**6304 Module 6**

**R Script File**

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

**#Read in data.**

**wages=rio::import("Wages Data.xlsx",which="Fixed Data")**

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

**names(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)**

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

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

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

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

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

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

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

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

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

**#Correlation matrix with p values.**

**library(Hmisc)**

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

**gilligan**

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

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

**summary(cont.out)**

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

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

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

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

**#Exploring binary variables.**

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

**reg.out=lm(wage~yearsed+experience+age+union,data=wages)**

**summary(reg.out)**

**points(wages$wage,reg.out$fitted.values,pch=19,col="red")**

**#Adding gender to the model.**

**reg.out=lm(wage~yearsed+experience+age+union+gender,**

**data=wages)**

**summary(reg.out)**

**points(wages$wage,reg.out$fitted.values,pch=19,col="green")**

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

**reg.out=lm(wage~yearsed+experience+age+union+gender+race,**

**data=wages)**

**summary(reg.out)**

**points(wages$wage,reg.out$fitted.values,pch=19,col="pink")**

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

**full.out=lm(wage~yearsed+experience+age+union+gender+**

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

**summary(full.out)**

**points(wages$wage,full.out$fitted.values,pch=19,**

**col="orange")**

**#Back to only continuous variables.**

**summary(cont.out)**

**#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)**

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

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

**corrplot(gilligan,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.**

**car::vif(cont.out)**

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

**summary(full.out)**

**car::vif(full.out)**

**#Drop Experience, Keep Age**

**noexp.out=lm(wage~yearsed+age+union+gender+race+**

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

**summary(noexp.out)**

**car::vif(noexp.out)**

**#Drop Age, Keep Experience**

**noage.out=lm(wage~yearsed+experience+union+gender+**

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

**summary(noage.out)**

**car::vif(noage.out)**

**#Arbitrary Choice**

**#Model with Experience and other**

**#continuous variables, Union and Gender**

**build.out=lm(wage~yearsed+experience+union+gender,**

**data=wages)**

**summary(build.out)**

**car::vif(build.out)**

**#Bringing in Occupation**

**build.out=lm(wage~yearsed+experience+union+gender**

**+occupation,data=wages)**

**summary(build.out)**

**car::vif(build.out)**

**#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.**

**build.out=lm(wage~yearsed+experience+union+gender+pm,**

**data=wages)**

**summary(build.out)**

**car::vif(build.out)**

**#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.**

**build.out=lm(wage~yearsed+experience+union+gender+pm,**

**data=wages)**

**summary(build.out)**

**car::vif(build.out)**

**#And evaluate the standardized residuals.**

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

**main="Standardized Residuals")**

**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,**

**main="Boxplot of Original Wages")**

**max(wages$wage)**

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

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

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

**wages[which.max(wages$wage),]**

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

**#except for the outlier.**

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

**max(reduced.wages$wage)**

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

**main="Reduced Data Set")**

**#And rerun the regression.**

**build.out=lm(wage~yearsed+experience+union+gender+pm,**

**data=reduced.wages)**

**summary(build.out)**

**car::vif(build.out)**

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

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

**qqnorm(build.out$residuals,pch=19)**

**qqline(build.out$residuals,col="red",lwd=3)**

**hist(build.out$residuals,col="red",**

**main="Basic Histogram of Residuals")**

**hist(build.out$residuals,col="red",**

**main="Histogram of Residuals",ylim=c(0,.15),**

**probability = TRUE)**

**curve(dnorm(x,mean(build.out$residuals),**

**sd(build.out$residuals)),**

**from=min(build.out$residuals),**

**to=max(build.out$residuals)**

**,lwd=3,col="blue",add=TRUE)**

**points(density(build.out$residuals),col="green",**

**type="l",lwd=3,**

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

**#Leverage of Points**

**lev=hat(model.matrix(build.out))**

**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.**

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

**no.leverage=reduced.wages[-gilligan,]**

**# OR**

**no.leverage=**

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

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

**build.out=lm(wage~yearsed+experience+union+gender+pm,**

**data=no.leverage)**

**summary(build.out)**

**car::vif(build.out)**

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

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

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

**qqnorm(build.out$residuals,pch=19)**

**qqline(build.out$residuals,col="red",lwd=3)**

**hist(build.out$residuals,col="red",**

**main="Histogram of Residuals",ylim=c(0,.15),**

**probability = TRUE)**

**curve(dnorm(x,mean(build.out$residuals),**

**sd(build.out$residuals)),**

**from=min(build.out$residuals),**

**to=max(build.out$residuals)**

**,lwd=3,col="blue",add=TRUE)**

**points(density(build.out$residuals),col="green",**

**type="l",lwd=3,**

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

**lev=hat(model.matrix(build.out))**

**plot(lev,pch=19,main="Leverages")**

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

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

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