Predictive Analytics
Homework 4
Chris Cirelli

```
# Question 1
' Redo the example for job training grants and firm scrap rates
  by taking the firm level heterogeneity into consideration.
'

# Clear namespace ------------------------------------
rm(list=ls())

# Load Libraries --------------------------------------
require(foreign)
library('plm')
library("nlme")

# Load Data -------------------------------------------
jtrain= read.dta('/home/cc2/Desktop/repositories/Time_Series/wk8_est_bias/data/jtrain1.dta',
convert.factors=FALSE)

# Inspect Data ----------------------------------------
head(jtrain)
summary(jtrain)

# Remove scrap na values from data.frame ---------------
jt <- jtrain[!is.na(jtrain$scrap),]
summary(jt)  # note neither scrap nor lscrap have na values now.


# Part A:  Include lscrap_1 in your model and calculate OLS estimator
' lscrap = B1 + B2I(year=1998) + B3I(year=1989) + B4grant + B5grant_1 +
        B6lscrap_1 + u

  Note:  By refering to OLS in the homework description Im assuming that you
      want us to compare the results from using OLS and a fixed effect model
      in order to illustrate the effects of contant variable c.
'

# OLS
ols <- lm(lscrap ~ d88+d89+grant+grant_1+lscrap_1, data=jt)
summary(ols)
```

```
# Fixed Effect Model
f.eff <- lm(lscrap ~ d88 + d89 + grant + grant_1 + lscrap_1 + factor(fcode)-1, data=jt)
summary(f.eff)

# Part B:  Compare results
' Compare the results obtained from part a.) with what we got in class from a
  fixed effect model.  What is the difference between the two versions of
  B4 and B5.  Why is there such a difference

  Results OLS -----------------------------------------------
  lm(formula = lscrap ~ d88 + d89 + grant + grant_1 + lscrap_1,
      data = jt)

  Residuals:
       Min       1Q   Median       3Q      Max
  -2.87588 -0.12525  0.07163  0.24906  1.88619

  Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)  -0.1525     0.1002  -1.522    0.131
  d88           0.1154     0.1199   0.962    0.338
  d89               NA         NA      NA       NA
  grant        -0.1724     0.1257  -1.371    0.173
  grant_1      -0.1073     0.1610  -0.666    0.507
  lscrap_1      0.8808     0.0358  24.606   <2e-16 ***

  # Fixed Effects Model ---------------------------------------
  lm(formula = lscrap ~ d88 + d89 + grant + grant_1 + lscrap_1 +
    factor(fcode) - 1, data = jt)

  Residuals:
      Min      1Q  Median      3Q     Max
  -1.4565 -0.1172  0.0000  0.1172  1.4565

  Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
  d88                2.625312   0.515470   5.093 5.39e-06 ***
  d89                2.390899   0.515981   4.634 2.60e-05 ***
  grant             -0.004311   0.213569  -0.020 0.983976
  grant_1            0.011470   0.326481   0.035 0.972113
  lscrap_1           0.258079   0.137640   1.875 0.066635 .

  i.) Why is there such a difference?
```

- We cam clearly see a difference in the importance of the d88 and d89 variables when we use the fixed-effect model. With ordinar OLS they are not signficant (pvalue), which changes when we fit with the fixed effect model.
- Regular OLS does not take into consideration the heterogeneity across the features or years.

'

```
# Part C:  Random Effects Model
' Include the lag of log(scrap)(lscrap_1) in your model and calculate B using
  the random effect model
'
r.eff <- plm(lscrap ~ d88 + d89 + grant + grant_1 + lscrap_1, data=jt, index=c("fcode", "year"),
model="random")
summary(r.eff)


# Results
'
Oneway (individual) effect Random Effect Model
   (Swamy-Aroras transformation)

Call:
  plm(formula = lscrap ~ d88 + d89 + grant + grant_1 + lscrap_1,
      data = jt, model = "random", index = c("fcode", "year"))

Balanced Panel: n = 54, T = 2, N = 108

Effects:
  var std.dev share
idiosyncratic 0.1597  0.3997 0.584
individual    0.1138  0.3374 0.416
theta: 0.3579

Residuals:
  Min.   1st Qu.    Median   3rd Qu.      Max.
-2.253770 -0.139917  0.064347  0.227121  1.342206

Coefficients: (1 dropped because of singularities)
Estimate Std. Error z-value Pr(>|z|)
(Intercept) -0.171152   0.108798 -1.5731   0.1157
d88          0.138110   0.099456  1.3886   0.1649
grant       -0.127571   0.125323 -1.0179   0.3087
```

```
grant_1     -0.037487   0.171848 -0.2181   0.8273
lscrap_1     0.847573   0.042913 19.7512   <2e-16 ***
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    95.955
Residual Sum of Squares: 19.576
R-Squared:      0.79598
Adj. R-Squared: 0.78806
Chisq: 401.861 on 4 DF, p-value: < 2.22e-16
'

# Part d:
' Which model should we use?
  - We should reject the null hypothesis and not use OLS.
  - We should use the Fixed Effect model.

'
```

```
# Question 2 -------------------------------------------------------------------
' Analyse the DoctorVisits data using a Poisson regression for the
  number of visits. Is the Possion model satisfactory?  If not,
  where are the problems and what could be done about them?
'

# Clear namespace ---------------------------------------------
rm(list=ls())

# Load Libraries ----------------------------------------------
library(AER)

# Load data ---------------------------------------------------
data(DoctorVisits)

# Inspect Data ------------------------------------------------
' Observations:

  Target is not normally distributed.
  Var > mean indicates over dispersion.
  Likely not a good candidate for a Poisson Regression model
  as the data deviates from the required assumptions.
'
col_names = names(DoctorVisits)
col_names
summary(DoctorVisits)
hist(DoctorVisits$visits)
mean(DoctorVisits$visits)
var(DoctorVisits$visits)    # Variance > mean


# Fit Poisson Model
m.poisson <- glm(visits ~ ., DoctorVisits, family=poisson(link='log'))
summary(m.poisson)

# Interpretation
' Significance:  Pvalues indicate that genderfemal, income, illenss,
                 reduced, health and freepooryes are significant and
                 affect the response variable.

  Disperson:      Residual Deviance > dof, which indicates over
                 dispersion.
```

```
  Model is not satisfactory.
'

# Check for overdispersion
dispersiontest(m.poisson, alternative='greater')
' Output

  data:  m.poisson
  z = 6.5386, p-value = 3.105e-11
  alternative hypothesis: true dispersion is greater than 1
  sample estimates:
  dispersion
  1.415602
'
```

```
# Question 3 ------------------------------------------------------------------------
' Fit the following four models for the Affairs data
  a.) Poisson,
  b.) Negative Binomial: glm.nb(dependend ~., data=data)
  c.) Hurdle Poisson: hurdle(formula, data, link=logit)
  d.) Hurdle Negative Binomial:
  Discuss results comparing Log likelihood, AIC,
  Prediction vs Actual, rootgram.

'

# Clear namespace ------------------------------------
rm(list=ls())

# Load Packages --------------------------------------
library(AER)
library(MASS)
library(dplyr)
library(pscl)

# Load data ------------------------------------------
data("Affairs")

# Inspect Data ---------------------------------------
' y = affairs
'
?Affairs
names(Affairs)
hist(Affairs$affairs)
summary(Affairs)
Affairs %>% group_by(gender) %>% summarise(affairs=sum(affairs))

affairs.mu <- mean(Affairs$affairs)
affairs.var <- var(Affairs$affairs)
if affairs.var > affair.mu:
  print('Variance is greater than mean')


# Fit Poisson Model ---------------------------------
' Results:
    Residual Deviance:  2359
    AIC:                2871
```

```
   Regressors:      Age, yearsmarried, religiousness, occupation and rating are all significant
based
                    on the p-value.
'
m.poisson <- glm(affairs ~., data=Affairs,
                family=poisson(link='log'))

summary(m.poisson)


# Fit Negative Binomial Regression -------------------
' Results:
   Residual Deviance:  339 (substantially less than the poisson model)
   AIC:                1476 (almost half the poisson model)
   Log-likelihood:     -1456
   Dispersion:         less than 1.
   Regressors:         only yearsmarried, religiousness and rating are significant based on
p-value.
   Overall:            seems to be a better model based on the Residual erro and AIC scores and
that the model
                       does not assume mu = var.
'

m.nb <- glm.nb(affairs ~., data=Affairs)
summary(m.nb)

# Fit Hurdle Model -----------------------------------
' Results:
   Model:            Fit both truncated poisson with log link and binomial with logit link.
   Log-likelihood:   -758.8
   Overall:          Appears to be a better fit based on teh log-likelihood value.


'
m.hurdle.poisson <- hurdle(affairs~., data=Affairs, dist='poisson')

summary(m.hurdle.poisson)
```