Abir Pattnaik MBA Starting Salaries Code

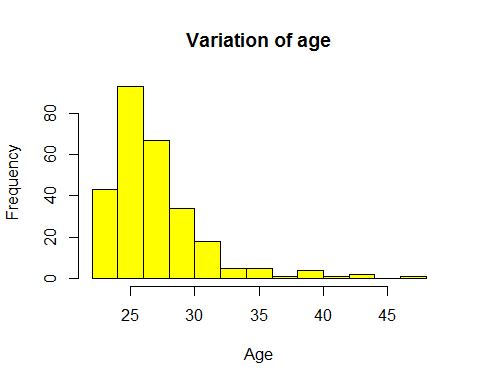
College:MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY

Phone No.8586896169

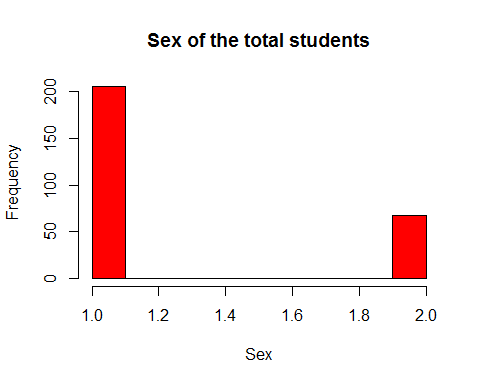
# Analysis of MBA Starting Salaries  
# NAME: ABIR PATTNAIK  
# EMAIL: abir.pattanaik@gmail.com  
# COLLEGE : Maharaja Agrasen Institute of Technology  
  
# NOTE:2 datasets have been used  
# Dataset1:StartingSalary.csv  
# Dataset2:StartingSalary\_job.csv  
  
# a) Load the dataset into RStudio and Viewing the dataframe for checking whether the   
# data set uploaded is correct or not.  
#  
# Choose your own location of where you have saved the dataset StartingSalary.csv  
  
StartingSalary.df<-read.csv("C:/Users/DRDO HQ/Desktop/DATA ANALYTICS INTERNSHIP/MBA SALARIES PROJECT/StartingSalary.csv")  
View(StartingSalary.df)  
  
# a1) Viewing the summary for the first 11 columns as the last 2 columns contain coded symbols which affect the data   
  
summary(StartingSalary.df[1:11])

## age sex gmat\_tot gmat\_qpc   
## Min. :22.00 Min. :1.000 Min. :450.0 Min. :28.00   
## 1st Qu.:25.00 1st Qu.:1.000 1st Qu.:580.0 1st Qu.:72.00   
## Median :27.00 Median :1.000 Median :620.0 Median :83.00   
## Mean :27.36 Mean :1.248 Mean :619.5 Mean :80.64   
## 3rd Qu.:29.00 3rd Qu.:1.000 3rd Qu.:660.0 3rd Qu.:93.00   
## Max. :48.00 Max. :2.000 Max. :790.0 Max. :99.00   
## gmat\_vpc gmat\_tpc s\_avg f\_avg   
## Min. :16.00 Min. : 0.0 Min. :2.000 Min. :0.000   
## 1st Qu.:71.00 1st Qu.:78.0 1st Qu.:2.708 1st Qu.:2.750   
## Median :81.00 Median :87.0 Median :3.000 Median :3.000   
## Mean :78.32 Mean :84.2 Mean :3.025 Mean :3.062   
## 3rd Qu.:91.00 3rd Qu.:94.0 3rd Qu.:3.300 3rd Qu.:3.250   
## Max. :99.00 Max. :99.0 Max. :4.000 Max. :4.000   
## quarter work\_yrs frstlang   
## Min. :1.000 Min. : 0.000 Min. :1.000   
## 1st Qu.:1.250 1st Qu.: 2.000 1st Qu.:1.000   
## Median :2.000 Median : 3.000 Median :1.000   
## Mean :2.478 Mean : 3.872 Mean :1.117   
## 3rd Qu.:3.000 3rd Qu.: 4.000 3rd Qu.:1.000   
## Max. :4.000 Max. :22.000 Max. :2.000

# This would ensure that data that has been collected is for both placed and unplaced students  
# All these data are collected before and is not required for people to fill data  
  
StartingSalary\_TotalStudents.df<-StartingSalary.df[1:11]  
  
# b) Visualising each variable and understanding the graph  
# Age has a positive skewed distribution with median at age 27 and mean at 27.36  
# Table for Male and Female shows that the Male:Female ratio is high  
# A normal distribution curve was made with respect to students count with mean at 619.5  
# and going high as 790.0   
# A positively distribution curve was established with most of the work experiences being with range of 0-5 yrs exp.  
  
attach(StartingSalary\_TotalStudents.df)  
hist(age,col="yellow",main="Variation of age",xlab="Age")



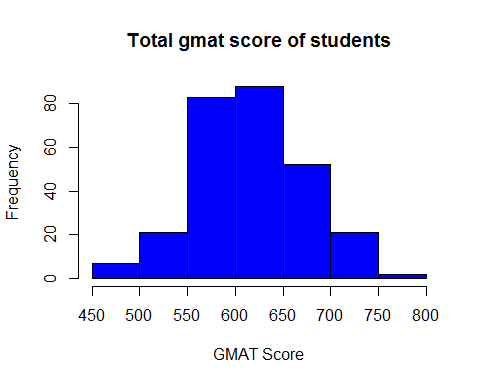
hist(sex,col="red",main="Sex of the total students",xlab="Sex")



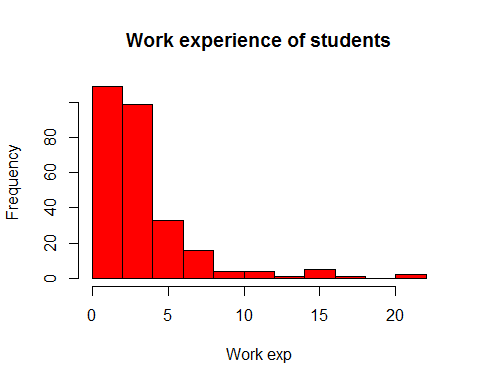
addmargins(table(sex))

## sex  
## 1 2 Sum   
## 206 68 274

hist(gmat\_tot,col="blue",main="Total gmat score of students",xlab="GMAT Score")



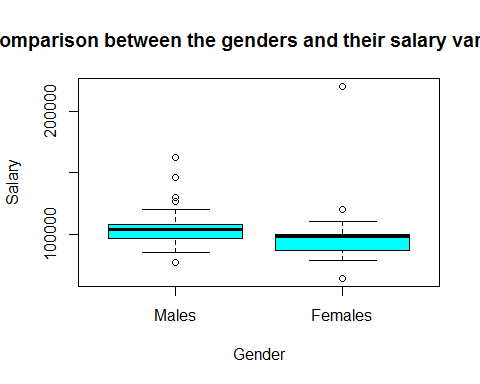
hist(work\_yrs,col="red",main="Work experience of students",xlab="Work exp")



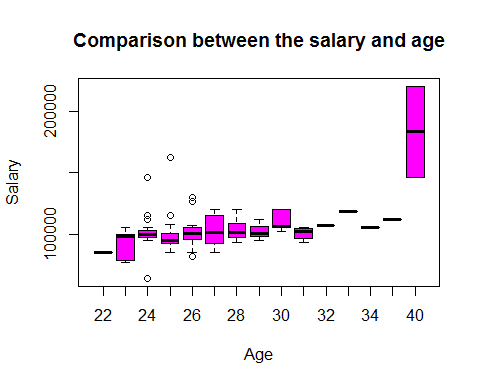
addmargins(table(frstlang))

## frstlang  
## 1 2 Sum   
## 242 32 274

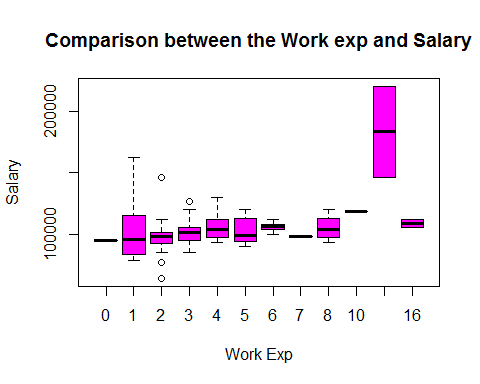
detach(StartingSalary\_TotalStudents.df)  
  
# c) Segregating the databases into 2 databases  
# PlacedStudents--Students who were placed and mentioned their salary  
# NotPlacedStudents--Students who weren't placed but took part in the survey  
# PlacedStudents--Students who were placed but some of them didn't disclose their salaries  
  
PlacedStudents.df<-StartingSalary.df[which(StartingSalary.df$salary!=999&StartingSalary.df$salary!=998&StartingSalary.df$salary!=0),]  
View(PlacedStudents.df)  
NotPlacedStudents.df<-StartingSalary.df[which(StartingSalary.df$salary!=999&StartingSalary.df$salary!=998&StartingSalary.df$salary==0),]  
View(NotPlacedStudents.df)  
PlacedStudents\_999.df<-StartingSalary.df[which(StartingSalary.df$salary!=998&StartingSalary.df$salary!=0),]  
View(PlacedStudents\_999.df)  
  
# d) Creating boxplots to understand the some basic relations of salaries and other factors  
# such as Gender or age or ranking in their college  
# Inferring from the boxplot it is clear that the male salary is higher than female salary   
# although it the mean is approximately same but it might not be true as there is an outlier in the   
# female salary that can affect the mean  
# On the basis of age it is clear that with age the salaries do increase.Further analysis should be done in this area  
# There is a fairly linear relationship in the comparison between the Salary and work exp  
# A negative linear relationship was established between the percentile and Salary   
  
attach(PlacedStudents.df)  
boxplot(salary~sex,main="Comparison between the genders and their salary variance",xlab="Gender",ylab="Salary",xaxt="n",col="cyan")  
axis(side=1, at=c(1,2), labels=c("Males", "Females"))



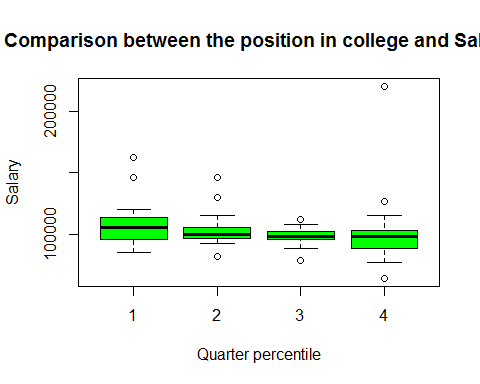
boxplot(salary~age,main="Comparison between the salary and age",xlab="Age",ylab="Salary",col="magenta")



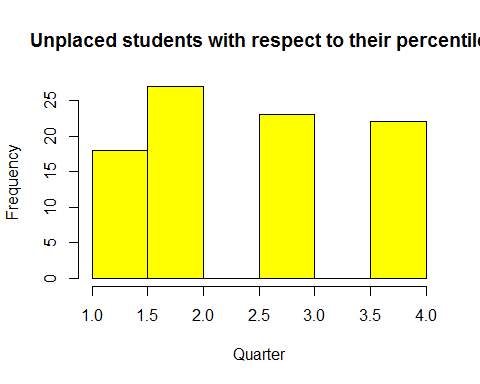
boxplot(salary~work\_yrs,main="Comparison between the Work exp and Salary",xlab="Work Exp",ylab="Salary",col="magenta")



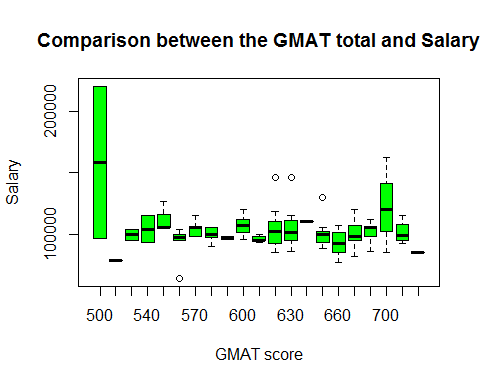
boxplot(salary~quarter,main="Comparison between the position in college and Salary",xlab="Quarter percentile",ylab="Salary",col="green")



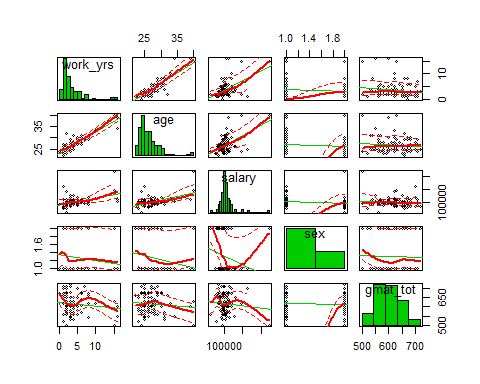
hist(NotPlacedStudents.df$quarter,main="Unplaced students with respect to their percentile",xlab="Quarter",col="yellow")



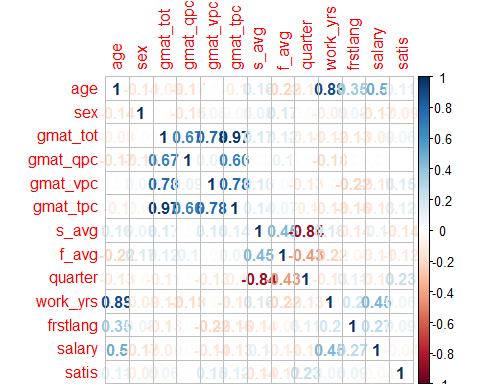
boxplot(salary~gmat\_tot,main="Comparison between the GMAT total and Salary",xlab="GMAT score",ylab="Salary",col="green")  
detach(PlacedStudents.df)  
  
# d) Creating a scatterplot matrix and understanding the pattern it makes with salary  
# 'car' package was used  
# The relationship between salary and age are linear relationship  
# Same is for the work\_exp and salary  
# However due to strong linear relationship between work\_exp and age chances are correlated  
  
  
library(car)



scatterplotMatrix(formula=~work\_yrs+age+salary+sex+gmat\_tot,cex=0.6,diagonal="histogram",data=PlacedStudents.df)



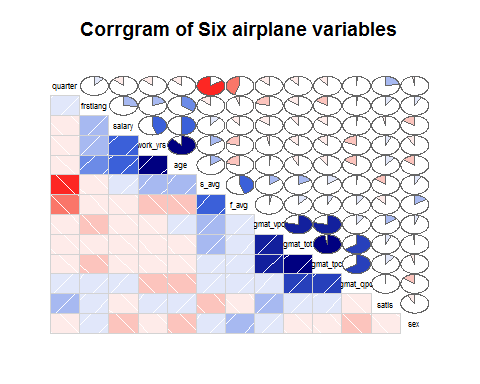
# e) For checking correlation corrplot was used  
# 'corrplot' package was used  
# 'corrgram' package was used  
# corrplot provides some very interesting insights of the correlation  
# work\_yrs and age have value of 0.88 that is almost close to 1  
# It is evident that gmat\_tot and gmat\_tpc would be correlated  
# Quarter percentile has a negative correlation of 0.84  
# corrgram depicts the same plots to understand correlationship better in   
  
library(corrplot)  
corrplot(cor(PlacedStudents.df),method="number")



library(corrgram)

## Warning: package 'corrgram' was built under R version 3.4.1

corrgram(PlacedStudents.df, main = "Corrgram of Six airplane variables", lower.panel = panel.shade, upper.panel = panel.pie, text.panel = panel.txt,order=TRUE)



# d) Creating a linear model using all factors   
# Three models are created.Details are given as following:-  
# i)lm.fit\_full:This model has all factors for the model  
# Residuals are high and Residual Standard error is too high(15430)  
# r-squared:0.3422 and adjusted r-squared:0.2545  
# ii)lm.fit:This model has factors as frstlang+age+sex+work\_yrs+work\_yrs:gmat\_tot  
# Residuals are high and similarly for RSE(15570)  
# r-squared:0.2702 and adjusted r-squared:0.2409  
# iii)lm.fit2:This factors considers only age as a element  
# Residuals are high and so is the RSE(15550)  
# r-squared:0.2496 and adjusted r-squared:0.2422   
# vif stands for variance inflation factor  
# 'visreg' to realise visualisations of regression made  
# anova was used as a comparison method  
  
lm.fit\_full=lm(salary~.,data=PlacedStudents.df)  
lm.fit=lm(salary~work\_yrs+frstlang+sex+work\_yrs:gmat\_tot,data=PlacedStudents.df)  
lm.fit2=lm(salary~age,data=PlacedStudents.df)  
summary(lm.fit\_full)

##   
## Call:  
## lm(formula = salary ~ ., data = PlacedStudents.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26489 -7983 -373 5923 70602   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 78005.66 52981.93 1.472 0.1444   
## age 1750.65 1130.92 1.548 0.1251   
## sex -3584.07 3595.85 -0.997 0.3216   
## gmat\_tot 16.19 178.85 0.090 0.9281   
## gmat\_qpc 796.55 496.78 1.603 0.1123   
## gmat\_vpc 546.31 501.97 1.088 0.2794   
## gmat\_tpc -1457.09 714.94 -2.038 0.0445 \*  
## s\_avg -931.53 8240.31 -0.113 0.9102   
## f\_avg -2222.82 3894.57 -0.571 0.5696   
## quarter -2336.56 2721.89 -0.858 0.3929   
## work\_yrs 749.66 1135.90 0.660 0.5110   
## frstlang 7719.42 7373.27 1.047 0.2979   
## satis -1086.54 2157.76 -0.504 0.6158   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15430 on 90 degrees of freedom  
## Multiple R-squared: 0.3422, Adjusted R-squared: 0.2545   
## F-statistic: 3.902 on 12 and 90 DF, p-value: 8.086e-05

summary(lm.fit)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + frstlang + sex + work\_yrs:gmat\_tot,   
## data = PlacedStudents.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39092 -7251 -2329 5540 64709   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92169.233 7981.077 11.548 < 2e-16 \*\*\*  
## work\_yrs 11131.813 3646.113 3.053 0.00292 \*\*   
## frstlang 10457.320 6287.385 1.663 0.09946 .   
## sex -6082.002 3307.728 -1.839 0.06898 .   
## work\_yrs:gmat\_tot -14.836 6.125 -2.422 0.01726 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15230 on 98 degrees of freedom  
## Multiple R-squared: 0.3018, Adjusted R-squared: 0.2733   
## F-statistic: 10.59 on 4 and 98 DF, p-value: 3.577e-07

summary(lm.fit2)

##   
## Call:  
## lm(formula = salary ~ age, data = PlacedStudents.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31454 -8533 -2182 4546 80886   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 29962.6 12697.8 2.360 0.0202 \*   
## age 2728.8 470.7 5.797 7.75e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15550 on 101 degrees of freedom  
## Multiple R-squared: 0.2496, Adjusted R-squared: 0.2422   
## F-statistic: 33.6 on 1 and 101 DF, p-value: 7.748e-08

anova(lm.fit,lm.fit\_full)

## Analysis of Variance Table  
##   
## Model 1: salary ~ work\_yrs + frstlang + sex + work\_yrs:gmat\_tot  
## Model 2: salary ~ age + sex + gmat\_tot + gmat\_qpc + gmat\_vpc + gmat\_tpc +   
## s\_avg + f\_avg + quarter + work\_yrs + frstlang + satis  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 98 2.2739e+10   
## 2 90 2.1423e+10 8 1.316e+09 0.6911 0.6984

vif(lm.fit)

## work\_yrs frstlang sex work\_yrs:gmat\_tot   
## 52.964329 1.111534 1.021803 52.359471

par(mfrow=c(2,2))  
library(visreg)

## Warning: package 'visreg' was built under R version 3.4.1

visreg(lm.fit)

## Warning: Note that you are attempting to plot a 'main effect' in a model that contains an  
## interaction. This is potentially misleading; you may wish to consider using the 'by'  
## argument.

## Conditions used in construction of plot  
## frstlang: 1  
## sex: 1  
## gmat\_tot: 620

## Warning: Note that you are attempting to plot a 'main effect' in a model that contains an  
## interaction. This is potentially misleading; you may wish to consider using the 'by'  
## argument.

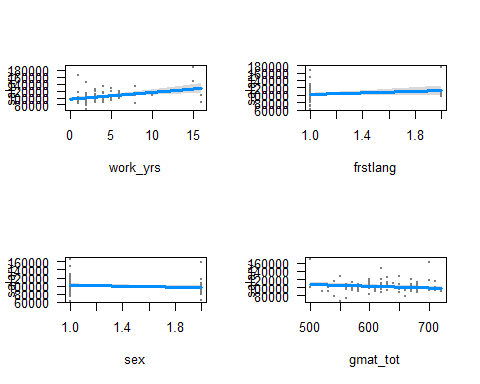
## Conditions used in construction of plot  
## work\_yrs: 3  
## sex: 1  
## gmat\_tot: 620

## Warning: Note that you are attempting to plot a 'main effect' in a model that contains an  
## interaction. This is potentially misleading; you may wish to consider using the 'by'  
## argument.

## Conditions used in construction of plot  
## work\_yrs: 3  
## frstlang: 1  
## gmat\_tot: 620

## Warning: Note that you are attempting to plot a 'main effect' in a model that contains an  
## interaction. This is potentially misleading; you may wish to consider using the 'by'  
## argument.

## Conditions used in construction of plot  
## work\_yrs: 3  
## frstlang: 1  
## sex: 1



# e) Linear model for people having job and satisfaction from college  
# i)lm.fit3:For Y as a job parameter it is evident that salary would be a factor and so is satisfaction from college   
# r^2=95.29,adjusted r^2=95.23  
# RSE=0.106  
# ii)lm.fit4:Not considering salary as one of the factors  
# r^2=0.1898 and adjusted r^2=0.176  
# RSE=0.4407  
  
  
StartingSalary\_job.df<-read.csv("C:/Users/DRDO HQ/Desktop/DATA ANALYTICS INTERNSHIP/MBA SALARIES PROJECT/StartingSalary\_job.csv")  
View(StartingSalary\_job.df)  
PlacedStudents\_job.df<-StartingSalary\_job.df[which(StartingSalary\_job.df$salary!=999),]  
View(PlacedStudents\_job.df)  
  
corrplot(cor(PlacedStudents\_job.df),method="number")  
lm.fit3\_full=lm(job~.,data=PlacedStudents\_job.df)  
lm.fit3=lm(job~salary+work\_yrs+satis,data=PlacedStudents\_job.df)  
lm.fit4=lm(job~age+gmat\_tot+work\_yrs+satis,data=PlacedStudents\_job.df)  
summary(lm.fit3)

##   
## Call:  
## lm(formula = job ~ salary + work\_yrs + satis, data = PlacedStudents\_job.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.93045 -0.04375 -0.00697 0.04162 0.36339   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.728e-02 1.414e-02 4.757 3.43e-06 \*\*\*  
## salary 9.086e-06 1.441e-07 63.035 < 2e-16 \*\*\*  
## work\_yrs -9.435e-03 2.044e-03 -4.615 6.46e-06 \*\*\*  
## satis 9.547e-04 1.936e-05 49.304 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.106 on 235 degrees of freedom  
## Multiple R-squared: 0.9529, Adjusted R-squared: 0.9523   
## F-statistic: 1585 on 3 and 235 DF, p-value: < 2.2e-16

summary(lm.fit3\_full)

##   
## Call:  
## lm(formula = job ~ ., data = PlacedStudents\_job.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.88664 -0.04446 -0.00105 0.04848 0.30486   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.602e-01 1.820e-01 0.880 0.3798   
## age -7.877e-03 3.761e-03 -2.094 0.0374 \*   
## sex 2.418e-02 1.590e-02 1.521 0.1297   
## gmat\_tot 1.944e-04 5.354e-04 0.363 0.7169   
## gmat\_qpc -1.842e-03 1.487e-03 -1.238 0.2168   
## gmat\_vpc -8.897e-04 1.419e-03 -0.627 0.5313   
## gmat\_tpc 1.780e-03 1.252e-03 1.421 0.1566   
## s\_avg -9.778e-03 2.967e-02 -0.330 0.7420   
## f\_avg 2.214e-02 1.494e-02 1.482 0.1397   
## quarter 9.197e-03 9.356e-03 0.983 0.3267   
## work\_yrs -1.448e-03 4.213e-03 -0.344 0.7314   
## frstlang -4.662e-02 2.533e-02 -1.840 0.0670 .   
## salary 9.053e-06 1.458e-07 62.111 <2e-16 \*\*\*  
## satis 9.552e-04 1.954e-05 48.885 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1038 on 225 degrees of freedom  
## Multiple R-squared: 0.9568, Adjusted R-squared: 0.9543   
## F-statistic: 382.9 on 13 and 225 DF, p-value: < 2.2e-16

summary(lm.fit4)

##   
## Call:  
## lm(formula = job ~ age + gmat\_tot + work\_yrs + satis, data = PlacedStudents\_job.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7000 -0.4839 0.0183 0.4098 0.7354   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.697e+00 4.730e-01 3.587 0.000407 \*\*\*  
## age -4.662e-02 1.509e-02 -3.090 0.002247 \*\*   
## gmat\_tot -5.712e-06 5.101e-04 -0.011 0.991075   
## work\_yrs 3.020e-02 1.734e-02 1.742 0.082831 .   
## satis 4.411e-04 7.378e-05 5.978 8.36e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4407 on 234 degrees of freedom  
## Multiple R-squared: 0.1898, Adjusted R-squared: 0.176   
## F-statistic: 13.71 on 4 and 234 DF, p-value: 4.655e-10

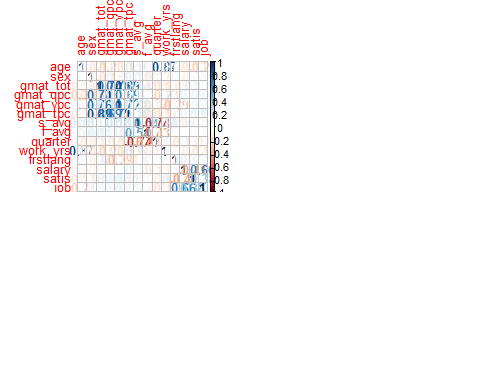
anova(lm.fit3,lm.fit3\_full)

## Analysis of Variance Table  
##   
## Model 1: job ~ salary + work\_yrs + satis  
## Model 2: job ~ age + sex + gmat\_tot + gmat\_qpc + gmat\_vpc + gmat\_tpc +   
## s\_avg + f\_avg + quarter + work\_yrs + frstlang + salary +   
## satis  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 235 2.6420   
## 2 225 2.4263 10 0.21573 2.0006 0.03433 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

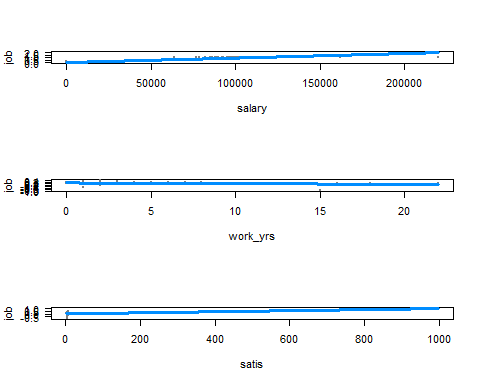
vif(lm.fit3)

## salary work\_yrs satis   
## 1.202599 1.017027 1.219718

par(mfrow=c(3,1))



visreg(lm.fit3)



# f) Perfroming t-tests on the following hypothesis  
# T test between Salary and Gender  
# Chi-square test between Work\_exp and Age  
# Dataset used:PlacedStudents.df  
  
attach(PlacedStudents.df)  
aggregate(salary,by=list(Gender=sex),FUN=mean)

## Gender x  
## 1 1 104970.97  
## 2 2 98524.39

t.test(salary~sex,alternative="greater")

##   
## Welch Two Sample t-test  
##   
## data: salary by sex  
## t = 1.3628, df = 38.115, p-value = 0.09047  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -1527.96 Inf  
## sample estimates:  
## mean in group 1 mean in group 2   
## 104970.97 98524.39

# Hypothesis is accepted and Male have greater salary than woman p>0.05 and it is within the confidence interval  
# For the chi-square test between age and work years these are dependent as x-squared is quite high  
  
mytable1 <- xtabs(age~work\_yrs, data=PlacedStudents.df)  
chisq.test(mytable1)

##   
## Chi-squared test for given probabilities  
##   
## data: mytable1  
## X-squared = 3496.5, df = 11, p-value < 2.2e-16

detach(PlacedStudents.df)  
  
attach(PlacedStudents\_job.df)  
mytable2 <- xtabs(job~frstlang, data=PlacedStudents\_job.df)  
mytable2

## frstlang  
## 1 2   
## 134 15

mytable3 <- xtabs(job~quarter, data=PlacedStudents\_job.df)  
mytable3

## quarter  
## 1 2 3 4   
## 46 36 39 28

detach(PlacedStudents\_job.df)