Homework 5

Juwon Lee, Economics and Statistics, UCLA

2023-02-23

tinytex::install_tinytex()

1.

The analyst was so impressed with your answers to Exercise 5 in Section 3.4 that your advice has been sought regarding the next stage in the data-anlysis, namely and analysis of the effects of different aspects of a car on its suggested retail price. Data are available for all 234 cars on the following variables:
Y = Suggested Retail Price, x_1 = Engine size, x_2 = Cylinders,
x_3 = Horse power, x_4= Highway mpg, x_5 = Weight, x_6 = Wheel Base, x_7 = Hyprid.

The model is
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + e$.

(a) Decide is it a valid model. Give reasons to support your answer.

```
car <- read.csv("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regression/cars

lm_hw5_1 <- lm(SuggestedRetailPrice~EngineSize+Cylinders+Horsepower+HighwayMPG+Weight+WheelBase+Hybrid,

summary(lm_hw5_1)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
##     Horsepower + HighwayMPG + Weight + WheelBase + Hybrid, data = car)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17436  -4134    173   3561  46392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68965.793  16180.381  -4.262 2.97e-05 ***
## EngineSize   -6957.457   1600.137  -4.348 2.08e-05 ***
## Cylinders     3564.755    969.633   3.676 0.000296 ***
## Horsepower     179.702     16.411  10.950  < 2e-16 ***
## HighwayMPG     637.939    202.724   3.147 0.001873 **
## Weight          11.911      2.658   4.481 1.18e-05 ***
## WheelBase       47.607    178.070   0.267 0.789444
## Hybrid         431.759   6092.087   0.071 0.943562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7533 on 226 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7751
## F-statistic: 115.7 on 7 and 226 DF,  p-value: < 2.2e-16
```

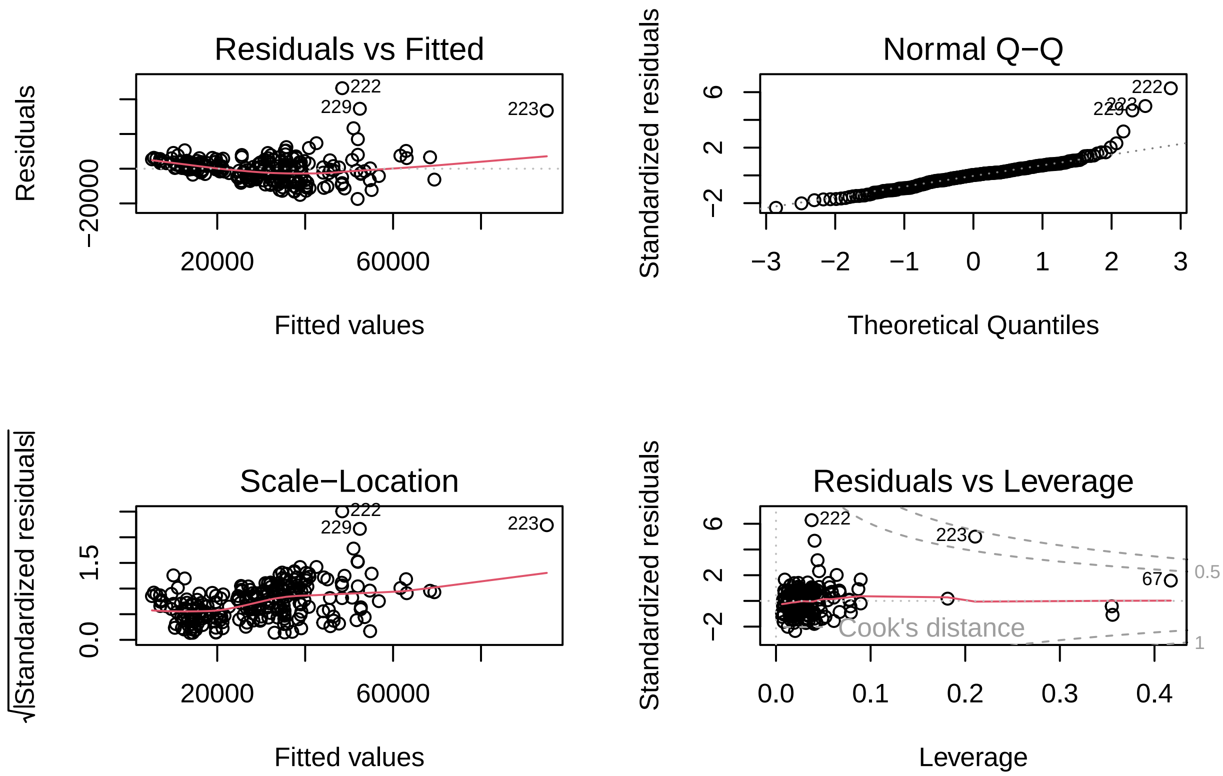
It is valid model, because $adj - R^2 = 0.7751$, very well.

(b) The plot of residuals against fitted values produces a curved pattern. Describe what, if anything can be learned about model from this plot.

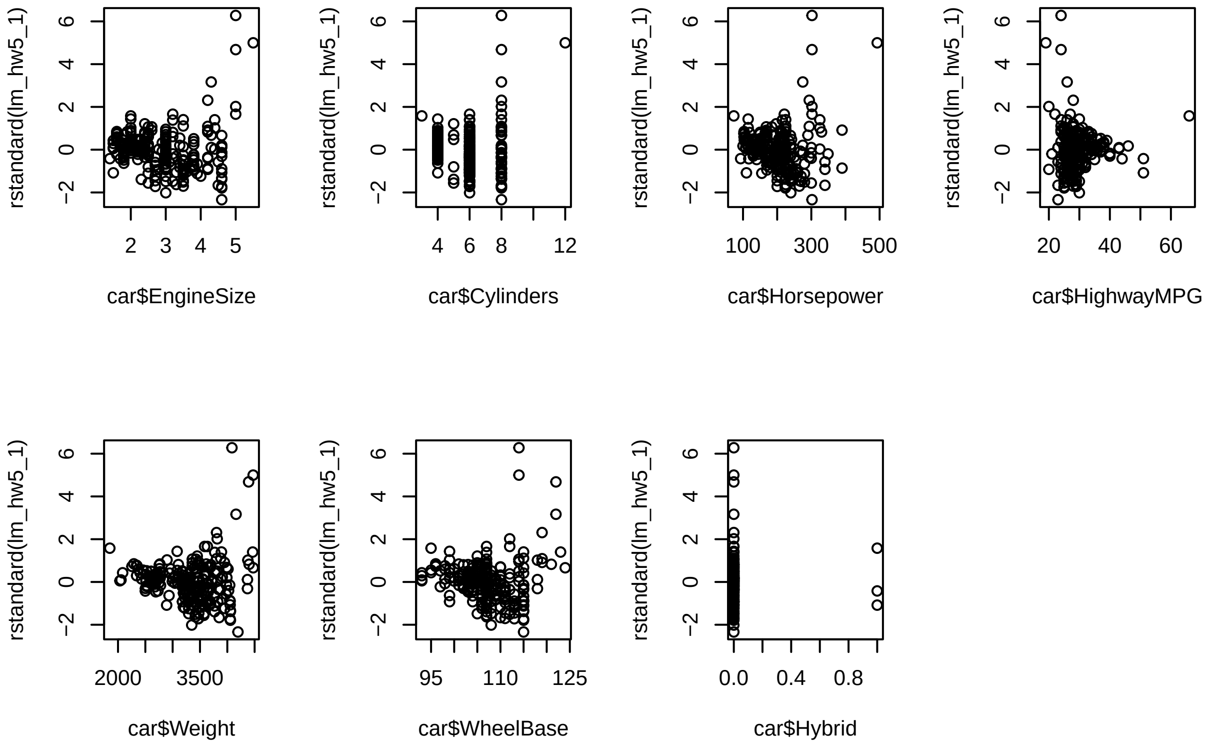
```
library(car)
```

```
## Loading required package: carData
```

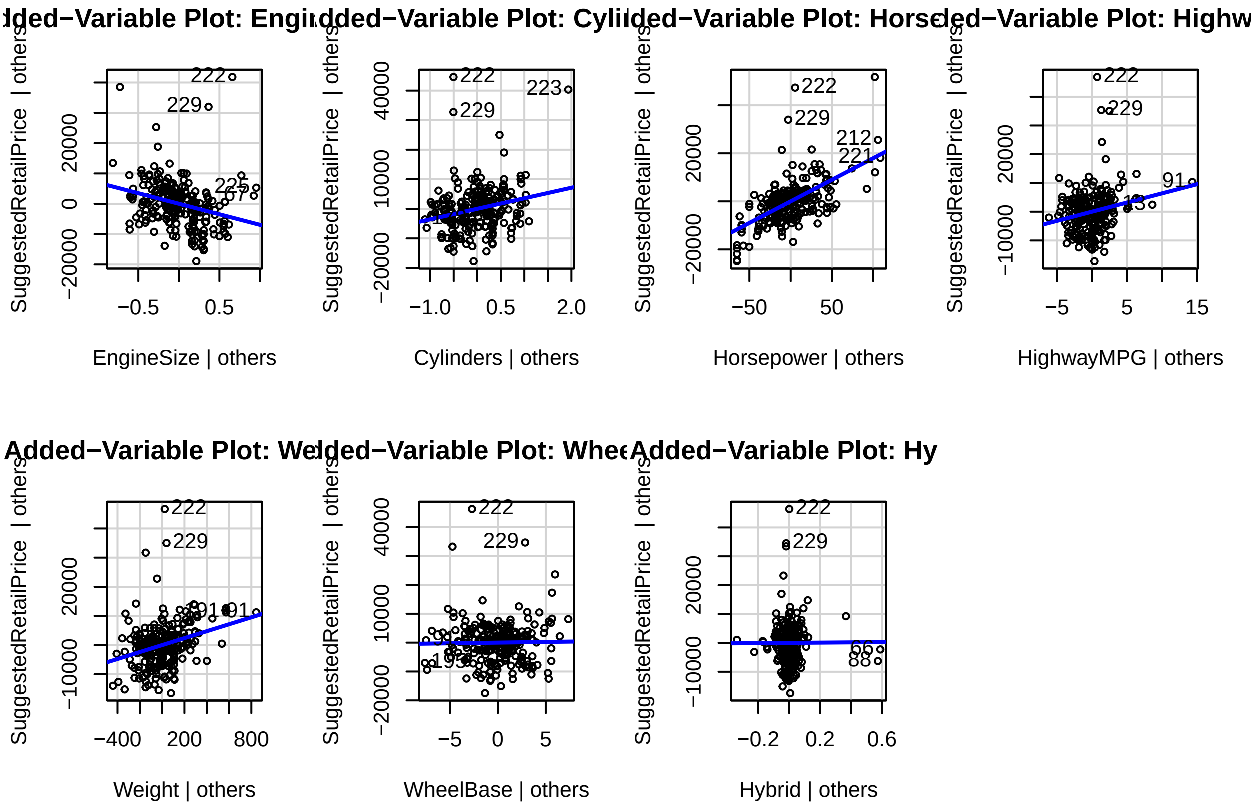
```
par(mfrow=c(2,2))
plot(lm_hw5_1)
```




```
par(mfrow=c(2,4))
plot(car$EngineSize, rstandard(lm_hw5_1))
plot(car$Cylinders, rstandard(lm_hw5_1))
plot(car$Horsepower, rstandard(lm_hw5_1))
plot(car$HighwayMPG, rstandard(lm_hw5_1))
plot(car$Weight, rstandard(lm_hw5_1))
plot(car$WheelBase, rstandard(lm_hw5_1))
plot(car$Hybrid, rstandard(lm_hw5_1))

par(mfrow=c(2,4))
```




```r
avPlot(lm_hw5_1, variable='EngineSize', ask=FALSE)
avPlot(lm_hw5_1, variable='Cylinders', ask=FALSE)
avPlot(lm_hw5_1, variable='Horsepower', ask=FALSE)
avPlot(lm_hw5_1, variable='HighwayMPG', ask=FALSE)
avPlot(lm_hw5_1, variable='Weight', ask=FALSE)
avPlot(lm_hw5_1, variable='WheelBase', ask=FALSE)
avPlot(lm_hw5_1, variable='Hybrid', ask=FALSE)
```



Thus, the data having small, or large fitted values have larger residuals.

(c) Identify any bad leverage points for model.

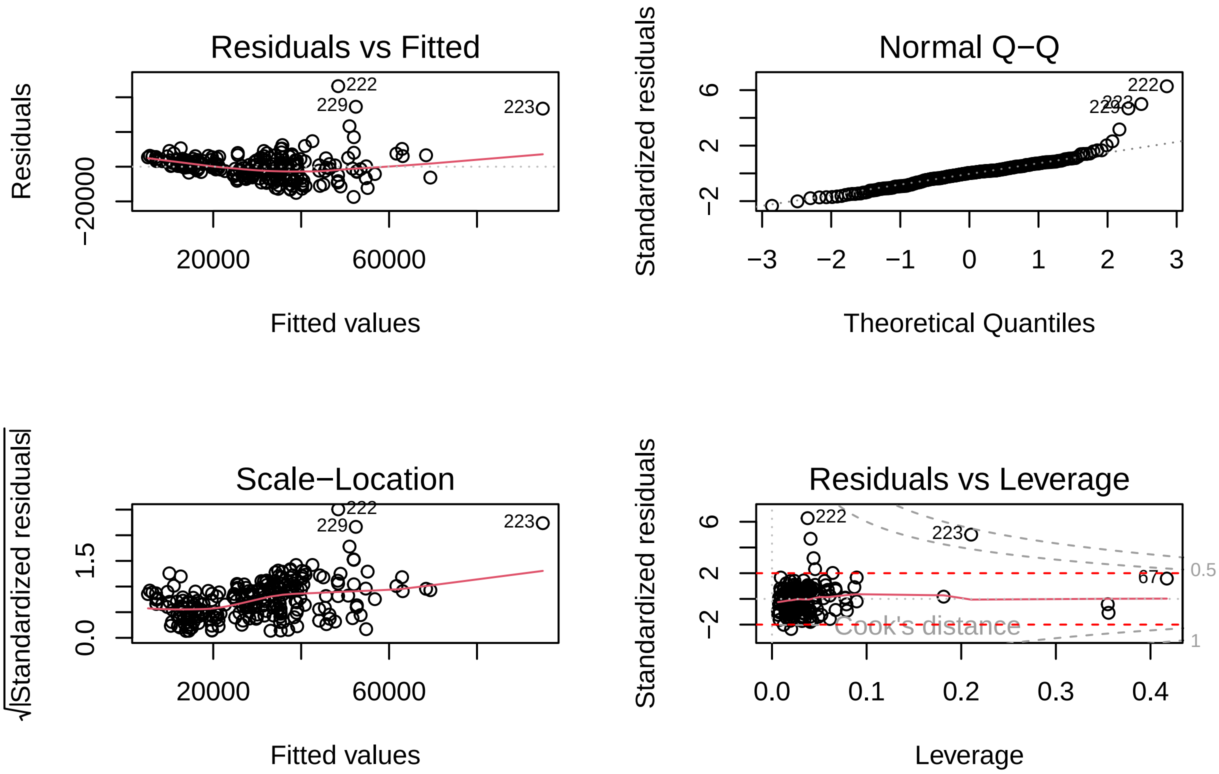
Bad leverage points $\rightarrow \begin{cases} h_{ii} > \frac{4}{n} \\ |\gamma_i| > 2 \end{cases}$

```
car[hatvalues(lm_hw5_1) > 2 * (7+1)/length(car) & abs(rstandard(lm_hw5_1)) > 2]
```

```
## data frame with 0 columns and 234 rows
```

Thus, there doesn't exist any bad leverage points.

```
par(mfrow=c(2,2))
plot(lm_hw5_1)
abline(-2, 0, col='red', lty='dashed')
abline(2, 0, col='red', lty='dashed')
abline(v=2 * (7+1)/length(car), col='blue', lty='dashed')
```



The multivariate version of the Box-Cox method was used to transform the predictors, while a log transformation was used for the response variable to improve interpretability. This resulted in the following model: $\log(Y) = \beta_0 + \beta_1 x_1^{0.25} + \beta_2 \log(x_2) + \beta_3 \log(x_3) + \beta_4(\frac{1}{x_4}) + \beta_5 x_5 + \beta_6 \log(x_6) + \beta_7 x_7 + e$.

(d) Decide whether this is a valid model.

```
car[5] <- car[5]^(1/4)

car[9] <- car[9]^(-1)

head(car)
```

```
##                  Vehicle.Name Hybrid SuggestedRetailPrice DealerCost EngineSize
## 1           Chevrolet Aveo 4dr      0                11690      10965   1.124683
## 2 Chevrolet Aveo LS 4dr hatch      0                12585      11802   1.124683
## 3       Chevrolet Cavalier 2dr      0                14610      13697   1.217883
## 4       Chevrolet Cavalier 4dr      0                14810      13884   1.217883
## 5    Chevrolet Cavalier LS 2dr      0                16385      15357   1.217883
## 6            Dodge Neon SE 4dr      0                13670      12849   1.189207
##   Cylinders Horsepower CityMPG HighwayMPG Weight WheelBase Length Width
## 1         4        103      28 0.02941176   2370        98    167    66
## 2         4        103      28 0.02941176   2348        98    153    66
## 3         4        140      26 0.02702703   2617       104    183    69
## 4         4        140      26 0.02702703   2676       104    183    68
## 5         4        140      26 0.02702703   2617       104    183    69
## 6         4        132      29 0.02777778   2581       105    174    67
```

```
lm_hw5_2 <- lm(SuggestedRetailPrice~EngineSize+Cylinders+Horsepower+HighwayMPG+Weight+WheelBase+Hybrid,

summary(lm_hw5_2)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
##     Horsepower + HighwayMPG + Weight + WheelBase + Hybrid, data = car)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18006  -3709    251   2945  45844
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28508.68   16287.21   1.750   0.0814 .
## EngineSize   -81838.75   13530.84  -6.048 6.01e-09 ***
## Cylinders      4001.50     881.74   4.538 9.22e-06 ***
## Horsepower      175.37      16.57  10.587  < 2e-16 ***
## HighwayMPG  -183402.09  196528.85  -0.933   0.3517
## Weight           10.21       2.54   4.020 7.93e-05 ***
## WheelBase       208.79     180.85   1.155   0.2495
## Hybrid         9644.02    4767.91   2.023   0.0443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7381 on 226 degrees of freedom
## Multiple R-squared:  0.7906, Adjusted R-squared:  0.7841
## F-statistic: 121.9 on 7 and 226 DF,  p-value: < 2.2e-16
```

It is a valid model, having $adj - R^2 = 0.7882$.

(e) To obtain a final model, the analyst wants to simply remove the two insignificant predictors $(1/x_4)$ and $\log(x_6)$. Perform a partial F-test to see if this is a sensible strategy.

```r
lm_hw5_3 <- lm(SuggestedRetailPrice~EngineSize+Cylinders+Horsepower+Weight+Hybrid, data=car)

summary(lm_hw5_3)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
##     Horsepower + Weight + Hybrid, data = car)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18384  -3953    332   3219  45501
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37708.453  11465.938   3.289  0.00117 **
## EngineSize  -76763.103  12882.854  -5.959 9.58e-09 ***
## Cylinders     3892.780    880.463   4.421 1.52e-05 ***
## Horsepower     168.833     15.672  10.773  < 2e-16 ***
## Weight          10.837      1.992   5.439 1.37e-07 ***
## Hybrid       11994.057   4405.552   2.722  0.00698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 7401 on 228 degrees of freedom
## Multiple R-squared:  0.7876, Adjusted R-squared:  0.7829
## F-statistic: 169.1 on 5 and 228 DF,  p-value: < 2.2e-16
```

```r
anova(lm_hw5_3, lm_hw5_2)
```

```
## Analysis of Variance Table
##
## Model 1: SuggestedRetailPrice ~ EngineSize + Cylinders + Horsepower +
##     Weight + Hybrid
## Model 2: SuggestedRetailPrice ~ EngineSize + Cylinders + Horsepower +
##     HighwayMPG + Weight + WheelBase + Hybrid
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    228 1.2489e+10
## 2    226 1.2314e+10  2 175602038 1.6115 0.2019
```

Then p-value $= 0.1929 > 0.05$, so that we cannot reject the null.
Thus, we cannot say that using a full model is better.

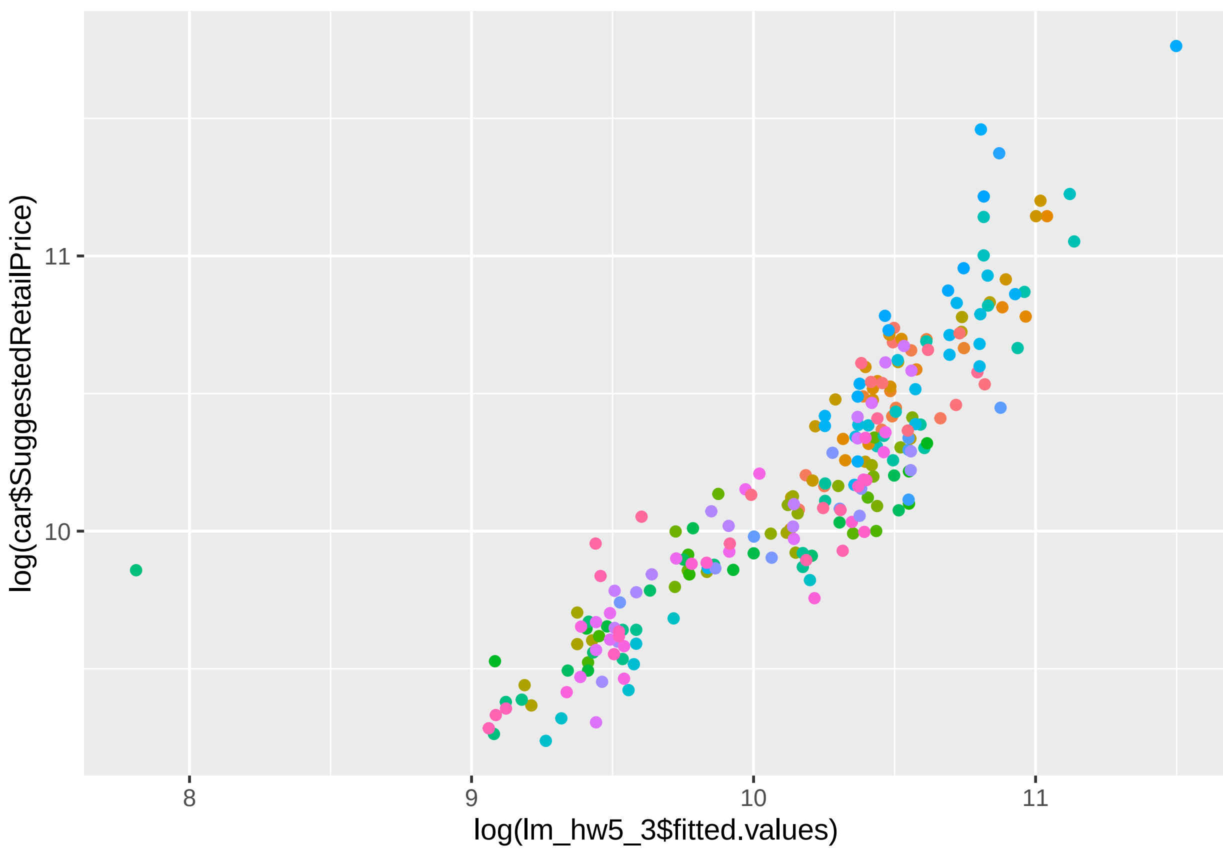
(f) The analyst's boss has complained about the model saying that it fails to take account of the manufacturer of the vehicle (e.g. BMW vs Toyota). Describe how model could be expanded in order to estimate the effect of manufacturer on suggested retail price.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.4      v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

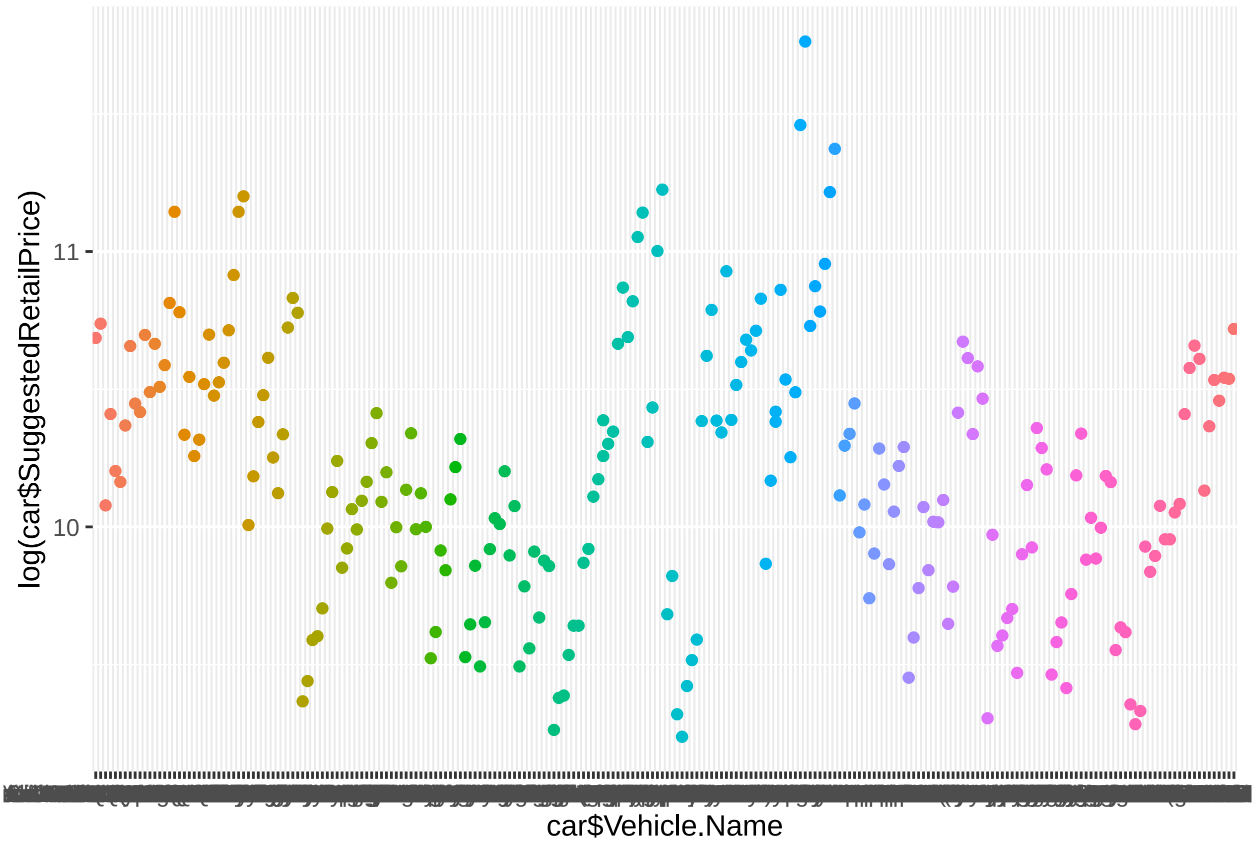
```r
par(mfrow=c(1,2))

ggplot(data=car) +
  geom_point(mapping=aes(x=log(lm_hw5_3$fitted.values), y=log(car$SuggestedRetailPrice), color=car$Vehi
```

```
## Warning: Use of `car$SuggestedRetailPrice` is discouraged.
## i Use `SuggestedRetailPrice` instead.

## Warning: Use of `car$Vehicle.Name` is discouraged.
## i Use `Vehicle.Name` instead.
```




```
ggplot(data=car) +
  geom_point(mapping=aes(x=car$Vehicle.Name, y=log(car$SuggestedRetailPrice), color=car$Vehicle.Name), s
```

```
## Warning: Use of `car$Vehicle.Name` is discouraged.
## i Use `Vehicle.Name` instead.

## Warning: Use of `car$SuggestedRetailPrice` is discouraged.
## i Use `SuggestedRetailPrice` instead.

## Warning: Use of `car$Vehicle.Name` is discouraged.
## i Use `Vehicle.Name` instead.
```



2.

An avid fan of the PGA tour with limited background in statistics has sought your help in answering one of the age-old questions in golf, namely, *what is the relative importance of each different aspect of the game on average prize money in professional golf?*

Y (Prize Money) = Average prize money per tournament.
x_1 (Driving Accuracy) = The percent of time a player is able to hit the fairway with his tee shot.
x_2 (GIR) = Greens in Regulation is the percent of time a player was able to hit the green in regulation.
x_3 (Putting Average) = Putting performance on those holes where the green is hit in regulation (GIR).
x_4 (Birdie Conversion) = The percent of time a player makes birdie or better after hitting the green in regulation.
x_5 (SandSaves) = The percent of time a player was able to get "up and down" once in a greenside sand bunker.
x_6 (Scrambling) = The percent of time that a player misses the green in regulation, but sill makes par or better.
x_7 (PuttsPerRound) = The average total number of putts per round.

(a) A statistician from Australia has recommended to the analyst that they not transform any of the predictor variables but that they transform Y using the $\log$ transformation. Do you agree with this recommendation? Give reasons to support your answer.

```
golf <- read.csv("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regression/pga

lm_hw5_2_1 <- lm(PrizeMoney~DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+SandSaves+Scrambling+Pu

summary(lm_hw5_2_1)
```

```
##
## Call:
## lm(formula = PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
##     data = golf)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81239 -26260  -6521  17539 420230
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1165233.1   587382.9  -1.984 0.048737 *
## DrivingAccuracy     -1835.8      889.2  -2.065 0.040326 *
## GIR                  9671.3     3309.4   2.922 0.003899 **
## PuttingAverage     -47435.3   521566.4  -0.091 0.927631
## BirdieConversion    10426.0     3049.6   3.419 0.000771 ***
## SandSaves            1182.1      744.8   1.587 0.114184
## Scrambling           4741.3     2400.8   1.975 0.049749 *
## PuttsPerRound        5267.5    35765.7   0.147 0.883070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50140 on 188 degrees of freedom
## Multiple R-squared:  0.4064, Adjusted R-squared:  0.3843
## F-statistic: 18.39 on 7 and 188 DF,  p-value: < 2.2e-16
```

```
library(MASS)
```

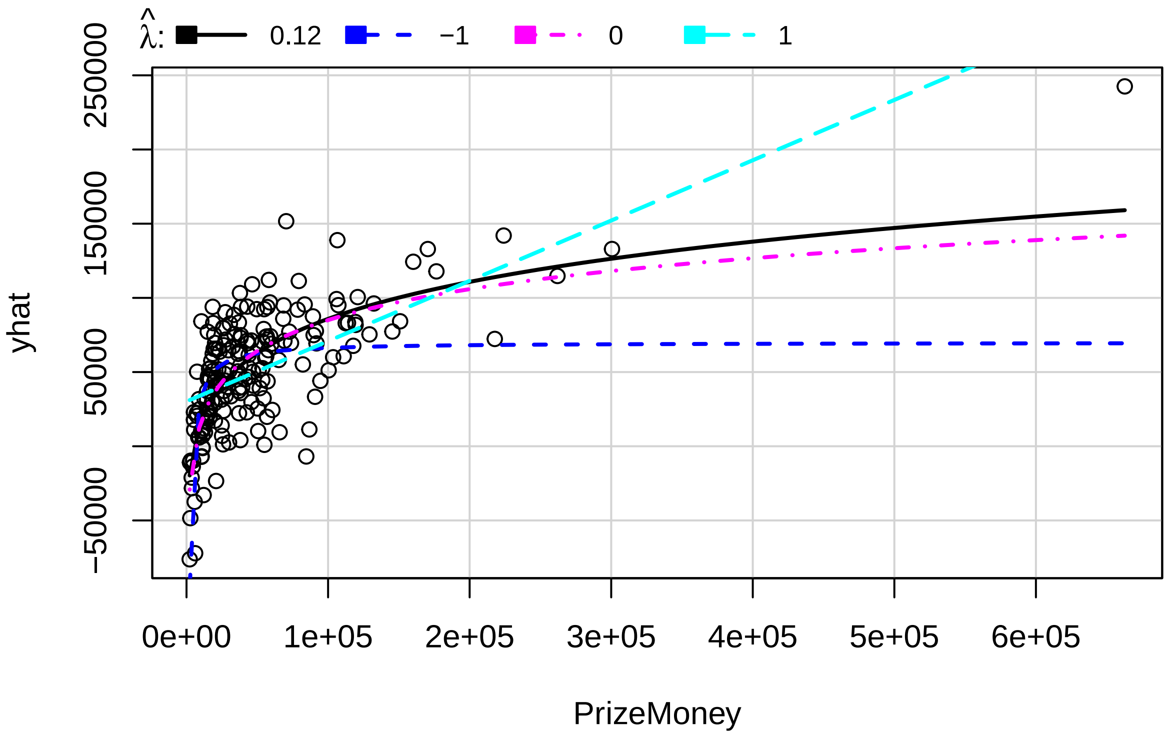
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
power_hw5_2_1 <- powerTransform(cbind(golf$DrivingAccuracy, golf$GIR, golf$PuttingAverage, golf$BirdieC

summary(power_hw5_2_1)
```

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.2751           1      -0.8984       1.4486
## Y2    1.7972           1       0.3116       3.2828
## Y3    1.0999           1      -3.4385       5.6384
## Y4    0.8033           1      -0.2707       1.8772
## Y5    1.0064           1       0.0634       1.9493
## Y6    0.7495           1      -0.6752       2.1742
## Y7    0.0079           1      -3.2327       3.2486
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                         LRT df     pval
## LR test, lambda = (0 0 0 0 0 0 0) 13.46843  7 0.061485
##
## Likelihood ratio test that no transformations are needed
```

```
##                                          LRT df    pval
## LR test, lambda = (1 1 1 1 1 1 1) 3.687514  7 0.81498
```

```
library(car)

inverseResponsePlot(lm_hw5_2_1)
```




```
##       lambda          RSS
## 1  0.1191664 153353617043
## 2 -1.0000000 202266980718
## 3  0.0000000 154049980760
## 4  1.0000000 192096985076
```

Thus, only transforming Y is a reasonable way.

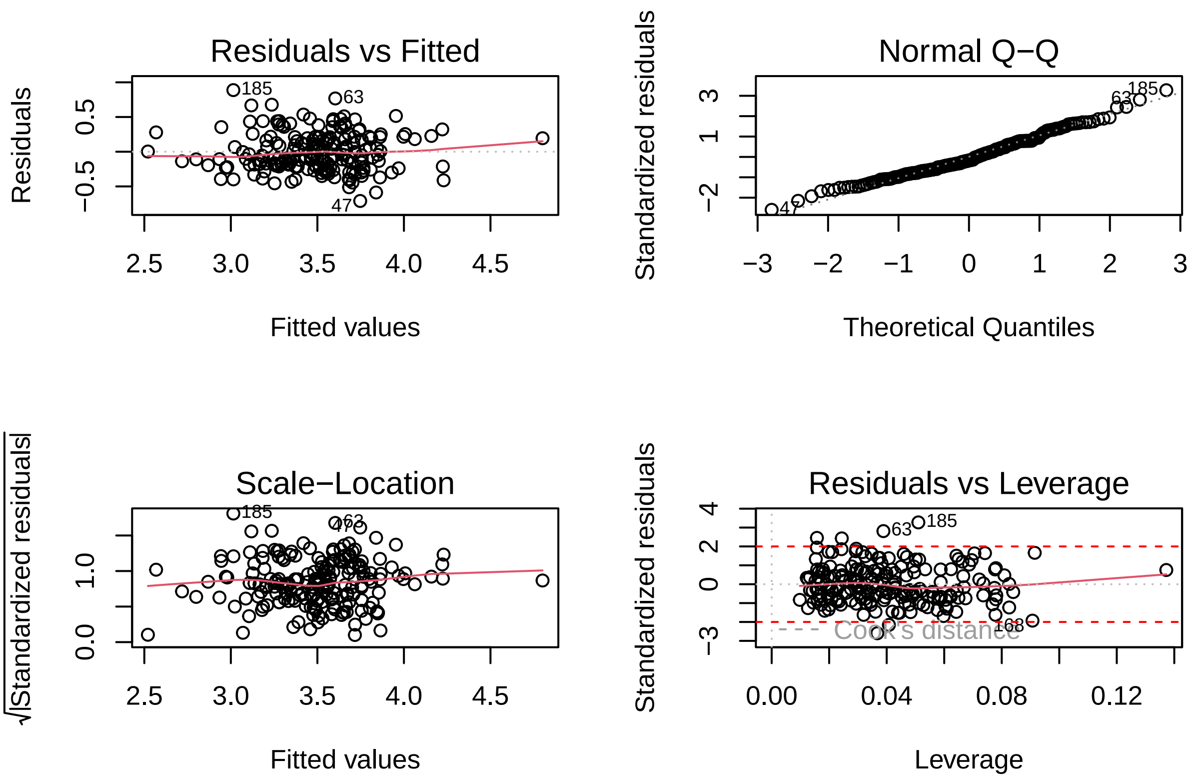
(b) Develop a valid full regression model containing all seven potential predictor variables listed above. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots plots of standardized residuals, and any other relevant diagnostic plots.

```
lm_hw5_2_2 <- lm((PrizeMoney)^(0.12)~DrivingAccuracy+GIR+PuttingAverage+BirdieConversion+SandSaves+Scran

summary(lm_hw5_2_2)
```

```
##
## Call:
## lm(formula = (PrizeMoney)^(0.12) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
##     data = golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.70953 -0.18983 -0.04663  0.19450  0.88757
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.002067   3.260968  -0.307 0.758962
## DrivingAccuracy  -0.003212   0.004936  -0.651 0.516020
## GIR               0.081528   0.018373   4.437 1.55e-05 ***
## PuttingAverage   -0.379049   2.895575  -0.131 0.895989
## BirdieConversion  0.065552   0.016931   3.872 0.000149 ***
## SandSaves         0.007052   0.004135   1.705 0.089773 .
## Scrambling        0.022693   0.013329   1.703 0.090300 .
## PuttsPerRound    -0.119381   0.198560  -0.601 0.548411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2784 on 188 degrees of freedom
## Multiple R-squared:  0.552,	Adjusted R-squared:  0.5354
## F-statistic:  33.1 on 7 and 188 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(lm_hw5_2_2)
abline(-2, 0, col='red', lty='dashed')
abline(2, 0, col='red', lty='dashed')
abline(v=2 * (7+1)/length(golf), col='blue', lty='dashed')
```

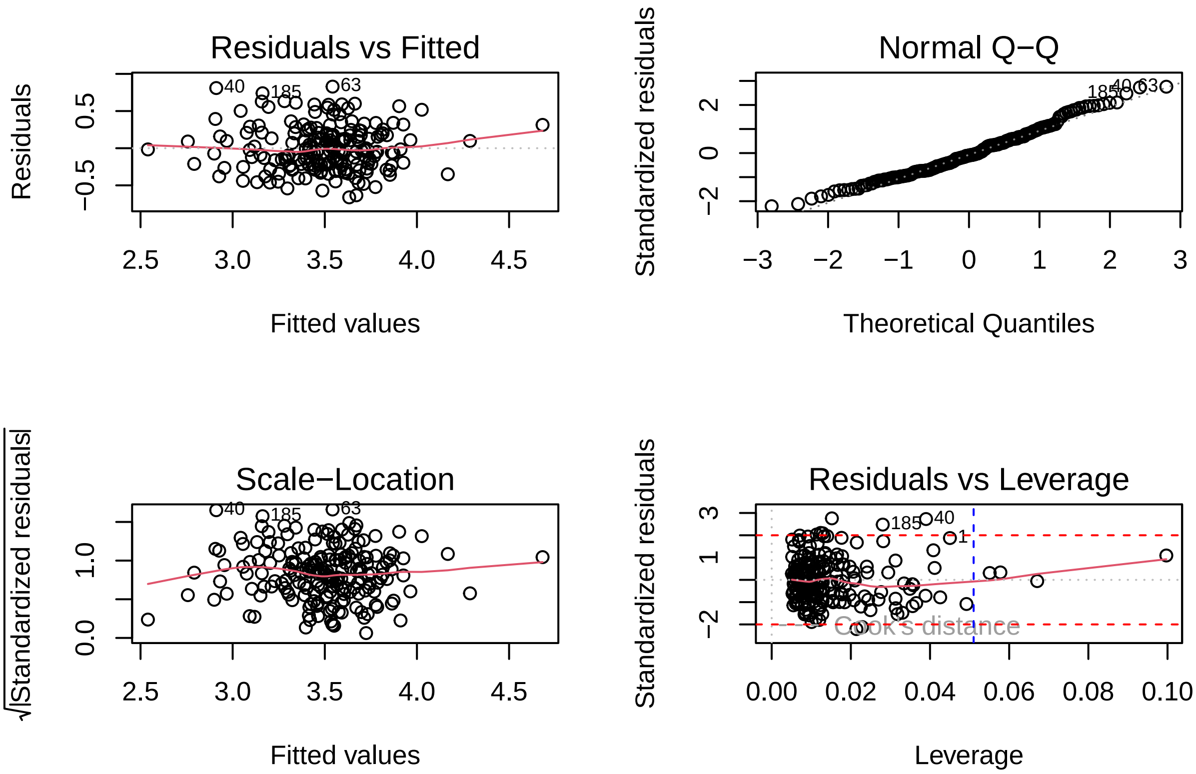



```r
lm_hw5_3 <- lm((PrizeMoney)^(0.12)~GIR+BirdieConversion, data=golf)
summary(lm_hw5_3)
```

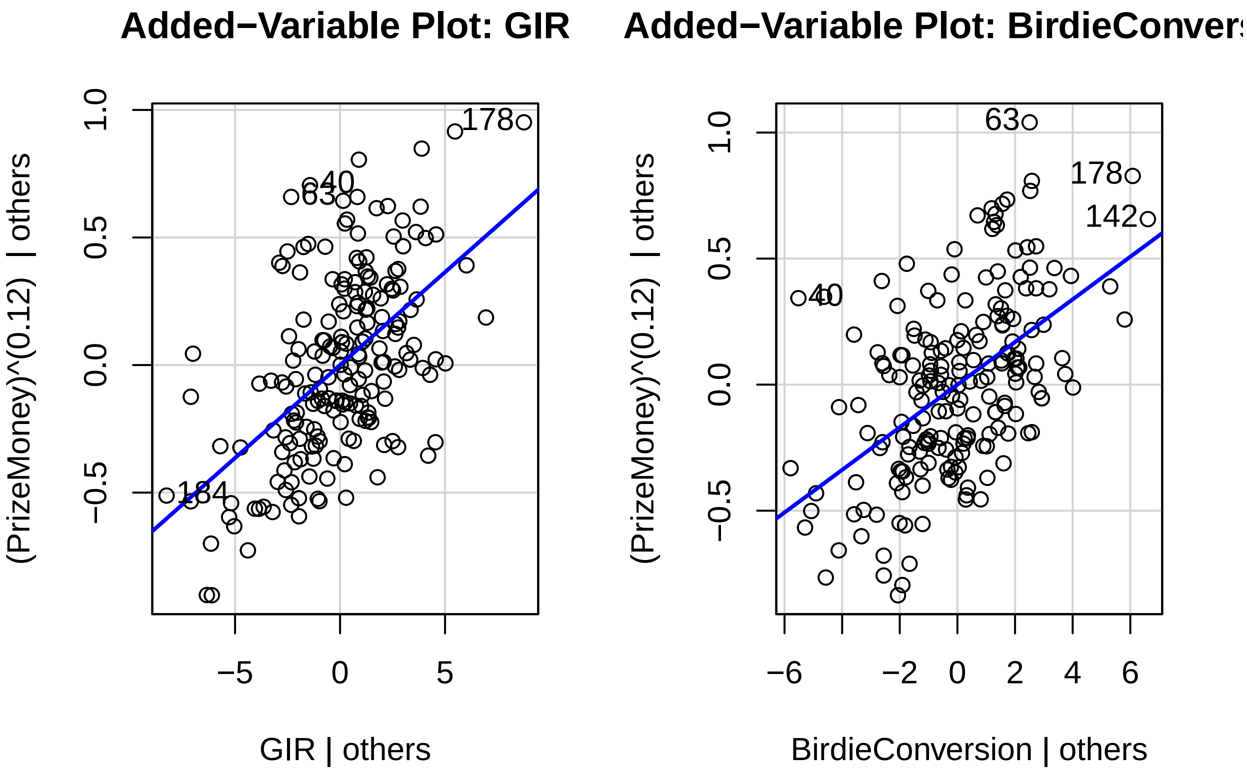
```
##
## Call:
```

```
## lm(formula = (PrizeMoney)^(0.12) ~ GIR + BirdieConversion, data = golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66110 -0.22379 -0.03036  0.18082  0.82852
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.700314   0.585824  -6.316 1.80e-09 ***
## GIR               0.072838   0.007965   9.145  < 2e-16 ***
## BirdieConversion  0.084545   0.009827   8.604 2.66e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3027 on 193 degrees of freedom
## Multiple R-squared:  0.4563, Adjusted R-squared:  0.4507
## F-statistic:    81 on 2 and 193 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(lm_hw5_3)
abline(2,0,col='red',lty='dashed')
abline(-2,0,col='red',lty='dashed')
abline(v=2*(4+1)/(length(golf$GIR)),col='blue',lty='dashed')
```




```r
library(car)

par(mfrow=c(1,2))
avPlot(lm_hw5_3, variable='GIR', ask=FALSE)
avPlot(lm_hw5_3, variable='BirdieConversion', ask=FALSE)
```



(c) Identify any points that should be investigated. Give one or more reasons to support each point chosen.

The point 40, 63, 185 should be investigated, because
(i) They have tail-parts on QQ-plot.
(ii) They are outliers for all of Added-Variable Plots.

(d) Describe any weaknesses in your model.

(i) Only two variables can reject the null on t-test.
(ii) For added-variable plot, three variables doesn't explain Y.

```
summary(lm_hw5_3)$adj.r.squared
```

```
## [1] 0.4507045
```

```
extractAIC(lm_hw5_3)[2]
```

```
## [1] -465.5016
```

```
extractAIC(lm_hw5_3)[2] + 2 * 2 * (2+2) * (2+3) / (length(golf$GIR)-2-1)
```

```
## [1] -465.0871
```

```
extractAIC(lm_hw5_3, k=log(length(golf$GIR)))[2]
```

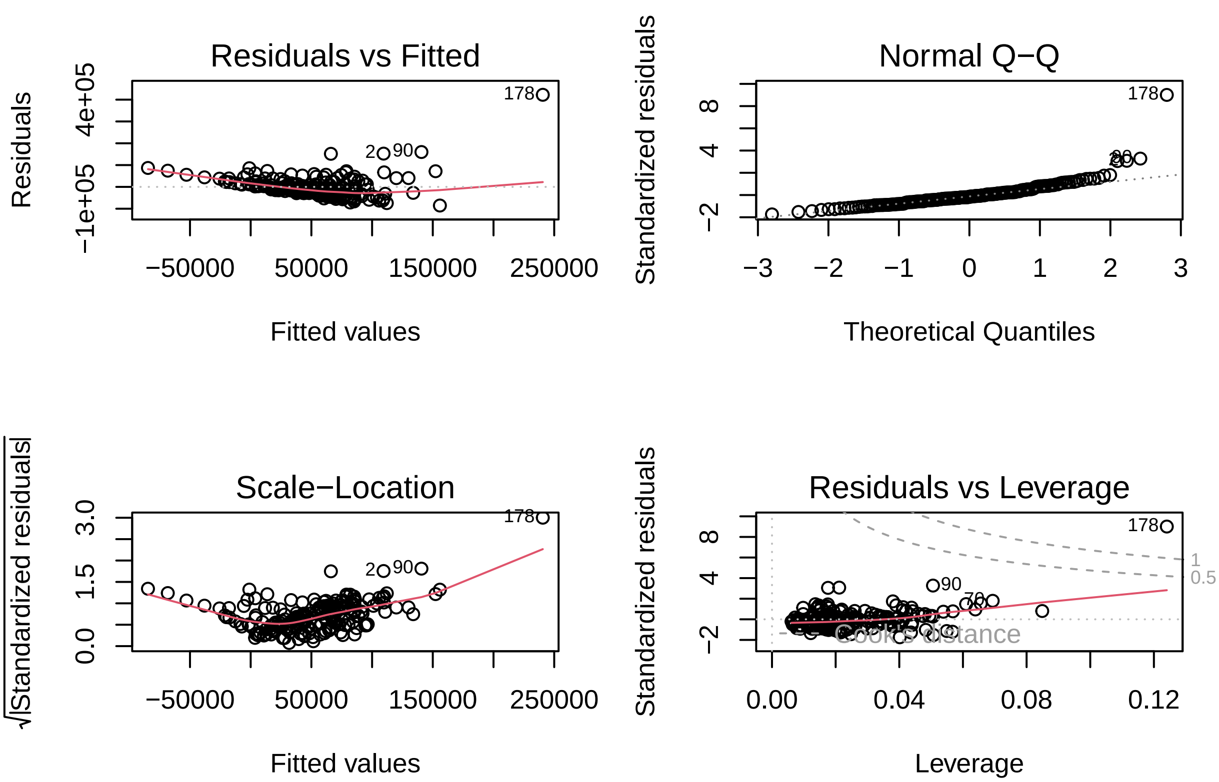
```
## [1] -455.6672
```

(e) The golf fan wants to remove all predictors with insignificant t-values from the full model in a single step. Explain why you would not recommend this approach.

```r
lm_hw5_4 <- lm(PrizeMoney~DrivingAccuracy+GIR+BirdieConversion+Scrambling, data=golf)

summary(lm_hw5_4)
```

```
##
## Call:
## lm(formula = PrizeMoney ~ DrivingAccuracy + GIR + BirdieConversion +
##     Scrambling, data = golf)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -85429 -27959  -7833  15674 422173
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1094996.9   109585.4  -9.992  < 2e-16 ***
## DrivingAccuracy      -1964.1      815.7  -2.408    0.017 *
## GIR                   9742.9     1465.9   6.646 3.06e-10 ***
## BirdieConversion     10670.5     1703.7   6.263 2.44e-09 ***
## Scrambling            5670.4     1239.4   4.575 8.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50080 on 191 degrees of freedom
## Multiple R-squared:  0.3984, Adjusted R-squared:  0.3858
## F-statistic: 31.62 on 4 and 191 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(lm_hw5_4)
```



```r
library(car)

par(mfrow=c(2,2))
avPlot(lm_hw5_4, variable='GIR', ask=FALSE)
avPlot(lm_hw5_4, variable='BirdieConversion', ask=FALSE)
avPlot(lm_hw5_4, variable='Scrambling', ask=FALSE)

summary(lm_hw5_4)$adj.r.squared
```

```
## [1] 0.3857836
```

```r
extractAIC(lm_hw5_4)[2]
```

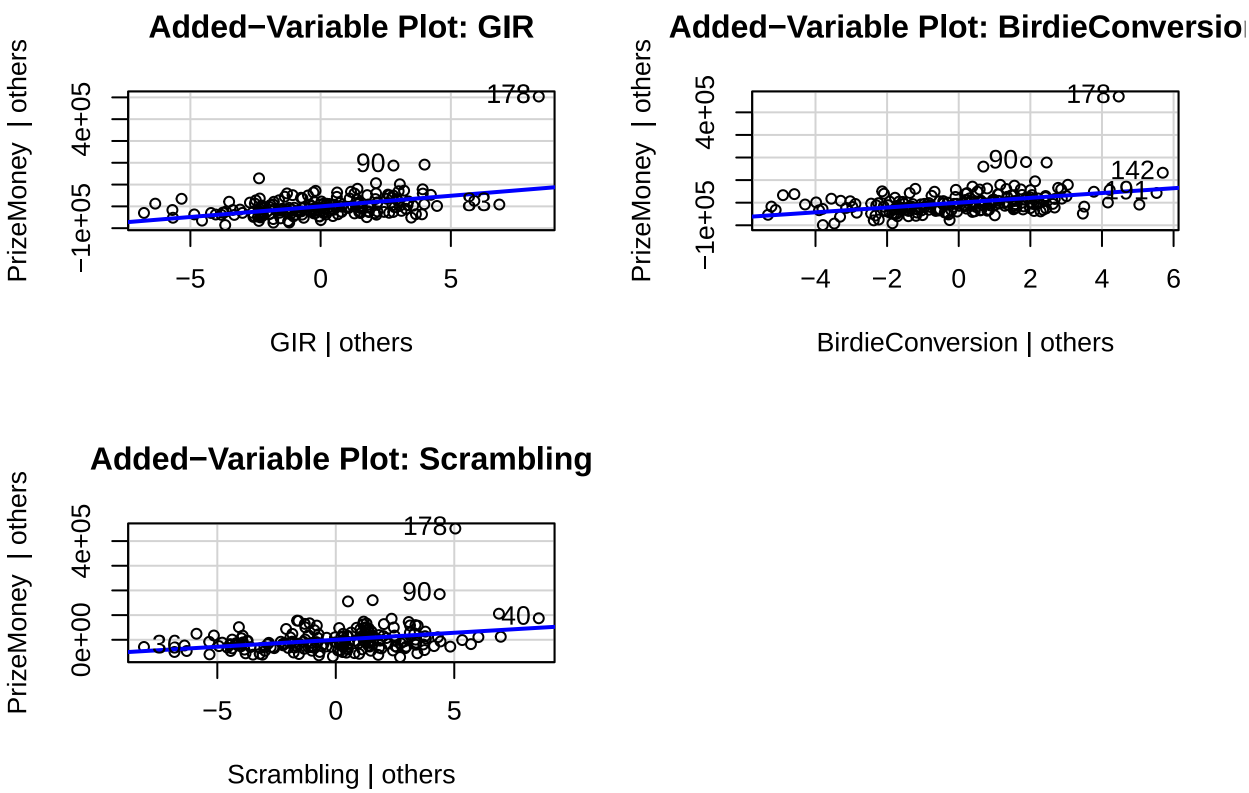
```
## [1] 4246.931
```

```r
extractAIC(lm_hw5_4)[2] + 2 * 3 * (3+2) * (3+3) / (length(golf$GIR)-3-1)
```

```
## [1] 4247.868
```

```r
extractAIC(lm_hw5_4, k=log(length(golf$GIR)))[2]
```

```
## [1] 4263.321
```



Adjusted R^2, AIC, AICc, BIC are poorer than the final model.
It has significant outlier, which is 178, too.

3.

The real data set in this question first appeared in Hald (1952). The data are given in Table 7.5 and can be found on the book web site in the file Haldcement.txt. Interest centers on using variable selection to choose a subset of the predictors to model Y. Throughout this question we shall assume that the full model below is a

valid model for the data
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e.$

```
cement <- read.table("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regressio

lm_hw5_5 <- lm(Y~x1+x2+x3+x4, data=cement)
summary(lm_hw5_5)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x4, data = cement)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -1.6709  0.2508  1.3783  3.9254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## x1            1.5511     0.7448   2.083   0.0708 .
## x2            0.5102     0.7238   0.705   0.5009
## x3            0.1019     0.7547   0.135   0.8959
## x4           -0.1441     0.7091  -0.203   0.8441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

```
summary(lm_hw5_5)$adj.r.squared
```

```
## [1] 0.9735634
```

```
extractAIC(lm_hw5_5)[2]
```

```
## [1] 26.94429
```

```
extractAIC(lm_hw5_5)[2] + 2 * 4 * (4+2) * (4+3) / (length(golf$GIR)-4-1)
```

```
## [1] 28.70345
```

```
extractAIC(lm_hw5_5, k=log(length(golf$GIR)))[2]
```

```
## [1] 43.33486
```

(a) Identify the optimal model or models on R^2_{adj}, AIC, $AICc$, BIC from the approach based on all possible subsets.

```
library(leaps)

xvalues <- cbind(cement$x1, cement$x2, cement$x3, cement$x4)

considerallsusbset <- regsubsets(as.matrix(xvalues), cement$Y)
summary(considerallsusbset)
```

```
## Subset selection object
## 4 Variables  (and intercept)
```

```
##   Forced in Forced out
## a     FALSE      FALSE
## b     FALSE      FALSE
## c     FALSE      FALSE
## d     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          a   b   c   d
## 1  ( 1 ) " " " " " " "*"
## 2  ( 1 ) "*" "*" " " " "
## 3  ( 1 ) "*" "*" " " "*"
## 4  ( 1 ) "*" "*" "*" "*"
```

```
lm_hw5_6_1 <- lm(Y~x4, data=cement)
summary(lm_hw5_6_1)$adj.r.squared
```

```
## [1] 0.6449549
```

```
lm_hw5_6_2 <- lm(Y~x1+x2, data=cement)
summary(lm_hw5_6_2)$adj.r.squared
```

```
## [1] 0.974414
```

```
lm_hw5_6_3 <- lm(Y~x1+x2+x3+x4, data=cement)
summary(lm_hw5_6_3)$adj.r.squared
```

```
## [1] 0.9735634
```

```
lm_hw5_6_4 <- lm(Y~x1+x2+x3+x4, data=cement)
summary(lm_hw5_6_4)$adj.r.squared
```

```
## [1] 0.9735634
```

Thus, it may be the optimal model that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

(b) Identify the optimal model or models based on AIC and BIC from the approach based on forward selection.

```
lm_hw5_6 <- lm(Y~1, data=cement)
forwardAIC <- step(lm_hw5_6, scope=list(lower=~1, upper=~x1+x2+x3+x4), direction='forward', data=cement
```

```
## Start:  AIC=71.44
## Y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 58.852
## + x2    1   1809.43  906.34 59.178
## + x1    1   1450.08 1265.69 63.519
## + x3    1    776.36 1939.40 69.067
## <none>              2715.76 71.444
##
## Step:  AIC=58.85
## Y ~ x4
##
##        Df Sum of Sq    RSS    AIC
## + x1    1    809.10  74.76 28.742
## + x3    1    708.13 175.74 39.853
```

```
## <none>              883.87 58.852
## + x2    1     14.99 868.88 60.629
##
## Step:  AIC=28.74
## Y ~ x4 + x1
##
##        Df Sum of Sq    RSS    AIC
## + x2    1    26.789 47.973 24.974
## + x3    1    23.926 50.836 25.728
## <none>              74.762 28.742
##
## Step:  AIC=24.97
## Y ~ x4 + x1 + x2
##
##        Df Sum of Sq    RSS    AIC
## <none>              47.973 24.974
## + x3    1   0.10909 47.864 26.944
```

Thus, it is an optimal model that $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2$.

(c) Identify the optimal model or models based on AIC and BIC from the approach based on backward elimination.

```r
lm_hw5_6 <- lm(Y~x1+x2+x3+x4, data=cement)
backAIC <- step(lm_hw5_6, direction='backward', data=cement)
```

```
## Start:  AIC=26.94
## Y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x3    1    0.1091 47.973 24.974
## - x4    1    0.2470 48.111 25.011
## - x2    1    2.9725 50.836 25.728
## <none>              47.864 26.944
## - x1    1   25.9509 73.815 30.576
##
## Step:  AIC=24.97
## Y ~ x1 + x2 + x4
##
##        Df Sum of Sq    RSS    AIC
## <none>               47.97 24.974
## - x4    1      9.93  57.90 25.420
## - x2    1     26.79  74.76 28.742
## - x1    1    820.91 868.88 60.629
```

Thus, $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2$.

(d) Carefully explain why the models chosen in (a), (b) & (c) are not all the same.

I think that (a) considers all of the results,
but (b) and (c) have steps, so that the x_4 cannot be eliminated, because
(b) x_4 contains at first, then we consider given x_4.
(c) x_4 contains, because just eliminating x_3 makes well model.

(e) Recommend a final model. Give detailed reasons to support your choice.

I think that $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$ is optimal. This is because x_4 does not increase AIC much, but two methods pick x_4.