Coursework - Applied Statistics

CID: 02091191

December 10, 2021

Abstract

In this document we are presenting the work conducted during the third Applied Statistics Coursework of the year.

1 Question 1

(a) Fitting a Poisson GLM to the data

To recall, consider a data set $\mathcal{D} = \{Y_i, x_i\}_{i=1}^n$ where $n = 500$. Consider a Poisson regression GLM to these data using the canonical link function (denoted g).

- **Random component:** The components of $\boldsymbol{Y}$ are independent and follow a Poisson distribution, which is a member of exponential family with $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$.
- **Systematic component:** Using the predictors $(x_1 \ldots, x_p)$ form the linear predictor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ where X is the design matrix and $\boldsymbol{\beta}$ is the vector of unknown parameters.
- **Link function:** The link between the random and systematic components is $\eta_i = g(\mu_i)$ for $i = 1, \ldots, n$. One can easily show that in the case of a Poisson GLM, we have $g = \log$.

One can fit this GLM in `R` using:

```
fit <- glm(y~x, family=poisson(), data=dfrm)
```

Therefore, this model would estimate $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 0.8738 \\ -3.4810 \end{pmatrix}$ with $\eta = \hat{\beta}_1 + \hat{\beta}_2 x$. Note that the reported z-statistics for each estimate show enough evidence to reject the null hypothesis that $\beta_1 = 0$ or $\beta_2 = 0$, with almost null p-values. Hence, in this Poisson GLM for these data, β_1 and β_2 are statistically significant.

(b) & (c) Comparison between Poisson GLM and true model

Using the Poisson GLM model fitted in (a), we may plot the response values y_i against the inverse link function of the estimated linear predictors $g^{-1}(\hat{\eta}_i)$. Recall that that we had for the negative binomial regression fitted in the previous coursework $\beta = \begin{pmatrix} -0.5 \\ -2.5 \end{pmatrix}$ with mean

$$\mathbb{E}[Y] = \mu = \frac{r}{e^{-(\beta_1 + \beta_2 x)} - 1} \iff x = -\frac{\log\left(\frac{r}{\mu} + 1\right) + \beta_1}{\beta_2}$$

where $\beta_1 = -0.5$, $\beta_2 = -2.5$ and $r = 2$. Then, we only need to predict the response of the Poisson GLM using these values of x in order to get the fitted regression line.

```
beta1 <- -0.5; beta2 <- -2.5
mu.hat <- fit$fitted.values
plot(mu.hat, dfrm$y, pch=16, xlab=TeX(r'(Fitted values $\g^{-1}(\hat{\eta})$)'),
     ylab=TeX(r'(Response $y_{i}$)'))
mu <- seq(0, 4, by=0.001)
x <- -(log(2/mu + 1) + beta1)/beta2
fit.val <- predict.glm(fit, data.frame(x), type="response")
lines(x=fit.val, y=mu, col="red", lwd=2)
library(MASS)
fit.nb <- glm.nb(y~x, data=dfrm)
mu <- seq(0, 4, by=0.001)
x <- -(log(2/mu + 1) + beta1)/beta2
fit.vals.nb <- predict.glm(fit.nb, data.frame(x), type="response")
lines(x=fit.vals.nb, y=mu, col="blue", lty=2, lwd=2)
legend("topleft", legend=c("Poisson model", "True model"), col=c("red", "blue"),
       lty=c(1,2))
```

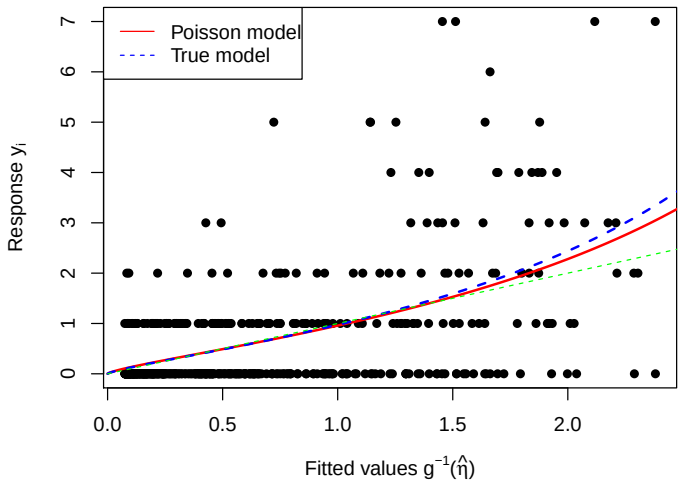

Figure 1.1: Plot of the response values y_i against the fitted values. Regression lines of the Poisson GLM and the negative binomial GLM model.

Figure 1.1 shows the scatter plot of the response against the fitted values. The red line corresponds to the regression line of the Poisson model. In a *true* model, we fit a GLM using the true distribution of the response values which is a negative binomial distribution with $r = 2$. The regression line is plotted in dashed blue on Figure 1.1. We may notice that for small response values, the Poisson model line is very close to the true model line, whereas the it is diverging for larger response values. As it seems quite unclear to what refers to "true model" (the GLM with the true distribution, which is a negative binomial distribution, or a model which would have exactly fitted values equal to response values), we also plotted on Figure 1.1 the line with equation $y = x$ in dashed green. This line would be the one of a "perfectly" fitted model, where all fitted and predicted values would be equal to the response values. The Poisson GLM regression line in red is very close to this line for low response values, and diverges for higher response values.

(d) 99% confidence interval for the mean response

Consider the Poisson regression model fitted in (a) and assume we are interested in calculating a 99% confidence interval for the mean response value for the covariate $x = 0.5$. According to the Lecture Notes [1], an approximate 99% confidence interval for the mean response of a predictor $\boldsymbol{x}_*$ can be written as

$$\left(g^{-1}\left(\hat{\eta}_* - q_{0.005}\sqrt{\boldsymbol{x}_*^T (X^T W X)^{-1} \boldsymbol{x}_*} \right),\ g^{-1}\left(\hat{\eta}_* + q_{0.005}\sqrt{\boldsymbol{x}_*^T (X^T W X)^{-1} \boldsymbol{x}_*} \right) \right)$$

where g^{-1} is the inverse link function, W is the weights matrix, X is the design matrix and $q_{0.005}$ is the 0.005% quantile of a normal distribution.

```
critical <- qnorm(0.995)
X.new <- c(1, 0.5)
W <- diag(fit$weights)
X <- model.matrix(fit)
beta <- fit$coefficients
cat(exp(X.new %*% beta -
          critical * sqrt(t(X.new) %*% solve(t(X) %*% W %*% X) %*% X.new)),
    exp(X.new %*% beta +
          critical * sqrt(t(X.new) %*% solve(t(X) %*% W %*% X) %*% X.new)))
```

Hence, a 99% confidence interval for the mean response for the covariate $x = 0.5$ is

$$(0.3410,\ 0.5182)$$

(e) Deviance residuals

Consider now the deviance residuals of this Poisson GLM. Recall the definition of deviance,

$$D = \sum_{i=1}^{n} d_i, \text{ where } d_i = 2\left\{ b'^{-1}(y_i) - \hat{\theta}_i \right\} y_i - b\left(b'^{-1}(y_i) \right) + b\left(\hat{\theta}_i \right)$$

for $i = 1, \ldots, n$. Then, deviance residuals are defined as

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

```
deviance.residuals <- residuals.glm(fit, type="deviance")
plot(dfrm$x, deviance.residuals, pch=16, xlab="Predictors",
     ylab="Deviance residuals")
sum(deviance.residuals <= 0) / length(deviance.residuals)
```

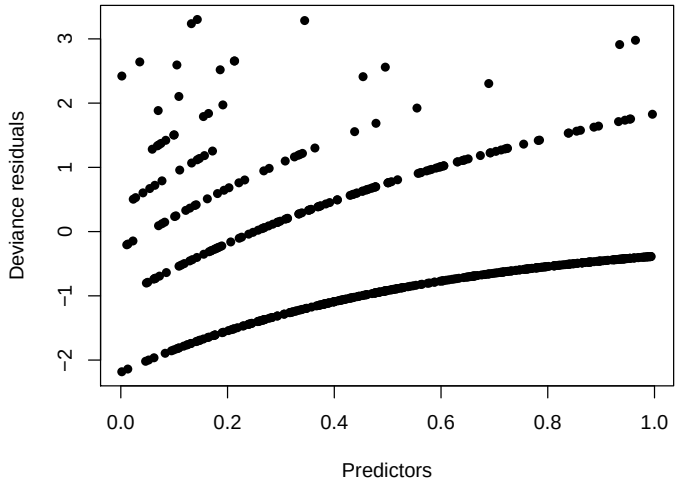

Figure 1.2: Plot of the deviance residuals $r_{D,i}$ against the predictor values x_i

Figure 1.2 reveals a trend between the deviance residuals and the predictors, which could be an issue regarding the Poisson GLM assumptions. Indeed, it seems that the deviance residuals increase as the predictor values increase. Moreover, we may notice that 69.6% of the deviance residuals are negative. Recall that $D^* \dot{\sim} \chi^2_{n-p-1}$ where $\dot{\sim}$ can be interpreted as "is approximately distributed as". Hence, this suggests that the deviance residuals $r_{D,i}$ are approximately normally distributed if the model holds according to the definition of D above. However, as we obtained almost 70% of deviance residuals being negative, this assumption does not hold in this case.

(f) Modelling the conditional distribution of x given y

Suppose we are now interested in modelling the conditional expectation of x given y. We may first fit a normal linear model for x with y treated as factor variable using the dummy coding contrast by default.

```
lm1 <- lm(x~as.factor(y)+0, data=dfrm)
summary(lm1)
```

	NLM Estimate	**NLM *p*-value**	**Confidence**	**NLMM Estimate**
$y = 0$	0.58111	$< 2 \times 10^{-16}$	***	0.57985
$y = 1$	0.44035	$< 2 \times 10^{-16}$	***	0.43779
$y = 2$	0.27549	1.29×10^{-9}	***	0.26973
$y = 3$	0.14995	0.0263	*	0.14297
$y = 4$	0.10678	0.1806		0.09995
$y = 5$	0.18924	0.0663	.	0.16990
$y = 6$	0.10505	0.6768		0.06241
$y = 7$	0.07825	0.5346		0.06683

Table 1: Estimates of unknown parameters in both normal linear model (NLM) and normal linear mixed model (NLMM). For NLM, p-values and confidence as given in R are reported.

Note that we fitted a linear model without any intercept term, so that we may interpret directly each unknown parameter estimate without doing any linear transformation. The estimates co-

efficients for low values of y in Table 1 are greater than for higher values of y (NLM estimate column). The summary given above (NLM p-value and Confidence columns) shows that some covariates may not be statistically significant to the linear model. Therefore, we would note that the means of x for observations where $y = 0$ or $y = 1$ or $y = 2$ are significantly not equal to zero; the mean of x for observations where $y = 3$ also seems not equal to zero but with less confidence; and all the other observations with different values of y may have a mean of x equal to zero (even if for $y = 5$, it is more unclear).

We may also consider the y factor as random effects and no longer as simple fixed effects, using the assumptions of normal linear mixed models. Then, we would fit a normal linear mixed model for x with y as a factor and a random effect term, without any fixed effect term.

```
library("lme4")
y.factor <- as.factor(dfrm$y)
x <- dfrm$x
nlm <- lmer(x~(1|y.factor)+0)
ranef(nlm)
```

The estimates are reported in Table 1 in the NLMM Estimate column. Notice that the estimates for each factor y are slightly different from the ones of the normal linear model as they are smaller than the linear model estimates. In fact, even if these two models seem similar, they are slightly different. The equation of the linear model can be written as

$$\boldsymbol{X} = Y\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$$

where Y denotes the design matrix containing the dummy coding for the "group" (y value) of each observation. The mixed model equation would be written as

$$\boldsymbol{X} = Z\boldsymbol{\nu} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n),\ \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \sigma_\nu^2 I_m)$$

where Z is the model matrix for the random effects. Hence, in the mixed model, the y factor is seen as a random variable and is not deterministic. The mixed model will therefore be more consistent as Y is a random variable (negative binomial), taking an infinite set of possible values (infinite possible levels). Hence, as Y has to be seen more as a random factor than a fixed factor, one would trust more the results of the mixed model than the linear model as it is more robust.

2 Question 2

(a) Scatter plot

Consider now a data set $\mathcal{D} = \{x_i, y_i, z_i\}_{i=1}^n$ where $n = 480$, y_i is a response variable, associated with an integer-valued predictor x_i and a category label A-H z_i variable.

```
plot(x=xyz$x, y=xyz$y, col=xyz$z, pch=as.numeric(xyz$z)+14, xlim=c(0, 110),
     xlab="x", ylab="y")
legend("bottomright", legend=levels(xyz$z), col=sort(unique(xyz$z)),
       pch=c(1,2,3,4,5,6,7,8)+14)
lm1 <- lm(y~x, data=xyz)
lines(x=c(0, 100), y=cbind(c(1,1), c(0, 100)) %*% lm1$coefficients, lwd=3)
```

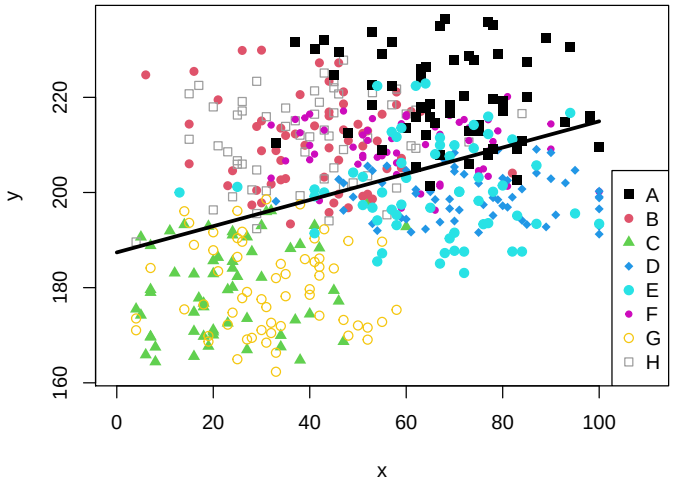

Figure 2.1: Scatter plot of y against x using different color and mark for each group z. Fitted regression line of y against x in bold black.

The black line on Figure 2.1 represents the regression line of a linear model for y against x with a global intercept term. Some aspects of this plot would make us think that a simple linear model is not appropriate to the data:

- Even if we can observe a global trend for y against x on this plot, a simple linear model seems inappropriate or too simple because of high residuals. We may notice on Figure 2.1 that most of the data points are relatively far from the black line, which would induce high RSS. Moreover, note that if we consider each group individually, this linear model would not fit at all to the data.
- For some groups like C and G, the residuals are (almost) only negative, which means that this linear model is very often over-estimating the response y. Hence, the fitted values for these groups are biased.
- A simple linear model like this would not provide any estimate of the between-group variability. Thus, we need to consider a more complex model for these data than a simple linear regression.

(b) Box-plots

```
boxplot(y~z, data=xyz, col=c("grey15", "#DF536B", "#61D04F", "#2297E6",
                             "#28E2E5", "#CD0BBC", "#F5C710", "gray62"))
t(aggregate(xyz$y, list(xyz$z), mean))
```

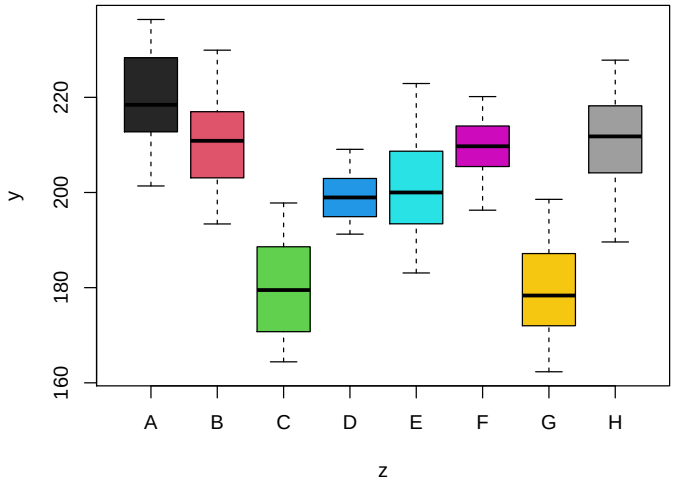

Figure 2.2: Box-plots of the response y for each group z (A-H)

A	B	C	D	E	F	G	H
219.7	210.7	179.9	199.3	200.7	209.3	179.8	211.1

Table 2: Sample mean of the response y for each group z (A-H)

Figure 2.2 shows that some groups do not have the same y distribution at all compared to other groups. For example, C and G groups have lower y values than the other groups. This is confirmed by Table 2 which reveals that groups C and G have almost the same mean of y, and same for groups B, F and H. Hence, it would be wise to use a regression model for y against x which would take into account the groups z.

(c) Normal linear mixed model

Assume a normal linear mixed model for y assuming a fixed effect for x with an intercept and a random effect for the z categories A-H.

```
lmm <- lmer(y~x+(1|z), data=xyz, REML=TRUE)
summary(lmm)
```

The restricted maximum likelihood estimates for the error variance $\hat{\sigma}^2_\varepsilon$ and for the variance of the random effects $\hat{\sigma^2_\nu}$ are

$$\hat{\sigma}^2_\varepsilon = 74.73 \text{ and } \hat{\sigma}^2_\nu = 213.03$$

(d) Removing some groups and fitting a normal linear mixed model

We now remove observations from categories A, C, D, E and G from the data and fit the same normal linear mixed model as previously on this new data set. However, we will not fit this model using restricted maximum likelihood but unrestricted maximum likelihood instead.

```
xyz.new <- xyz[((xyz$z == "B") | (xyz$z == "F") | (xyz$z == "H")),]
lmm.new <- lmer(y~1+x+(1|z), data=xyz.new, REML=FALSE)
logLik(lmm.new)
```

The reported unrestricted log-likelihood is -634.7416 (with 4 degrees of freedom). Note that the only remaining groups are B, F and H which, according to the box-plots in Figure 2.2, are similarly distributed. Hence, when fitting the normal linear mixed model in `R`, we get a warning saying the Z matrix is singular.

(e) Fitting a simple normal linear model

Using the same data set as in (d), we may also fit a simple normal linear model for y against x.

```
lm2 <- lm(y~x+1, data=xyz.new)
logLik(lm2)
d <- as.numeric(2*(logLik(lmm.new)-logLik(lm2)))
```

The reported log-likelihood for this linear model is -634.7416, which is exactly the same (with 3 degrees of freedom). The reported deviance parameter is 2.2737×10^{-13}, which is almost zero. Therefore, these two models are very similar. Thus, we may wonder if we should include or not the category variables in our model on this new data set.

(f) Significance of category variables

We now consider two models: one null model which is the linear model for y against x which does not take into account the categories z and one normal linear mixed model for y against x as fixed effects and z as random effects. We may then test H_0: do not include the random effects (null model) against H_1: do include the random effects (mixed model). Using parametric bootstrap, under the null, $y \sim \mathcal{N}(\mu, \sigma^2)$ so we can use this distribution to generate data under this model, using estimates for μ and σ^2.

```
ds <- numeric(1000)
for (i in 1:1000) {
  y <- unlist(simulate(lm2))
  nullmod <- lm(y ~ 1+xyz.new$x)
  altmod <- lmer(y ~ 1 + xyz.new$x + (1 | xyz.new$z), REML=FALSE)
  ds[i] <- as.numeric(2 * (logLik(altmod) - logLik(nullmod)))
}
phat <- mean(ds>d)
sqrt((phat*(1-phat))/1000)
```

Using parametric bootstrap after 1000 simulations, we obtain a p-value of 0.269 and a standard error on this estimate of 0.014. Hence, we can be fairly sure that the p-value is above 0.05, which would make us fail to reject the null hypothesis H_0. Therefore, we should not take into account the random effects in our model. This is consistent to what we could have expected since the three groups taken into account in the new data set have the same distribution. Then taking into account random effects for each group would make no sense as it would not bring any additional information to the model. This explains why we got warnings in `R` about singularity. Hence, in this case, fitting a simple normal linear model would be more appropriate.

References

[1] Dr Din-Houn Lau (Autumn 2020) *MATH70071 - Applied Statistics lecture notes*, Imperial College London MSc Statistics resources