PS 1

Redmond Xia

June 19, 2020

Libraries

```
library(DataAnalytics)
library(data.table)
library(ggplot2)
library(lfe)
```

```
## Loading required package: Matrix
```

```
library(foreign)
```

Problem Set 1

Question 1 : On ggplot2 and regression planes

Use the imports-85.csv dataset available at CCLE (Week 1). The data is taken from the UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php.

Reading in the data

```
imports <- as.data.table(read.csv("imports-85.csv"))
imports$price <- as.numeric(as.character(imports$price))
```

```
## Warning: NAs introduced by coercion
```

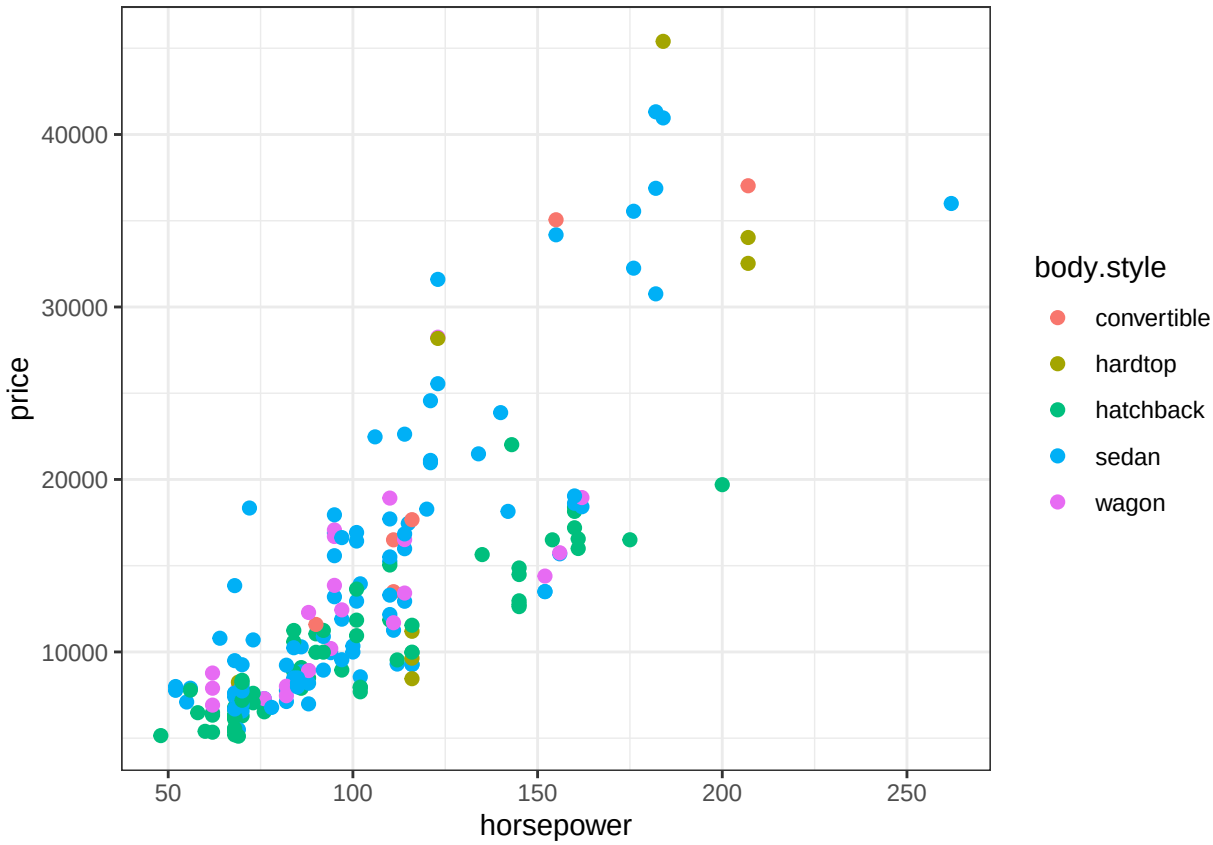
```
imports$horsepower <- as.numeric(as.character(imports$horsepower))
```

```
## Warning: NAs introduced by coercion
```

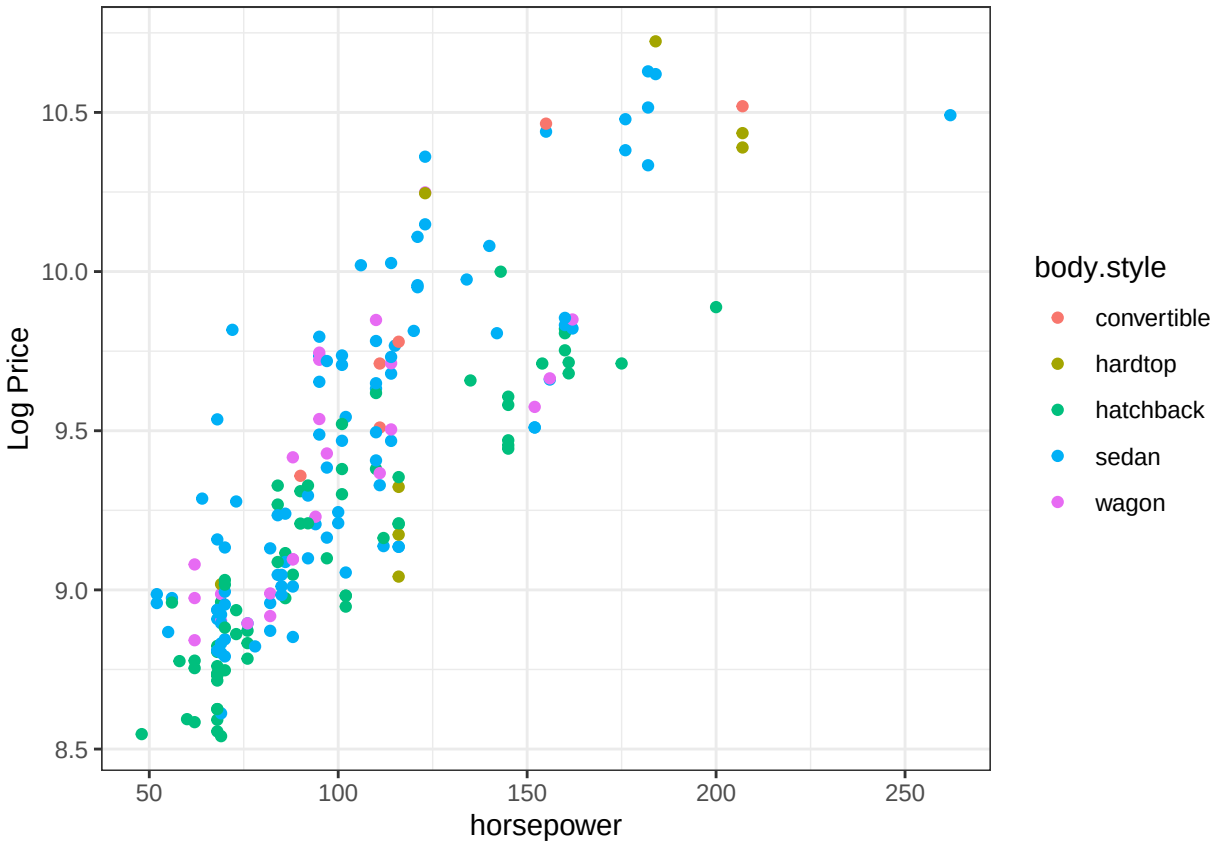
```
imports$city.mpg <- as.numeric(as.character(imports$city.mpg))
imports <- imports[!is.na(horsepower) & ! is.na(price) & ! is.na(city.mpg)]
```

1. Use ggplot2 to visualize the relationship between price and horsepower and body style. Price is the dependent variable. Consider both the "log()" and "^2" transformations of price as dependent variables. Does the body style variable appear to be relevant for car prices , above and beyond horsepower?

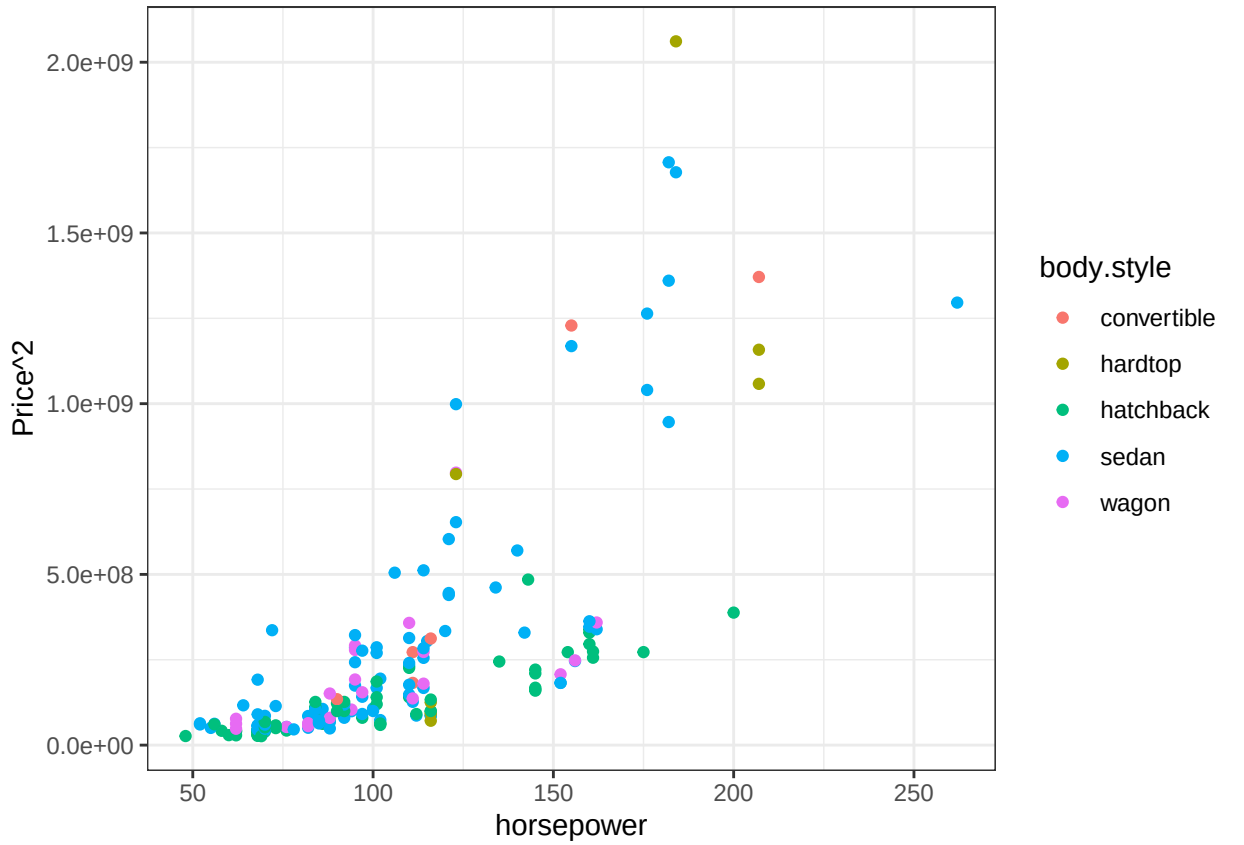
```
ggplot(imports,aes(x = horsepower, y = price, color = body.style)) + geom_point(size = 2) + theme_bw()
```




```
ggplot(imports, aes(x = horsepower, y = log(price), color = body.style)) + geom_point() + ylab("Log Pri
```




```
ggplot(imports,aes(x = horsepower, y = price^2, color = body.style)) + geom_point() + ylab("Price^2") +
```



It seems that body style have little relevance for car prices using horsepower as an independent variable. We would not want to include body type when trying to predict the price of the car.

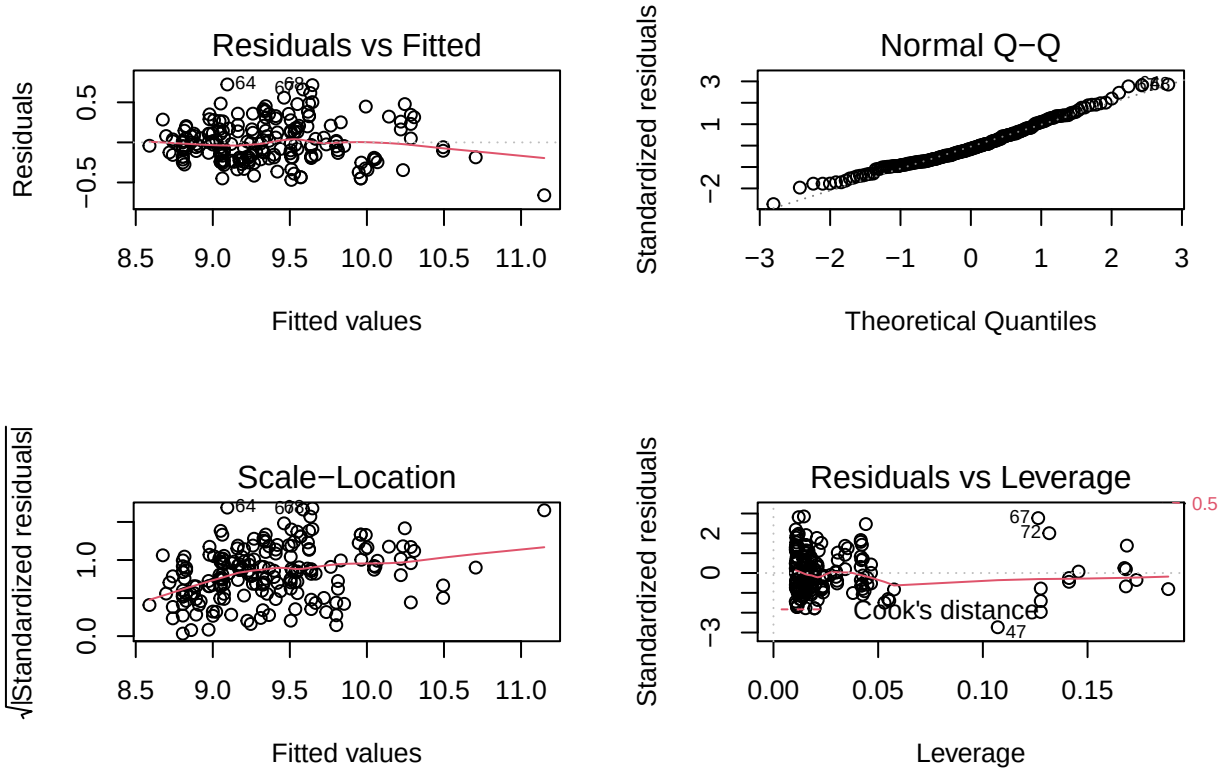
2. Run a regression of your preferred specification. Perform residual diagnostics as you learned in Econometrics. What do you conclude from your regression diagnostic plots of residuals vs. fitted and residuals vs. horsepower?

```
reg <- lm(log(price) ~ horsepower + body.style, data = imports)
summary(reg)
```

```
##
## Call:
## lm(formula = log(price) ~ horsepower + body.style, data = imports)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6594 -0.1863 -0.0421  0.1620  0.7227
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.4655114  0.1234053  68.599  < 2e-16 ***
## horsepower           0.0108230  0.0005014  21.584  < 2e-16 ***
## body.stylehardtop   -0.2111930  0.1380238  -1.530 0.127624
## body.stylehatchback -0.3959712  0.1101542  -3.595 0.000412 ***
## body.stylesedan     -0.1504265  0.1084385  -1.387 0.166979
## body.stylewagon     -0.1639401  0.1177993  -1.392 0.165618
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2554 on 193 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:  0.7437
## F-statistic: 115.9 on 5 and 193 DF,  p-value: < 2.2e-16
```

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

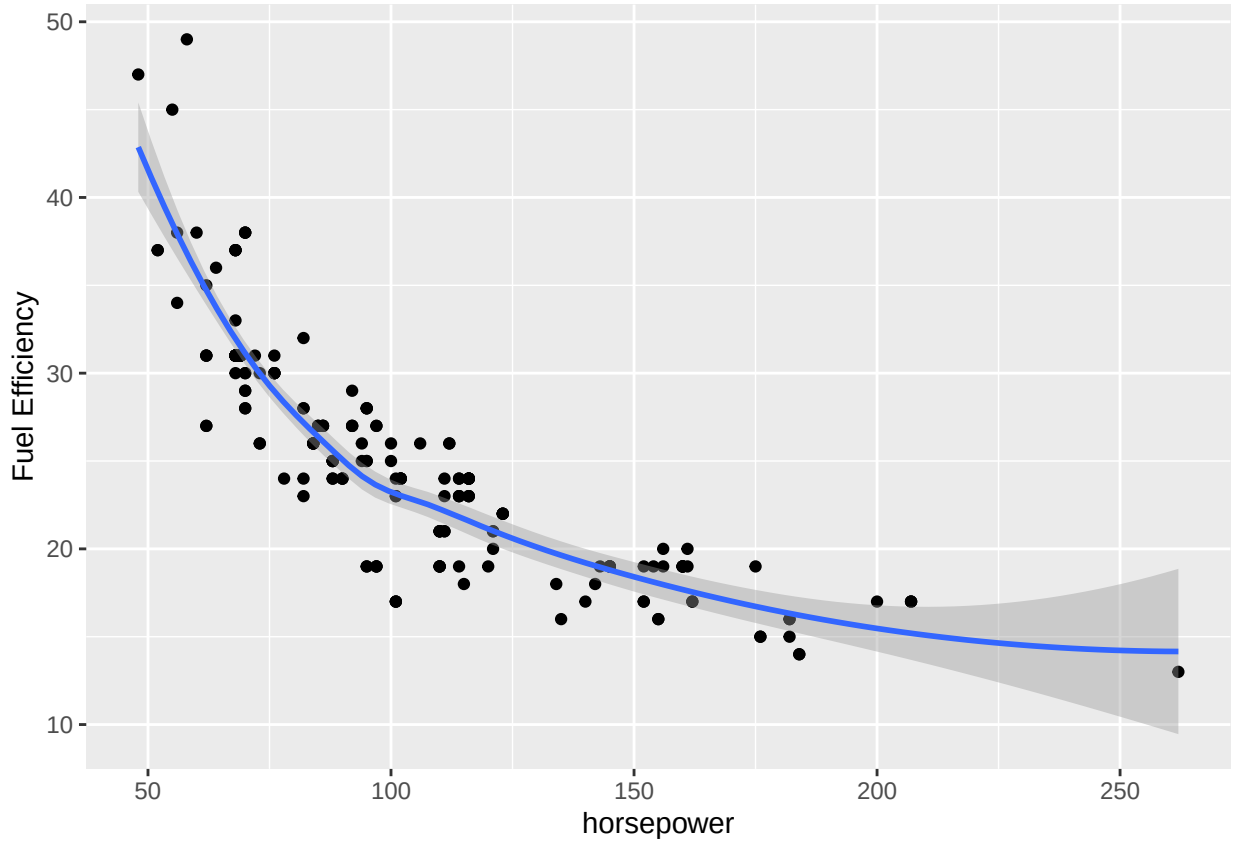


Based on the residual we see that the linear regression does seem to fit well for the log price based on horse power. However, only the hatchback body type is significant variable. The residual also do seem fairly normal

3. Now use ggplot2 to visualize the relationship between fuel efficiency (city-mpg) and horsepower. Now regress city.mpg on horsepower. Is the regression result consistent with the conclusion you would draw based on the plot? More on this next week.

```
ggplot(imports, aes(x = horsepower, y = city.mpg)) + geom_point() + ylab("Fuel Efficiency") + geom_smoo
```

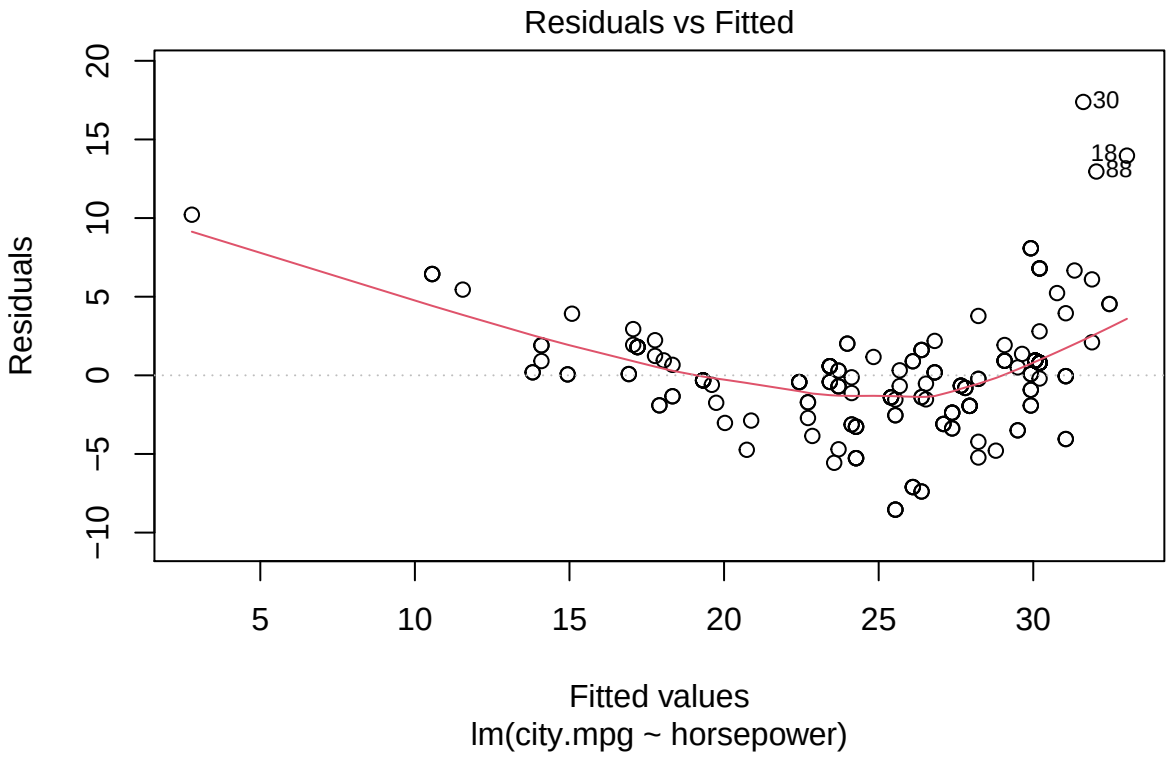
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

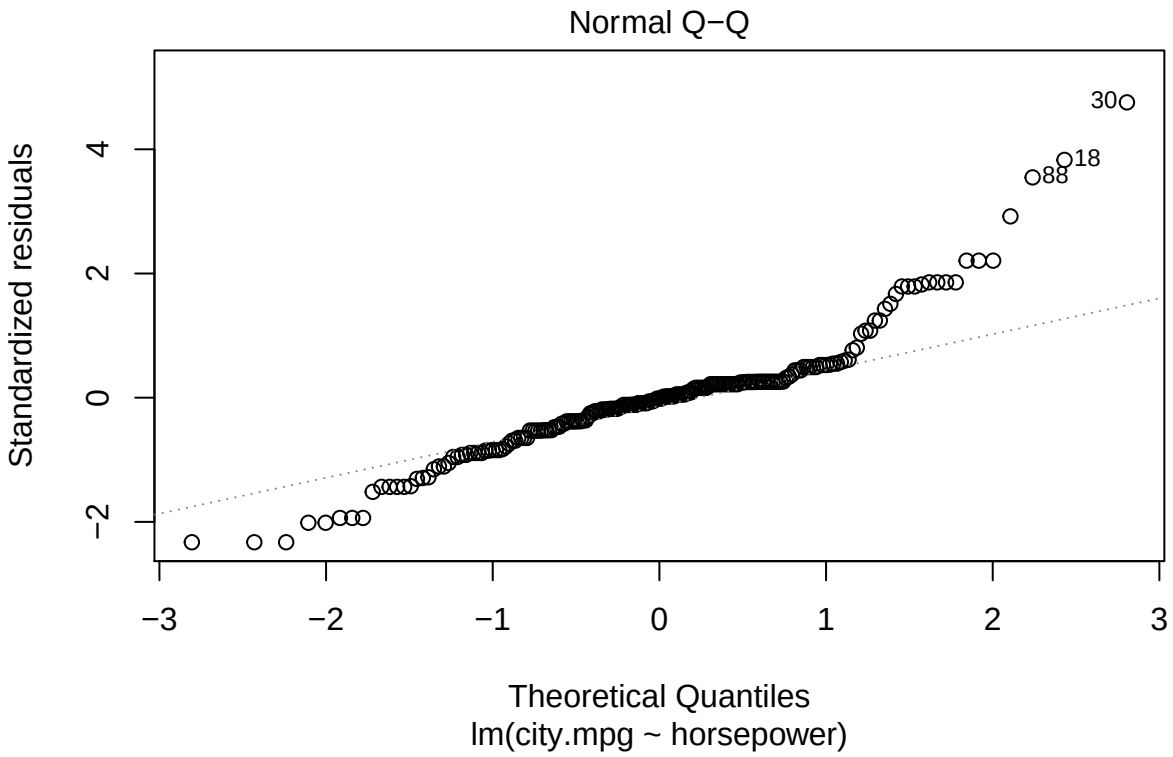


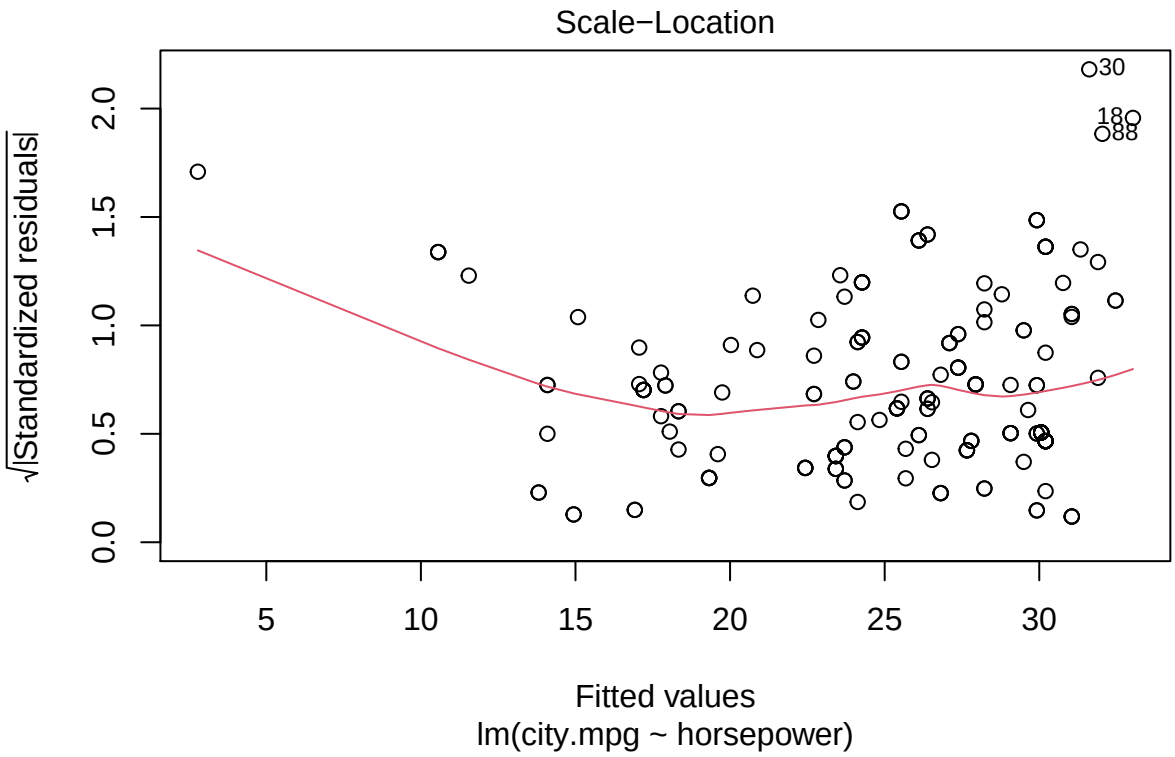

```
reg2 <- lm(city.mpg ~ horsepower, data = imports)
summary(reg2)
```

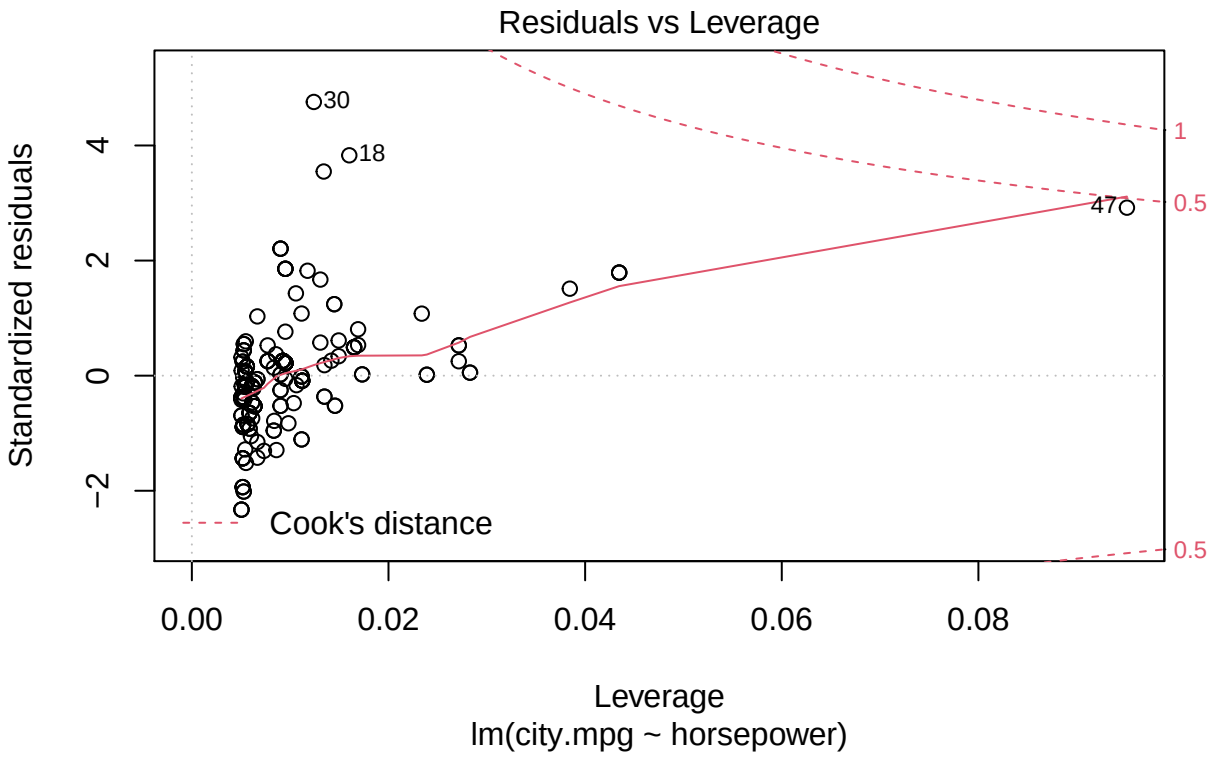
```
##
## Call:
## lm(formula = city.mpg ~ horsepower, data = imports)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5398 -1.9145 -0.0515  0.9378 17.3832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.81381    0.76540   52.02   <2e-16 ***
## horsepower  -0.14133    0.00696  -20.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 197 degrees of freedom
## Multiple R-squared:  0.6767, Adjusted R-squared:  0.6751
## F-statistic: 412.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
plot(reg2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(city.mpg ~ horsepower)
30
18
88
20
15
10
5
0
−5
−10
5
10
15
20
25
30


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(city.mpg ~ horsepower)
30
18
88
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
30
30
18
88
Fitted values
lm(city.mpg ~ horsepower)



We can see a strong negative relation between low horsepower cars and fuel efficiency. The relationship does not look non-linear and a lot more parabolic. The residual is not normal and does not look i.i.d.

Question 2 : Nonlinear relations

A common concern is that the relationship between a predictive variable (X) and the outcome we are trying to predict (Y) is nonlinear. On the surface, this seems to invalidate linear regressions, such as the Fama-MacBeth regression. However, this is not generally the case. For instance, if Y = f(X) + noise, where f(.) is not linear in X, simply define a transformation of X as, generally, Z = a + bf(X). Now, it is clear that Y = a1 + b1*Z, for constants a, a1, b, and b1. In other words, one could include squared values of X in the regression, perhaps max(0,X), etc.

We will see this in action for the case of Issuance (lnIssue). This is the average amount of stock issuance in the last 36 months, normalized by market equity. Generally, firms that issue a lot of equity have low returns going forward.

a. Construct decile sorts (10 portfolios) as in the class notes, but now based on the issuance variable lnIssue. Give the average return to each decile portfolio, valueweighting stocks within each portfolio each year, equal-weighting across years.

```
remove(list = ls())

StockRetAcct_DT <- as.data.table(read.dta("StockRetAcct_insample.dta"))
setkey(StockRetAcct_DT, FirmID, year)

StockRetAcct_DT[,ExRet := exp(lnAnnRet) - exp(lnRf)]
# The data has been winsored so we add a little amount of noise
```

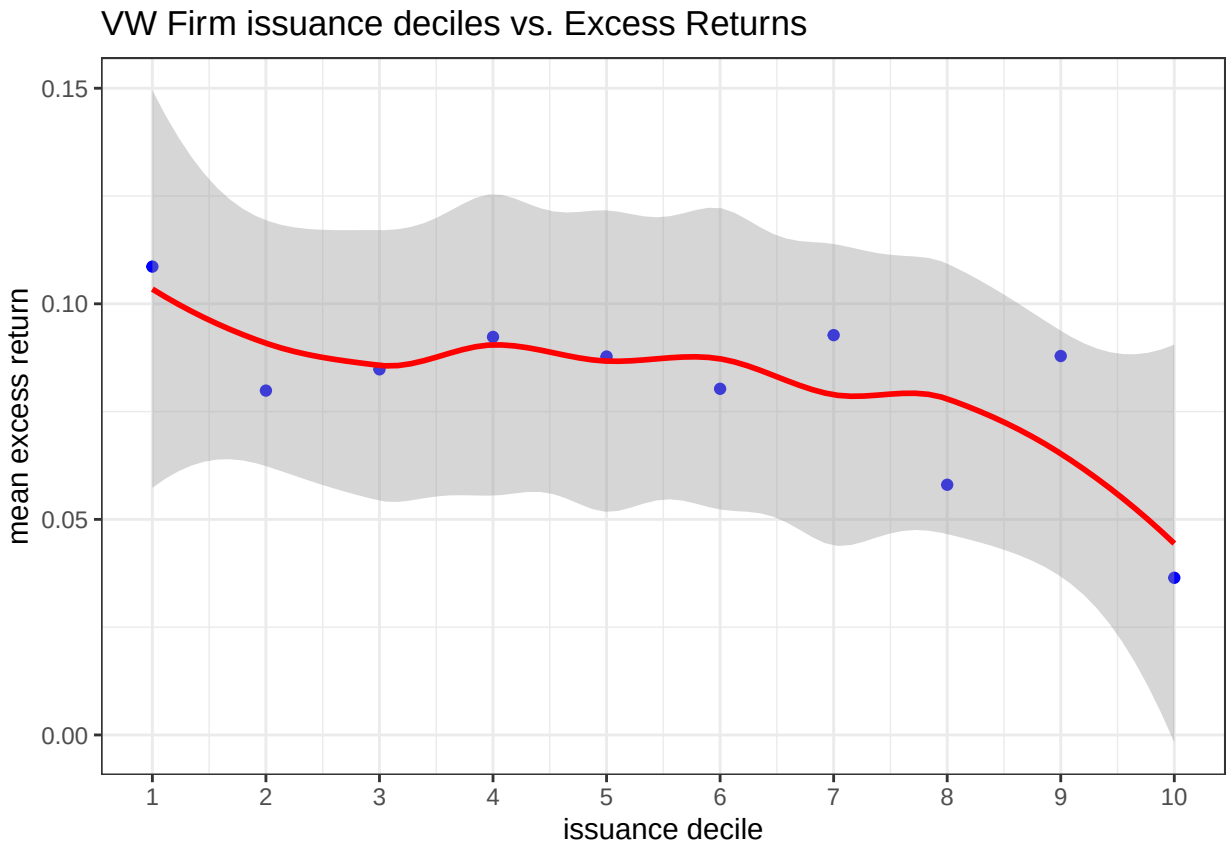
```
StockRetAcct_DT[,lnIssue := jitter(lnIssue, amount = 0)]
StockRetAcct_DT[,lnBM := jitter(lnBM, amount = 0)]
StockRetAcct_DT[ ,lnME := jitter(lnME, amount = 0)]

# Create the deciel shorts based on ln issues
StockRetAcct_DT[,issue_decile_yr:=
                  cut(lnIssue,breaks=quantile(lnIssue,probs=c(0:10)/10,na.rm=TRUE,include.lowest=TRUE),
                      labels=FALSE), by = year]

# get the average return for each portfolio (VW across stocks, EW across years)
EW_ISSUE_MutualFunds_yr <- StockRetAcct_DT[,.(MeanExRetYr = weighted.mean(ExRet, MEwt)),
                                           by = .(issue_decile_yr, year)]
# then average across years
EW_ISSUE_MutualFunds_yr <- EW_ISSUE_MutualFunds_yr[,.(MeanExRet = mean(MeanExRetYr)),
                                                   by = issue_decile_yr]
setkey(EW_ISSUE_MutualFunds_yr,issue_decile_yr)
EW_ISSUE_MutualFunds_yr[!is.na(issue_decile_yr)]
```

```
##     issue_decile_yr  MeanExRet
##  1:               1 0.10862872
##  2:               2 0.07987265
##  3:               3 0.08479732
##  4:               4 0.09232022
##  5:               5 0.08777017
##  6:               6 0.08028680
##  7:               7 0.09273709
##  8:               8 0.05802911
##  9:               9 0.08789102
## 10:              10 0.03645501
```

b. Plot the average return to these 10 portfolios, similar to what we did in the Topic 1(e-f) notes. Discuss whether the pattern seems linear or not.

```
ggplot(EW_ISSUE_MutualFunds_yr[!is.na(issue_decile_yr)],aes(x=issue_decile_yr,y=MeanExRet)) +
geom_point(col="blue") +
geom_smooth(col="red") +
theme_bw() + scale_x_continuous(breaks = 1:10) +
xlab("issuance decile") +
ylab("mean excess return") +
ggtitle("VW Firm issuance deciles vs. Excess Returns")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Based off the graph, we conclude that the firms reissue (net issuance) more performs worse.

c. Since most of the 'action' is in the extreme portfolios, consider a model where expected returns to stocks is linear in a transformed issuance-characteristic that takes three values: -1 if the stock's issuance is in Decile 1, 1 if the stock's issuance is in decile 10, and 0 otherwise.

```
StockRetAcct_DT[,trans_issue_decile_yr:= ifelse(issue_decile_yr == 1, -1, ifelse(issue_decile_yr == 10,

# Fama-MacBeth Regressions
port_ret = StockRetAcct_DT[, .(lambda = felm(ExRet ~ trans_issue_decile_yr,
                                             na.action = na.omit)$coef[2]), by = year]
fm_output = list(MeanReturn = mean(port_ret$lambda), StdReturn = sqrt(var(port_ret$lambda)),
                 SR_Return = mean(port_ret$lambda)/sqrt(var(port_ret$lambda)),
                 tstat_MeanRet = sqrt(1+2014-1980)*mean(port_ret$lambda)/sqrt(var(port_ret$lambda)))
fm_output
```

```
## $MeanReturn
## [1] -0.03443594
##
## $StdReturn
## [1] 0.06126816
##
## $SR_Return
## [1] -0.5620528
##
## $tstat_MeanRet
## [1] -3.325149
```

This portfolio goes long on stocks in decile 10 and short on stocks in decile 1 and takes no positions on the stocks in the other 8 deciles. Based on the results of the Fama-Macbeth regression, we would want to hold the opposite; go long on stocks in decile 1, and short on stocks in decile 1.

Question 3

In the lecture notes we saw that the value spread is much larger for small stocks. Using this fact, I proposed a model where expected returns are linear in the book-to-market ratio as well as the interaction between book-to-market and size. In other words, holding size constant there is a linear relation between expected stock returns and book-to-market.

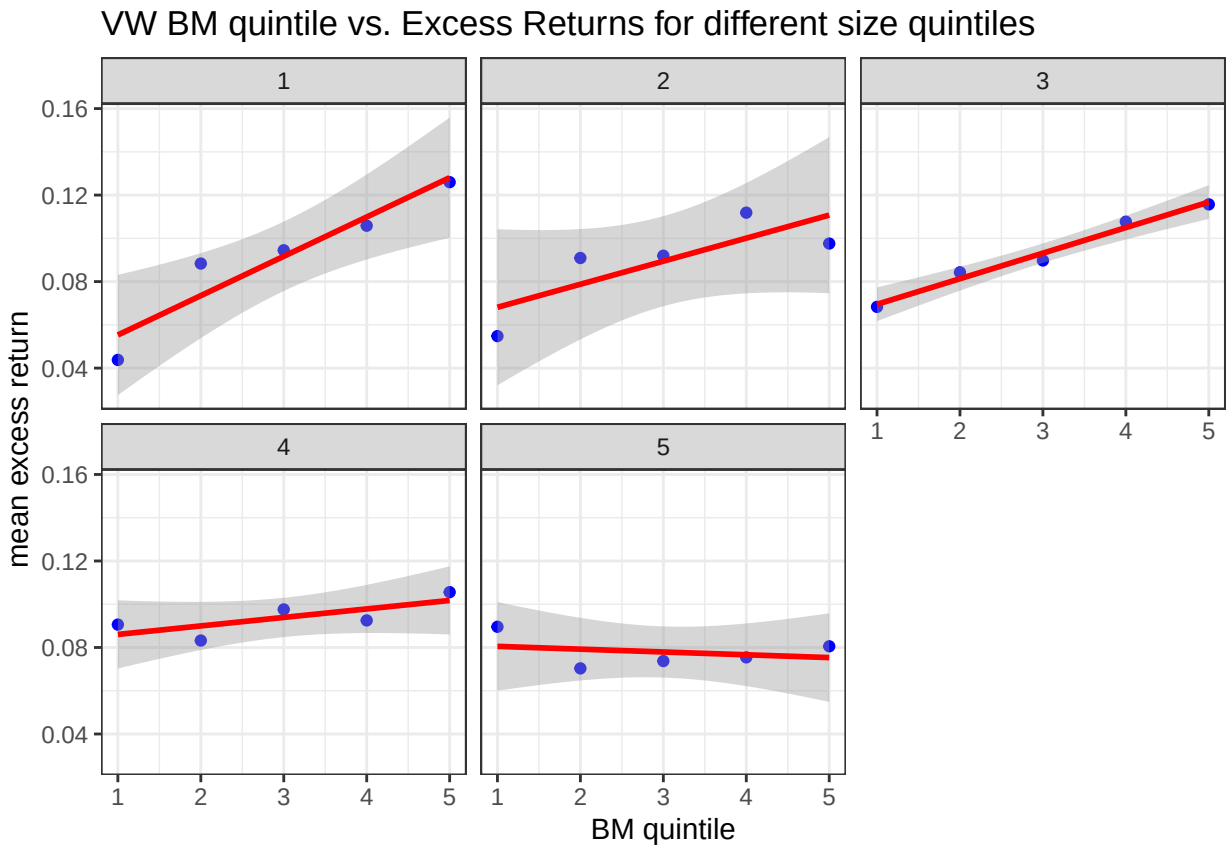
In this question, we will dig deeper into whether this is a reasonable assumption or not based on visual analysis.

a. Create independent quintile sorts based on book-to-market (lnBM) and size (lnME). That is create a quintile variable by year for book-to-market and then create a quintile variable by year for size.

```
StockRetAcct_DT[,`:=`(bm_quintile_yr= cut(lnBM, breaks =
                                            quantile(lnBM,probs=c(0:5)/5, na.rm=TRUE,
                                                     include.lowest=TRUE),labels=FALSE)
                      , size_quintile_yr= cut(lnME,breaks= quantile(lnME,probs=c(0:5)/5,na.rm=TRUE,
                                                                    include.lowest=TRUE),labels=FALSE))
```

```
# get the average return for each portfolio (VW across stocks, EW across years)
EW_SIZE_BM_MutualFunds_yr <- StockRetAcct_DT[!is.na(bm_quintile_yr) & !is.na(size_quintile_yr),
                                             .(MeanExRetYr = weighted.mean(ExRet, MEwt)),
                                             by = .(bm_quintile_yr,size_quintile_yr, year)]

# then average across years
EW_SIZE_BM_MutualFunds_yr <- EW_SIZE_BM_MutualFunds_yr[,.(MeanExRet = mean(MeanExRetYr)),
                                                       by = .(bm_quintile_yr,size_quintile_yr)]

ggplot(EW_SIZE_BM_MutualFunds_yr, aes(x=bm_quintile_yr,y=MeanExRet)) +
geom_point(col="blue") +
geom_smooth(col="red",method="lm") +
theme_bw() +
xlab("BM quintile") +
ylab("mean excess return") +
facet_wrap(~ size_quintile_yr)+
ggtitle("VW BM quintile vs. Excess Returns for different size quintiles")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the plots above, we see that the conditional linearity assumption (that expected returns are linear in the BM ratio as well as the interaction between BM and size) seems to be a pretty good assumption.