University of Edinburgh
School of Mathematics

Generalised Regression Models

GRM: Case Study — GLMs **Semester 1, 2022–2023**

Problem and Data

The file `Books.dat` contains the data from a survey of pupils from three schools, asking how many books they they have read outside of school over the last year may. The variable in the data set are:

Variable	Description
books	The number of books read in the last year outside of school.
school	The school that the pupil attends (`Sch1/Sch2/Sch3`).
gender	Gender of the pupil identified at birth (`M/F`).
math	Standardized maths test scores.
eng	Standardized English test scores.

The objective is construct a generalised linear regression model that best describes the number of books a pupil may read within a year. Use this model to predict how many books a pupil reads who is at the 75^{th} percentile in both English and maths (i.e. a standardized score of $\Phi^{-1}(0.75) = 0.674$) for each school and both gender identities.

Statistical Model

The number of read books is a count variable that can be modelled using a Poisson distribution. Therefore we need to fit the following generalised linear regression model:

$$Y_i \sim \text{Poisson}(\lambda_i) \quad \text{for} \quad i = 1, \ldots, n \quad \text{where} \tag{1}$$

$$\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} \tag{2}$$

Here, we assume that number of books (Y_i) are independent Poisson random variables with some unique rate parameter λ_i. The goal of the model is to construct a relationship between the rate parameter and the covariates $(x_{1,i}, \ldots, x_{p,i})$ by regressing through the mean. The expectation for the Poisson distribution is strictly positive ($\mathbb{E}[Y_i] = \lambda_i$) and so need to be transformed onto the same real-space as the linear predictor. We achieve this through the log-transformation which is the canonical link function for Poisson regression model. The aim is then to estimate the regression coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ in order to evaluate how important each covariate is in explaining the variation seen in the number of books read by the pupils.

R analysis

Download and save the data file and read it into R:

```
Book <- read.table(file = "Book.dat")
```

Explore the data

Quiz 1: Briefly explore the the data to identify what relationships exists within the data by creating plots and evaluating appropriate summary statistics.

Building the Poisson regression model

To fit a generalised linear regression model we use the `glm()` command. The input are:

`formula`	A symbolic description of the model to be fitted in the form `y ~ x` where `y` is the response variable and `x` is some linear combination of the explanatory variable.
`data`	A data frame containing the variables in the model.
`family`	A description of the distribution and link function to be used in the model.

Using the `glm()` command is very similar to using the `lm()` for normal linear regression models. The only difference is that we need to specify the distribution family. The canonical link function for each distribution is used by default, but alternatives could be specified.

Quiz 2: Run `help(family)`. What family of distributions are available for the `glm()` command. What link functions can be used for the Poisson model?

Let's begin by fitting a simple Poisson regression model with maths scores as the only covariate:

```
Model1 <- glm(formula = books ~ math, data = Book, family = poisson)
summary(Model1)
```

From the results we can write the equation for the fitted curve as:

$$\log(\mathbb{E}[Y]) = 1.208 + 0.403x \quad \text{or} \quad \mathbb{E}[Y] = e^{1.208+0.403x} = 3.345\ e^{0.403x}$$

From the summary we see that the maths score covariate is important in this model with a significant co-efficient. To interpret this estimate, consider a similar pupil who has one extra unit increase in their maths score, then their expected number of read books is:

$$\mathbb{E}[Y'] = 3.345e^{0.403(x+1)} = 3.345e^{0.403x} \times e^{0.403} = \mathbb{E}[Y] \times 1.496.$$

In otherwords, with all other variables held constant, a unit in a pupil's math score increases increases their expected number of read books by a *multiplicative factor* of $e^{\hat{\beta}_1} = 1.496$.

Copy and run the following code to evaluate which covariates are important in describing the variation in the number of books read:

```
ModelAll <- glm(formula = books ~ ., data = Book, family = poisson)
ModelBest <- step(ModelAll)
summary(ModelBest)
```

Quiz 3: What are the covariates that are in the best model? Is this what you would have anticipated from exploring the data.

Quiz 4: Write down the formula for the fitted regession line and interpret the estimates.

Before using the fitted model, we first need to validate the modelling assumptions:

```
par(mfrow = c(2, 2))
plot(ModelBest)
```

Unlike with the normal linear regression model, we can see some clear patterns in these plots where a sequence of points seem to follow a curve. However, this is a pattern that we may expect to see for the Poisson regression model. For example, if the value of the linear predictor is low, say 0.5, then the expected value for the number of books read is $\mathbb{E}[Y] = e^{0.5} \approx 1.649$ which means that the probability of reading at one or no books is 0.509 (`ppois(1, exp(0.5))`). The number of unique outcomes that we are likely to see for the Poisson distribution reduced for lower or negative values of the linear predictor, leading to the observed patterns the plots.

Despite this, the overall structure in the residuals vs predicted illustrate no broad pattern and so we may suggest that the variation in the number of books read is being captured by the linear predictor. The Q-Q plot is of the deviance residuals and appear to satisfy the approximate normality assumption.

Prediction

We can now address the original question and predict predict how many books a pupil reads who is at the 75^{th} percentile in both English and maths for each school and both gender identities. For this, we first need to construct a data frame containing the information for all six scenarios:

```
PredictData <- data.frame(
  school = c("Sch1", "Sch2", "Sch3", "Sch1", "Sch2", "Sch3"),
  gender = c("M", "M", "M", "F", "F", "F"),
  math   = rep(qnorm(0.75), rep = 6),
  eng    = rep(qnorm(0.75), rep = 6)
)
```

For prediction we use the `predict()` command where we supply the results from fitting the best model and data frame of the new covariates that we have just defined:

```
predict(ModelBest, newdata = PredictData)
```

This command returns six values, one for each row in `PredictData`. However, this is a prediction according to the linear predictor, not for the expected number of books. To get what we require
This function returns the predicted value according to the fitted linear predictor, and so must be converted to the original scale in order to interpret the values with respect to the original scale of the response variable. To do this we can add the `type` argument into the `predict()` command as follows:

```
predict(ModelBest, newdata = PredictData, type = "response")
```

Quiz 5: Create a new data frame to predict the number of books read by:

A) a pupil at school 2 at the 80^{th} percentile in both subjects, and

B) a pupil at school 3 at the 25^{th} percentile in both subjects.

Quiz 6: Which pupil is expected to read more books outside of school?