University of Edinburgh
School of Mathematics

Generalised Regression Models

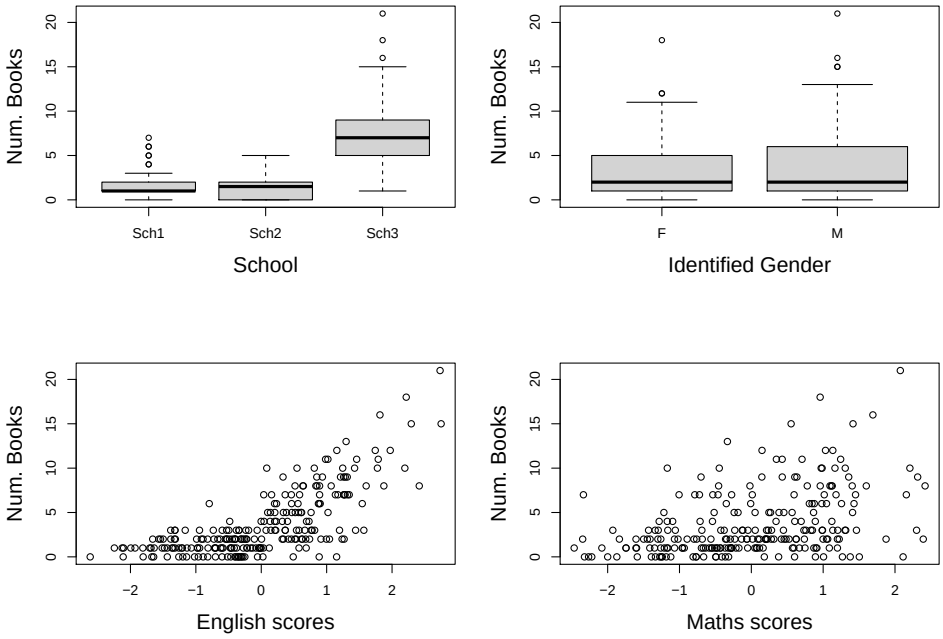
GRM: Case Study — GLMs **Semester 1, 2022–2023**

FEEDBACK

Explore the data

Quiz 1: Load the data and explore the data to identify any relationships between the number of read books against the other variables, and amongst the covariates:

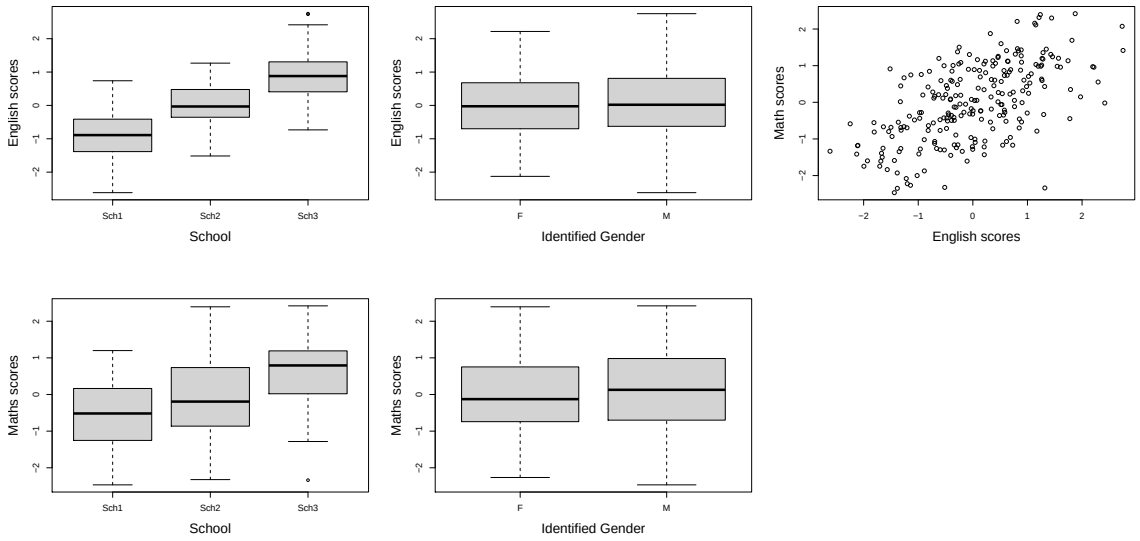
```
Book <- read.table(file = "Books.dat")

boxplot(Book[, "books"] ~ Book[, "school"])
boxplot(Book[, "books"] ~ Book[, "gender"])
plot(Book[, "eng"], Book[, "books"])
plot(Book[, "math"], Book[, "books"])
```



There is very little difference between schools 1 and 2, but the pupils at school 3 read many more books. There is no difference in the number of books between the identified male or female pupils. The number of books read is positively related with both English and maths test scores, but more strongly with English scores than for Maths scores.

```
boxplot(Book[, "eng"] ~ Book[, "school"])
boxplot(Book[, "math"] ~ Book[, "school"])
boxplot(Book[, "eng"] ~ Book[, "gender"])
boxplot(Book[, "math"] ~ Book[, "gender"])
plot(Book[, "eng"], Book[, "math"])
```

English and Maths test scores are positively correlated (correlation: 0.593), and the scores for both subjects are higher on average at school 3 and lower on average at school 1. There are no difference in score for both subjects with respect to pupil's identified gender at birth.



Building the Poisson regression model

Quiz 2: Run and read `help(family)`.

The `glm()` can fit a range of different models. The common distributions are `gaussian` for the normal linear regression model (i.e. `lm()`), `poisson` for the Poisson regression model, and `binomial` for the logistic regression model.

The available link functions for the Poisson distribution are `"log"` (the canonical link function), `"identity"` (no transformation of the expectation, but beware of out-of-range negative predictions!) and `"sqrt"` (e.g., $\mathbb{E}[Y] = (\beta_0 + \beta_1 x)^2$).

The results from the first Poisson regression model based on the standardized maths scores are:

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


Call:
glm(formula = books ~ math, family = poisson(link = "log"), data = Book)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9587  -1.3355  -0.5679   0.6676   4.5308

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.20754    0.03695   32.68   <2e-16 ***
math         0.40286    0.03345   12.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 801.52  on 236  degrees of freedom
Residual deviance: 651.66  on 235  degrees of freedom
AIC: 1274.5
Number of Fisher Scoring iterations: 5
```

Both intercept and maths covariate are significant, resulting in the fitted model for the expected number of books being $\mathbb{E}[Y] = e^{1.208+0.403x} = 3.345e^{0.403x}$.

To evaluate the best Poisson regression model, we first fit a model using all of the available covariates and use the `step()` function to perform variable selection:

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

Call:
glm(formula = books ~ school + eng, family = poisson(link = log),
    data = Book)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4622  -0.7803  -0.0685   0.5757   2.5689

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.92124    0.08783  10.489  < 2e-16 ***
schoolSch2  -0.51954    0.13314  -3.902 9.54e-05 ***
schoolSch3   0.54290    0.12513   4.339 1.43e-05 ***
eng          0.50075    0.04678  10.705  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 801.52  on 236  degrees of freedom
Residual deviance: 238.94  on 233  degrees of freedom
AIC: 865.8
Number of Fisher Scoring iterations: 5
```

Quiz 3: The covariates in the best model are the pupil's school and the standardized English scores. For the variables that are not included, the data exploration showed that the number of books read did not appear to be related with pupil's identified gender at birth, and the standardised English and maths scores are correlated.

Quiz 4: The fitted regression curve are:

$$\mathbb{E}[Y] = \begin{cases} \exp(0.921 + 0.501x) & \text{if at school 1} \\ \exp(0.402 + 0.501x) & \text{if at school 2} \\ \exp(1.464 + 0.501x) & \text{if at school 3} \end{cases}$$

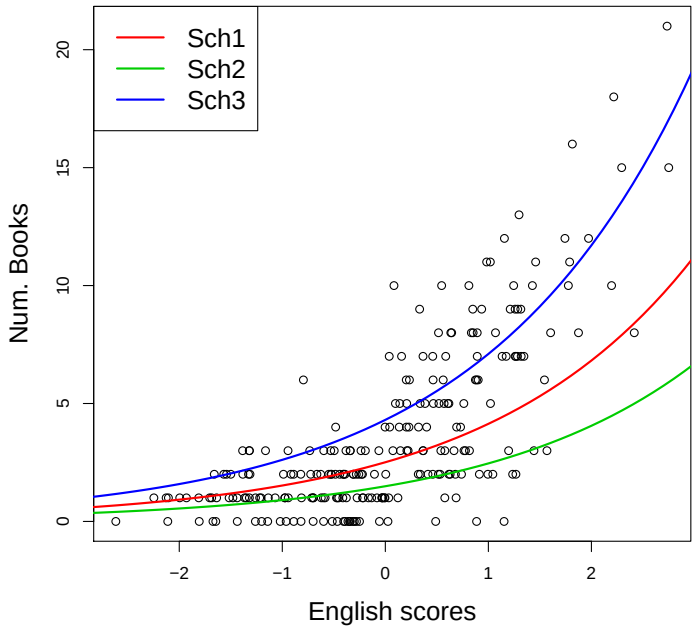
where x represents the pupil's English scores. The following creates an image of the regression function for each school:

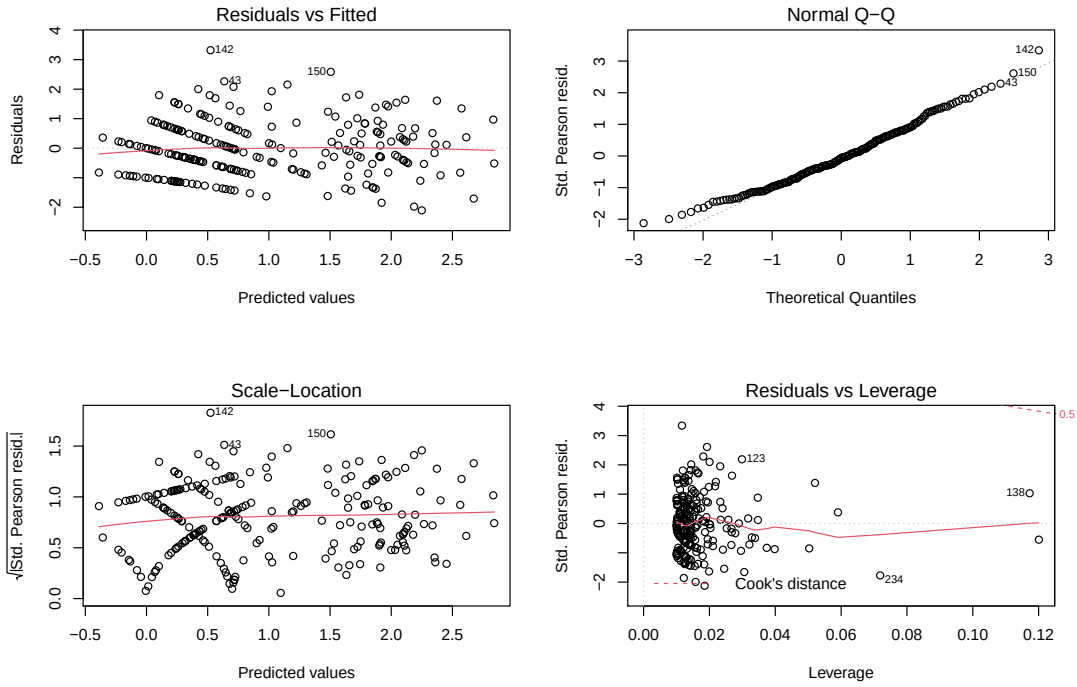
```
plot(Book[,"eng"], Book[,"books"])
x <- seq(from = -3, to = 3, by = 0.1)
lines(x, exp(0.92 + 0.50*x), col = 2, lwd = 2)
lines(x, exp(0.92 - 0.52 + 0.50*x), col = 3, lwd = 2)
lines(x, exp(0.92 + 0.54 + 0.50*x), col = 4, lwd = 2)
legend("topleft", legend = c("Sch1","Sch2","Sch3"),
       col = c(2,3,4), lty = 1,lwd = 2, cex = 1.5)
```

Using the plot command on the best model allows us to assess the modelling assumptions.

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

The lines that appear in these plots are a result of the discrete nature of the response variable. Despite this, there are no concerns that the modelling assumptions are being violated.





Prediction

To answer the original question, we first need to construct a data frame containing the covariates for each of the 6 cases that we want to predict for:

```
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)
)
```

The following evaluates the regression line (i.e $\log(\mathbb{E}[Y])$) for each row in `PredictData`:

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

```
       1        2        3        4        5        6
1.258993 0.739457 1.801892 1.258993 0.739457 1.801892
```

The following predicts the expected number of books ($\mathbb{E}[Y]$) for each row in `PredictData`:

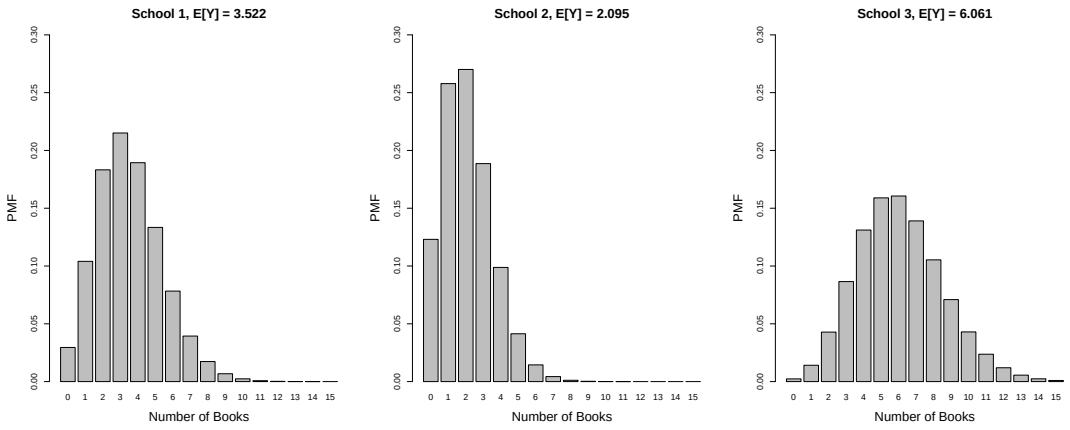
```
predict(ModelBest, newdata = PredictData, type = "response")
```

```
       1        2        3        4        5        6
3.521874 2.094798 6.061104 3.521874 2.094798 6.061104
```

Note that the best model does not use the identified gender covariate, so the predicted value for cases 1-3 are identical for cases 4-6.

The following barcharts helps to illustrates the difference in probability mass function in predicting the number of books read for the three pupils at different schools, all at the 75^{th} percentile in standardized English scores.

```
par(mfrow=c(1,3))
barplot(dpois(0:15, 3.522), names = 0:15)
barplot(dpois(0:15, 2.095), names = 0:15)
barplot(dpois(0:15, 6.061), names = 0:15)
```



Quiz 5: The expected number of books for the two pupils can be evaluated as follows:

```
PredictData2 <- data.frame(
  eng = c(qnorm(0.8), qnorm(0.25)),
  school = c("Sch2", "Sch3")
)
predict(ModelBest, newdata = PredictData2, type = "response")
```

```
       1        2
2.277661 3.084483
```

Quiz 6: The expected number if books read outside of school by pupil A is 2.3 whilst pupil B is anticipated to read more with an expectation of 3.1.