

Homework 03 Solutions

9/14/2020

1 Problem 1

1.1 (a) EDA

```
data <- read.table("../cars04.dat",header=T)
num_of_NA_vals <- sum(is.na(data))
```

Our data contains 234 unique vehicles' information and 13 features about each vehicle, and the data is very clean, with 0 missing data values across all vehicles. Table 1 contains information about all the variables in the dataset.

Table 1: Details about all variables in `car04.dat` data frame

Variable name	Description
Hybrid	Is the vehicle a hybrid? (Yes = 1, No = 0)
SuggestedRetailPrice	Suggested retail price, what the manufacturer thinks the vehicle is worth, including adequate profit for the automaker and the dealer (U.S. dollars)
DealerCost	Price dealer had to pay manufacturer (U.S. dollars)
EngineSize	Engine Size (liters)
Cylinders	Number of cylinders
Horsepower	Gross Horsepower
CityMPG	City miles per gallon
HighwayMPG	Highway miles per gallon
Weight	Weight of vehicle (pounds)
WheelBase	Wheelbase - horizontal distance between front and rear wheels (inches)
Length	Length of vehicle (inches)
Width	Width of vehicle (inches)

```
par(mfrow = c(3,4))
for (c_name in colnames(data)){
  if(length(unique(data[,c_name])) <= 6){
    c_vals <- names(table(data[,c_name]))
    c_counts <- table(data[,c_name])
    barplot(names.arg = c_vals,height= c_counts, main = c_name)
  }else {
    hist(data[,c_name], main = c_name)
  }
}
```

```
## or
# summary(data)
# table(data$Hybrid)
#
```

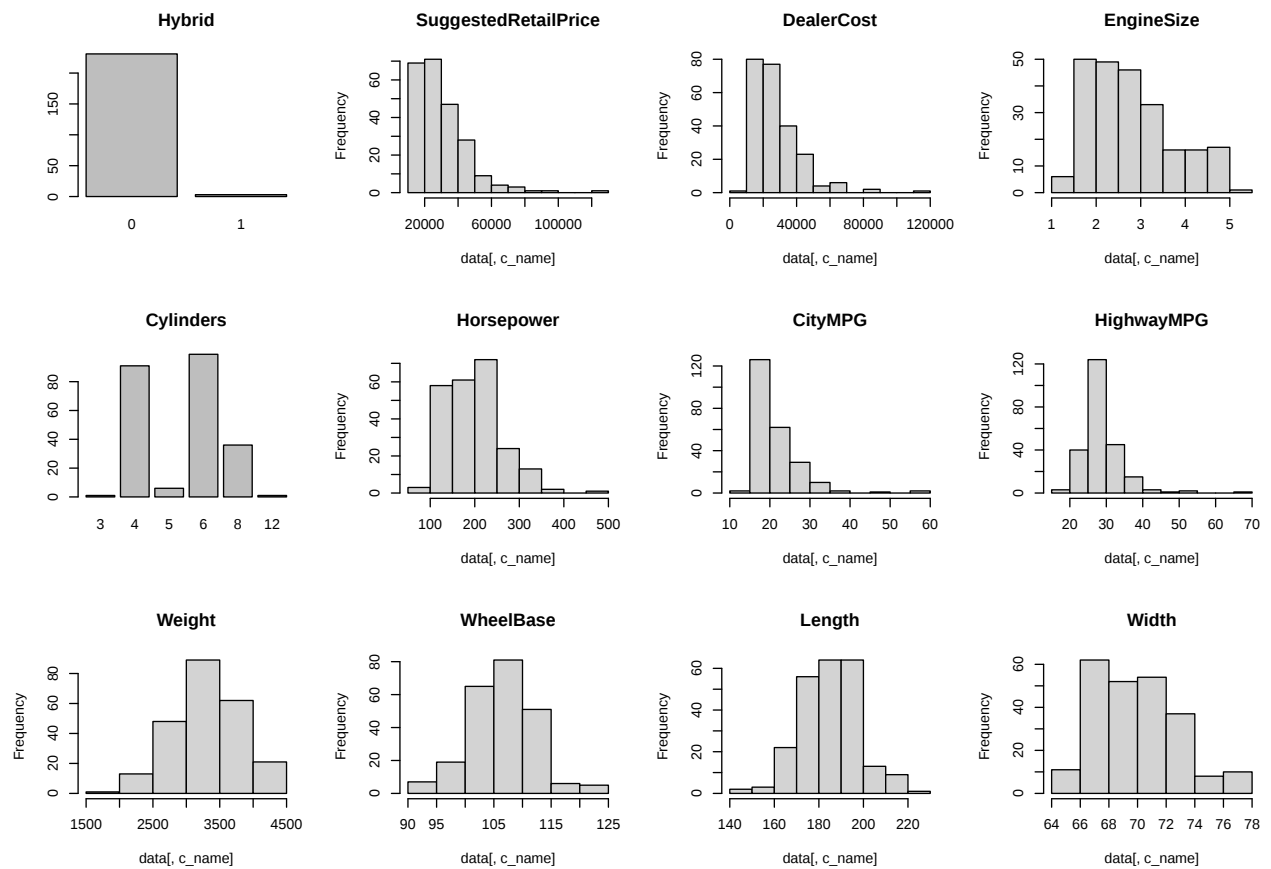


Figure 1: Univariate Distribution of Variables in `car04.dat`

```

# par(mfrow=c(3,4))
#
# for (i in 1:12) {
#   hist(data[,i],main=names(data)[i])
# }
#
# for (i in 1:12) {
#   boxplot(data[,i],main=names(data)[i])
# }

par(mar = c(4, 4, .1, .1))
library(psych)

pairs.panels(data[,2:3])
log_price <- log10(data[,2:3])
names(log_price) <- paste0("log10(",names(log_price), ")")
pairs.panels(log_price)

```

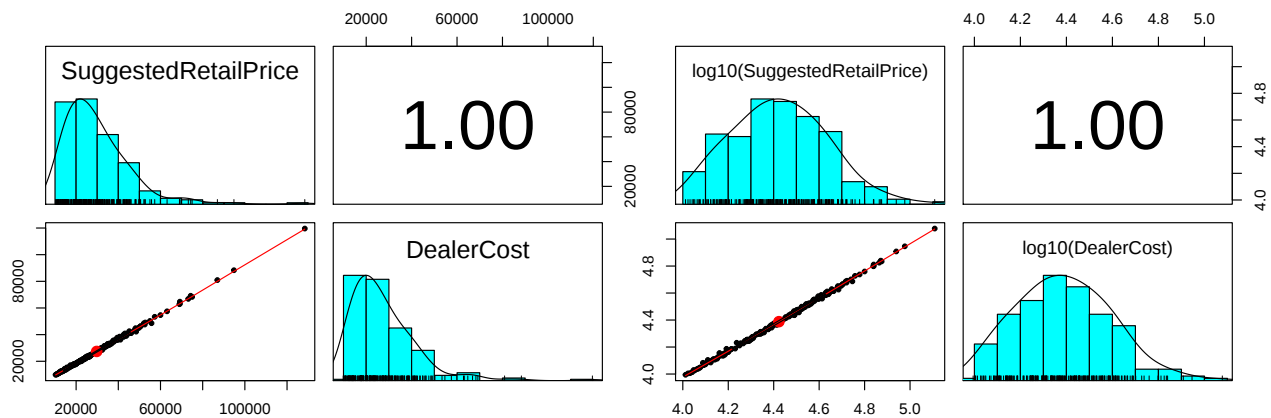


Figure 2: Distributions of Prices and $\log_{10}(\text{Prices})$ relative to each other

1.2 (b) Model Fitting

```

lm.1 <- lm(SuggestedRetailPrice ~ DealerCost, data=data)

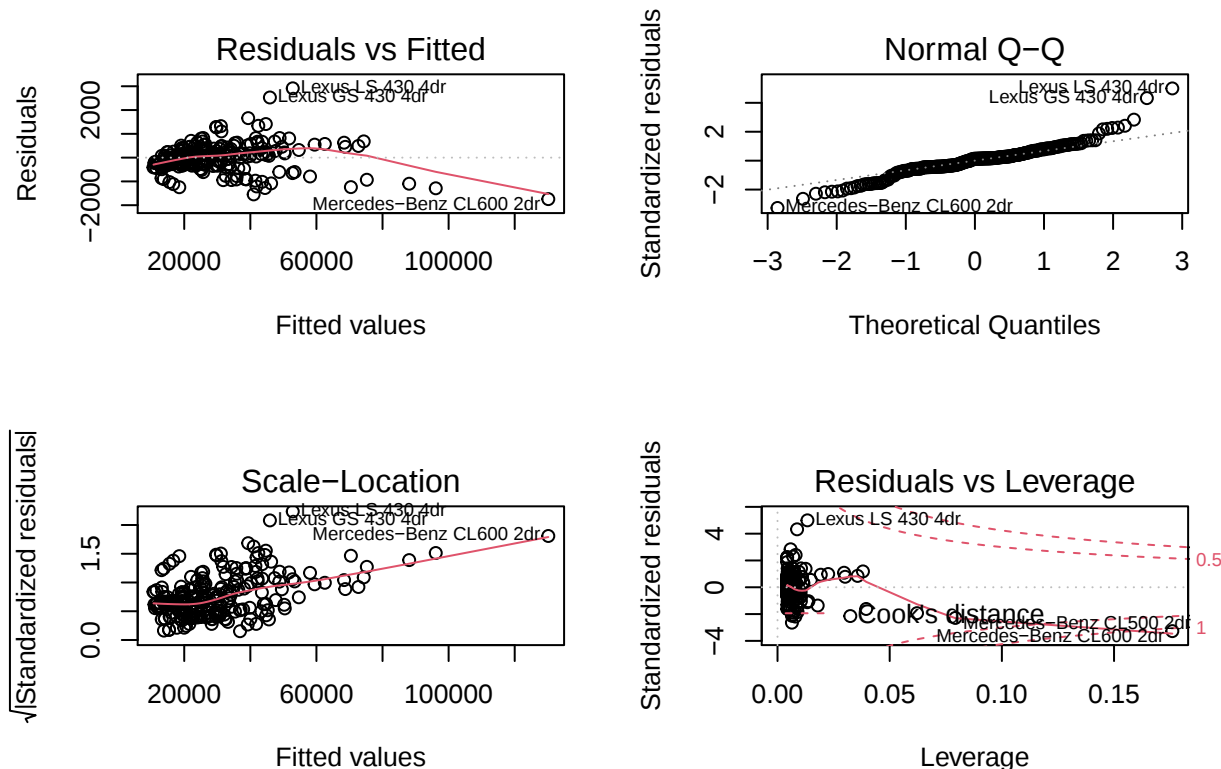
summary(lm.1)

##
## Call:
## lm(formula = SuggestedRetailPrice ~ DealerCost, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1743.52  -262.59   74.92   265.98  2912.72
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.904248  81.801381  -0.757    0.45
## DealerCost   1.088841   0.002638 412.768 <2e-16 ***
## ---

```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 587 on 232 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.704e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(lm.1)
```



In examining the relationship in the Section 1.1 there was a very linear relationship between the dealer's cost and the suggested retail price. After fitting, the very expensive cars (say with predicted sales cost above \$60,000) seem to be underestimated. Additionally, there appears to be natural clusters in the residual vs fitted plot and with future examination of some of these clusters (using `identify`) we find that certain clusters relate to vehicles made by the same brand - (e.g. lower left 5 points of the plot are all Suzuki cars). The residuals vs normal QQ plot suggests that the residuals have slightly heavier tails (i.e. more extreme values of residuals than might be expected). The scale-location plot shows us that as the estimated value increases we see an increase in the variability (suggesting heteroskedasticity of the errors)¹ - though it should be acknowledged that there are less observations with high estimated fitted values. Our Cook's distance figure shows that we have the Mercedes-Benz CL600 2dr has a very high impact on the model (having a Cook's distance above 1 and a high leverage score - with the average expected score being $1/234$).

The fitted regression would suggest that the retail price is supposed to be around 8-9% larger than the dealer cost, and it is natural to see that as the price goes up the difference in predicting the truth also scales (think as a proportion of the predicted price).

¹This observation can also be gathered for the residual-fitted visualization.

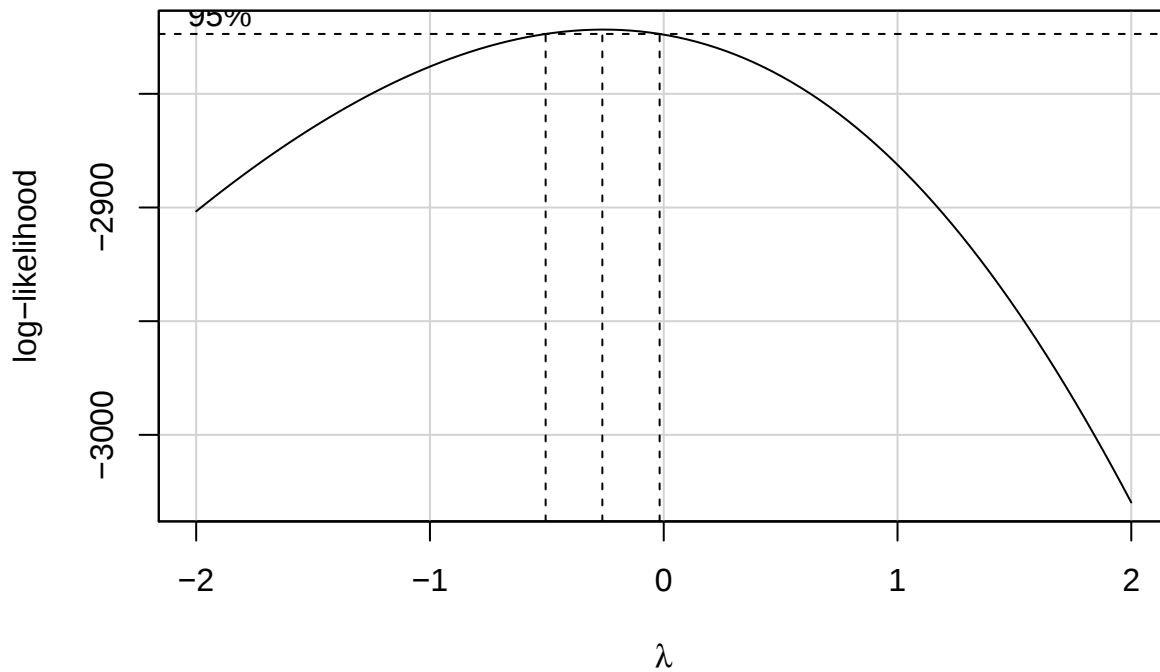
2 Problem 2 Box-Cox

```
library(car)
```

2.1 (a) Transform Exploration

2.1.1 i.

```
boxCox(lm(DealerCost~1,data=data))
```



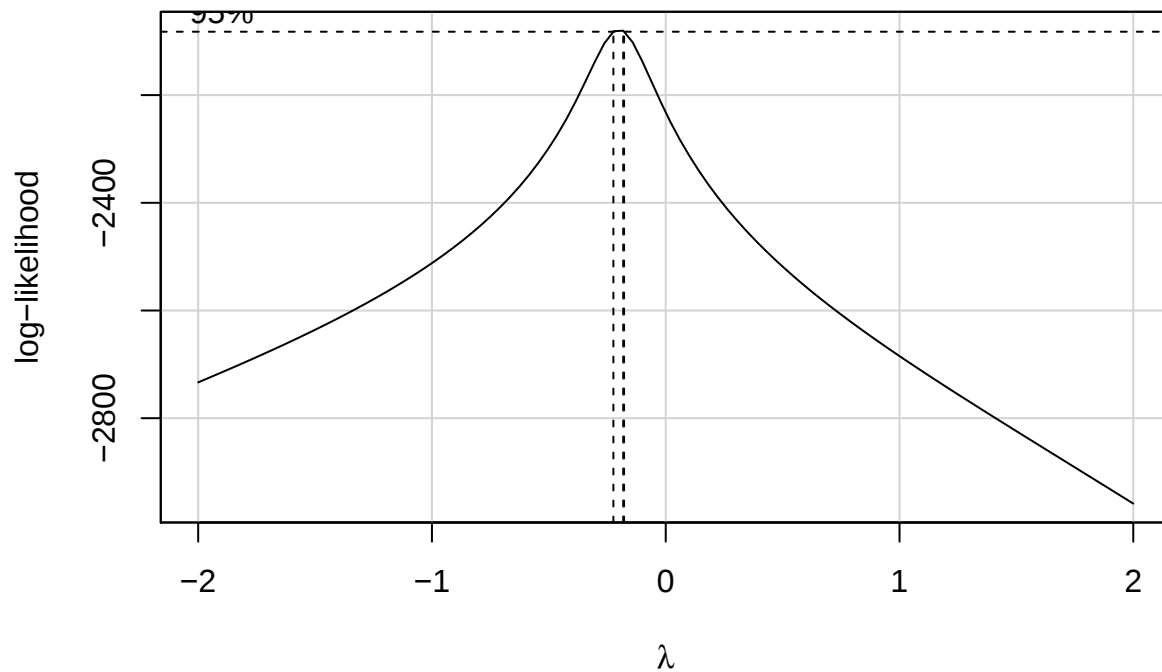
```
powerTransform(lm(DealerCost~1,data=data))
```

```
## Estimated transformation parameter  
##      Y1  
## -0.2576098
```

Based on the figure above we might transform DealerCost by raising it to a power of $-.25$.

2.1.2 ii.

```
data$TransCost <- data$DealerCost^(-.25)  
  
boxCox(lm(SuggestedRetailPrice~TransCost,data=data))
```



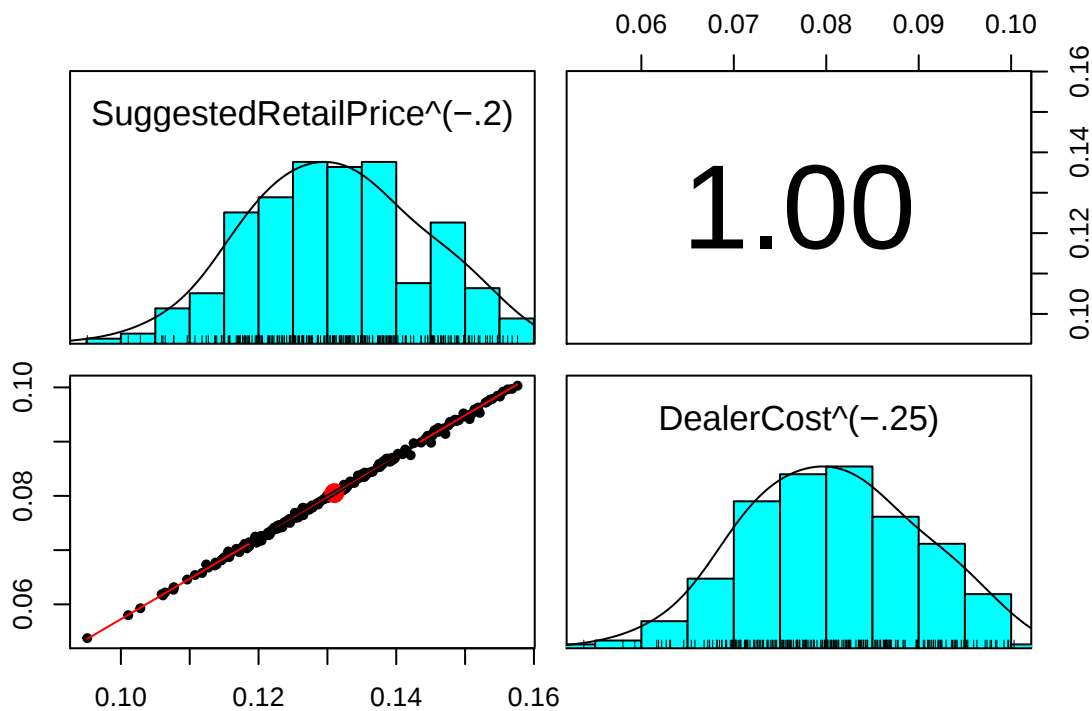
```
powerTransform(lm(SuggestedRetailPrice~TransCost,data=data))
```

```
## Estimated transformation parameter
##      Y1
## -0.2001193
```

2.2 (b) Transformed Model

```
data$srp.20 <- data$SuggestedRetailPrice^(-0.20)
data$dc.25 <- data$DealerCost^(-0.25)

par(mfrow=c(1,1))
pairs.panels(data.frame(`SuggestedRetailPrice^(-.2)` = data$srp.20,
                        `DealerCost^(-.25)` = data$dc.25,check.names = F))
```

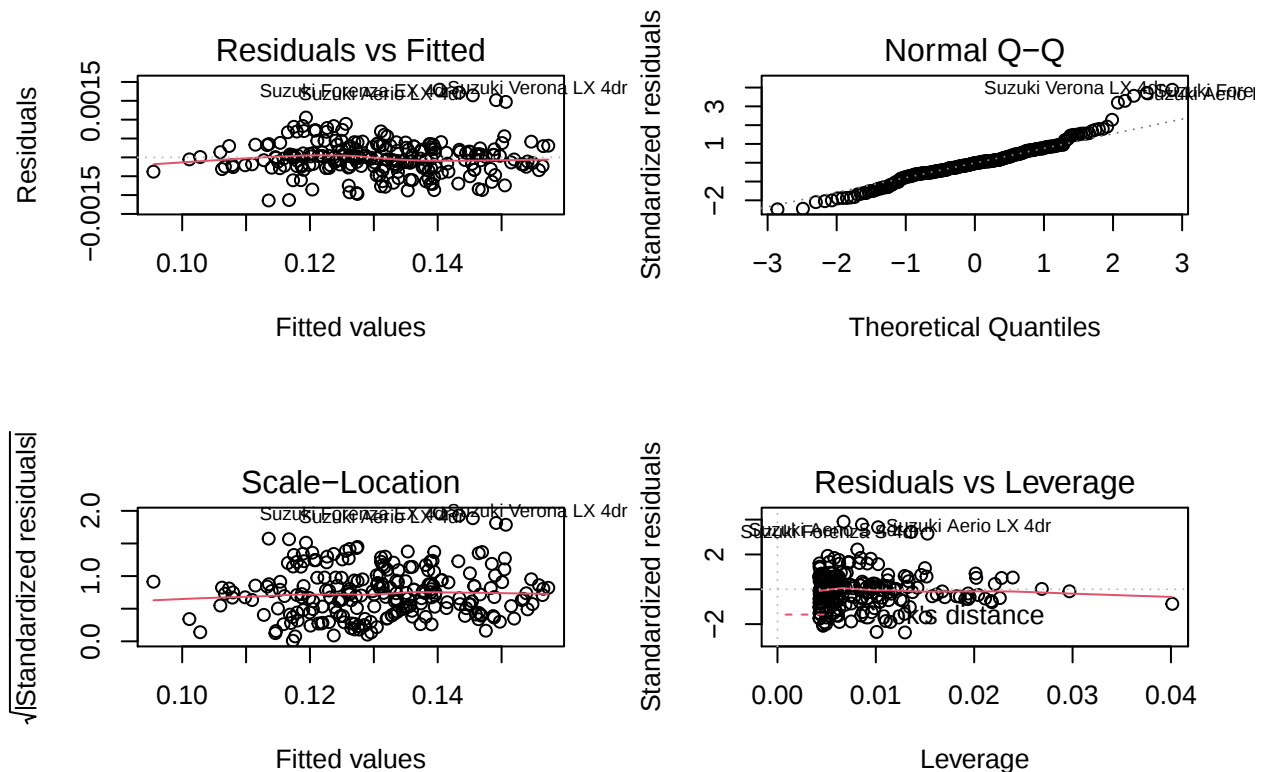


```
lm.2 <- lm(srp.20 ~ dc.25, data = data)
```

```
summary(lm.2)
```

```
##
## Call:
## lm(formula = srp.20 ~ dc.25, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.143e-03 -2.482e-04 -1.312e-05  2.365e-04  1.796e-03
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0240895  0.0002663   90.44  <2e-16 ***
## dc.25        1.3279968  0.0032863  404.10  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0004641 on 232 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.633e+05 on 1 and 232 DF, p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.2)
```



Relative to the original data, the transformed **Suggested RetailPrice** and **DealerCost** have distributions that are much more symmetric. In the residuals vs Fitted plot we still see some outliers² when the fitted value is higher but overall the distribution of residuals looks much better - it has a conditional average residual value around zero and the variability doesn't look too different as we look across the range of fitted values. The residuals now only have a heavy positive tail³. The scale-Location plot confirms our observation from the residual vs Fitted that this new model with the transformations seems to have homoscedastic residuals. The Cook's distance plot also shows less impact from highly influential points, there is a point with leverage at .04 (compared with the average 0.004, but this plot looks better than the original model).

I'm pretty pleased with this new model, most of the linear model assumptions look well met except for a few outliers (which were also identified in our examination of the first model).

3 Problem 3: Log-Log

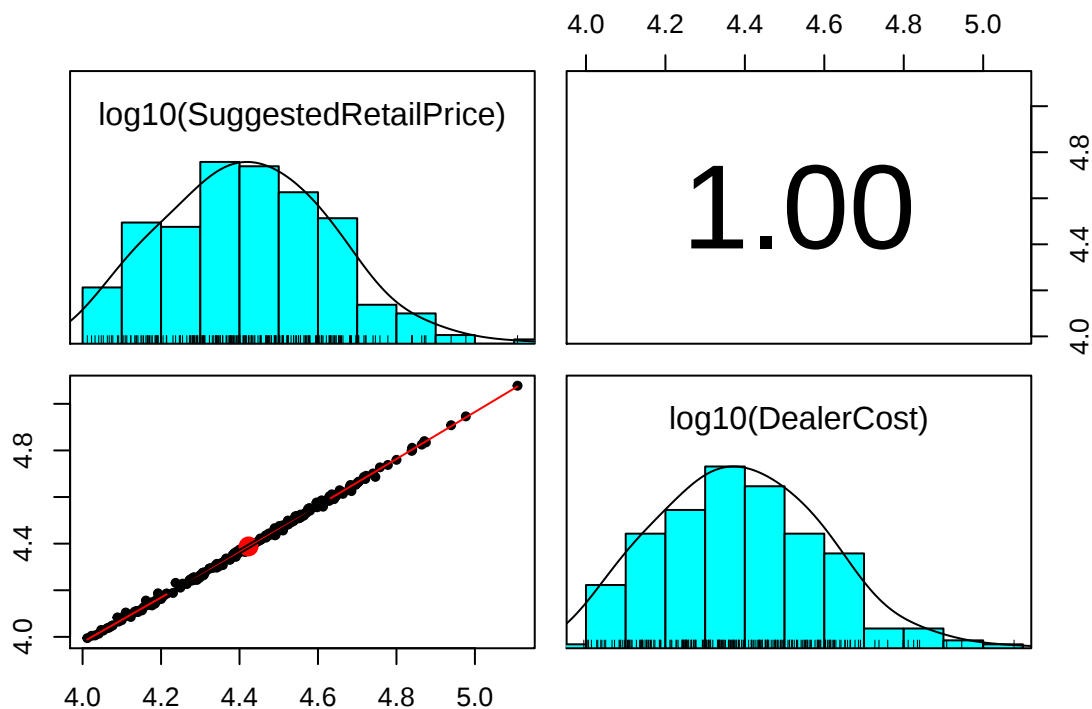
Note that we used log-based 10 (this is just a good thing to be thinking about when you're transforming and are interested in interpretation).

```
lm.3 <- lm(log10(SuggestedRetailPrice) ~ log10(DealerCost), data=data)

log_price <- log10(data[,2:3])
names(log_price) <- paste0("log10(", names(log_price), ")")
pairs.panels(log_price)
```

²Looks like 2 Suzuki branded cars

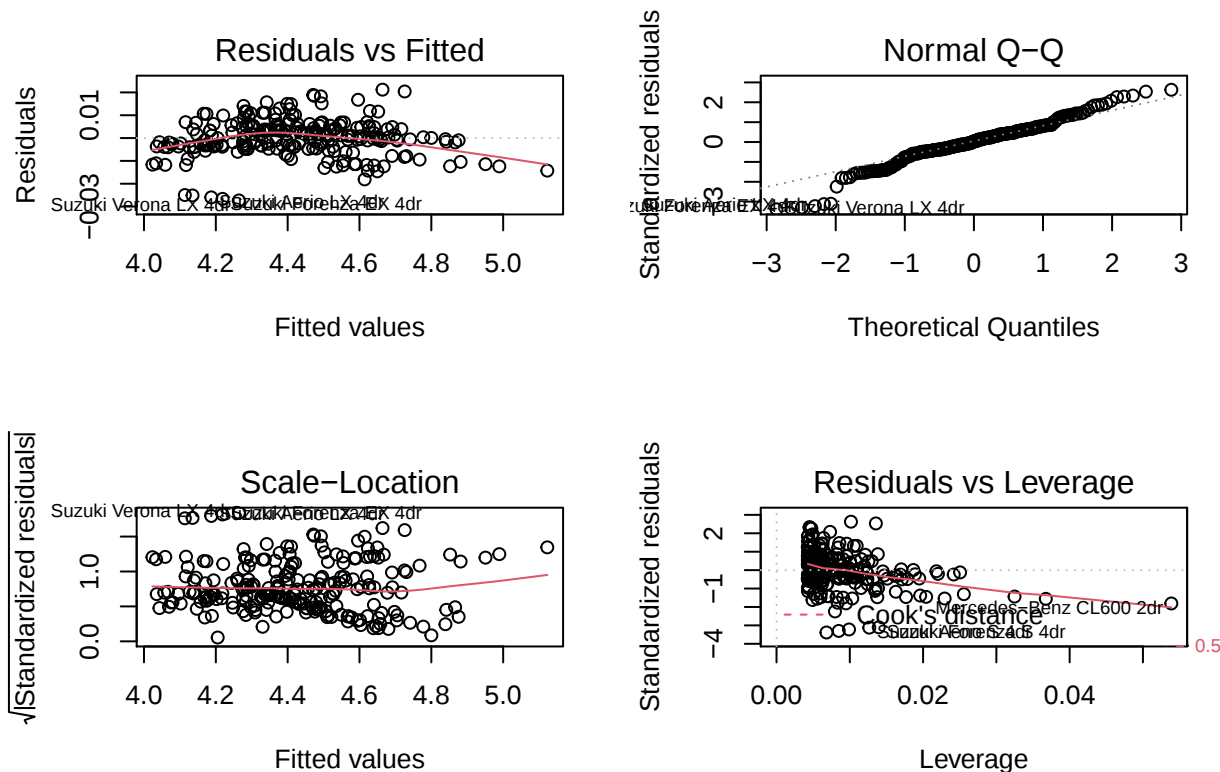
³The heavy positive tail on the QQ plot is directly related to those large residuals in the residual vs fitted plot.



```
summary(lm.3)
```

```
##
## Call:
## lm(formula = log10(SuggestedRetailPrice) ~ log10(DealerCost),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0273260 -0.0037760  0.0002708  0.0046125  0.0211929
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.030166   0.011491  -2.625  0.00924 **
## log10(DealerCost)  1.014836   0.002616 387.942 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008101 on 232 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 1.505e+05 on 1 and 232 DF, p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.3)
```



For our “Log-Log” model, there appears to be a slight trend in the average residual as we increase the fitted value, with a slightly positive average residual for the majority of the fitted value values with $\log(\text{fitted})$ around 4.2 to 4.6 but negative residual for fitted values on the the extremes. There is also a set of cars with very low residuals with $\log(\text{fitted})$ values between 4.2 and 4.6. The normal QQ plot looks decent, with a slightly heavy tail with negative residuals. For the extremely large fitted values we might also be seeing an indication for increased variability of residuals in the Scale-Location plot (although with so few points, this claim is probably premature). There is a group of points with larger leverage than the average, but their Cook’s distances aren’t that large.

This fit looks much better than the original regression, but there are still some odd groupings in the residual-fitted plot. Overall it looks like a decent model.

4 Problem 4

4.1 (a) Best Model (based on assumptions)

The Box-Cox transformed model seems to best meet the assumptions of the linear regression, specifically in the true linear relationship (examined with $E[\hat{e}|x] = 0$). Additionally, the second model seems well meet the assumption of constant noise variance.

4.2 (b) “Complete the Sentence”

- For the model in #1(b), a \$1 increase in `DealerCost` results in a \$1.09 increase in the expected value of `SuggestedRetailPrice`.
- For the model in #2(b), a 1 unit increase in $(\text{DealerCost})^{-.25}$ results in a 1.33% increase in the expected value of $(\text{SuggestedRetailPrice})^{-.2}$.

- For the model in #3, a 1% increase in DealerCost results in a 1.01% increase in the expected value of SuggestedRetailPrice.

4.3 (c) Best Model (for consulting)

Giving the difficulty of translating the relationship between DealerCost to SuggestedRetailPrice for the Box-Cox transformed model, I would probably use the Log-Log model for a consulting project. The Log-Log model is still interpretable and more closely meets the OLS assumptions.

5 Problem 5

```
pyth_data <- read.table("../pyth.dat",header = T)
```

5.1 (a)

$$M1 : y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$M2 : y = \beta_0 + \beta_1 x_2 + \varepsilon$$

```
lm.1 <- lm(y~x1, data=pyth_data)
lm.2 <- lm(y~x2, data=pyth_data)
summary(lm.1)

##
## Call:
## lm(formula = y ~ x1, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.7409 -4.5056  0.7114  4.3739  7.7547
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.0633     1.5526   6.481 1.25e-07 ***
## x1              0.6559     0.2499   2.625  0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.921 on 38 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  0.1535, Adjusted R-squared:  0.1312
## F-statistic:  6.89 on 1 and 38 DF,  p-value: 0.01242

summary(lm.2)

##
## Call:
## lm(formula = y ~ x2, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

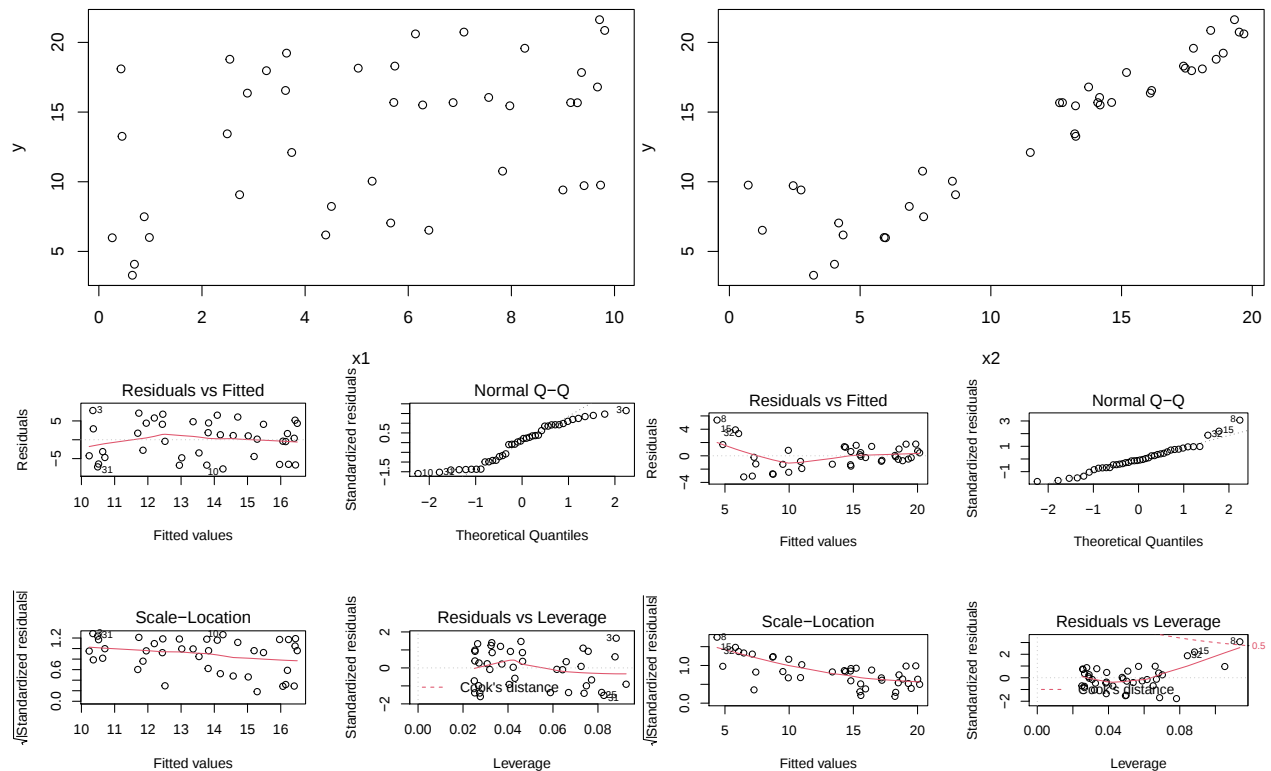
```
## -3.1751 -1.2352 -0.1867  1.0899  5.3755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.78532    0.66037   5.732 1.33e-06 ***
## x2           0.83223    0.05017  16.589 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.863 on 38 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  0.8787, Adjusted R-squared:  0.8755
## F-statistic: 275.2 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
plot(y~x1, data = pyth_data)
plot(y~x2, data = pyth_data)
```

```
par(mfrow=c(2,2))
```

```
plot(lm.1)
```

```
plot(lm.2)
```



$M2$ is better, it reduces more of the variability of $y - \hat{y}$ by much more than $M1$.

5.2 (b)

$$M3: y2 = \beta_0 + \beta_1 x12 + \varepsilon$$

$$M4: y2 = \beta_0 + \beta_1 x22 + \varepsilon$$

```
pyth_data$y2 <- pyth_data$y^2
pyth_data$x12 <- pyth_data$x1^2
pyth_data$x22 <- pyth_data$x2^2

## OR
# attach(pyth_data)
#
# y2 <- y^2
# x12 <- x1^2
# x22 <- x2^2
#
# detach()

lm.3 <- lm(y2~x12, data=pyth_data)

lm.4 <- lm(y2~x22, data=pyth_data)

summary(lm.3)

##
## Call:
## lm(formula = y2 ~ x12, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -189.324 -125.674   4.988  131.052  214.089
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  161.7831    31.7873   5.090   1e-05 ***
## x12           1.2971     0.6242   2.078   0.0445 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.1 on 38 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  0.102, Adjusted R-squared:  0.07841
## F-statistic: 4.318 on 1 and 38 DF, p-value: 0.04452

summary(lm.4)

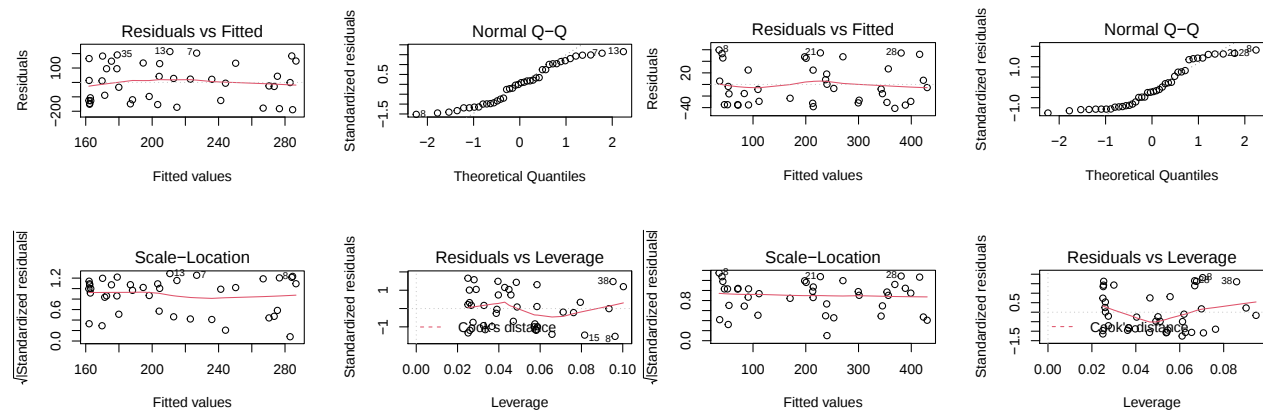
##
## Call:
## lm(formula = y2 ~ x22, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -41.280 -31.224  -7.463   25.422   59.571
```

```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.1583     9.0306   3.893 0.000387 ***
## x22          1.0198     0.0419  24.338 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.96 on 38 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  0.9397, Adjusted R-squared:  0.9381
## F-statistic: 592.3 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
```

```
plot(lm.3)
```

```
plot(lm.4)
```



$M4$ is better, it reduces more of the variability of $y^2 - \hat{y}^2$ by much more than $M3$.

5.3 (c)

$$M5 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$M6 : y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \varepsilon$$

```
lm.5 <- lm(y ~ x1 + x2, data=pyth_data)
```

```
lm.6 <- lm(y2 ~ x12 + x22, data=pyth_data)
```

```
summary(lm.5)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## x1           0.51481    0.04590  11.216 1.84e-13 ***
## x2           0.80692    0.02434  33.148 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 37 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9709
## F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
```

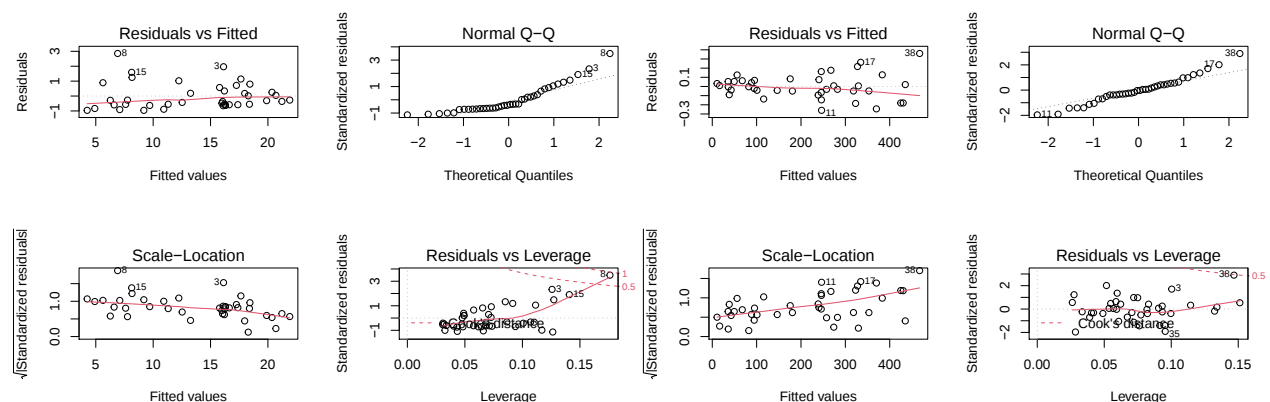
```
summary(lm.6)
```

```
##
## Call:
## lm(formula = y2 ~ x12 + x22, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26020 -0.05391 -0.00396  0.06367  0.35990
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0026691  0.0422669   0.063    0.95
## x12          0.9999672  0.0006419 1557.713 <2e-16 ***
## x22          0.9998685  0.0001663 6011.909 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1344 on 37 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  1, Adjusted R-squared:  1
## F-statistic: 2.013e+07 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
```

```
plot(lm.5)
```

```
plot(lm.6)
```



Both $M5$ and $M6$ represent a lot their respectively output, but $M6$ basically captures all variability in y^2 .

5.4 (d)

$$M7 : y = \beta_0 + \beta_1 \sqrt{x_{12} + x_{22}} + \varepsilon$$

```
pyth_data$x3 <- sqrt(pyth_data$x12 + pyth_data$x22)

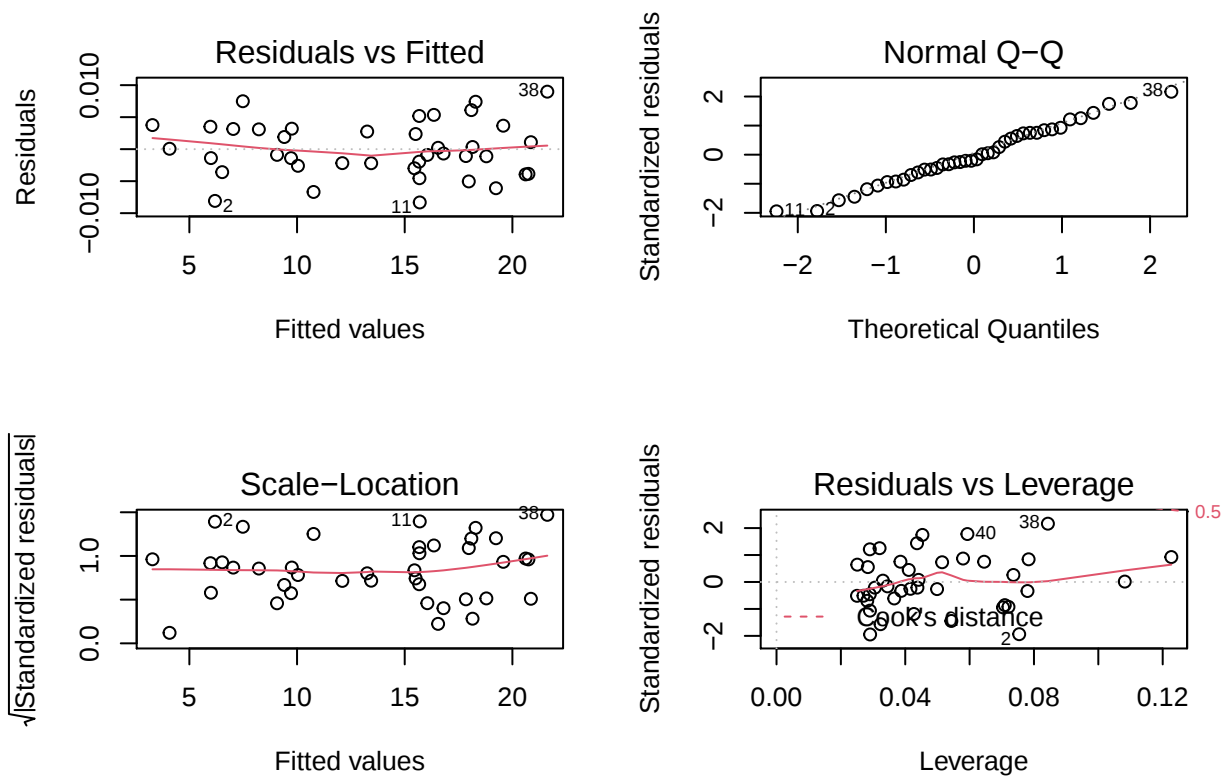
lm.7 <- lm(y ~ x3, data=pyth_data)

summary(lm.7)

##
## Call:
## lm(formula = y ~ x3, data = pyth_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0083283 -0.0027000 -0.0007907  0.0031643  0.0089809
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.0018422   0.0019159    0.962   0.342
## x3           0.9998313   0.0001316 7596.431 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00434 on 38 degrees of freedom
## (20 observations deleted due to missingness)
## Multiple R-squared:  1, Adjusted R-squared:  1
## F-statistic: 5.771e+07 on 1 and 38 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))

plot(lm.7)
```



Based on observations made on $M6$, $M7$ seems to capture basically all the variability of y . This relates to the fact that $M6$ was basically suggesting that $y^2 \approx 0 + 1x_1^2 + 1x_2^2$, which could be re-expressed as

$$y^2 \approx 0 + 1x_1^2 + 1x_2^2$$

$$\Leftrightarrow y \approx \sqrt{x_1^2 + x_2^2} .$$