Analysis 4 – Baldridge Data
Yagna Venkitasamy

Part 1: Panel Models and Instrumental Variable Estimation

For this assignment, please use the same dataset that was provided to you via Canvas for the mid-term exam. To answer the questions, please use R script or R markdown 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(dplyr)

##
## Attaching package: 'dplyr'

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

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

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

library(moments)
library(ggplot2)
library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(readxl)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

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

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following object is masked from 'package:psych':
##
##     describe

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

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

library(corrplot)

## corrplot 0.84 loaded

library(tidyr)
library(effects)

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

## Use the command
##     lattice::trellis.par.set(effectsTheme())
##   to customize lattice options for effects plots.
## See ?effectsTheme for details.
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(Rcpp)
library(gvlma)
library(plm)
```

```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sem)
library(systemfit)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Please cite the 'systemfit' package as:
## Arne Henningsen and Jeff D. Hamann (2007). systemfit: A Package for
```

```
Estimating Systems of Simultaneous Equations in R. Journal of Statistical
Software 23(4), 1-40. http://www.jstatsoft.org/v23/i04/.
##
## If you have questions, suggestions, or comments regarding the 'systemfit'
package, please use a forum or 'tracker' at systemfit's R-Forge site:
## https://r-forge.r-project.org/projects/systemfit/
```

```
library(lme4)
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:rio':
##
##     factorize
```

```
library(ggplot2)
library(car)
library(AER)
```

```
## Loading required package: sandwich
```

```
library(plm)

rm(list = ls())
library(readxl)
setwd("C:/Users/yagna/Documents/R/R workings")
df <- import("Balridge_data_prep_25Jan2020.csv")

colnames(df) <- tolower(colnames(df))
df$period <- ifelse(df$year<1995,0,
                        ifelse(df$year>=1995&df$year<=1998, 1,
                               ifelse(df$year>=1999&df$year<=2006,2,NA)))
```

1. First run an OLS pooled regression of y1pccust on inf1pcma +ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc; and include sector dummies (with manufacturing as the omitted or base group) and period dummies (you will need to create a period variable that divides 1990-2006 into 3 periods: before 1995, 1995-1998, and 1999-2006, make before 1995 as the base period).

```
#Pooled OLS
#Removes missing values
df<-df[!(df$permanentid==""),]
#remove non-profit sector
df<-df[!(df$sector==6),]

#OLS using lm
m1=lm(y1pccust ~
inf1pcma+as.factor(sector)+as.factor(period)+ld1pcla+st1pcdev+cs1pccmk+
          hr3pcsat+pm1pcvc, data=df)

#OLS Pooled Regression using plm
```

```
m2 <- plm(y1pccust ~ inf1pcma + as.factor(period)+as.factor(sector)+
           ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc, data = df,
model = "pooling")
stargazer(m2,type='text',summary=FALSE)
```

```
## 
## ==============================================
##                         Dependent variable:   
##                     ---------------------------
##                              y1pccust         
## ----------------------------------------------
## inf1pcma                    0.256***          
##                              (0.061)          
##                                               
## as.factor(period)1           -3.134*          
##                              (1.704)          
##                                               
## as.factor(period)2          -6.147***         
##                              (1.582)          
##                                               
## as.factor(sector)2            0.494           
##                              (1.733)          
##                                               
## as.factor(sector)3            2.093           
##                              (1.483)          
##                                               
## as.factor(sector)4          20.622***         
##                              (2.159)          
##                                               
## as.factor(sector)5           4.630**          
##                              (1.840)          
##                                               
## ld1pcla                       0.082           
##                              (0.068)          
##                                               
## st1pcdev                    0.246***          
##                              (0.059)          
##                                               
## cs1pccmk                    0.273***          
##                              (0.063)          
##                                               
## hr3pcsat                      0.030           
##                              (0.052)          
##                                               
## pm1pcvc                      0.144**          
##                              (0.063)          
##                                               
## Constant                    -6.877**          
##                              (2.768)          
## 
```

```
## ----------------------------------------------
## Observations                   537
## R2                            0.556
## Adjusted R2                   0.546
## F Statistic         54.640*** (df = 12; 524)
## ==============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

2. Next, run First Difference (FD), Between, Fixed Effects (FE) and Random Effects (RE) models and show OLS, FD, BE, FE and RE results side-by-side using stargazer. You will need to use plm library for running these models.

```
#First Difference

df.p <- pdata.frame(df, index=c("permanentid","year"))
pdim(df.p)

## Unbalanced Panel: n = 199, T = 2-9, N = 537

m3 <- plm(y1pccust ~ inf1pcma + as.factor(period)+as.factor(sector)+
           ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc, data = df.p,
model = "fd")

#Between

m4 <- plm(y1pccust ~ inf1pcma + as.factor(period)+as.factor(sector)+
            ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc, data = df.p,
model = "between")

#Within
m5 <- plm(y1pccust ~ inf1pcma + as.factor(period)+as.factor(sector)+
            ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc, data = df.p,
model = "within")

#Random
m6 <- plm(y1pccust ~ inf1pcma + as.factor(period)+as.factor(sector)+
            ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc, data = df.p,
model = "random")

#displays
stargazer(m2, m3, m4, m5, m6, type = 'text', column.labels=c("OLS","FD",
"BW","FE","RE") , digits = 2)

##
##
=============================================================================
===============================================
##                                                                Dependent
variable:
##                    --------------------------------------------------------
--------------------------------------------------
```

```
##
y1pccust
##                              OLS                     FD
BW                      FE               RE
##                              (1)                     (2)
(3)                     (4)              (5)
## --------------------------------------------------------------------------
--------------------------------------------------
## inf1pcma                   0.26***                 0.47***
0.13                   0.39***          0.32***
##                            (0.06)                  (0.07)
(0.12)                  (0.06)           (0.06)
##
## as.factor(period)1         -3.13*                   -3.20
-3.72                   -2.11           -2.93*
##                            (1.70)                  (2.47)
(3.03)                  (2.06)           (1.61)
##
## as.factor(period)2        -6.15***                  -4.47
-6.70***                  -2.83          -5.56***
##                            (1.58)                  (3.45)
(2.54)                  (2.72)           (1.71)
##
## as.factor(sector)2          0.49                    0.50
0.38                    0.69             0.17
##                            (1.73)                  (10.04)
(2.39)                  (9.68)           (2.26)
##
## as.factor(sector)3          2.09                    -0.88
2.57                    -0.46            0.74
##                            (1.48)                  (5.06)
(2.17)                  (4.03)           (1.75)
##
## as.factor(sector)4        20.62***
22.84***                                 19.62***
##                            (2.16)
(3.22)                                   (2.69)
##
## as.factor(sector)5         4.63**
5.66**                                    2.77
##                            (1.84)
(2.75)                                   (2.27)
##
## ld1pcla                     0.08                    -0.05
0.14                    0.003            0.05
##                            (0.07)                  (0.07)
(0.13)                  (0.07)           (0.06)
##
## st1pcdev                   0.25***                 0.16**
0.23**                  0.20***          0.24***
```

```
##                            (0.06)                  (0.06)
(0.11)                  (0.06)           (0.05)
##
## cs1pccmk                   0.27***                  0.01
0.45***                  0.04            0.16***
##                            (0.06)                  (0.07)
(0.12)                  (0.06)           (0.06)
##
## hr3pcsat                    0.03                    0.05
0.02                    0.01             0.01
##                            (0.05)                  (0.06)
(0.09)                  (0.06)           (0.05)
##
## pm1pcvc                    0.14**                   0.11
0.15                    0.14*           0.15**
##                            (0.06)                  (0.07)
(0.11)                  (0.07)           (0.06)
##
## Constant                   -6.88**                  0.16
-13.19***                                  -1.46
##                            (2.77)                  (0.80)
(4.22)                                   (2.87)
##
## ---------------------------------------------------------------------------
--------------------------------------------------
## Observations                 537                     338
199                     537              537
## R2                           0.56                    0.33
0.64                    0.43             0.52
## Adjusted R2                  0.55                    0.31
0.62                    0.06             0.50
## F Statistic        54.64*** (df = 12; 524) 16.23*** (df = 10; 327)
27.39*** (df = 12; 186) 24.33*** (df = 10; 328) 557.55***
##
=============================================================================
===============================================
## Note:
*p<0.1; **p<0.05; ***p<0.01
```

3. Explain which model should be preferred and why. You are expected to use Hausman test for choosing between RE and FE models.

```
pFtest(m5,m1)

##
##  F test for individual effects
##
## data:  y1pccust ~ inf1pcma + as.factor(period) + as.factor(sector) +  ...
## F = 3.0819, df1 = 196, df2 = 328, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

```
phtest(m5,m6)

##
##  Hausman Test
##
## data:  y1pccust ~ inf1pcma + as.factor(period) + as.factor(sector) +  ...
## chisq = 29.409, df = 10, p-value = 0.001069
## alternative hypothesis: one model is inconsistent

?coeftest

## starting httpd help server ... done

coeftest(m5)

##
## t test of coefficients:
##
##                        Estimate Std. Error t value  Pr(>|t|)
## inf1pcma              0.3931997  0.0610634  6.4392 4.269e-10 ***
## as.factor(period)1 -2.1093271  2.0632740 -1.0223  0.307383
## as.factor(period)2 -2.8300161  2.7183050 -1.0411  0.298598
## as.factor(sector)2  0.6918385  9.6812638  0.0715  0.943074
## as.factor(sector)3 -0.4608358  4.0291596 -0.1144  0.909010
## ld1pcla             0.0025708  0.0717637  0.0358  0.971445
## st1pcdev            0.1960147  0.0604128  3.2446  0.001297 **
## cs1pccmk            0.0386397  0.0633560  0.6099  0.542362
## hr3pcsat            0.0123885  0.0602705  0.2055  0.837271
## pm1pcvc             0.1419677  0.0740020  1.9184  0.055924 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: We can see from the table that based on the F test for the individual effects and the hausman test (P value less tha 0.05) the fixed effects model is the better model in estimating the predictor variable.

4. Based on your preferred model, which variable has the strongest effect on y1pccust. Discuss whether the difference in coefficient of this variable compared to the second most important variable is statistically significant or not based on a formal test.

Bazed on the model comparison and stastical significance, we can see that for the fixed effects model, the strongest effect on y1pccust is made by the variables inf1pcma and st1pcdev followed by pm1pcvc. From the coeftest of th emodel to determinate the coefficients, we can see that inf1pcma is far more significant with p value of 4.26e-10 while the pvalue of the st1pcdev which is 0.001. This marks inf1pcma is the strongest variable that affects the y1pccust scores.

5. Discuss how your findings based on panel models are different from the ones that you obtained earlier based on Homework C. Which model seems more plausible and robust to you and why?

The first OLS model does not take into account the number of different years by the same company which leads to dependence and multi-collinearity. So it is not a good model. The second model is better but involves a lot of code and computation even though it gives results. Also, it only considers between companies variability and ignores the different years by the same company (within permenantid variability). Hence, the third model that considers the within effect would be the most appropriate.

6. Finally, use an instrumental variable approach using ivreg, and estimate the effect of inf1pcma on y1pccust by instrumenting inf1pcma by 1999-2006 period dummy. Also show the first stage regression testing relevance of the instrument. How does IV estimate for inf1pcma compare with FE estimate above and what does that say about the robustness of the effect of inf1pcma across different estimators and specifications?

```
iv=ivreg(y1pccust~inf1pcma+(inf1pcma|period==2 )+ period+sector+
            ld1pcla + st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc,data=df.p )
summary(iv)

##
## Call:
## ivreg(formula = y1pccust ~ inf1pcma + (inf1pcma | period == 2) +
##     period + sector + ld1pcla + st1pcdev + cs1pccmk + hr3pcsat +
##     pm1pcvc, data = df.p)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.129  -7.850  -1.201   5.789  68.079
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.707501   3.042825  -2.533 0.011597 *
## inf1pcma     0.194450   0.064646   3.008 0.002756 **
## period      -1.251594   0.812444  -1.541 0.124031
## sector       1.556309   0.476719   3.265 0.001167 **
## ld1pcla      0.121251   0.071948   1.685 0.092528 .
## st1pcdev     0.270026   0.063275   4.268 2.34e-05 ***
## cs1pccmk     0.237588   0.067580   3.516 0.000476 ***
## hr3pcsat     0.008604   0.056174   0.153 0.878319
## pm1pcvc      0.153301   0.067470   2.272 0.023480 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.72 on 528 degrees of freedom
## Multiple R-Squared: 0.4705,  Adjusted R-squared: 0.4625
## Wald test: 58.65 on 8 and 528 DF,  p-value: < 2.2e-16
```

Part 2: Mixed-Level Models

For this assignment, please use the same dataset that was provided to you via Canvas for the mid-term exam.

We want to now model both the intercept of y1pccust and the relationship between inf1pcma and y1pccust by edu sector (all other sector are used as a base) using mixed effects models. You will have in your model level 1 variables such as ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc. At level 2 you will only have edu sector variable to explain variance in intercept and slope (you may have to include appropriate interaction term(s) in the model syntax to allow for cross-level effects.

Please use a multi-level model using the lme4 packages in R following the example discussed in class, and based on your multi-level analysis, provide an interpretation of your mixed-effects results for random intercepts and random slopes.

Which model will you choose among mixed-effects model and any other models that you have run so far (including the models that you estimated in Homework C and the panel models in this homework), and why? What are pros and cons of different models?

```
df.p$mfg <- ifelse(df.p$sector == 0, 1, 0)
df.p$serv <- ifelse(df.p$sector == 0, 1, 0)
df.p$small <- ifelse(df.p$sector == 0, 1, 0)
df.p$educ <- ifelse(df.p$sector == 1, 1, 0)
df.p$hc <- ifelse(df.p$sector == 0, 1, 0)

#Varying intercept model with multiple individual-level and group-level
predictors
m7 <-lmer(y1pccust ~ inf1pcma + ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat +
pm1pcvc + I(inf1pcma*educ)+
            (1 + inf1pcma|permanentid ),data=df.p)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
control$checkConv, :
## Model failed to converge with max|grad| = 0.104988 (tol = 0.002, component
1)

## Varying intercept and varying slope model--intercept and inf1pcma being
treated as random
m8 <-lmer(y1pccust ~ inf1pcma + ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat +
pm1pcvc +
            (1 + inf1pcma| educ),data=df.p)

## boundary (singular) fit: see ?isSingular

summary(m7)

## Linear mixed model fit by REML ['lmerMod']
## Formula: y1pccust ~ inf1pcma + ld1pcla + st1pcdev + cs1pccmk + hr3pcsat +
##     pm1pcvc + I(inf1pcma * educ) + (1 + inf1pcma | permanentid)
##    Data: df.p
##
## REML criterion at convergence: 4152.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0678 -0.5457 -0.0600  0.4965  5.3157 
## 
## Random effects:
##  Groups      Name        Variance Std.Dev. Corr
##  permanentid (Intercept) 30.50381 5.5230       
##              inf1pcma     0.01055 0.1027   0.75
##  Residual                75.42903 8.6850       
## Number of obs: 537, groups:  permanentid, 199
## 
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          1.08243    2.38037   0.455
## inf1pcma             0.27310    0.05422   5.037
## ld1pcla              0.07222    0.05944   1.215
## st1pcdev             0.26747    0.05334   5.014
## cs1pccmk             0.13389    0.05579   2.400
## hr3pcsat             0.01017    0.04921   0.207
## pm1pcvc              0.13277    0.06106   2.174
## I(inf1pcma * educ) -0.04604    0.03287  -1.401
## 
## Correlation of Fixed Effects:
##             (Intr) inf1pc ld1pcl st1pcd cs1pcc hr3pcs pm1pcv
## inf1pcma     0.081                                          
## ld1pcla     -0.068 -0.257                                   
## st1pcdev    -0.143 -0.227 -0.155                            
## cs1pccmk    -0.355 -0.197 -0.112 -0.321                     
## hr3pcsat     0.060 -0.059 -0.444  0.073  0.046              
## pm1pcvc     -0.313 -0.138 -0.223 -0.144 -0.103 -0.335       
## I(nf1p*edc)  0.150 -0.116  0.062 -0.007 -0.124  0.018 -0.119
## convergence code: 0
## Model failed to converge with max|grad| = 0.104988 (tol = 0.002, component 
1)
```

```
#intercept only model
m9 = lmer(y1pccust ~ 1 +(1|educ),data=df.p)
```

```
summary(m8)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: y1pccust ~ inf1pcma + ld1pcla + st1pcdev + cs1pccmk + hr3pcsat +  
##     pm1pcvc + (1 + inf1pcma | educ)
##    Data: df.p
## 
## REML criterion at convergence: 4276.2
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.4351 -0.6358 -0.0595  0.4629  5.3354 
## 
## Random effects:
```

```
##  Groups   Name        Variance  Std.Dev. Corr 
##  educ     (Intercept) 2.612e+00  1.61631      
##           inf1pcma    5.582e-03  0.07471 -1.00
##  Residual             1.626e+02 12.74960      
## Number of obs: 537, groups:  educ, 2
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) -3.34318    2.75374  -1.214
## inf1pcma     0.16223    0.08322   1.949
## ld1pcla      0.13330    0.06783   1.965
## st1pcdev     0.29047    0.06307   4.605
## cs1pccmk     0.22102    0.06678   3.310
## hr3pcsat     0.01476    0.05308   0.278
## pm1pcvc      0.12700    0.06606   1.923
## 
## Correlation of Fixed Effects:
##          (Intr) inf1pc ld1pcl st1pcd cs1pcc hr3pcs
## inf1pcma -0.193                                   
## ld1pcla  -0.068 -0.119                            
## st1pcdev -0.048 -0.182 -0.201                     
## cs1pccmk -0.295 -0.189 -0.168 -0.342              
## hr3pcsat  0.005 -0.115 -0.414  0.049  0.064       
## pm1pcvc  -0.280 -0.161 -0.234 -0.140 -0.100 -0.226
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

Interpretation:

In the mixed level model, we see how the inf1pcmc variable is effecting the significance of the other estimators for the dependant variable, ie y1pccust scores. The interaction term between inf1pcmc and education sector is not significant.

The first OLS model does not take into account the number of different years by the same company which leads to dependence and multi-collinearity. So it is not a good model.

The second model is better but involves a lot of code and computation even though it gives results. Also, it only considers between companies variability and ignores the different years by the same company (within permenantid variability). Hence, the third model that considers the within effect would be the most appropriate.

Part 3: Logit/Probit Models

For this assignment, please use the same dataset that was provided to you via Canvas for the mid-term exam.

Create a dependent variable y2hicust that is 1 when y1pccust>50, else zero. Now run a logit model of y2hicust on inf1pcma +ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat + pm1pcvc; and include sector dummies (with manufacturing as the omitted or base group) and period dummies (you will need to create a period variable that divides 1990-2006 into 3 periods: before 1995, 1995-1998, and 1999-2006, make before 1995 as the base period).

Interpret the output and discuss effects of statistically significant variables in terms of odds and probabilities.

Plot at least two charts that show non-linear effects of key explanatory variables on probability scale.

Create a 2x2 classification (or confusion) matrix and discuss how good your model is.

Extra points: Come up with a better model and explain why that model is better.

```r
df$y2hicust <- ifelse(df$y1pccust>50 , 1, 0)

logit <- glm(y2hicust ~ inf1pcma +ld1pcla+ st1pcdev + cs1pccmk + hr3pcsat +
pm1pcvc
                + sector+period, data = df, family = "binomial")

summary(logit)
```

```
##
## Call:
## glm(formula = y2hicust ~ inf1pcma + ld1pcla + st1pcdev + cs1pccmk +
##     hr3pcsat + pm1pcvc + sector + period, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2721  -0.7228  -0.2383   0.6951   2.8808
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.0345092  0.9342277  -9.671  < 2e-16 ***
## inf1pcma     0.0353869  0.0132779   2.665   0.0077 **
## ld1pcla      0.0075530  0.0160722   0.470   0.6384
## st1pcdev     0.0415113  0.0130768   3.174   0.0015 **
## cs1pccmk     0.0683483  0.0153048   4.466 7.98e-06 ***
## hr3pcsat     0.0001424  0.0123573   0.012   0.9908
## pm1pcvc      0.0121298  0.0143179   0.847   0.3969
## sector       0.0699856  0.0957943   0.731   0.4650
## period      -0.1859572  0.1670721  -1.113   0.2657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 729.62  on 536  degrees of freedom
## Residual deviance: 481.39  on 528  degrees of freedom
## AIC: 499.39
```

```
##
## Number of Fisher Scoring iterations: 5
```

```
S(logit)
```

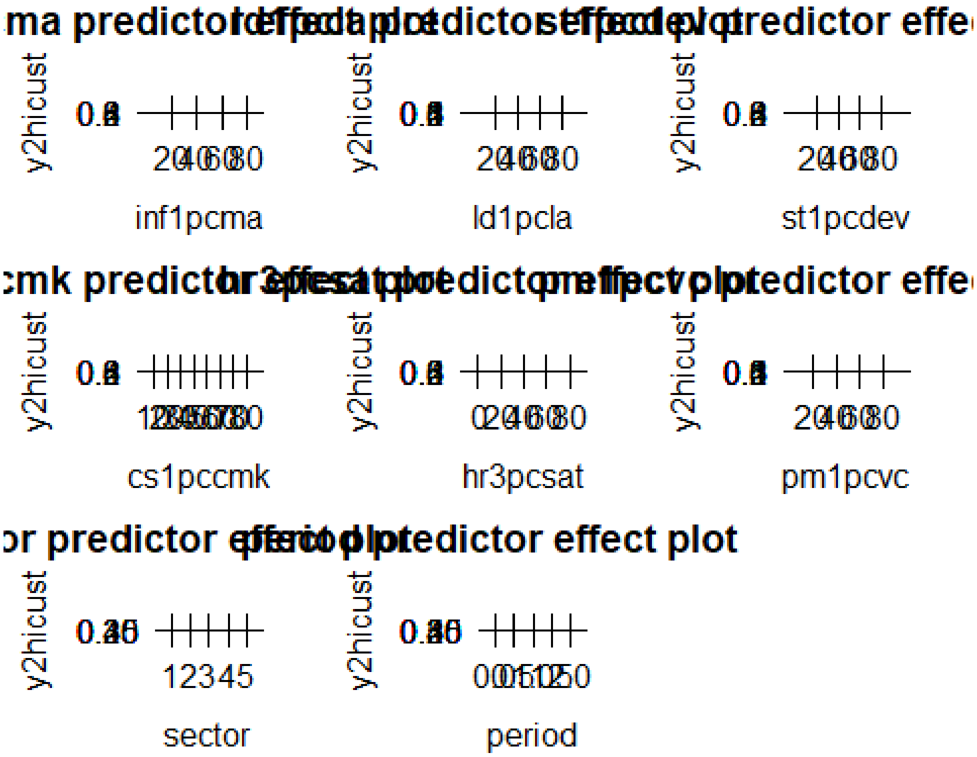
```
## Call: glm(formula = y2hicust ~ inf1pcma + ld1pcla + st1pcdev + cs1pccmk +
##           hr3pcsat + pm1pcvc + sector + period, family = "binomial", data
= df)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.0345092  0.9342277  -9.671  < 2e-16 ***
## inf1pcma     0.0353869  0.0132779   2.665   0.0077 **
## ld1pcla      0.0075530  0.0160722   0.470   0.6384
## st1pcdev     0.0415113  0.0130768   3.174   0.0015 **
## cs1pccmk     0.0683483  0.0153048   4.466 7.98e-06 ***
## hr3pcsat     0.0001424  0.0123573   0.012   0.9908
## pm1pcvc      0.0121298  0.0143179   0.847   0.3969
## sector       0.0699856  0.0957943   0.731   0.4650
## period      -0.1859572  0.1670721  -1.113   0.2657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 729.62  on 536  degrees of freedom
## Residual deviance: 481.39  on 528  degrees of freedom
##
##  logLik      df     AIC     BIC
## -240.69       9  499.39  537.96
##
## Number of Fisher Scoring iterations: 5
##
## Exponentiated Coefficients and Confidence Bounds
##                 Estimate        2.5 %       97.5 %
## (Intercept) 0.0001192237 1.752136e-05 0.0006862016
## inf1pcma    1.0360204814 1.009834e+00 1.0639800485
## ld1pcla     1.0075815940 9.761949e-01 1.0398152923
## st1pcdev    1.0423849024 1.016330e+00 1.0698868381
## cs1pccmk    1.0707382129 1.039661e+00 1.1040907515
## hr3pcsat    1.0001424398 9.761272e-01 1.0246769747
## pm1pcvc     1.0122036825 9.841126e-01 1.0411131527
## sector      1.0724927370 8.897458e-01 1.2963008794
## period      0.8303091164 5.972547e-01 1.1512596821
```

```
logitscalar <- mean(dlogis(predict(logit, type="link"))) #average marginal
effects to interpret how we interpret regular OLS
logitscalar*coef(logit)
```

```
##   (Intercept)      inf1pcma       ld1pcla      st1pcdev      cs1pccmk
## -1.322162e+00  5.178725e-03  1.105349e-03  6.074998e-03  1.000249e-02
##      hr3pcsat       pm1pcvc        sector        period
##  2.084397e-05  1.775148e-03  1.024210e-02 -2.721405e-02

#PLOT

plot(predictorEffects(logit))
```



```
pred.data=expand.grid(
  sector    =unique(df$sector),
  period    =unique(df$period),

  inf1pcma=quantile(df$inf1pcma,c(.25,.5,.75,1.0)),
  ld1pcla =quantile(df$ld1pcla,c(.25,.5,.75,1.0)),
  st1pcdev =quantile(df$st1pcdev,c(.25,.5,.75,1.0)),
  cs1pccmk =quantile(df$cs1pccmk,c(.25,.5,.75,1.0)),
  pm1pcvc =quantile(df$pm1pcvc,c(.25,.5,.75,1.0)),
  hr3pcsat =quantile(df$hr3pcsat,c(.25,.5,.75,1.0))

)

my.sample.predictions=predict(logit,newdata=pred.data,type = "response")
my.sample.predictions=cbind(pred.data,my.sample.predictions)
head(my.sample.predictions)

##   sector period inf1pcma ld1pcla st1pcdev cs1pccmk pm1pcvc hr3pcsat
## 1      1      0    37.14   45.71       40       40      46       40
## 2      2      0    37.14   45.71       40       40      46       40
## 3      3      0    37.14   45.71       40       40      46       40
## 4      4      0    37.14   45.71       40       40      46       40
## 5      5      0    37.14   45.71       40       40      46       40
## 6      1      1    37.14   45.71       40       40      46       40
##   my.sample.predictions
## 1            0.08730822
## 2            0.09304851
## 3            0.09912522
## 4            0.10555259
## 5            0.11234475
## 6            0.07358298
```

```r
my.sample.predictions[which(my.sample.predictions$my.sample.predictions==max(
my.sample.predictions$my.sample.predictions)),]
```

```
##       sector period inf1pcma ld1pcla st1pcdev cs1pccmk pm1pcvc hr3pcsat
## 61430      5      0       80   87.14    91.43       80   89.09       80
##       my.sample.predictions
## 61430             0.9942901
```

```r
my.sample.predictions[which(my.sample.predictions$my.sample.predictions==min(
my.sample.predictions$my.sample.predictions)),]
```

```
##    sector period inf1pcma ld1pcla st1pcdev cs1pccmk pm1pcvc hr3pcsat
## 11      1      2    37.14   45.71       40       40      46       40
##    my.sample.predictions
## 11            0.06186914
```

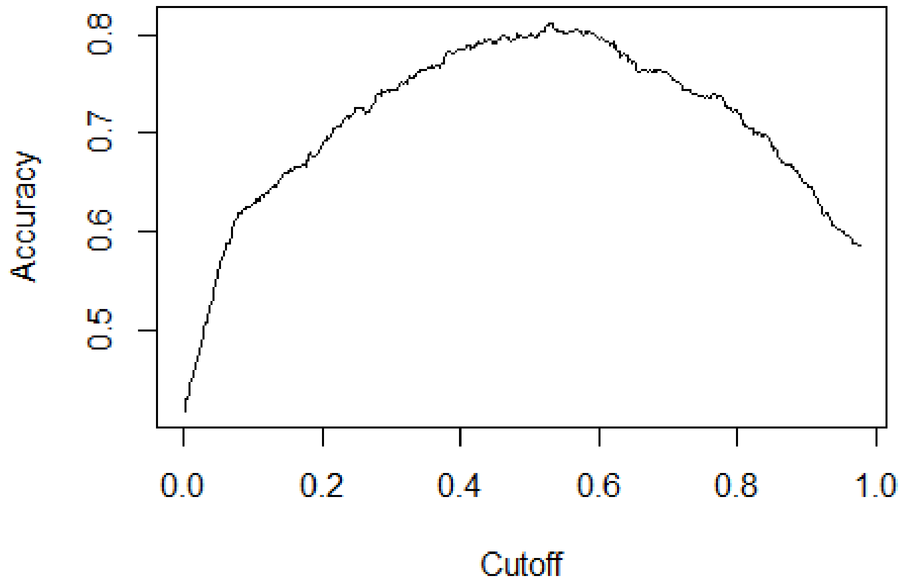
```r
library("ROCR")
```

```
## Loading required package: gplots
```

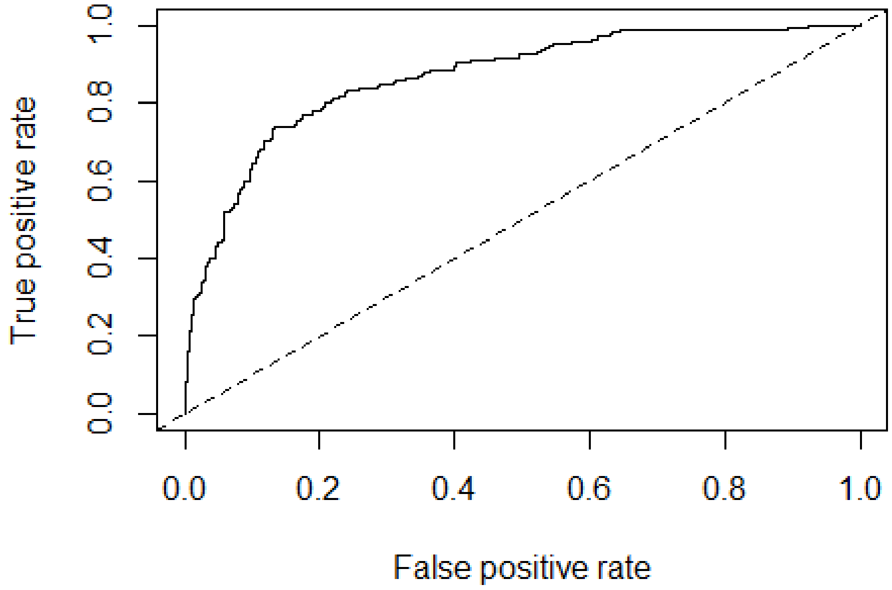
```
## 
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
## 
##     lowess
```

```r
pred <- prediction(fitted(logit),
                   df$y2hicust) #Accuracy curve
plot(performance(pred, "acc"))
```



```r
plot(performance(pred, "tpr", "fpr")) #ROC curve
abline(0, 1, lty = 2)
```



```
#Confusion Matrix
table(true = df$y2hicust,
      pred = round(fitted(logit)))
```

```
##     pred
## true   0   1
##    0 260  53
##    1  55 169
```

We could see few beta as significant. Model shows two deviances one is Null deviance and other is Residual deviance.Deviance is a measure of goodness of fit of a generalized linear model. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean). Here, residual deviance is large from Null deviance which is good indication of model

Interpretation in terms of probability: inf1pcma: The coefficient estimate of the variable inf1pcma is b = 0.03157, which is positive. This means that an increase in inf1pcma is associated with increase in the probability of being y2hicust as 1. st1pcdev: The coefficient estimate of the variable st1pcdev is b = 0.02892, which is positive. This means that an increase in st1pcdev is associated with increase in the probability of being y2hicust as 1. cs1pccmk: The coefficient estimate of the variable cs1pccmk is b = 0.06760, which is positive. This means that an increase in cs1pccmk is associated with increase in the probability of being y2hicust as 1. pm1pcvc: The coefficient estimate of the variable pm1pcvc is b = 0.03975, which is positive. This means that an increase in pm1pcvc is associated with increase in the probability of being y2hicust as 1. sectoreducation: The coefficient estimate of the variable sectoreducation is b = 2.07622, which is positive (highest positive). This means that if company belongs to education, it adversly influences in increasing probability of being y2hicust as 1 compared to mfg sector. time_period95to98: The coefficient estimate of the variable time_period95to98 is b = -.94663, which is negative. This means that an company belonging from 1995-1998 will be associated with a decreased probability of being y2hicust as 1. time_period99to06: The coefficient estimate of the variable time_period99to06 is b = -.68953, which is negative. This means that an company belonging from 1996-2006 will be associated with a decreased probability of being y2hicust as 1.

Interpretation in terms of odds ratio. An odds ratio measures the association between a predictor variable (x) and the outcome variable (y). It represents the ratio of the odds that an event will occur (event = 1) given the presence of the predictor x (x = 1), compared to the odds of the event occurring in the absence of that predictor (x = 0)

inf1pcma: The regression coefficient for inf1pcma is 0.03157. This indicate that one unit increase in the inf1pcma score will increase the odds of being y2hicust as 1 by exp(0.03157) 1.03 times. st1pcdev: The regression coefficient for st1pcdev is 0.02892. This indicate that one unit increase in the st1pcdev score will increase the odds of being y2hicust as 1 by exp(0.02892) 1.029 times. cs1pccmk: The regression coefficient for cs1pccmk is 0.06760. This indicate that one unit increase in the cs1pccmk score will increase the odds of being y2hicust as 1 by exp(0.06760) 1.07 times. pm1pcvc: The regression coefficient for pm1pcvc is 0.03975. This indicate that one unit increase in the pm1pcvc score will increase the odds of being y2hicust as 1 by exp(0.03975) 1.04 times. sectoreducation: The regression coefficient for sectoreducation is 2.07622. This indicate that if company belongs to sectoreducation increase the odds of being y2hicust as 1 by exp(2.07622) 7.98 times compared to mfg. time_period95to98: The coefficient estimate of the variable time_period95to98 is b = 0.94663, which is negative. This means that an if company is from time_period95to98 decrease the odds of being y2hicust as 1 by exp(0.94663) 2.58 times compared to companies hailing below 1995. time_period99to06: The coefficient estimate of the variable time_period99to06 is b = 0.68953, which is negative. This means that an if company is from time_period99to06 decrease the odds of being y2hicust as 1 by exp(0.68953) 2 times compared to companies hailing below 1995.