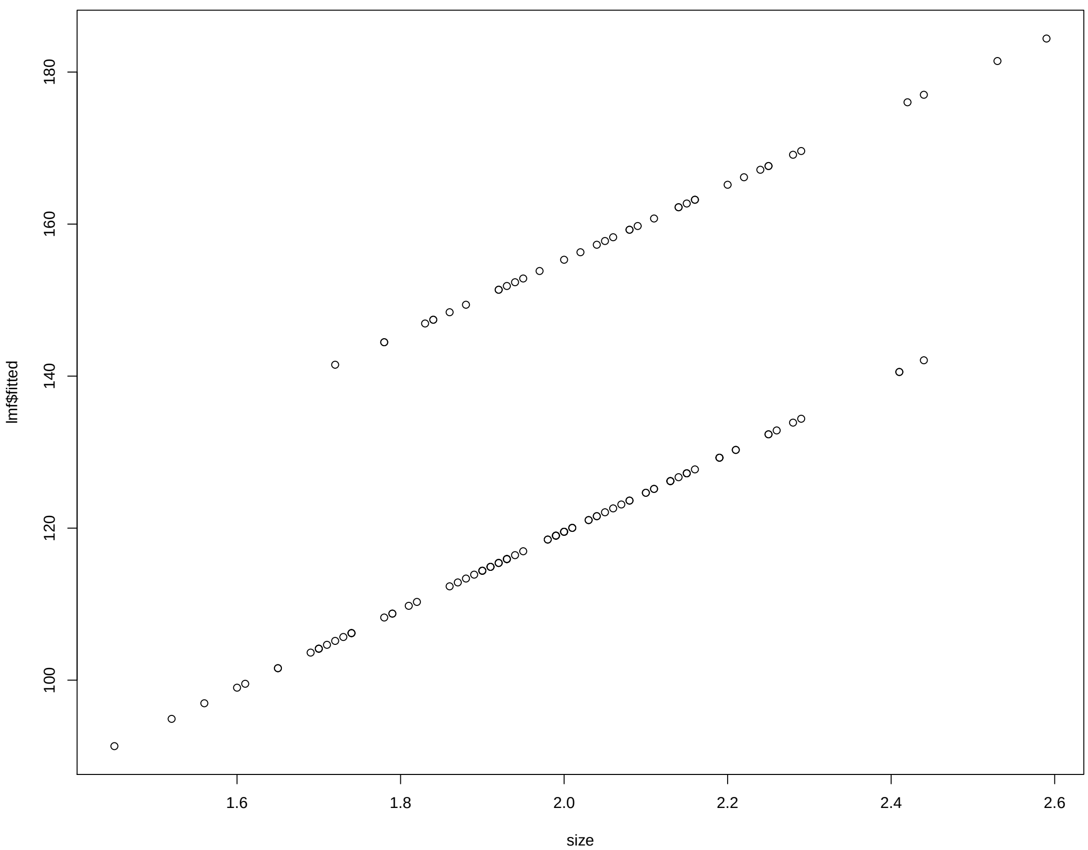
$$price = \beta_0 + \beta_1 size + \beta_2 n3 + \beta_3 size \times n3 + \epsilon$$

Here is the regression output where n3size = n3 × size.

```
            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.

1.2 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).

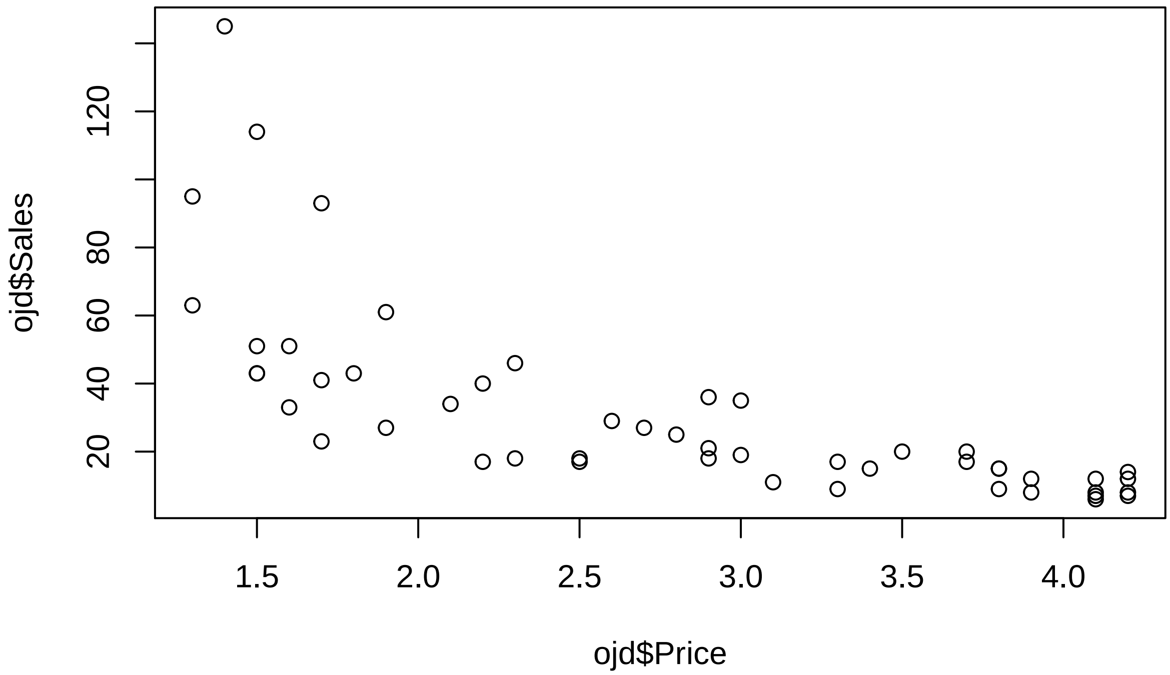
What is your prediction for sales give price=3.0?

Solution

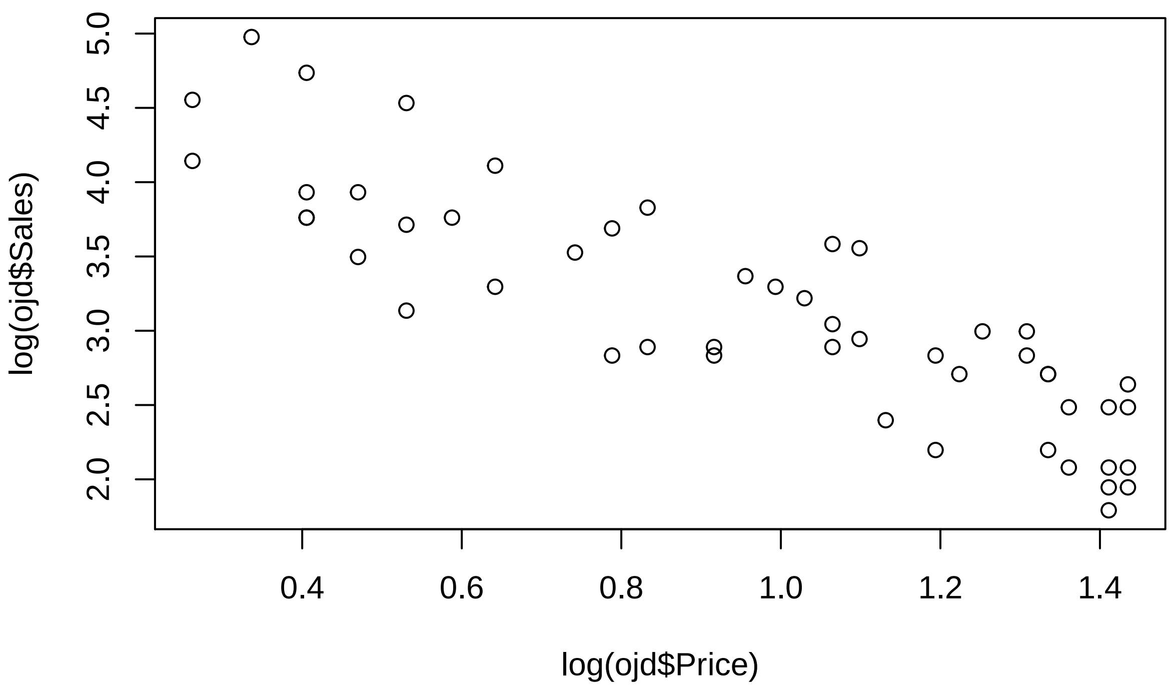
(a)

```
ojd = read.csv("http://www.rob-mcculloch.org/data/OJ.csv")
```

```
plot(ojd$Price,ojd$Sales)
```




```
plot(log(ojd$Price),log(ojd$Sales))
```



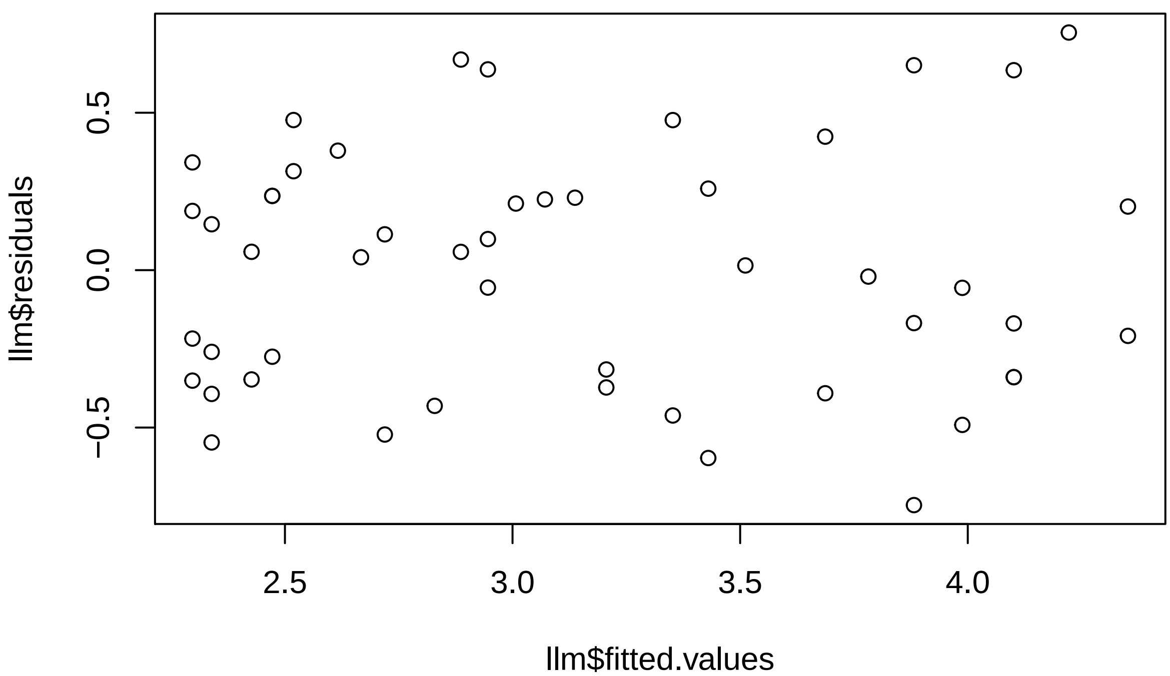
Logging both Price and Sales really helps.

(b)

```
ldf = data.frame(lPrice = log(ojd$Price), lSales = log(ojd$Sales))
llm = lm(lSales~lPrice,ldf)
summary(llm)
```

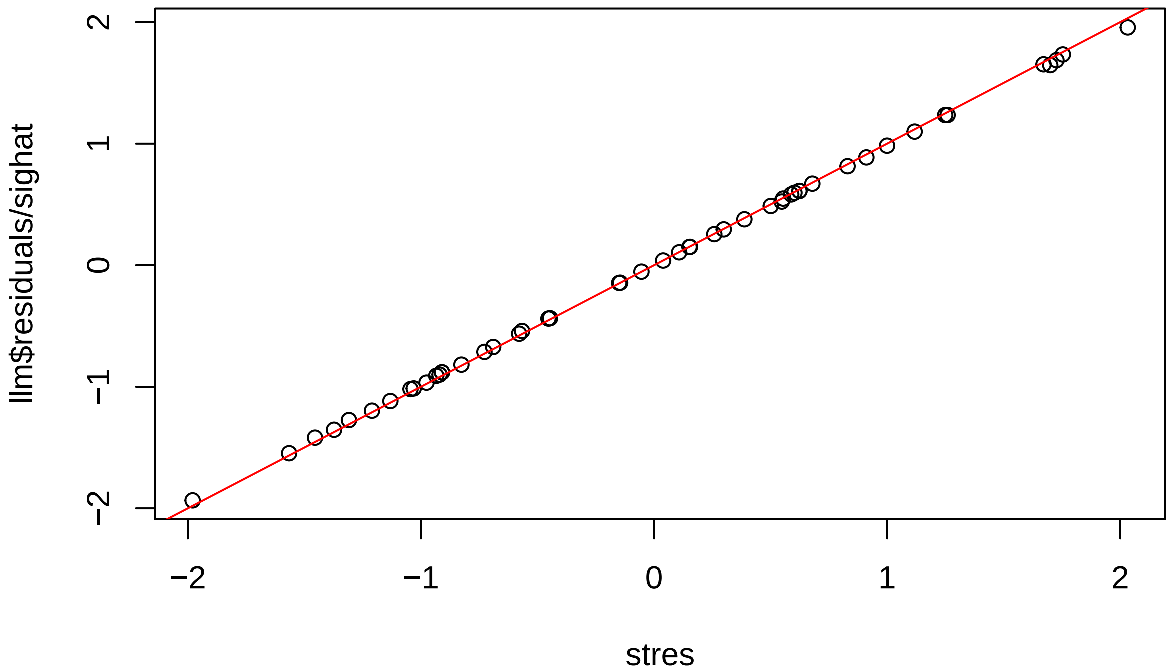
```
##
## Call:
## lm(formula = lSales ~ lPrice, data = ldf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7463 -0.3399  0.0279  0.2358  0.7547
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.812      0.148   32.50  < 2e-16 ***
## lPrice        -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
```

```
plot(llm$fitted.values,llm$residuals)
```



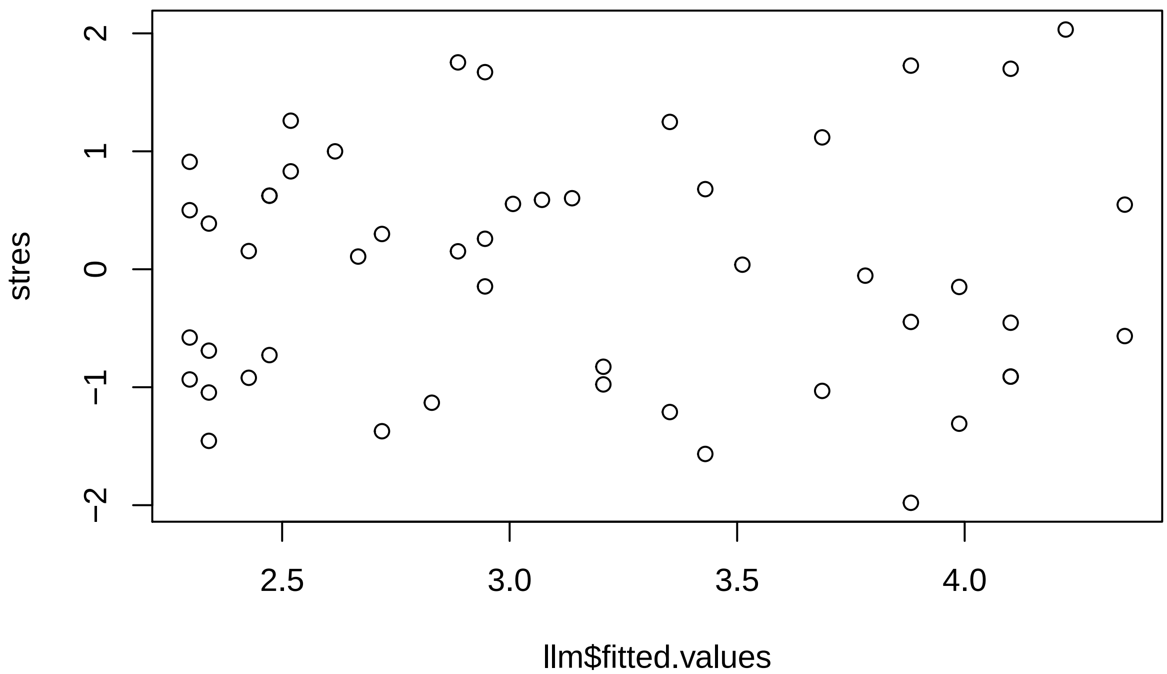
To get the standardized residuals in `R` we can use the `rstandard` function.
This basically devides the resids by $\hat{\sigma}$ as discussed in the notes.

```
stres = rstandard(llm)
sighat = summary(llm)$sigma
plot(stres,llm$residuals/sighat)
abline(0,1,col="red")
```



Now let's plot:

```
plot(llm$fitted.values,stres)
```



Do not see any outliers or obvious pattern !!

(c)

```
lshat = llm$coef[1] + llm$coef[2]*log(3.0)
shat = exp(lshat)
cat("plug in prediction for Sales is ",shat,"\n")
```

```
## plug in prediction for Sales is  17.92966
```

1.3 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.

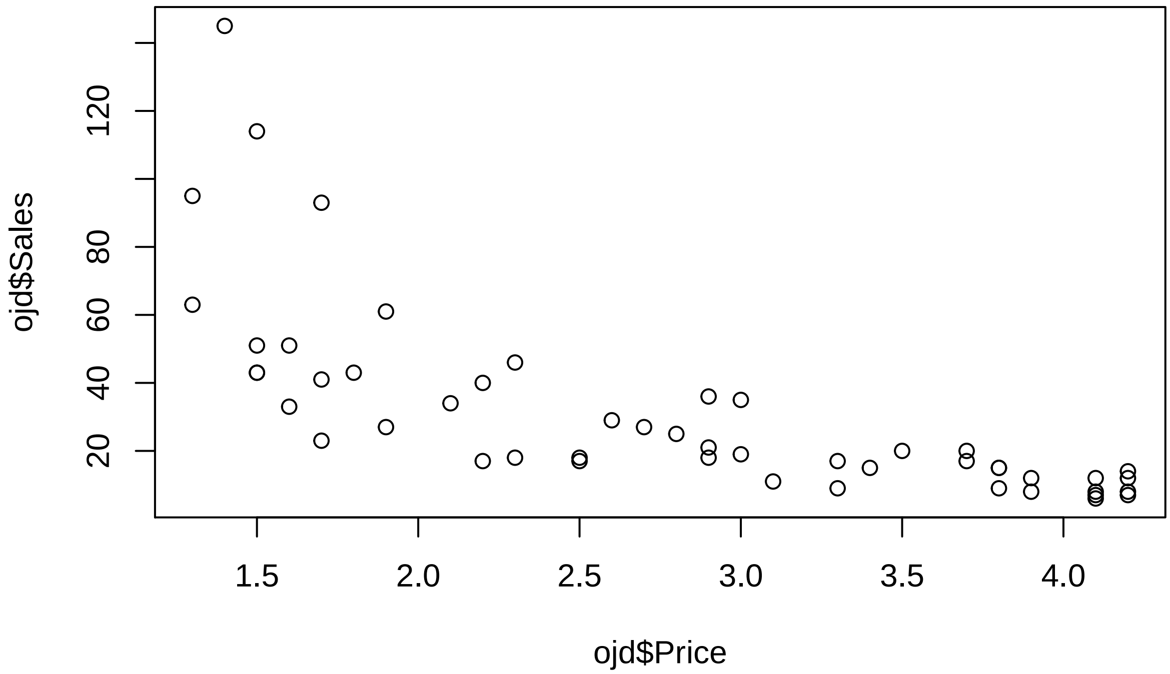
Plot Price vs. Sales.
Clearly the relationship is not linear!!

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

```
ojd = read.csv("http://www.rob-mcculloch.org/data/OJ.csv")
plot(ojd$Price,ojd$Sales)
```

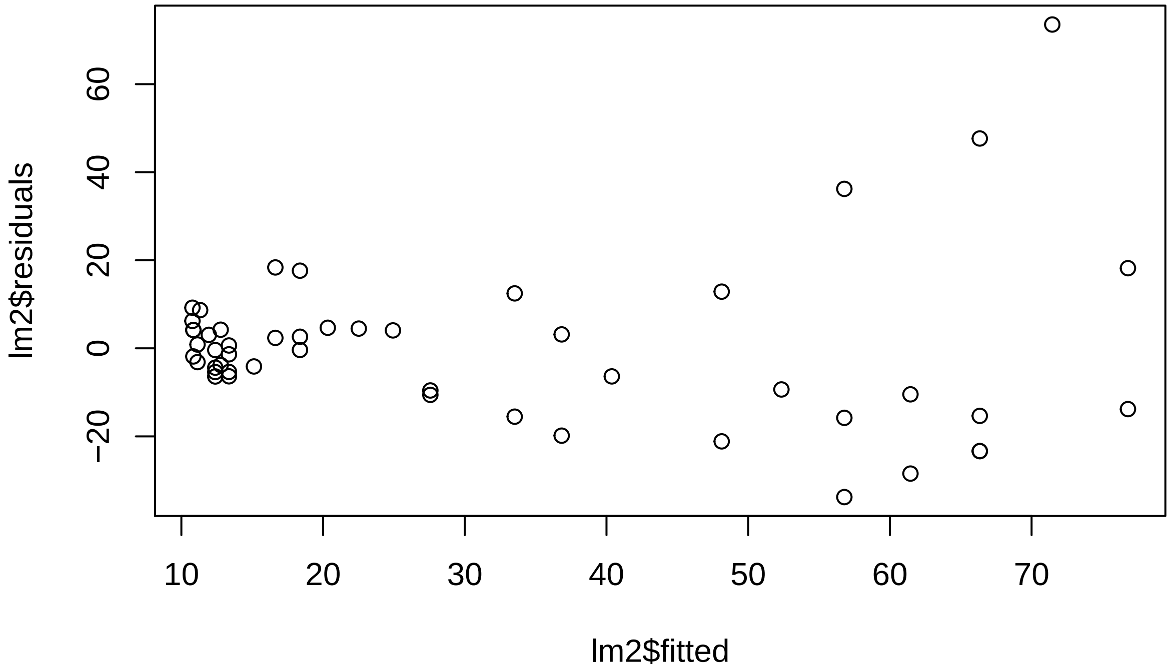


Not linear! Plausible a quadratic fit might work.

```
ojd$PriceSq = ojd$Price^2
lm2 = lm(Sales~.,ojd)
summary(lm2)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = ojd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.786  -9.514  -0.876   4.417  73.541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  166.740     25.914   6.434  5.9e-08 ***
## Price        -83.827     20.306  -4.128 0.000148 ***
## PriceSq       11.264      3.605   3.125 0.003048 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.49 on 47 degrees of freedom
## Multiple R-squared:  0.6003, Adjusted R-squared:  0.5833
## F-statistic: 35.29 on 2 and 47 DF,  p-value: 4.38e-10
```

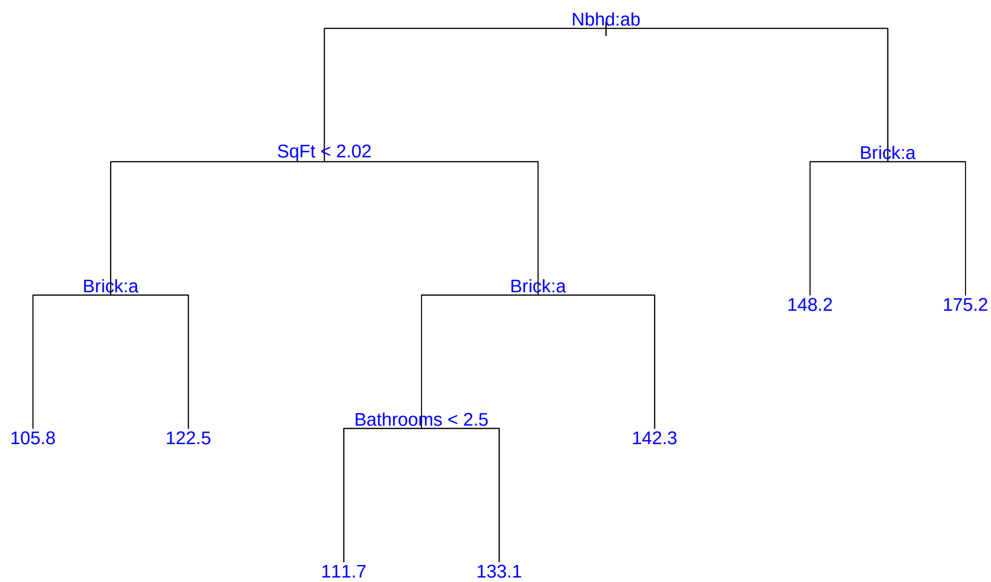
```r
plot(lm2$fitted,lm2$residuals)
```



Hmm. Not so good. We have *heteroskedasticiy*. The variance of the residuals seems way bigger for bigger prices.
Looks like log(Sales) on log(Price) is a better way to go!!

1.4 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 Neighborhood c=3 ?

(b)

According to the tree, what seems to be the highest price 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.