Practical session on logistic regression

1 Study of Premature Birth Data

As part of a study on prenatal factors related to premature delivery in women already in preterm labor, there are 13 explanatory variables for 388 women included in the study.

The response variable (`PREMATURE`) is premature delivery (1=yes ; 0=no).
The data contain the following variables :

Table 1 – Variables available for childbirth data

Var	Description	Comment
`GEST`	gestational age at study entry	in weeks
`DILATE`	cervical dilation	in cm
`EFFACE`	cervical effacement	in %
`CONSIS`	cervical consistency	1 : soft 2 : medium 3 : firm
`CONTR`	presence of contractions	1 : yes 2 : no
`MEMBRAN`	membranes	1 : ruptured 2 : not ruptured 3 : uncertain
`AGE`	mother's age	in years
`STRAT`	pregnancy period	1-4
`GRAVID`	gravidity	number of previous pregnancies including the current one
`PARIT`	parity	number of previous full-term pregnancies
`DIAB`	diabetes issue	1 : present 2 : absent
`TRANSF`	transfer to a specialized hospital	1 : yes 2 : no
`GEMEL`	pregnancy type	1 : single 2 : multiple

The objective is to define the predictive factors for premature delivery (Y). For each considered model, denote π as the probability of premature delivery given the variables $X_1, \cdots, X_p$ included.

1. Load the dataset into a table `prema`, obtain the summary, and check that nominal qualitative variables are indeed factors (necessary for logistic regression). If needed, use the `as.factor()` command.

```
load("prema.RData")
str(prema)
prema$DIAB = as.factor(prema$DIAB)
attach(prema)
```

Study of a Binary Variable

2. Construct the contingency table PREMATURE/GEMEL.
3. Calculate the probability of premature delivery for a multiple pregnancy.
4. Fit the model explaining premature delivery by pregnancy type GEMEL (`model1`).

```
model1 <- glm(PREMATURE ~ GEMEL, family = "binomial", data = prema)
summary(model1)
```

5. Is the coefficient associated with the variable GEMEL significant ? Retrieve the odds ratio associated in two different ways.

Study of a Quantitative Variable

6. What is the average cervical effacement in patients who delivered prematurely ? In others ? (You can use the `by` function for this.)
7. Fit the model explaining premature delivery by cervical effacement (`model2`).
8. Express $\pi(x) = P(\text{PREMATURE} = 1/\text{EFFACE} = x)$ as a function of x and write an R function to perform this calculation.
9. What is the probability of premature delivery when the cervix is effaced at 60% ?
10. Use the previously written function to calculate the π score associated with the women in the study. Compare this score to the results returned by the following commands :

```
pi_hat = predict(model2, prema, type = "response")
model2$fitted.values
```

11. Create a graph illustrating the dependence between cervical effacement and premature delivery. For example, plot two densities corresponding to the score distributions in the two groups using the following commands :

```
library(lattice)
gS = densityplot(~pi_hat, data = data.frame(prema, pi_hat), groups = PREMATURE,
    plot.points = FALSE, ref = TRUE, auto.key = list(columns = 1))
print(gS)
```

Study of Multiple Explanatory Variables

12. Fit the model explaining premature delivery by pregnancy type and cervical effacement (`model3`) :

```
model3 <- glm(PREMATURE ~ GEMEL + EFFACE, family = "binomial",
    data = prema)
summary(model3)
```

13. Compare the two models `model2` and `model3` using the likelihood ratio test :

```
anova(model2, model3, test = "LRT")
```

14. Which model do you choose ?
15. Estimate the complete model (`fullmodel`) :

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

```
fullmodel <- glm(PREMATURE ~ ., family = "binomial", data = prema)
summary(fullmodel)
```

16. Evaluate the significance of each coefficient in `fullmodel`. Use the `step` function for automatic variable selection in the model and interpret. Call `reduced` the model reduced to the selected variables. Compare the two models (full and reduced).
17. Interpret the coefficients of `reduced`. What are the risk factors for premature delivery ? What are the protective factors ?
Hint :

```
exp(cbind(OR=coef(reduced), confint(reduced)))
```

Evaluation of the Decision Rule

18. Calculate the predicted values of the probabilities of interest using the `predict` function or the `fitted.values` field of `reduced`. Name this new score S. Visualize and comment on the prediction quality (e.g., plot boxplots).
19. Calculate the confusion matrix for a decision threshold of 0.5.
20. Arbitrarily decide to assign all values with an S score higher than the score of the last row to group 1 and the others to group 0. Then calculate sensitivity and specificity for this threshold.
21. Plot the ROC curve associated with the S score using the `prediction` and `performance` functions from the `ROCR` package.

```
library(ROCR)
pred = prediction(S, prema$PREMATURE)
perf = performance(pred, "tpr", "fpr")
plot(perf)
```

22. Explore the objects that allow you to calculate the ROC curve :

```
perf@x.values[[1]]
perf@y.values[[1]]
perf@alpha.values[[1]]
```

23. Calculate the area under the ROC curve using the following commands :

```
AUC = performance(pred, "auc")
attr(AUC, "y.values")[[1]]
```

24. Calculate the threshold closest to the ideal point for the ROC curve related to the S score. Calculate the new confusion matrix associated with this decision threshold.