Regression Modeling with Actuarial and Financial Applications
Exercise 3.5

Note: See the last four pages of this document for code that is usable in a script file within R

Download "NAICExpense" data from Jed Frees' website
Follow the 'Data' link to find 'NAICExpense'

A preliminary inspection of the data showed that many firms did not report any insurance losses incurred in 2005. For this exercise, we consider the 384 companies with some losses in the file "NAICExpense"

a) Produce summary statistics of the response variable and the non-binary explanatory variables. Note the pattern of skewness for each variable. Note that many variables have negative values.

Choose NAICExpense.csv from its saved location

```
NAICExpense <- read.table(choose.files(), header=TRUE,
sep=",")

summary(NAICExpense)
```

Notes

- The assignment operator **<-** assigns the chosen file to the variable `NAICExpense`
- **`read.table`** reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.
- **`choose.files`** uses a Windows file dialog to choose a list of zero or more files interactively (this allows for more universal code, rather than using a specific location)
- **`header=TRUE`** tells `read.table` that the data has headers as its first row
- **`sep=","`** tells `read.table` the table is in comma separated format (.csv)
- **`summary`** is a generic function used to produce result summaries of the results of various model fitting functions. The results will vary depending on the class of first argument
- In this case, summary's output will be descriptive statistics because the data are numerical

b) Transform each non-binary variable through the modified logarithm transform, ln(1+x). Produce summary statistics of these modified non-binary explanatory variables. Let LNEXPENSES (=ln(1+EXPENSES)) denote the modified expense variable. For subsequent analysis, use only the modified variables described in b)

```
attach(NAICExpense)
```

Don't forget to use `help()` to find out more about specific functions in R!

```
NAICExpense$LNEXPENSES <- log(1+EXPENSES)
NAICExpense$LNLONGLOSS <- log(1+LONGLOSS)
NAICExpense$LNSHORTLOSS <- log(1+SHORTLOSS)
NAICExpense$LNGPWPERSONAL <- log(1+GPWPERSONAL)
NAICExpense$LNGPWCOMM <- log(1+GPWCOMM)
NAICExpense$LNASSETS <- log(1+ASSETS)
NAICExpense$LNCASH <- log(1+CASH)
NAICExpense$LNSTAFFWAGE <- log(1+STAFFWAGE)
NAICExpense$LNAGENTWAGE <- log(1+AGENTWAGE)
```

Notes

- **attach** attaches a database to the **R** search path. This means that the database is searched by **R** when evaluating a variable, so objects in the database can be accessed by simply giving their names
- `NAICExpense$LNEXPENSES` specifies a new variable, `LNEXPENSES`, in the data frame `NAICExpense` for the assignment operator to create
- **log** computes logarithms, by default natural logarithms

c) Produce a table of correlations for the non-binary variables. What three variables are most highly correlated with LNEXPENSES?

```
cor(NAICExpense[, c("LNEXPENSES", "LNLONGLOSS",
"LNSHORTLOSS", "LNGPWPERSONAL","LNGPWCOMM", "LNASSETS",
"LNCASH","LNSTAFFWAGE","LNAGENTWAGE")], use="complete.obs")
```

LNLONGLOSS=.918, LNSHORTLOSS=.929, LNGPWCOMM=.864

Notes

- **cor** computes the correlation of x and y if these are vectors. If x and y are matrices then the correlations between the columns of x and the columns of y are computed
- **c** combines its arguments to form a vector
- **use="complete.obs"** tells `cor` that missing values should be handled by casewise deletion (and if there are no complete cases, that gives an error)

d) Provide a box plot of LNEXPENSES by level of GROUP. Which level of group has higher expenses?

```
boxplot(LNEXPENSES ~ NAICExpense$GROUP, data=NAICExpense)
```

Group 1, affiliated companies

Don't forget to use `help()` to find out more about specific functions in R!

Notes

- **`boxplot`** produces box-and-whisker plot(s) of the given (grouped) values
- Use `help(boxplot)` to find out more about the arguments involved in boxplot
- `~` denotes a formula
- **`data=NAICExpense`** loads specified data sets, in this case `NAICExpense`

e) Fit a linear model of LNEXPENSES on all eleven explanatory variables. Summarize the fit of this model by citing the residual standard deviation, s; the coefficient of determination, R^2; and its adjusted version, Ra^2

```
NAIC_lm1 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS
+LNGPWPERSONAL+LNGPWCOMM+LNASSETS+LNCASH+GROUP
+STOCK+MUTUAL+LNSTAFFWAGE+LNAGENTWAGE, data=NAICExpense)

summary(NAIC_lm1)
```

s=0.02235, R^2=0.9422, Ra^2=0.9404

Notes

- **`lm`** is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance
- For more information about using `lm`, use `help(lm)`

f) Fit a linear model of LNEXPENSES on a reduced model using eight explanatory variables, dropping CASH, STOCK, and MUTUAL. For the explanatory variables, include assets, GROUP, both versions of losses and gross premiums, as well as the two BLS variables.

```
NAIC_lm2 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS
+LNGPWPERSONAL+LNGPWCOMM+LNASSETS+GROUP+LNSTAFFWAGE
+LNAGENTWAGE, data=NAICExpense)
```

f(i) Summarize the fit of this model by citing s, R^2, and Ra^2.

```
summary(NAIC_lm2)
```

s=0.02242, R^2=0.9414, Ra^2=0.94

f(ii) Interpret the coefficient associated with commercial lines gross premiums on the logarithmic scale.

Don't forget to use `help()` to find out more about specific functions in R!

LNGPWCOMM coefficient = .0688575
When LNGPWCOMM increases by 1, LNEXPENSES increases by .0688575

f(iii) Suppose that GPWCOMM increases by $1.00, how much do we expect EXPENSES to increase? Use your answer in part f(ii) and median values of GPWCOMM and EXPENSES for this question. **(Why use median values?)**

When GPWCOMM increases by $2.72, EXPENSES increases by $1.07
When GPWCOMM increases by $1, EXPENSES increases by $0.39

g) Square each of the two loss and the two gross premium variables. Fit a linear model of LNEXPENSES on a reduced model using twelve explanatory variables, the eight variables in part(f), and the four additional squared terms just created.

```
attach(NAICExpense)

NAICExpense$LNLONGLOSS_sq <- LNLONGLOSS^2
NAICExpense$LNSHORTLOSS_sq <- LNSHORTLOSS^2
NAICExpense$LNGPWPERSONAL_sq <- LNGPWPERSONAL^2
NAICExpense$LNGPWCOMM_sq <- LNGPWCOMM^2

NAIC_lm3 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS
+LNGPWPERSONAL+LNGPWCOMM+LNASSETS+GROUP +LNSTAFFWAGE
+LNAGENTWAGE+LNLONGLOSS_sq+LNSHORTLOSS_sq
+LNGPWPERSONAL_sq+LNGPWCOMM_sq, data=NAICExpense)
```

g(i) Summarize the fit of this model by citing s, R^2, and Ra^2.

```
summary(NAIC_lm3)
```

s=0.02099, R^2=0.9492, Ra^2=0.9474

g(ii) Do the quadratic variables appear to be useful explanatory variables?

Since s decreased and R^2 increased, yes.

h) Now omit the two BLS variables, so you are fitting a model of LNEXPENSES on assets, GROUP, both versions of losses and gross premiums, as well as quadratic terms. Summarize the fit of the model by citing s, R^2, and Ra^2.

Don't forget to use `help()` **to find out more about specific functions in R!**

```
NAIC_lm4 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS
+LNGPWPERSONAL+LNGPWCOMM+LNASSETS+GROUP +LNLONGLOSS_sq
+LNSHORTLOSS_sq+LNGPWPERSONAL_sq+LNGPWCOMM_sq,
data=NAICExpense)

summary(NAIC_lm4)
```

s=0.02177, R^2=0.9424, Ra^2=0.9409

i) Drop the quadratic terms in part(g) and add interaction terms with the dummy variable GROUP. Thus, there are now 11 variables, assets, GROUP, both versions of losses and gross premiums, as well as interactions of GROUP with assets and both versions of losses and gross premiums. Summarize the fit of this model by citing s, R^2, and Ra^2.

```
attach(NAICExpenses)

GROUPc <- GROUP - mean(GROUP)
LNASSETSc <- LNASSETS - mean(LNASSETS)
LNLONGLOSSc <- LNLONGLOSS - mean(LNLONGLOSS)
LNSHORTLOSSc <- LNSHORTLOSS - mean(LNSHORTLOSS)
LNGPWPERSONALc <- LNGPWPERSONAL - mean(LNGPWPERSONAL)
LNGPWCOMMc <- LNGPWCOMM - mean(LNGPWCOMM)

NAICExpense$GROUPintLNASSETS <- GROUPc*LNASSETSc
NAICExpense$GROUPintLNLONGLOSS <- GROUPc*LNLONGLOSSc
NAICExpense$GROUPintLNSHORTLOSS <- GROUPc*LNSHORTLOSSc
NAICExpense$GROUPintLNGPWPERSONAL <- GROUPc*LNGPWPERSONALc
NAICExpense$GROUPintLNGPWCOMM <- GROUPc*LNGPWCOMMc

NAIC_lm5 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS
+LNGPWPERSONAL+LNGPWCOMM+LNASSETS+GROUP+GROUPintLNASSETS
+GROUPintLNLONGLOSS+GROUPintLNSHORTLOSS
+GROUPintLNGPWPERSONAL+GROUPintLNGPWCOMM, data=NAICExpense)

summary(NAIC_lm5)

# s=0.02323, R^2=0.9347, Ra^2=0.9327
```

Notes

- `mean` is a generic function for the (trimmed) arithmetic mean

Don't forget to use `help()` to find out more about specific functions in R!

Follow these steps to copy the following into a script file in order to copy code into R more easily:

- Open R
- Select File -> New Script
- Copy and paste this into the new script
- R will ignore text following #, which allows notes to be made in script files
- Highlighting lines of code and right-clicking allows the selection to be run in R

```
#Regression Modeling with Actuarial and Financial Applications
#Exercise 3.5
#This Exercise considers insurance company data from the NAIC

#A Preliminary inspection of the data showed that many firms did not
#report any inusrance losses incurred in 2005. For this exercise,
#we consider the 384 companies with some losses in the file "NAICExpense"

#a)Produce summary statistics of the response variable and the
#nonbinary explanatory variables. Note the pattern of skewness for
#each variable. Note that many variables have negative values.

NAICExpense <- read.table(choose.files(), header=TRUE, sep=",")
summary(NAICExpense)

#b)Transform each nonbinary variable through the modified logarithm
#transform, ln(1+x). Produce summary statistics of these modified
#nonbinary explanatory variables. Let LNEXPENSES (=ln(1+EXPENSES))
#denote the modified expense variable.
#For subsequent analysis, use only the modified variables described in b)

attach(NAICExpense)

NAICExpense$LNEXPENSES <- log(1+EXPENSES)
NAICExpense$LNLONGLOSS <- log(1+LONGLOSS)
NAICExpense$LNSHORTLOSS <- log(1+SHORTLOSS)
NAICExpense$LNGPWPERSONAL <- log(1+GPWPERSONAL)
NAICExpense$LNGPWCOMM <- log(1+GPWCOMM)
NAICExpense$LNASSETS <- log(1+ASSETS)
NAICExpense$LNCASH <- log(1+CASH)
NAICExpense$LNSTAFFWAGE <- log(1+STAFFWAGE)
NAICExpense$LNAGENTWAGE <- log(1+AGENTWAGE)

#c)Produce a table of correlations for the nonbinary variables. What three
#variables are most highly correlated with LNEXPENSES?

cor(NAICExpense[, c("LNEXPENSES","LNLONGLOSS","LNSHORTLOSS","LNGPWPERSONAL",
"LNGPWCOMM","LNASSETS","LNCASH","LNSTAFFWAGE","LNAGENTWAGE")], use="complete.obs")
```

Don't forget to use `help()` to find out more about specific functions in R!

```
        # LNLONGLOSS=.918, LNSHORTLOSS=.929, LNGPWCOMM=.864

#d)Provide a boxplot of LNEXPENSES by level of GROUP. Which level of group
#has higher expenses?

boxplot(LNEXPENSES ~ NAICExpense$GROUP, data=NAICExpense)

        #Group 1, affliated companies

#e)Fit a linear model of LNEXPENSES on all eleven explanatory variables.
#Summarize the fit of this model by citing the residual standard
#deviation,s;the coefficient of determination, R^2; and its adjusted
#version, Ra^2

NAIC_lm1 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS+LNGPWPERSONAL+LNGPWCOMM
+LNASSETS+LNCASH+GROUP+STOCK+MUTUAL+LNSTAFFWAGE+LNAGENTWAGE,
data=NAICExpense)

summary(NAIC_lm1)

        # s=0.02235, R^2=0.9422, Ra^2=0.9404

#f)Fit a linear model of LNEXPENSES on a reduced model using eight
#explanatory variables, dropping CASH, STOCK, and MUTUAL. For the
#explanatory variables, include assets, GROUP, both versions of losses and
#gross premiums, as well as the two BLSvariables.

NAIC_lm2 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS+LNGPWPERSONAL+LNGPWCOMM
+LNASSETS+GROUP+LNSTAFFWAGE+LNAGENTWAGE, data=NAICExpense)

#f(i) Summarize the fit of this model by citing s, R^2, and Ra^2.

summary(NAIC_lm2)

        # s=0.02242, R^2=0.9414, Ra^2=0.94

#f(ii) Interpret the coefficient associated with commercial lines gross
#premiums on the logarithmic scale.

        #LNGPWCOMM coefficent = .0688575
        #When LNGPWCOMM increases by 1, LNEXPENSES increases by .0688575

#f(iii) Suppose that GPWCOMM increases by $1.00, how much do we expect
#EXPENSES to increase? Use your answer in part f(ii) and median values
#of GPWCOMM and EXPENSES for this question. (Why use median values?)

        #When GPWCOMM increases by $2.71828183, EXPENSES increases
                by $1.07128354
        #When GPWCOMM increases by $1, EXPENSES increases by $0.39410319
```

Don't forget to use `help()` to find out more about specific functions in R!

```
#g)Square each of the two loss and the two gross premium variables. Fit
#a linear model of LNEXPENSES on a reduced model using twelve explanatory
#variables, the eight variables in part(f), and the four additional
#squared terms just created.

attach(NAICExpense)

NAICExpense$LNLONGLOSS_sq <- LNLONGLOSS^2
NAICExpense$LNSHORTLOSS_sq <- LNSHORTLOSS^2
NAICExpense$LNGPWPERSONAL_sq <- LNGPWPERSONAL^2
NAICExpense$LNGPWCOMM_sq <- LNGPWCOMM^2

NAIC_lm3 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS+LNGPWPERSONAL+LNGPWCOMM
+LNASSETS+GROUP+LNSTAFFWAGE+LNAGENTWAGE+LNLONGLOSS_sq+LNSHORTLOSS_sq
+LNGPWPERSONAL_sq+LNGPWCOMM_sq, data=NAICExpense)

#g(i) Summarize the fit of this model by citing s, R^2, and Ra^2.

summary(NAIC_lm3)

        #s=0.02099, R^2=0.9492, Ra^2=0.9474

#g(ii) Do the quadratic variables appear to be useful explanatory
#variables?

        #Since s decreased and R^2 increased, I think so.

#h)Now omit the two BLS variables, so you are fitting a model of
#LNEXPENSES on assets, GROUP, both versions of losses and gross premiums,
#as well as quadratic terms. Summarize the fit of the model by citing
#s, R^2, and Ra^2.

NAIC_lm4 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS+LNGPWPERSONAL+LNGPWCOMM
+LNASSETS+GROUP+LNLONGLOSS_sq+LNSHORTLOSS_sq+LNGPWPERSONAL_sq+LNGPWCOMM_sq,
data=NAICExpense)

summary(NAIC_lm4)

        # s=0.02177, R^2=0.9424, Ra^2=0.9409

#i)Drop the quadratic terms in part(g) and add interaction terms with
#the dummy variable GROUP. Thus, there are now 11 variables, assets,
#GROUP, both versions of losses and gross premiums, as well as
#interactions of GROUP with assets and both versions of losses and gross
#premiums. Summarize the fit of this model by citing s, R^2, and Ra^2.

attach(NAICExpense)
```

Don't forget to use `help()` to find out more about specific functions in R!

```
GROUPc <- GROUP - mean(GROUP)
LNASSETSc <- LNASSETS - mean(LNASSETS)
LNLONGLOSSc <- LNLONGLOSS - mean(LNLONGLOSS)
LNSHORTLOSSc <- LNSHORTLOSS - mean(LNSHORTLOSS)
LNGPWPERSONALc <- LNGPWPERSONAL - mean(LNGPWPERSONAL)
LNGPWCOMMc <- LNGPWCOMM - mean(LNGPWCOMM)

NAICExpense$GROUPintLNASSETS <- GROUPc*LNASSETSc
NAICExpense$GROUPintLNLONGLOSS <- GROUPc*LNLONGLOSSc
NAICExpense$GROUPintLNSHORTLOSS <- GROUPc*LNSHORTLOSSc
NAICExpense$GROUPintLNGPWPERSONAL <- GROUPc*LNGPWPERSONALc
NAICExpense$GROUPintLNGPWCOMM <- GROUPc*LNGPWCOMMc

NAIC_lm5 <- lm(LNEXPENSES ~ LNLONGLOSS+LNSHORTLOSS+LNGPWPERSONAL+LNGPWCOMM
+LNASSETS+GROUP+GROUPintLNASSETS+GROUPintLNLONGLOSS+GROUPintLNSHORTLOSS
+GROUPintLNGPWPERSONAL+GROUPintLNGPWCOMM, data=NAICExpense)

summary(NAIC_lm5)

        # s=0.02323, R^2=0.9347, Ra^2=0.9327
```