**6304 Wages Data Live Lecture**

**R Script File**

**rm(list=ls())**

**library(rio)**

**#Read in data.**

**wages=import("Our Wages Data.xlsx",which="Our Data")**

**colnames(wages)=tolower(make.names(colnames(wages)))**

**attach(wages)**

**#Copy the continuous variables to a new data object.**

**some.of.wages=subset(wages,select=c("wage","yearsed",**

**"experience","age"))**

**#Correlation analysis of the continuous variables.**

**plot(some.of.wages,main="Some of Everything with**

**Some of Everything")**

**cor(some.of.wages)**

**round(cor(some.of.wages),3)**

**#First put a correlation matrix into an object.**

**library(corrplot)**

**xx=cor(some.of.wages)**

**corrplot(xx,method="circle")**

**corrplot(xx,method="pie")**

**corrplot(xx,method="ellipse")**

**corrplot(xx,method="color")**

**corrplot(xx,method="number")**

**corrplot(xx,method="square")**

**corrplot(xx,method="circle",type="upper")**

**corrplot(xx,method="circle",type="lower")**

**#Correlation matrix with p values.**

**library(Hmisc)**

**xx=rcorr(as.matrix(some.of.wages))**

**xx**

**#Conducting a Regression -- Continuous Variables Only**

**regout=lm(wage~yearsed+experience+age,data=some.of.wages)**

**summary(regout)**

**#Verifying the r^2 value.**

**cor(regout$fitted.values,some.of.wages$wage)^2**

**plot(some.of.wages$wage,regout$fitted.values,pch=19,**

**main="Actual v. Fitted Values")**

**#Exploring binary variables.**

**#Using the Union variable -- two levels.**

**regout=lm(wage~yearsed+experience+age+union,data=wages)**

**summary(regout)**

**#Adding gender to the model.**

**regout=lm(wage~yearsed+experience+age+union+gender,**

**data=wages)**

**summary(regout)**

**#Adding race to the model -- three levels.**

**regout=lm(wage~yearsed+experience+age+union+gender+race,**

**data=wages)**

**summary(regout)**

**#All Variables -- the "kitchen sink" model.**

**regout=lm(wage~yearsed+experience+age+union+gender+**

**race+marr+south+occupation+sector,data=wages)**

**summary(regout)**

**#Back to only continuous variables.**

**regout=lm(wage~yearsed+experience+age,data=some.of.wages)**

**summary(regout)**

**#Variance Inflation Factors (VIF)**

**#Measure of Multicollinearity –**

**#correlation of independents.**

**#How much the variance of a beta coefficient is**

**#being inflated by multicollinearity.**

**#Evidence of Multicollinearity.**

**plot(some.of.wages)**

**xx=cor(some.of.wages)**

**corrplot(xx,method="number")**

**corrplot(xx,method="ellipse")**

**#Variance Inflation Factors (VIF)**

**#Measure of Multicollinearity –**

**#correlation of independents.**

**#How much the variance of a beta coefficient is being**

**#inflated by multicollinearity.**

**library(car)**

**vif(regout)**

**#Back to the kitchen sink model.**

**regout=lm(wage~yearsed+experience+age+union+**

**gender+race+marr+south+occupation+sector,**

**data=wages)**

**summary(regout)**

**#Dump Experience, Keep Age**

**regout=lm(wage~yearsed+age+union+gender+race+**

**marr+south+occupation+sector,data=wages)**

**summary(regout)**

**vif(regout)**

**#Dump Age, Keep Experience**

**regout=lm(wage~yearsed+experience+union+gender+**

**race+marr+south+occupation+sector,data=wages)**

**summary(regout)**

**#Model with Experience and other**

**#continuous variables, Union and Gender**

**regout=lm(wage~yearsed+experience+union+gender,data=wages)**

**summary(regout)**

**#Bringing in Occupation**

**regout=lm(wage~yearsed+experience+union+gender+occupation,**

**data=wages)**

**summary(regout)**

**#Only two levels of Occupation seem to have a contribution.**

**#Now we collapse Occupation to "Professional & Management"**

**#and "Other"**

**wages$pm=NA**

**for(i in 1:length(wages$occupation)){**

**if(wages$occupation[i]=="Management"|**

**wages$occupation[i]=="Professional"){**

**wages$pm[i]="ProfMgt"}**

**else{**

**wages$pm[i]="Other"**

**}**

**}**

**#And conduct a regression with the new variable.**

**regout=lm(wage~yearsed+experience+union+gender+pm,**

**data=wages)**

**summary(regout)**

**#Let's separate out Professional and Management.**

**for(i in 1:length(wages$occupation)){**

**wages$pm[i]="Another"**

**if(wages$occupation[i]=="Management"){**

**wages$pm[i]="Management"}**

**if (wages$occupation[i]=="Professional"){**

**wages$pm[i]="Professional"**

**}**

**}**

**#And re-run the regression.**

**regout=lm(wage~yearsed+experience+union+gender+pm,**

**data=wages)**

**summary(regout)**

**#And evaluate the standardized residuals.**

**stdresids=rstandard(regout)**

**plot(regout$fitted.values,stdresids,pch=19)**

**abline(0,0,col="red",lwd=3)**

**#We have an outlier. Can we get rid of it?**

**#We have to find it first.**

**boxplot(wages$wage,col="red",ylim=c(0,50),pch=19)**

**max(wages$wage)**

**#This statement finds the data frame row**

**#that's the max value.**

**which(wages$wage==44.5)**

**wages[171,]**

**#Or combine the statements.**

**wages[which(wages$wage==44.5),]**

**#Now we create a new data frame that's a copy**

**#except for the outlier.**

**reduced.wages=wages[-171, ]**

**#Or...**

**reduced.wages=wages[-which(wages$wage==44.5),]**

**boxplot(reduced.wages$wage,col="red",ylim=c(0,50),pch=19)**

**#And rerun the regression.**

**regout=lm(wage~yearsed+experience+union+gender+pm,**

**data=reduced.wages)**

**summary(regout)**

**plot(regout$fitted.values,rstandard(regout),pch=19)**

**abline(0,0,col="red",lwd=3)**

**qqnorm(regout$residuals,pch=19)**

**qqline(regout$residuals,col="red",lwd=3)**

**hist(regout$residuals,col="red")**

**plot(density(regout$residuals),lwd=3,**

**main="Density Plot of Residuals")**

**#Leverage of Points**

**lev=hat(model.matrix(regout))**

**plot(lev,pch=19)**

**abline(3\*mean(lev),0,col="red",lwd=3)**

**reduced.wages[lev>(3\*mean(lev)),]**

**reduced.wages[lev>(3\*mean(lev)),1]**

**#So let's get rid of the high leverage data points.**

**no.leverage=reduced.wages**

**gilligan=reduced.wages[lev>(3\*mean(lev)),1]**

**no.leverage=no.leverage[-xx,]**

**# OR**

**no.leverage=reduced.wages**

**no.leverage=**

**no.leverage[-(reduced.wages[lev>(3\*mean(lev)),1]),]**

**#And re-run the regression.**

**regout=lm(wage~yearsed+experience+union+gender+pm,**

**data=no.leverage)**

**summary(regout)**

**#And look again at the residuals and leverage.**

**plot(regout$fitted.values,rstandard(regout),pch=19)**

**abline(0,0,col="red",lwd=3)**

**qqnorm(regout$residuals,pch=19)**

**qqline(regout$residuals,col="red",lwd=3)**

**hist(regout$residuals,col="red")**

**plot(density(regout$residuals),lwd=3,**

**main="Density Plot of Residuals")**

**lev=hat(model.matrix(regout))**

**plot(lev,pch=19)**

**abline(3\*mean(lev),0,col="red",lwd=3)**

**reduced.wages[lev>(3\*mean(lev)),1]**