STATISTICAL MODELLING Part IIC

Practical 7: Poisson regression IAC/Lent 2011

Comments and corrections to ioana@statslab.cam.ac.uk

Download the `AidsData` from the course web page and save it in your `RWork` directory in a file called `aids.txt`. It gives the number of reported new cases of AIDS in the UK for 36 consecutive months up to November 1985. Open R, read in the data.

```
> y <- scan("aids.txt", comment.char = "#")
> Month <- 1:36
```

Exercise: Plot the data as a function of `Month`.

Fit a generalised linear model with

```
> PoiMod <- glm(y ~ Month, family = poisson)
> summary(PoiMod)
```

```
Call:
glm(formula = y ~ Month, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4196  -1.1553  -0.2742   0.7264   2.8500

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03966    0.21200   0.187    0.852
Month        0.07957    0.00771  10.321   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 190.17  on 35  degrees of freedom
Residual deviance:  62.36  on 34  degrees of freedom
```

```
AIC: 177.69

Number of Fisher Scoring iterations: 5
```

Presumably Poisson is not offended that his name must be entered with a lower case 'p'. Write down the model that is being fitted here. Much of the information is very similar to that presented in the binomial regression examples in Practical 6. How would you compute the estimates of the parameters? By evaluating the Fisher information matrix at the maximum likelihood estimators of the parameters, verify the calculations leading to the standard errors. How are the z- and p-values obtained?

```
> X <- model.matrix(y ~ Month)
> W <- diag(PoiMod$weights)
```

The standard errors are computed as

```
> sd.error <- sqrt(diag(solve(t(X) %*% W %*% X)))
```

Then the z-values and p-values follow

```
> est <- coef(PoiMod)
> z <- est/sd.error
> apply(matrix(z, nrow = 2), 1, function(q) {
+     2 * (1 - pnorm(q, 0, 1))
+ })

[1] 0.8515975 0.0000000
```

The residual deviance is $D(y; \hat{\mu})$. In lectures, we derived the expression

$$D(y; \hat{\mu}) := \sum_{i=1}^{n} d_i = 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right] = 2 \sum_{i=1}^{n} y_i \log \frac{y_i}{\hat{\mu}_i}$$

for the deviance, where the right-hand side is a simplification for the model containing an intercept term. When we try to use R to verify this formula, however, it complains about having to evaluate terms in the sum for which $y_i = 0$.

Exercise: What should the contribution to the sum from such terms be? Now go ahead and use R to verify that the residual deviance is what you would expect. How does Pearson's χ^2 statistic compare? Recall that Pearson's χ^2 statistic is defined as

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

Exercise: Why might it not be a very good approximation to say that the residual deviance has a χ^2_{34} distribution if our model with $\log \mu_i = \alpha + \beta i$ is correct? In such circumstances, the residual plots are even more useful, so examine these. In particular, we want to obtain a plot of fitted values against standardised deviance residuals. Recall that the standardised deviance residuals are defined as

$$r_i = \text{sign}(y_i - \hat{\mu}_i)\frac{\sqrt{d_i}}{\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n,$$

where h_{ii} is the i*th* diagonal element in the hat matrix H for regression in the *IWLS* algorithm,

$$H = W^{1/2}X\left(X^TWX\right)^{-1}X^TW^{1/2}.$$

Check that your computations agree with the output of the command `rstandard(PoiMod)`. An alternative would be to combine consecutive months in some way to ensure each of our fitted values is at least 5, say.

The null deviance is $D(y; \hat{\hat{\mu}}_0)$, where $\hat{\hat{\mu}}_0 = \exp(\hat{\hat{\alpha}})$, and $\hat{\hat{\alpha}}$ is the maximum likelihood estimator of α in the model in which $Y_1, \ldots, Y_n$ are assumed to be independent with $Y_i \sim \text{Poi}(\mu_i)$, and $\log(\mu_i) = \alpha$ for all i. Verify the calculation.

Exercise: Return to your initial plot of `Month` versus `y`. Add the fitted line to your plot. What would you have concluded in November 1985?

The next data set is `MissingData`.

Copy the file `missing.txt` from the course web page and save it in your `Rwork` directory. Read in the data as a table and attach the column headings.

```
> MissingData <- read.table("missing.txt", header = TRUE)
> names(MissingData)
```

```
[1] "Sex"   "Age"   "n"    "Still"
```

```
> attach(MissingData, warn.conflicts = FALSE)
```

Here, `n` is the number of people in a particular age and sex category reported missing in a year, and `Still` is the number still missing at the end of the year. The three age categories are: 1) 13 years and under, 2) 14-18 years, 3) 19 years and older.

The first command below doesn't plot the points, due to the `type="n"` option. It does, however, set up the plotting window for the next command.

```
> plot(Age, Still/n, type = "n", main = "MissingData",
+     xlab = "Age", ylab = "Still/n")
> text(Age, Still/n, c("F", "M")[Sex])
> is.factor(Age)

[1] TRUE

> is.factor(Sex)

[1] TRUE

> Age <- factor(Age)
```

Figure 1 shows the plot.

EXERCISES: We compare a Poisson regression model with a binomial logistic regression model.

1. Write down the model being fitted below, explaining why we need to include an offset.

```
> PoiMod1 <- glm(Still ~ Age + Sex, family = poisson,
+     offset = log(n))
> summary(PoiMod1)

Call:
glm(formula = Still ~ Age + Sex, family = poisson, offset = log(n))

Deviance Residuals:
       1          2          3          4          5          6
```

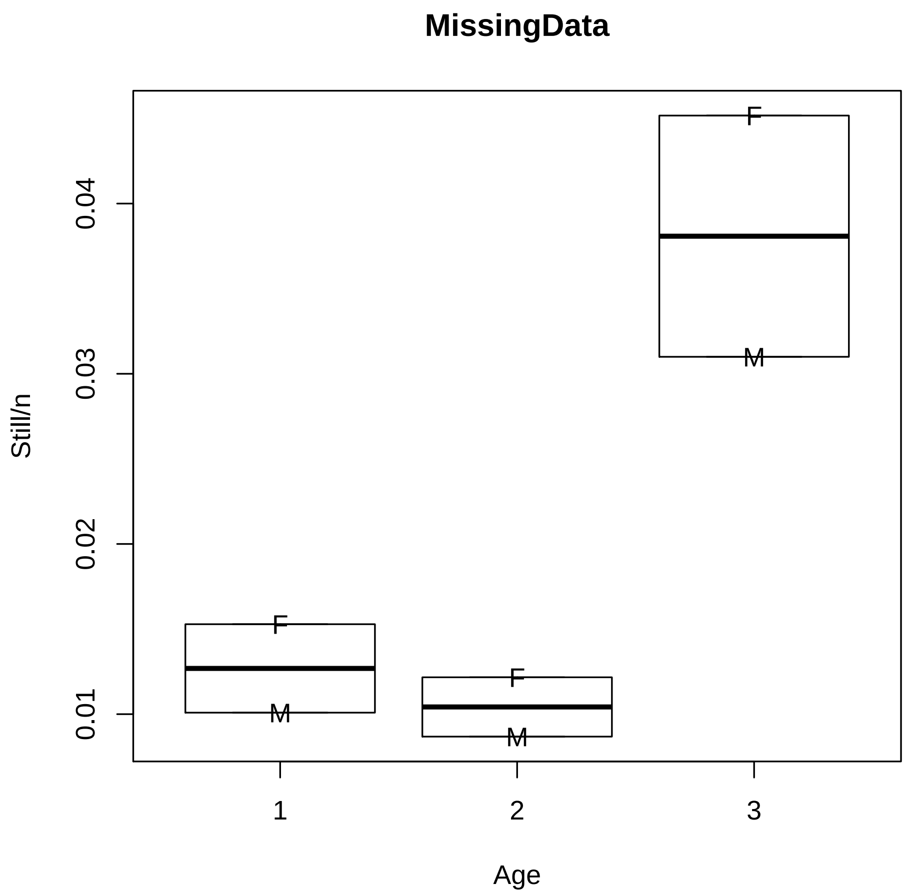

Figure 1: Plot of `MissingData`

```
-0.13819   0.16462  -0.03965   0.13074  -0.12437   0.03949

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2021     0.1255 -33.484  < 2e-16 ***
Age2         -0.1950     0.1415  -1.378    0.168
Age3          1.1017     0.1313   8.387  < 2e-16 ***
SexM         -0.3703     0.0857  -4.320 1.56e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 217.10061  on 5  degrees of freedom
Residual deviance:   0.08189  on 2  degrees of freedom
AIC: 45.209

Number of Fisher Scoring iterations: 3

> text(as.real(Age), fitted(PoiMod1)/n, c("f", "m")[Sex])
```

2. Now fit a binomial logistic regression model and compare the results. Do either/both of the models fit the data? How would you interpret the output? Can you quantify the change in odds of still being missing at the end of the year if you are female as opposed to male? What if you are 19 years old or over?

3. Could we have the same parameter for two of the age categories? Create a new factor, taking values `"Young"` and `"Old"` and compare Poisson and logistic binomial additive regression models. Do they still fit the data satisfactorily? Furthermore, perform a likelihood ratio test of the reduced model with 2 age categories against the model with 3 age categories. This can be done using the `anova` command by specifying that you're performing a χ^2 test, i.e., `anova(model1, model2, test='Chisq')`.