Project\_Education\_data

Frances Lin

5/2019

## I. DESCRIPTION

library(GGally)

## Loading required package: ggplot2

library(ggplot2)  
library(gridExtra)  
library(car)

## Loading required package: carData

# Load data   
ScoreData <- read.csv("https://raw.githubusercontent.com/franceslinyc/Regression-Analysis-of-Education-data-in-R-2019/master/scores.csv")  
str(ScoreData)

## 'data.frame': 3236 obs. of 8 variables:  
## $ school.id: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ gender : Factor w/ 2 levels "boy","girl": 2 2 2 1 1 1 1 1 1 1 ...  
## $ employ : int 9 9 9 2 2 2 2 2 9 9 ...  
## $ t.score : int 23 23 23 15 15 22 22 22 14 14 ...  
## $ ID : int 1 1 1 2 2 3 3 3 4 4 ...  
## $ english : int 72 80 39 7 17 88 89 83 12 25 ...  
## $ math : int 23 24 23 14 11 36 32 39 24 26 ...  
## $ grade : int 3 4 5 3 4 3 4 5 3 4 ...

#View(ScoreData)

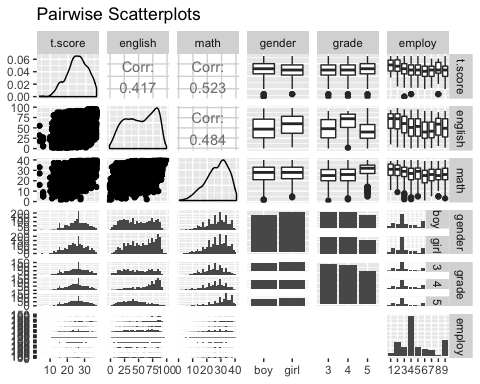
# Change variable(s) to factor  
ScoreData$school.id <- as.factor(ScoreData$school.id)  
ScoreData$employ <- as.factor(ScoreData$employ)  
ScoreData$ID <- as.factor(ScoreData$ID)  
ScoreData$grade <- as.factor(ScoreData$grade)

# Attach data set   
attach(ScoreData)

# Reorder columns   
ScoreData <- ScoreData[c(1,5,4,6,7,2,8,3)]

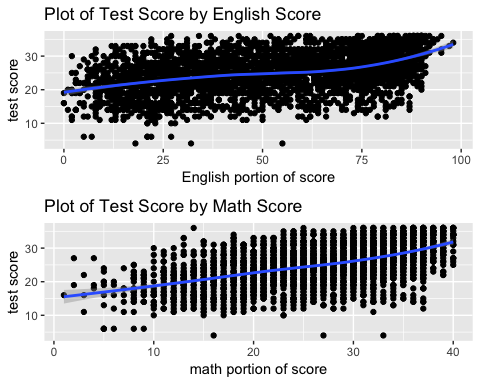
# Pairwise Scatterplots using ggplot2  
ggpairs(ScoreData[c(-1,-2)]) + ggtitle("Pairwise Scatterplots")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

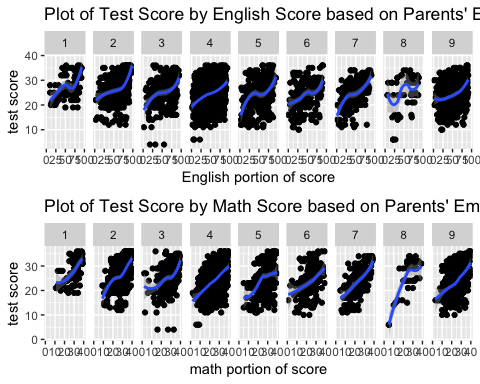


## II. ANALYSIS

# Ques 1a: x=english, y=t.score, based on employ  
g1 <- ggplot(ScoreData, aes(x=english, y=t.score)) +   
 labs(title="Plot of Test Score by English Score", x="English portion of score", y="test score")  
geng0 <- g1 + geom\_point() + geom\_smooth(method="loess")  
geng <- g1 + geom\_point() + geom\_smooth(method="loess") +   
 facet\_grid(.~employ) + ggtitle("Plot of Test Score by English Score based on Parents' Employment")  
  
# Ques 1b: x=math, y=t.score, based on employ  
g11 <- ggplot(ScoreData, aes(x=math, y=t.score)) +   
 labs(title="Plot of Test Score by Math Score", x="math portion of score", y="test score")  
gmath0 <- g11 + geom\_point() + geom\_smooth(method="loess")  
gmath <- g11 + geom\_point() + geom\_smooth(method="loess") +   
 facet\_grid(.~employ) + ggtitle("Plot of Test Score by Math Score based on Parents' Employment")  
  
# Combine Plots of Test Score by English Score and by Math Score  
grid.arrange(geng0, gmath0, nrow=2)



grid.arrange(geng, gmath, nrow=2)



# Fit simple linear regression model for Ques 1  
lmeng <- lm(t.score~english)  
lmmath <- lm(t.score~math)  
summary(lmeng, type=3)

##   
## Call:  
## lm(formula = t.score ~ english)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.3717 -3.4553 0.2956 3.8827 13.3744   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19.977539 0.217829 91.71 <2e-16 \*\*\*  
## english 0.098075 0.003755 26.12 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.277 on 3234 degrees of freedom  
## Multiple R-squared: 0.1742, Adjusted R-squared: 0.174   
## F-statistic: 682.3 on 1 and 3234 DF, p-value: < 2.2e-16

summary(lmmath, type=3)

##   
## Call:  
## lm(formula = t.score ~ math)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.6481 -3.2603 0.3316 3.5458 15.9133   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.51510 0.31592 45.95 <2e-16 \*\*\*  
## math 0.39797 0.01139 34.94 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.948 on 3234 degrees of freedom  
## Multiple R-squared: 0.274, Adjusted R-squared: 0.2738   
## F-statistic: 1221 on 1 and 3234 DF, p-value: < 2.2e-16

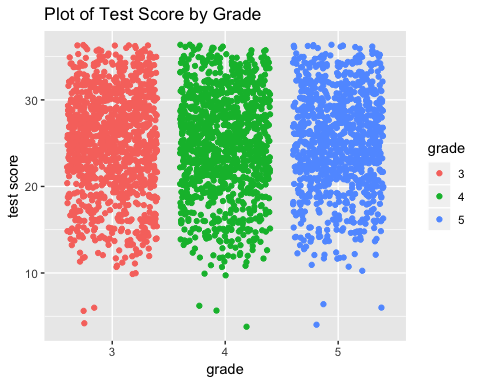
lmengc <- lm(t.score~english+employ)  
lmmathc <- lm(t.score~math+employ)  
Anova(lmengc, type=3)

## Anova Table (Type III tests)  
##   
## Response: t.score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 40089 1 1462.3065 < 2.2e-16 \*\*\*  
## english 15772 1 575.3048 < 2.2e-16 \*\*\*  
## employ 1624 8 7.4054 8.154e-10 \*\*\*  
## Residuals 88441 3226   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

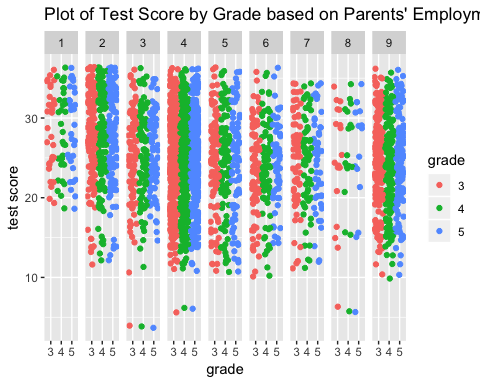
Anova(lmmathc, type=3)

## Anova Table (Type III tests)  
##   
## Response: t.score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 18683 1 775.4661 < 2.2e-16 \*\*\*  
## math 26490 1 1099.5289 < 2.2e-16 \*\*\*  
## employ 1459 8 7.5721 4.514e-10 \*\*\*  
## Residuals 77722 3226   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

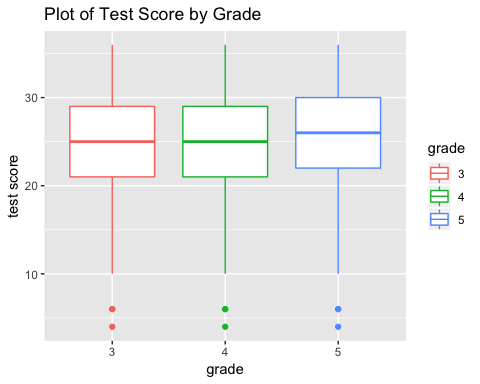
# Ques 2: x=grade, y=t.score, based on employ  
g2 <- ggplot(ScoreData, aes(x=grade, y=t.score, color=grade)) +   
 labs(title="Plot of Test Score by Grade", y="test score")  
g2 + geom\_jitter()



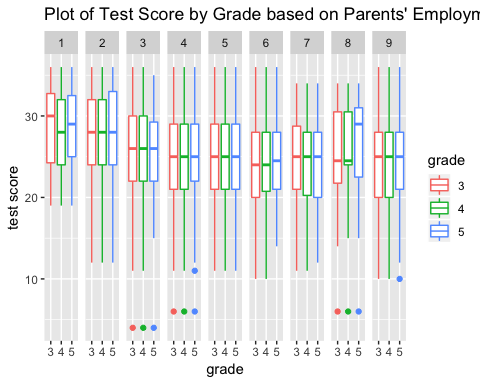
g2 + geom\_jitter() + facet\_grid(.~employ) +   
 ggtitle("Plot of Test Score by Grade based on Parents' Employment")



g2 + geom\_boxplot()



g2 + geom\_boxplot() + facet\_grid(.~employ) +  
 ggtitle("Plot of Test Score by Grade based on Parents' Employment")



# Fit simple linear regression model for Ques 2  
lmgrade <- lm(t.score~grade)  
lmgradec <- lm(t.score~grade+employ)  
Anova(lmgrade, type=3)

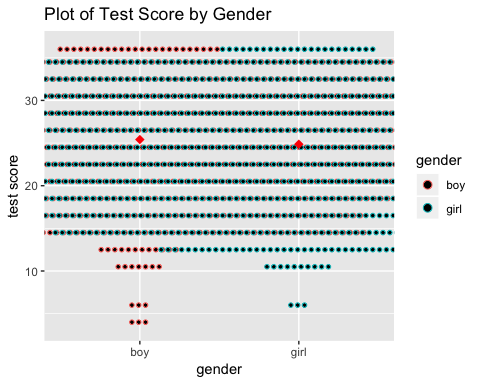
## Anova Table (Type III tests)  
##   
## Response: t.score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 723652 1 21461.9483 <2e-16 \*\*\*  
## grade 58 2 0.8573 0.4244   
## Residuals 109010 3233   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova(lmgradec, type=3)

## Anova Table (Type III tests)  
##   
## Response: t.score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 71747 1 2221.2322 <2e-16 \*\*\*  
## grade 44 2 0.6855 0.5039   
## employ 4841 8 18.7357 <2e-16 \*\*\*  
## Residuals 104169 3225   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

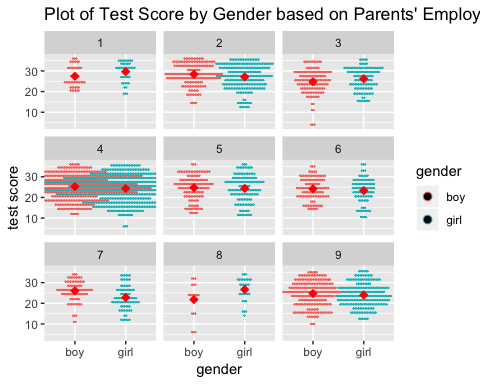
# Ques 3: x=gender, y=t.score, based on employ  
g3 <- ggplot(ScoreData, aes(x=gender, y=t.score, color=gender)) +   
 labs(title="Plot of Test Score by Gender", y="test score")  
g3 + geom\_dotplot(binaxis='y', stackdir='center', stackratio=1.5, dotsize=0.5) +   
 stat\_summary(fun.y=mean, geom="point", shape=18, size=3, color="red")

## `stat\_bindot()` using `bins = 30`. Pick better value with `binwidth`.



g3 + geom\_dotplot(binaxis='y', stackdir='center', stackratio=1.5, dotsize=0.5) +   
 stat\_summary(fun.y=mean, geom="point", shape=18, size=3, color="red") +   
 facet\_wrap(~employ) +   
 ggtitle("Plot of Test Score by Gender based on Parents' Employment")

## `stat\_bindot()` using `bins = 30`. Pick better value with `binwidth`.



# Fit simple linear regression model for Ques 3  
lmgender <- lm(t.score~gender)  
lmgenderc <- lm(t.score~gender+employ)  
Anova(lmgender, type=3)

## Anova Table (Type III tests)  
##   
## Response: t.score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 1001537 1 29763.1545 < 2.2e-16 \*\*\*  
## gender 243 1 7.2189 0.007251 \*\*   
## Residuals 108825 3234   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova(lmgenderc, type=3)

## Anova Table (Type III tests)  
##   
## Response: t.score  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 75936 1 2357.852 < 2.2e-16 \*\*\*  
## gender 317 1 9.853 0.001711 \*\*   
## employ 4929 8 19.132 < 2.2e-16 \*\*\*  
## Residuals 103896 3226   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## III. RESULTS

## IV. R CODE