PROG8430

Assignment02

Fei Yun

8680643

Data Transformation  
1. As demonstrated in class, transform any variables that are required to conduct the regression analysis.

names(Diamond)<-c('Price\_FY','Carat.Size\_FY','Source\_FY','Year\_FY','Clar\_FY','Col\_FY','Cut\_FY','Val\_FY')

Descriptive Data Analysis

1. Create numeric and graphical summaries of the data (as demonstrated in class).

summary(Diamond)

A close up of text on a white background

Description automatically generated

par(mfrow=c(3,3))

sapply(names(Diamond), function(cname){

if(is.numeric(Diamond[[cname]]))

print(hist(Diamond[[cname]],main=cname))

})

par(mfrow=c(1,1))

A close up of a organ

Description automatically generated

1. Comment on anything noteworthy or unusual. You are looking for

distributions that seem reasonable and reflective of the data you are analysing.

From the Histgram we can see the price of Diamond is not following normal distribution.

Outliers

1. Create boxplots of all relevant variables to determine outliers.

par(mfrow=c(3,3))

sapply(names(Diamond), function(cname){

if(is.numeric(Diamond[[cname]]))

print(boxplot(Diamond[[cname]],main=cname))

})

par(mfrow=c(1,1))

Box and whisker chart

Description automatically generated

1. Comment on any outliers you see.

From Cut and Clar and Carat.Size and Val outliers we can see they have extreme values out of max and min values in this variables.

Exploratory Analysis

1. Create QQNorm plots and numeric tests for normality of data and

identify data that seems to be normal and not anything else that seems

remarkable. If none do, state that.

par(mfrow=c(3,3))

sapply(names(Diamond), function(cname){

if(is.numeric(Diamond[[cname]]))

print(qqnorm(Diamond[[cname]],main=cname))

})

par(mfrow=c(1,1))

Shape

Description automatically generated

NumDiamd<-Diamond[-c(3)]

DiaNrm<-lapply(NumDiamd,shapiro.test)

DiaNrm

str(DiaNrm[[4]])

DiaRes<-sapply(DiaNrm, '[',c("statistic","p.value"))

DiaRest<-t(DiaRes)

DiaRest

Text

Description automatically generated

From this statistic and p.value we can see none of those variables is following normal disturibute.

2 Correlations: Create both numeric and graphical correlations (as demonstrated) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?

library(corrgram)

corrgram(NumDiamd,order=TRUE,lower.panel=panel.shade,upper.panel=panel.pie,

text.panel=panel.txt,main="Diamond stats")

A picture containing diagram

Description automatically generated

res<-cor(NumDiamd,method = "spearman")

round(res,2)

Table

Description automatically generated

From correlation we can see Price have strong positive relation with carat.size. And carat size has strong positive relation with its value. The year of first cut almost has nothing relation with anything. But clarity has negative correlation with price and carat.size. And we can see the price the diamond sold for almost euqual its value for insurance.

Model Development  
As demonstrated in class, create three models using three automatic variable selection techniques discussed in class (Full, Forward, Stepwise). For each model interpret and comment on the five main measures we discussed in class:

Fullmodel

Dia\_lm=lm(Price\_FY~Carat.Size\_FY+Year\_FY+Clar\_FY+Col\_FY+Cut\_FY+Val\_FY,data=NumDiamd,na.action = na.omit)

summary(Dia\_lm)

Residuals:

Min 1Q Median 3Q Max

-2771.6 -494.6 -75.4 340.4 3770.0

A close up of text on a white background

Description automatically generated

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 88.36% of variation is explained by the model.
3. The residuals look not approximately symmetrical.
4. 4 variables (and the intercept) look significant (p-values of t-test <0.001)
5. Variable color is negatively correlated with price instead of positively, and clarity is positive correlated with price instead of negative, years seems negative correlative with price.

Stepwise

Stp\_Dia\_lm=step(Dia\_lm,details=TRUE)

summary(Stp\_Dia\_lm)A close up of text on a white background

Description automatically generated

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 88.36% of variation is explained by the model.
3. The residuals look not approximately symmetrical.
4. 4 variables (and the intercept) look very significant (p-values of t-test <0.001), year was dropped out.
5. Variable color is negatively correlated with price instead of positively, clarity still positive with price, but year does not has relation with price.

Forward

min\_model<-lm(Price\_FY~1,data=NumDiamd,action=na.omit)

Fwd\_Dia\_lm=step(min\_model,direction = "forward",scope = (~Carat.Size\_FY+Year\_FY+Clar\_FY+Col\_FY+Cut\_FY+Val\_FY),details=TRUE)

summary(Fwd\_Dia\_lm)

A screenshot of text

Description automatically generated

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 88.36% of variation is explained by the model.
3. The residuals look not approximately symmetrical.
4. 4 variables (and the intercept) look very significant (p-values of t-test <0.001), year was dropped out.
5. Variable color is negatively correlated with price instead of positively, clarity still positive with price, but year does not has relation with price.

Model Evaluation

#Model Evaluation

DiaFit<-predict(Dia\_lm)

FullDiaRes<-residuals(Dia\_lm)

StpDiaFit<-predict(Stp\_Dia\_lm)

StpDiaRes<-residuals(Stp\_Dia\_lm)

FwdDiaFit<-predict(Fwd\_Dia\_lm)

FwdDiaRes<-residuals(Fwd\_Dia\_lm)

#Numercial

shapiro.test(FullDiaRes)

shapiro.test(StpDiaRes)

shapiro.test(FwdDiaRes)

Independence of Predictors

From correlation result we can see the insurance value of diamond almost equal to diamond price, but from models we can see the value of insurance does not have strong significant (p-value>0.1)

Distribution of Error Terms

data: FullDiaRes

W = 0.93852, p-value < 2.2e-16

data: StpDiaRes

W = 0.93875, p-value < 2.2e-16

data: FwdDiaRes

W = 0.93875, p-value < 2.2e-16

From shapiro test we can see all these three model are following normal distribute.

**Non-AutoCorrelation and Homoscedasticity**

Based on Residuals vs. Fitted (form a "horizontal band" around the 0 line

) and Scale-Location, the variances of the error terms are equal. There appears to be no explicit pattern to the residuals. Therefore, no there is no appearance of autocorrelation.

Based on Residuals vs. Leverage and Cook’s Distance, , there is no data point exerting undue influence or leverage on the model.

Diagram, engineering drawing

Description automatically generated

Final Model

Base on above,

The final model is

Price=8242.452\*Carat.Size+375.908\*Clarity-376.916\*Color-142.923\*Cut+80.575\*Value