

506CourseWork

Table of contents

For the love of god don't forget	1
Question 1	1
Question 1 a)	2
Question 1 b)	2
TODO GET BETTER WORDING FOR THE REPLACE PART	2
Question 1 c)	3
Question 1 d)	4
Question 1 e)	4
Question 1 f)	5
Question 1 g)	5
Question 2	5
Question 2 a)	5
Question 2 b)	6

For the love of god don't forget

Do not display too much raw R output (e.g. don't display the full output of 'summary(model)'), but edit this down to the essentials. Ensure to include justification for each step of your analyses, providing comments alongside your R code to explain what you are doing and add appropriate titles and labelled axes to your plots.

Question 1

We have the model:

$$Y_i \sim N\left(\frac{\theta_1 x_i}{\theta_2 + x_i}, \sigma^2\right)$$

Question 1 a)

Due to the visible non-linearity of the model, we would be required to significantly transform our data to get a linear model that would have an acceptable fit of the data. We can also see that the response data seems to be only positive while a normal distribution goes from $]-\infty, \infty[$. Such arbitrary transformation increases the complexity of the model, making it less interpretable and not respect the nature of the data.

Linear regression models are based on the assumption that the relationship between the independent and dependent variables is linear. If the relationship between the variables is non-linear, a linear regression model may not be appropriate to use. In such cases, transforming the data to make the relationship linear may not result in an accurate representation of the true relationship, and can lead to overfitting or underfitting. Additionally, transforming the data can result in a loss of interpretability of the results, as it can be difficult to understand the meaning of the transformed variables.

Another issue with using a linear regression model for non-linear data is that the residuals, which represent the difference between the observed and predicted values, may not be normally distributed, which is another assumption of linear regression models. This can lead to biased or incorrect results.

In conclusion, when the data is non-linear, a linear regression model may not be the best choice for modelling the relationship between the variables, and alternative methods need to be considered.

// make a graph to show the data is not linear

Question 1 b)

The Y_i are independent so the likelihood is a product of the individual pdfs.

TODO GET BETTER WORDING FOR THE REPLACE PART

Likelihood of a normal distribution where $L(\mu, \sigma^2, x_1 \dots x_n) =$

$$\begin{aligned} &= \prod_{i=1}^n f_X(y_i, \mu, \sigma^2) \\ &= \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} * \exp\left(-\frac{1}{2} * \frac{(y_i - \mu)^2}{\sigma^2}\right) = \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} * \exp\left(-\frac{1}{2\sigma^2} * \sum_{i=1}^n (y_i - \mu)^2\right) \end{aligned}$$

Replacing the μ with the respective θ s and n , we have the likelihood as:

$$\begin{aligned} &L(\beta_0, \beta_1, \sigma^2; x, y) = \\ &= \prod_{i=1}^n p(\beta_0, \beta_1, \sigma^2; x, y) = \end{aligned}$$

$$= (2\pi\sigma^2)^{-\frac{100}{2}} * \exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^{100} (yi - \frac{\theta 1 x_i}{\theta 2 + x_i})^2)$$

The log-likelihood of a normal distribution is:

$$\begin{aligned} l(\mu, \sigma^2, x_1 \dots x_n) &= \\ &= \ln(L(\mu, \sigma^2, x_1 \dots x_n)) = \\ &= \ln((2\pi\sigma^2)^{-\frac{n}{2}} * \exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^n (yi - \mu)^2)) = \\ &= \ln((2\pi\sigma^2)^{-\frac{n}{2}}) + \ln(\exp(-\frac{1}{2\sigma^2} * \sum_{i=1}^n (yi - \mu)^2)) = \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^n (yi - \mu)^2 = \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^n (yi - \mu)^2 \end{aligned}$$

Once again replacing the μ with the respective θ s and n, we have the log-likelihood as:

$$\begin{aligned} l(\beta_0, \beta_1, \sigma^2; x, y) &= \\ &= -\frac{100}{2} \ln(2\pi) - \frac{100}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{100} (yi - \frac{\theta 1 x_i}{\theta 2 + x_i})^2 = \\ &= -50 \ln(2\pi) - 50 \ln(\sigma^2) - \frac{1}{2\sigma^2} * \sum_{i=1}^{100} (xi - \frac{\theta 1 x_i}{\theta 2 + x_i})^2 \end{aligned}$$

Question 1 c)

```
n = nrow(nlmodel)

### Create a function to evaluate minus the log-likelihood
myLike = function(variables) {

  theta1 = variables[1] #theta1
  theta2 = variables[2] #theta2
  sigma = variables[3] #sigma

  mu = ((theta1 * nlmodel$x)/(theta2 + nlmodel$x))

  # Log-likelihood
  result = (-(n/2) * log(2 * pi)) - ((n/2) * log(sigma^2)) - (1/(2 * (sigma^2))) *
    (sum((nlmodel$y - mu)^2))

  # Returning negative log-likelihood
  return(-result)
}
```

Question 1 d)

From the graph data we can clearly see that the deviation is approximately 15 d see that when $x \rightarrow 1$ y is approximately 210 and as $x \rightarrow 0$ y is approximately to 50

```
# Estimating the MLE
out <- nlm(myLike,
  p = c(210,0.02,15), #the values seem to vary by around 15 units, ##the professor said th
  hessian = T,
  iterlim = 10000,
  steptol = 1e-10)

# Reporting estimates
variableEstimates= out$estimate
out$estimate
```

```
[1] 214.64918216  0.06353253  13.61564101
```

Question 1 e)

```
# Invert the negated Hessian to obtain the Observed Information Matrix
OIM <- solve(out$hessian)

# The diagonal entries are the variances of beta0 and beta1
# respectively so # obtain them
VarianceBeta <- diag(OIM)

# and then square root them to obtain standard errors
stand_error <- sqrt(VarianceBeta)

# reporting standard errors
stand_error
```

```
[1] 2.67474443 0.00514019 0.96302341
```

The formula to calculate a 99% confidence interval is: $\pm 2.576 * SE()$

```
# Estimating CIs
CIs <- cbind(variableEstimates - 2.576 * stand_error, variableEstimates +
  2.576 * stand_error)
```

```
# Reporting the CIs
CIs
```

```
          [,1]      [,2]
[1,] 207.75904050 221.53932382
[2,]  0.05029141  0.07677366
[3,] 11.13489269 16.09638932
```

Question 1 f)

Question 1 g)

Question 2

Model 1:

$$Y_i \sim \text{Pois}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

Model 2:

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\log(\mu_i) = \gamma_0 + \gamma_1 x_i$$

Question 2 a)

As we can from the graph and what we can determine from the nature of the data represented in such graph the recorded number of AIDS cases is a count variable and the counts are non-negative integers.

The first model, a Poisson distribution, would be a more appropriate choice. The Poisson distribution is a discrete distribution that models count data which respects the nature of the data

The second model, a Normal distribution, would not be the best fit since its range is from $]-\infty, \infty[$ and expects continuous values, not respecting the nature of the data.

The log-link function in both models ensures that the predicted values are always positive.
//TODO redo this pls

Question 2 b)

The Y_i are independent so the likelihood is a product of the individual pdfs.

$$L(\theta_1, \theta_2, \sigma^2; y, x)$$

```
# Fitting in R model <- glm(< response > < covariates >, data = <data>,
# family = gaussian(link='identity')) model <- glm(<response>
# <covariates>, data = <data>, family = poisson(link='log'))

# Fitting model 2 pois.model <- glm(ca ~ offset(logcells) + doseamt +
# doseate, data = dicentric, family = poisson(link='log'))

model2 = glm(cases ~ date, data = aids, family = poisson(link = "log"))

# Summarise the model
summary(model2)
```

Call:

```
glm(formula = cases ~ date, family = poisson(link = "log"), data = aids)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-7.768	-4.042	-0.335	3.048	7.281

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-16.875879	0.350353	-48.17	<2e-16 ***
date	0.247169	0.003856	64.10	<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: 5738.16 on 44 degrees of freedom
Residual deviance: 854.02 on 43 degrees of freedom
AIC: 1153.9

Number of Fisher Scoring iterations: 5

```
# Fitting model 1
model1 = glm(cases ~ date, data = aids, family = gaussian(link = "identity"))

# Summarise the model
summary(model1)
```

Call:

```
glm(formula = cases ~ date, family = gaussian(link = "identity"),
    data = aids)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-75.018	-21.703	-4.756	25.350	75.824

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3711.854	134.163	-27.67	<2e-16 ***
date	44.210	1.515	29.18	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

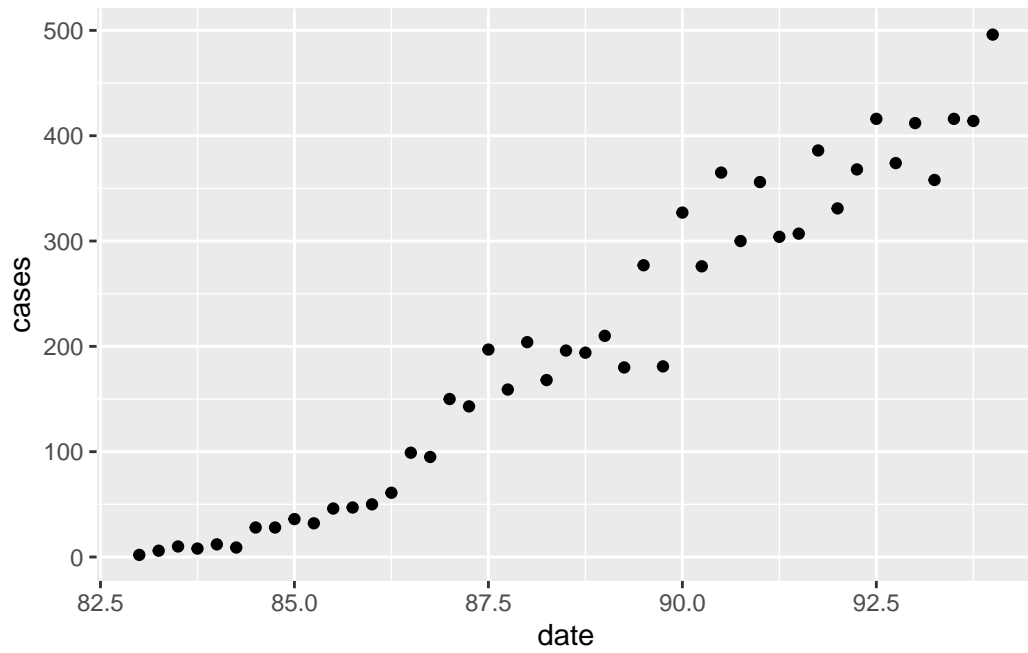
(Dispersion parameter for gaussian family taken to be 1088.729)

Null deviance: 974004 on 44 degrees of freedom
 Residual deviance: 46815 on 43 degrees of freedom
 AIC: 446.33

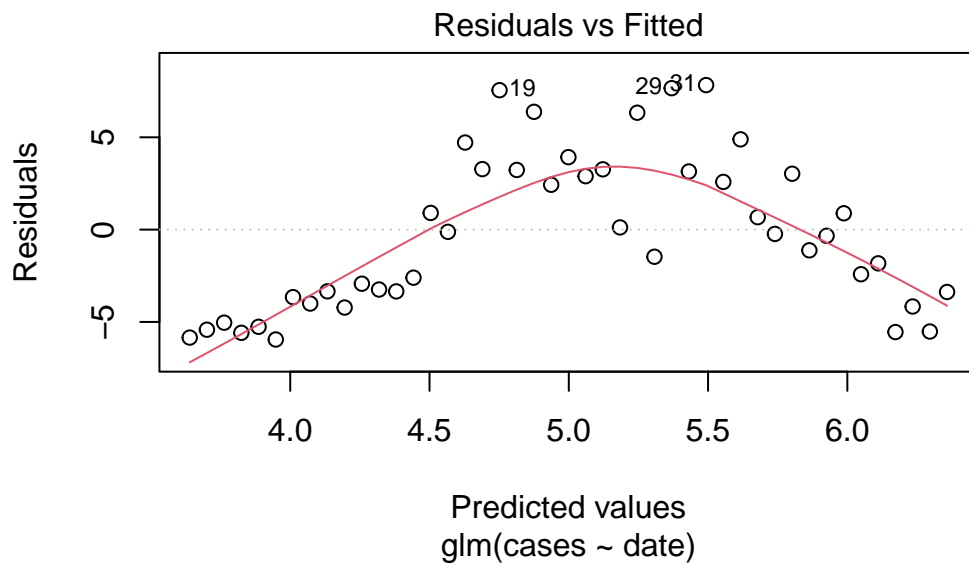
Number of Fisher Scoring iterations: 2

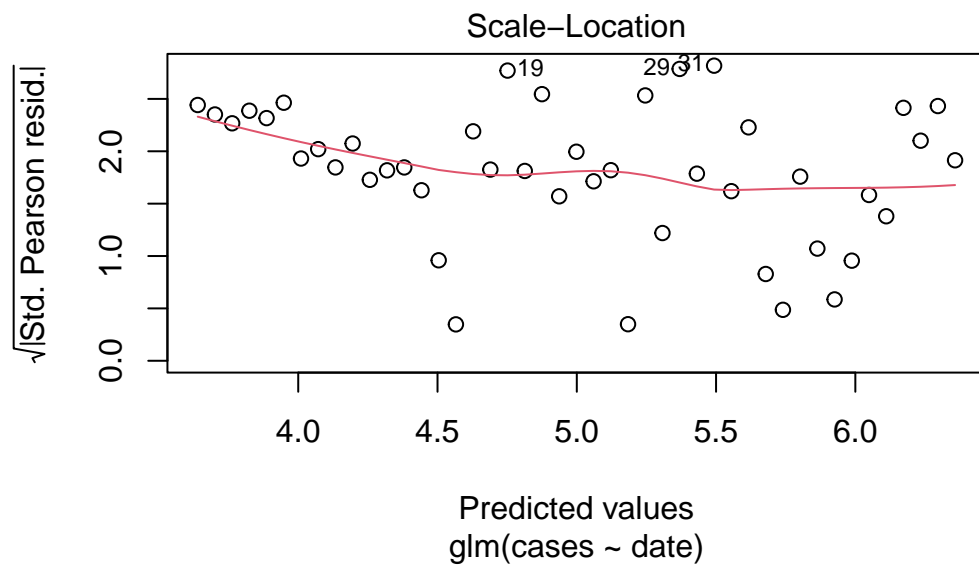
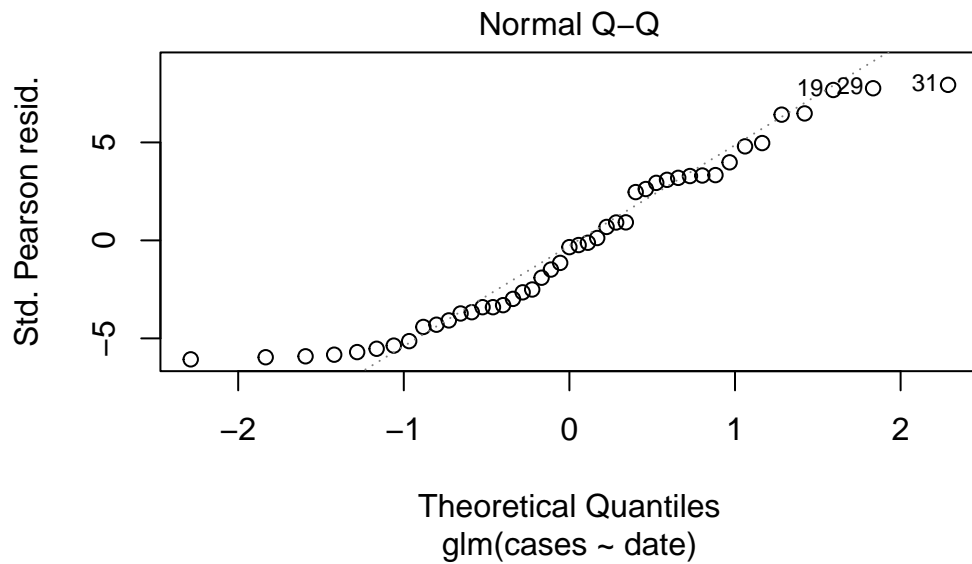
```
## plotting the models
```

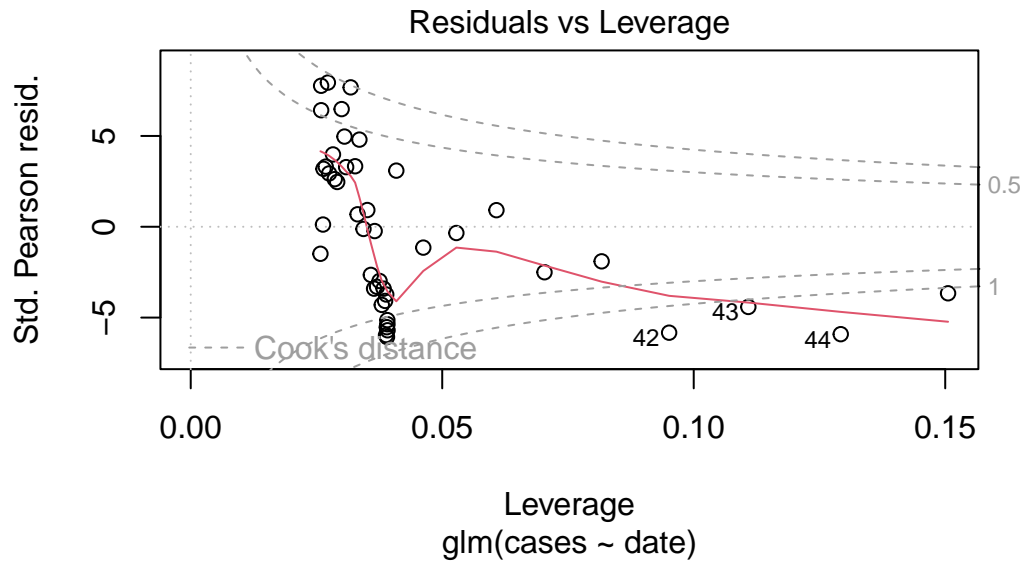
```
## this is the original data to then plot the models on top of
ggplot(aids, aes(x = date, y = cases)) + geom_point()
```



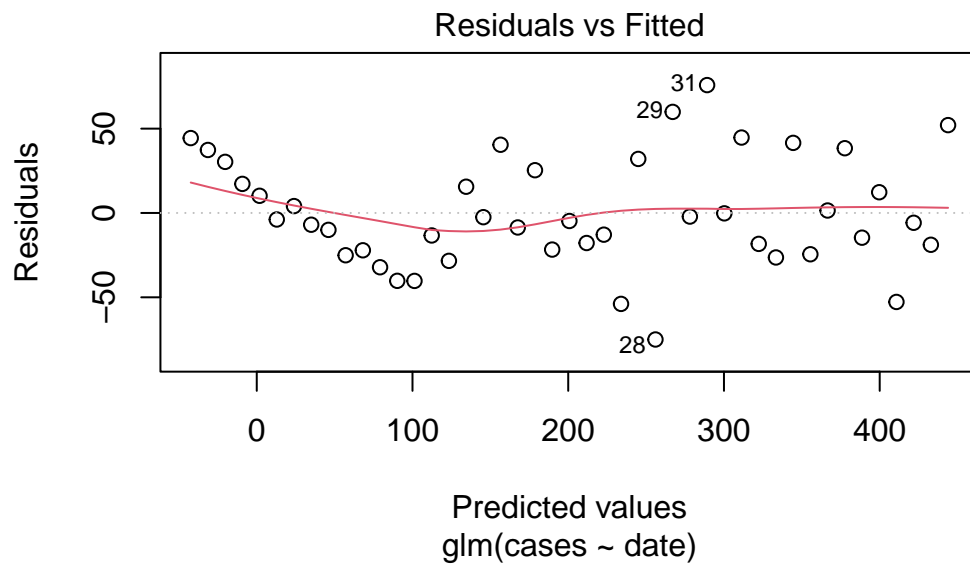
```
plot(model2)
```

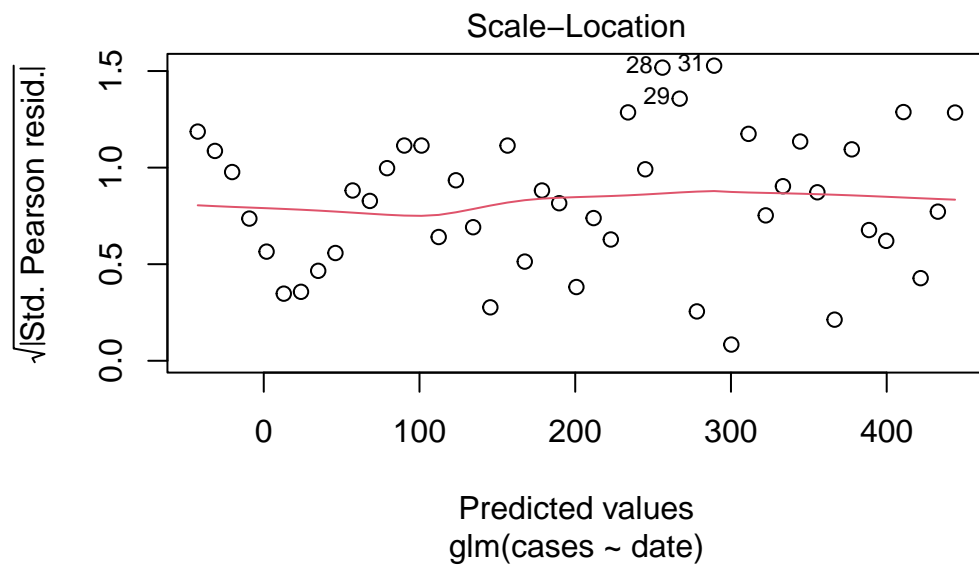
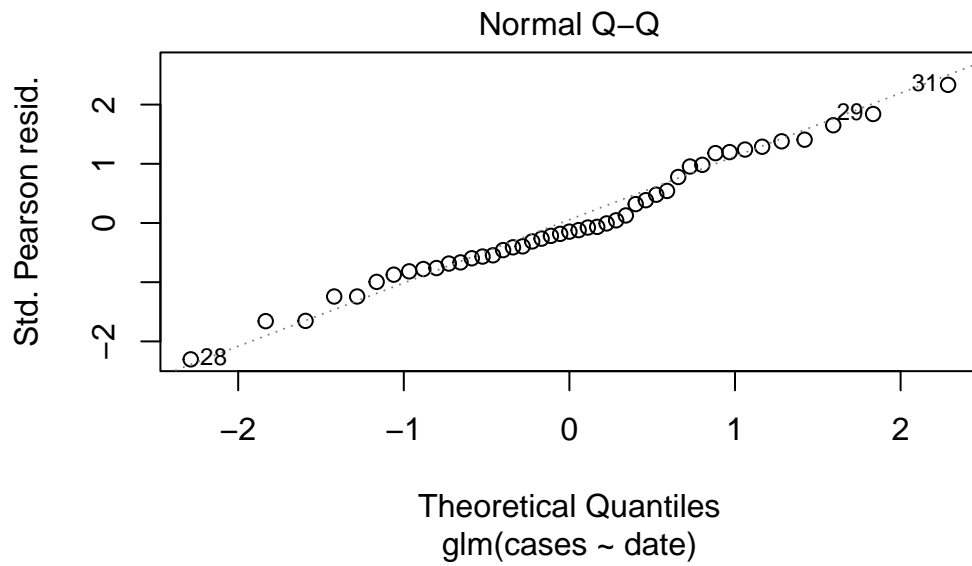


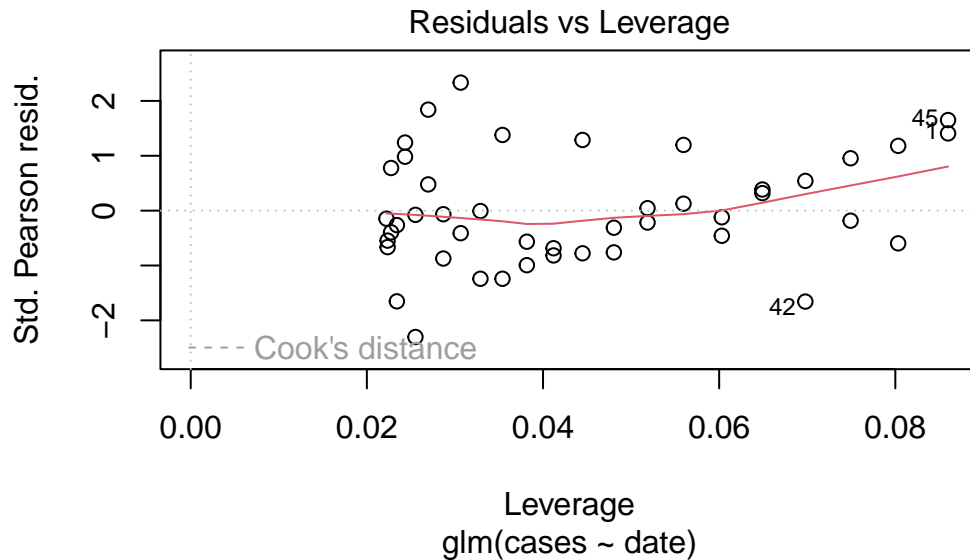




```
plot(model1)
```







```
confint(model1, level = 1 - 0.05)
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	-3974.80950	-3448.89836
date	41.24102	47.17953

```
confint(model2, level = 1 - 0.05)
```

Waiting for profiling to be done...

	2.5 %	97.5 %
(Intercept)	-17.564979	-16.191587
date	0.239636	0.254751

```
# As we are modelling an unbounded count we use Poisson distribution.
# The data increases exponentially so we use a log-link with a model
# linear in time.
```