Poisson Regression in R

Julian

2023-05-21

1. The Poisson distribution

We often use the Poisson distribution to model count data. If $Y \sim Poi(\lambda)$ with $\lambda > 0$, then the PMF is given by

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots$$

In addition, for Poisson distributed random variables, we have that $E[Y] = var(Y) = \lambda$. Eventually, we have that $\sum_{i=1}^{n} y_i \sim Poi(\sum_{i=1}^{n} \lambda_i)$. The code below shows how to draw the first plot on the next page.

```
set.seed(2023)
s1 <- data.frame('data' = rpois(n = 1000, lambda = 0.5))
s2 <- data.frame('data' = rpois(n = 1000, lambda = 2))
s3 <- data.frame('data' = rpois(n = 1000, lambda = 10))

p1 <- s1 %>% ggplot() +
  geom_bar(aes(x = data, y = stat(count / sum(count))), width = 0.5,
              fill = 'firebrick3') +
  labs(x = 'y', y = 'proportion', title = lambda~ '= 0.5') +
  theme_minimal()
  scale_color_gradient(low="firebrick1", high="firebrick4")
```

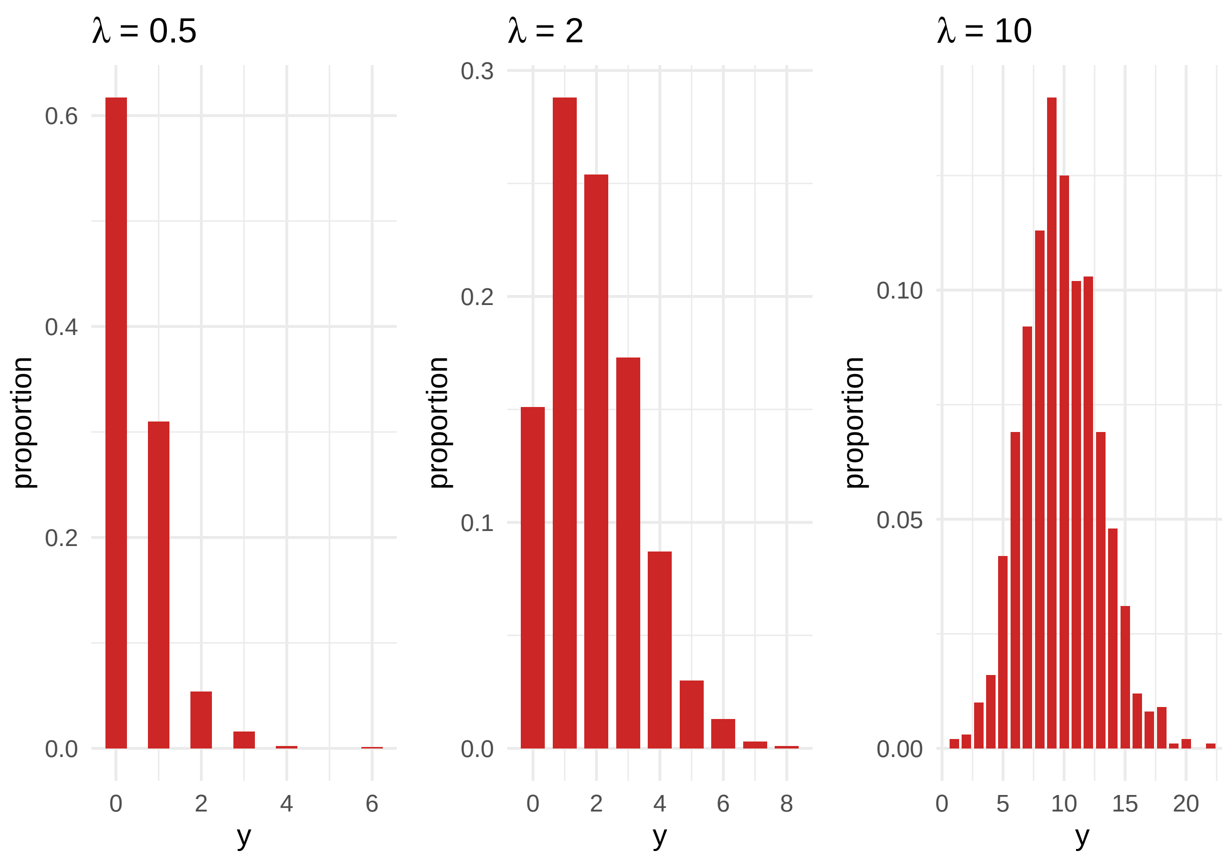
```
## <ScaleContinuous>
##  Range:
##  Limits:    0 --    1
```

```
p2 <- s2 %>% ggplot() +
  geom_bar(aes(x = data, y = stat(count / sum(count))), width = 0.75,
         fill = 'firebrick3') +
  labs(x = 'y', y = 'proportion', title = lambda~ '= 2') +
  theme_minimal()
p3 <- s3 %>% ggplot() +
  geom_bar(aes(x = data, y = stat(count / sum(count))), width = 0.75,
         fill = 'firebrick3') +
  labs(x = 'y', y = 'proportion', title = lambda~ '= 10') +
  theme_minimal()
```

Finally we can plot different artificially generated Poisson distributed data.

```r
grid.arrange(p1, p2, p3, nrow = 1)
```

```
## Warning: `stat(count / sum(count))` was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count / sum(count))` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



2. Poisson regression model

Consider n independent observations $y_1, ..., y_n$ for which we assume a Poisson distribution conditionally on a set of p categorical or numerical covariates x_j, for $j = 1, ..., p$. The model is given by

$$ln\Big(E[y_i \mid x_i]\Big) = ln(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta}$$

with $i = 1, ..., n$, with $\mathbf{x}_i^T = (1, x_{i1}, ..., x_{ip})^T$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$.

The natural link function is the log link. It ensures that $\lambda_i \geq 0$. It follows that

$$E\big[y_i \mid x_i\big] = \lambda_i = e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}} = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

The Poisson GLM is suitable for modeling count data as response variable Y when a set of assumptions are met.

3. Parameter estimation

The log-likelihood function is given by

$$l(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i^T \boldsymbol{\beta} - e^{\mathbf{x}_i^T \boldsymbol{\beta}} - ln(y_i!) \right)$$

Differentiating with respect to $\boldsymbol{\beta}$ and setting the new function equal to 0 yields the ***Maximum Likelihood equations***

$$\sum_{i=1}^{n} \left(y_i - e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right) x_{ij} = 0$$

with $j = 0, ..., p$ and x_{i0} = 1.

There is no closed-form solution for the Maximum Likelihood equations. We therefore have to resort to numerical optimization, for example the Iteratively Weighted Least Squares (IWLS) algorithm or the Newton-Raphson algorithm to obtain estimates of the regression coefficients.

4. Model assumptions

(i) **Count response**: The response variable is a count (non-negative integers), i.e. the number of times an event occurs in an homogeneous time interval or a given space (e.g. the number of goal scored during a football game). It is suitable for grouped or ungrouped data since the sum of Poisson distributed observations is also Poisson. When the reponse is a category (a ranking), we should consider a Multinomial GLM instead.

(ii) **Independent events**: The counts, i.e. the events, are assumed to be independent of each other. When this assumption does not hold, we should consider a Generalized Linear Mixed Model (GLMM) instead.

(iii) **Constant variance**: The factors affecting the mean are also affecting the variance. The variance is assumed to be equal to the mean. When this assumption does not hold, we should consider a Quasipoisson GLM for overdispersed (or underdispersed) data or a Negative Binomial GLM instead.

5. Parameter interpretation

(i) β_0 represents the change in the log of the mean when all covariates x_j are equal to 0. Thus e^{β_0} represents the change in the mean.

(ii) β_j, for $j > 0$ represents the change in the log of the mean when x_j increases by one unit and all other covariates are held constant. Thus e^{β_j} represents the change in the mean.

6. Practical example using the 'Affairs' dataset

We will fit a Poisson regression model to a subset of the 'Affairs' dataset (after W. H. Greene).

There are $n = 20$ observations and 8 variables in the reduced dataset. The variable 'affairs' is the number of extramarital affairs in the past year and is our response variable. We will include as covariates the variables 'gender', 'age', 'yearsmarried', 'children', 'religiousness', 'education' and 'rating' in our analysis. 'religiousness' ranges from 1 (anti) to 5 (very) and 'rating' is a self rating of the marriage, ranging from 1 (very unhappy) to 5 (very happy).

```r
data(Affairs, package = 'AER')
set.seed(2023)
data <- Affairs[sample(nrow(Affairs), size = 20, replace = FALSE),-c(8)]
head(data)
```

```
##      affairs gender age yearsmarried children religiousness education rating
## 174       12 female  42           15      yes             5         9      1
## 1895       0 female  32           15      yes             2        14      4
## 1540       0   male  32           10      yes             3        20      5
## 1226       0 female  32           15      yes             4        18      4
## 1445      12   male  37           15      yes             5        17      2
## 526       12 female  42           15      yes             4        12      1
```

```r
dim(data)
```

```
## [1] 20  8
```

```r
class(data)
```

```
## [1] "data.frame"
```

7. Fitted Poisson model

```r
# Poisson model
poisson.model <- glm(affairs ~ .,
                     family = 'poisson', data = data)
summary(poisson.model)
```

```
##
## Call:
## glm(formula = affairs ~ ., family = "poisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8004  -0.9069  -0.5699   0.0426   3.3917
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.55732 2433.69469   0.000 0.999817
## gendermale      2.90347    0.81913   3.545 0.000393 ***
## age            -0.17067    0.05807  -2.939 0.003293 **
## yearsmarried    0.11233    0.08247   1.362 0.173180
## childrenyes    14.46413 2433.69273   0.006 0.995258
## religiousness  -0.07241    0.17862  -0.405 0.685188
## education      -0.50119    0.12543  -3.996 6.45e-05 ***
## rating         -0.80337    0.17495  -4.592 4.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 150.341  on 19  degrees of freedom
## Residual deviance:  41.005  on 12  degrees of freedom
## AIC: 88.102
##
## Number of Fisher Scoring iterations: 15
```

8. Deviance and goodness-of-fit

The **deviance** of the model (also called G-statistic) is given by

$$D_{model} = 2\sum_{i=1}^{n}\left(y_i ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right)$$

where $\hat{\lambda}_i = e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}$ is the fitted value of λ_i.

The deviance can be used as a goodness-of-fit test. We test H_0: 'The model is appropriate' versus H_1: 'The model is not appropriate'. Under H_0, we have that

$$D_{model} \sim \chi^2_{1-\alpha,n-(p+1)}$$

where $p+1$ is the number of parameters of the model and $1-\alpha$ is a quantile of the χ^2 distribution.

```r
# p-value of Residual deviance goodness-of-fit test
1 - pchisq(deviance(poisson.model), df = poisson.model$df.residual)
```

```
## [1] 4.890236e-05
```

Our model does not fit the data very well.} Since our p-value is 0.085, H_0 is just not rejected.

9. Pearson goodness-of-fit

The ***Pearson goodness-of-fit statistic*** is given by

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

where $\hat{\lambda}_i = e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}$ is the fitted value of λ_i.

We test H_0: 'The model is appropriate' versus H_1: 'The model is not appropriate'. Under H_0, we have that

$$X^2 \sim \chi^2_{1-\alpha,n-(p+1)}$$

where $p+1$ is the number of parameters of the model and $1-\alpha$ is a quantile of the χ^2 distribution.

```r
# Pearson's goodness-of-fit
Pearson <- sum((data$affairs - poisson.model$fitted.values)^2
               / poisson.model$fitted.values)
1 - pchisq(Pearson, df = poisson.model$df.residual)
```

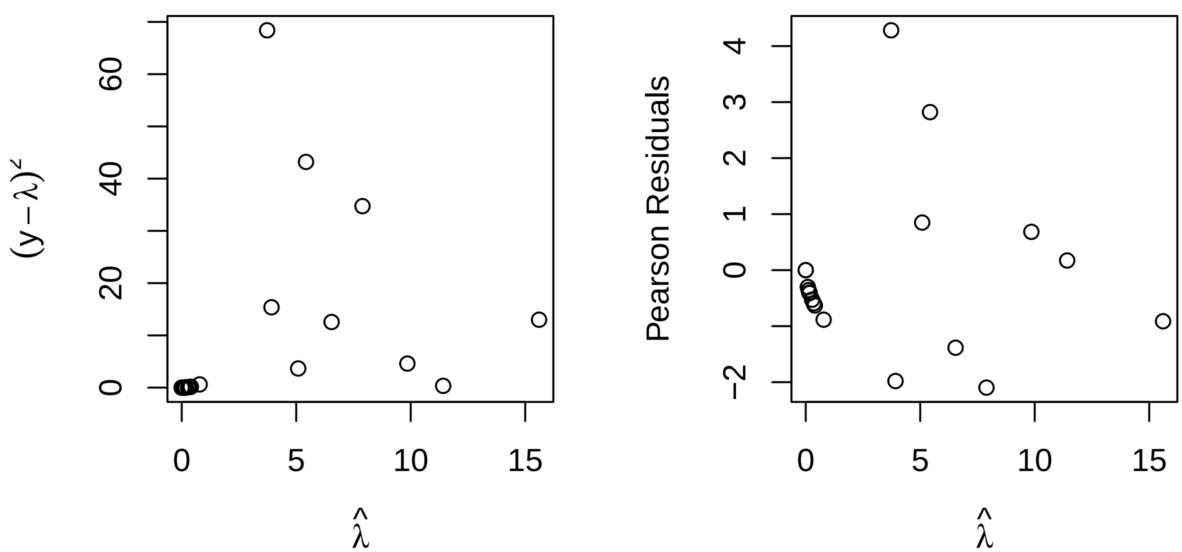
```
## [1] 4.702802e-05
```

The fit is not much better. Our p-value is 0.1054 and H_0 is not rejected.

10. Checking $E[Y] = var(Y)$ assumption

The variance of y_i is approximated by $(y_i - \hat{\lambda}_i)^2$. From the first graph we can see that the range of the variance differs from the range of the mean. Moreover, from the second graph, we see that the residuals show some kind of pattern. $E[Y] = var(Y)$ ***seems not to hold***. Let us examine the dispersion of the data and try a Quasipoisson in case of overdispersion.

```
lambdahat <-fitted(poisson.model)
par(mfrow=c(1,2), pty="s")
plot(lambdahat,(data$affairs-lambdahat)^2,
     xlab=expression(hat(lambda)), ylab=expression((y-hat(lambda))^2 ))
plot(lambdahat, resid(poisson.model,type="pearson"),
     xlab=expression(hat(lambda)), ylab="Pearson Residuals")
```



11. Assessing overdispersion

The variance of Y must be somewhat proportional to its mean. We can write

$$var(Y) = E[Y] = \phi\lambda$$

where ϕ is a scale parameter of dispersion and is equal to 1 if the equality $E[Y] = var(Y)$ holds. If $\phi > 1$, the data are **overdispersed** and if $\phi < 1$, the data are underdispersed. If a Poisson model is fitted under overdispersion of the response, then the standard errors of the estimated coefficients are underestimated. The scale parameter ϕ can be estimated as

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}}{n - (p+1)} = \frac{X^2}{n - (p+1)}$$

```
# Estimated dispersion parameter
Pearson / poisson.model$df.residual
```

```
## [1] 3.425568
```

The dispersion parameter is roughly equal to 1.53 for our data. Let us try a Quasipoisson regression model.

12. Fitted Quasipoisson model

The fitted Quasipoison model yields the following R output. However, the fit seems not to have improved based on the deviance goodness-of-fit test.

```
# Quasipoisson model
quasipoisson.model <- glm(affairs ~ .,
                          family = 'quasipoisson', data = data)
summary(quasipoisson.model)
```

```
##
## Call:
## glm(formula = affairs ~ ., family = "quasipoisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8004  -0.9069  -0.5699   0.0426   3.3917
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.55732 4504.35291   0.000   0.9999
## gendermale      2.90347    1.51608   1.915   0.0796 .
## age            -0.17067    0.10748  -1.588   0.1383
## yearsmarried    0.11233    0.15264   0.736   0.4759
## childrenyes    14.46413 4504.34928   0.003   0.9975
## religiousness  -0.07241    0.33059  -0.219   0.8303
## education      -0.50119    0.23216  -2.159   0.0518 .
## rating         -0.80337    0.32380  -2.481   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.425568)
##
##     Null deviance: 150.341  on 19  degrees of freedom
## Residual deviance:  41.005  on 12  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 15
```

```r
# p-value of Residual deviance goodness-of-fit test
1 - pchisq(deviance(quasipoisson.model), df = quasipoisson.model$df.residual)
```

```
## [1] 4.890236e-05
```

13. Variable selection using BIC

Some variables may not be relevant to the model or have low explanatory power. **Stepwise model selection** provides one possible solution to select our covariates based on Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) reduction (not available for Quasipoisson models).

```r
# variable selection using BIC
library(MASS)
```

```
##
## Attachement du package : 'MASS'

## L'objet suivant est masqué depuis 'package:dplyr':
##
##     select
```

```r
stepAIC(poisson.model, direction = 'both', k = log(dim(data)[1]))
```

```
## Start:  AIC=96.07
## affairs ~ gender + age + yearsmarried + children + religiousness +
##     education + rating
##
##                 Df Deviance     AIC
## - religiousness  1   41.169  93.236
## - children       1   41.231  93.298
## - yearsmarried   1   43.037  95.104
## <none>               41.005  96.068
## - age            1   51.372 103.439
## - gender         1   57.363 109.430
## - education      1   63.560 115.627
## - rating         1   77.775 129.841
##
## Step:  AIC=93.24
## affairs ~ gender + age + yearsmarried + children + education +
##     rating
##
##                 Df Deviance     AIC
## - children       1   41.407  90.478
## - yearsmarried   1   43.038  92.109
## <none>               41.169  93.236
## + religiousness  1   41.005  96.068
## - age            1   53.378 102.448
## - gender         1   57.369 106.440
## - education      1   63.894 112.965
## - rating         1   78.697 127.768
##
```

```
## Step:  AIC=90.48
## affairs ~ gender + age + yearsmarried + education + rating
##
##                  Df Deviance     AIC
## - yearsmarried   1   43.478  89.553
## <none>                41.407  90.478
## + children       1   41.169  93.236
## + religiousness  1   41.231  93.298
## - age            1   54.563 100.638
## - gender         1   58.957 105.032
## - education      1   65.960 112.036
## - rating         1   79.587 125.662
##
## Step:  AIC=89.55
## affairs ~ gender + age + education + rating
##
##                  Df Deviance     AIC
## <none>                43.478  89.553
## + yearsmarried   1   41.407  90.478
## + children       1   43.038  92.109
## + religiousness  1   43.478  92.549
## - gender         1   59.877 102.956
## - education      1   66.504 109.584
## - age            1   77.193 120.272
## - rating         1   88.952 132.032
```

```
##
## Call:  glm(formula = affairs ~ gender + age + education + rating, family = "poisson",
##     data = data)
##
## Coefficients:
## (Intercept)   gendermale          age    education       rating
##     12.2632       2.6951      -0.1084      -0.4541      -0.8337
##
## Degrees of Freedom: 19 Total (i.e. Null);  15 Residual
## Null Deviance:       150.3
## Residual Deviance: 43.48     AIC: 84.57
```

```
# Step:  AIC=61.42
# affairs ~ yearsmarried + children + religiousness + rating
#
#                 Df Deviance    AIC
# <none>               20.753 61.423
# + age            1   19.461 63.128
# - children       1   25.501 63.176
# + gender         1   19.879 63.546
# + education      1   20.750 64.417
# - yearsmarried   1   32.187 69.862
# - religiousness  1   32.965 70.640
# - rating         1   57.142 94.817
```

It appears that the variables 'yearsmariried', 'children', 'religiousness' and 'rating' are the most relevant to our analysis. The next step is to select the best Quasipoisson model between one including all covariates and one for which only those four covariates are incorporated in the model.

14. Model selection using Crossvalidation

We will select the best model in terms of predictions using leave-one-out Crossvalidation (LOOCV). The model with the lowest Root Mean Squared Error (RMSE) will be preferred.

```
# Leave-one-out crossvalidation (LOOCV)
pred.cv.mod_1 <- pred.cv.mod_2 <- numeric(dim(data)[1])
for(i in 1:dim(data)[1]) {
  mod_1 = glm(affairs ~ .,
              family = 'quasipoisson', data = data, subset = -i)
  mod_2 = glm(affairs ~ children + yearsmarried + religiousness + rating,
              family = 'quasipoisson', data = data, subset = -i)
  pred.cv.mod_1[i] = predict.glm(mod_1, data[i,], type = 'response' )
  pred.cv.mod_2[i] = predict.glm(mod_2, data[i,], type = 'response')
}

error.mod_1 = (1/dim(data)[1]) * sum((data$affairs - pred.cv.mod_1)^2)
error.mod_2 = (1/dim(data)[1]) * sum((data$affairs - pred.cv.mod_2)^2)

# Root Mean Squared Error (RMSE)
sqrt(c(error.mod_1, error.mod_2))
```

```
## [1] 14557.069929     5.789588
```

Clearly, the model with four covariates yields better predictions than the complete model and should be preferred. However, the RMSE remains relatively large indicating potential outliers in the dataset.

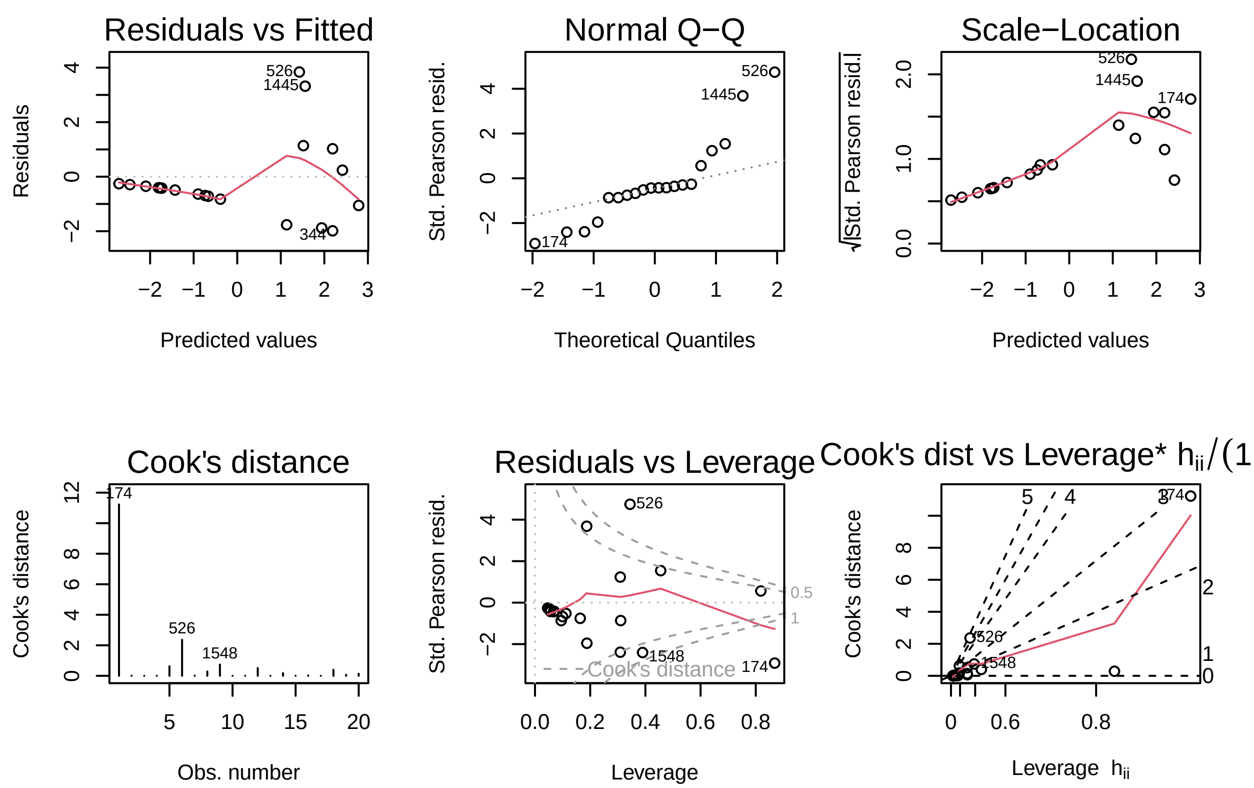
15. Diagnostic plots

```
# Diagnostic plots
quasipoisson.model.2 <- stepAIC(poisson.model, direction = 'both', k = log(dim(data)[1]))
```

```
## Start:  AIC=96.07
## affairs ~ gender + age + yearsmarried + children + religiousness +
##     education + rating
##
##                 Df Deviance     AIC
## - religiousness  1   41.169  93.236
## - children       1   41.231  93.298
## - yearsmarried   1   43.037  95.104
## <none>               41.005  96.068
## - age            1   51.372 103.439
## - gender         1   57.363 109.430
## - education      1   63.560 115.627
## - rating         1   77.775 129.841
##
## Step:  AIC=93.24
## affairs ~ gender + age + yearsmarried + children + education +
##     rating
##
##                 Df Deviance     AIC
```

```
## - children       1   41.407  90.478
## - yearsmarried   1   43.038  92.109
## <none>               41.169  93.236
## + religiousness  1   41.005  96.068
## - age            1   53.378 102.448
## - gender         1   57.369 106.440
## - education      1   63.894 112.965
## - rating         1   78.697 127.768
##
## Step:  AIC=90.48
## affairs ~ gender + age + yearsmarried + education + rating
##
##                 Df Deviance     AIC
## - yearsmarried   1   43.478  89.553
## <none>               41.407  90.478
## + children       1   41.169  93.236
## + religiousness  1   41.231  93.298
## - age            1   54.563 100.638
## - gender         1   58.957 105.032
## - education      1   65.960 112.036
## - rating         1   79.587 125.662
##
## Step:  AIC=89.55
## affairs ~ gender + age + education + rating
##
##                 Df Deviance     AIC
## <none>               43.478  89.553
## + yearsmarried   1   41.407  90.478
## + children       1   43.038  92.109
## + religiousness  1   43.478  92.549
## - gender         1   59.877 102.956
## - education      1   66.504 109.584
## - age            1   77.193 120.272
## - rating         1   88.952 132.032
```

```r
par(mfrow = c(2,3))
plot(quasipoisson.model.2, which = 1:6)
```



Based on the Cook's distance, the observation 1218 appears to be atypical and have a strong influence on the parameter estimates as well as on the predictions. This observation should be removed.

16 Our final model

```
# Final model
# 10. Remove outlier
round(cooks.distance(quasipoisson.model.2)) # observation 1218 is atypical
```

```
##  174 1895 1540 1226 1445  526  903 1138 1548  648  374  344 1671  929  227 1654
##   11    0    0    0    1    2    0    0    1    0    0    1    0    0    0    0
##  670  635 1341  516
##    0    0    0    0
```

```
data2 <- data[ - which.max(round(cooks.distance(quasipoisson.model.2))), ]

quasipoisson.model.3 = glm(affairs ~ children + yearsmarried + religiousness + rating,
              family = 'quasipoisson', data = data2, maxit = 100)
summary(quasipoisson.model.3)
```

```
##
## Call:
## glm(formula = affairs ~ children + yearsmarried + religiousness +
```

```
##     rating, family = "quasipoisson", data = data2, maxit = 100)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8931  -1.8296  -0.9807   0.3044   4.0127
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -11.76079 3344.36345  -0.004   0.9972
## childrenyes     16.45699 3344.36313   0.005   0.9961
## yearsmarried    -0.08343    0.07988  -1.044   0.3140
## religiousness   -0.14587    0.29007  -0.503   0.6229
## rating          -0.83127    0.34010  -2.444   0.0284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.058069)
##
##     Null deviance: 137.169  on 18  degrees of freedom
## Residual deviance:  68.519  on 14  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 14
```

```
# p-value of Residual deviance goodness-of-fit test
1 - pchisq(deviance(quasipoisson.model.3), df = quasipoisson.model.3$df.residual)
```

```
## [1] 3.574799e-09
```

```
# Pearson's goodness-of-fit
Pearson <- sum((data2$affairs - quasipoisson.model.3$fitted.values)^2
               / quasipoisson.model.3$fitted.values)
1 - pchisq(Pearson, df = quasipoisson.model.3$df.residual)
```

```
## [1] 1.374603e-09
```

Once the outlier has be removed, the fit is much better and the standard errors are much lower compared to the parameter estimates. This is our best model.

17. Conclusions

(i) The problems of overdispersion, covariate selection and influence of outliers have been addressed. Our final Quasipoisson model is a good fit for the data. About 86% of the deviance is explained by the model.

(ii) The level of religiousness and the number of years of marriage seem to be positively related to the average number of affairs, whereas having children and a happy self rated marriage seem to be negatively related to the average number of affairs. Caution however since the dataset only contains 19 observations.

(iii) If an individual has one child or more, the change in the mean response given all other covariates held constant is $e^{-2.75} \approx 0.064$, hence a decrease of 93.6% of the average number of affairs in the past year.

(iv) For one more year of marriage, the change in the mean response given all other covariates held constant is $e^{0.304} \approx 1.36$, hence an increase of 36% of the average number of affairs in the past year.

(v) When the self rating of the marriage changes from unhappy to happy, the change in the mean response given all other covariates held constant is $e^{-2.034} \approx 0.13$, hence a decrease of 87% of the average number of affairs in the past year.