Assignment 08

Regression Analysis

Dean D'souza

H.U. ID: 168424

***Redo examples 9.10, 9.11, 9.15, and 9.16 from the chapter and compare your answers with the book's result. The dataset required is supplied in the unit***

For the purpose of this assignment I chose to perform the examples using R through R studio.

# Solutions:

### Example 9.10: Interpreting Regression Results for the Colleges and Universities Data

We perform the required commands for obtaining the required statistics as follows (Names of each variable were shortened for convenience):

# Importing the data  
cau <- read\_excel("~/GitHub/CISC\_520-50\_FA2016/Assignment8/data/Colleges and Universities.xlsx", skip = 1)  
  
# Cleaning up the data frame  
cau <- cau[,-8]  
cau <- cau[c(1:49),]  
names(cau) <- c("School","Type","SAT","Acceptance","Expenditure","Top.10.HS","GraduationPercent")  
  
# Summary Statistics  
summary(cau)

## School Type SAT Acceptance   
## Length:49 Length:49 Min. :1109 Min. :0.170   
## Class :character Class :character 1st Qu.:1225 1st Qu.:0.280   
## Mode :character Mode :character Median :1260 Median :0.360   
## Mean :1263 Mean :0.381   
## 3rd Qu.:1300 3rd Qu.:0.480   
## Max. :1400 Max. :0.670   
## Expenditure Top.10.HS GraduationPercent  
## Min. : 15904 Min. :47.0 Min. :61.00   
## 1st Qu.: 20179 1st Qu.:65.0 1st Qu.:77.00   
## Median : 24718 Median :76.0 Median :85.00   
## Mean : 30060 Mean :74.2 Mean :83.24   
## 3rd Qu.: 37137 3rd Qu.:85.0 3rd Qu.:89.00   
## Max. :102262 Max. :98.0 Max. :93.00

# Building the multiple Regression Model  
cau.lm <- lm(GraduationPercent ~ SAT + Acceptance + Expenditure + Top.10.HS, data=cau)  
  
# Model Statistics  
summary(cau.lm)

##   
## Call:  
## lm(formula = GraduationPercent ~ SAT + Acceptance + Expenditure +   
## Top.10.HS, data = cau)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.8674 -2.0462 0.6193 3.6417 11.2090   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.792e+01 2.456e+01 0.730 0.469402   
## SAT 7.201e-02 1.798e-02 4.004 0.000236 \*\*\*  
## Acceptance -2.486e+01 8.315e+00 -2.990 0.004560 \*\*   
## Expenditure -1.356e-04 6.593e-05 -2.057 0.045600 \*   
## Top.10.HS -1.628e-01 7.934e-02 -2.051 0.046214 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.308 on 44 degrees of freedom  
## Multiple R-squared: 0.5344, Adjusted R-squared: 0.4921   
## F-statistic: 12.63 on 4 and 44 DF, p-value: 6.332e-07

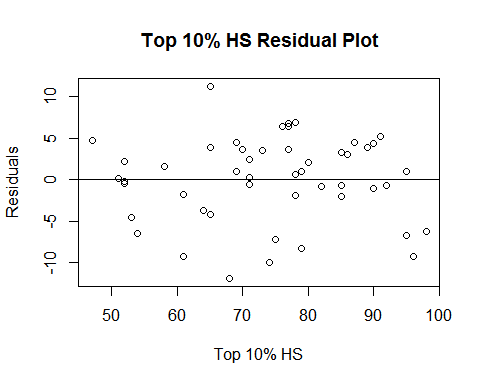
anova(cau.lm)

## Analysis of Variance Table  
##   
## Response: GraduationPercent  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SAT 1 847.55 847.55 30.0780 1.92e-06 \*\*\*  
## Acceptance 1 185.59 185.59 6.5864 0.013756 \*   
## Expenditure 1 271.49 271.49 9.6346 0.003333 \*\*   
## Top.10.HS 1 118.58 118.58 4.2081 0.046214 \*   
## Residuals 44 1239.85 28.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

confint(cau.lm)

## 2.5 % 97.5 %  
## (Intercept) -3.157088e+01 6.741279e+01  
## SAT 3.576208e-02 1.082505e-01  
## Acceptance -4.161739e+01 -8.101078e+00  
## Expenditure -2.685259e-04 -2.773789e-06  
## Top.10.HS -3.226729e-01 -2.856120e-03

# Plot for Top 10% HS  
plot(cau$Top.10.HS, cau.lm$residuals, xlab = "Top 10% HS", ylab = "Residuals", main = "Top 10% HS Residual Plot")  
abline(0,0)



As we can see from the above statistics, we obtain a similar value for R-squared of 0.53 and also the p-value for each of the variables indicate that they are all significant. We also plot a single residual plot for TOP 10% HS which confirms that assumptions seem to be met.

### Example 9.11: Identifying the best Regression Model (Banking Data)

We perform the commands for obtaining the first model as follows (Names of each variable were shortened for convenience):

# Importing the data  
bd <- read\_excel("~/GitHub/CISC\_520-50\_FA2016/Assignment8/data/Banking Data.xlsx", skip = 2)  
  
# Cleaning up the data  
bd <- bd[-103,]  
names(bd) <- c("MedAge","MedEdu","MedInc","MedHomeVal","MedWealth","AvgBal")  
  
# Summary of data  
summary(bd)

## MedAge MedEdu MedInc MedHomeVal   
## Min. :19.50 Min. :11.00 Min. : 7741 Min. : 40313   
## 1st Qu.:33.92 1st Qu.:12.40 1st Qu.: 35078 1st Qu.: 83017   
## Median :36.10 Median :12.70 Median : 47656 Median : 97744   
## Mean :35.45 Mean :12.98 Mean : 48811 Mean :106845   
## 3rd Qu.:37.58 3rd Qu.:13.20 3rd Qu.: 60157 3rd Qu.:121791   
## Max. :43.10 Max. :16.10 Max. :111548 Max. :276139   
## MedWealth AvgBal   
## Min. : 24999 Min. : 5956   
## 1st Qu.: 70263 1st Qu.:20036   
## Median :102348 Median :24661   
## Mean :109026 Mean :24888   
## 3rd Qu.:142518 3rd Qu.:29180   
## Max. :331009 Max. :56569

# Building the first model  
bd.lm <- lm(AvgBal ~ MedAge + MedEdu + MedInc + MedHomeVal + MedWealth, data=bd)  
summary(bd.lm)

##   
## Call:  
## lm(formula = AvgBal ~ MedAge + MedEdu + MedInc + MedHomeVal +   
## MedWealth, data = bd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5376.9 -1110.8 -77.2 872.3 7732.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.071e+04 4.261e+03 -2.514 0.013613 \*   
## MedAge 3.187e+02 6.099e+01 5.225 1.01e-06 \*\*\*  
## MedEdu 6.219e+02 3.190e+02 1.950 0.054135 .   
## MedInc 1.463e-01 4.078e-02 3.588 0.000527 \*\*\*  
## MedHomeVal 9.183e-03 1.104e-02 0.832 0.407505   
## MedWealth 7.433e-02 1.119e-02 6.643 1.85e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2056 on 96 degrees of freedom  
## Multiple R-squared: 0.9469, Adjusted R-squared: 0.9441   
## F-statistic: 342.4 on 5 and 96 DF, p-value: < 2.2e-16

We can see that Median Home Value (MedHomeVal) is not significant (as the p-value is greater than 0.05 and is the largest among all variables). Hence we build the new model as follows:

# Building the second model  
bd.lm2 <- lm(AvgBal ~ MedAge + MedEdu + MedInc + MedWealth, data=bd)  
summary(bd.lm2)

##   
## Call:  
## lm(formula = AvgBal ~ MedAge + MedEdu + MedInc + MedWealth, data = bd)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5413.6 -1214.2 -62.3 981.7 7439.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.243e+04 3.719e+03 -3.343 0.00118 \*\*   
## MedAge 3.251e+02 6.040e+01 5.382 5.13e-07 \*\*\*  
## MedEdu 7.734e+02 2.614e+02 2.958 0.00389 \*\*   
## MedInc 1.598e-01 3.739e-02 4.272 4.52e-05 \*\*\*  
## MedWealth 7.299e-02 1.105e-02 6.603 2.16e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2052 on 97 degrees of freedom  
## Multiple R-squared: 0.9465, Adjusted R-squared: 0.9443   
## F-statistic: 429.2 on 4 and 97 DF, p-value: < 2.2e-16

We can thus verify that the Adjusted R-squared value increases slightly while the R-squared value decreases slightly. Additionally, we can also see that the p-value for Median Years of Education (MedEdu) has decreased below the significance level of 0.05 and hence all terms are significant.

### Example 9.15: A Regression Model with Multiple Levels of Categorical Variables

We perform the commands for obtaining the required statistics as follows (Names of each variable were shortened for convenience):

# Importing the data  
sf <- read\_excel("~/GitHub/CISC\_520-50\_FA2016/Assignment8/data/Surface Finish.xlsx", skip = 1)  
  
# Cleaning up and transforming data  
sf <- sf[-36,]  
names(sf) <- c("Part","SurFin","RPM","CT")  
  
CT\_A <- NULL  
for (i in 1:nrow(sf)){  
 if(sf[i,]$CT == 'A'){  
 CT\_A[i] <- 1  
 } else {  
 CT\_A[i] <- 0  
 }  
}  
  
CT\_B <- NULL  
for (i in 1:nrow(sf)){  
 if(sf[i,]$CT == 'B'){  
 CT\_B[i] <- 1  
 } else {  
 CT\_B[i] <- 0  
 }  
}  
  
CT\_C <- NULL  
for (i in 1:nrow(sf)){  
 if(sf[i,]$CT == 'C'){  
 CT\_C[i] <- 1  
 } else {  
 CT\_C[i] <- 0  
 }  
}  
  
CT\_D <- NULL  
for (i in 1:nrow(sf)){  
 if(sf[i,]$CT == 'D'){  
 CT\_D[i] <- 1  
 } else {  
 CT\_D[i] <- 0  
 }  
}  
  
sf <- sf[,-4]  
  
new\_sf <- data.frame(sf, CT\_A, CT\_B, CT\_C, CT\_D)  
  
# Summary Statistics  
summary(new\_sf)

## Part SurFin RPM CT\_A   
## Min. : 1.0 Min. :20.50 Min. :200.0 Min. :0.0000   
## 1st Qu.: 9.5 1st Qu.:27.15 1st Qu.:222.5 1st Qu.:0.0000   
## Median :18.0 Median :32.29 Median :237.0 Median :0.0000   
## Mean :18.0 Mean :34.15 Mean :235.6 Mean :0.2857   
## 3rd Qu.:26.5 3rd Qu.:43.41 3rd Qu.:249.0 3rd Qu.:1.0000   
## Max. :35.0 Max. :52.26 Max. :265.0 Max. :1.0000   
## CT\_B CT\_C CT\_D   
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.2857 Mean :0.2857 Mean :0.1429   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

It should be noted that the 'Part' variable refers to a unique identifier for each part and hence can be left out of the regression model. We build the regression model as follows:

# Building the model  
sf.lm <- lm(SurFin ~ RPM + CT\_A + CT\_B + CT\_C + CT\_D, data=new\_sf)  
summary(sf.lm)

##   
## Call:  
## lm(formula = SurFin ~ RPM + CT\_A + CT\_B + CT\_C + CT\_D, data = new\_sf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0165 -0.7805 0.1189 0.6170 2.0916   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.54237 2.51655 -0.613 0.545   
## RPM 0.09776 0.01040 9.400 1.89e-10 \*\*\*  
## CT\_A 26.03675 0.59689 43.621 < 2e-16 \*\*\*  
## CT\_B 12.72618 0.59717 21.311 < 2e-16 \*\*\*  
## CT\_C 5.54975 0.59689 9.298 2.42e-10 \*\*\*  
## CT\_D NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.089 on 30 degrees of freedom  
## Multiple R-squared: 0.9889, Adjusted R-squared: 0.9874   
## F-statistic: 669.7 on 4 and 30 DF, p-value: < 2.2e-16

As we can see from the above summary statistics, most of the new columns created (except for parts of type 'D' which we cannot verify here due to singularities) as well as 'RPM' are significant and we may choose to build additional models for each of the Cutting Tools as follows:

# Building model for Type="A"  
sf.lma <- lm(SurFin ~ RPM + CT\_A, data=new\_sf)  
summary(sf.lma)

##   
## Call:  
## lm(formula = SurFin ~ RPM + CT\_A, data = new\_sf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8174 -2.3954 -0.6611 4.4283 7.6285   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.06942 9.97955 0.809 0.4247   
## RPM 0.08799 0.04219 2.086 0.0451 \*   
## CT\_A 18.72540 1.65499 11.314 1.02e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.423 on 32 degrees of freedom  
## Multiple R-squared: 0.8052, Adjusted R-squared: 0.793   
## F-statistic: 66.12 on 2 and 32 DF, p-value: 4.309e-12

# Building model for Type="B"  
sf.lmb <- lm(SurFin ~ RPM + CT\_B, data=new\_sf)  
summary(sf.lmb)

##   
## Call:  
## lm(formula = SurFin ~ RPM + CT\_B, data = new\_sf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.1164 -7.3752 -0.4959 11.5870 15.5795   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 13.67024 22.35064 0.612 0.545  
## RPM 0.08683 0.09438 0.920 0.364  
## CT\_B 0.07978 3.70225 0.022 0.983  
##   
## Residual standard error: 9.891 on 32 degrees of freedom  
## Multiple R-squared: 0.02577, Adjusted R-squared: -0.03512   
## F-statistic: 0.4232 on 2 and 32 DF, p-value: 0.6586

# Building model for Type="C"  
sf.lmc <- lm(SurFin ~ RPM + CT\_C, data=new\_sf)  
summary(sf.lmc)

##   
## Call:  
## lm(formula = SurFin ~ RPM + CT\_C, data = new\_sf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.9667 -3.1061 -0.6482 8.7008 12.7326   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.70340 19.63082 0.851 0.40116   
## RPM 0.08613 0.08299 1.038 0.30716   
## CT\_C -9.95659 3.25555 -3.058 0.00447 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.701 on 32 degrees of freedom  
## Multiple R-squared: 0.2461, Adjusted R-squared: 0.199   
## F-statistic: 5.223 on 2 and 32 DF, p-value: 0.01089

# Building model for Type="D"  
sf.lmd <- lm(SurFin ~ RPM + CT\_D, data=new\_sf)  
summary(sf.lmd)

##   
## Call:  
## lm(formula = SurFin ~ RPM + CT\_D, data = new\_sf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.874 -7.893 -1.775 9.762 13.086   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.92158 18.67209 0.692 0.493912   
## RPM 0.09907 0.07910 1.252 0.219525   
## CT\_D -14.77367 4.00591 -3.688 0.000834 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.286 on 32 degrees of freedom  
## Multiple R-squared: 0.3163, Adjusted R-squared: 0.2736   
## F-statistic: 7.403 on 2 and 32 DF, p-value: 0.002278

### Example 9.16: Modeling Beverage Sales Using Curvilinear Regression

We perform the commands to obtain the required statistics as follows:

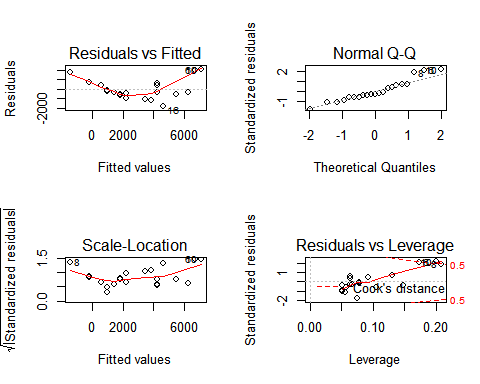
# Importing the data  
bsal <- read\_excel("~/GitHub/CISC\_520-50\_FA2016/Assignment8/data/Beverage Sales.xlsx", skip = 1)  
  
# Cleaning up the data  
bsal <- bsal[-22,]  
  
# Summary Statistics  
summary(bsal)

## Temperature Sales   
## Min. :76.00 Min. : 266   
## 1st Qu.:82.00 1st Qu.:1018   
## Median :85.00 Median :1810   
## Mean :86.62 Mean :2882   
## 3rd Qu.:90.00 3rd Qu.:4615   
## Max. :97.00 Max. :9138

# Building the model  
bsal.lm <- lm(Sales ~ Temperature, data=bsal)  
summary(bsal.lm)

##   
## Call:  
## lm(formula = Sales ~ Temperature, data = bsal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1765.6 -598.4 -293.0 562.0 2014.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -32511.25 3408.72 -9.538 1.12e-08 \*\*\*  
## Temperature 408.60 39.27 10.406 2.76e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1041 on 19 degrees of freedom  
## Multiple R-squared: 0.8507, Adjusted R-squared: 0.8429   
## F-statistic: 108.3 on 1 and 19 DF, p-value: 2.761e-09

# Plotting  
par(mfrow=c(2,2))  
plot(bsal.lm)

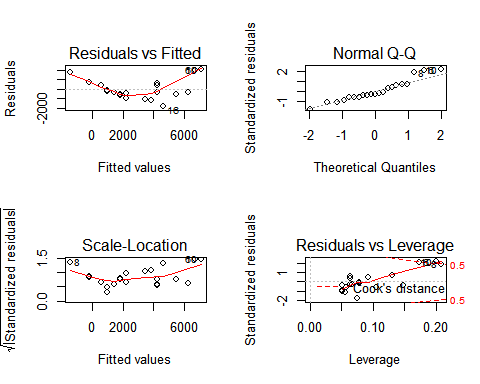


We also build another model taking into account the curvilinear nature of the data.

# Building the curvilinear model  
bsal.lm2 <- lm(Sales ~ Temperature + (Temperature\*Temperature), data=bsal)  
summary(bsal.lm2)

##   
## Call:  
## lm(formula = Sales ~ Temperature + (Temperature \* Temperature),   
## data = bsal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1765.6 -598.4 -293.0 562.0 2014.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -32511.25 3408.72 -9.538 1.12e-08 \*\*\*  
## Temperature 408.60 39.27 10.406 2.76e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1041 on 19 degrees of freedom  
## Multiple R-squared: 0.8507, Adjusted R-squared: 0.8429   
## F-statistic: 108.3 on 1 and 19 DF, p-value: 2.761e-09

# Plotting  
par(mfrow=c(2,2))  
plot(bsal.lm2)



As we can see form the above summary statistics, the algorithm had automatically adjusted for the curvilinear nature of the data and hence we obtain the best model as shown above.

# References

[1] "Assessing Regression" Lecture Slides, Stephen Penn, DM, PMP, Harrisburg University of Science and Technology, ANLY-510, Summer 2016