Regression Homework 5
Three Major Japanese Automakers–Price of Automobile

Janet Ye: jy1151

April 30, 2014

1 Introduction

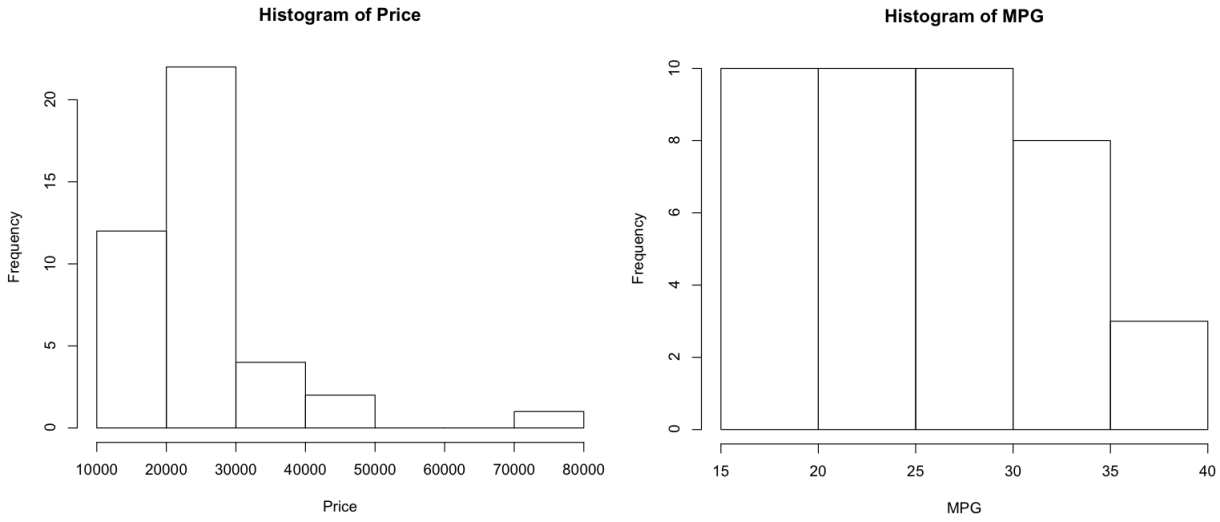
The three major Japanese automakers are Toyota, Honda, and Nissan. What determines the price of a Japanese automobile? Naturally, one would think that the types–with levels car, Truck, SUV, and the Miles Per Gallon (MPG) are the main factors that affect the prices of automobile. In this analysis, we examine the relationship between the prices of the cars and the two predictor variables–type and MPG.

2 The Data

The data are manually recorded from the automakers' websites. All cars in the data set are gasoline fueled. We excluded hybrid models, since Toyota has hybrid cars and hybrid SUV, but Honda and Civic have hybrid cars only. The prices recorded down are manufactures suggested price. For MPG, I averaged city and highway MPG.

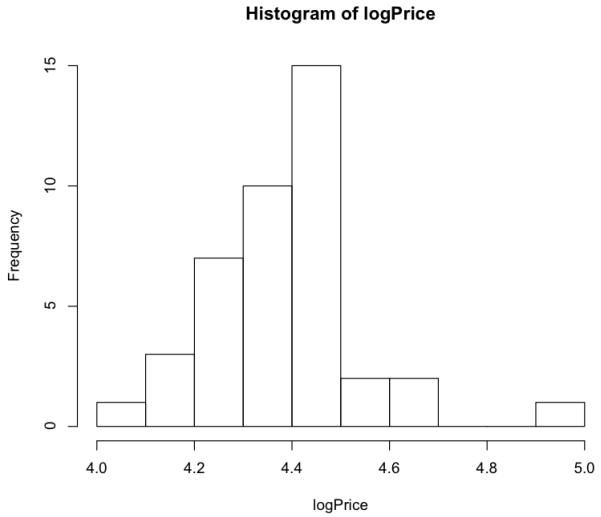
Below are a quick summary statistics.

```
     Price          Brand       Type         MPG       
 Min.   :11900   Honda :11   Car  :17   Min.   :15.50  
 1st Qu.:19170   Nissan:16   SUV  :19   1st Qu.:21.00  
 Median :23625   Toyota:14   Truck: 5   Median :26.50  
 Mean   :26056                          Mean   :25.73  
 3rd Qu.:29215                          3rd Qu.:30.50  
 Max.   :79605                          Max.   :37.50  
```



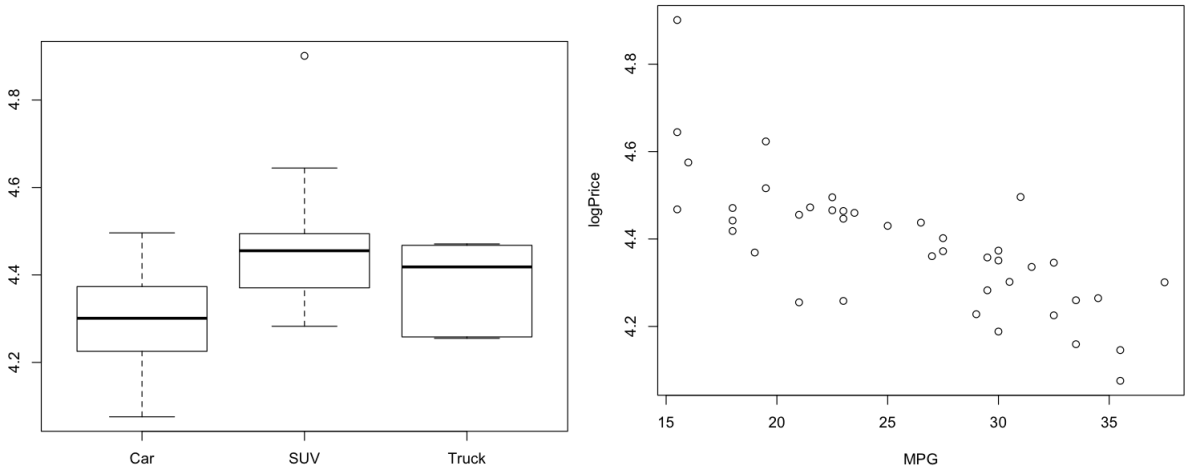
The prices of the cars are skewed right. We will therefore work in the log scale. The histogram of MPG looks okay.

Below is the histogram of logged price. The plot looks more normal.



Side-by-side box plots show that types of vehicles have a marginal effect on the price, with car having the lowest mean price, trucks slightly lower than SUVs.

There also appears to be a moderate to strong, negative, linear relationship between MPG and logged Price. This makes sense as higher MPG vehicles are mostly cars, while lower MPG vehicles are SUVs and trucks. Cars generally cost less compare to SUVs and trucks.



3 ANCOVA Model

The ANCOVA model relates types of the automobile and MPG to logged prices of the automobile. Here is the output.

```
> summary(lm(logPrice~Type+MPG))

Call:
lm(formula = logPrice ~ Type + MPG)

Residuals:
      Min        1Q    Median        3Q       Max
-0.182064 -0.044264  0.003937  0.039539  0.270231

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.009820   0.107147  46.757  < 2e-16 ***
TypeSUV     -0.026529   0.041594  -0.638  0.52752
TypeTruck   -0.201291   0.060718  -3.315  0.00206 **
MPG         -0.022747   0.003358  -6.774 5.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0883 on 37 degrees of freedom
Multiple R-squared:  0.6842,    Adjusted R-squared:  0.6586
F-statistic: 26.72 on 3 and 37 DF,  p-value: 2.272e-09

> library(car)
> Anova(d1,type=3)
Anova Table (Type III tests)

Response: logPrice
             Sum Sq Df   F value    Pr(>F)
(Intercept) 17.0458  1 2186.1879 < 2.2e-16 ***
Type         0.1181  2    7.5715  0.001752 **
MPG          0.3578  1   45.8896 5.678e-08 ***
Residuals    0.2885 37
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both types of the automobile and MPG are significant predictors for logged Price. They have F values 7.57 and 45.99, respectively; p-values are below any reasonable α. The output shows that there is a negative relationship between MPG and logged Price, with 10 additional MPG associated with a decrease in logged Price of $(10)(0.022747) = \$0.22747$, that is, 10 additional MPG is associated with multiplying the Price by $10^{-0.22747} = \$0.592284$, or a reduction in price of 69%, given the type of automobile is held fixed.

The following output gives the the least squares means.

```
> library(lsmeans)
> d1.rg<-ref.grid(d1)
> d1.lsm<-lsmeans(d1.rg,"Type")
> d1.lsm
 Type    lsmean         SE df lower.CL upper.CL
 Car   4.424496 0.02835215 37 4.367049 4.481943
 SUV   4.397967 0.02293957 37 4.351487 4.444447
 Truck 4.223206 0.04533553 37 4.131347 4.315064

Confidence level used: 0.95
```

The entries under Least Squares Means for logged price show that given MPG, trucks cost the least, followed by SUVs, and cars cost the most. The difference between trucks and SUVs is $4.397967 - 4.223206 = 0.174761$, which means that given that two automobile have the same MPG, the expected price of an SUV is a multiplicative factor of $10^{0.174761} = 1.49541$, that is, 49.54%, higher than that of a truck.

The difference between cars and SUVs is $4.424496 - 4.397967 = 0.026529$, which means that given that two automobile have the same MPG, the expected price of a car is a multiplicative factor of $10^{0.026529} = 1.062990$, that is, 6.3%, higher than that of an SUV.

The difference between cars and trucks is $4.424496 - 4.223206 = 0.20129$. In order words, given that two automobile have the same MPG, the expected price of a car is a multiplicative factor of $10^{0.20129} = 1.58961$, that is, 58.96% higher than than that of a truck.

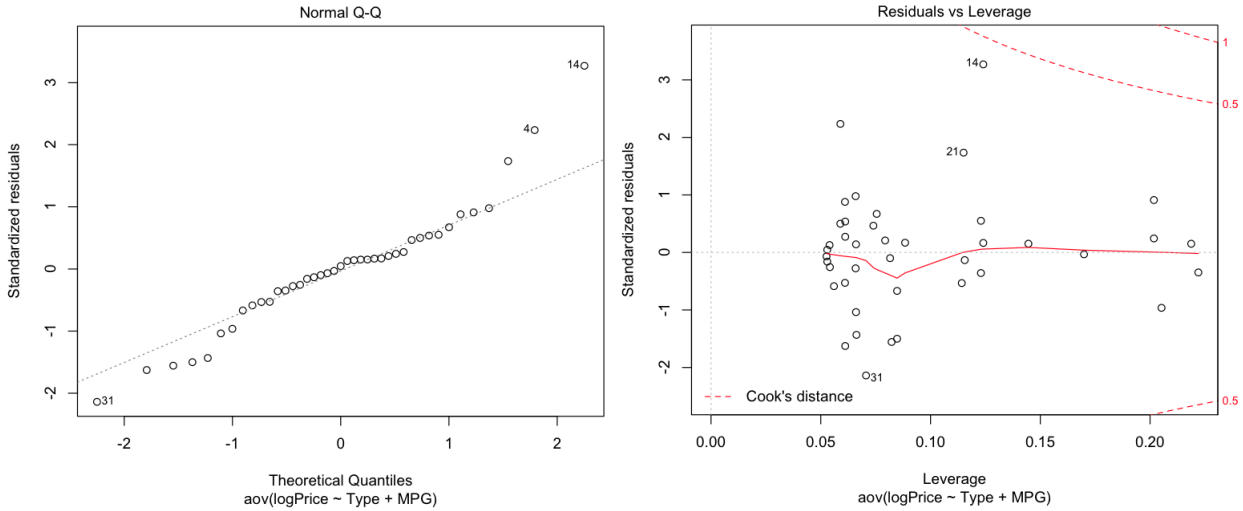
Since this model does not include an interaction effect, we can compare the least squares means to see which types of automobile is significantly different from each other. Below is Tukey comparison.

```
> summary(glht(d1,linfct=mcp(Type="Tukey")))

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = logPrice ~ Type + MPG)

Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
SUV - Car == 0   -0.02653    0.04159  -0.638  0.79253
Truck - Car == 0 -0.20129    0.06072  -3.315  0.00551 **
Truck - SUV == 0 -0.17476    0.04585  -3.812  0.00134 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

The output shows that the logged prices of trucks are significantly lower than cars and SUVs. SUVs and cars do not have a significant difference. One may think of it as trucks being in one group, and SUV and car are in another another group.

4 Regression Diagnostics

Regression diagnostics check assumptions.



QQ plot shows that there are a few outliers above the line to the far right. There are also some points below the lines to the far left.

We expect 95% of the points lie within ±2.5 standard residuals band. Observation 14 lies outside this range.

Cook's distance measures influence. One should expect Cook's distance to be less than 1. The plot shows that all points lie within this range (shown by the dotted curves marked by 1).

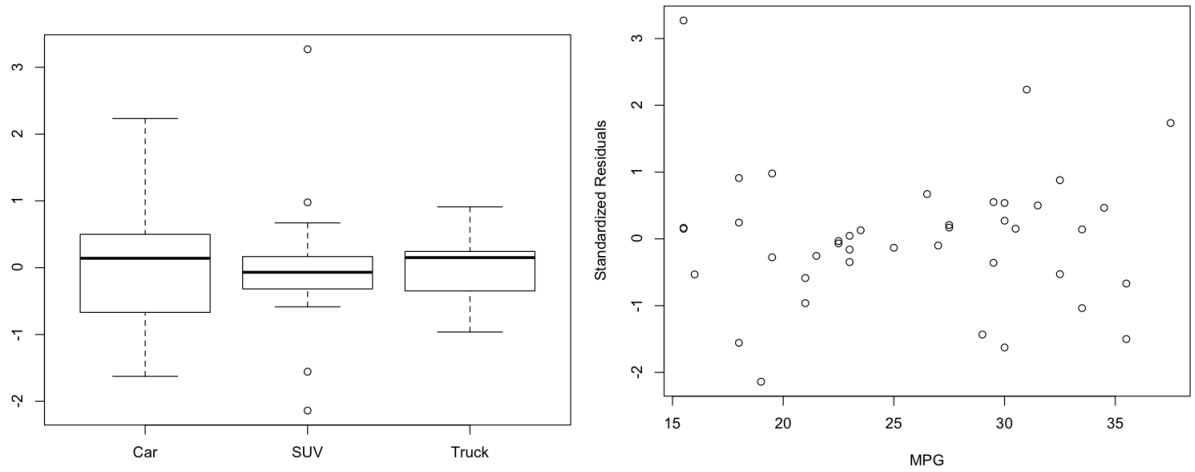
A good guideline for what constitutes as a large leverage value is $2.5\left(\frac{p+1}{n}\right) = 2.5\left(\frac{3+1}{41}\right) = 0.098$. Almost half of the data points have leverage greater than this value. A detailed output is provided below.

```
> d1.diag=ls.diag(lsfit(model.matrix(d1)[,2:4],logPrice))
> d1.diag
$std.dev
[1] 0.08830098

$hat
 [1] 0.06604923 0.06103027 0.06113660 0.05892486 0.11557955 0.22199580 0.20174983 0.08840553
 [9] 0.05295606 0.05263258 0.08225944 0.06587618 0.12402626 0.12402626 0.05890359 0.06113660
[17] 0.07396047 0.06604923 0.07546638 0.08157442 0.11504788 0.06113660 0.05400261 0.05415483
[25] 0.20174983 0.12296068 0.12296068 0.14457668 0.05600057 0.06587618 0.07061420 0.05295606
[33] 0.11422676 0.20522057 0.21874198 0.08476399 0.08476399 0.06624063 0.06103027 0.07931970
[41] 0.16991616

$std.res
 [1] -1.03735946 -0.52857918  0.27265127  2.23490222 -0.13325285 -0.34751287  0.24363735
 [8]  0.16899718 -0.15970589 -0.06833637 -1.55660732 -0.27632156  0.16552834  3.26982084
[15]  0.49935779  0.53725595  0.46531849  0.14112874  0.67067800 -0.09960325  1.73500574
[22] -1.62670346  0.12851559 -0.25511864  0.91066583 -0.35826802  0.55016583  0.15178510
[29] -0.58627369  0.97813453 -2.13875310  0.04581207 -0.53262193 -0.96299262  0.15129014
[36] -0.66856317 -1.50040839 -1.43291200  0.87921314  0.20825393 -0.03235511

$stud.res
 [1] -1.03845785 -0.52336707  0.26921214  2.37027875 -0.13147135 -0.34334537  0.24051541
 [8]  0.16676216 -0.15758724 -0.06741083 -1.58831570 -0.27284357  0.16333664  3.82497895
[15]  0.49423172  0.53202528  0.46033624  0.13924602  0.66561100 -0.09826122  1.78557521
[22] -1.66522155  0.12679530 -0.25186910  0.90851459 -0.35400797  0.54491367  0.14976653
[29] -0.58100175  0.97754746 -2.25354891  0.04519003 -0.52740077 -0.96202246  0.14927786
[36] -0.66348643 -1.52718267 -1.45434450  0.87645441  0.20554091 -0.03191534

$cooks
 [1] 1.902577e-02 4.539979e-03 1.210191e-03 7.818642e-02 5.801165e-04 8.614789e-03 3.750610e-03
 [8] 6.924313e-04 3.565557e-04 6.486046e-05 5.429551e-02 1.346150e-03 9.698561e-04 3.784517e-01
[15] 3.901857e-03 4.698956e-03 4.323254e-03 3.521395e-04 9.179076e-03 2.202912e-04 9.783645e-02
[22] 4.307802e-02 2.357091e-04 9.316259e-04 5.240012e-02 4.498869e-03 1.060900e-02 9.734540e-04
[29] 5.097550e-03 1.686791e-02 8.688750e-02 2.933900e-05 9.145835e-03 5.986324e-02 1.602134e-03
[36] 1.034911e-02 5.212394e-02 3.641401e-02 1.256094e-02 9.341113e-04 5.357208e-05

$dfits
 [1] -0.27615999 -0.13342993  0.06869800  0.59311178 -0.04752720 -0.18340561  0.12091495
 [8]  0.05193216 -0.03726437 -0.01588904 -0.47552103 -0.07245623  0.06146033  1.43926357
[15]  0.12364725  0.13576309  0.13009489  0.03703008  0.19016732 -0.02928446  0.64380977
[22] -0.42493399  0.03029462 -0.06026754  0.45673993 -0.13255211  0.20403342  0.06157060
[29] -0.14151021  0.25959711 -0.62117646  0.01068601 -0.18939277 -0.48884632  0.07898856
[36] -0.20191621 -0.46476149 -0.38735756  0.22344785  0.06033015 -0.01443962

$correlation
           Intercept    TypeSUV  TypeTruck        MPG
Intercept  1.000000 -0.7941450 -0.7296800 -0.9798210
TypeSUV   -0.794145  1.0000000  0.6562166  0.7054659
TypeTruck -0.729680  0.6562166  1.0000000  0.6727556
MPG       -0.979821  0.7054659  0.6727556  1.0000000

$std.err
                [,1]
Intercept 0.107146602
TypeSUV   0.041593571
TypeTruck 0.060717666
MPG       0.003357924

$cov.scaled
              Intercept       TypeSUV    TypeTruck           MPG
Intercept  0.0114803942 -3.539195e-03 -0.004747073 -3.525299e-04
TypeSUV   -0.0035391945  1.730025e-03  0.001657252  9.853103e-05
TypeTruck -0.0047470733  1.657252e-03  0.003686635  1.371650e-04
MPG       -0.0003525299  9.853103e-05  0.000137165  1.127565e-05

$cov.unscaled
            Intercept     TypeSUV   TypeTruck          MPG
Intercept  1.47239976 -0.45391378 -0.60882835 -0.045213162
TypeSUV   -0.45391378  0.22188163  0.21254820  0.012636941
TypeTruck -0.60882835  0.21254820  0.47282352  0.017591876
MPG       -0.04521316  0.01263694  0.01759188  0.001446141
```

Observation 14 has a standard residual of 3.82. It is Toyota Land Cruiser, an SUV with 15.5 MPG costing $79,605. The points that have high leverage values are observation 5(0.11557955), 6(0.22199580), 7(0.20174983), 13(0.12402626), 14(0.12402626), 21(0.11504788), 25(0.20174983), 26(0.12296068), 27(0.12296068), 28(0.14457668), 33(0.11422676), 34(0.20522057), 35(0.21874198), and 41(0.16991616). These points are Toyota Sienna, Toyota Tacoma, Toyota Tundra, Toyota Sequoia, Toyota Land Cruiser, Honda CRZ, Honda Ridgeline, Nissan Juke, Nissan Rogue, Nissan Rogue Select, Nissan Armada, Nissan Frontier, Nissan Titan, and Nissan Maxima, respectively.

Most vehicles have border line leverage value, around 0.10. The ones have very high leverage values, say, around 0.20, are all trucks. This is expected as we only have five trucks in the data. There is nothing to do about them. Taking out these points is not really something we want to do, since that is taking out an entire level in the variable type.

Below are a side-by-side box plot of Type versus standard residuals and a plot of MPG versus standard residuals.



Toyota Land Cruiser, the SUV with standard residual of 3.82 is apparent in the side-by-side box plots and far left in the MPG versus standard residuals plot. We can see that cars have more variability than SUVs and trucks, but its residuals lie within ± 2.5 range. We will later check with Levene's test to see if there is non-constant variance. The plot of MPG versus standard residuals looks okay.

Below is the output of Levene's test.

```
> absres=abs(d1.diag$std.res)
> Anova(aov(absres~Type),type=3)
Anova Table (Type III tests)

Response: absres
             Sum Sq Df F value    Pr(>F)
(Intercept) 11.4197  1 21.4161 4.216e-05 ***
Type         0.4677  2  0.4386    0.6482
Residuals   20.2627 38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistics is 0.4386, giving a p-value of 0.6482. The test is not significant, so non-constant variance is addressed after we took log.

5 Prediction Interval

Below is Our regression model allows us to write down a model for prediction. The predicted price for a certain type of vehicle i with x MPG equals

$$\text{Least Squares Mean}_i + \hat{\beta}_1(x - \bar{x})$$

$\bar{x}$ is the means of MPG, which in this case is 25.73171. Let us predict the price of a truck with 20 MPG. It has logged price

$$4.223206 + (-0.022747)(20 - 25.73171) = 4.35358520737$$

or $10^4.35358520737 = \$22573$ after anti-logging. A rough prediction 95% interval is calculated as

$$4.35358520737 \pm 2 \times 0.0883 = (4.157499, 4.549673)$$

```
> nd=data.frame(type="Truck",MPG=20)
> predict(d1,nd,interval="predict")
       fit      lwr      upr
1 4.353586 4.157499 4.549673
```

or (14371,35455) in the original scale.

6 Adding an Interaction

The above model does not include an interaction effect. Here, we investigate whether different slopes for MPG for each type of automobile would improve the model. Below are the output of the regression.

```
> d2<-aov(logprice~type+MPG+type*MPG)
> coef(d2)
  (Intercept)       typeSUV     typeTruck           MPG   typeSUV:MPG typeTruck:MPG
   4.94861357    0.03869861    0.07899249   -0.02078950   -0.00213621   -0.01342767
> summary(lm(logprice~type+MPG+type*MPG))

Call:
lm(formula = logprice ~ type + MPG + type * MPG)

Residuals:
      Min        1Q    Median        3Q       Max
-0.182694 -0.045429  0.006604  0.033206  0.268977

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.948614   0.183341  26.991  < 2e-16 ***
typeSUV        0.038699   0.209381   0.185  0.85443
typeTruck      0.078992   0.348570   0.227  0.82204
MPG           -0.020789   0.005823  -3.571  0.00106 **
typeSUV:MPG   -0.002136   0.007295  -0.293  0.77138
typeTruck:MPG -0.013428   0.016443  -0.817  0.41968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08993 on 35 degrees of freedom
Multiple R-squared:  0.6901,    Adjusted R-squared:  0.6459
F-statistic: 15.59 on 5 and 35 DF,  p-value: 4.52e-08

> Anova(d2,type=3)
Anova Table (Type III tests)

Response: logprice
             Sum Sq Df  F value    Pr(>F)
(Intercept) 5.8920  1 728.5281 < 2.2e-16 ***
type        0.0005  2   0.0298  0.970678
MPG         0.1031  1  12.7487  0.001059 **
type:MPG    0.0054  2   0.3355  0.717244
Residuals   0.2831 35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

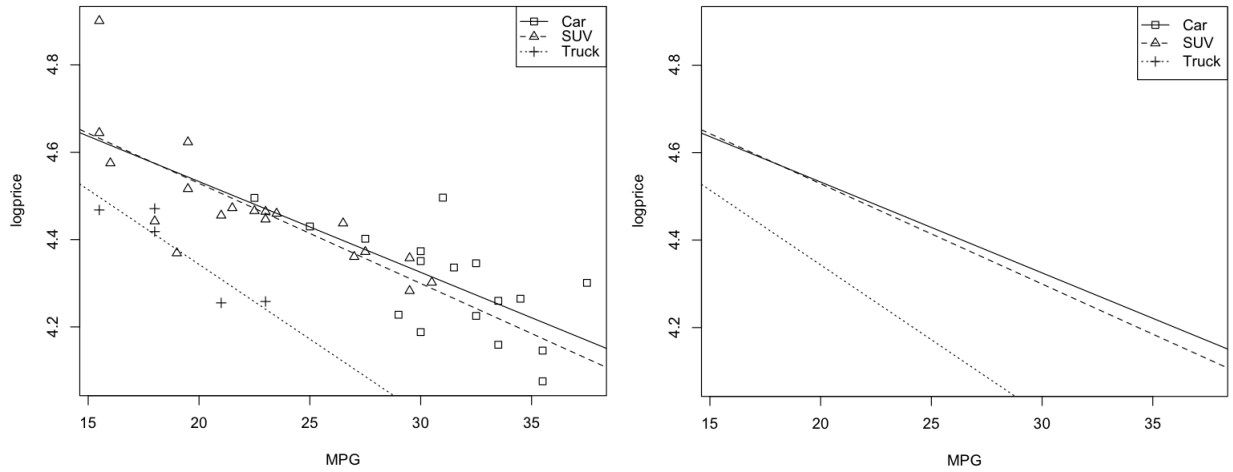
Notice that type : MPG has an F value of 0.3355, which gives p-value of 0.717244, not statistically significant at all. Adding the interaction of type of automobile and MPG does not help.

The t-test for typeSUV:MPG shows that the coefficient for typeSUV is not significantly different from the overall coefficient for the variable Type, as it has a p-value of 0.77138. The t-test for typeTruck:MPG is not significant either.

The missing coefficient for typeCar:MPG is calculated as $0 - (-0.002136 - 0.013428) = 0.01556388$ since their coefficient must sum to zero.

This model that includes the interaction effect between type of vehicle and MPG corresponds to separate regression line for each type of automobile. We plot separate lines on the same plot below

```
> plot(logprice~MPG,data=d,type='n')
> points(cars$MPG,cars$logprice,pch=0)
> points(suvs$MPG,suvs$logprice,pch=2)
> points(trucks$MPG,trucks$logprice,pch=3)
> abline(reg1,lty=1)
> abline(reg2,lty=2)
> abline(reg3,lty=3)
> legend("topright",c("Car","SUV","Truck"),lty=c(1,2,3),pch=c(0,2,3))
```



Notice that the slopes are fairly parallel, with the exception of trucks having a steeper slopes.

7 Conclusion

In this analysis, we conducted an ANCOVA model of price of vehicles built by three major Japanese automakers based on one categorical, one numerical variable–type of vehicles and MPG, respectively. By examining price versus type box plot, cars have the least price, followed by trucks and SUVs. Interestingly, the least squares means show that given MPG, trucks cost the least, followed by SUVs, then cars.

Finally, interaction effect between type and MPG is not significant.