Generalized Linear Modeling

The purpose of this TP is to practice using **R** for two types of generalized linear modeling: logistic regression and Poisson regression. We will apply GLM to two problems with data in the **HSAUR3** package. To carry out these analyses, first load **R** and the **HSAUR3** package.

As usual, make sure that you read the **help** for any new functions and data that you use.

Blood plasma data

Load the **plasma** data and read the help for this data set so that you understand the scientific background and question of interest as well as the variables included in the data. Then start off making a graphical exploration of the data. Here, we will look at **conditional density plots** of the response variable ESR given each each of the (numerical) explanatory variables fibrinogen and gamma globulin change.

```
data("plasma", package = "HSAUR3")
layout(matrix(1:2, nrow = 2))
cdplot(ESR ~ fibrinogen, data = plasma)
cdplot(ESR ~ globulin, data = plasma)
```

The conditional density for each ESR category (smaller or larger than 20) is shown in a different shade, with the numerical value on the right hand vertical scale. Here, we can see that **higher** levels of each protein are associated with ESR values above 20, since the conditional probability for ESR bigger than 20 in both plots is larger as we move to the right of the plot horizontally (larger values of fibrinogen/gamma globulin).

Now, fit a logistic regression model to the data with the **glm** function, including only the single variable **fibrinogen**:

```
plasma.glm.1 <- glm(ESR ~ fibrinogen, data = plasma, family = binomial())
```

As is the case for the **lm** function, an intercept term is automatically included in the model. The **family** argument for **glm** specifies the distribution of the response - in this case, a binomial distribution. The default **link** function for the binomial family is **logistic**.

Apply the **summary** function to your fitted model to obtain a description of the fitting. Is the coefficient for **fibrinogen** significant at the 5% level? Interpret the meaning of this coefficient.

You can obtain a 95% confidence interval for the coefficient (make sure to read the help):

```
confint(plasma.glm.1, parm = "fibrinogen")
```

Since these values correspond to the **log-odds**, we can get the **odds** by exponentiating the estimate (subsetting **coef** according to the value that we want; here it is for **fibrinogen** and not the intercept) and confidence interval:

```
exp(coef(plasma.glm.1)["fibrinogen"])
exp(confint(plasma.glm.1, parm = "fibrinogen"))
```

We can see that the confidence interval is very wide. Can you think of any reason why this might be?

Now, fit a logistic regression model using both explanatory variables:

```
plasma.glm.2 <- glm(ESR ~ fibrinogen + globulin, data = plasma, family = binomial())
```

and output the results using **summary**. Is the coefficient for gamma globulin significantly different from 0?

You can perform a likelihood ratio (chi-square) test by subtracting the residual deviance of the second (bigger) model from that of the first (smaller) model, then comparing the result to a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the two models:

```
anova(plasma.glm.1, plasma.glm.2, test = "Chisq")
```

Here, we see that the *p*-value is large, meaning that we do **not** reject the null hypothesis that the coefficient for **globulin** is zero.

Even though we would not include **globulin** in our final model, for purposes of illustration we will use model 2 to obtain predicted values for each observation, then plot these against the values of **both** explanatory variables using the **symbols** to create a **bubbleplot**. The estimated conditional probability of an ESR value larger than 20 is obtained by:

```
prob <- predict(plasma.glm.2, type = "response")
```

and then assign a larger circle to observations with larger probability:

```
plot(globulin ~ fibrinogen, data = plasma, xlim = c(2,6), ylim = c(25,55), pch = "*")
symbols(plasma$fibrinogen, plasma$globulin, circles = prob, add = TRUE)
```

This plot shows an increasing probability of ESR bigger than 20 (larger circles) with increasing fibrinogen and, to a lesser extent, with increasing gamma globulin.

Briefly summarize your findings.

Colon polyps data

Load the **polyps** data and read the help.

You will see that the response variable is a *count*. We have already used multiple regression to model a numerical response that is *continuous*, but there are problems with using this approach for count data: a count can only take on *positive* values, and a count is unlikely to be *normally distributed*. So here, we will fit a GLM with a *log link function*, thus ensuring that the fitted values are positive, and with a *Poisson error* distribution - i.e., **Poisson regression**:

```
data("polyps", package = "HSAUR3")
polyps.glm.1 <- glm(number ~ treat + age, data = polyps, family = poisson())
```

By default, the **link** function for the Poisson family is the **log** function.

Look at the results of the fitting (**summary**). Does there seem to be a larger variance in observed counts than expected from the Poisson assumption? How can you tell?

Now use the *quasi-likelihood* approach to deal with this over-dispersion, by specifying the **quasipoisson** family:

```
polyps.glm.2 <- glm(number ~ treat + age, data = polyps, family = quasipoisson())
```

then look at the results of the fitting. How do these results compare with those of the first fitting? Can you explain why they are different?

Briefly summarize your results and conclusions: does the drug treatment appear to be effective in reducing the number of polyps? Does the age of the patient have a large effect?