

UNIVERSITY OF WOLLONGONG  
School of Mathematics and Applied Statistics  
**STAT902. Advanced Data Analysis**

ASSIGNMENT 6

**Due:** 5:00pm Tuesday 27th April, 2010.

Local students: please put under lecturer's door.

Remote students: please e-mail or post by this date.

1. This question deals with logistic regression modelling when there is a high number of candidate predictors. The predictors are also of various types: continuous, binary and ternary (3 categories). The data correspond to a study on the survival of patients following admission to an adult intensive care unit (ICU). The variables are:

variable	description
died	Vital status (0 = Lived, 1 = Died)
age	Patient's age in years
female	Patient's sex (0 = Male, 1 = Female)
race	Patient's race (1 = White, 2 = Black, 3 = Other)
surgical	Service at ICU admission (0 = Medical, 1 = Surgical)
cancer	Is cancer part of the present problem? (0 = No, 1 = Yes)
chRenFail	History of chronic renal failure (0 = No, 1 = Yes)
infection	Infection probable at ICU admission (0 = No, 1 = Yes)
CPR	CPR prior to ICU admission (0 = No, 1 = Yes)
SBP	Systolic blood pressure at ICU admission (in mm Hg)
heartRate	Heart rate at ICU admission (beats/min)
prevAdmin	Previous admission to an ICU within 6 months (0 = No, 1 = Yes)
emergency	Type of admission (0 = Elective, 1 = Emergency)
fracture	Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes)
hiPO2	PO2 from initial blood gases ( $0 \leq 60$ , $1 > 60$ )
hiPH	PH from initial blood gases ( $0 \leq 7.25$ , $1 > 7.25$ )
hiPCO2	PCO2 from initial blood gases ( $0 \leq 45$ , $1 > 45$ )
hiBIC	Bicarbonate from initial blood gases ( $0 \leq 18$ , $1 > 18$ )
hiCRE	Creatinine from initial blood gases ( $0 \leq 2.0$ , $1 > 2.0$ )
LOC	Level of consciousness at admission (1 = no coma or stupor, 2 = deep stupor, 3 = coma)

Please work through the following instructions. You may wish to save some of the longer commands in a file so you don't have re-type them if you make an error. Note, also, that the  $\uparrow$  key can be used to recall previous commands in R. Your assignment submission should include the summary outputs of each model fit, the final partial residual plots and the solution to part (k).

- Obtain the data file `ICU.txt` from the course web-site (Computer Code and Data page).
- Start an R session and, after sorting out directory/folder issues, import the data via the command `ICU <- read.table("ICU.txt", header=TRUE)`
- Issue the command `summary(ICU)`. It gives a feeling for the nature of the data.
- The response variable is `died`. Out of the 19 candidate predictors two are continuous (`age` and `SBP`), two are ternary (`race` and `LOC`) and the rest are binary. It is the opinion of the lecturer of this course that multi-category variables be

converted to binary indicator form, at least at the learning stage. This allows easier interpretation of the output. Issue the following commands to perform this conversion:

```
raceBlack <- as.numeric(ICU$race==2)
raceOther <- as.numeric(ICU$race==3)
deepStupor <- as.numeric(ICU$LOC==2)
coma <- as.numeric(ICU$LOC==3)
myICU <- ICU[, -c(4, 20)]
myICU <- data.frame(myICU, raceBlack, raceOther, deepStupor, coma)
```

The new data frame `myICU` has replaced the ternary variables by appropriate binary indicator variables.

- (e) Fit the full model via the command:

```
fullFit <- glm(died~., data=myICU, family=binomial)
```

(Note that the 'dot' after the `~` tells R to include all variables in the `myICU` as predictors except for `died`). Ignore any warning messages you may get.

- (f) Issue the command `print(summary(fullFit))`. You should see that several predictors are not significant.
- (g) Next we will use the stepwise regression algorithm available in R to automatically build a model using the AIC model selection criterion. Enter commands:

```
stepOut <- step(glm(died~., data=myICU, family=binomial))
print(summary(stepOut))
```

Your assignment should include the output from the second of these commands; not the first.

- (h) The model with the lowest AIC value should have 10 predictors (compared with the original 19). However, some of them are still not very significant. As a final parsimonious model (in this assignment) we will omit the non-significant variables from `stepOut` and fit:

```
parsFit <- glm(died~age+cancer+SBP+emergency+hiPH
              +hiPCO2+coma, data=myICU, family=binomial)
print(summary(parsFit))
```

- (i) A final step is to check the linearity assumption for the continuous predictors:

```
partialResids <- residuals(parsFit, type="partial")
```

```
age <- myICU$age
ageg <- seq(min(age), max(age), length=101)
SBP <- myICU$SBP
SBPg <- seq(min(SBP), max(SBP), length=101)
```

```
par(mfrow=c(1, 2))
plot(age, partialResids[, 1], main="age",
      xlab="age", ylab="partial residuals")
fitSS1 <- smooth.spline(age, partialResids[, 1], df=3)
lines(ageg, predict(fitSS1, ageg)$y, col="red")
```

```
plot(SBP, partialResids[, 3], main="SBP",
```

```

      xlab="SBP",ylab="partial residuals")
fitSS3 <- smooth.spline(SBP,partialResids[,3],df=3)
lines(SBPg,predict(fitSS3,SBPg)$y,col="red")

```

- (j) The plots obtained in the previous step have the problem that the high partial residuals inflate the frame size and make it hard to see what is happening near the ‘action’. The following modifications set the vertical frame size to be between -4 and 4 to improve matters:

```

par(mfrow=c(1,2))
plot(age,partialResids[,1],main="age",
      xlab="age",ylab="partial residuals",ylim=c(-4,4))
lines(ageg,predict(fitSS1,ageg)$y,col="red")

plot(SBP,partialResids[,3],main="SBP",
      xlab="SBP",ylab="partial residuals",ylim=c(-4,4))
lines(SBPg,predict(fitSS3,SBPg)$y,col="red")

```

These plots suggest that the linearity assumption for age is reasonable, but there is some pronounced non-linearity in SBP. Later in the course we will investigate non-linear models for the effect of SBP.

Note that the entire model building in this assignment question has assumed that the predictors impact the response *additively*. There may well be interactions, but statistical techniques for such models are a bit complicated when there are so many predictors. So we will stick with additivity here.

- (k) Kerry Bryant enters the intensive care unit with the following predictor values corresponding to the final model:

age	38
cancer	0
SBP	150
emergency	1
hiPH	0
hiPCO2	0
coma	1

Estimate Kerry’s survival probability.

The aim of the next two questions is to teach you some simple longitudinal data analyses in R. Two longitudinal data sets are analysed. The first is the pig weight data discussed in class. The second is one on orthodontic measurements for several young subjects. This involves comparison of two groups (females and males) so is like 'longitudinal meets ANOVA'.

You should hand in the summary output and graphics.

2.
  - (a) Download the text file named `pigweights.txt` from the course web-site (Computer Code and Data page).
  - (b) Download the text file named `pigweights.Rs` from the course web-site (Computer Code and Data page). (If using Windows then make sure it gets saved as a ordinary text file.)
  - (c) Open the file `pigweights.Rs` and start an R session.
  - (d) Ignoring, for now, the parts of the R script involving `wait()`, cut and paste each line uncommented line (i.e. without the hash symbol (#) at the front) into the R session. This will give you a step-by-step appreciation of what the script does.
  - (e) By the end of the script you will have fit 2 models: an random intercept and a random intercept & slope model. The output corresponds to:
    - i. A plot of the data using 'trellis-type' graphics. Here the data for each pig is shown in a separate panel. Comparisons can be made using the 'graph paper' background.
    - ii. ML/REML estimates and 95% confidence intervals for what the notes calls  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ ,  $\sigma_u$  (not  $\sigma_u^2$ !) and  $\sigma_\varepsilon$  for both models, and  $\sigma_v$  for the second model. Note that none of the confidence intervals contain zero, so all parameters are statistically significant. In particular, the confidence interval for  $\sigma_v$  is statistically significant indicating a need for the random slope extension.
    - iii. The likelihood ratio statistic for comparing the two models. The null hypothesis corresponds to the random intercept model. The statistic is about 293 which should lead to overwhelming rejection of the random intercept model in favour of the random intercept & slope model (the distribution theory to justify this is a bit complicated, though).
    - iv. Residual plots for each model. The residual plot for the first model shows a suspicious 'bow tie' pattern. The second one is much more 'random looking'.
  - (f) To quickly go through the analysis again make sure that R is pointing to the directory containing `pigweights.Rs` and type `source("pigweights.Rs")`
3.
  - (a) Download the text file named `Orthodont.Rs` from the course web-site (Computer Code and Data page). (If using Windows then make sure it gets saved as a ordinary text file.)
  - (b) Ignoring, for now, the parts of the R script involving `wait()`, cut and paste each line uncommented line (i.e. without the hash symbol (#) at the front) into the R session. This will give you a step-by-step appreciation of what the script does. The command starting with `pobj <-` is spread over about 10-11 lines, so this should be cut-and-pasted all together.
  - (c) By the end of the script you will have:
    - i. Plotted the data in such a way that comparison among females and males is easy.

- ii. Fit a random intercept model to the data. However, the model has an additional fixed effect (compared with the random intercept model that you just fit to the pigs data) corresponding to a 'sex' indicator.
- iii. ML/REML estimates and 95% confidence intervals for the 5 parameters. The main conclusion that can be drawn from this is that the mean distance between from the pituitary to the pterygomaxillary fissure is between 0.75mm and 3.89mm lower for females than it is for males (with 95% confidence).
- iv. A residual plot for the model. It isn't too bad, but the points corresponding the two extreme standardised residuals (close to  $\pm 4$ ) may be having an undue effect on the results (especially the precision of the confidence intervals) and should be looked at more closely.

(d) To quickly go through the analysis again make sure that R is pointing to the directory containing `Orthodont.Rs` and type `source("Orthodont.Rs")`

4. The simple random intercept & slope model for longitudinal data is

$$y_{ij} = \beta_0 + U_i + (\beta_1 + V_i)x_{ij} + \varepsilon_{ij}.$$

where

$$U_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_U^2), V_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_V^2) \quad \text{independently of} \quad \varepsilon_{ij} \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2).$$

Assuming that the covariance between  $U_i$  and  $V_i$  can be taken to be negligible, derive an expression for

$$\text{Cov}(y_{ij}, y_{ij'}), \quad j \neq j',$$

and simplify as much as possible. This corresponds to the covariance between repeated measures on the same subject.