**SDM Homework B: Spring 2020**

**Yagna Dheepika Venkitasamy**

SDM Homework B: Baldrige dataset Simple Regression, and Multiple Regression with Dummy Variables and Interactions

Download the dataset “baldrige2011.xlsx” posted on Canvas. You have used this dataset earlier but to refresh your memory, I have provided details about the dataset towards the end of this document. To answer the questions, please use R markdown to execute the R code and document it with appropriate comments and observations wherever it is required. Please provide professional looking tables and charts wherever requested so that they are self-explanatory when printed in black and white; you can use “stargazer” library for showing the output tables and “ggplot2” for graphs, or other R packages.

rm(list=ls())  
library(rio)  
library(moments)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

df = import("baldridge2011.xlsx")  
setwd("C:/Users/yagna/Documents/R/R workings")  
str(df)

## 'data.frame': 1099 obs. of 177 variables:  
## $ slnoskm17mar11: num 1 2 3 4 5 6 7 8 9 10 ...  
## $ year : num 1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...  
## $ sector : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ applicant : num 1 2 5 7 9 10 12 15 16 17 ...  
## $ permanentid : chr "B-179" NA NA NA ...  
## $ i1.1score : num 24 8 24 9 27 17 15 23 18 21 ...  
## $ i1.1max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i1.2score : num 16 0 14 0 18 0 0 14 0 0 ...  
## $ i1.2max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ i1.3score : num 21 0 21 0 24 0 0 21 0 0 ...  
## $ i1.3max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i1.4score : num 14 0 10 0 14 0 0 12 0 0 ...  
## $ i1.4max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ icat1total : num 75 8 69 9 83 17 15 70 18 21 ...  
## $ icat1max : num 100 100 100 100 100 100 100 100 100 100 ...  
## $ i2.1score : num 25 14 26 9 18 19 14 21 21 16 ...  
## $ i2.1max : num 35 35 35 35 35 35 35 35 35 35 ...  
## $ i2.2score : num 18 0 18 0 15 0 0 15 0 0 ...  
## $ i2.2max : num 25 25 25 25 25 25 25 25 25 25 ...  
## $ i2.3score : num NA NA NA NA NA NA NA NA NA NA ...  
## $ i2.3max : num NA NA NA NA NA NA NA NA NA NA ...  
## $ icat2total : num 43 14 44 9 33 19 14 36 21 16 ...  
## $ icat2max : num 60 60 60 60 60 60 60 60 60 60 ...  
## $ i3.1score : num 28 8 28 6 16 24 12 28 24 20 ...  
## $ i3.1max : num 40 40 40 40 40 40 40 40 40 40 ...  
## $ i3.2score : num 15 0 15 0 10 0 0 15 0 0 ...  
## $ i3.2max : num 25 25 25 25 25 25 25 25 25 25 ...  
## $ i3.3score : num 15 0 18 0 10 0 0 13 0 0 ...  
## $ i3.3max : num 25 25 25 25 25 25 25 25 25 25 ...  
## $ icat3total : num 58 8 61 6 36 24 12 56 24 20 ...  
## $ icat3max : num 90 90 90 90 90 90 90 90 90 90 ...  
## $ i4.1score : num 21 0 18 0 21 0 0 15 0 0 ...  
## $ i4.1max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i4.2score : num 32 8 22 8 20 22 8 24 18 20 ...  
## $ i4.2max : num 40 40 40 40 40 40 40 40 40 40 ...  
## $ i4.3score : num 28 2 22 12 28 20 6 22 16 16 ...  
## $ i4.3max : num 40 40 40 40 40 40 40 40 40 40 ...  
## $ i4.4score : num 10 0 14 0 12 0 0 10 0 0 ...  
## $ i4.4max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ i4.5score : num 16 0 12 0 14 0 0 10 0 0 ...  
## $ i4.5max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ icat4total : num 107 10 88 20 95 42 14 81 34 36 ...  
## $ icat4max : num 150 150 150 150 150 150 150 150 150 150 ...  
## $ i5.1score : num 24 15 23 14 18 23 14 24 15 18 ...  
## $ i5.1max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i5.2score : num 18 6 16 9 18 15 13 13 15 16 ...  
## $ i5.2max : num 25 25 25 25 25 25 25 25 25 25 ...  
## $ i5.3score : num 10 0 18 0 18 0 0 15 0 0 ...  
## $ i5.3max : num 25 25 25 25 25 25 25 25 25 25 ...  
## $ i5.4score : num 11 0 12 0 9 0 0 9 0 0 ...  
## $ i5.4max : num 15 15 15 15 15 15 15 15 15 15 ...  
## $ i5.5score : num 6 0 8 0 6 0 0 6 0 0 ...  
## $ i5.5max : num 10 10 10 10 10 10 10 10 10 10 ...  
## $ i5.6score : num 18 0 20 0 10 0 0 15 0 0 ...  
## $ i5.6max : num 25 25 25 25 25 25 25 25 25 25 ...  
## $ i5.7score : num 16 0 16 0 10 0 0 8 0 0 ...  
## $ i5.7max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ icat5total : num 103 21 113 23 89 38 27 90 30 34 ...  
## $ icat5max : num 150 150 150 150 150 150 150 150 150 150 ...  
## $ i6.1score : num 35 10 30 10 35 23 8 30 20 28 ...  
## $ i6.1max : num 50 50 50 50 50 50 50 50 50 50 ...  
## $ i6.2score : num 21 16 23 9 21 21 9 19 19 18 ...  
## $ i6.2max : num 35 35 35 35 35 35 35 35 35 35 ...  
## $ i6.3score : num 25 14 18 9 21 14 5 19 14 12 ...  
## $ i6.3max : num 35 35 35 35 35 35 35 35 35 35 ...  
## $ i6.4score : num 15 0 15 0 15 0 0 12 0 0 ...  
## $ i6.4max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i6.5score : num NA NA NA NA NA NA NA NA NA NA ...  
## $ i6.5max : num NA NA NA NA NA NA NA NA NA NA ...  
## $ icat6total : num 96 40 86 28 92 58 22 80 53 58 ...  
## $ icat6max : num 150 150 150 150 150 150 150 150 150 150 ...  
## $ i7.1score : num 40 20 40 13 35 33 20 33 30 33 ...  
## $ i7.1max : num 50 50 50 50 50 50 50 50 50 50 ...  
## $ i7.2score : num 18 0 24 0 24 0 0 24 0 0 ...  
## $ i7.2max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i7.3score : num 6 0 14 0 8 0 0 14 0 0 ...  
## $ i7.3max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ i7.4score : num 14 0 16 0 14 0 0 14 0 0 ...  
## $ i7.4max : num 20 20 20 20 20 20 20 20 20 20 ...  
## $ i7.5score : num 21 0 24 0 18 0 0 18 0 0 ...  
## $ i7.5max : num 30 30 30 30 30 30 30 30 30 30 ...  
## $ i7.6score : num 35 0 40 0 25 0 0 30 0 0 ...  
## $ i7.6max : num 50 50 50 50 50 50 50 50 50 50 ...  
## $ i7.7score : num 35 8 40 8 30 30 8 30 20 28 ...  
## $ i7.7max : num 50 50 50 50 50 50 50 50 50 50 ...  
## $ i7.8score : num 35 18 45 8 30 23 8 33 23 23 ...  
## $ i7.8max : num 50 50 50 50 50 50 50 50 50 50 ...  
## $ icat7total : num 204 46 243 29 184 86 36 196 73 84 ...  
## $ icat7max : num 300 300 300 300 300 300 300 300 300 300 ...  
## $ iirtotal : num 686 147 704 124 612 284 140 609 253 269 ...  
## $ iirmax : num 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 ...  
## $ c1.1score : chr "21" NA "23" NA ...  
## $ c1.1max : num 30 NA 30 NA 30 NA NA 30 NA NA ...  
## $ c1.2score : chr "14" NA "13" NA ...  
## $ c1.2max : num 20 NA 20 NA 20 NA NA 20 NA NA ...  
## $ c1.3score : chr "21" NA "20" NA ...  
## $ c1.3max : num 30 NA 30 NA 30 NA NA 30 NA NA ...  
## $ c1.4score : num 14 NA 12 NA 14 NA NA 12 NA NA ...  
## $ c1.4max : num 20 NA 20 NA 20 NA NA 20 NA NA ...  
## [list output truncated]

library(car)

## Loading required package: carData

1. SLR: Run a simple regression model m1 with “ccat7total” as the response variable and “icat4total” as the (only) explanatory variable for the 1999-2006 period only. Interpret the coefficient of icat4total. Also calculate the correlation between “ccat7total” and “icat4total”, and explain how this relates to R squared of the model m1. Add any further observations that you can infer from this analysis.

df$ccat7total <- as.numeric(df$ccat7total)

## Warning: NAs introduced by coercion

#subsetting the variables for the required years only  
df1 <- subset(df[c("ccat7total","icat1total","icat2total","icat3total",  
 "icat4total","icat5total","icat6total", "year","sector")], df$year >= 1999)  
#running the linear regression model  
m1 <- lm(ccat7total ~ icat4total, data=df1)  
stargazer(m1,type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## -----------------------------------------------  
## icat4total 2.423\*\*\*   
## (0.439)   
##   
## Constant 70.745\*\*\*   
## (22.013)   
##   
## -----------------------------------------------  
## Observations 243   
## R2 0.112   
## Adjusted R2 0.108   
## Residual Std. Error 49.136 (df = 241)   
## F Statistic 30.420\*\*\* (df = 1; 241)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

summary(m1)

##   
## Call:  
## lm(formula = ccat7total ~ icat4total, data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -123.276 -33.932 5.301 35.011 115.376   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 70.7453 22.0129 3.214 0.00149 \*\*   
## icat4total 2.4229 0.4393 5.515 8.95e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 49.14 on 241 degrees of freedom  
## (221 observations deleted due to missingness)  
## Multiple R-squared: 0.1121, Adjusted R-squared: 0.1084   
## F-statistic: 30.42 on 1 and 241 DF, p-value: 8.948e-08

#correlation on the 2 variables  
cor(df1$ccat7total,df1$icat4total,use="complete.obs")

## [1] 0.3347802

Interpretation: We can interpret that with the unit increase in the icat4total, the ccat7total score increases by 2.432 units. From the correlation of 0.33 we can see that they have a weak positive correlation between the 2 variables. The R squared score of about only 10% signifies that the model might not be the best way to explain the correlation between the variables.

1. MLR: Now, we would like to investigate the impact of multiple variable using multiple linear regression. To that end, run a regression model m2 with “ccat7total” as the response variable and six individual category totals (icat1total, icat2total, icat3total, icat4total, icat5total, icat6total) as predictor variables only for the 1999-2006 period. Interpret the coefficient of icat1total and icat3total. Also comment on other variables that are statistically significant. Add any further observations that you can infer from this analysis.

#running the linear regression model  
m2 <- lm(ccat7total ~ icat1total+icat2total+icat3total+icat4total+icat5total+ icat6total, data=df1)  
  
stargazer(m2,type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## -----------------------------------------------  
## icat1total 1.682\*\*\*   
## (0.500)   
##   
## icat2total 0.349   
## (0.599)   
##   
## icat3total 0.994\*   
## (0.593)   
##   
## icat4total -0.297   
## (0.560)   
##   
## icat5total 0.254   
## (0.583)   
##   
## icat6total 0.953\*   
## (0.520)   
##   
## Constant -29.754   
## (24.227)   
##   
## -----------------------------------------------  
## Observations 243   
## R2 0.288   
## Adjusted R2 0.269   
## Residual Std. Error 44.476 (df = 236)   
## F Statistic 15.879\*\*\* (df = 6; 236)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Interpretation: Co-efficient of icat1total: with every unit increase in the score of icat1total, there will be an increase of 1.682 units in the ccat7total score. Co-efficient of icat3total: with every unit increase in the score of icat3total, there will be an increase of 0.994 units in the ccat7total score. Other significant variable: with every unit increase in the score of icat6total, there will be an increase of 0.953 units in the ccat7total score.

1. MLR with Dummy Variables: Now, run a regression model m3 with “ccat7total” as the response variable, six individual category totals (icat1total, icat2total, icat3total, icat4total, icat5total, icat6total), year dummies for 2000-2006, and dummies for services, healthcare, small, education and nonprofits for the 1999-2006 period. Interpret the coefficient of icat1total and icat6total. Also interpret the coefficient of period dummies and sector dummies.

##multiple numeric and dummy vars  
  
m3 <- lm(ccat7total ~ icat1total+icat2total+icat3total+icat4total+icat5total+ icat6total+as.factor(year)+ as.factor(sector), data=df1)  
stargazer(m3,type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## -----------------------------------------------  
## icat1total 0.996\*   
## (0.552)   
##   
## icat2total 0.573   
## (0.606)   
##   
## icat3total 0.947   
## (0.609)   
##   
## icat4total 0.209   
## (0.657)   
##   
## icat5total 0.955   
## (0.626)   
##   
## icat6total 1.245\*\*   
## (0.560)   
##   
## as.factor(year)2000 11.382   
## (14.569)   
##   
## as.factor(year)2001 18.064   
## (15.717)   
##   
## as.factor(year)2002 17.704   
## (15.143)   
##   
## as.factor(year)2003 2.185   
## (14.533)   
##   
## as.factor(year)2004 -8.527   
## (14.586)   
##   
## as.factor(year)2005 -18.734   
## (14.165)   
##   
## as.factor(year)2006 -10.741   
## (14.207)   
##   
## as.factor(sector)2 12.987   
## (11.529)   
##   
## as.factor(sector)3 28.691\*\*   
## (11.454)   
##   
## as.factor(sector)4 21.050\*\*   
## (10.224)   
##   
## as.factor(sector)5 15.327   
## (9.563)   
##   
## as.factor(sector)6 2.968   
## (18.387)   
##   
## Constant -75.410\*\*   
## (31.914)   
##   
## -----------------------------------------------  
## Observations 243   
## R2 0.352   
## Adjusted R2 0.300   
## Residual Std. Error 43.539 (df = 224)   
## F Statistic 6.760\*\*\* (df = 18; 224)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

summary(m3)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat2total + icat3total +   
## icat4total + icat5total + icat6total + as.factor(year) +   
## as.factor(sector), data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -112.008 -29.775 5.163 27.789 91.650   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -75.4096 31.9141 -2.363 0.0190 \*  
## icat1total 0.9964 0.5525 1.804 0.0726 .  
## icat2total 0.5735 0.6059 0.946 0.3450   
## icat3total 0.9469 0.6089 1.555 0.1213   
## icat4total 0.2089 0.6574 0.318 0.7509   
## icat5total 0.9548 0.6261 1.525 0.1287   
## icat6total 1.2455 0.5598 2.225 0.0271 \*  
## as.factor(year)2000 11.3817 14.5693 0.781 0.4355   
## as.factor(year)2001 18.0643 15.7172 1.149 0.2516   
## as.factor(year)2002 17.7042 15.1426 1.169 0.2436   
## as.factor(year)2003 2.1849 14.5330 0.150 0.8806   
## as.factor(year)2004 -8.5269 14.5856 -0.585 0.5594   
## as.factor(year)2005 -18.7343 14.1650 -1.323 0.1873   
## as.factor(year)2006 -10.7411 14.2070 -0.756 0.4504   
## as.factor(sector)2 12.9866 11.5293 1.126 0.2612   
## as.factor(sector)3 28.6906 11.4540 2.505 0.0130 \*  
## as.factor(sector)4 21.0497 10.2237 2.059 0.0407 \*  
## as.factor(sector)5 15.3272 9.5631 1.603 0.1104   
## as.factor(sector)6 2.9682 18.3865 0.161 0.8719   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.54 on 224 degrees of freedom  
## (221 observations deleted due to missingness)  
## Multiple R-squared: 0.352, Adjusted R-squared: 0.2999   
## F-statistic: 6.76 on 18 and 224 DF, p-value: 1.785e-13

AIC(m3)

## [1] 2543.822

BIC(m3)

## [1] 2613.683

Interpretation: Co-efficient of icat1total: For the year 1999 and a company in the manufacturing sector, with every unit increase in the score of icat1total, there will be an increase of 0.996 units in the ccat7total score.

Co-efficient of icat6total: For the year 1999 and a company in the manufacturing sector, with every unit increase in the score of icat6total, there will be an increase of 1.245 units in the ccat7total score.

Co-efficient of period dummies: The year categories are not statistically significant. For a company in mfg sector, each year change, changes the ccat7total by the coefficient of the year dummy.

Co-efficient of sector dummies: Most of the sector dummies are not statistically significant. For a company in small and education sectors will have an increased score of 28.69 and 21.05 units more than the manufacturing sector score.

1. MLR with interactions among continuous variables: Now, run a regression model m4a with “ccat7total” as the response variable, six individual category totals (icat1total, icat2total, icat3total, icat4total, icat5total, icat6total), year dummies for 2000-2006, and dummies for services, healthcare, small, education and nonprofits, and the interaction between icat1total and icat6total for the 1999-2006 period. Compare this model with the nested model m3 with the same variables but without the interaction term. Then, answer the following questions:
2. Which of the two models (m3 and m4a) is better?
3. Are icat1total and icat6total important predictors of ccat7total? Why or why not?
4. How will you interpret the coefficient of icat1total, icat6total, and the interaction term?
5. Run a model m4b where you include interaction involving the mean-centered icat1total and mean-centered icat6total variable (you will continue to use icat1total and icat6total variables without mean-centering as before), instead of simply using multiplication of icat1total and icat6total as the interaction term. Interpret the coefficient of icat1total, icat6total, and the interaction terms.

m4 <- lm(ccat7total ~ icat1total+icat2total+icat3total+icat4total+icat5total+ icat6total+as.factor(year)+ as.factor(sector)+(icat1total\*icat6total), data=df1)  
stargazer(m4,type = "text")

##   
## =================================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## -------------------------------------------------  
## icat1total -2.797   
## (1.876)   
##   
## icat2total 0.419   
## (0.606)   
##   
## icat3total 0.995   
## (0.605)   
##   
## icat4total 0.046   
## (0.657)   
##   
## icat5total 0.887   
## (0.622)   
##   
## icat6total -4.024   
## (2.554)   
##   
## as.factor(year)2000 9.798   
## (14.477)   
##   
## as.factor(year)2001 18.238   
## (15.597)   
##   
## as.factor(year)2002 12.344   
## (15.239)   
##   
## as.factor(year)2003 -0.366   
## (14.472)   
##   
## as.factor(year)2004 -10.240   
## (14.497)   
##   
## as.factor(year)2005 -22.866   
## (14.192)   
##   
## as.factor(year)2006 -11.587   
## (14.104)   
##   
## as.factor(sector)2 13.520   
## (11.444)   
##   
## as.factor(sector)3 31.460\*\*\*   
## (11.442)   
##   
## as.factor(sector)4 22.177\*\*   
## (10.159)   
##   
## as.factor(sector)5 17.958\*   
## (9.571)   
##   
## as.factor(sector)6 0.975   
## (18.270)   
##   
## icat1total:icat6total 0.080\*\*   
## (0.038)   
##   
## Constant 186.766   
## (128.002)   
##   
## -------------------------------------------------  
## Observations 243   
## R2 0.365   
## Adjusted R2 0.311   
## Residual Std. Error 43.206 (df = 223)   
## F Statistic 6.739\*\*\* (df = 19; 223)   
## =================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

summary(m4)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat2total + icat3total +   
## icat4total + icat5total + icat6total + as.factor(year) +   
## as.factor(sector) + (icat1total \* icat6total), data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -125.371 -27.901 5.419 27.681 98.247   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 186.76615 128.00178 1.459 0.14595   
## icat1total -2.79721 1.87647 -1.491 0.13746   
## icat2total 0.41877 0.60575 0.691 0.49008   
## icat3total 0.99501 0.60467 1.646 0.10126   
## icat4total 0.04558 0.65698 0.069 0.94475   
## icat5total 0.88732 0.62212 1.426 0.15519   
## icat6total -4.02437 2.55404 -1.576 0.11652   
## as.factor(year)2000 9.79797 14.47715 0.677 0.49924   
## as.factor(year)2001 18.23837 15.59710 1.169 0.24351   
## as.factor(year)2002 12.34367 15.23916 0.810 0.41881   
## as.factor(year)2003 -0.36614 14.47219 -0.025 0.97984   
## as.factor(year)2004 -10.23979 14.49665 -0.706 0.48070   
## as.factor(year)2005 -22.86575 14.19184 -1.611 0.10855   
## as.factor(year)2006 -11.58702 14.10398 -0.822 0.41221   
## as.factor(sector)2 13.52038 11.44386 1.181 0.23868   
## as.factor(sector)3 31.45964 11.44155 2.750 0.00646 \*\*  
## as.factor(sector)4 22.17697 10.15943 2.183 0.03009 \*   
## as.factor(sector)5 17.95843 9.57117 1.876 0.06192 .   
## as.factor(sector)6 0.97459 18.27017 0.053 0.95751   
## icat1total:icat6total 0.08019 0.03793 2.114 0.03563 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.21 on 223 degrees of freedom  
## (221 observations deleted due to missingness)  
## Multiple R-squared: 0.3647, Adjusted R-squared: 0.3106   
## F-statistic: 6.739 on 19 and 223 DF, p-value: 6.984e-14

AIC(m4)

## [1] 2541

BIC(m4)

## [1] 2614.354

mean(df1$icat1total,na.rm = TRUE)

## [1] 58.9676

mean(df1$icat6total,na.rm = TRUE)

## [1] 41.18143

m4b <- lm(ccat7total ~ icat1total+icat2total+icat3total+icat4total+icat5total+ icat6total+as.factor(year)+ as.factor(sector)+I((icat1total-58.96)\*(icat6total-41.18)), data=df1)  
stargazer(m4b,type = "text")

##   
## ==========================================================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## --------------------------------------------------------------------------  
## icat1total 0.505   
## (0.596)   
##   
## icat2total 0.419   
## (0.606)   
##   
## icat3total 0.995   
## (0.605)   
##   
## icat4total 0.046   
## (0.657)   
##   
## icat5total 0.887   
## (0.622)   
##   
## icat6total 0.704   
## (0.612)   
##   
## as.factor(year)2000 9.798   
## (14.477)   
##   
## as.factor(year)2001 18.238   
## (15.597)   
##   
## as.factor(year)2002 12.344   
## (15.239)   
##   
## as.factor(year)2003 -0.366   
## (14.472)   
##   
## as.factor(year)2004 -10.240   
## (14.497)   
##   
## as.factor(year)2005 -22.866   
## (14.192)   
##   
## as.factor(year)2006 -11.587   
## (14.104)   
##   
## as.factor(sector)2 13.520   
## (11.444)   
##   
## as.factor(sector)3 31.460\*\*\*   
## (11.442)   
##   
## as.factor(sector)4 22.177\*\*   
## (10.159)   
##   
## as.factor(sector)5 17.958\*   
## (9.571)   
##   
## as.factor(sector)6 0.975   
## (18.270)   
##   
## I((icat1total - 58.96) \* (icat6total - 41.18)) 0.080\*\*   
## (0.038)   
##   
## Constant -7.930   
## (44.966)   
##   
## --------------------------------------------------------------------------  
## Observations 243   
## R2 0.365   
## Adjusted R2 0.311   
## Residual Std. Error 43.206 (df = 223)   
## F Statistic 6.739\*\*\* (df = 19; 223)   
## ==========================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

summary(m4b)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat2total + icat3total +   
## icat4total + icat5total + icat6total + as.factor(year) +   
## as.factor(sector) + I((icat1total - 58.96) \* (icat6total -   
## 41.18)), data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -125.371 -27.901 5.419 27.681 98.247   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -7.92954 44.96616 -0.176  
## icat1total 0.50496 0.59552 0.848  
## icat2total 0.41877 0.60575 0.691  
## icat3total 0.99501 0.60467 1.646  
## icat4total 0.04558 0.65698 0.069  
## icat5total 0.88732 0.62212 1.426  
## icat6total 0.70355 0.61185 1.150  
## as.factor(year)2000 9.79797 14.47715 0.677  
## as.factor(year)2001 18.23837 15.59710 1.169  
## as.factor(year)2002 12.34367 15.23916 0.810  
## as.factor(year)2003 -0.36614 14.47219 -0.025  
## as.factor(year)2004 -10.23979 14.49665 -0.706  
## as.factor(year)2005 -22.86575 14.19184 -1.611  
## as.factor(year)2006 -11.58702 14.10398 -0.822  
## as.factor(sector)2 13.52038 11.44386 1.181  
## as.factor(sector)3 31.45964 11.44155 2.750  
## as.factor(sector)4 22.17697 10.15943 2.183  
## as.factor(sector)5 17.95843 9.57117 1.876  
## as.factor(sector)6 0.97459 18.27017 0.053  
## I((icat1total - 58.96) \* (icat6total - 41.18)) 0.08019 0.03793 2.114  
## Pr(>|t|)   
## (Intercept) 0.86018   
## icat1total 0.39739   
## icat2total 0.49008   
## icat3total 0.10126   
## icat4total 0.94475   
## icat5total 0.15519   
## icat6total 0.25143   
## as.factor(year)2000 0.49924   
## as.factor(year)2001 0.24351   
## as.factor(year)2002 0.41881   
## as.factor(year)2003 0.97984   
## as.factor(year)2004 0.48070   
## as.factor(year)2005 0.10855   
## as.factor(year)2006 0.41221   
## as.factor(sector)2 0.23868   
## as.factor(sector)3 0.00646 \*\*  
## as.factor(sector)4 0.03009 \*   
## as.factor(sector)5 0.06192 .   
## as.factor(sector)6 0.95751   
## I((icat1total - 58.96) \* (icat6total - 41.18)) 0.03563 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.21 on 223 degrees of freedom  
## (221 observations deleted due to missingness)  
## Multiple R-squared: 0.3647, Adjusted R-squared: 0.3106   
## F-statistic: 6.739 on 19 and 223 DF, p-value: 6.984e-14

Interpretation: 4a: Model 4a renders most of the variables and the dummy variables as insignificant where as m3 has some variables as significant. Also there is no significant difference between the 2 models in terms of R squared, AIC and BIC values. So its better to conclude M3 is better fit than M4A.

1. Scores icat1total and icat6total seems to be an important predictor if the ccat7total as they remain statistically significant in most of the models and there seems to be higher fit than the other variables in the models that significantly affect the ccat7total scores.

c.Co-efficient of icat1total: For the year 1999 and a company in the manufacturing sector, indicates that one unit increase in icat1total score is not statistically significantly associated with an average decrease of 2.78 units in ccat7total score.

Co-efficient of icat6total: For the year 1999 and a company in the manufacturing sector, indicates that one unit increase in icat6total score is not statistically significantly associated with an average decrease of 4.02 units in ccat7total score.

Co-efficient of the interaction term: For the year 1999 and a company in the manufacturing sector,indicates that the effect of a unit increase for both icat1total and icat6total is associated with an average decrease of 6 in ccat7total score.

d.Co-efficient of icat1total: For the year 1999 and a company in the manufacturing sector, indicates that one unit increase in icat1total score is not statistically significantly associated with an average increase of 0.5 units in ccat7total score.

Co-efficient of icat6total: For the year 1999 and a company in the manufacturing sector, indicates that one unit increase in icat6total score is not statistically significantly associated with an average increase of 0.7 units in ccat7total score.

Co-efficient of the interaction term: For the year 1999 and a company in the manufacturing sector,indicates that the effect of a unit increase for both icat1total and icat6total is associated with an average increase of 0.68 units in ccat7total score.

1. MLR with interactions among continuous and binary (nonprofits sector) variable: Consider a model m5 only for 1999-2006 period, with “ccat7total” as the response variable. Examine if nonprofits dummy variable has a significant interaction effect with icat6total. You will also have six individual category totals (icat1total, icat2total, icat3total, icat4total, icat5total, icat6total), year dummies for 2000-2006, and dummies for services, healthcare, small, education and nonprofits in the model. How do you interpret the interaction term. Comment on differences in coefficients from model m3.

#create binary variable for nonprofit sector  
  
  
#interactions among continuous and binary (nonprofit)  
m5 <- lm(ccat7total ~ icat1total + icat2total + icat3total + icat4total +   
 icat5total + icat6total + as.factor(year) + as.factor(sector)+  
 I(icat6total\*(sector==6)), data=df1)  
summary(m5)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat2total + icat3total +   
## icat4total + icat5total + icat6total + as.factor(year) +   
## as.factor(sector) + I(icat6total \* (sector == 6)), data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -108.64 -30.34 5.74 28.11 85.19   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -59.3087 32.6735 -1.815 0.0708 .  
## icat1total 1.0917 0.5507 1.982 0.0487 \*  
## icat2total 0.5501 0.6019 0.914 0.3618   
## icat3total 0.7594 0.6117 1.241 0.2158   
## icat4total 0.1637 0.6533 0.251 0.8023   
## icat5total 0.9952 0.6221 1.600 0.1111   
## icat6total 1.0198 0.5670 1.799 0.0734 .  
## as.factor(year)2000 10.7629 14.4723 0.744 0.4578   
## as.factor(year)2001 16.8614 15.6203 1.079 0.2816   
## as.factor(year)2002 16.3062 15.0542 1.083 0.2799   
## as.factor(year)2003 0.7670 14.4500 0.053 0.9577   
## as.factor(year)2004 -8.6025 14.4853 -0.594 0.5532   
## as.factor(year)2005 -18.0289 14.0719 -1.281 0.2015   
## as.factor(year)2006 -9.9463 14.1147 -0.705 0.4817   
## as.factor(sector)2 12.5128 11.4524 1.093 0.2758   
## as.factor(sector)3 27.2686 11.3968 2.393 0.0176 \*  
## as.factor(sector)4 18.9827 10.2044 1.860 0.0642 .  
## as.factor(sector)5 13.8191 9.5263 1.451 0.1483   
## as.factor(sector)6 -137.2035 71.4782 -1.920 0.0562 .  
## I(icat6total \* (sector == 6)) 3.0116 1.4848 2.028 0.0437 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.24 on 223 degrees of freedom  
## (221 observations deleted due to missingness)  
## Multiple R-squared: 0.3638, Adjusted R-squared: 0.3095   
## F-statistic: 6.71 on 19 and 223 DF, p-value: 8.126e-14

stargazer(m5,type = "text")

##   
## =========================================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## ---------------------------------------------------------  
## icat1total 1.092\*\*   
## (0.551)   
##   
## icat2total 0.550   
## (0.602)   
##   
## icat3total 0.759   
## (0.612)   
##   
## icat4total 0.164   
## (0.653)   
##   
## icat5total 0.995   
## (0.622)   
##   
## icat6total 1.020\*   
## (0.567)   
##   
## as.factor(year)2000 10.763   
## (14.472)   
##   
## as.factor(year)2001 16.861   
## (15.620)   
##   
## as.factor(year)2002 16.306   
## (15.054)   
##   
## as.factor(year)2003 0.767   
## (14.450)   
##   
## as.factor(year)2004 -8.602   
## (14.485)   
##   
## as.factor(year)2005 -18.029   
## (14.072)   
##   
## as.factor(year)2006 -9.946   
## (14.115)   
##   
## as.factor(sector)2 12.513   
## (11.452)   
##   
## as.factor(sector)3 27.269\*\*   
## (11.397)   
##   
## as.factor(sector)4 18.983\*   
## (10.204)   
##   
## as.factor(sector)5 13.819   
## (9.526)   
##   
## as.factor(sector)6 -137.204\*   
## (71.478)   
##   
## I(icat6total \* (sector == 6)) 3.012\*\*   
## (1.485)   
##   
## Constant -59.309\*   
## (32.673)   
##   
## ---------------------------------------------------------  
## Observations 243   
## R2 0.364   
## Adjusted R2 0.310   
## Residual Std. Error 43.240 (df = 223)   
## F Statistic 6.710\*\*\* (df = 19; 223)   
## =========================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Interpretation: With keeping the nonprofit sector as binary and its interaction with the icat6total score, there is a significant change in the coefficients of all the variables as compared to M3. The coefficient of the interaction terms signifies that for a company in the non profit sector it has statistically significant effect by an unit increase in the icat6total, decreases the ccat7total by 133 units.

1. MLR with interactions among continuous and binary (year 2006) variable: Consider a model m6 only for 1999-2006 period. Examine if year 2006 interacts with icat6total. You will also have six individual category totals (icat1total, icat2total, icat3total, icat4total, icat5total, icat6total), year dummies for 2000-2006, and dummies for services, healthcare, small, education and nonprofits in the model.  
   How do you interpret the interaction term. Comment on differences in coefficients from model m3.

#interactions among continuous and binary (2006)  
m6 <- lm(ccat7total ~ icat1total + icat2total + icat3total + icat4total +   
 icat5total + icat6total + as.factor(year) + as.factor(sector) +  
 I(icat6total\*(year==2006)), data=df1)  
summary(m6)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat2total + icat3total +   
## icat4total + icat5total + icat6total + as.factor(year) +   
## as.factor(sector) + I(icat6total \* (year == 2006)), data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -106.913 -28.969 7.125 27.489 83.332   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -49.3651 34.1507 -1.446 0.1497   
## icat1total 1.0452 0.5491 1.903 0.0583 .  
## icat2total 0.5417 0.6019 0.900 0.3691   
## icat3total 0.9047 0.6050 1.496 0.1362   
## icat4total 0.1006 0.6550 0.154 0.8781   
## icat5total 0.8758 0.6229 1.406 0.1611   
## icat6total 0.9531 0.5740 1.661 0.0982 .  
## as.factor(year)2000 10.1371 14.4795 0.700 0.4846   
## as.factor(year)2001 15.8822 15.6430 1.015 0.3111   
## as.factor(year)2002 14.7827 15.1037 0.979 0.3288   
## as.factor(year)2003 -0.5528 14.4927 -0.038 0.9696   
## as.factor(year)2004 -9.1417 14.4861 -0.631 0.5286   
## as.factor(year)2005 -17.4991 14.0783 -1.243 0.2152   
## as.factor(year)2006 -122.3979 56.3679 -2.171 0.0310 \*  
## as.factor(sector)2 11.0101 11.4889 0.958 0.3389   
## as.factor(sector)3 26.1199 11.4426 2.283 0.0234 \*  
## as.factor(sector)4 18.4711 10.2297 1.806 0.0723 .  
## as.factor(sector)5 12.5378 9.5932 1.307 0.1926   
## as.factor(sector)6 8.3157 18.4433 0.451 0.6525   
## I(icat6total \* (year == 2006)) 2.2573 1.1033 2.046 0.0419 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.23 on 223 degrees of freedom  
## (221 observations deleted due to missingness)  
## Multiple R-squared: 0.364, Adjusted R-squared: 0.3098   
## F-statistic: 6.716 on 19 and 223 DF, p-value: 7.88e-14

stargazer(m6,type = "text")

##   
## ==========================================================  
## Dependent variable:   
## ---------------------------  
## ccat7total   
## ----------------------------------------------------------  
## icat1total 1.045\*   
## (0.549)   
##   
## icat2total 0.542   
## (0.602)   
##   
## icat3total 0.905   
## (0.605)   
##   
## icat4total 0.101   
## (0.655)   
##   
## icat5total 0.876   
## (0.623)   
##   
## icat6total 0.953\*   
## (0.574)   
##   
## as.factor(year)2000 10.137   
## (14.480)   
##   
## as.factor(year)2001 15.882   
## (15.643)   
##   
## as.factor(year)2002 14.783   
## (15.104)   
##   
## as.factor(year)2003 -0.553   
## (14.493)   
##   
## as.factor(year)2004 -9.142   
## (14.486)   
##   
## as.factor(year)2005 -17.499   
## (14.078)   
##   
## as.factor(year)2006 -122.398\*\*   
## (56.368)   
##   
## as.factor(sector)2 11.010   
## (11.489)   
##   
## as.factor(sector)3 26.120\*\*   
## (11.443)   
##   
## as.factor(sector)4 18.471\*   
## (10.230)   
##   
## as.factor(sector)5 12.538   
## (9.593)   
##   
## as.factor(sector)6 8.316   
## (18.443)   
##   
## I(icat6total \* (year == 2006)) 2.257\*\*   
## (1.103)   
##   
## Constant -49.365   
## (34.151)   
##   
## ----------------------------------------------------------  
## Observations 243   
## R2 0.364   
## Adjusted R2 0.310   
## Residual Std. Error 43.233 (df = 223)   
## F Statistic 6.716\*\*\* (df = 19; 223)   
## ==========================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Interpretation: With keeping the year 2006 as binary and its interaction with the icat6total score, there is a significant change in the coefficients of all the variables as compared to M3. The coefficient of the interaction terms signifies that for any company in the year 2006 it has statistically significant effect by an unit increase in the icat6total, decreases the ccat7total by 119 units.

1. Create a table listing the multiple R-squared, adjusted R-squared, AIC, and BIC of the models above. Also, create a table listing the variance inflation factor (VIF) of the models above. Which is the best model and why? Write your key observations based on this Table.

a=data.frame("Model 1",summary(m1)$r.squared,summary(m1)$adj.r.squared,AIC(m1),BIC(m1))  
names(a)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
b=data.frame("Model 2",summary(m2)$r.squared,summary(m2)$adj.r.squared,AIC(m2),BIC(m2))  
names(b)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
c=data.frame("Model 3",summary(m3)$r.squared,summary(m3)$adj.r.squared,AIC(m3),BIC(m3))  
names(c)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
d=data.frame("Model 4",summary(m4)$r.squared,summary(m4)$adj.r.squared,AIC(m4),BIC(m4))  
names(d)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
d\_b=data.frame("Model 4b",summary(m4b)$r.squared,summary(m4b)$adj.r.squared,AIC(m4b),BIC(m4b))  
names(d\_b)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
e=data.frame("Model 5",summary(m5)$r.squared,summary(m5)$adj.r.squared,AIC(m5),BIC(m5))  
names(e)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
f=data.frame("Model 6",summary(m6)$r.squared,summary(m6)$adj.r.squared,AIC(m6),BIC(m6))  
names(f)=c("Model number","R squared","Adj. Rsquared","AIC","BIC")  
  
comparison=rbind(a,b,c,d,d\_b,e,f)  
  
stargazer(comparison,type="text",summary=FALSE,digits=2)

##   
## ========================================================  
## Model number R squared Adj. Rsquared AIC BIC   
## --------------------------------------------------------  
## 1 Model 1 0.11 0.11 2,586.37 2,596.85  
## 2 Model 2 0.29 0.27 2,542.85 2,570.80  
## 3 Model 3 0.35 0.30 2,543.82 2,613.68  
## 4 Model 4 0.36 0.31 2,541.00 2,614.35  
## 5 Model 4b 0.36 0.31 2,541.00 2,614.35  
## 6 Model 5 0.36 0.31 2,541.38 2,614.73  
## 7 Model 6 0.36 0.31 2,541.30 2,614.66  
## --------------------------------------------------------

stargazer(vif(m2),type="text",digits=2)

##   
## =================================================================  
## icat1total icat2total icat3total icat4total icat5total icat6total  
## -----------------------------------------------------------------  
## 2.70 2.28 2.19 1.99 1.90 1.97   
## -----------------------------------------------------------------

stargazer(vif(m3),vif(m4),vif(m4b),vif(m5),vif(m6),type="text",digits=2)

##   
## ==================================  
## GVIF Df GVIFDf))  
## ----------------------------------  
## icat1total 3.43 1 1.85   
## icat2total 2.43 1 1.56   
## icat3total 2.41 1 1.55   
## icat4total 2.85 1 1.69   
## icat5total 2.28 1 1.51   
## icat6total 2.38 1 1.54   
## as.factor(year) 3.56 7 1.09   
## as.factor(sector) 2.09 5 1.08   
## ----------------------------------  
##   
## ========================================  
## GVIF Df GVIFDf))  
## ----------------------------------------  
## icat1total 40.20 1 6.34   
## icat2total 2.46 1 1.57   
## icat3total 2.41 1 1.55   
## icat4total 2.89 1 1.70   
## icat5total 2.29 1 1.51   
## icat6total 50.40 1 7.10   
## as.factor(year) 3.84 7 1.10   
## as.factor(sector) 2.17 5 1.08   
## icat1total:icat6total 140.32 1 11.85   
## ----------------------------------------  
##   
## ===============================================================  
## GVIF Df GVIFDf))  
## ---------------------------------------------------------------  
## icat1total 4.05 1 2.01   
## icat2total 2.46 1 1.57   
## icat3total 2.41 1 1.55   
## icat4total 2.89 1 1.70   
## icat5total 2.29 1 1.51   
## icat6total 2.89 1 1.70   
## as.factor(year) 3.84 7 1.10   
## as.factor(sector) 2.17 5 1.08   
## I((icat1total - 58.96) \* (icat6total - 41.18)) 3.59 1 1.90   
## ---------------------------------------------------------------  
##   
## ===============================================  
## GVIF Df GVIFDf))  
## -----------------------------------------------  
## icat1total 3.46 1 1.86   
## icat2total 2.43 1 1.56   
## icat3total 2.46 1 1.57   
## icat4total 2.86 1 1.69   
## icat5total 2.29 1 1.51   
## icat6total 2.48 1 1.57   
## as.factor(year) 3.62 7 1.10   
## as.factor(sector) 40.81 5 1.45   
## I(icat6total \* (sector == 6)) 24.43 1 4.94   
## -----------------------------------------------  
##   
## =================================================  
## GVIF Df GVIFDf))  
## -------------------------------------------------  
## icat1total 3.44 1 1.85   
## icat2total 2.43 1 1.56   
## icat3total 2.41 1 1.55   
## icat4total 2.87 1 1.69   
## icat5total 2.29 1 1.51   
## icat6total 2.54 1 1.59   
## as.factor(year) 161.70 7 1.44   
## as.factor(sector) 2.26 5 1.09   
## I(icat6total \* (year == 2006)) 62.36 1 7.90   
## -------------------------------------------------

Interpretation: Based on the comparison chart involving the R squared, adj R squared, AIC, BIC and VIF values model 4 and 4b seem to be the best models with the least AIC and BIC values and VIF values less than 5. The R squared values at 36% and adj R squared at 31%. But considering the VIF value for the interaction term greater than 10 for model 4b, model 4 seem to the best model for the model fitting.

1. Create a table of summary stats (N, mean, sd, min, max) and correlations for ccat7total, icat1total, icat2total, icat3total, icat4total, icat5total, icat6total, and binary variables for sectors for the 1999-2006 period. The output should be neatly formatted in a Table and the values should be rounded to 2 decimal places. The correlations should indicate significance levels. The output should be neatly formatted in a Table and the values should be rounded to 2 decimal places.

library(plyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(reshape2)  
  
#Summary Stats  
  
  
df1$sector=as.factor(df1$sector)  
  
dfm <- melt(df1, id.vars = c("sector"),na.rm=TRUE)  
  
dis=lapply(list(.(sector, variable)),   
 ddply, .data = dfm, .fun = summarise,   
 mean = mean(value),   
 sd = sd(value),  
 N=length(value),  
 se=sd/sqrt(N))  
  
stargazer(dis,type="text",summary=FALSE,digits=2)

##   
## =============================================  
## sector variable mean sd N se   
## ---------------------------------------------  
## 1 1 ccat7total 189.56 52.05 36 8.67   
## 2 1 icat1total 63.28 14.86 54 2.02   
## 3 1 icat2total 38.93 11.36 54 1.55   
## 4 1 icat3total 44.04 10.41 54 1.42   
## 5 1 icat4total 45.24 10.13 54 1.38   
## 6 1 icat5total 43.11 8.22 54 1.12   
## 7 1 icat6total 45.57 10.69 54 1.46   
## 8 1 year 2,001.89 1.91 55 0.26   
## 9 2 ccat7total 195.92 54.45 26 10.68  
## 10 2 icat1total 62.02 14.37 46 2.12   
## 11 2 icat2total 38.50 11.63 46 1.71   
## 12 2 icat3total 42.89 10.17 46 1.50   
## 13 2 icat4total 43.50 12.64 46 1.86   
## 14 2 icat5total 41.43 8.74 46 1.29   
## 15 2 icat6total 43.78 10.33 46 1.52   
## 16 2 year 2,002.11 2.44 46 0.36   
## 17 3 ccat7total 216.13 49.66 30 9.07   
## 18 3 icat1total 55.44 18.74 77 2.14   
## 19 3 icat2total 34 12.00 77 1.37   
## 20 3 icat3total 39.51 11.86 77 1.35   
## 21 3 icat4total 37.61 12.38 77 1.41   
## 22 3 icat5total 38.13 10.22 77 1.16   
## 23 3 icat6total 37 11.67 77 1.33   
## 24 3 year 2,002.27 2.26 77 0.26   
## 25 4 ccat7total 186.94 46.52 52 6.45   
## 26 4 icat1total 54.53 15.27 115 1.42   
## 27 4 icat2total 34.03 11.72 115 1.09   
## 28 4 icat3total 36.21 10.01 115 0.93   
## 29 4 icat4total 38.09 12.47 115 1.16   
## 30 4 icat5total 37.42 9.80 115 0.91   
## 31 4 icat6total 38.58 11.15 115 1.04   
## 32 4 year 2,002.74 2.34 115 0.22   
## 33 5 ccat7total 188.27 48.63 89 5.15   
## 34 5 icat1total 61.37 12.09 161 0.95   
## 35 5 icat2total 38.84 9.70 161 0.76   
## 36 5 icat3total 42.35 8.81 161 0.69   
## 37 5 icat4total 45.20 10.01 161 0.79   
## 38 5 icat5total 42.45 7.96 161 0.63   
## 39 5 icat6total 42.56 9.66 161 0.76   
## 40 5 year 2,003.81 2.12 161 0.17   
## 41 6 ccat7total 151.10 81.18 10 25.67  
## 42 6 icat1total 61.20 10.36 10 3.28   
## 43 6 icat2total 42.80 8.34 10 2.64   
## 44 6 icat3total 40.90 10.14 10 3.21   
## 45 6 icat4total 48.50 9.08 10 2.87   
## 46 6 icat5total 45.40 8.33 10 2.63   
## 47 6 icat6total 45.40 10.31 10 3.26   
## 48 6 year 2,006 0 10 0   
## ---------------------------------------------

#Correlation  
  
df1 = na.omit(df1)  
df1 <- df1 %>% mutate\_if(is.character, as.numeric)  
  
  
library(xtable)  
  
corstars <-function(x, method=c("pearson", "spearman"), removeTriangle=c("upper", "lower"),  
 result=c("none", "html", "latex")){  
 #Compute correlation matrix  
 require(Hmisc)  
 x <- as.matrix(x)  
 correlation\_matrix<-rcorr(x, type=method[1])  
 R <- correlation\_matrix$r # Matrix of correlation coeficients  
 p <- correlation\_matrix$P # Matrix of p-value   
   
 ## Define notions for significance levels; spacing is important.  
 mystars <- ifelse(p < .0001, "\*", ifelse(p < .001, "\* ", ifelse(p < .01, "\* ", ifelse(p < .05, "\* ", " "))))  
   
 ## trunctuate the correlation matrix to two decimal  
 R <- format(round(cbind(rep(-1.11, ncol(x)), R), 2))[,-1]  
   
   
 Rnew <- matrix(paste(R, mystars, sep=""), ncol=ncol(x))  
 diag(Rnew) <- paste(diag(R), " ", sep="")  
 rownames(Rnew) <- colnames(x)  
 colnames(Rnew) <- paste(colnames(x), "", sep="")  
   
 ## remove upper triangle of correlation matrix  
 if(removeTriangle[1]=="upper"){  
 Rnew <- as.matrix(Rnew)  
 Rnew[upper.tri(Rnew, diag = TRUE)] <- ""  
 Rnew <- as.data.frame(Rnew)  
 }  
   
 ## remove lower triangle of correlation matrix  
 else if(removeTriangle[1]=="lower"){  
 Rnew <- as.matrix(Rnew)  
 Rnew[lower.tri(Rnew, diag = TRUE)] <- ""  
 Rnew <- as.data.frame(Rnew)  
 }  
   
 ## remove last column and return the correlation matrix  
 Rnew <- cbind(Rnew[1:length(Rnew)-1])  
 if (result[1]=="none") return(Rnew)  
 else{  
 if(result[1]=="html") print(xtable(Rnew), type="html")  
 else print(xtable(Rnew), type="latex")   
 }  
}   
corr=corstars(df1, result="none") # \*\* p < .0001,\*\*p < .001,\* p < .01, \* p < .05

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:xtable':  
##   
## label, label<-

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:plyr':  
##   
## is.discrete, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

stargazer(corr, type = "text", title="Correlations",   
 summary=FALSE)

##   
## Correlations  
## =============================================================================================  
## ccat7total icat1total icat2total icat3total icat4total icat5total icat6total year   
## ---------------------------------------------------------------------------------------------  
## ccat7total   
## icat1total 0.51\*   
## icat2total 0.41\* 0.68\*   
## icat3total 0.43\* 0.67\* 0.56\*   
## icat4total 0.33\* 0.56\* 0.59\* 0.59\*   
## icat5total 0.38\* 0.61\* 0.52\* 0.57\* 0.51\*   
## icat6total 0.40\* 0.58\* 0.60\* 0.46\* 0.57\* 0.55\*   
## year -0.14\* -0.06 0.13\* 0.09 0.33\* 0.18\* 0.03   
## sector -0.10 -0.14\* -0.02 -0.17\* 0.00 0.02 -0.15\* 0.40\*  
## ---------------------------------------------------------------------------------------------

cat("\n\*\* p < .0001,\*\*p < .001,\* p < .01, \* p < .05")

##   
## \*\* p < .0001,\*\*p < .001,\* p < .01, \* p < .05

1. This is an open ended question and will require you to use your own creativity, imagination and what you know about organizations and management. Come up with a model m7 that you think is best in explaining ccat7total for the 1999-2006 period only. Make sure that you have reasonable VIF values for your “best” model. Make a case why this is the best model that makes sense based on your intuition, your own theories and statistical/econometric considerations. Let this be your answer: it is unlikely that two students in a class will have exactly same “best” model and same explanation for choosing that so I really expect to see each student providing a different answer and be surprised if two of you come up with the same answer by sheer chance! This question will have more weight than others.

m0 <- lm(ccat7total ~.,  
 data = df1)  
  
best\_mod <- step(m0)

## Start: AIC=1847.54  
## ccat7total ~ icat1total + icat2total + icat3total + icat4total +   
## icat5total + icat6total + year + sector  
##   
## Df Sum of Sq RSS AIC  
## - sector 5 12942.3 450558 1844.6  
## - icat2total 1 1083.9 438699 1846.1  
## - icat4total 1 1736.7 439352 1846.5  
## - icat5total 1 2085.8 439701 1846.7  
## <none> 437615 1847.5  
## - icat6total 1 5521.0 443136 1848.6  
## - icat3total 1 5612.9 443228 1848.6  
## - icat1total 1 6264.0 443879 1849.0  
## - year 1 14611.4 452227 1853.5  
##   
## Step: AIC=1844.62  
## ccat7total ~ icat1total + icat2total + icat3total + icat4total +   
## icat5total + icat6total + year  
##   
## Df Sum of Sq RSS AIC  
## - icat4total 1 759.1 451317 1843.0  
## - icat2total 1 1752.9 452311 1843.6  
## - icat5total 1 2118.1 452676 1843.8  
## - icat6total 1 3435.3 453993 1844.5  
## <none> 450558 1844.6  
## - icat3total 1 4922.4 455480 1845.3  
## - icat1total 1 10472.3 461030 1848.2  
## - year 1 16286.1 466844 1851.2  
##   
## Step: AIC=1843.03  
## ccat7total ~ icat1total + icat2total + icat3total + icat5total +   
## icat6total + year  
##   
## Df Sum of Sq RSS AIC  
## - icat5total 1 2094.3 453411 1842.2  
## - icat2total 1 2112.2 453429 1842.2  
## <none> 451317 1843.0  
## - icat6total 1 4930.4 456247 1843.7  
## - icat3total 1 6504.8 457822 1844.5  
## - icat1total 1 11436.6 462753 1847.1  
## - year 1 16084.6 467401 1849.5  
##   
## Step: AIC=1842.15  
## ccat7total ~ icat1total + icat2total + icat3total + icat6total +   
## year  
##   
## Df Sum of Sq RSS AIC  
## - icat2total 1 2092.0 455503 1841.3  
## <none> 453411 1842.2  
## - icat6total 1 7200.3 460611 1844.0  
## - icat3total 1 8678.6 462090 1844.8  
## - year 1 14299.0 467710 1847.7  
## - icat1total 1 15421.2 468832 1848.3  
##   
## Step: AIC=1841.27  
## ccat7total ~ icat1total + icat3total + icat6total + year  
##   
## Df Sum of Sq RSS AIC  
## <none> 455503 1841.3  
## - icat3total 1 10080 465583 1844.6  
## - icat6total 1 11138 466641 1845.1  
## - year 1 12707 468210 1846.0  
## - icat1total 1 24247 479750 1851.9

summary(best\_mod)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat3total + icat6total +   
## year, data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -119.85 -29.48 4.65 31.16 95.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6558.7261 2557.1028 2.565 0.010935 \*   
## icat1total 1.6089 0.4520 3.559 0.000449 \*\*\*  
## icat3total 1.2455 0.5427 2.295 0.022610 \*   
## icat6total 1.0900 0.4518 2.412 0.016606 \*   
## year -3.2895 1.2766 -2.577 0.010578 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.75 on 238 degrees of freedom  
## Multiple R-squared: 0.3049, Adjusted R-squared: 0.2932   
## F-statistic: 26.1 on 4 and 238 DF, p-value: < 2.2e-16

all\_vifs <- car::vif(best\_mod)  
print(all\_vifs)

## icat1total icat3total icat6total year   
## 2.275211 1.893195 1.538628 1.040639

m7 <- lm(ccat7total ~ icat1total + icat3total + icat6total+ as.factor(sector)+as.factor(year),data = df1)  
summary(m7)

##   
## Call:  
## lm(formula = ccat7total ~ icat1total + icat3total + icat6total +   
## as.factor(sector) + as.factor(year), data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -119.122 -28.431 5.757 27.908 90.302   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -62.72482 30.79552 -2.037 0.04283 \*   
## icat1total 1.45132 0.48502 2.992 0.00308 \*\*  
## icat3total 1.31662 0.56432 2.333 0.02052 \*   
## icat6total 1.60967 0.48855 3.295 0.00114 \*\*  
## as.factor(sector)2 13.55243 11.43277 1.185 0.23710   
## as.factor(sector)3 28.32628 11.38389 2.488 0.01356 \*   
## as.factor(sector)4 22.41656 10.15659 2.207 0.02831 \*   
## as.factor(sector)5 16.65056 9.50624 1.752 0.08120 .   
## as.factor(sector)6 6.05421 18.29170 0.331 0.74096   
## as.factor(year)2000 7.04112 14.12511 0.498 0.61863   
## as.factor(year)2001 17.62918 14.30015 1.233 0.21893   
## as.factor(year)2002 14.03016 14.05368 0.998 0.31918   
## as.factor(year)2003 0.03384 13.38032 0.003 0.99798   
## as.factor(year)2004 -8.04826 13.21655 -0.609 0.54316   
## as.factor(year)2005 -16.84406 13.15014 -1.281 0.20153   
## as.factor(year)2006 -7.97888 13.09092 -0.609 0.54280   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 43.58 on 227 degrees of freedom  
## Multiple R-squared: 0.3422, Adjusted R-squared: 0.2987   
## F-statistic: 7.872 on 15 and 227 DF, p-value: 3.573e-14

AIC(m7)

## [1] 2541.478

BIC(m7)

## [1] 2600.86

vif(m7)

## GVIF Df GVIF^(1/(2\*Df))  
## icat1total 2.639993 1 1.624806  
## icat3total 2.062825 1 1.436254  
## icat6total 1.812921 1 1.346448  
## as.factor(sector) 1.967960 5 1.070044  
## as.factor(year) 2.208118 7 1.058213

To come up with the best model, I first did the stepwise regression to compare the variables that are most significant in the kitchen sink model. By analyzing the AIC of the models, it is understood that the variables with highest significance in influencing the ccat7total are icat1,icat3 and icat6 along with the sectors as factors. This means that the variables such as Leadership, customer focus and the process management plays a higher significance in the ccat7total scores though it is commonly known that other variables such as strategy, information analysis and the workforce focus is also equally important in the determination of the final ccat7total scores. By running the model, we can find that the adj R square is at 30% and has all 3 variables significant. The AIC values are the least for this model and also has the least VIF values within 2.

In Management point of view, not focusing on the other variables might significantly affect the final results, but according to this model, we can derive that focusing more on the variables such as Leadership, customer focus and the process management will fetch a higher score in the final result than the focusing on the other 3 variables.