Exerciese

1. For the `Birthwt` data set, fit a suitable linear model and state your conclusions.

   Examine the residuals and report if there is any concern about the assumptions underlying your analysis.

   Submit your answer as a working `R` script, with your conclusions included as comments in the appropriate places.

2. With the `menarche` data, fit a binomial model (as done in lectures) but use three link functions, namely the `logistic`, `probit` and `cauchit`. Compare your predictions graphically, including the relative frequencies and fitted lines on the same diagram. Include a legend in the top left hand corner.

   Now consider the analysis with the data presented in *binary* form, that is with one entry for each student in the sample. [Hint: One way to get the data in binary form is as follows:

   ```
   menarche_binary <- with(menarche,
   rbind(data.frame(Age = rep(Age, Menarche), Men = T),
         data.frame(Age = rep(Age, Total-Menarche), Men = F)))
   ```

   Then the models may be fitted with `Men` as the binary response and `Age` as the predictor.]

   Show computationally that fitting the model in this form,

   (a) The estimated coefficients, their standard errors and $t-$statistics are the exactly the same as for the same model fitted with the data in frequency form,

   (b) The Deviance is *not* the same, but

   (c) If you fit sub-models, *differences* of deviance are the same for the data in both forms.

   (d) For the data in binary form, fir the model `Men ~ factor(Age)` and test the straight line model as a sub-model. What do you notice?

   (You need only do this with one of the link functions.)

   If you can, give a theoretical explaination of these results you have observed from the computation.

3. For the gamma distribution, defined as having probability densithy function

$$f_Y(y;\alpha,\phi) = \frac{e^{-y/\alpha} y^{\phi-1}}{\alpha^{\phi}\Gamma(\phi)}, \qquad 0 < y < \infty$$

   (a) Show that it belongs to the generalized linear modelling family and find the key functions $\theta(\mu)$ and $b(\theta)$;

   (b) Hence write down the Cumuland Generating Function, $K_Y(t)$ and find the mean and variance in terms of the original parameters,

   (c) Verify that $\theta(\mu)$ is an *increasing* function of μ.

   (d) Find the *natural* link.

4. The 'credit card' data set `CC` comes from a commercial bank in Switzerland. The response of inerest is the variable `credit.card.owner`, which is a binary response stating whether or not the person has a credit card with the bank. There is also a large set of candidate predictors from which to build a predictive model for the binary response, which was the purpose for which the data were collected.

```
> Attach() # your data sets
> with(CC, table(credit.card.owner))
credit.card.owner
  no  yes
 609 1011
```

   (a) Split the data into two parts of about 800 observations each, a 'training' and 'test' set. Build models from the training set and test them on the remainder. [Hint: one way to do this is

   ```
   set.seed(12354)  # choose a suitable seed
   ind <- sample(1:nrow(CC), 800)
   CCTrain <- CC[ind, ]
   CCTest <- CC[-ind, ]
   ```

   and check the sizes of both.]

   (b) Starting with any suitable model, use automatic stepwise techniques to arrive at a suitable logistic regression model. You need only consider main effect terms.

   Compare the result of using AIC and BIC as your selection crierion.

   (c) Construct two other predictive models, namely

      i. A tree model, fitted by `rpart` from the `rpart` package, and pruned by the 'One standard error' rule,
      ii. A random forest model, fitted by `randomForest` from the `randomForest` package.

   (d) For each fitted model find the 'confusion matrix' when testing it on the test set and compare each of the models by their crude error rates.

   The models you should consider are

      i. Your original logistic regression,
      ii. The stepwise model got by using AIC,
      iii. The stepwise model got by using BIC,
      iv. The tree model,
      v. The random forest model.

   For the regression models, predict '`yes`' if the predicted probability equals or exceeds 0.5. For the tree and random forest models predict with `type = "class"`.

   Submit your exercise as before as an annotated working `R` script.