# Data Analytics and Machine Learning PS1 Group 7

*Jiaming Huang, Hogun Kim, Yichu Li, An Yang*

*2019/4/6*

## Question 1 : On ggplot2 and regression planes

The classic dataset, diamonds, (you must load the ggplot2 package to access this data) has about 50,000 prices of diamonds along with weight (carat) and quality of cut (cut).

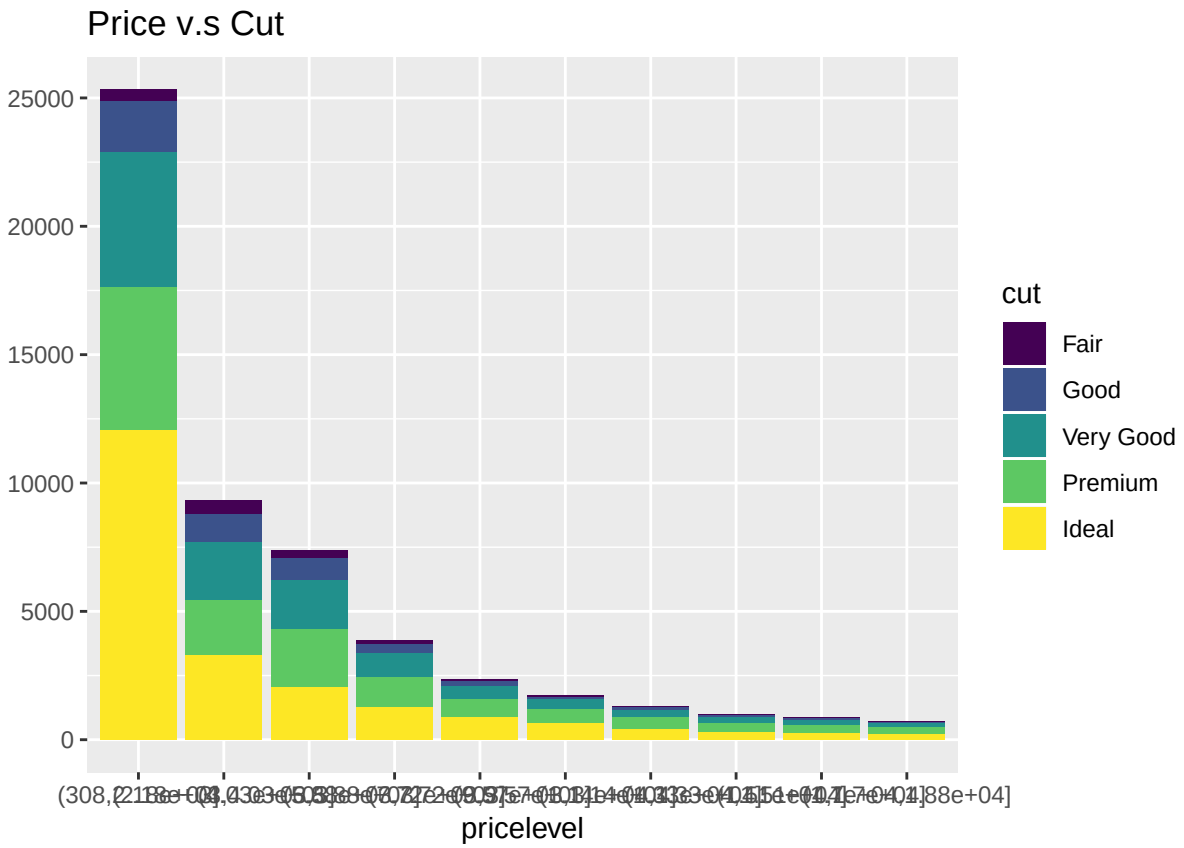
1. Use ggplot2 to visualize the relationship between price and carat and cut. price in the dependent variable. Consider both the log() and sqrt() transformation of price.

a. Initially, we discover whether for different price span, there is some tendency of weight or cut through bar graph:

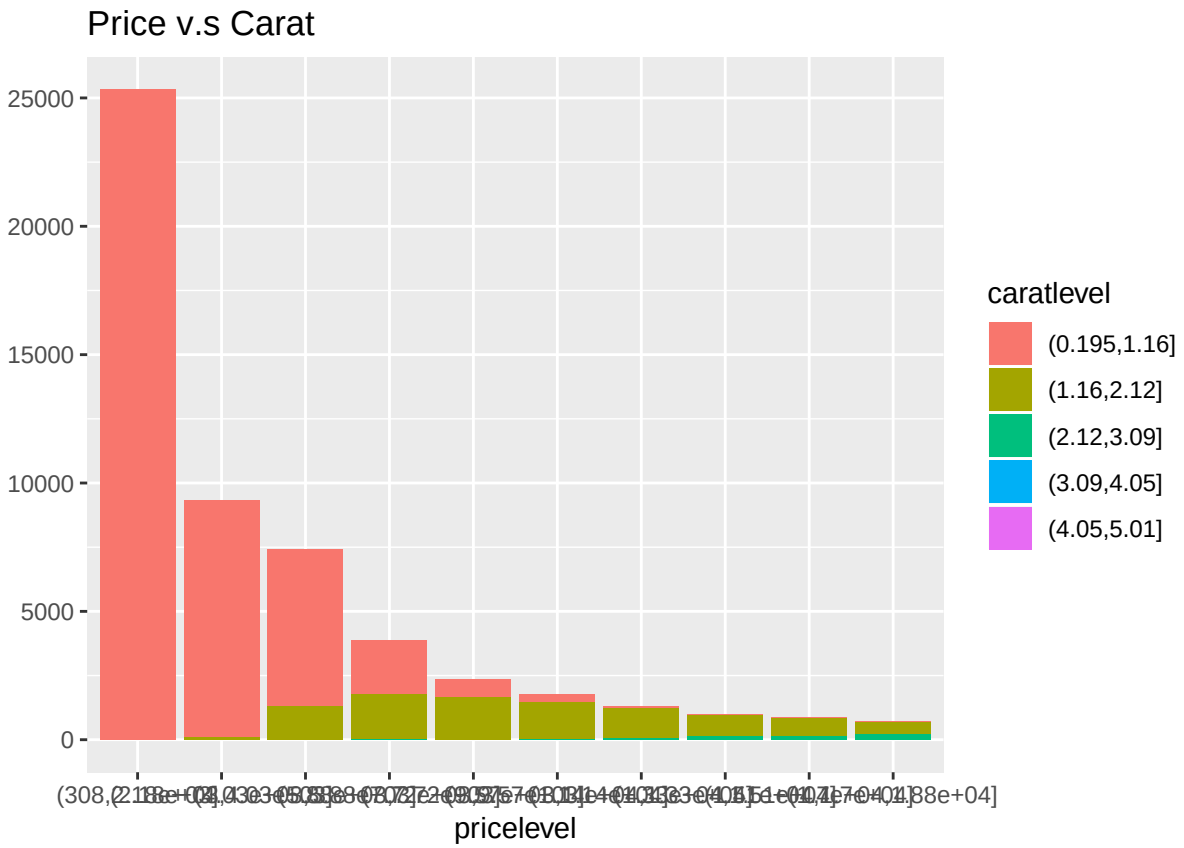
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
cutf=as.character(diamonds$cut)
cutf=as.factor(cutf)
diamonds$cutf = cutf
diamonds$caratlevel = cut(diamonds$carat,breaks = 5)
diamonds$pricelevel = cut(diamonds$price,breaks = 10)

qplot(pricelevel, data=diamonds, fill=cut, geom="bar", main = "Price v.s Cut")
```



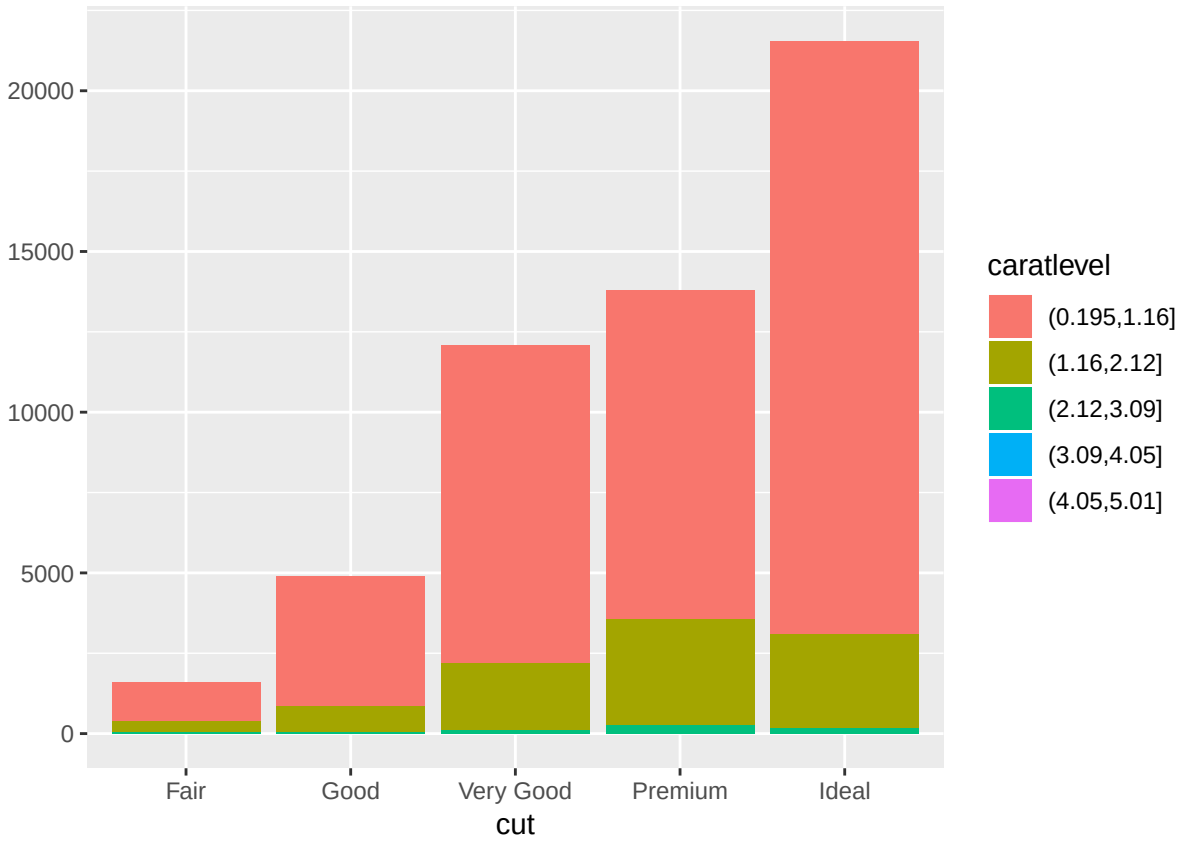

```
qplot(pricelevel, data=diamonds, fill=caratlevel, geom="bar",main = "Price v.s Carat")
```



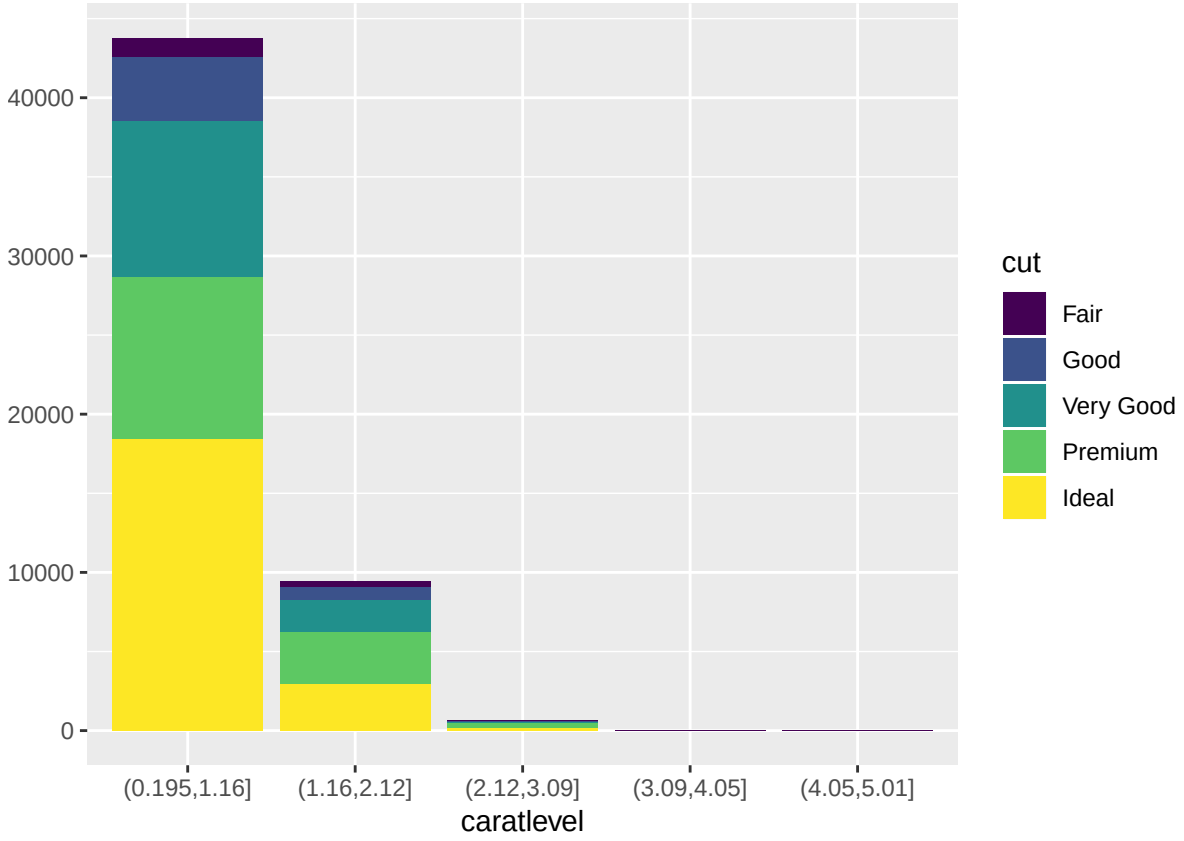

As we can see from above that it seems like more ideal the cut is and more weight diamond is, more expensive the diamond is.

b. Then, we explore the relationship between cut and weight, we want to see if there is some connection between those two effects.

```
qplot(cut, data=diamonds, fill=caratlevel, geom="bar")
```



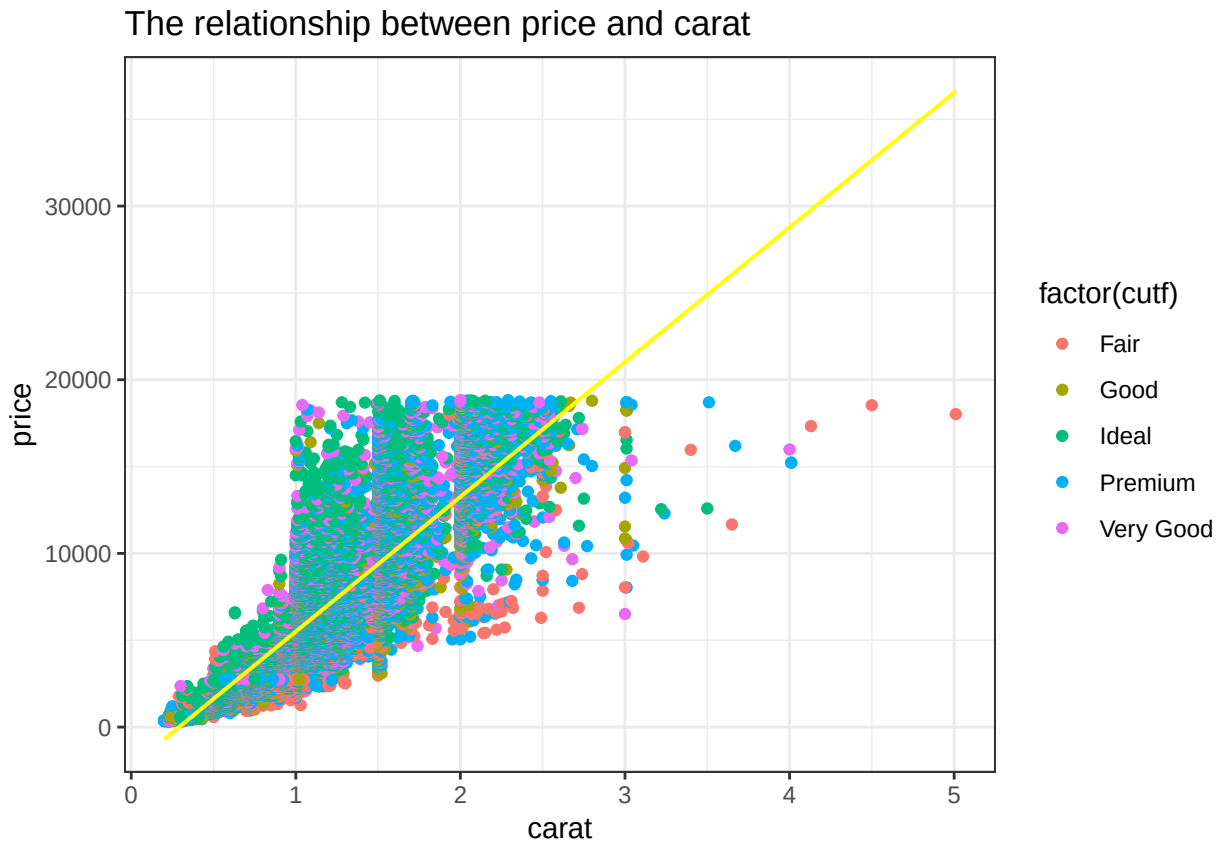

```
qplot(caratlevel, data=diamonds, fill=cut, geom="bar")
```



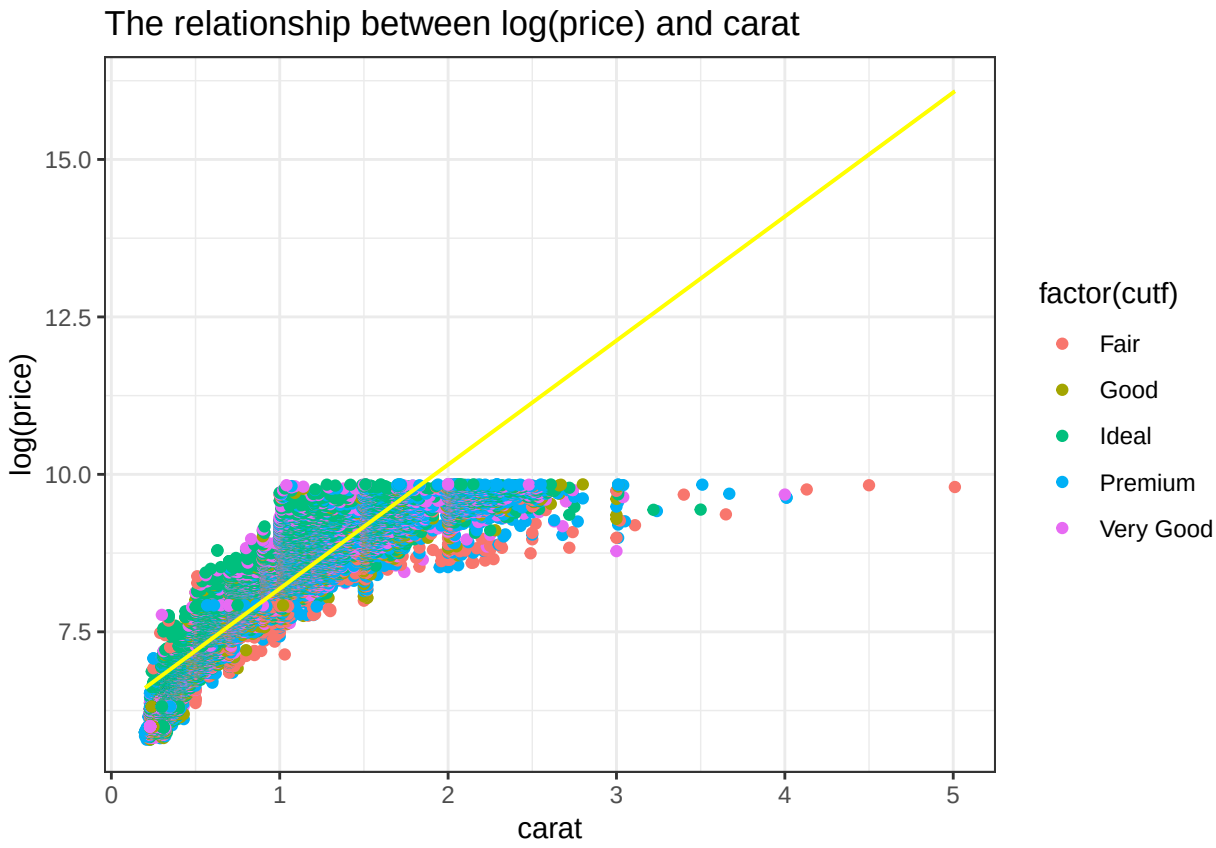

from these two graphs, it seems that those two characteristics are uncorrelated.

c. then we explore the relationship between price and weight.

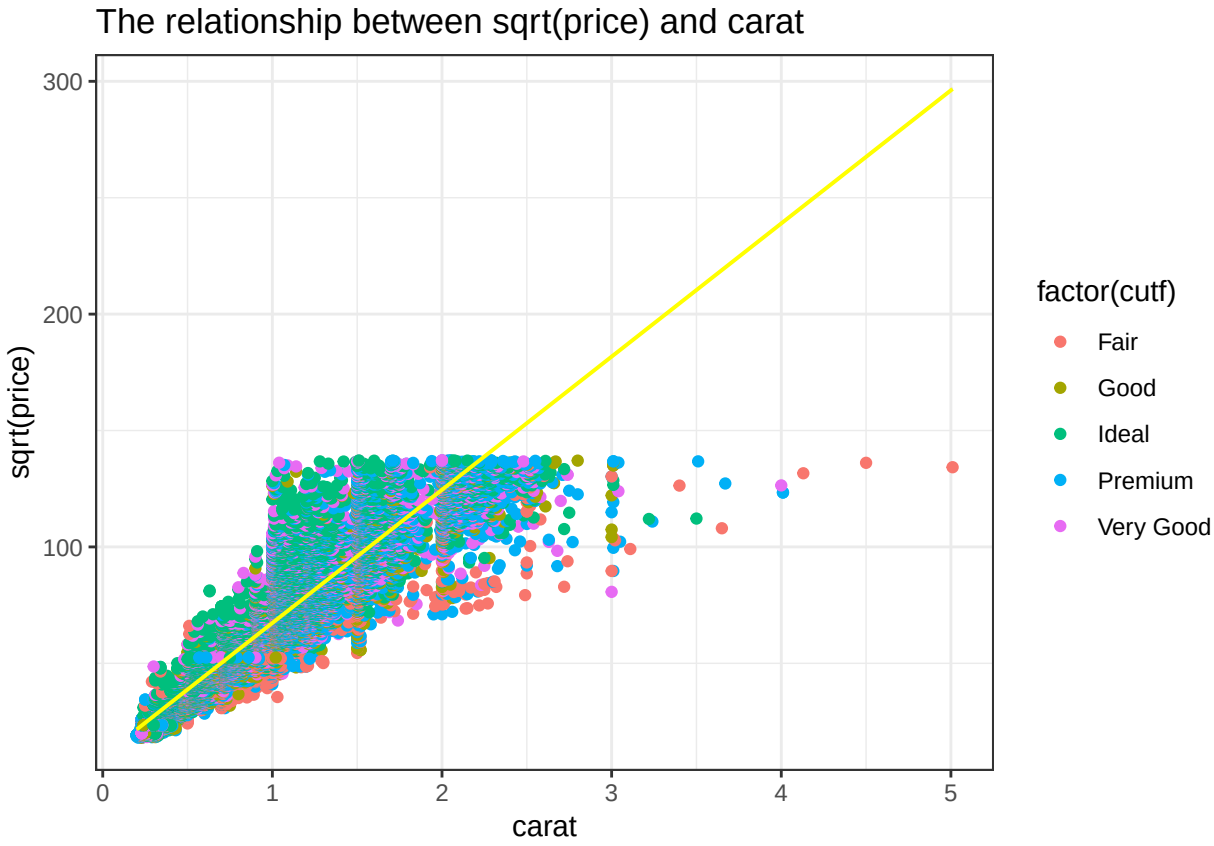
```
qplot(carat,price,data = diamonds, colour = factor(cutf),main = "The relationship between price and cara
```




```
qplot(carat,log(price),data = diamonds, colour = factor(cutf),main = "The relationship between log(price
```

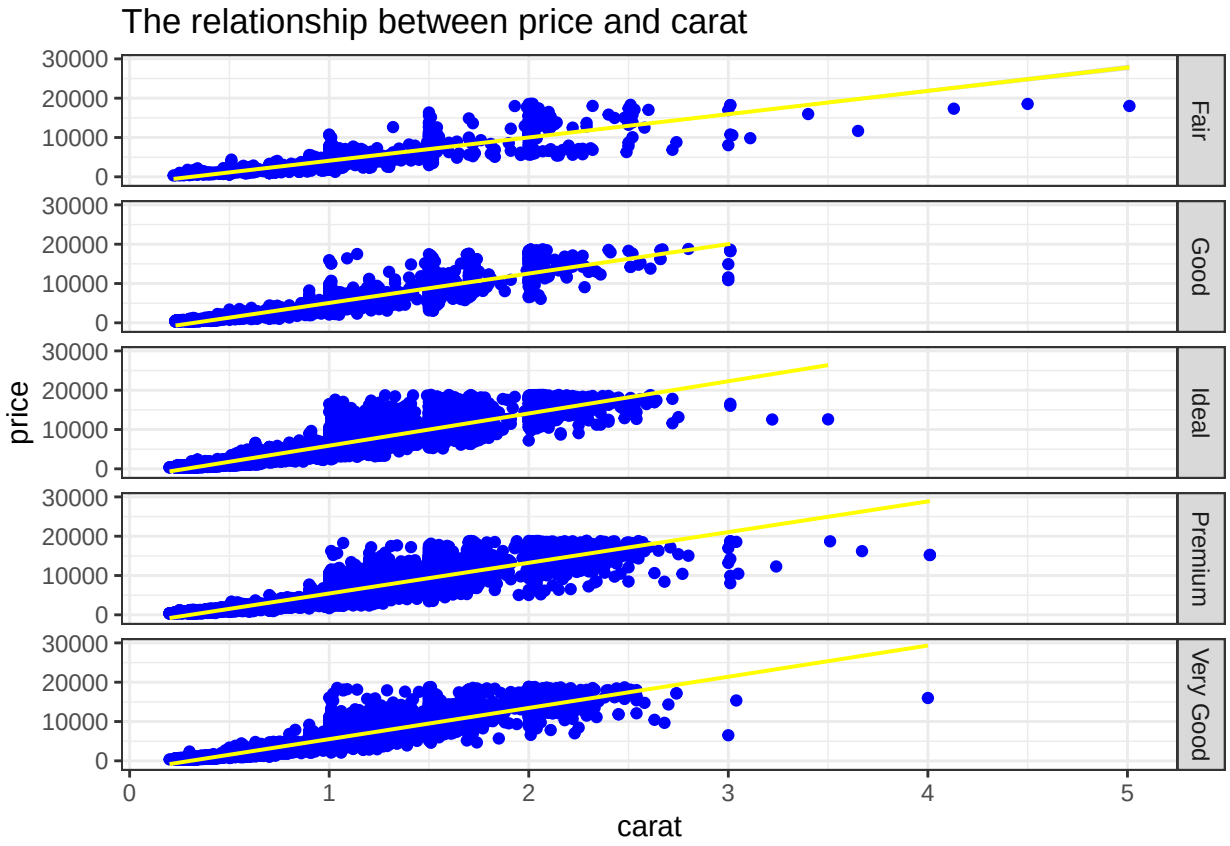



```
qplot(carat,sqrt(price),data = diamonds, colour = factor(cutf),main = "The relationship between sqrt(pr
```




we can see from the graphs above that a. there seems to be a linear correlation between price and weight. b. price has highest limitation when weight increases, price will not increase if it touches the limitation. c. sqrt(price) and log(price) seem to have lower standard error than price.

```
qplot(carat,price,data = diamonds,facets = cutf~., col = I("blue"),main = "The relationship between pri
```

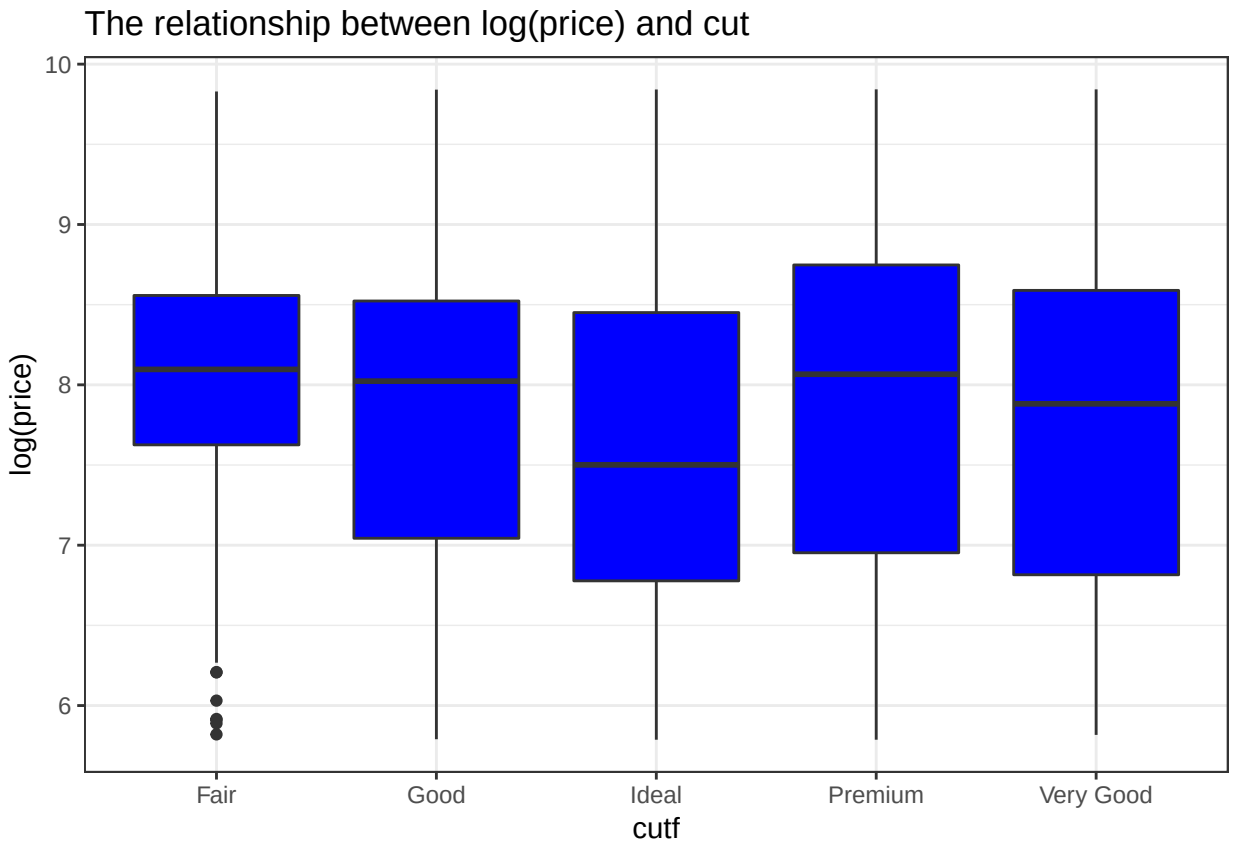



Also, for different cut, the line coefficients are same, which means cut may not effect the relation between weight and price.

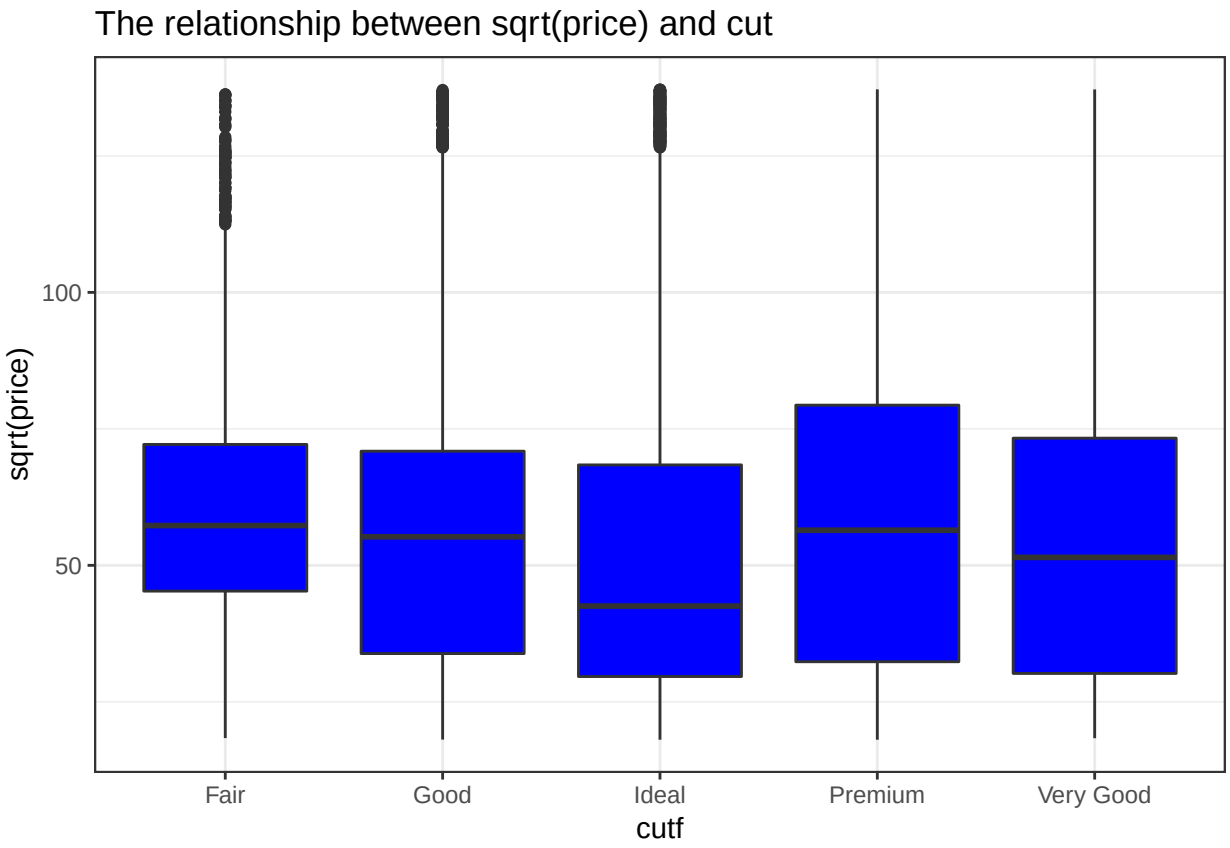
d. Eventually, we explore the relation between cut and price.

```
qplot(cutf, price, data = diamonds,geom = "boxplot",fill = I("blue"),main = "The relationship between pr
```


```
qplot(cutf, log(price), data = diamonds,geom = "boxplot",fill = I("blue"),main = "The relationship betw
```

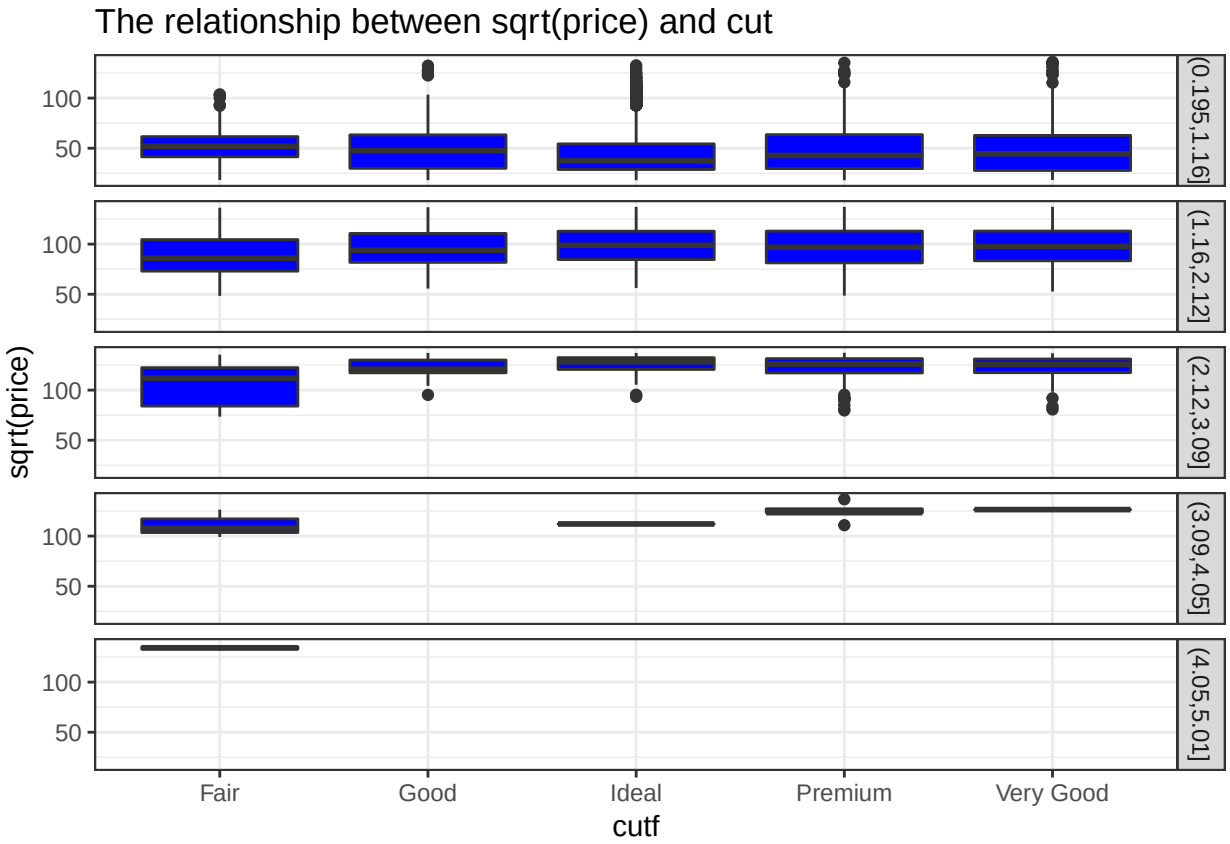



```
qplot(cutf, sqrt(price), data = diamonds,geom = "boxplot",fill = I("blue"),main = "The relationship betw
```




From above graphs, we consider there is no correlation between cut and price, but we still want to see if it is because of disturbance of weight. Then we seperate the characteristic of weight.

```
qplot(cutf, sqrt(price), data = diamonds,facets = caratlevel~.,geom = "boxplot",fill = I("blue"),main =
```




We now know that for small weight diamonds(weight less than 1), cut is not an influential factor for the price of diamond. However, for large weight diamonds(weight more than 1), cut is likely to be linear correlated with price.

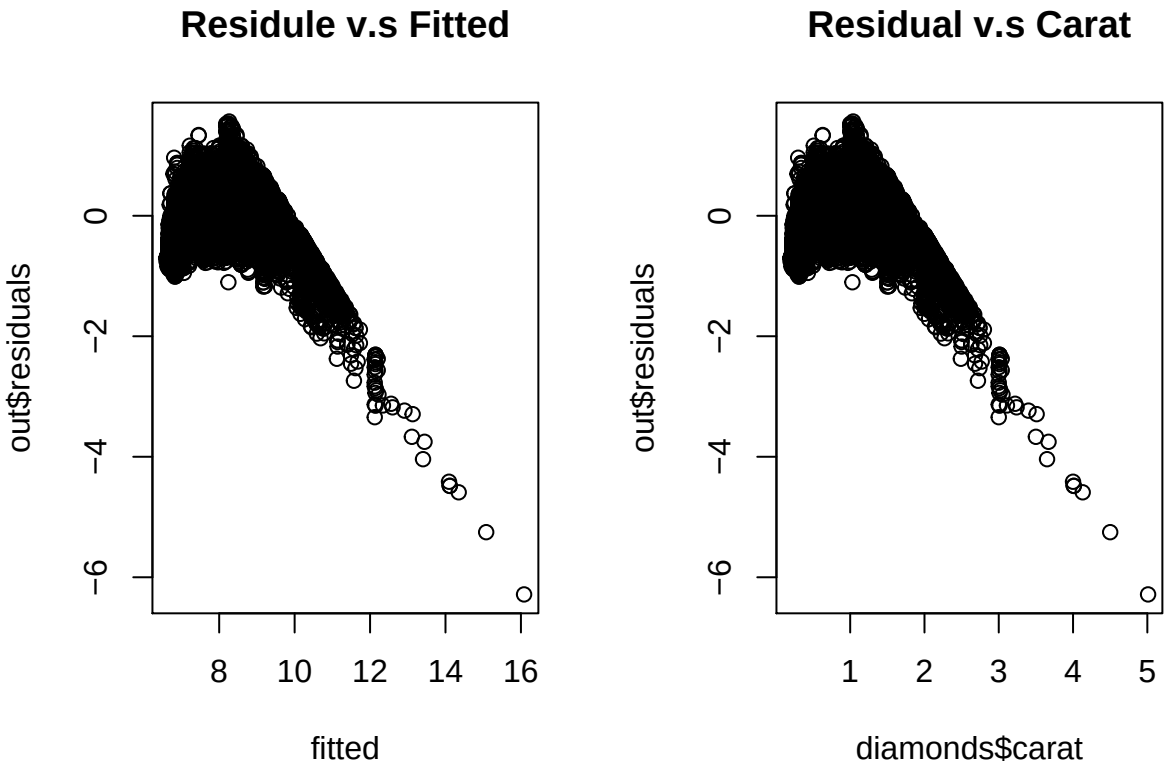
2. Run a regression of your preferred specification. Perform residual diagnostics as you learned in 237Q,1. What do you conclude from your regression diagnostic plots of residuals vs. fitted and residuals vs. carat?

```
par(mfrow = c(1,2))
out = lm(log(price)~carat, data = diamonds)
summary(out)
```

```
##
## Call:
## lm(formula = log(price) ~ carat, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2844 -0.2449  0.0335  0.2578  1.5642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.215021   0.003348    1856   <2e-16 ***
## carat       1.969757   0.003608     546   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3972 on 53938 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8468
```

```
## F-statistic: 2.981e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

```
fitted = out$coefficients[1]+out$coefficients[2]*diamonds$carat
plot(x = fitted, y = out$residuals, main = "Residule v.s Fitted", type = "p")
plot(x = diamonds$carat, y = out$residuals, main = "Residual v.s Carat", type = "p")
```




The conclusion is that the residual is negatively linear with fitted value and carat(weight), as though price is linear correlated with carat, as carat becomes so large, the price touches the highest limitation, price will remain around highest price and cannot continously grow.

## 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, value??? weighting stocks within each portfolio each year, equal???weighting across years.

```
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.4.4
```

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

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
setwd("/Users/jiaminghuang/Downloads")
StockRetAcct = read.dta("StockRetAcct_insample.dta")
StockRetAcct_DT = as.data.table(StockRetAcct)
StockRetAcct_DT =  na.omit(StockRetAcct_DT)
StockRetAcct_DT$lnAnnRet = exp(StockRetAcct_DT$lnAnnRet) - exp(StockRetAcct_DT$lnRf)
StockRetAcct_DT[,lnIssue:=jitter(lnIssue, amount = 0)]
StockRetAcct_DT[,lnIssue_vingtile:=cut(StockRetAcct_DT$lnIssue,breaks = quantile(StockRetAcct_DT$lnIssu
for (i in 1980:2014)
{
  StockRetAcct_DT[year == i,issue_vingtile_yr:=cut(StockRetAcct_DT[year == i,]$lnIssue,
  breaks=quantile(StockRetAcct_DT[year == i,]$lnIssue,probs=c(0:10)/10,na.rm=TRUE),
  include.lowest=TRUE, labels=FALSE)]
}
Mean1 = StockRetAcct_DT[, list(WeightedMean = weighted.mean(lnAnnRet, MEwt)), by = list(issue_vingtile_y
Mean2 = Mean1[, list(Mean = mean(WeightedMean)), by = issue_vingtile_yr]
Mean2
```

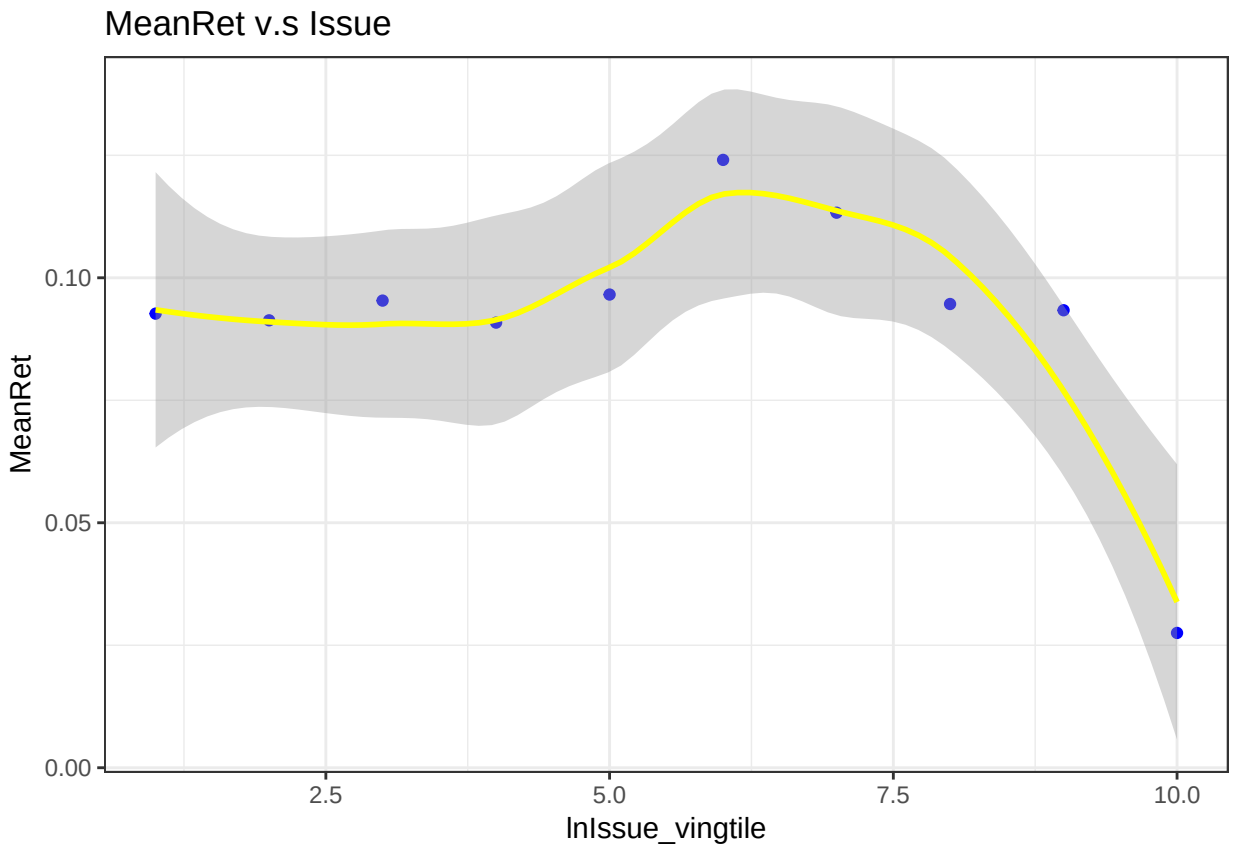
```
##     issue_vingtile_yr       Mean
##  1:                 5 0.07971176
##  2:                 7 0.08576898
##  3:                 1 0.11264212
##  4:                 8 0.05784967
##  5:                 3 0.07389554
##  6:                 2 0.08733117
##  7:                10 0.03682874
##  8:                 9 0.08120383
##  9:                 4 0.08013198
## 10:                 6 0.09412326
```

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.

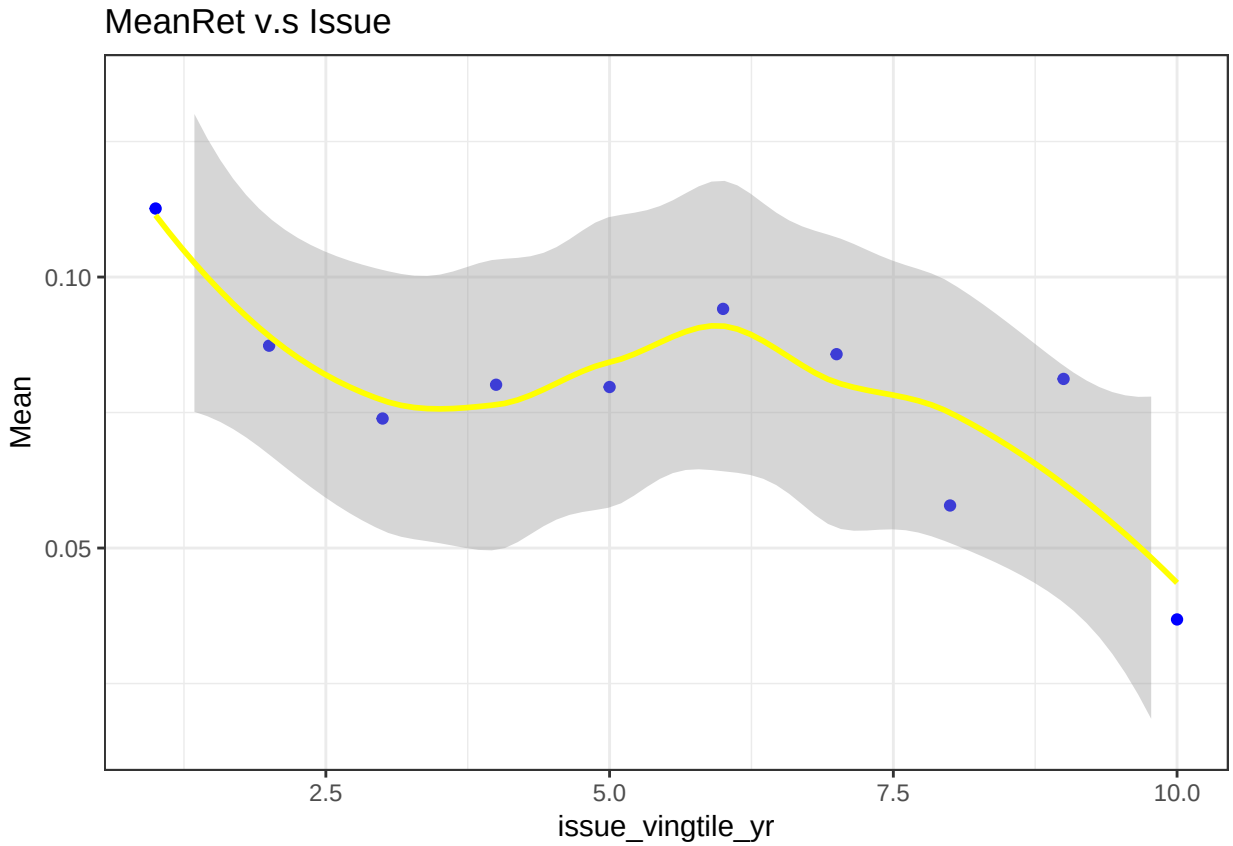
```
Mean3 = StockRetAcct_DT[,list(MeanRet = mean(lnAnnRet)), by = lnIssue_vingtile]
qplot(lnIssue_vingtile,MeanRet, data = Mean3,na.rm = TRUE, col = I("blue"),main = "MeanRet v.s Issue")+g
```

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




```
qplot(issue_vingtile_yr,Mean, data = Mean2,na.rm = TRUE, col = I("blue"),main = "MeanRet v.s Issue",yli
```

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




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.

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 3.4.4
## Loading required package: Formula
## Warning: package 'Formula' was built under R version 3.4.4
##
## Attaching package: 'plm'
## The following object is masked from 'package:data.table':
##
##     between
```

```
temp = c(-1,0,0,0,0,0,0,0,0,1)
StockRetAcct_DT$transformed_issuance = temp[StockRetAcct_DT$issue_vingtile_yr]

lambdas = c()
count = 1

years = unique(StockRetAcct_DT$year)

for(y in years){
  test_data = StockRetAcct_DT[year == y]
```

```
  out = lm(lnAnnRet~transformed_issuance, test_data,na.action = na.omit)
  lambdas[count] = out$coefficients[2]
  count = count + 1
}

lambda = mean(lambdas)
lambda
```

```
## [1] -0.0333771
```

As we can see from the coefficient, which is negative, the nature of portfolio is to long 1 decile and short 10 decile and no postion for others.

## Question 3 : Double sorts and functional forms

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.

We will first set up the StockRetAcct-dataset.

```
rm(list=ls())
setwd("/Users/jiaminghuang/Downloads")

suppressMessages(require(foreign))
suppressWarnings(suppressMessages(require(data.table)))
suppressWarnings(suppressMessages(require(ggplot2)))

sra_dt <- as.data.table(read.dta("StockRetAcct_insample.dta"))

head(sra_dt)
```

```
##    FirmID year   lnAnnRet       lnRf         MEwt     lnIssue       lnMom
## 1:      6 1980  0.3636313 0.07894428 2.814308e-04  0.03134417  0.07535515
## 2:      6 1981 -0.2904088 0.13019902 3.214631e-04  0.04421350  0.51265192
## 3:      6 1982  0.1866300 0.13070259 2.663127e-04 -0.06819496 -0.22050542
## 4:      6 1983  0.4898190 0.08983046 1.699149e-04 -0.07177968  0.04621762
## 5:     10 1991 -0.5080047 0.06121579 3.269729e-05  0.11520413  1.34105313
## 6:     12 2000 -1.3568472 0.06197736 1.219181e-05  0.16523764  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
## 1: 0.08413413  0.6333913      3
## 2: 0.05638131  0.3567226      3
## 3: 0.06207170  0.7794052      3
## 4: 0.07695480  0.7021134      3
## 5: 0.37436786 -2.1609421     10
## 6: 1.06719565 -3.8155227      6
```

```
sra_dt[, excess := exp(lnAnnRet) - exp(lnRf)]
```

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 quintilevariable by year for size.

Create independent quintile sorts based on book-to-market (lnBM) and size (lnME).

```
setorder(sra_dt, year)

# lnBM quintile sort
sra_dt[,bm_quintile := cut(sra_dt$lnBM, breaks = quantile(sra_dt$lnBM, probs = c(0:5)/5, na.rm=TRUE), la

# lnME quintile sort
sra_dt[,me_quintile := cut(sra_dt$lnME, breaks = quantile(sra_dt$lnME, probs = c(0:5)/5, na.rm=TRUE), la

head(sra_dt)
```

```
##    FirmID year    lnAnnRet       lnRf          MEwt     lnIssue
## 1:      6 1980  0.36363131 0.07894428 0.0002814308  0.03134417
## 2:     50 1980  0.16006657 0.07894428 0.0001000092 -0.02015643
## 3:    120 1980 -0.00523882 0.07894428 0.0006449434  0.15793896
## 4:    128 1980  0.15910992 0.07894428 0.0015726771  0.17260452
## 5:    135 1980  0.12482896 0.07894428 0.0003092261  0.05916636
## 6:    143 1980  0.19028714 0.07894428 0.0025189146  0.24339491
##           lnMom     lnME    lnProf        lnEP       lnInv   lnLever
## 1:  0.075355150 12.58147 0.2017671 0.146411210 0.093626112 0.6960014
## 2:  0.306288451 11.54685 0.2938228 0.162321284 0.174244955 0.6668934
## 3: -0.001933197 13.41075 0.1699867 0.157097325 0.045985684 0.9570380
## 4:  0.545399904 14.30212 0.4446782 0.007621429 0.265525669 1.0578818
## 5: -0.297930151 12.67566 0.2358481 0.173495725 0.003995264 0.7851831
## 6:  0.248956278 14.77317 0.2602473 0.139581934 0.125844866 0.7962804
##          lnROE         rv       lnBM ff_ind      excess bm_quintile
## 1: 0.095294215 0.08413413  0.6333913      3  0.35639972           5
## 2: 0.182228252 0.10430538  0.1612592      3  0.09144497           5
## 3: 0.096518815 0.04695285  0.5531039      8 -0.08736915           5
## 4: 0.007153527 0.11101584  0.1333171      5  0.09032279           5
## 5: 0.121454217 0.06324774  0.8440748      3  0.05081064           5
## 6: 0.217954814 0.11710954 -0.3082972     12  0.12745284           4
##    me_quintile
## 1:           2
## 2:           1
## 3:           3
## 4:           4
## 5:           2
## 6:           4
```

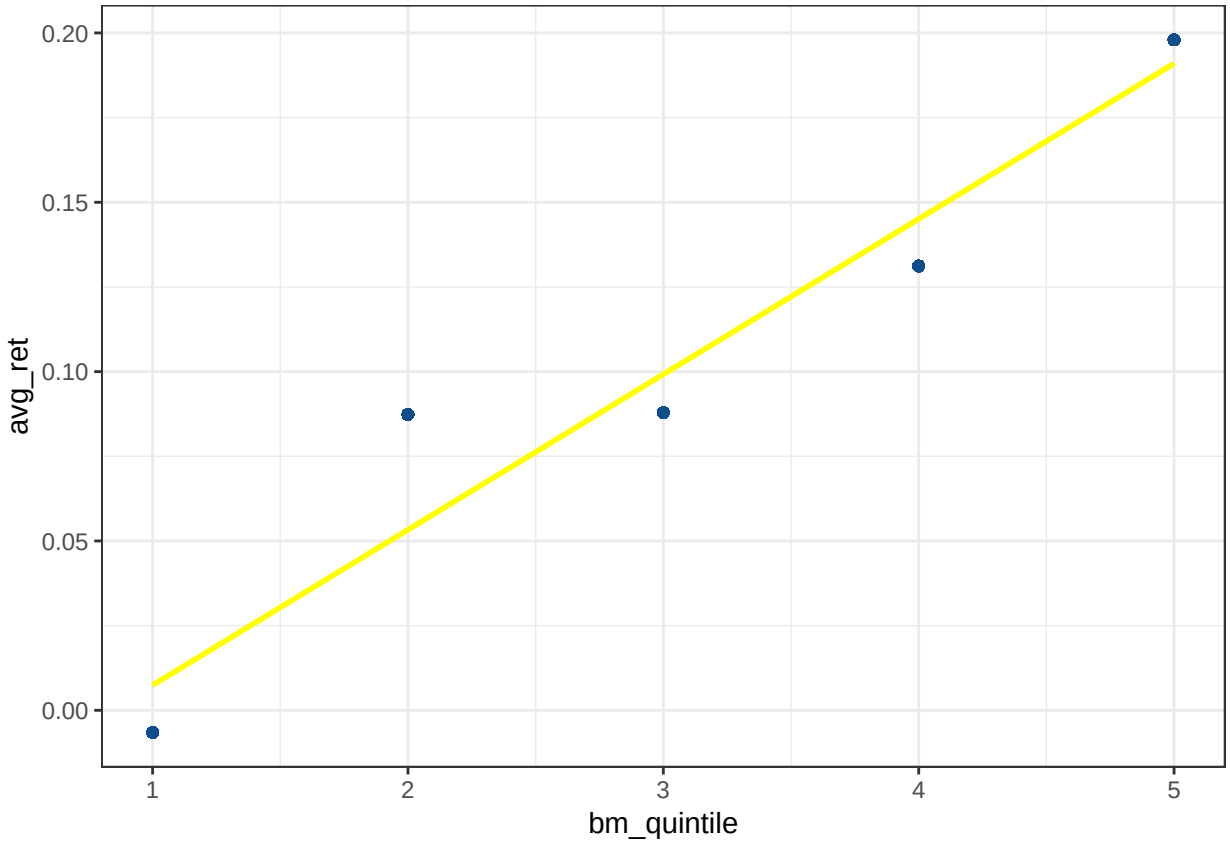
We can see that the quintile variables from lnBM and lnME have been set.

b. For each size quintile, plot the average returns to the five book???to???market quintile portfolios. So, for size quintile 1, and book???to???market quintile 3, the stocks in this portfolio all have size quintile equal to 1 and book???to???market quintile equal to 3. Thus, I'm looking for five plots here, one for each size quintile.

```
p1<-sra_dt[me_quintile==1][,avg_ret:=mean(excess), by=bm_quintile]
p2<-sra_dt[me_quintile==2][,avg_ret:=mean(excess), by=bm_quintile]
p3<-sra_dt[me_quintile==3][,avg_ret:=mean(excess), by=bm_quintile]
```

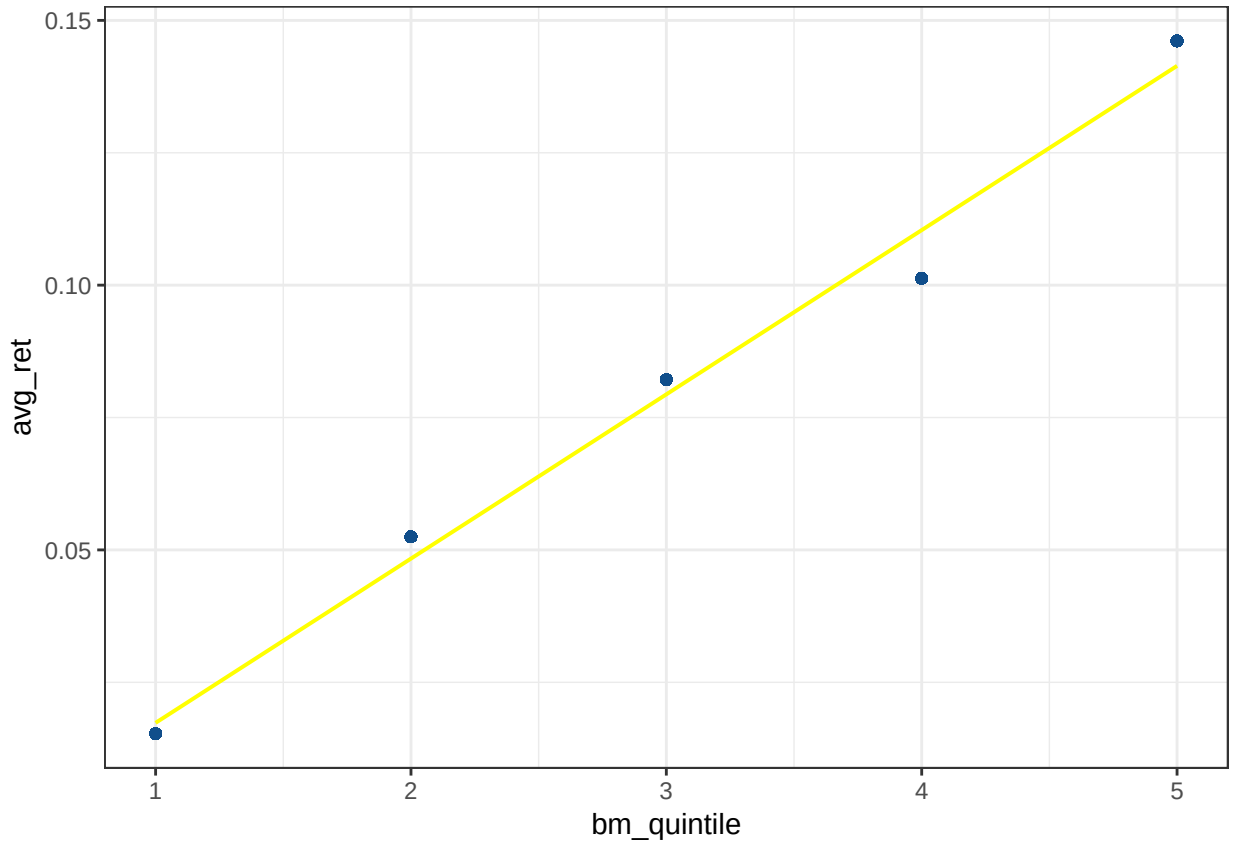
```
p4<-sra_dt[me_quintile==4][,avg_ret:=mean(excess), by=bm_quintile]
p5<-sra_dt[me_quintile==5][,avg_ret:=mean(excess), by=bm_quintile]

qplot(bm_quintile,avg_ret,data = p1, col=I("dodgerblue4"), na.rm = TRUE)+geom_smooth(method = "lm",col=
```

```
## Warning: Removed 725 rows containing non-finite values (stat_smooth).
```



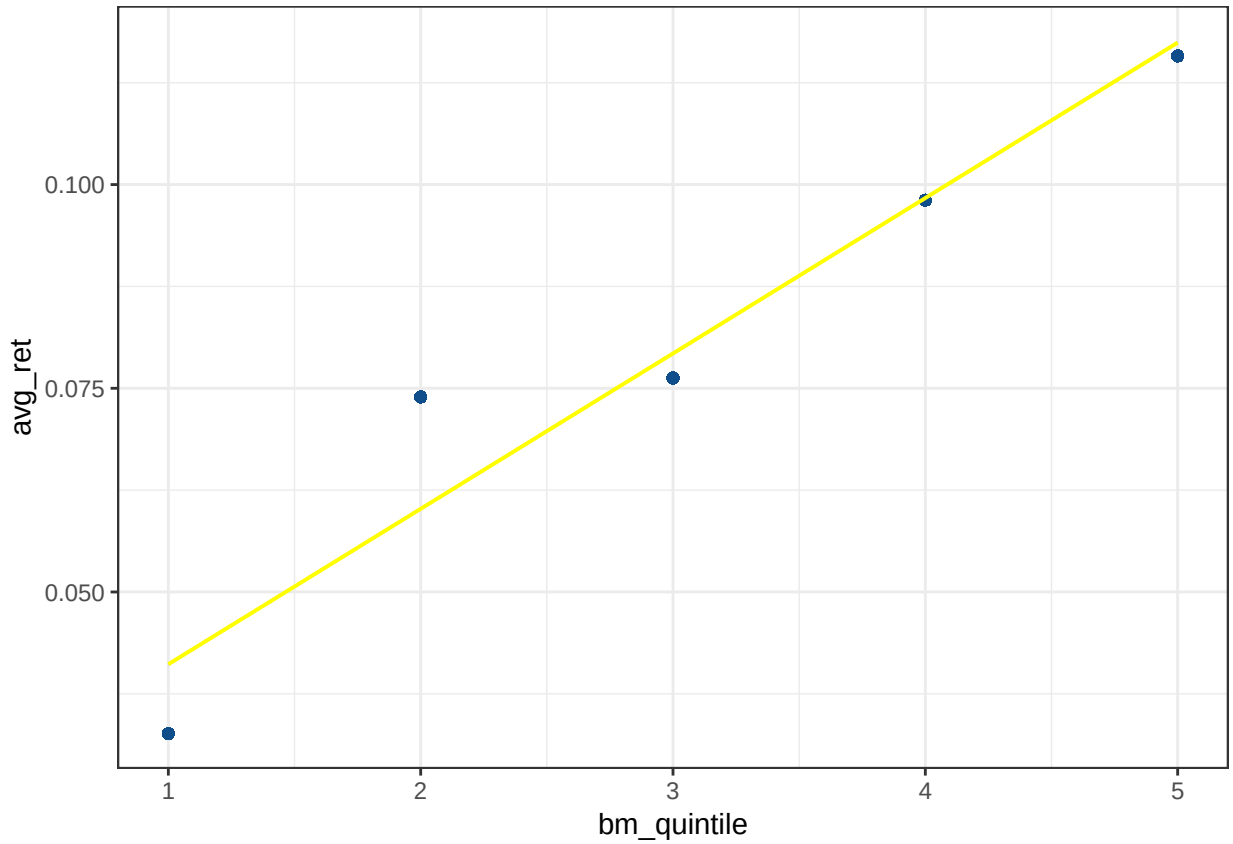

```
qplot(bm_quintile,avg_ret,data = p2, col=I("dodgerblue4"), na.rm = TRUE)+geom_smooth(method = "lm",col=
```

```
## Warning: Removed 769 rows containing non-finite values (stat_smooth).
```



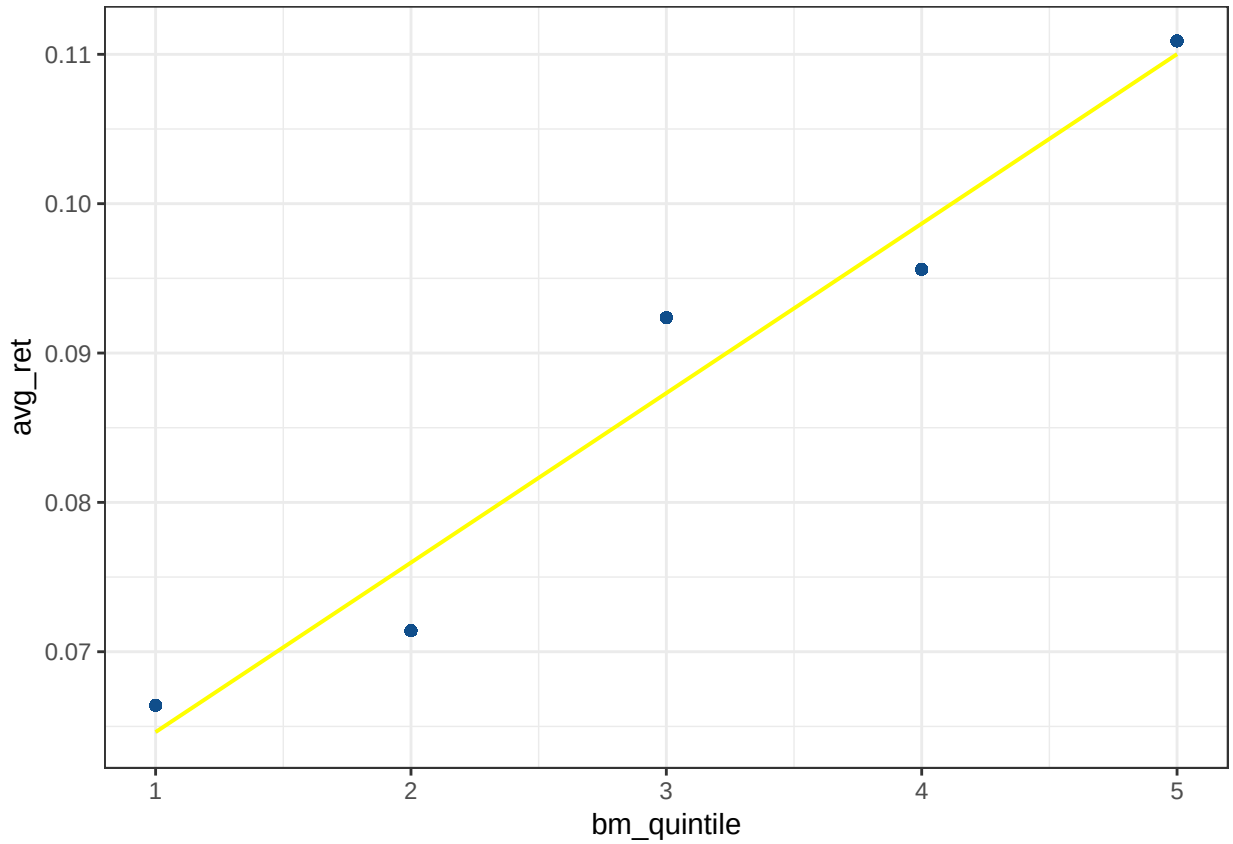

```
qplot(bm_quintile,avg_ret,data = p3, col=I("dodgerblue4"), na.rm = TRUE)+geom_smooth(method = "lm",col=
```

```
## Warning: Removed 701 rows containing non-finite values (stat_smooth).
```



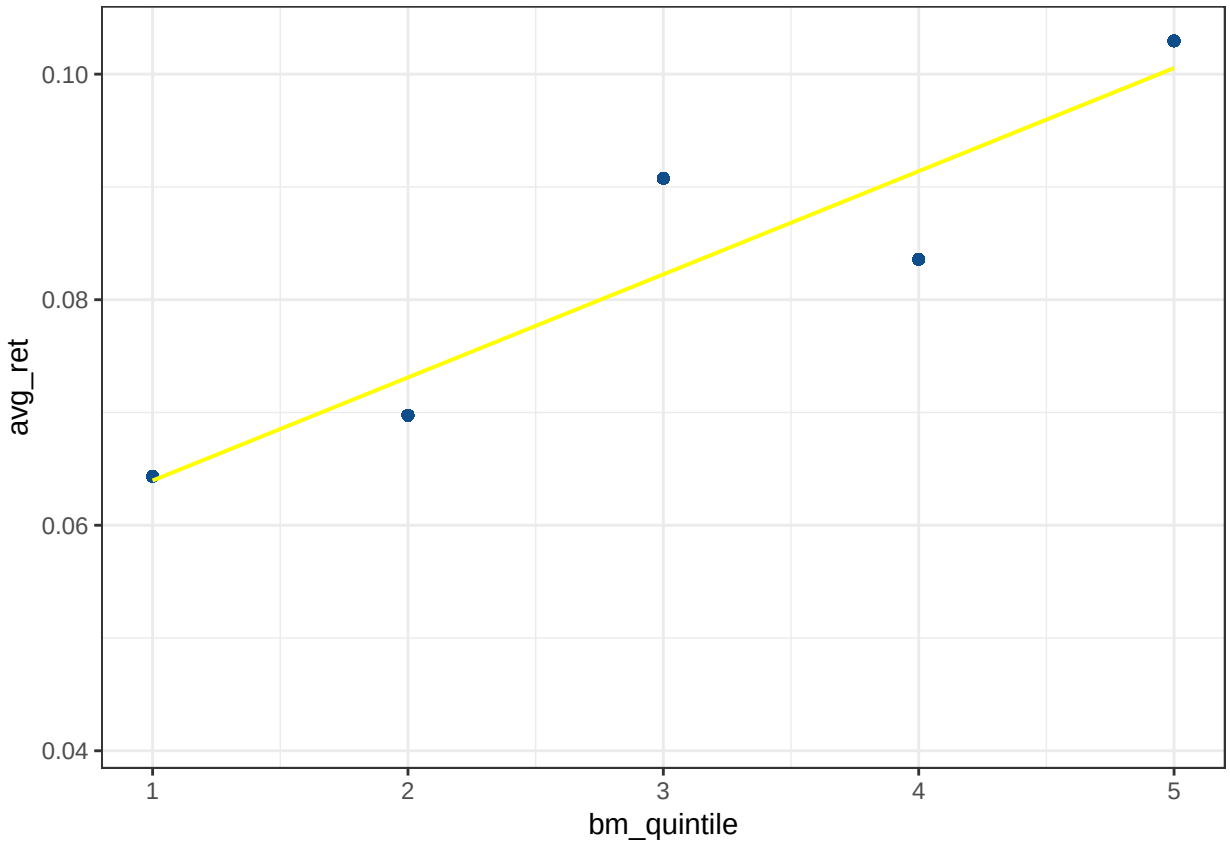

```
qplot(bm_quintile,avg_ret,data = p4, col=I("dodgerblue4"), na.rm = TRUE)+geom_smooth(method = "lm",col=
```

```
## Warning: Removed 515 rows containing non-finite values (stat_smooth).
```




```
qplot(bm_quintile,avg_ret,data = p5, col=I("dodgerblue4"), na.rm = TRUE)+geom_smooth(method = "lm",col=
```

```
## Warning: Removed 328 rows containing non-finite values (stat_smooth).
```




Based on these five plots, using eyeball econometics, the assumption of conditional linearity seems to hold up.