MAS61001-GENERALIZED LINEAR MODELS PROJECT

JAMES HALLAM - 200234131

1. Introduction

In an experiment to monitor and improve student attendance in the tutorials of Level 1 in the School of Mathematics and Statistics of the University of Sheffield, a study was conducted for a single module. The number of students attending each tutorial of the module in Weeks $1-11$ was recorded together with an index measuring how on average the students performed in the homework questions (minimum 0 for a poor performance and maximum 2.5 for excellent performance). The aim of this project is to fit a generalized linear model that accurately models the data and can predict the response variable accurately.

2. Exploratory Data Analysis

All of the variables are numerical and we are not supplied with the knowledge that the data has missing values and so this can not be explored. As the data is grouped into weeks, we are unsure on the individual scores of each student each week. Because of this we can not explore the possibility of outliers in the data. However, we can explore possible trends and relationships between variables.

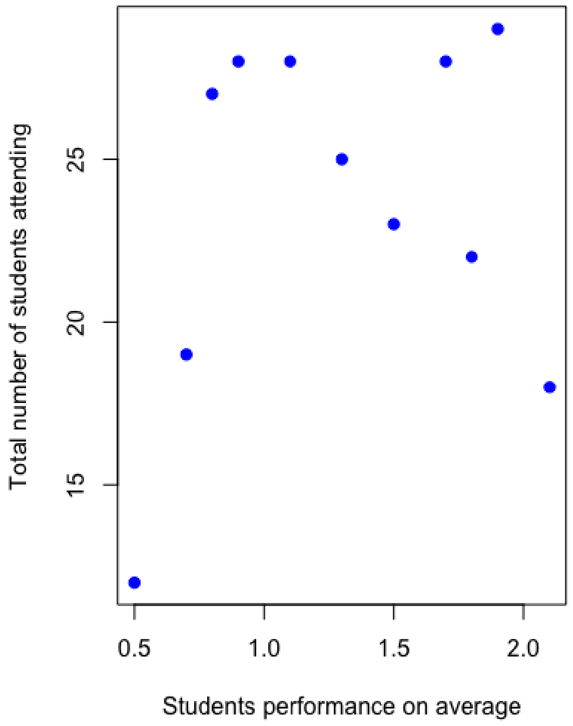

Figure 1. Plot of student attendance against student mean performance

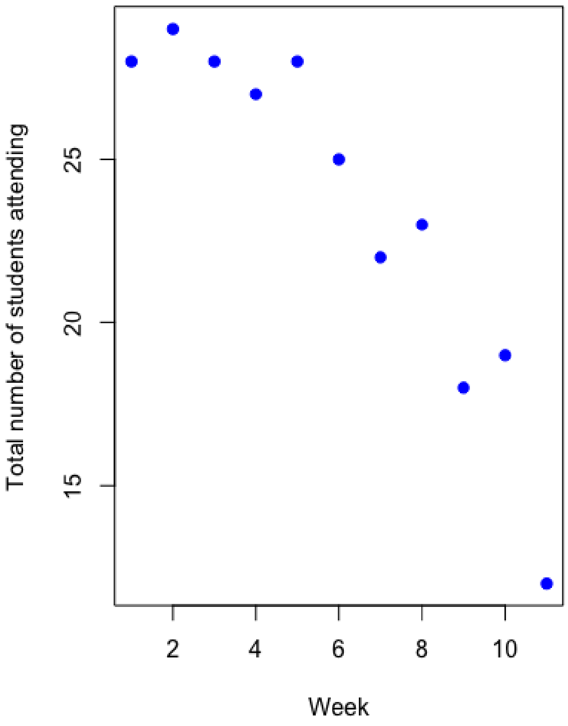

Figure 2. Plot of student attendance the week

Figure 2 shows some linearity between week and total number of students, however there is some curvature. There is a correlation of -0.92, agreeing with the fairly strong negative correlation seen in the plot. Figure 1 shows no visual correlation between the index and the total number of students, this has a correlation of 0.25. The correlation between week and index is -0.31. This suggests that the variable week is the only variable that has a strong correlation with the total number of students and so it will be the first predictor variable added to the linear GLM.

3. Fit a suitable model

A Poisson distribution is appropriate as the data is count data and is not proportional. A binomial model is not appropriate for this data. With a Poisson distribution model we're trying to figure

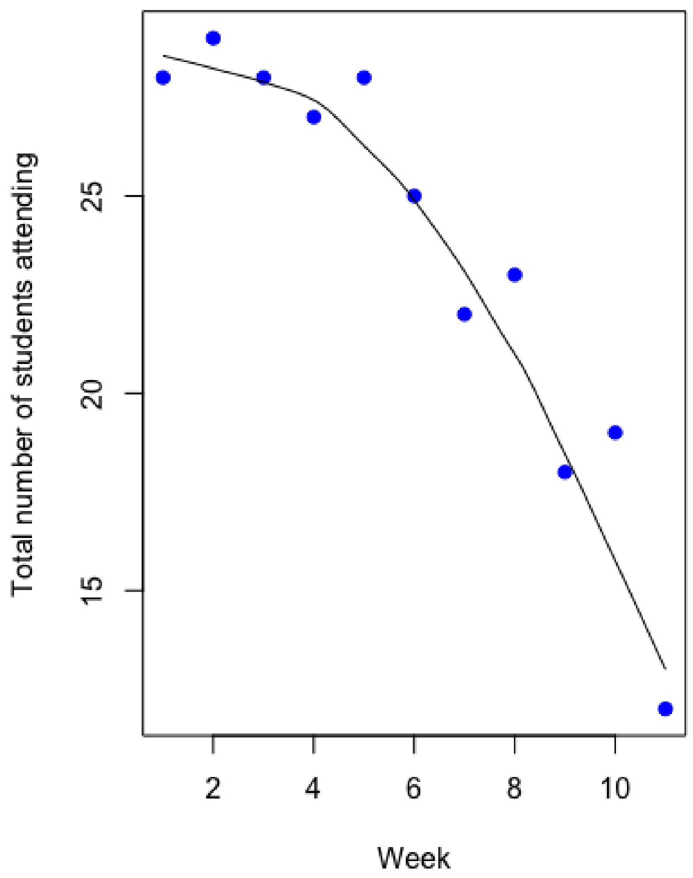

Figure 3. Week values against total number of students with a regression line

out how the predictor variables affect a response variable, here the response variable is the total number of students each week and the predictor variables are week and index. The default link function for a Poisson model is *log*.

The null model was fitted and the null deviance was 13.52 to $2.d.p$ (on 10 df).

```
poisson.null <- glm(total ~ 1, family=poisson(link="log"), data=df)
```

Due to the correlation between the variables week and total number of students, the GLM with the single predictor variable week was fitted first.

```
poisson1 <- glm(total ~ week, family=poisson(link="log"), data=df)
```

This produced a deviance of 3.09 to $2.d.p$ (on 9 df). The change in deviance from the null model to the linear model is 10.43 (on 1 df). $\chi^2_{1,0.95} = 3.84$ and so this is evidence to continue with the linear model over the null model. This model is a suitable fit for the data.

4. Compare and contrast models

We can assess the fit of the models using the residual deviance since $\phi = 1$. Using the `anova` function in R to perform deviance analysis. If we consider the variable index as a covariate, then using a GLM with a single explanatory variable being index gives a residual deviance of 12.74 to $2.d.p$ (on 9 df).

```
poisson2 <- glm(total ~ index, family=poisson(link="log"), data=df)
```

Then we compare this to the use both week and index as variables in the Poisson model.

```
poisson3 <- glm(total ~ index + week, family=poisson(link="log"), data=df)
```

This gives a deviance of 3.08 to $2.d.p$ (on 8 df). The change in deviance is 9.66 $2.d.p$ (on 1 df). This is significant as $\chi^2_{1,0.95} = 3.84 < 9.66$ and so provides evidence to continue with this model over the first model. Adding a quadratic term $week^2$, we get a residual deviance of 1.45.

```
poisson4 <- glm(total ~ index + I(week^2), family=poisson(link="log"), data=df)
```

This is lower than the previous model and in Figure 6 there is a clear improvement in the fitted values. However the change in deviance $(3.09 - 1.45 = 1.64 < \chi^2_{1,0.95} = 3.84)$ is not significant enough to suggest that there is evidence to change from the first model. This decrease in deviance may suggest that the model is becoming over fit.

Model	**Residual Deviance**
index	12.74
index + week	3.08
index + $week^2$	1.45

TABLE 1. Model results

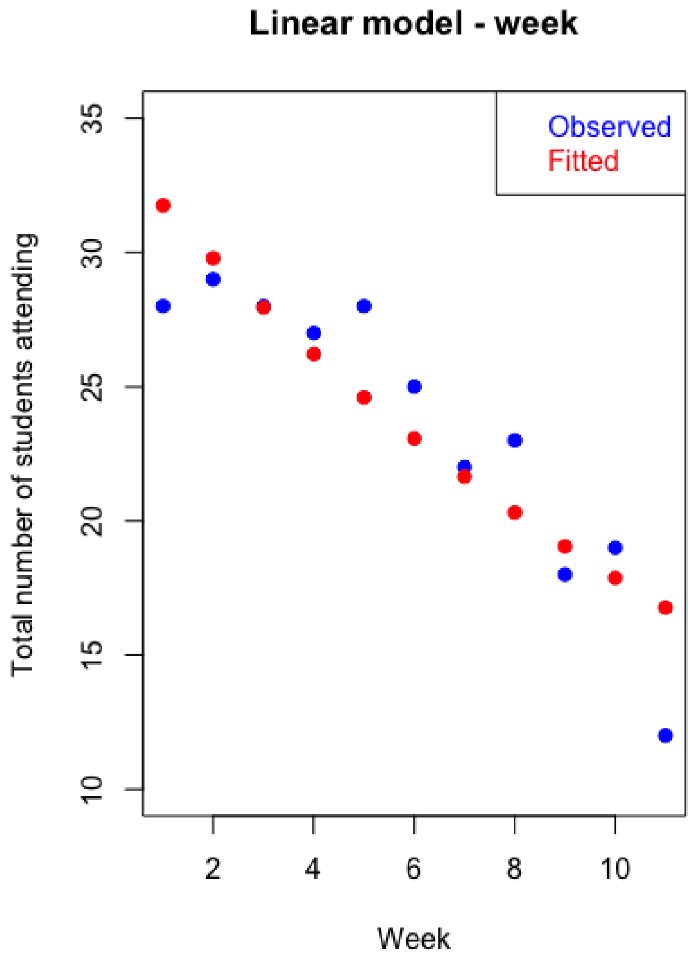

FIGURE 4. Observed values plotted against fitted values for the first GLM

The log-likelihood of the null model is $1 - \frac{1}{2}70.248 = -34.124$ and the log-likelihood of the linear model is $3 - \frac{1}{2}63.812 = -28.906$. The pseudo-$R^2$ of the final model was calculated to be

$$(-34.124 - (-28.906))/(-34.124) = 0.153$$

to 3.*d.p.* This means that there is a 15.3% increase in the log-likelihood for the linear model compared to the null model.

5. PREDICTIONS

If the value of the Index for tutorial of Week 12 was 1.5, we can calculate a prediction of the number of students attending the tutorial in Week 12 using the model

```
poisson3 <- glm(total ~ index + week, family=poisson(link="log"), data=df)
```

The predicted value is 15.6, and so we can estimate that at least 15 people would attend the week 12 tutorial and so we can make a prediction of 16 people to attend week 12.

If the actual number of students attending the tutorial of Week 12 was 8, then we can calculate the Pearson residual for that week using

$$e_{P,i} = \frac{y_i - \hat{\mu_i}}{\sqrt{\hat{\mu_i}}}$$

$$\frac{8 - 15.6314}{\sqrt{15.6314}} = -1.93$$

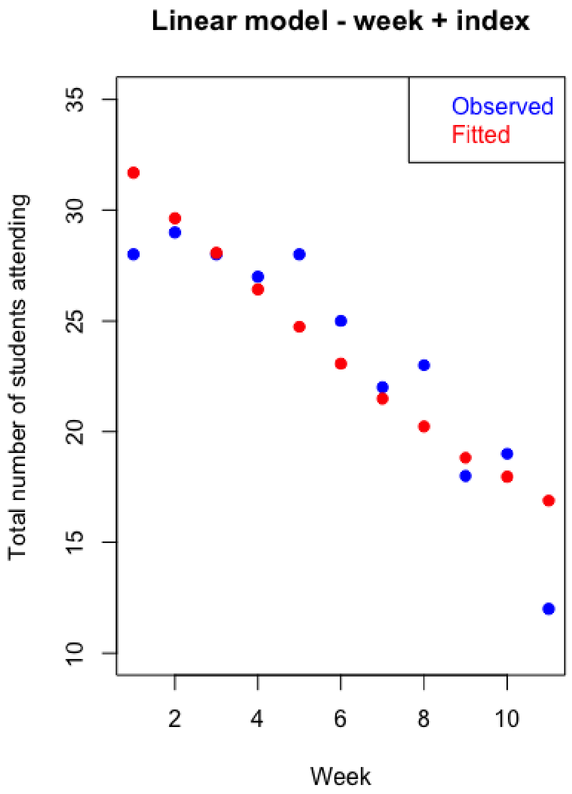

FIGURE 5. Observed values plotted against fitted values for the second GLM

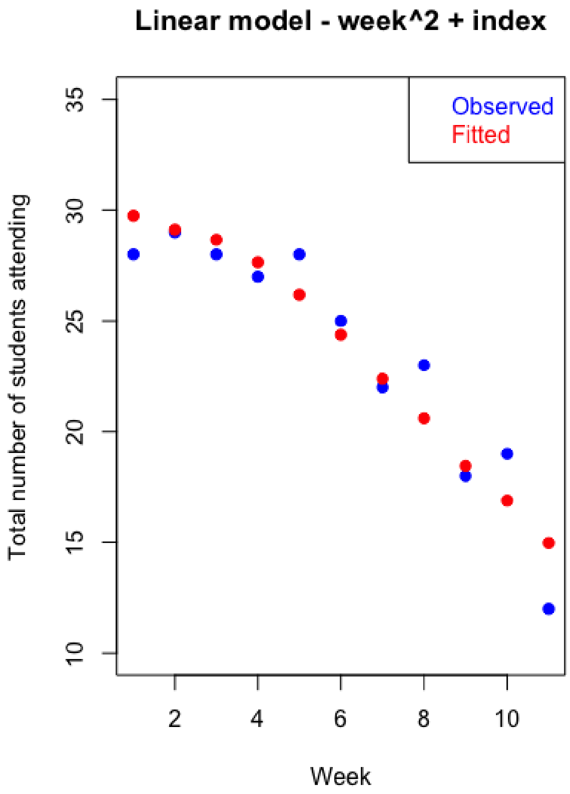

FIGURE 6. Observed values plotted against fitted values for the third GLM

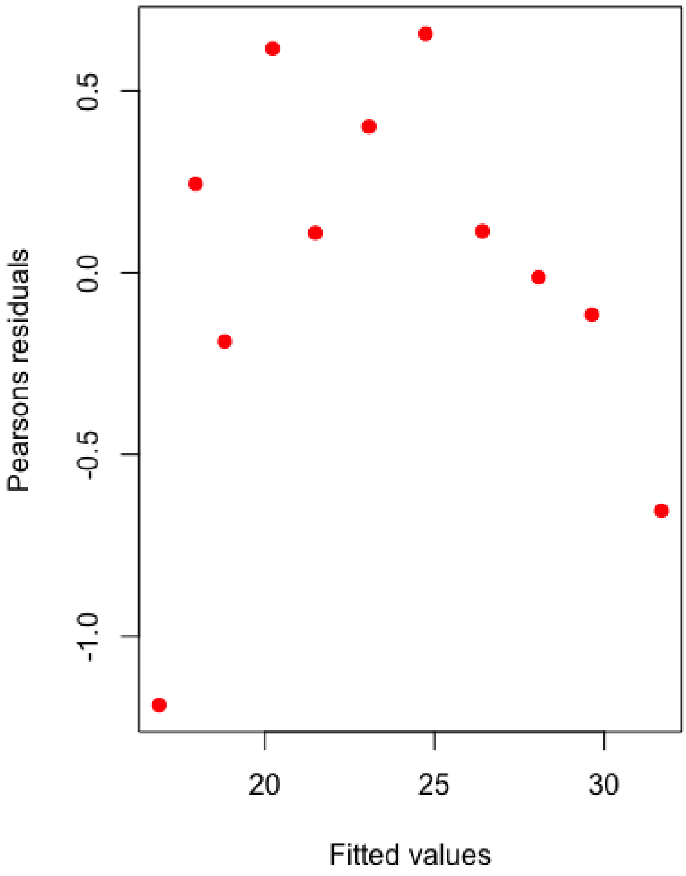

FIGURE 7. Pearson residuals against the fitted values

to 2.*d.p.* Figure 5 shows the fitted values plotted against the Pearson residuals using this model up to week 11.

6. Ways to improve model

One way to improve the model would be to collect more data in the 3 variables. This allows the "data to tell for itself," instead of relying on assumptions and weak correlations. Another way would be to add additional variables, ones in which have a stronger relationship with the response variable. This could be done from creating new variables from the original variables.

A way in which I could improve the model would be to explore transformations on the variables. This includes logarithms and quadratic terms. As seen in the analysis, the deviance of the model significantly decreases when quadratic terms are added to the model. Another transformation is standardizing based on the scale or potential range of the data (so that coefficients can be more directly interpreted and scaled).