MA 575 HW

Benjamin Draves

October 24, 2017

Exerices 6.1

(a)

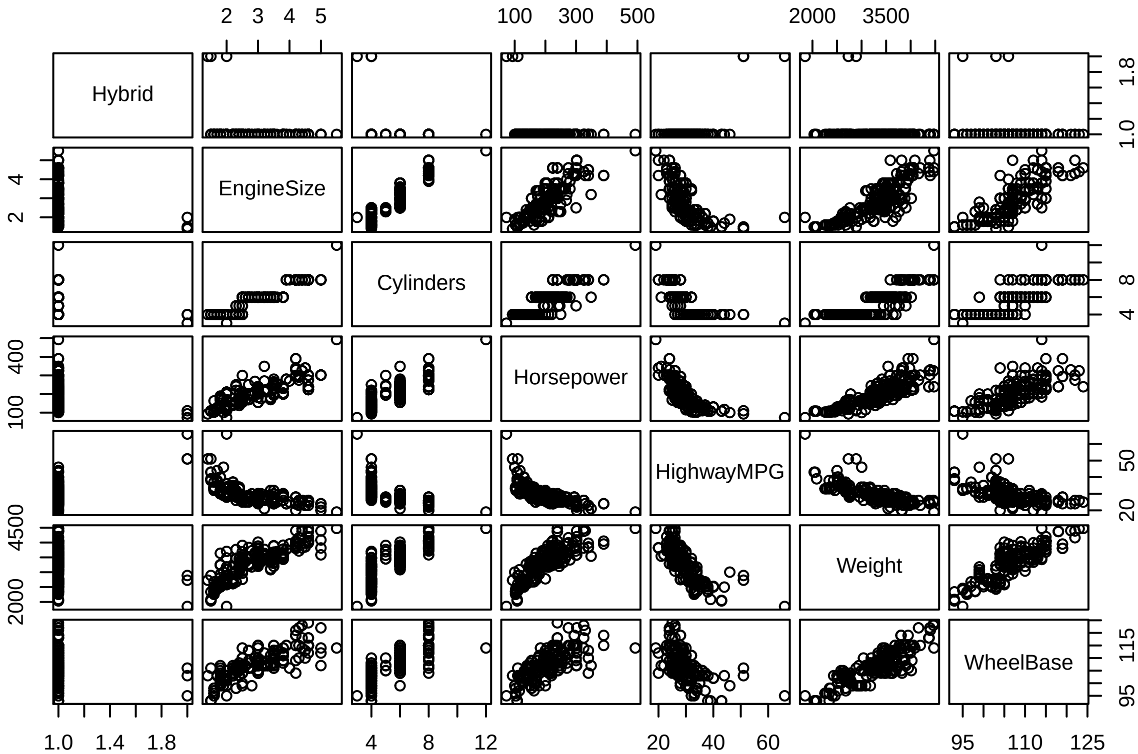
```r
#read ind data
dat = read.csv("~/Desktop/Courses/MA 575/book_data/cars04.csv", header = TRUE)

#take a peak
str(dat)
```

```
## 'data.frame':    234 obs. of  13 variables:
##  $ Vehicle.Name        : Factor w/ 232 levels "Acura 3.5 RL 4dr",..: 43 44 45 46 47 69 70 79
##  $ Hybrid              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SuggestedRetailPrice: int  11690 12585 14610 14810 16385 13670 15040 13270 13730 15460 .
##  $ DealerCost          : int  10965 11802 13697 13884 15357 12849 14086 12482 12906 14496 .
##  $ EngineSize          : num  1.6 1.6 2.2 2.2 2.2 2 2 2 2 2 ...
##  $ Cylinders           : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ Horsepower          : int  103 103 140 140 140 132 132 130 110 130 ...
##  $ CityMPG             : int  28 28 26 26 26 29 29 26 27 26 ...
##  $ HighwayMPG          : int  34 34 37 37 37 36 36 33 36 33 ...
##  $ Weight              : int  2370 2348 2617 2676 2617 2581 2626 2612 2606 2606 ...
##  $ WheelBase           : int  98 98 104 104 104 105 105 103 103 103 ...
##  $ Length              : int  167 153 183 183 183 174 174 168 168 168 ...
##  $ Width               : int  66 66 69 68 69 67 67 67 67 67 ...
```

```r
#cast Hybrid as a factor
dat$Hybrid = as.factor(dat$Hybrid)

#build model
model = lm(SuggestedRetailPrice ~ EngineSize + Cylinders + Horsepower + HighwayMPG + Weight + W
```

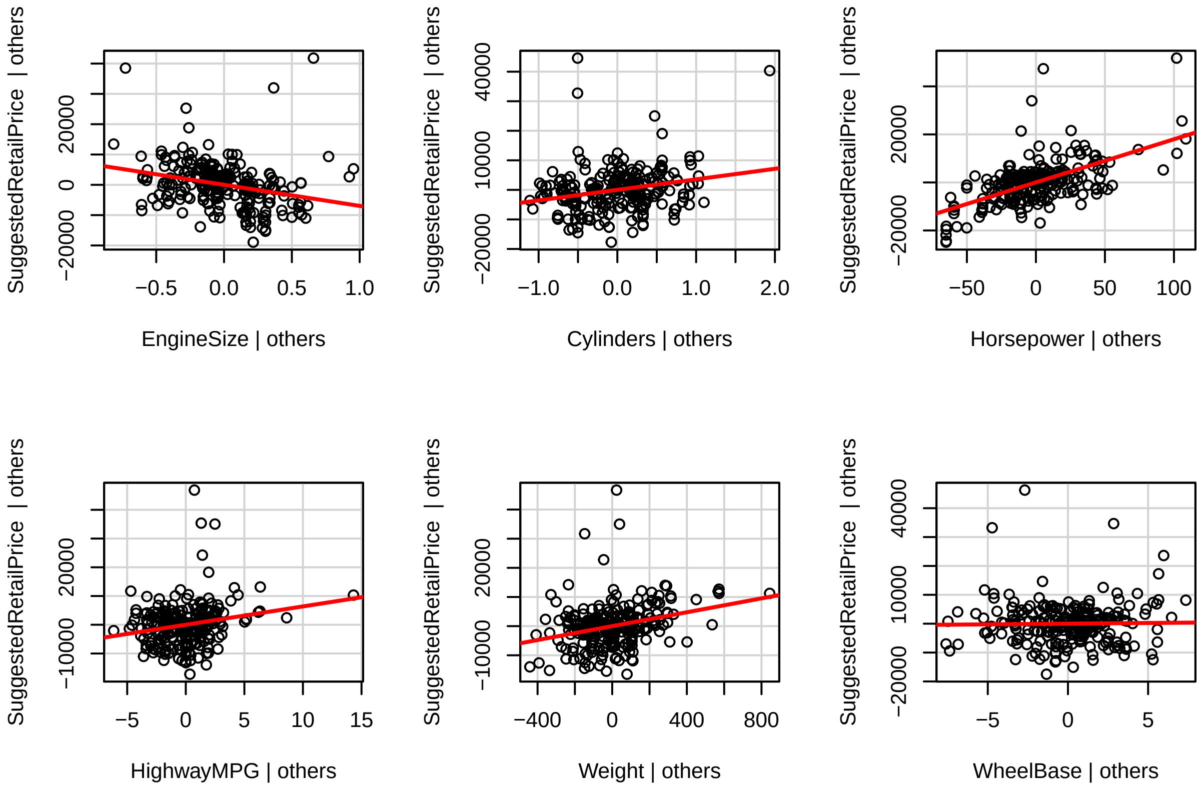
Having built the model we will now look at some diagnostics - scatter matrix of covariates, added variable plots, diagnostic plots etc.

```r
#Scatter matrix
scat.mat = dat[,c(2,5:7, 9:11)]

#Scatter matrix plot
pairs(scat.mat, gap = 0.4)
```



It appears that most of our covariates are related. Specifically weight and wheel base are strongly linearly related. We should be careful when testing significance of either of these variables with high VIF. There are other trends that can be seen - Horsepower vs Weight, Enginesize vs HighwayMPG, Engine Size vs Wheel base. These are all closely related to the overall size of the vehicle. (Some PC regression would do really well here..). Next we'll look at added variable plots.

```r
#added variable plots
library(car)

par(mfrow=c(2,3))
avPlots(model, ~EngineSize)
avPlots(model, ~Cylinders)
avPlots(model, ~Horsepower)
avPlots(model, ~HighwayMPG)
avPlots(model, ~Weight)
avPlots(model, ~WheelBase)
```



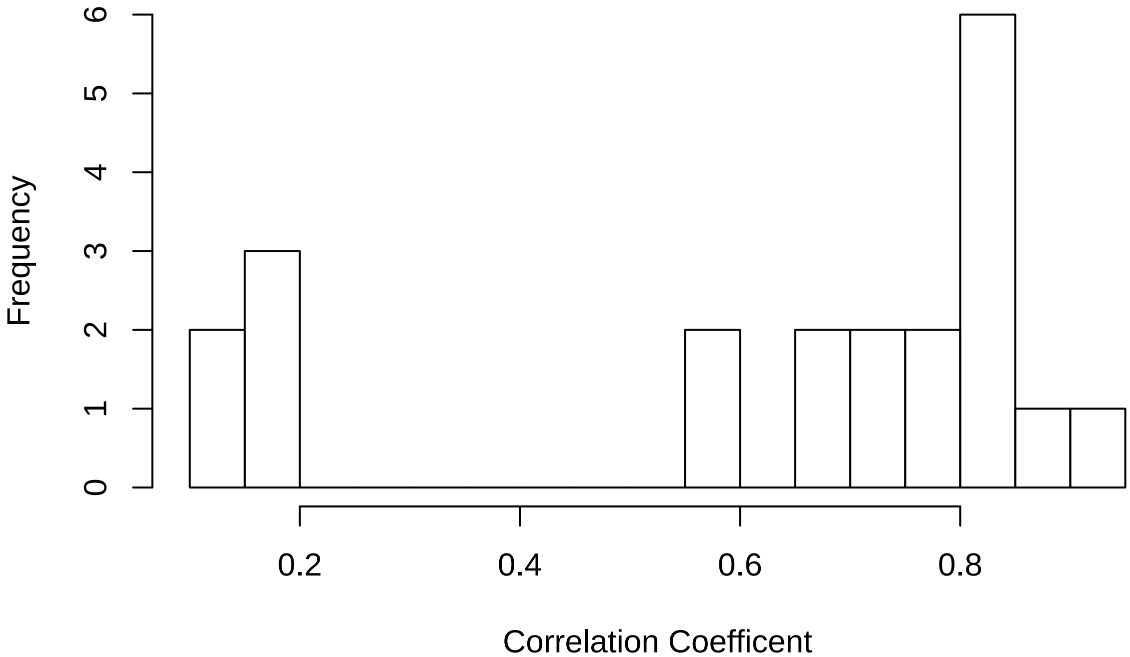
It looks like all variables add a significant amount of information to our regression (i.e. each regression explains additional variance in the price) expect for Wheel Base which is almost entirely constant after removing affects of other variables. Notice that weight does not have this effect but we noted how closely related they were. We may need to consider removing that variable to improve the colinearity issue.

```r
#Recast hybrid as a numeric for correlation purposes
scat.mat$Hybrid = as.numeric(as.character(scat.mat$Hybrid))

#cast scat.mat as a matrix
scat.mat = as.matrix(scat.mat)

#correlation matrix
x = cor(scat.mat)
x
```

```
##                Hybrid EngineSize  Cylinders Horsepower HighwayMPG
## Hybrid      1.0000000 -0.1562051 -0.1436261 -0.1922594  0.5655500
## EngineSize -0.1562051  1.0000000  0.9275933  0.8246932 -0.6564508
## Cylinders  -0.1436261  0.9275933  1.0000000  0.8405416 -0.6608553
## Horsepower -0.1922594  0.8246932  0.8405416  1.0000000 -0.7189530
## HighwayMPG  0.5655500 -0.6564508 -0.6608553 -0.7189530  1.0000000
## Weight     -0.1782540  0.8447530  0.8319864  0.8312084 -0.7605876
## WheelBase  -0.1134199  0.8171100  0.7725715  0.7283060 -0.5835760
##                Weight  WheelBase
## Hybrid     -0.1782540 -0.1134199
```

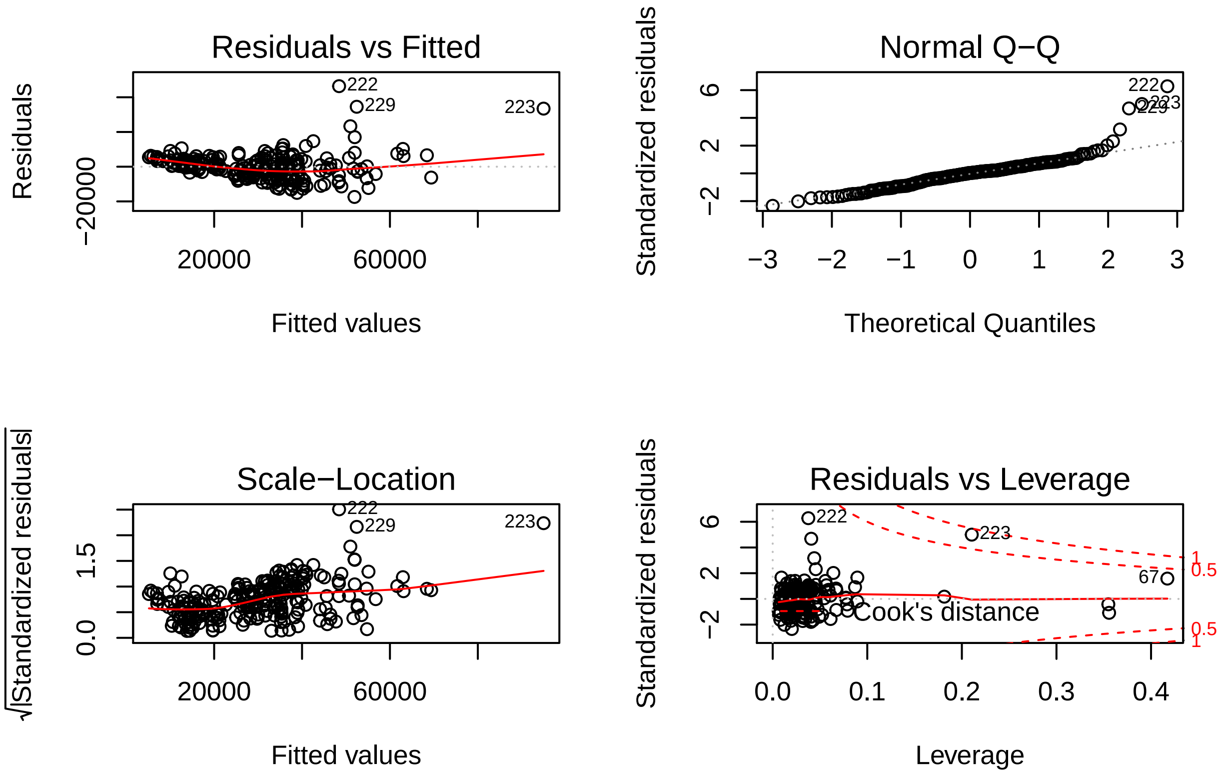
```
## EngineSize  0.8447530  0.8171100
## Cylinders   0.8319864  0.7725715
## Horsepower  0.8312084  0.7283060
## HighwayMPG -0.7605876 -0.5835760
## Weight      1.0000000  0.8524790
## WheelBase   0.8524790  1.0000000
```

```
#visual representaion of pairiwse
hist(abs(x[upper.tri(x, diag = FALSE)]), main = "", xlab = "Correlation Coefficent", breaks = 
```



As we noted in the pairwise plots - we have colinearity issues with our covariates.

```
#take a look at the standardized residuals
par(mfrow = c(2,2))
plot(model)
```



The residuals appear to have some heteroskedasiticy issues. As the fitted values grow, so does the variance. Specifically for estimated price over $40,000, we see that the variance is quite high. I would say that this is *not* a valid model.

(b)

As stated above, we can say that there are issues with estimating the tails of the distribution. As see in the Q-Q plot, our model underestimates prices for low price vehicles and for high price vehicles. This implies that we have some skewness issues - which could be corrected by taking a log of the response variable price.

(c)

It appears that points labeled 222, 223, and 229 are all bad leverage points. They all are quite high on the theoretical quartile and have high residuals. This is again due to the fact that we underestimate high price vehicles.

```
dat[c(222,223, 229),]
```

```
##                 Vehicle.Name Hybrid SuggestedRetailPrice DealerCost
## 222 Mercedes-Benz CL500 2dr      0                94820      88324
## 223 Mercedes-Benz CL600 2dr      0               128420     119600
## 229  Mercedes-Benz S500 4dr      0                86970      80939
##     EngineSize Cylinders Horsepower CityMPG HighwayMPG Weight WheelBase
## 222        5.0         8        302      16         24   4085       114
## 223        5.5        12        493      13         19   4473       114
## 229        5.0         8        302      16         24   4390       122
```

```
##     Length Width
## 222    196    73
## 223    196    73
## 229    203    73
```

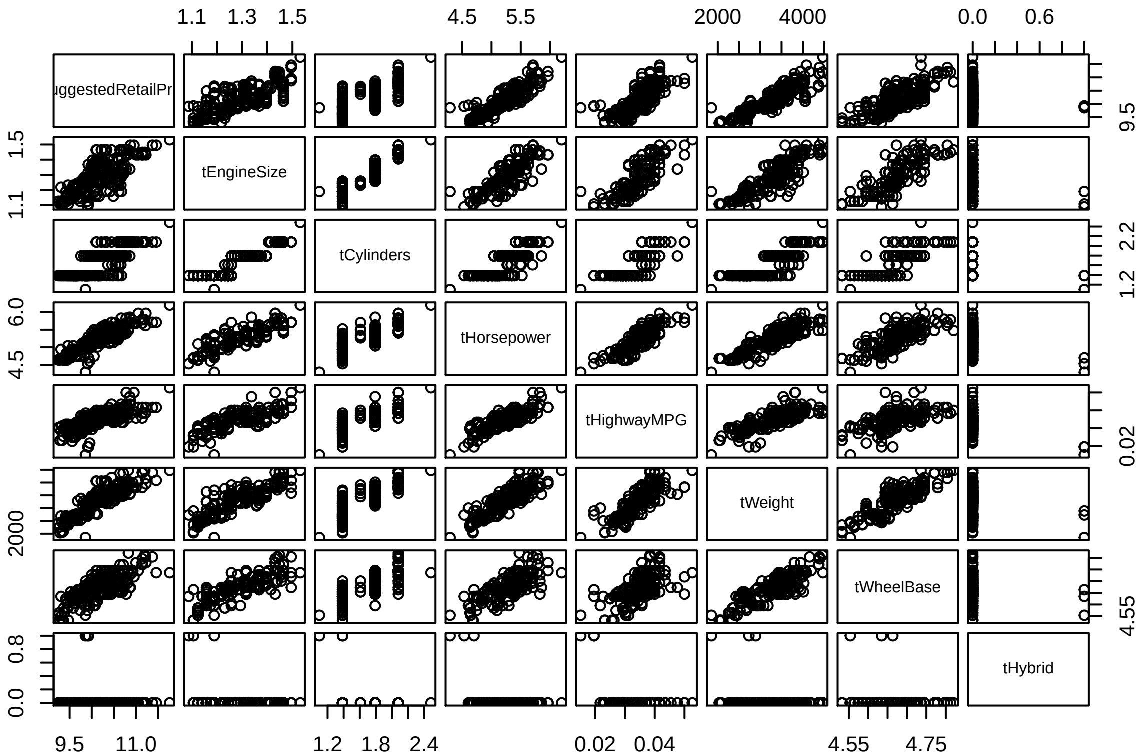
Here we see that all three vehicles are luxury Mercedes-Benz with the lowest suggested retail price of 86,970 (almost triple the mean of Suggested Retail Price!). We severely underestimate this value and these leverage points skew our regression line.

(d)

```
#create new data frame with transformed variables
bcdat = data.frame(tSuggestedRetailPrice = log(dat$SuggestedRetailPrice), tEngineSize = (dat$En

#build box-cox model
bcmodel = lm(tSuggestedRetailPrice ~ tEngineSize + tCylinders + tHorsepower + tHighwayMPG + tWe
```

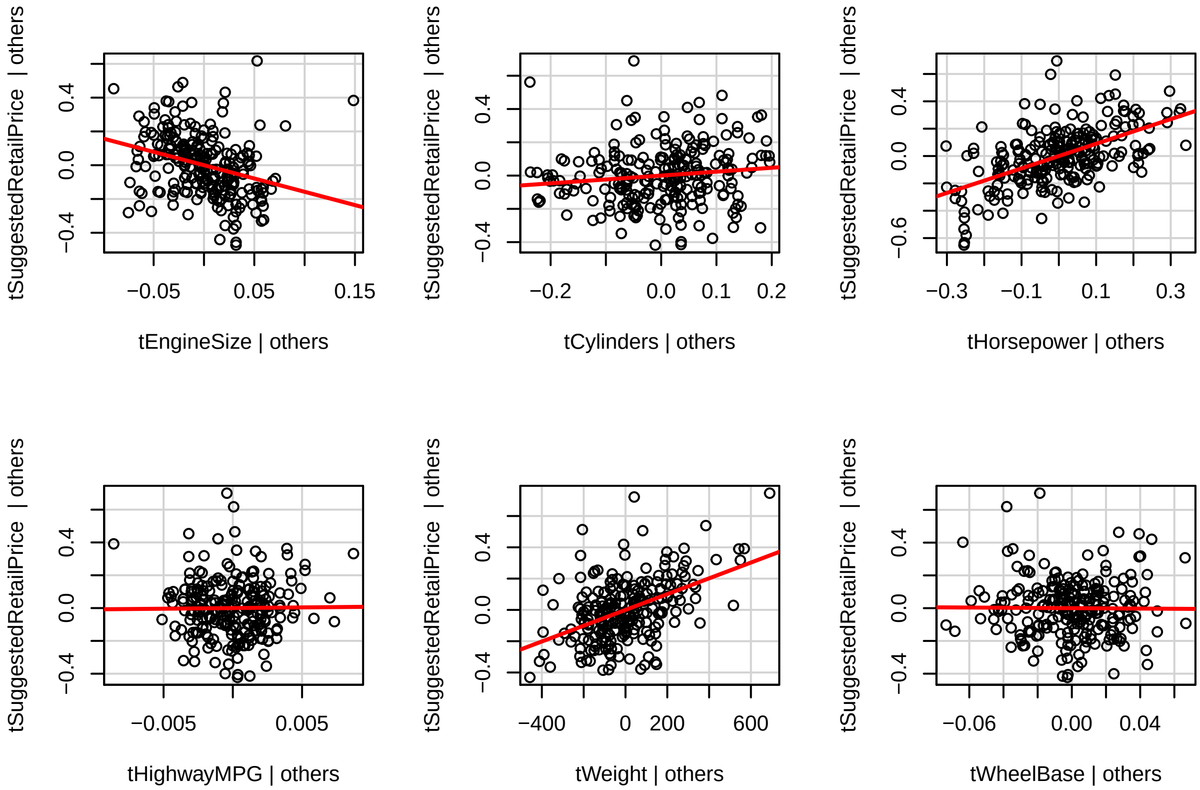
We complete the same procedure as above. This time however, we will produce all plots then discuss.

```
#Scatter matrix
bcdat$tHybrid = as.numeric(as.character(bcdat$tHybrid))
bcdat.mat = as.matrix(bcdat)
pairs(bcdat.mat, gap = 0.4)
```

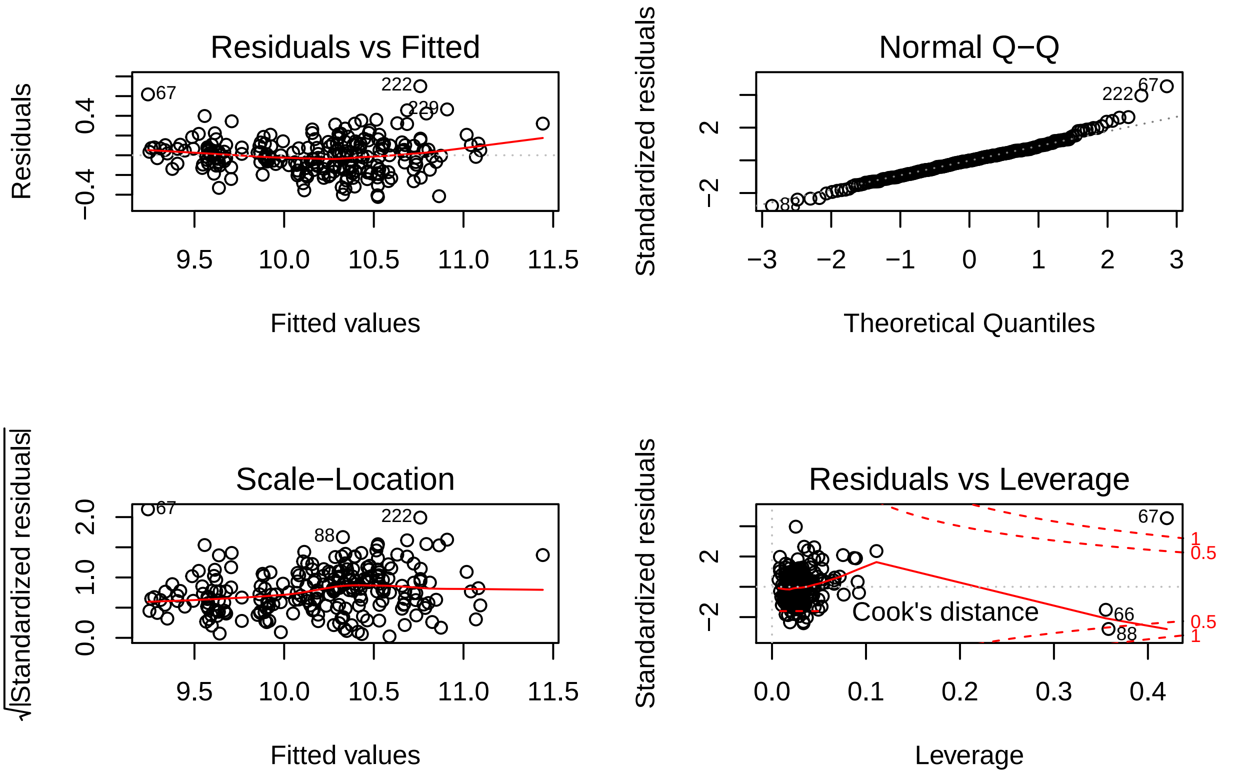



```
#Added variable plots
par(mfrow=c(2,3))
avPlots(bcmodel, ~tEngineSize)
avPlots(bcmodel, ~tCylinders)
```

```r
avPlots(bcmodel, ~tHorsepower)
avPlots(bcmodel, ~tHighwayMPG)
avPlots(bcmodel, ~tWeight)
avPlots(bcmodel, ~tWheelBase)
```




```r
#model diagnostics
par(mfrow = c(2,2))
plot(bcmodel)
```



Here we see that we still have colinearity but less so (most of our variables underwent nonlinear transformations). The added variable plots here show that tWheelBase's and tHighwayMPG could be adding little additional information to our model. Lastly, we see that we have significant improvement in our skewness issues. While there is still slight deviation in the Normal Q-Q plot there still could be like heteroskedasity issues. While this model isn't perfect, it could very *useful* at this point.

(e)

```r
#look at the summary of the model
summary(bcmodel)
```

```
##
## Call:
## lm(formula = tSuggestedRetailPrice ~ tEngineSize + tCylinders +
##     tHorsepower + tHighwayMPG + tWeight + tWheelBase + tHybrid,
##     data = bcdat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42288 -0.10983 -0.00203  0.10279  0.70068
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.703e+00  2.010e+00   2.838  0.00496 **
## tEngineSize -1.575e+00  3.332e-01  -4.727 4.01e-06 ***
```

```
## tCylinders   2.335e-01  1.204e-01   1.940  0.05359 .
## tHorsepower  8.992e-01  8.876e-02  10.130  < 2e-16 ***
## tHighwayMPG  8.029e-01  4.758e+00   0.169  0.86614
## tWeight      5.043e-04  6.367e-05   7.920 1.07e-13 ***
## tWheelBase  -6.385e-02  4.715e-01  -0.135  0.89240
## tHybrid1     6.422e-01  1.150e-01   5.582 6.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1789 on 226 degrees of freedom
## Multiple R-squared:  0.8621, Adjusted R-squared:  0.8578
## F-statistic: 201.8 on 7 and 226 DF,  p-value: < 2.2e-16
```

```r
#define reduced model without insignifcant materials
bcreduced = lm(tSuggestedRetailPrice ~ tEngineSize + tCylinders + tHorsepower + tWeight + tHyb

#preform parital F
anova(bcreduced, bcmodel)
```

```
## Analysis of Variance Table
##
## Model 1: tSuggestedRetailPrice ~ tEngineSize + tCylinders + tHorsepower +
##     tWeight + tHybrid
## Model 2: tSuggestedRetailPrice ~ tEngineSize + tCylinders + tHorsepower +
##     tHighwayMPG + tWeight + tWheelBase + tHybrid
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    228 7.2358
## 2    226 7.2337  2 0.0021769 0.034 0.9666
```

Here we see that the paritial F-test shows that we do not have sufficent evidence to suggest either tHighwayMPG or tWheel base is not zero. This implies that is a sensible strategy to remove both variables in this case.

(f)

We could simply add in another factor variable - Tayota Yes or No - to our regression that would model this covariance.

Exercise 6.2

```r
#Read in data
dat = read.table("~/Desktop/Courses/MA 575/book_data/krafft.txt", header = TRUE)

#take a peak
str(dat)
```

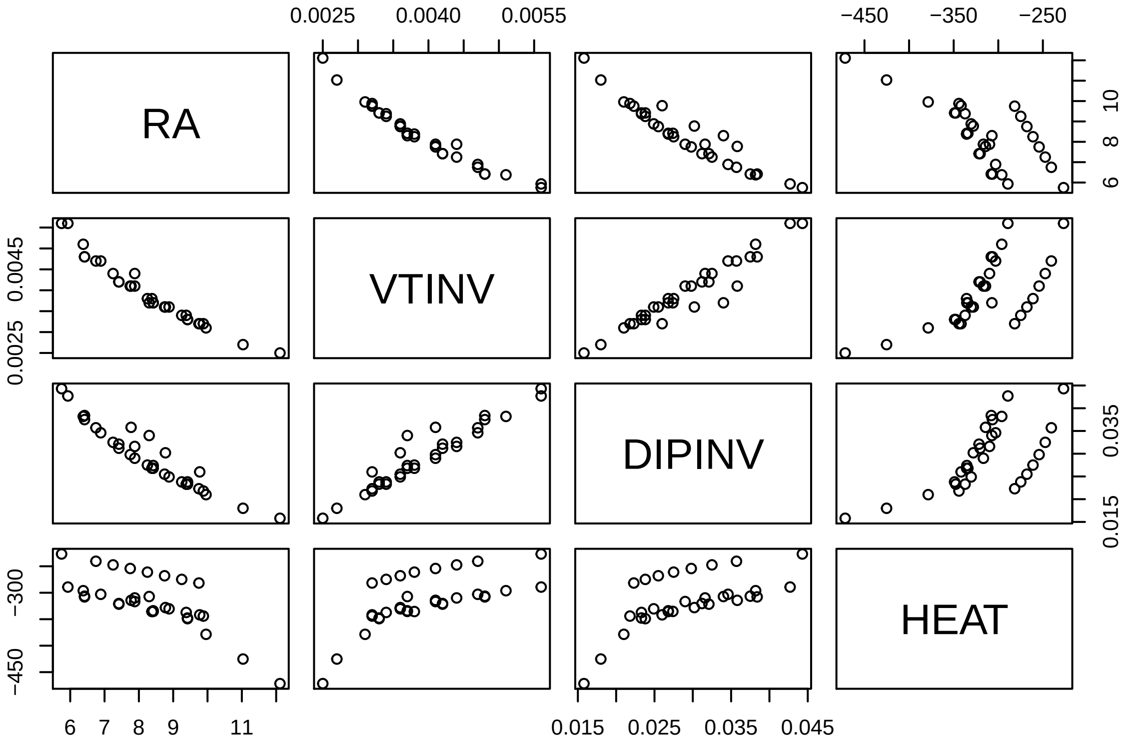
```
## 'data.frame':    32 obs. of  6 variables:
##  $ RA    : num  6.38 6.89 7.77 7.88 7.88 ...
```

```
##  $ VTINV : num  0.0051 0.0047 0.0041 0.0044 0.0041 0.0038 0.0036 0.0036 0.0034 0.0032 ...
##  $ DIPINV: num  0.0382 0.0346 0.0358 0.0316 0.029 0.0268 0.0249 0.0302 0.0233 0.0218 ...
##  $ HEAT  : num  -296 -303 -314 -310 -317 ...
##  $ KPOINT: num  7 16 11 20.8 21 31.5 31 25 38.2 40.5 ...
##  $ GROUP : int  1 1 1 1 1 1 1 1 1 1 ...
```

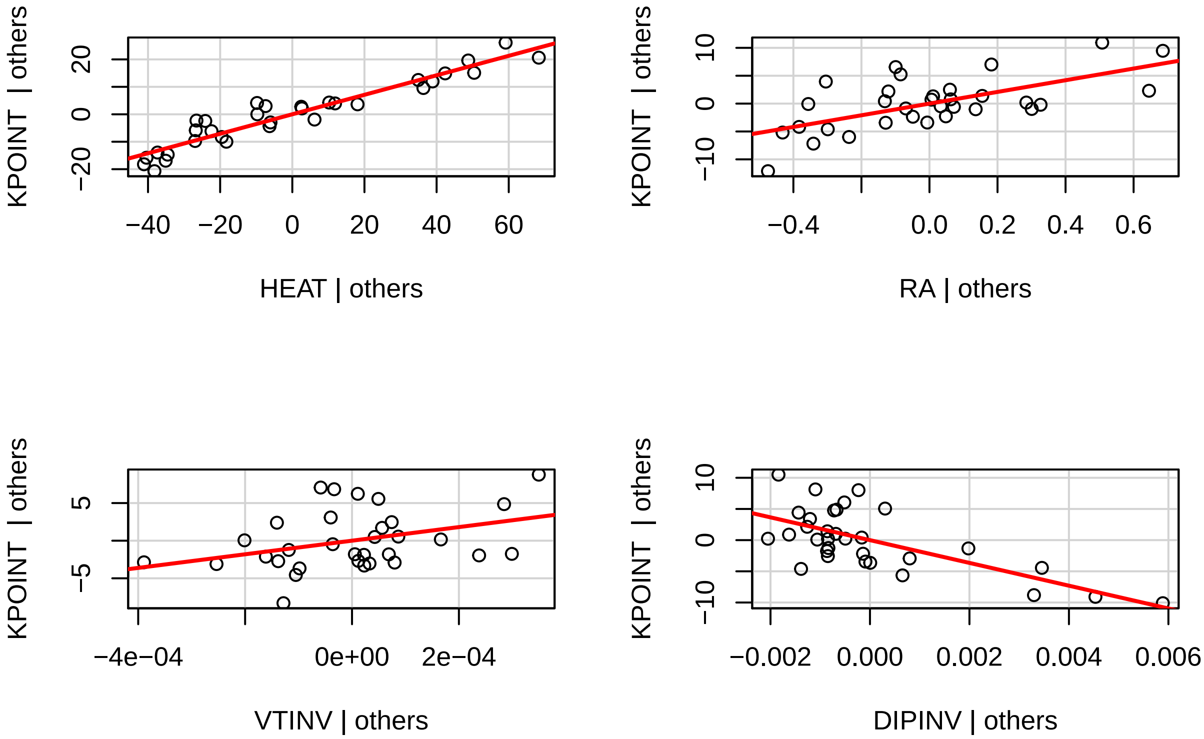
```r
#build initial model
model = lm(KPOINT ~ RA + HEAT + VTINV + DIPINV, data = dat)

#Take a look at model summary
summary(model)
```

```
##
## Call:
## lm(formula = KPOINT ~ RA + HEAT + VTINV + DIPINV, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1451 -2.7920 -0.4552  1.9715  7.5934
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.031e+01  3.368e+01   2.088 0.046369 *
## RA           1.047e+01  2.418e+00   4.331 0.000184 ***
## HEAT         3.550e-01  2.176e-02  16.312 1.66e-15 ***
## VTINV        9.038e+03  4.409e+03   2.050 0.050217 .
## DIPINV      -1.826e+03  3.765e+02  -4.850 4.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.919 on 27 degrees of freedom
## Multiple R-squared:  0.9446, Adjusted R-squared:  0.9363
## F-statistic:   115 on 4 and 27 DF,  p-value: < 2.2e-16
```

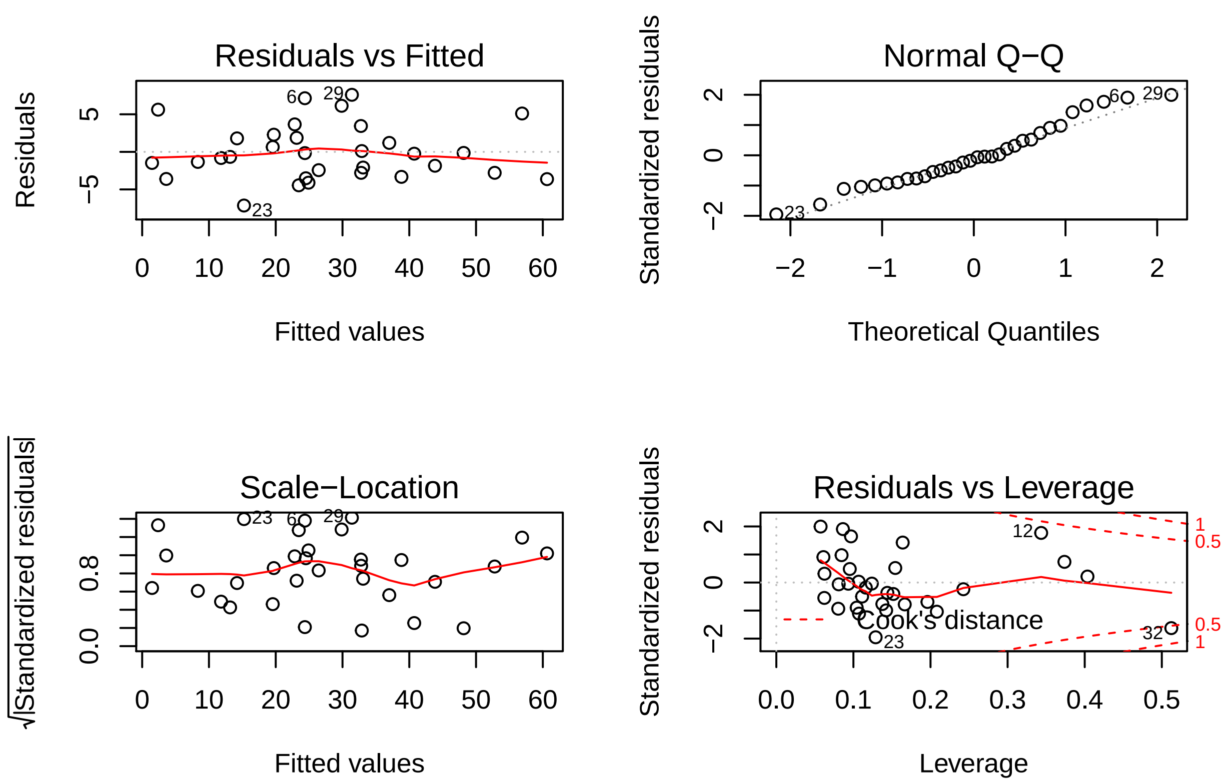
```r
#scatter matrix
scat.mat = dat[,1:4]
pairs(scat.mat)
```




```r
#Added variable plots
library(car)
par(mfrow=c(2,2))
avPlots(model, ~HEAT)
avPlots(model, ~RA)
avPlots(model, ~VTINV)
avPlots(model, ~DIPINV)
```




```r
#Model diagnostics
par(mfrow = c(2,2))
plot(model)
```



The covariates clearly have some colinearity issues. Also, for Heat/RA vs VTINIV/DIPINV there appears to be bands or groups of points that are not accounted for in the data. The added variable plots suggest that each covariate successfully explains some variance in the Krafft point, with relatively small sample sizes, we note that this effect is not difficult to achieve. Lastly, the model diagnostic plots suggest that there are some moderate issues with this model. The normal Q-Q plots show some skewness which is also evident in the residual plots. It actually appears that the highest leverage points are in the center of the theoretical quantiles. This suggests that the skewness may not be affecting the residual plots that drastically. With all this being said, unless we can account for this banding/grouping - the model can be useful.

(b)

This suggests that the skewness in the response variable is most evident for both RA and VTINV. This implies that the residual error in our model could be drastic against these two. Therefore, I suggest we consider a log transformation of the responses or some power transformation for RA and VTINV.

(c)

Using r as a model selection criterion is flawed. We see that there may be some nonlinear behavior between the covariates and the response at the boundary. Thus, if we only consider the linear correlation, represented by r, we do not penalize/reward models that fail/succeed in explaining this tail behavior. Moreover, r would simply choose a more complex model in this case because a more complex model would be guaranteed to explain more behavior than a simple one. (This is why we must consider regularization ideas).

For the same reason as r, standard deviation s, measures deviation from a linear relationship. So for all the reasons that r would not be a good criterion s would also be a poor measure of model performance.

Using F statistics and ANOVA/ANCOVA type analysis is only valid if our assumptions are valid. Here we see that without exploring transforming any variables that this methodology could be flawed. This framework, however, is the most powerful of those suggested thus far.

This last suggestion is a good one. As we discussed above, there needs to be some form regularization to the method. By restraining our model to a relatively simple model via examining the proportion of covariate and the number of samples, we allow models to be constructed that explain the variability in the tails of our response in addition to the high variance in the center of the residuals. I would use a combination of this method and ANOCA framework explained above.

Exericse 6.3

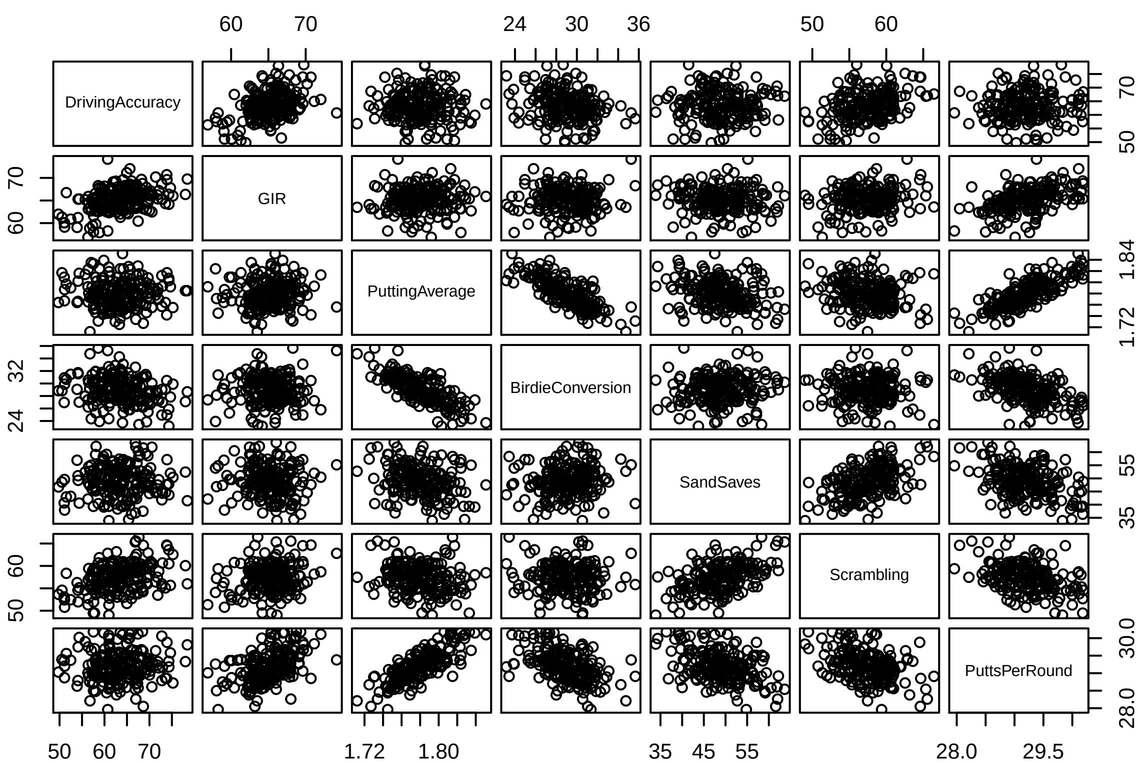
(a)

```
#read in PGA data
pga = read.csv("~/Desktop/Courses/MA 575/book_data/pgatour2006.csv")

#take a peak
str(pga)
```

```
## 'data.frame':    196 obs. of  12 variables:
##  $ Name              : Factor w/ 196 levels "Aaron Baddeley",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ TigerWoods        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PrizeMoney        : int  60661 262045 3635 17516 16683 107294 50620 57273 86782 23396 ..
##  $ AveDrivingDistance: num  288 301 303 289 288 ...
##  $ DrivingAccuracy   : num  60.7 62 51.1 66.4 63.2 ...
##  $ GIR               : num  58.3 69.1 59.1 67.7 64 ...
##  $ PuttingAverage    : num  1.75 1.77 1.79 1.78 1.76 ...
##  $ BirdieConversion  : num  31.4 30.4 29.9 29.3 29.3 ...
##  $ SandSaves         : num  54.8 53.6 37.9 45.1 52.4 ...
##  $ Scrambling        : num  59.4 57.9 50.8 54.8 57.1 ...
##  $ BounceBack        : num  19.3 19.4 16.8 17.1 18.2 ...
##  $ PuttsPerRound     : num  28 29.3 29.2 29.5 28.9 ...
```

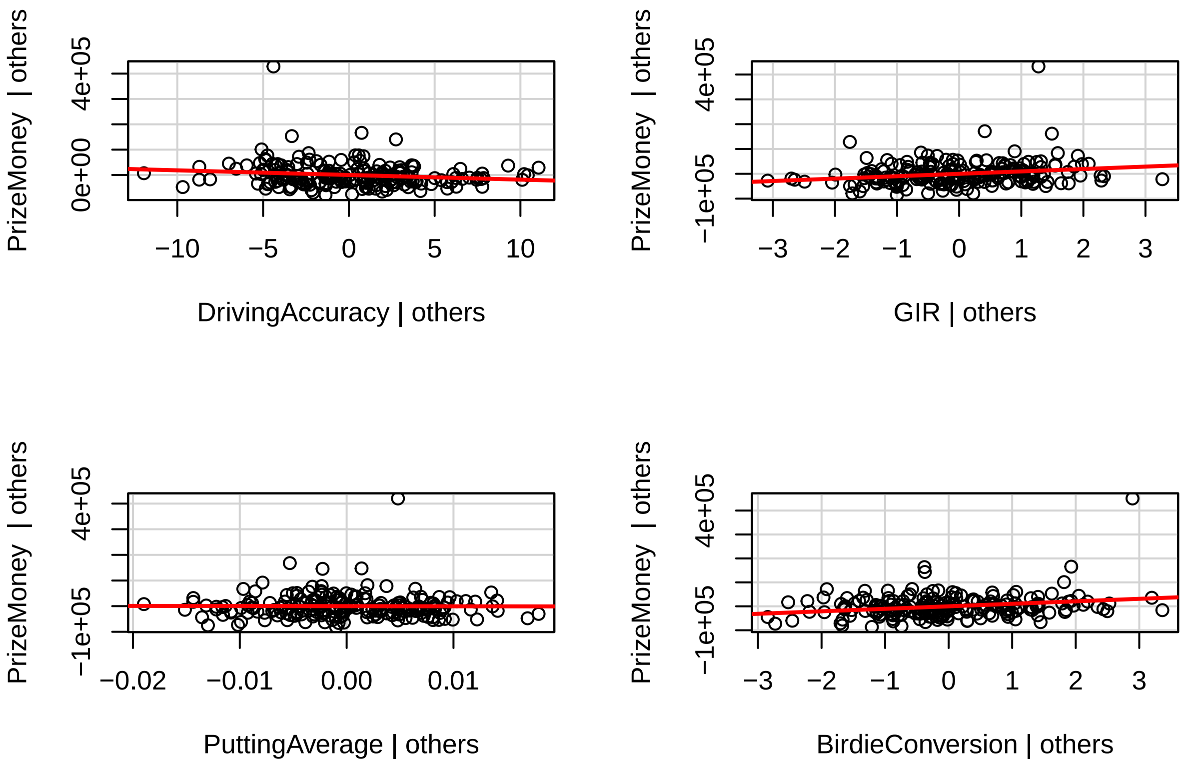
```r
#omg Tiger was so good he get's his own degree or freedom...

#Build model
model = lm(PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + SandSaves +
```

```r
#normal diagnostic stuff
scat.mat = pga[,c(5:10, 12)]
pairs(scat.mat, gap = 0.4)
```

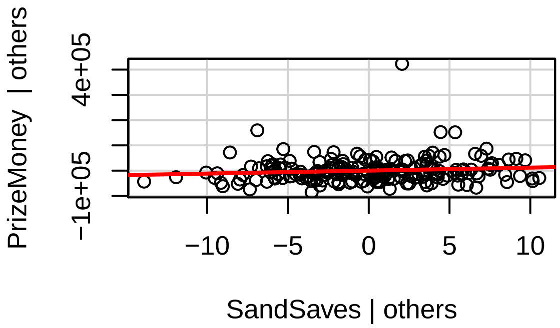


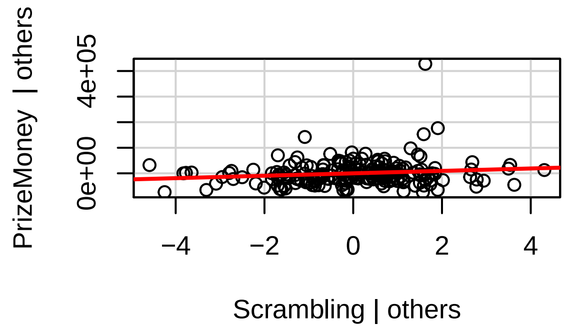

```r
#Added variable plots
library(car)
par(mfrow=c(2,2))
avPlots(model, ~DrivingAccuracy)
```

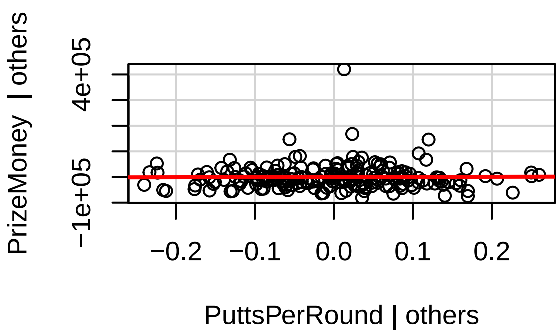
```r
avPlots(model, ~GIR)
avPlots(model, ~PuttingAverage)
avPlots(model, ~BirdieConversion)
```



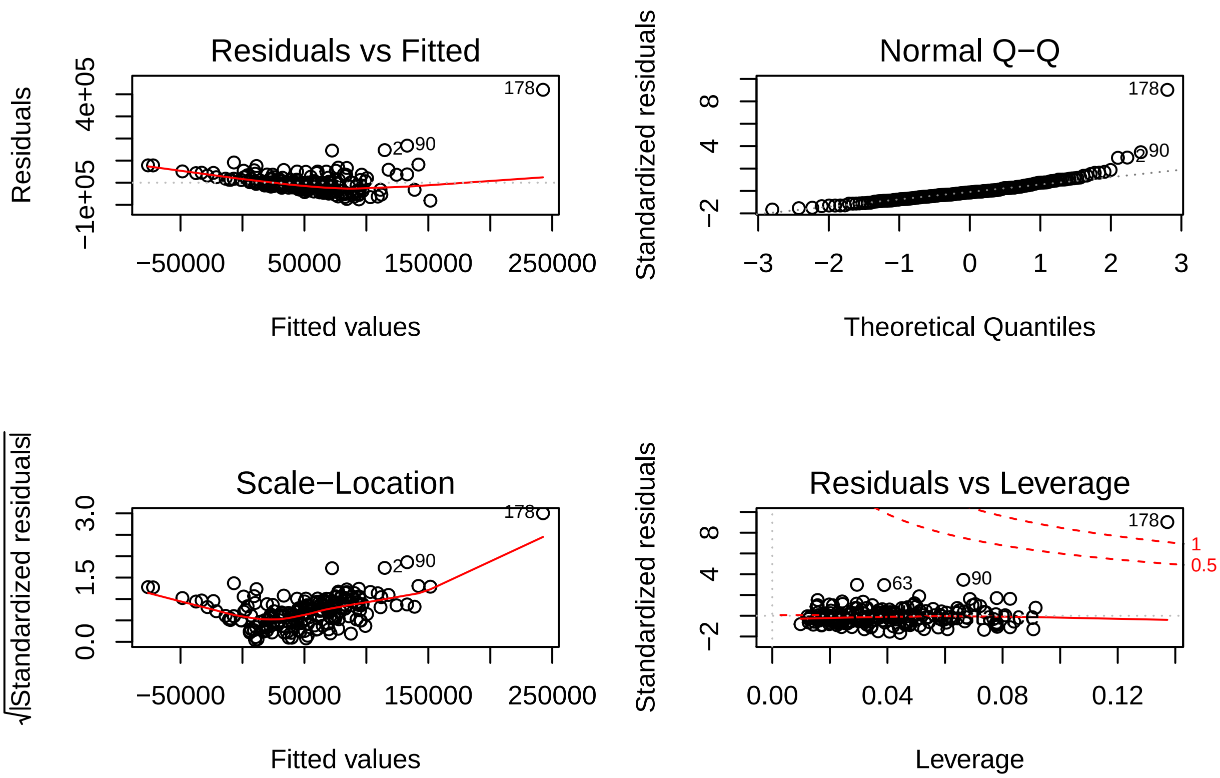

```r
par(mfrow=c(2,2))
avPlots(model, ~SandSaves)
avPlots(model, ~Scrambling)
avPlots(model, ~PuttsPerRound)

#Model diagnostics
par(mfrow = c(2,2))
```








```r
plot(model)
```



The scatter matrix shows that there is linear relationships between Putting Average and Birdie Conversion, Putting Average and Putts Per Round, and small correlations between Sand Saves and Scrambling. All three of these relationships does not come as a surprise. Other covariates are relatively uncorrelated. From the added variable plots, its hard to distinguish any clearly significant pairwise variables. It appears that GIR, Birdie Conversions, Sand Saves, and Scrambling all explain additional variance in prize money while Putts Per Round, Putting Average add no additional information and Driving accuracy could be either. There are some very clear issues in our model diagnostics. First, there is some serious heteroskedasticty issues. Secondly, it appears that participant 178 (Tiger Woods) is a horrible, horrible leverage point.

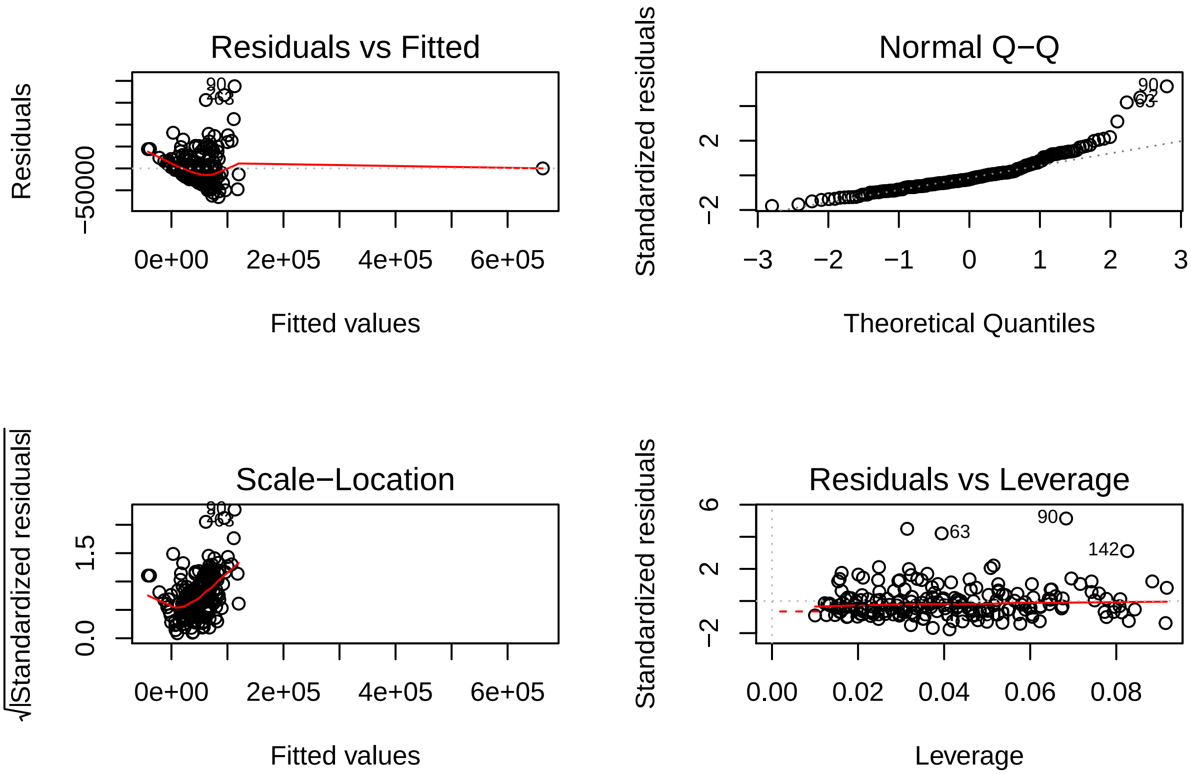
I initial disagree with the analyst. I believe we should add a factor variable for Tiger Woods (model Alien Golfers vs Human Golfers) and then possibly apply a log transformation to account for the larger purse events.

(b)

We will follow the suggestions above.

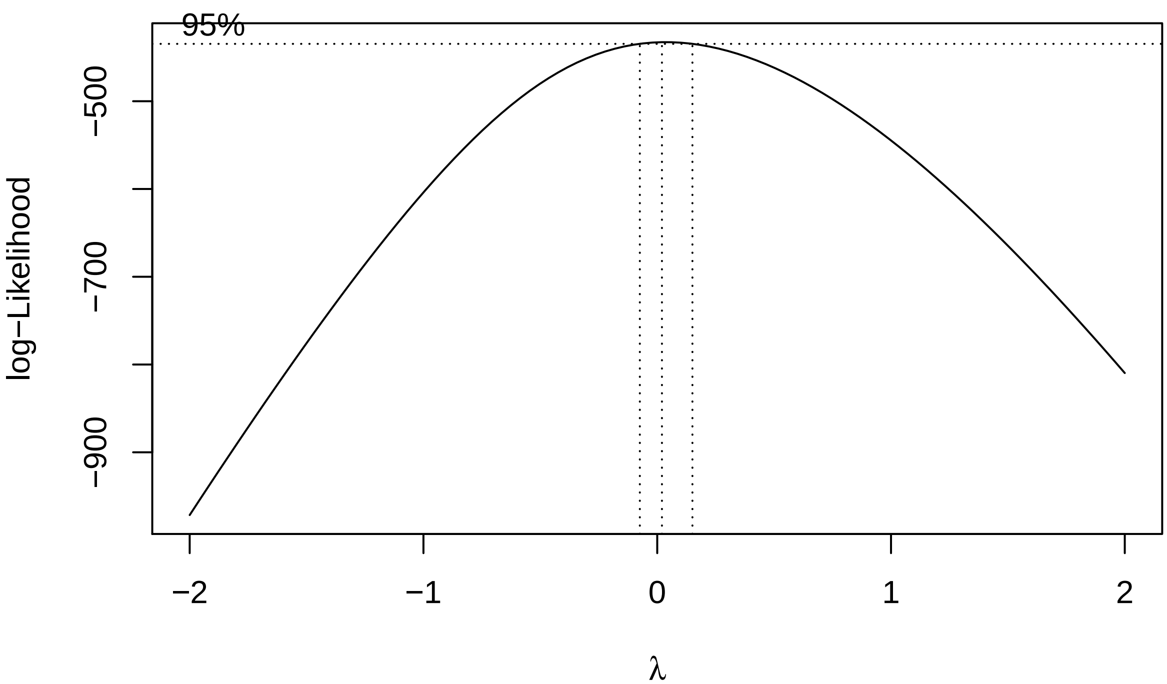
```
#tiger model
tigermodel = lm(PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + SandSa

#model diagnostics
par(mfrow=c(2,2))
plot(tigermodel)
```

```
## Warning: not plotting observations with leverage one:
##   178

## Warning: not plotting observations with leverage one:
##   178
```



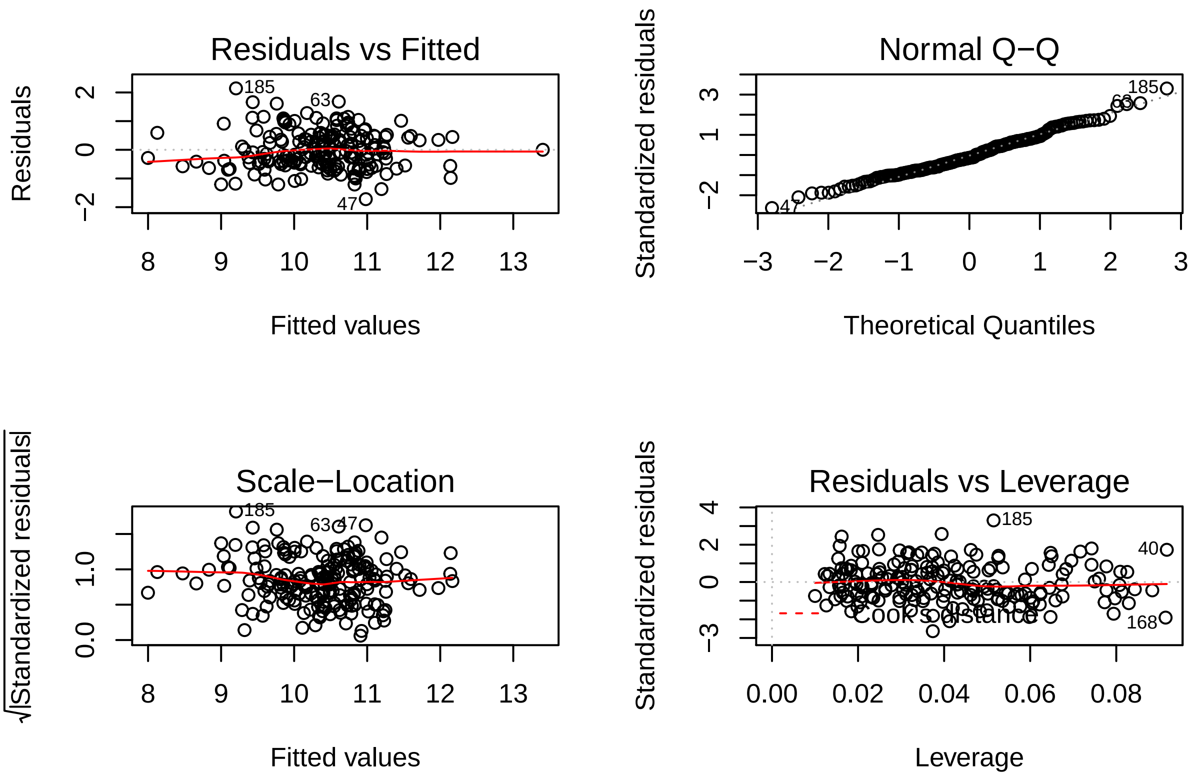
Here we see we fix some issues with the leverage due to Tiger Woods. Now, we still see some Skewness in the Normal Q-Q plot and a huge good leverage point in the Prize Money. We will try to remedy this problem with taking a log transformation.

```r
#Box Cox to confirm log transform
library(MASS)
boxcox(tigermodel)
```



```
#tiger model
logtigermodel = lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion

#model diagnostics
par(mfrow=c(2,2))
plot(logtigermodel)
```

```
## Warning: not plotting observations with leverage one:
##   178

## Warning: not plotting observations with leverage one:
##   178
```




```
#summary of model
summary(logtigermodel)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound +
##     TigerWoods, data = pga)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72152 -0.48488 -0.09094  0.44418  2.14193
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.025605   7.919751   0.003 0.997424
## DrivingAccuracy  -0.003650   0.011845  -0.308 0.758289
## GIR               0.199796   0.044113   4.529 1.05e-05 ***
## PuttingAverage   -0.420963   6.933873  -0.061 0.951654
## BirdieConversion  0.158275   0.041205   3.841 0.000168 ***
## SandSaves         0.015214   0.009893   1.538 0.125770
## Scrambling        0.051839   0.031983   1.621 0.106739
## PuttsPerRound    -0.342557   0.474818  -0.721 0.471535
## TigerWoods       -0.087245   0.716639  -0.122 0.903234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6657 on 187 degrees of freedom
## Multiple R-squared:  0.5577, Adjusted R-squared:  0.5388
## F-statistic: 29.48 on 8 and 187 DF,  p-value: < 2.2e-16
```

We see here that we have a very strong model. The Normal Q-Q plot is almost perfect, the residuals plots are relatively uncorrelated. There appears to be more variance in lower prize events as to be expected (not all elite golfers may attend these events). Therefore, on average, this is a strong model but for lower prized events we could see more variance in our estimates.

(c)

Here made Tiger Woods his own group. In any chance of analyzing normal of average performance, we needed to remove him from the equation. You can see this almost immediately after we added his factor group. There is also a single outlier in the prize money events. It appears to be a good leverage point however, so we include it in the model. This may however affect the higher variance in the lower priced events. We maybe should consider a weighted MLR that takes into account the size of the event as a proxy for the Prize Money. This may account for the small - nonconstant variance.

(d)

As stated above, higher variance in lower purse events could be an issue for prediction. Moreover, having Tiger Woods as his own group we have no sense of variability for his performance. This implies that any ANCOVA is impossible between these sets of golfers. This is a significant issue with the model. We do not have the framework/tools to predict or estimate how his performance affect other golfers in this framework. Instead we opt to consider the game as one that the "Tiger Effect" is modeled separately than the field.

(e)

Our model output suggests that GIR ($\widehat{\beta} \approx 0.199796$) and BirdieConversion ($\widehat{\beta} \approx 0.199796$) is the most important aspect of expected Prize Money. As we saw in the scatter matrix, these variables are clearly not independent of the other variables in the model. Thus if we remove the insignificant variables, you actually remove some of overall affect of the golfer's performance. For instance, Birdie

Conversion and Putting Average are highly correlated. Both serve as proxies for a golfer's putting game. By removing a piece of information about the golfer's putting game, our proxies become weaker because we cannot model the overall affect of these variables.