**Wages Data**

**R Script File**

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

**library(rio)**

**# Read in data.**

**wages=import("6304 Wages Model Building**

**Example 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,pch=19,main="All Continuous Variables")**

**# Evaluating Correlations**

**# Reporting the Correlation Matrix**

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

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

**# Prettier ways to express correlations.**

**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")**

**rm(gilligan)**

**# Correlation matrix with p values.**

**library(Hmisc)**

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

**maryann**

**rm(maryann)**

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

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

**summary(regout)**

**#-----------------------------**

**par(mfrow=c(2,2))**

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

**main=paste("Actual v. Fitted, Wages, r=",**

**round(cor(some.of.wages$wage,**

**regout$fitted.values),3)))**

**abline(0,1,col="red",lwd=3)**

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

**main="Residuals QQ Plot, Wages")**

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

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

**main="Residuals, Wages",freq=FALSE)**

**curve(dnorm(x,mean(regout$residuals),**

**sd(regout$residuals)),**

**from=min(regout$residuals),**

**to=max(regout$residuals),**

**lwd=3,add=TRUE)**

**plot(wages$wage,rstandard(regout),pch=19,**

**main="Standardized Residuals, Wages")**

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

**par(mfrow=c(1,1))**

**moments::skewness(regout$residuals)**

**moments::kurtosis(regout$residuals)**

**#-----------------------------**

**# Exploring binary variables.**

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

**regout2=lm(wage~yearsed+experience+age+union,data=wages)**

**summary(regout)**

**summary(regout2)**

**# Adding gender to the model.**

**regout3=lm(wage~yearsed+experience+age+union+gender,**

**data=wages)**

**summary(regout)**

**summary(regout3)**

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

**regout4=lm(wage~yearsed+experience+age+union+gender**

**+race,data=wages)**

**summary(regout)**

**summary(regout4)**

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

**regout5=lm(wage~yearsed+experience+age+union**

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

**data=wages)**

**summary(regout5)**

**# Let's look again at only continuous variables.**

**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,pch=19,main="All Continuous Variables")**

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

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

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

**rm(gilligan)**

**# Variance Inflation Factors (VIF)**

**# Measure of Multicollinearity**

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

**# is being inflated by multicollinearity.**

**library(car)**

**vif(regout)**

**# Delete Experience, Keep Age**

**regout1a=lm(wage~yearsed+age,data=wages)**

**summary(regout1a)**

**vif(regout1a)**

**# Delete Age, Keep Experience**

**regout1b=lm(wage~yearsed+experience,data=wages)**

**summary(regout1b)**

**vif(regout1b)**

**# Model with Experience and other continuous variables,**

**# Union and Gender**

**regout6=lm(wage~yearsed+experience+union+gender,data=wages)**

**summary(regout6)**

**# Bringing in Occupation**

**regout7=lm(wage~yearsed+experience+union+gender+occupation,**

**data=wages)**

**summary(regout7)**

**# 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.**

**regout8=lm(wage~yearsed+experience+union+**

**gender+pm,data=wages)**

**summary(regout8)**

**# Now 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.**

**regout9=lm(wage~yearsed+experience+union**

**+gender+pm,data=wages)**

**summary(regout9)**

**# And evaluate the standardized residuals.**

**#-----------------------------**

**par(mfrow=c(2,2))**

**plot(wages$wage,regout9$fitted.values,pch=19,**

**main=paste("Actual v. Fitted, Regout9, r=",**

**round(cor(wages$wage,**

**regout9$fitted.values),3)))**

**abline(0,1,col="red",lwd=3)**

**qqnorm(regout9$residuals,pch=19,**

**main="Residuals QQ Plot, Regout9")**

**qqline(regout9$residuals,col="red",lwd=3)**

**hist(regout9$residuals,col="red",**

**main="Residuals, Regout9",freq=FALSE)**

**curve(dnorm(x,mean(regout9$residuals),**

**sd(regout9$residuals)),**

**from=min(regout9$residuals),**

**to=max(regout9$residuals),**

**lwd=3,add=TRUE)**

**plot(wages$wage,rstandard(regout9),pch=19,**

**main="Standardized Residuals, Regout9")**

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

**par(mfrow=c(1,1))**

**#-----------------------------**

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

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

**boxplot(wages$wage,col="red",pch=19,**

**main="Boxplot of Wages",ylim=c(0,50))**

**max(wages$wage)**

**max(rstandard(regout9))**

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

**# with the max wage value.**

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

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

**#Or combine the statements.**

**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),]**

**boxplot(reduced.wages$wage,col="red",pch=19,**

**main="Boxplot of Wages",ylim=c(0,50))**

**#And rerun the regression.**

**regout9a=lm(wage~yearsed+experience+union+gender+pm,**

**data=reduced.wages)**

**summary(regout9a)**

**#-----------------------------**

**par(mfrow=c(2,2))**

**plot(reduced.wages$wage,regout9a$fitted.values,pch=19,**

**main=paste("Actual v. Fitted, Regout9a, r=",**

**round(cor(reduced.wages$wage,**

**regout9a$fitted.values),3)))**

**abline(0,1,col="red",lwd=3)**

**qqnorm(regout9a$residuals,pch=19,**

**main="Residuals QQ Plot, Regout9a")**

**qqline(regout9a$residuals,col="red",lwd=3)**

**hist(regout9a$residuals,col="red",**

**main="Residuals, Regout9a",freq=FALSE)**

**curve(dnorm(x,mean(regout9a$residuals),**

**sd(regout9a$residuals)),**

**from=min(regout9a$residuals),**

**to=max(regout9a$residuals),**

**lwd=3,add=TRUE)**

**plot(reduced.wages$wage,rstandard(regout9a),pch=19,**

**main="Standardized Residuals, Regout9a")**

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

**par(mfrow=c(1,1))**

**#-----------------------------**

**#Leverage of Points**

**leverage=hat(model.matrix(regout9a))**

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

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

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

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

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

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

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

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

**regout9b=lm(wage~yearsed+experience+union+gender+pm,**

**data=no.leverage)**

**summary(regout9b)**

**#-----------------------------**

**par(mfrow=c(2,2))**

**plot(no.leverage$wage,regout9b$fitted.values,pch=19,**

**main=paste("Actual v. Fitted, Regout9b, r=",**

**round(cor(no.leverage$wage,**

**regout9b$fitted.values),3)))**

**abline(0,1,col="red",lwd=3)**

**qqnorm(regout9b$residuals,pch=19,**

**main="Residuals QQ Plot, Regout9b")**

**qqline(regout9b$residuals,col="red",lwd=3)**

**hist(regout9b$residuals,col="red",**

**main="Residuals, Regout9b",freq=FALSE)**

**curve(dnorm(x,mean(regout9b$residuals),**

**sd(regout9b$residuals)),**

**from=min(regout9a$residuals),**

**to=max(regout9b$residuals),**

**lwd=3,add=TRUE)**

**plot(no.leverage$wage,rstandard(regout9b),pch=19,**

**main="Standardized Residuals, Regout9b")**

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

**par(mfrow=c(1,1))**

**moments::skewness(regout9b$residuals)**

**moments::kurtosis(regout9b$residuals)**

**#-----------------------------**

**# What's going on in the wages data?**

**plot(no.leverage$experience,no.leverage$wage,pch=19,**

**main="Experience v. Wage, No Leverage")**

**plot(no.leverage$age,no.leverage$wage,pch=19,**

**main="Age v. Wage, No Leverage")**