ML Hmwk 1

Mina Yuan, Yuting Shan, Yi Hu, Redmond Xia, Yushi Wei

03/04/2020

Question 1: On ggplot2 and regression planes

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```r
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.3
```

```r
library(DataAnalytics)

data = as.data.table(read.csv("imports-85.csv",stringsAsFactors = F))

head(data)
```

```
##    ï..symboling normalized.losses        make fuel.type aspiration
## 1:            3               N/A alfa-romero       gas        std
## 2:            3               N/A alfa-romero       gas        std
## 3:            1               N/A alfa-romero       gas        std
## 4:            2               164        audi       gas        std
## 5:            2               164        audi       gas        std
## 6:            2               N/A        audi       gas        std
##    num.of.doors  body.style drive.wheels engine.location wheel.base length
## 1:          two convertible          rwd           front       88.6  168.8
## 2:          two convertible          rwd           front       88.6  168.8
## 3:          two   hatchback          rwd           front       94.5  171.2
## 4:         four       sedan          fwd           front       99.8  176.6
## 5:         four       sedan          4wd           front       99.4  176.6
## 6:          two       sedan          fwd           front       99.8  177.3
##    width height curb.weight engine.type num.of.cylinders engine.size
## 1:  64.1   48.8        2548        dohc             four         130
## 2:  64.1   48.8        2548        dohc             four         130
## 3:  65.5   52.4        2823        ohcv              six         152
## 4:  66.2   54.3        2337         ohc             four         109
## 5:  66.4   54.3        2824         ohc             five         136
## 6:  66.3   53.1        2507         ohc             five         136
##    fuel.system bore stroke compression.ratio horsepower peak.rpm city.mpg
## 1:        mpfi 3.47   2.68               9.0     111.00  5000.00       21
## 2:        mpfi 3.47   2.68               9.0     111.00  5000.00       21
## 3:        mpfi 2.68   3.47               9.0     154.00  5000.00       19
## 4:        mpfi 3.19    3.4              10.0     102.00  5500.00       24
## 5:        mpfi 3.19    3.4               8.0     115.00  5500.00       18
## 6:        mpfi 3.19    3.4               8.5     110.00  5500.00       19
##    highway.mpg    price
## 1:          27 13495.00
## 2:          27 16500.00
## 3:          26 16500.00
```

```
## 4:          30 13950.00
## 5:          22 17450.00
## 6:          25 15250.00
```

```r
str(data)
```

```
## Classes 'data.table' and 'data.frame':   205 obs. of  26 variables:
##  $ ï..symboling     : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ normalized.losses: chr  "N/A" "N/A" "N/A" "164" ...
##  $ make             : chr  "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
##  $ fuel.type        : chr  "gas" "gas" "gas" "gas" ...
##  $ aspiration       : chr  "std" "std" "std" "std" ...
##  $ num.of.doors     : chr  "two" "two" "two" "four" ...
##  $ body.style       : chr  "convertible" "convertible" "hatchback" "sedan" ...
##  $ drive.wheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
##  $ engine.location  : chr  "front" "front" "front" "front" ...
##  $ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ length           : num  169 169 171 177 177 ...
##  $ width            : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
##  $ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
##  $ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
##  $ engine.type      : chr  "dohc" "dohc" "ohcv" "ohc" ...
##  $ num.of.cylinders : chr  "four" "four" "six" "four" ...
##  $ engine.size      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuel.system      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ bore             : chr  "3.47" "3.47" "2.68" "3.19" ...
##  $ stroke           : chr  "2.68" "2.68" "3.47" "3.4" ...
##  $ compression.ratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower       : chr  "111.00" "111.00" "154.00" "102.00" ...
##  $ peak.rpm         : chr  "5000.00" "5000.00" "5000.00" "5500.00" ...
##  $ city.mpg         : num  21 21 19 24 18 19 19 19 17 16 ...
##  $ highway.mpg      : num  27 27 26 30 22 25 25 25 20 22 ...
##  $ price            : chr  "13495.00" "16500.00" "16500.00" "13950.00" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
data[,price:=as.numeric(price)]
```

```
## Warning in eval(jsub, SDenv, parent.frame()): NAs introduced by coercion
```

```r
data[,horsepower:=as.numeric(horsepower)]
```

```
## Warning in eval(jsub, SDenv, parent.frame()): NAs introduced by coercion
```

```r
#str(data)
```

Questions 1 Part 1:

Observation from pair-wise. horsepower and price has strong positive relationship.

prices of hatchback and sedan is the lowest. prices of convertible and hardtop are higher in general, while sedan could be as expensive as these 2 types.

horsepower of convertible and hardtop are higher in gneral, while hatchback and sedan could be as high as these 2 types sometimes. Wagon is the lowest among the types.

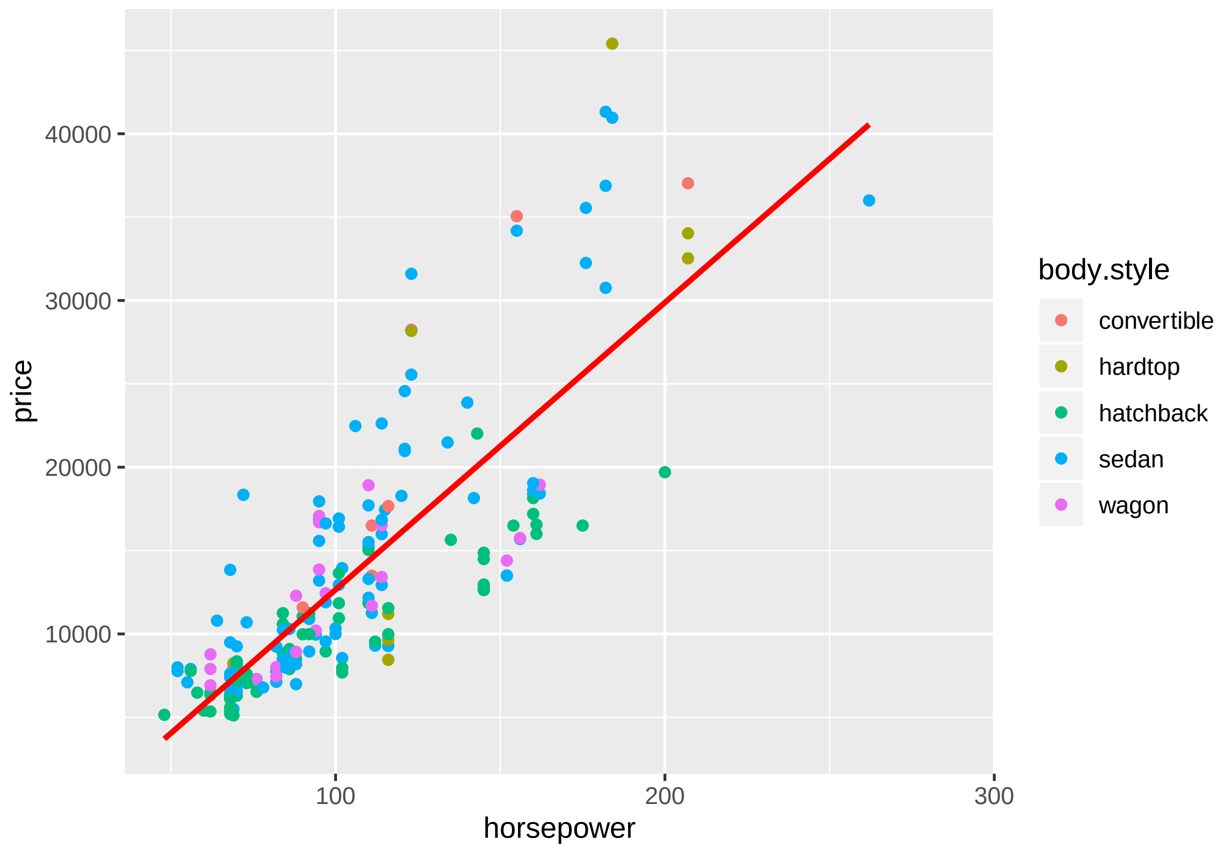
Observation from price and horsepower relationship, by body type horsepower and price has strong positive relationship among all types, with sedan, convertible, and hardtop being the strongest. note: convertible and hardtop has significantly less data points.

When log(price), relationship of horsepower and log(price) becomes more similar between all body types when price^2, relationship bw horsepower and price^2 shows clearer distinction of stronger positive relaitonship among with body types with price^2, and weaker.

We also compare the simple model with model including body style as factors, and the resulting p-value of F-stats is sufficiently low. Therefore, the model with body styles is better than the simple model. Body style variable appears relevant for car prices, beyond horsepower. Note that convertible and hardtop have significantly less data than the other body type.

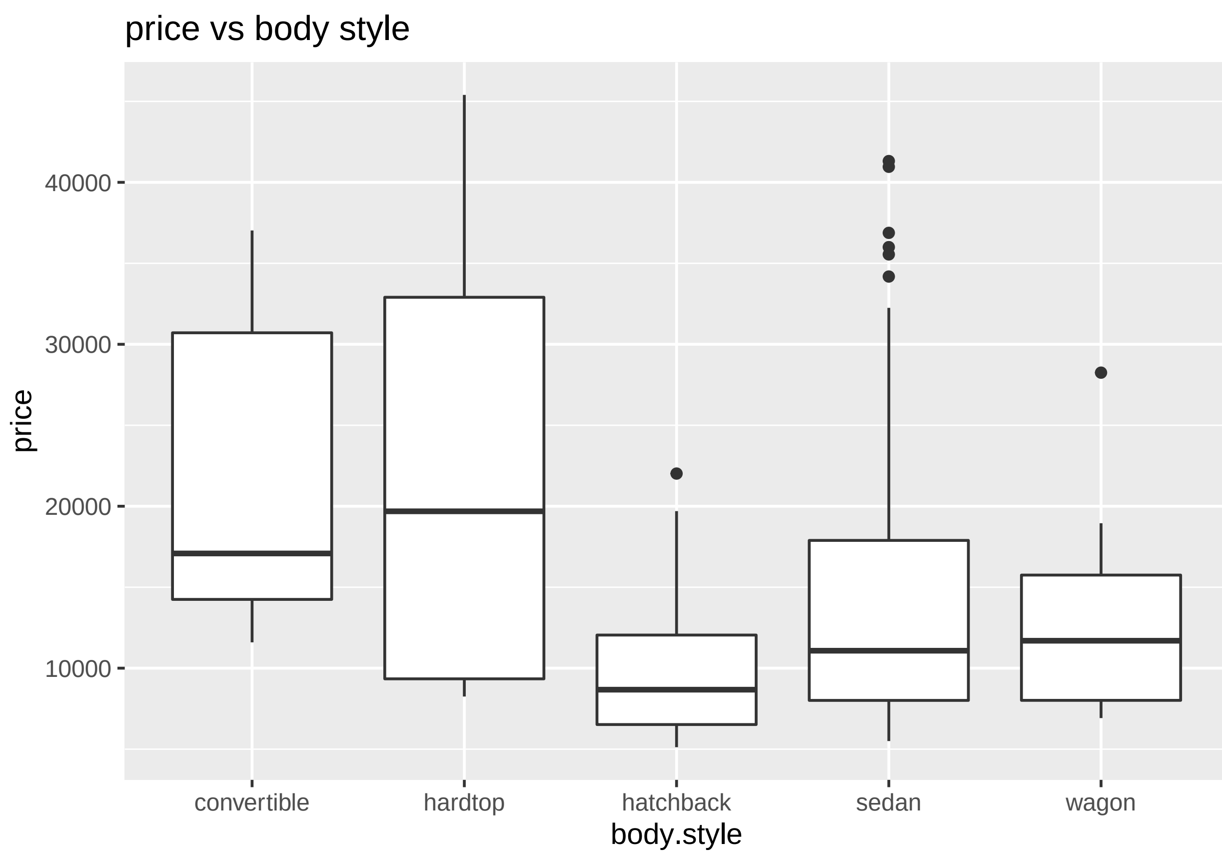
```
# pair-wise relationship
qplot(y=price,x=horsepower,data=data,col = body.style)+geom_smooth(method='lm',col = I('RED'),se = F,mai
```

```
## Warning: Ignoring unknown parameters: main

## Warning: Removed 6 rows containing non-finite values (stat_smooth).

## Warning: Removed 6 rows containing missing values (geom_point).
```



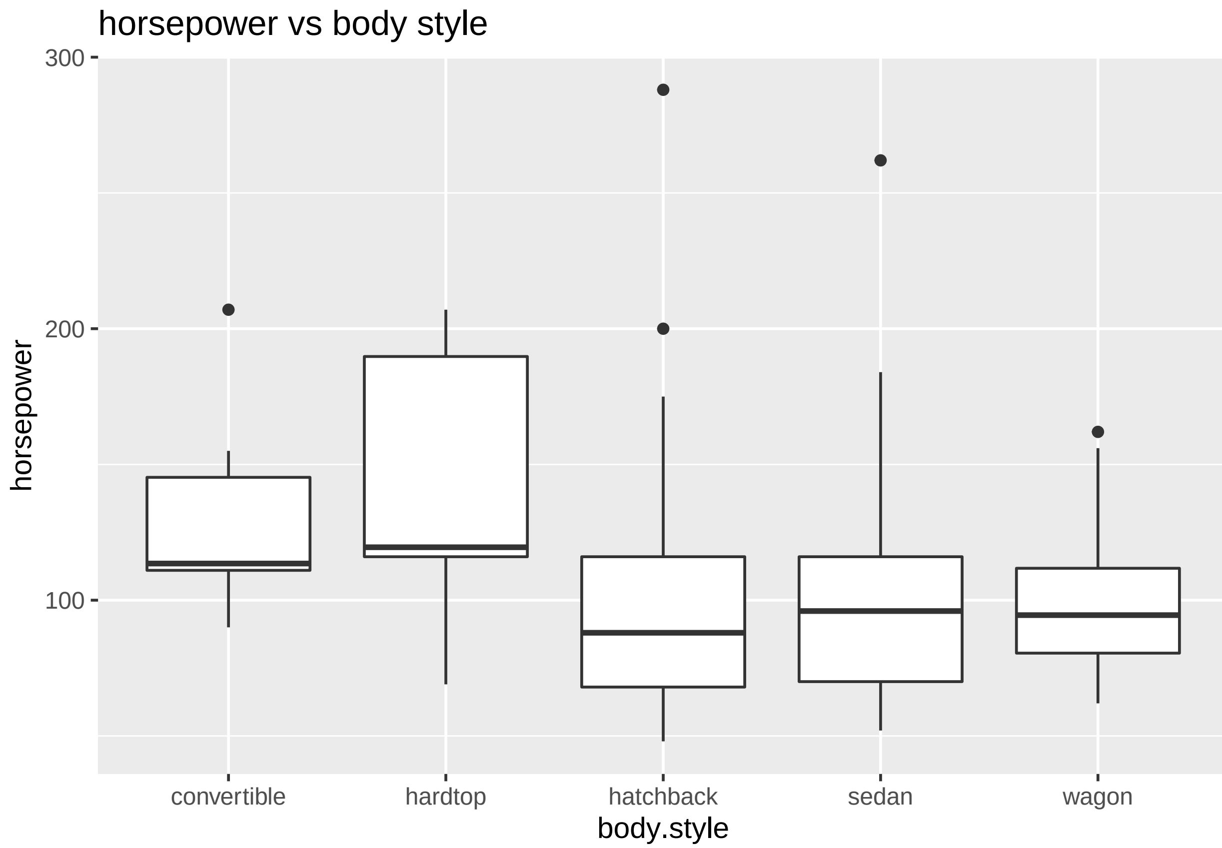

```
qplot(y=price,x=body.style,data=data,geom='boxplot',main = 'price vs body style')
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



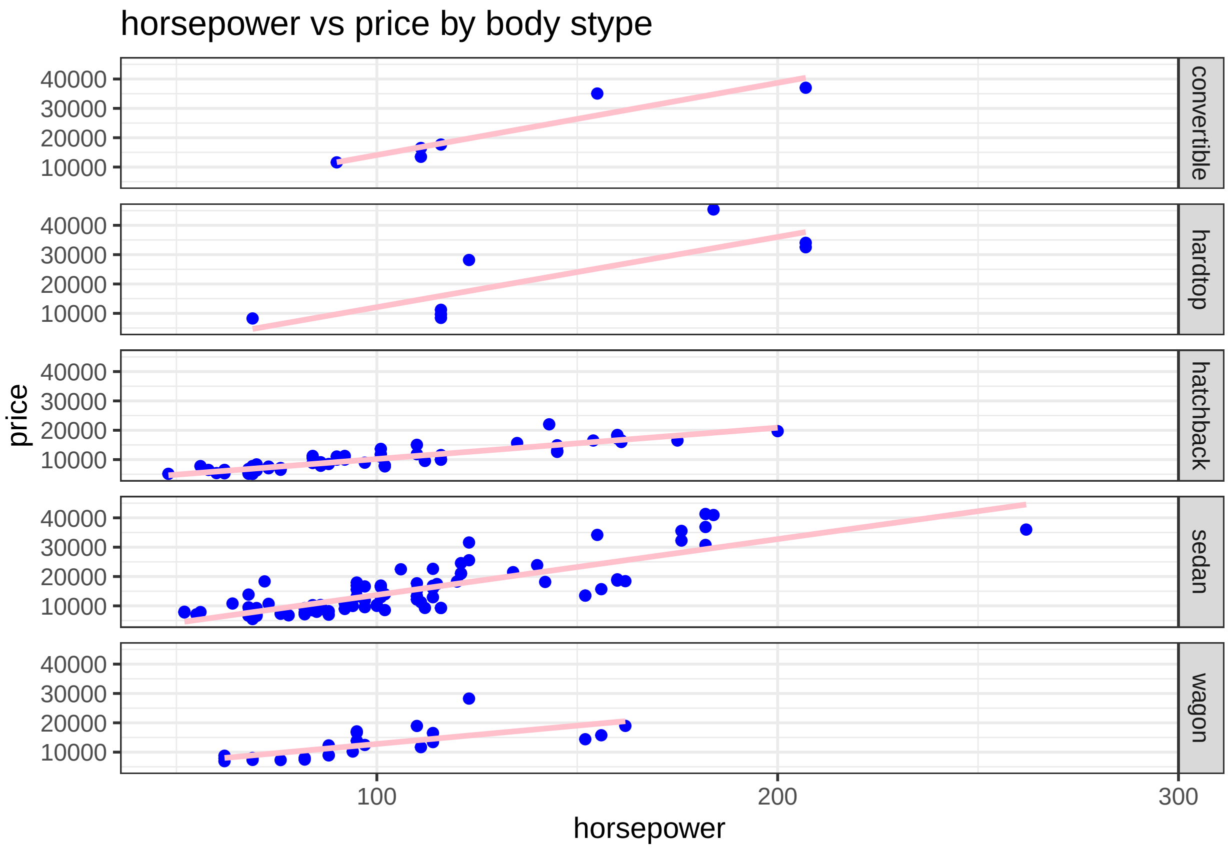

```r
qplot(y = horsepower,x = body.style,data=data,geom='boxplot',main = 'horsepower vs body style')
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```




```
# price and horsepower relationship, by body type
qplot(x = horsepower,y = price,data=data,facets = body.style~.,col = I("blue"),main = 'horsepower vs pri
  geom_smooth(method = 'lm',col = I('pink'),se = F)+ theme_bw()
```

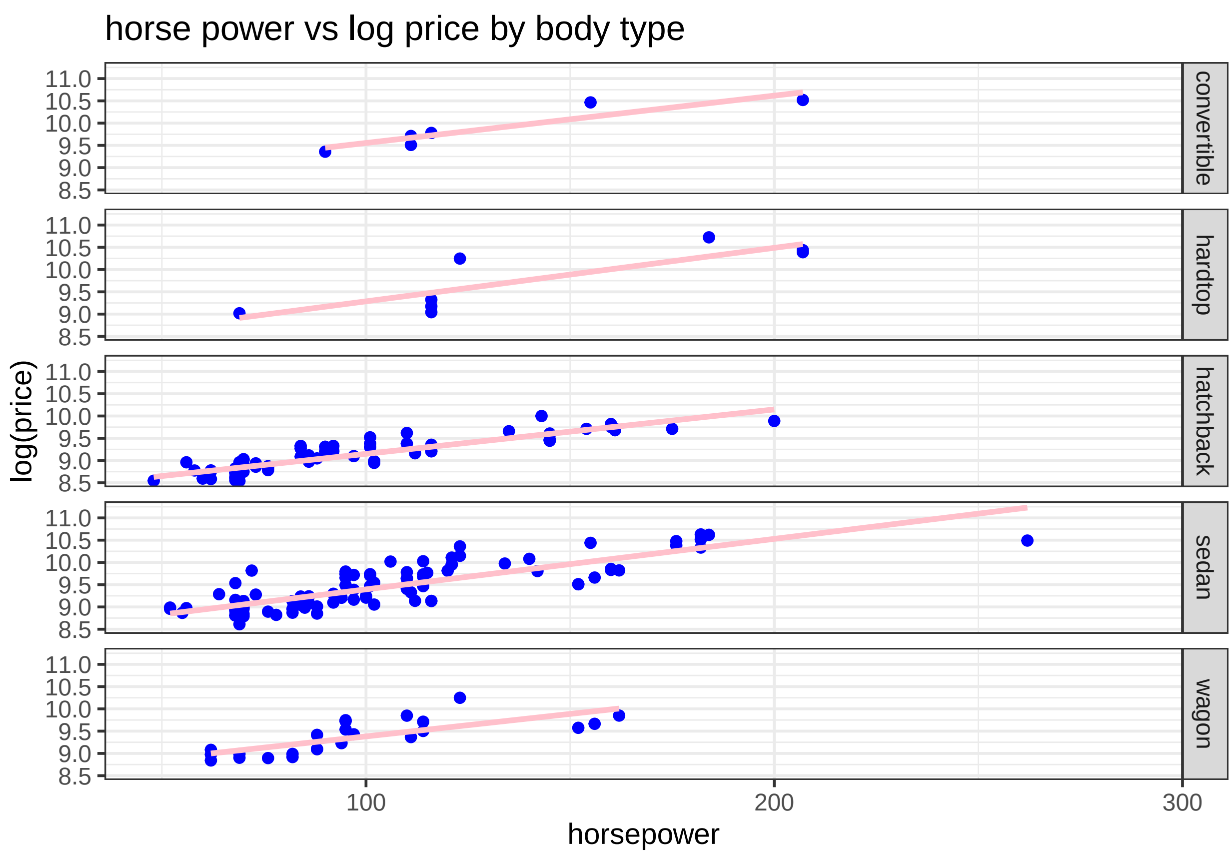
```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).

## Warning: Removed 6 rows containing missing values (geom_point).
```




```
# Transform prices using log and ^2
qplot(x = horsepower,y = log(price),data=data,facets = body.style~.,col = I("blue"),main='horse power vs
  geom_smooth(method = 'lm',col = I('pink'),se = F)+ theme_bw()
```

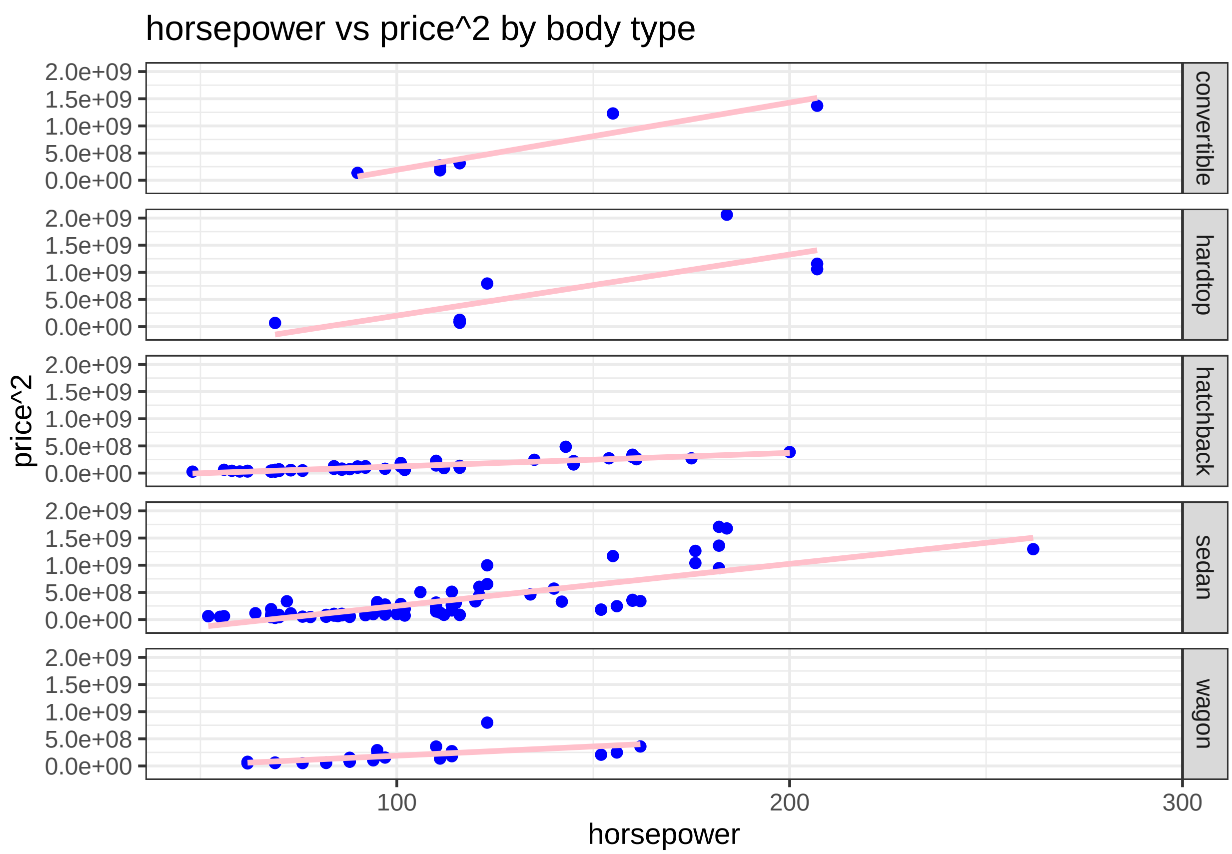
```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```




```
qplot(x = horsepower,y = price^2,data=data,facets = body.style~.,col = I("blue"),main = 'horsepower vs
  geom_smooth(method = 'lm',col = I('pink'),se = F)+ theme_bw()
```

```
## Warning: Removed 6 rows containing non-finite values (stat_smooth).

## Warning: Removed 6 rows containing missing values (geom_point).
```




```
model1 = lm(price ~ horsepower + as.factor(body.style),data = data)
model2 = lm(price ~horsepower, data = data)
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ horsepower + as.factor(body.style)
## Model 2: price ~ horsepower
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    193 3698635524
## 2    197 4323845281 -4 -625209757 8.1561 4.259e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 1 Part 2:

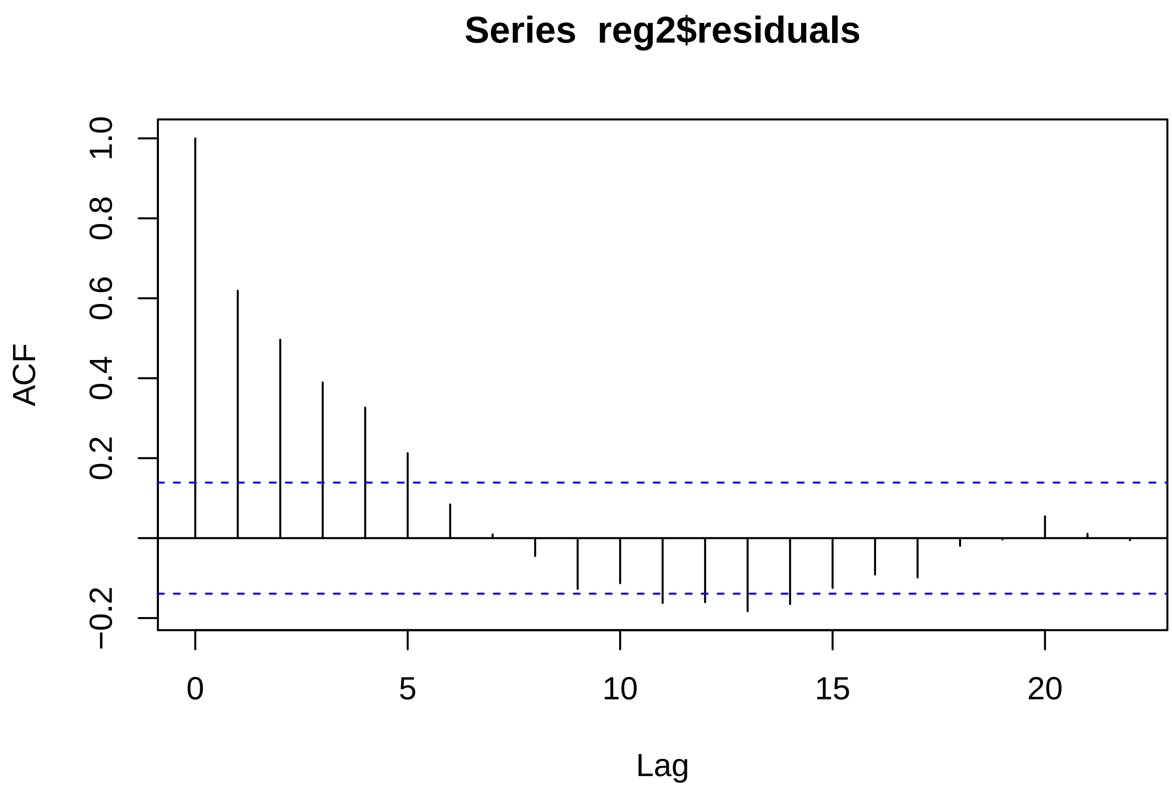
Regression of horsepower(x) and price(y): Statistically significance positive beta, with relatively small standard error. Although horsepower captures price well, but not completely because the residuals has autocorrelations, thus not white noises. Residuals display heteroskedasticity based on the p-value in Breusch-Pagen Test. The p-value is smaller than 0.05, thus reject the null where residuals are homoskedasticity.

```
#str(data)
#summary(data)

# my personal idea: explore relationship between engine size and horsepower (dont need to run)
#data2 = data[!is.na(horsepower) & !is.na(engine.size),c('horsepower','engine.size')]
```

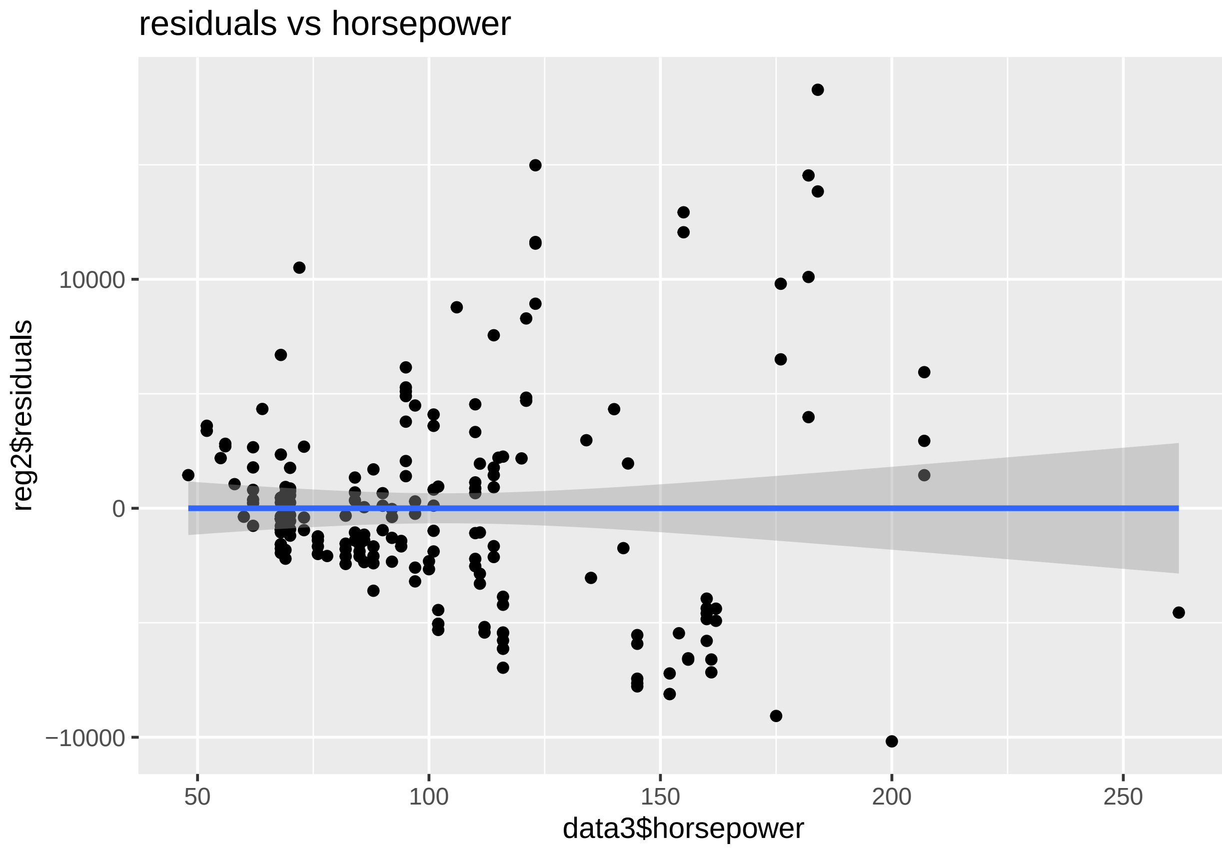
```r
#summary(data2)

#reg = lm(horsepower~engine.size,data = data2)
#lmSumm(reg)
#acf(reg$residuals)
#qplot(x = data2$engine.size,y=reg$residuals)+geom_smooth(method='lm')


# question ask for: explore relationship between horsepower and price
data3 = data[!is.na(horsepower) & !is.na(price),c('horsepower','price')]
#summary(data3)

reg2 = lm(price~horsepower,data = data3)
lmSumm(reg2)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 199 observations
##
## lm(formula = price ~ horsepower, data = data3)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)  -4562.0   975.000   -4.68       0
## horsepower     172.2     8.866   19.42       0
## ---
## Standard Error of the Regression:  4685
## Multiple R-squared:  0.657  Adjusted R-squared:  0.655
## Overall F stat: 377.28 on 1 and 197 DF, pvalue= 0
```

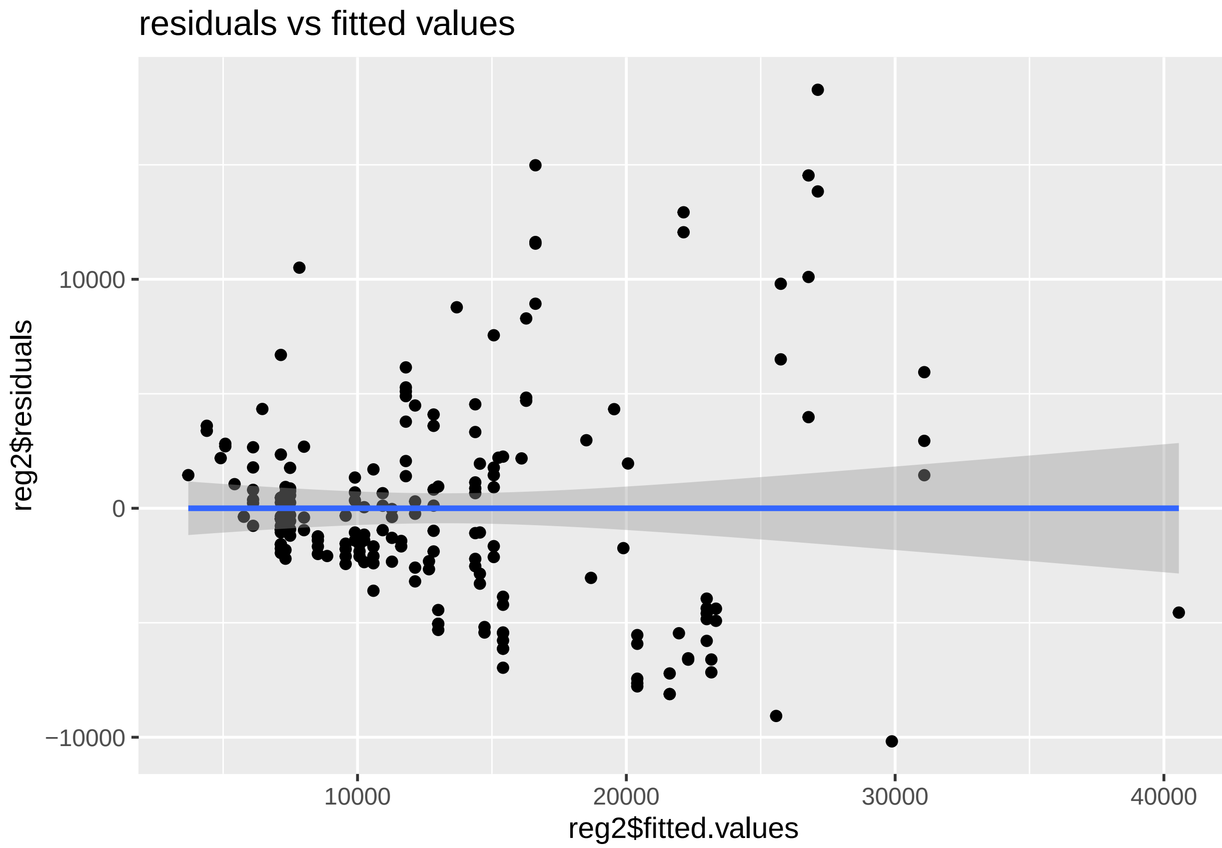
```r
acf(reg2$residuals)
```




```
qplot(x = data3$horsepower,y=reg2$residuals,main = 'residuals vs horsepower')+geom_smooth(method='lm')
```




```
qplot(x = reg2$fitted.values,y=reg2$residuals,main = 'residuals vs fitted values')+geom_smooth(method='l
```

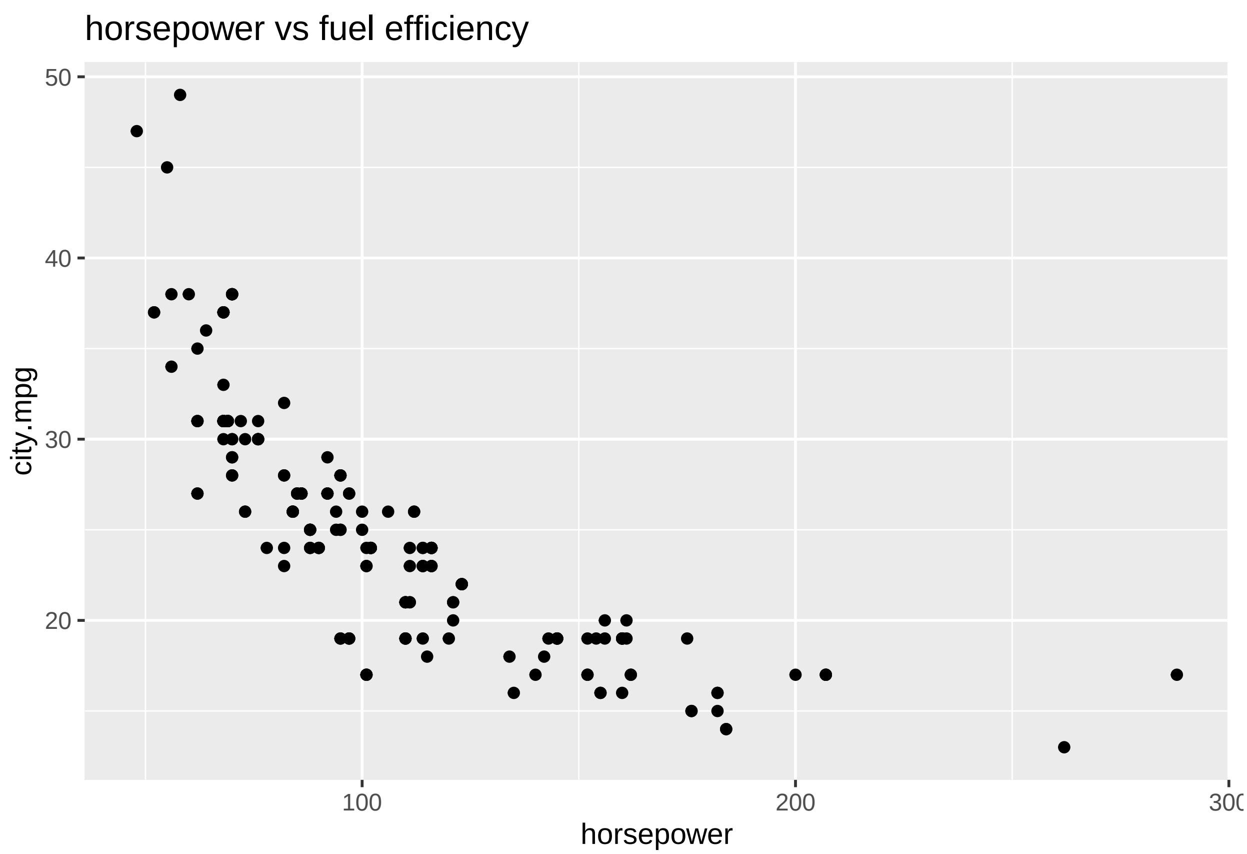


Question 1 Part 3:

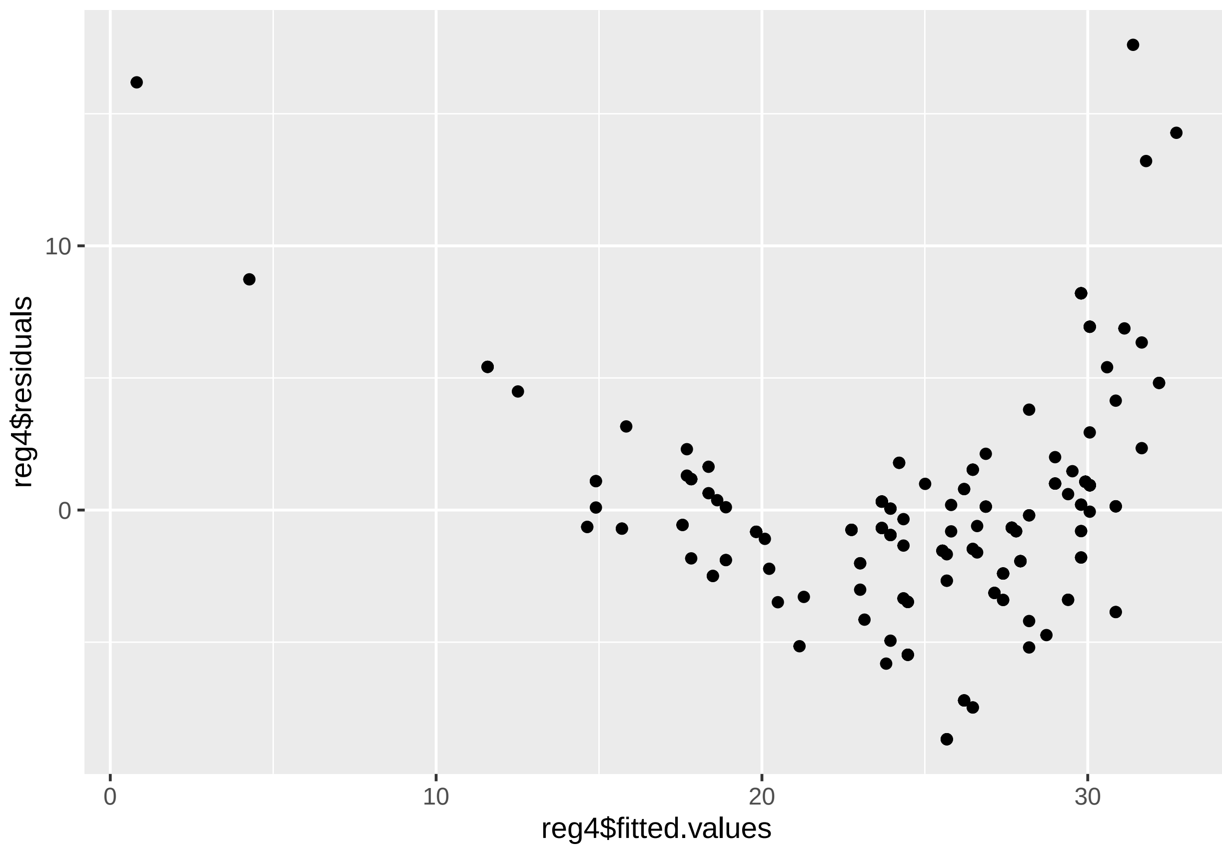
From the plot, increase in horsepower impacts city.mpg (fuel efficiency) negatively, which is, a higher horsepower indicates lower fuel efficiency. It displays a non-linear relationship, ie, negative exponential. A better way to regress is transforming x before linear regression. The regression demonstrates the same results, with beta of -0.133, statistically significant. This beta is small because horse power and city.mpg are on very different scale, but the beta is still negative. The residuals is not uniformly distributed, therefore this could be a nonlinear relationship.

```
#str(data)
qplot(x = horsepower, y = city.mpg, data = data,main = 'horsepower vs fuel efficiency')
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
reg4 = lm(city.mpg~horsepower,data = data)
qplot(y=reg4$residuals,x=reg4$fitted.values)
```




```r
lmSumm(reg4)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 203 observations
##
## lm(formula = city.mpg ~ horsepower, data = data)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)   39.100  0.774600   50.48       0
## horsepower    -0.133  0.006945  -19.14       0
## ---
## Standard Error of the Regression:  3.92
## Multiple R-squared:  0.646  Adjusted R-squared:  0.644
## Overall F stat: 366.48 on 1 and 201 DF, pvalue= 0
```

```r
#plot(exp(seq(1,-1,-0.001)))
```

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.

```
library(foreign) # for read.dta()
```

```
## Warning: package 'foreign' was built under R version 3.5.3
```

```
q2data = as.data.table(read.dta("StockRetAcct_insample.dta"))
```

Question 2 Part 1:

The result of this make sense because the lowest decile portfolio (least stock issuance), has the highest avg return across years (value weighted by firm within each year), vice versa.

```
#summary(q2data)

# assume lnIssue is already lagged by 1 period
q2data[,lnIssue:=jitter(lnIssue, amount = 0)]
for (i in 1980:2014){
  q2data[year == i,
         lnIssueDecile := cut(q2data[year == i,lnIssue],
            breaks = quantile(q2data$lnIssue,probs = seq(0,1,0.1),na.rm = T),
            include.lowest = T,
            labels = F)]
}
q2data = q2data[!is.na(lnIssueDecile),]

q2data[,ExRet:=exp(lnAnnRet) - exp(lnRf)]

# assume MEwt is already lagged by 1 period, mentioned in class, ExRet is current period
decile.ret.yr = q2data[,list(vwret=weighted.mean(ExRet, MEwt)),by=list(year,lnIssueDecile)]

setkey(decile.ret.yr,lnIssueDecile,year)
head(decile.ret.yr)
```

```
##    year lnIssueDecile       vwret
## 1: 1980             1  0.45921566
## 2: 1981             1 -0.17771907
## 3: 1982             1  0.67034888
## 4: 1983             1  0.05385209
## 5: 1984             1  0.21217610
## 6: 1985             1  0.26400801
```

```
decil.ret = decile.ret.yr[,list(ewyr.ret = mean(vwret)),by = lnIssueDecile]
setkey(decil.ret,lnIssueDecile)
decil.ret
```

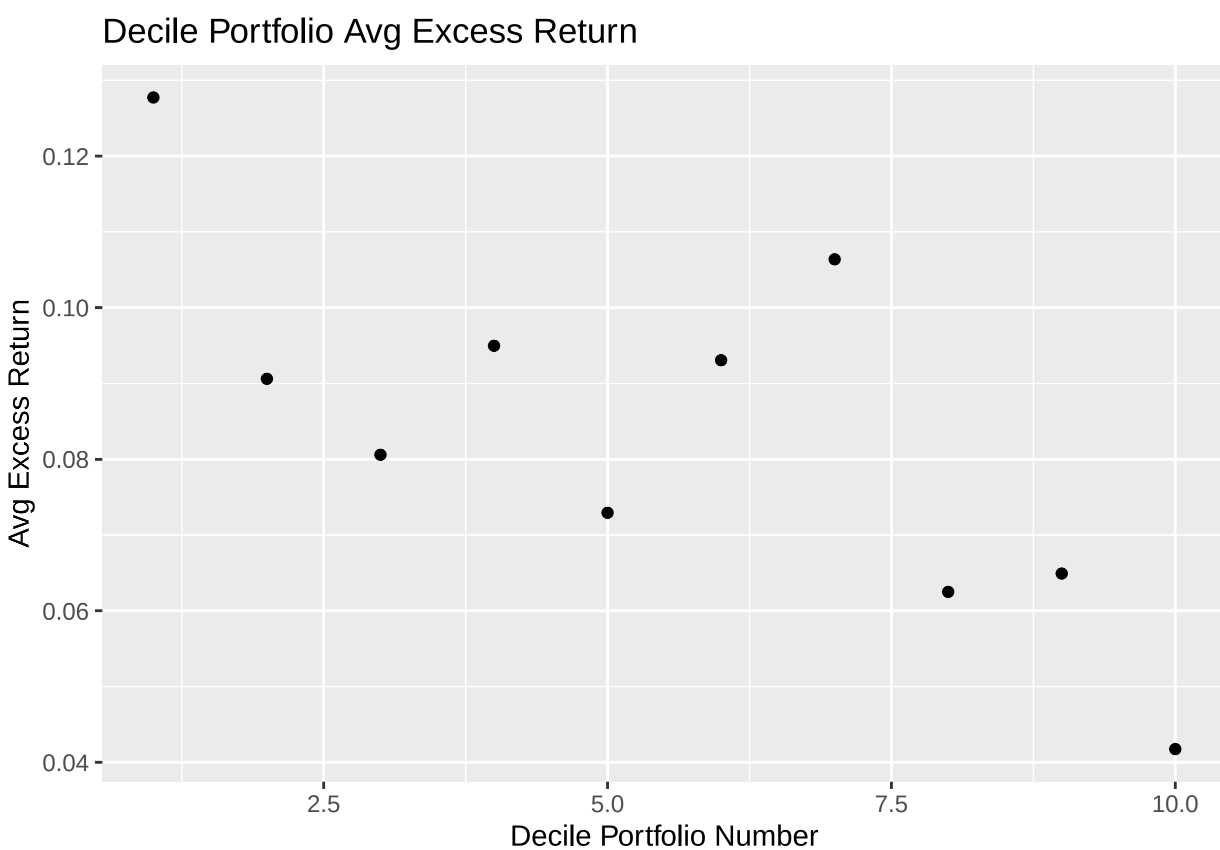
```
##     lnIssueDecile   ewyr.ret
##  1:             1 0.12773413
##  2:             2 0.09061307
##  3:             3 0.08058933
##  4:             4 0.09497489
##  5:             5 0.07293344
##  6:             6 0.09305916
##  7:             7 0.10637763
```

```
## 8:              8 0.06249066
## 9:              9 0.06491555
## 10:            10 0.04174718
```

Question 2 Part 2:

The decile portfolio is created by the quantile of number of issuance, where decile 1 corresponds to companies have the lowest 10% number of issuance. The parttern does not look linear. Although if it's linear, we could get a negative beta in a linear regression of decile vs average return. While portfolio 1 has extremely high return and portfolio 10 has significantly lower return, between decile 2 to 9, the plot almost indicates there's no relationship between the two variables as the points are scatter around its mean evenly.

```
qplot(x = lnIssueDecile,y = ewyr.ret, data = decil.ret,main = 'Decile Portfolio Avg Excess Return',xlab
```



Question 2 Part 3:

Fama Macbeth Regression of excess return on transformed issurance variable: The mean of risk premium of the issurance var is 0.0001542316, a positive relationship. This could indicate that higher decile have higher expected return. This is opposite of what Question 2 Part 2 shown. The fama macbeth coefficient suggests us to long the highest decile portfolio (characteristic = 1), don't invest in decile 2-9 portfolios (characteristic = 0), and short the lowerest decile portfolio (characteristic = -1), however, the regression may be insignificant.

```
q2data[lnIssueDecile=='1', characteristics := -1]
q2data[lnIssueDecile=='10',characteristics := 1]
q2data[lnIssueDecile!= '1' & lnIssueDecile != '10',characteristics := 0]
```

```r
# regression 1: across year for each firm
q2data[, reg1 := (lm(ExRet ~characteristics))$coef[2], by = FirmID]
reg1 = q2data[, list(reg1 = (lm(ExRet ~characteristics))$coef[2]), by = list(FirmID)]
onefirm = q2data[FirmID == 4326, c('ExRet','characteristics') ]
out = lm(ExRet~characteristics, data = onefirm)
lmSumm(out)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 13 observations
##
## lm(formula = ExRet ~ characteristics, data = onefirm)
##
## Coefficients:
##                  Estimate Std Error t value p value
## (Intercept)     -0.033420    0.1069   -0.31   0.760
## characteristics -0.004367    0.2225   -0.02   0.985
## ---
## Standard Error of the Regression:  0.338
## Multiple R-squared:  0  Adjusted R-squared:  -0.091
## Overall F stat: 0 on 1 and 11 DF, pvalue= 0.985
```

```r
# regression 2: across firms for each year
q2data[,reg2 := (lm(ExRet~reg1))$coef[2],by=year]
reg2 = q2data[,list(reg2 = (lm(ExRet~reg1))$coef[2]),by=list(year)]
setkey(reg2,year)
reg2
```

```
##     year         reg2
##  1: 1980  0.098620993
##  2: 1981 -0.037697569
##  3: 1982  0.282116125
##  4: 1983 -0.025110739
##  5: 1984 -0.009425911
##  6: 1985 -0.055295593
##  7: 1986 -0.001979419
##  8: 1987 -0.047739064
##  9: 1988 -0.068010054
## 10: 1989 -0.033222688
## 11: 1990 -0.036400301
## 12: 1991 -0.043754059
## 13: 1992 -0.031304134
## 14: 1993  0.112377194
## 15: 1994  0.077720954
## 16: 1995  0.038462624
## 17: 1996  0.002892592
## 18: 1997  0.004121591
## 19: 1998 -0.089510924
## 20: 1999 -0.091120942
## 21: 2000 -0.067787955
## 22: 2001  0.014809543
## 23: 2002 -0.071281332
## 24: 2003  0.030193456
## 25: 2004 -0.016931320
## 26: 2005  0.033192461
## 27: 2006  0.024757153
```

```
## 28: 2007 -0.043750804
## 29: 2008 -0.058750100
## 30: 2009 -0.025169292
## 31: 2010  0.086000501
## 32: 2011  0.056058293
## 33: 2012 -0.012722597
## 34: 2013  0.055983035
## 35: 2014 -0.022206342
##     year         reg2
```

```
reg2[,mean(reg2)]
```

```
## [1] 0.0008038679
```

Question 3: Double-sorts and functional forms

Question 3 Part 1:

```
for (i in 1980:2014){
  q2data[year == i,
         bmDecile := cut(q2data[year == i,lnBM],
            breaks = quantile(q2data$lnBM,probs = seq(0,1,0.2),na.rm = T),
            include.lowest = T,
            labels = F)]
}

for (i in 1980:2014){
  q2data[year == i,
         sizeDecile := cut(q2data[year == i,lnME],
            breaks = quantile(q2data$lnME,probs = seq(0,1,0.2),na.rm = T),
            include.lowest = T,
            labels = F)]
}

head(q2data)
```

```
##    FirmID year   lnAnnRet       lnRf         MEwt     lnIssue       lnMom
## 1:      6 1980  0.3636313 0.07894428 2.814308e-04  0.02677333  0.07535515
## 2:      6 1981 -0.2904088 0.13019902 3.214631e-04  0.01077732  0.51265192
## 3:      6 1982  0.1866300 0.13070259 2.663127e-04 -0.03092885 -0.22050542
## 4:      6 1983  0.4898190 0.08983046 1.699149e-04 -0.05867182  0.04621762
## 5:     10 1991 -0.5080047 0.06121579 3.269729e-05  0.10440025  1.34105313
## 6:     12 2000 -1.3568472 0.06197736 1.219181e-05  0.20096522  0.25174579
##        lnME     lnProf        lnEP       lnInv   lnLever       lnROE
## 1: 12.58147  0.2017671  0.14641121  0.09362611 0.6960014  0.09529421
## 2: 12.90800  0.2156609  0.10255504  0.08724214 0.7098430  0.08217967
## 3: 12.55777  0.1840875  0.11954752  0.11166344 0.7309716  0.07951558
## 4: 12.56195  0.1655312  0.11592383 -0.03311720 0.7108847  0.05537406
## 5: 11.56583  0.2397878  0.02314729  0.30005118 0.4187644  0.14682838
## 6: 12.27575 -0.3823268 -0.02378274 -0.17460629 0.8244712 -0.59177256
##            rv       lnBM ff_ind lnIssueDecile      ExRet characteristics
## 1: 0.08413413  0.6333913      3             5  0.3563997               0
## 2: 0.05638131  0.3567226      3             4 -0.3910973               0
## 3: 0.06207170  0.7794052      3             3  0.0655525               0
```

```
## 4: 0.07695480  0.7021134      3             2  0.5380321                0
## 5: 0.37436786 -2.1609421     10             7 -0.4614334                0
## 6: 1.06719565 -3.8155227      6             8 -0.8064670                0
##    reg1        reg2 bmDecile sizeDecile
## 1:   NA  0.09862099        5          1
## 2:   NA -0.03769757        5          2
## 3:   NA  0.28211612        5          1
## 4:   NA -0.02511074        5          1
## 5:   NA -0.04375406        1          1
## 6:   NA -0.06778795        1          1
```

Question 3 Part 2:

The assumption I am testing here is *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.* From the plot, I observed a weak linear relationship between b-m ratio and average expected returns. The relationship is stronger for small and medium-small companies. However, the linear relationship for medium, medium-large, and large companies are neutral and slightly negative. Therefore, the assumption of conditional linearity seem to be applicable here.

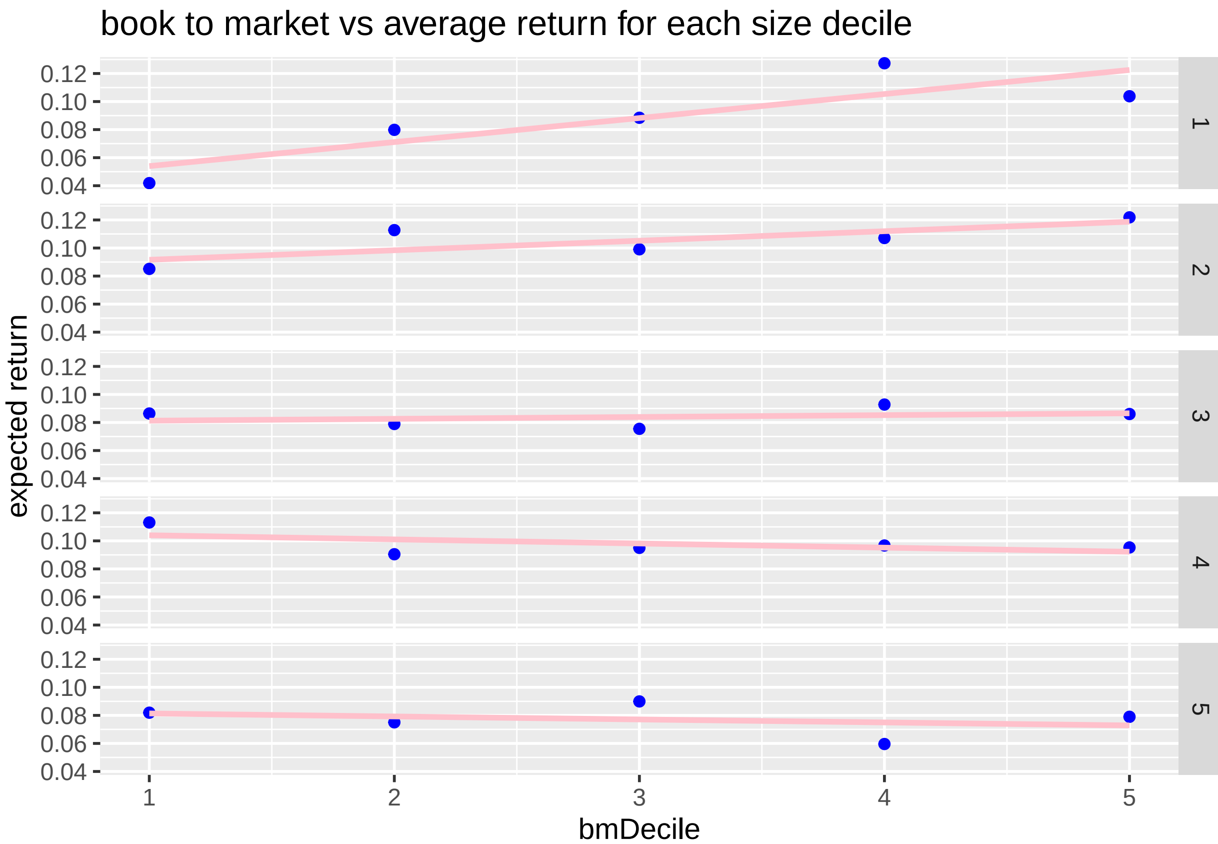
```
q2data = q2data[!is.na(lnIssueDecile) & !is.na(sizeDecile) & !is.na(bmDecile),]



dbsorted.ret = q2data[,list(vwret=weighted.mean(ExRet, MEwt)),by=list(year,sizeDecile,bmDecile)]

setkey(dbsorted.ret,sizeDecile,bmDecile,year)
head(dbsorted.ret)
```

```
##    year sizeDecile bmDecile       vwret
## 1: 1980          1        1  0.47812930
## 2: 1981          1        1 -0.42843120
## 3: 1982          1        1  0.88658821
## 4: 1983          1        1 -0.40853566
## 5: 1984          1        1  0.05162629
## 6: 1985          1        1  0.39453224
```

```
db.avgret = dbsorted.ret[,list(ewyr.ret = mean(vwret)),by = list(sizeDecile,bmDecile)]
setkey(db.avgret,sizeDecile,bmDecile)
head(db.avgret)
```

```
##    sizeDecile bmDecile   ewyr.ret
## 1:          1        1 0.04185860
## 2:          1        2 0.07982026
## 3:          1        3 0.08845356
## 4:          1        4 0.12730952
## 5:          1        5 0.10380049
## 6:          2        1 0.08510046
```

```
qplot(x = bmDecile, y = ewyr.ret,data = db.avgret,facets = sizeDecile~.,col = I('BLUE'),ylab='expected :
```


book to market vs average return for each size decile
expected return
bmDecile
0.12
0.10
0.08
0.06
0.04
1
2
3
4
5