6 Generalized Linear Models

This chapter provides an overview of generalized linear models (GLMs). We shall see that these models extend the linear modelling framework to variables that are not normally distributed. GLMs are most commonly used to model binary, positive continuous or count data.

In a linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

the response y_i, $i = 1, \ldots, n$ is modelled by a linear function of explanatory variables x_j, $j = 1, \ldots, p$ plus an error term.

We assume that the errors ϵ_i are independent and identically distributed such that

$$\mathrm{E}[\epsilon_i] = 0 \quad \text{and } \mathrm{Var}(\epsilon_i) = \sigma^2.$$

Typically we assume

$$\epsilon_i \sim N(0, \sigma^2)$$

as a basis for inference, e.g. t-tests on parameters.

Although a very useful framework, there are some situations where linear models are not appropriate:

- the range of y is restricted (e.g. binary, count, positive)
- the variance of y depends on the mean

Generalized linear models extend the linear model framework to address both of these issues.

A generalized linear model is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

and two functions:

- a link function that describes how the mean, $\mathrm{E}(y_i) = \mu_i$, depends on the linear predictor

  $$g(\mu_i) = \eta_i$$

- a variance function that describes how the variance, $\mathrm{Var}(y_i)$, depends on the mean

  $$\mathrm{Var}(y_i) = \phi V(\mu_i)$$

  where the dispersion parameter ϕ is a constant.

For the linear model with $y_i \sim N(0, \sigma^2)$ we have the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

the link function

$$g(\mu_i) = \mu_i$$

and the variance function

$$V(\mu_i) = 1.$$

6.1 Binomial Data

Suppose

$$y_i \sim \mathit{Binomial}\left(n_i, \frac{\mu_i}{n_i}\right)$$

and we wish to model the proportions y_i/n_i. Then

$$\mathrm{E}(y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(y_i) = \mu_i\left(1 - \frac{\mu_i}{n_i}\right).$$

The variance function is

$$V(\mu_i) = \mu_i\left(1 - \frac{\mu_i}{n_i}\right).$$

Our link function must map from $(0, 1) \to (-\infty, \infty)$. A common choice is

$$g(\mu_i) = \log\left(\frac{\mu_i}{n_i - \mu_i}\right).$$

6.2 Positive Continuous Data

A gamma GLM is of the form

$$y_i \sim \mathit{Gamma}(\mu_i, \nu),$$

with

$$\mathrm{E}(y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(y_i) = \frac{\mu_i^2}{\nu}.$$

The variance function is

$$V(\mu_i) = \mu_i^2.$$

Our link function must map from $(0, \infty) \to (-\infty, \infty)$. A common choice is

$$g(\mu_i) = \log(\mu_i).$$

6.3 Count Data

Suppose

$$y_i \sim \mathit{Poisson}(\mu_i)$$

then

$$\mathrm{E}(y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(y_i) = \mu_i.$$

So our variance function is

$$V(\mu_i) = \mu_i.$$

Our link function must map from $(0, \infty) \to (-\infty, \infty)$. A natural choice is

$$g(\mu_i) = \log(\mu_i).$$

6.4 Transformation versus GLM

In some situations a response variable can be transformed to improve linearity and homogeneity of the variance so that a general linear model can be applied. This approach has some drawbacks:

- the response variable has changed
- transformation must simultaneously improve linearity and homogeneity of variance.

For example, a common remedy for the variance increasing with the mean is to apply the log transform, e.g.

$$\begin{aligned} \log(y_i) &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \Rightarrow \mathrm{E}(\log y_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

This is a linear model for the mean of $\log y$ which may not always be appropriate. E.g. if y is income perhaps we are really interested in the mean income of population subgroups, in which case it would be better to model $\mathrm{E}(y)$ using a glm:

$$\log \mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$$

with a gamma distribution and a log link.

When we transform the data in a linear model, we are no longer claiming that y is normally distributed around a mean, given the values of x_1 - we are claiming that our new outcome variable, $\log(y_i)$, is normally distributed.

In fact, we often make this transformation specifically because the values of y do not appear to be normally distributed around their average.

In the case of the gamma model, however, the link function does not change the distribution of the actual observations in some way to make them something other than gamma distributed. Instead, the link function defines the relationship of x_1 directly to the mean of the gamma distributed y. The individual observations then vary around this expected value accordingly.

As you may know, if you have used this kind of data transformation in a linear model before, you cannot simply take the exponent of the mean of $\log(y)$ to get the mean of y.

You might be surprised to know, though, that you can do this with a link function. If you have specific values of your x variable, you can calculate the predicted average count, based on those x_1 values by inverting the natural log:

$$\mathrm{E}(y_i) = \exp(\beta_0 + \beta_1 x_i).$$

This ability to back-transform means (and regression coefficients) to a more intuitive scale is part of what makes generalized linear models so useful.

6.5 Exponential Family

Most of the commonly used statistical distributions, e.g. normal, binomial, gamma and Poisson, are members of the exponential family of distributions whose densities can be written in the form

$$f(y; \theta, \phi) = \exp\left\{[y\theta - b(\theta)]/\phi + c(y, \phi)\right\}$$

for functions $b(\cdot)$, $c(\cdot)$ and parameters θ (canonical parameter) and ϕ (dispersion parameter).

It can be shown that

$$\mathrm{E}(Y) = b'(\theta) \quad \text{and} \quad \mathrm{Var}(Y) = \phi b''(\theta).$$

For any particular problem, it is possible that there may be several plausible candidates for the link function. An interesting challenge is to propose the most suitable candidate for the problem at hand.

However, for any distribution, we can associate with it a default, or *canonical* link function. The canonical link is defined to be that function which maps the mean μ to the canonical parameter, θ. That is, that function g for which

$$g(\mu) \; = \; \theta \; = \; \eta.$$

Canonical links lead to desirable statistical properties of the glm hence tend to be used by default.

6.6 Estimation of the Model Parameters

An iterative algorithm can be used to estimate the parameters of an exponential family glm using maximum likelihood.

The log-likelihood for the sample $y_1, \ldots, y_n$ is

$$\ell = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi). \right\} = \sum_{i=1}^{n} \ell_i.$$

The maximum likelihood estimates are obtained by solving the score equations

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0$$

for parameters β_j. In general the score equations are non-linear in the β_j, and require numerical iterative techniques for their solution.

$\hat{\beta}_j$ have the usual properties of maximum likelihood estimators.

There are practical difficulties in estimating the dispersion ϕ by maximum likelihood. Therefore it is usually estimated by method of moments.

6.7 The `glm` Function

Generalized linear models can be fitted in `R` using the `glm` function, which is similar to the `lm` function for fitting linear models. The arguments to a glm call are as follows

```
glm(formula, family = gaussian, data,...)
```

6.7.1 Formula Argument

The formula is specified to glm as, e.g.

```
y ~ x1 + x2
```

where `x1`, `x2` are the names of

1. numeric vectors (continuous variables)
2. factors (categorical variables)

All specified variables must be in the workspace or in the data frame passed to the `data` argument.

6.7.2 Family Argument

The `family` argument takes (the name of) a family function which specifies the link function.

The exponential family functions available in R are

- `binomial(link = "logit")`
- `gaussian(link = "identity")`
- `Gamma(link = "inverse")`
- `inverse.gaussian(link = "1/mu2")`
- `poisson(link = "log")`

6.7.3 Extractor Functions

The `glm` function returns an object of class `c("glm", "lm")`. There are several `glm` or `lm` methods available for accessing/displaying components of the `glm` object, including:

- `residuals()`
- `fitted()`
- `predict()`
- `coef()`
- `deviance()`
- `formula()`
- `summary()`

6.8 Lab: Comparing `lm()` and `glm()`

We present a dataset on food expenditure for households that have three family members. We consider two variables, the logarithm of expenditure on food and the household income:

```
dat <- read.table("GHJ_food_income.txt", header = TRUE)
attach(dat)
plot(Food ~ Income, xlab = "Weekly Household Income ($)",
ylab = "Weekly Household Expenditure on Food (Log $)")
```

It would seem that a simple linear model would fit the data well.

We will first fit the model using `lm`, then compare to the results using `glm`.

```
foodLM <- lm(Food ~ Income)
summary(foodLM)
foodGLM <- glm(Food ~ Income)
summary(foodGLM)
```

The default family for `glm` is `"gaussian"` so the arguments of the call are unchanged. A five-number summary of the deviance residuals is given. Since the response is assumed to be normally distributed these are the same as the residuals returned from `lm`.

The estimated coefficients are unchanged. Partial t-tests test the significance of each coefficient in the presence of the others. The dispersion parameter for the gaussian family is equal to the residual variance.

Different model summaries are reported for GLMs. First we have the deviance of two models:

```
Null deviance: 4.4325 on 39 degrees of freedom
Residual deviance: 2.9073 on 38 degrees of freedom
```

The first refers to the null model in which all of the terms are excluded, except the intercept if present. The degrees of freedom for this model are the number of data points n minus 1 if an intercept is fitted. The second two refers to the fitted model, which has $n - p$ degrees of freedom, where p is the number of parameters, including any intercept.

The deviance of a model is defined as

$$D = 2\phi(l_{sat} - l_{mod})$$

where l_{mod} is the log-likelihood of the fitted model and l_{sat} is the log-likelihood of the saturated model. In the saturated model, the number of parameters is equal to the number of observations, so $\hat{y} = y$. For linear regression with normal data, the deviance is equal to the residual sum of squares.

Finally we have:

```
AIC: 14.649
Number of Fisher Scoring iterations: 2
```

The AIC is a measure of fit that penalizes for the number of parameters p

$$AIC = -2l_{mod} + 2p$$

Smaller values indicate better fit and thus the AIC can be used to compare models (not necessarily nested).

Several kinds of residuals can be defined for GLMs:

- response: $y_i - \hat{\mu}_i$.
- working: from the working response in the iterative algorithm
- Pearson
$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$
- deviance r_i^D

These definitions are all equivalent for normal models.

Deviance residuals are the default used in `R`, since they reflect the same criterion as used in the fitting. For example we can plot the deviance residuals against the fitted values (on the response scale) as follows:

```
plot(residuals(foodGLM) ~ fitted(foodGLM),
xlab = expression(hat(y)[i]),
ylab = expression(r[i]))
abline(0, 0, lty = 2)
```

The `plot` function gives the usual choice of residual plots, based on the deviance residuals. By default

- deviance residuals versus fitted values
- normal Q-Q plot of deviance residuals standardised to unit variance
- scale-location plot of standardised deviance residuals
- standardised deviance residuals versus leverage with Cook's distance contours

For the food expenditure data the residuals do not indicate any problems with the modelling assumptions:

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

6.9 Modelling Count Data

Examples of count data include

- number of household burglaries in a city in a given year

- number of customers served by a saleperson in a given month
- number of train accidents in a given year

In such situations, the counts can be assumed to follow a Poisson distribution, say

$$y_i \sim Poisson(\mu_i),$$

where

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

In many cases we are making comparisons across observation units $i = 1, \ldots, n$ with different levels of exposure to the event and hence the measure of interest is the rate of occurrence, e.g.

- number of household burglaries per 10,000 households in city i in a given year
- number of customers served per hour by salesperson i in a given month
- number of train accidents per billion train-kilometers in year i

Since the counts are Poisson distributed, we would like to use a GLM to model the expected rate, μ_i/t_i, where t_i is the exposure for unit i. Typically explanatory variables have a multiplicative effect rather than an additive effect on the expected rate, therefore a suitable model is

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

that is:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \log(t_i),$$

i.e. Poisson GLM with the canonical log link.

The standardizing term $\log(t_i)$ is an example of an offset: a term with a fixed coefficient of 1. Offsets are easily specified to `glm` using the offset argument in the formula, e.g. `offset(time)`.

If all the observations have the same exposure, the model does not need an offset term and we can model $\log(\mu_i)$ directly.

The `ships` data from the `MASS` package concern a type of damage caused by waves to the forward section of cargo-carrying vessels. The variables are number of damage incidents (`incidents`), aggregate months of service (`service`), period of operation (`period`): 1960-74, 75-79, year of construction (`year`): 1960-64, 65-69, 70-74, 75-79, and type (`type`): "A" to "E". Here it makes sense to model the expected number of incidents per aggregate months of service.

Let us consider a GLM Poisson model with log link including all the variables:

$$\begin{aligned}\log(\mu_i) = \beta_0 &+ \beta_1\texttt{typeB} + \beta_2\texttt{typeC} + \beta_3\texttt{typeD} + \beta_4\texttt{typeE} + \beta_5\texttt{year65} + \beta_6\texttt{year70} + \beta_7\texttt{year75} \\ &+ \beta_8\texttt{period75} + \log(\texttt{service}_i)\end{aligned} \tag{29}$$

We notice that the deviance is somewhat larger than the degrees of freedom which indicates lack of fit (deviance = 38.695 on 25 degrees of freedom). Note also that the residual deviance can be used as goodness-of-fit test only when ϕ is known. If ϕ is unknown (gamma, normal distributions for example), we cannot carry out a goodness-of-fit test, but can only estimate ϕ.

Lack of fit may be due to inadequate specification of the model, but another possibility when modelling discrete data is overdispersion. Under the Poisson model, we have a fixed mean-variance relationship

$$\mathrm{Var}(y_i) = \mathrm{E}(y_i) = \mu_i.$$

Overdispersion occurs when

$$\mathrm{Var}(y_i) > \mu_i.$$

This may occur due to correlated responses or variability between observational units.

We can adjust for over-dispersion by estimating a dispersion parameter ϕ

$$\mathrm{Var}(y_i) = \phi V(\mu_i).$$

This changes the assumed distribution of our response, to a distribution for which we do not have the full likelihood.

This approach used to estimate the parameters is known as quasi-likelihood estimation. Whilst estimating ϕ does not affect the parameter estimates, it will change inference based on the model. The theory for maximum likelihood also applies to quasi-likelihood.

In the `ships` data, it is likely that there is inter-ship variability in accident-proneness. Therefore we might expect some over-dispersion. We can switch to a quasi-likelihood estimation using the corresponding quasi-family.

The dispersion parameter is estimated as 1.69, much larger than the value of 1 assumed under the Poisson model.

Another possible remedy is to consider a more flexible distribution that does not impose equality of mean and variance. The most widely used distribution in this context is the negative binomial. It may be considered a mixture distribution arising from a Poisson distribution with random scale, the latter following a gamma distribution. The negative binomial distribution can be parametrised in terms of the mean μ and the shape parameter τ. The variance of the negative binomial distribution is given by

$$\mathrm{Var}(y) = \mu + \frac{\mu^2}{\tau}.$$

The Poisson distribution with parameter μ arises for $\tau \to \infty$.

Let's now see how to interpret the results form a Poisson GLM. We have the model expressed in equation (29) Consider ships of type C and E. We have

$$\log(\mu_i^E) - \log(\mu_i^C) = \log(\texttt{service}_i^E) - \log(\texttt{service}_i^C) + \beta_4 - \beta_2$$

and

$$\beta_4 - \beta_2 = \log\left(\frac{\mu_i^E}{\texttt{service}_i^E}\right) - \log\left(\frac{\mu_i^C}{\texttt{service}_i^C}\right) = \log\left(\frac{r_i^E}{r_i^C}\right).$$

So $\exp(\beta_4 - \beta_2)$ is the ratio of the rates (expected number of damages per month in service). Using the estimates of the coefficients reported in Table 12 we can conclude the following

- Types B and C have the lowest risk, E the highest (as compared to A). The rate for E is $\exp(0.33 - (-0.69)) = 2.75$ times that for C.
- The incident rate increased by a factor of $\exp(0.38) = 1.47$ after 1974

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	-6.40590	0.21744	-29.460	< 0.00000
`typeB`	-0.54334	0.17759	-3.060	0.00222
`typeC`	-0.68740	0.32904	-2.089	0.03670
`typeD`	-0.07596	0.29058	-0.261	0.79377
`typeE`	0.32558	0.23588	1.380	0.16750
`year65`	0.69714	0.14964	4.659	< 0.00000
`year70`	0.81843	0.16977	4.821	< 0.00000
`year75`	0.45343	0.23317	1.945	0.05182
`period75`	0.38447	0.11827	3.251	0.00115

Table 12: Coefficients of the Poison GLM model.

	ATP-preventable		
Period	Yes	No	**Total**
1967-1971	9	16	25
1972-1976	7	7	14
1977-1981	5	5	10
1982-1986	3	8	11
1987-1991	4	6	10
1992-1996	2	4	6
1997-2001	2	1	3
2002-2006	0	1	1
Total	32	48	80

Table 13: Occurrences of fatal train collisions, derailments and overruns in the UK from 1967 to 2006.

- The ships built between 1960 and 1964 seem to be the safest, with ships built between 1965 and 1974 having the highest risk

Let's consider another example. The data are on occurrences of fatal train collisions, derailments and overruns in the UK from 1967 to 2006 (see Table 11). The accidents are categorized into 5-year time periods, and into whether they would have been prevented if Automatic Train Protection had been installed.

We fit a generalized linear model for these data, in which the data are assumed to be Poisson, and a logarithmic link function is used. On the scale of the linear predictor the full model is fitted, in which the single available explanatory variable is time (`period.adj`), but a separate intercept and slope are allowed for ATP-preventable accidents and non-ATP-preventable accidents. Mathematically the model can be expressed as

$$\log(\mu_i) = \beta_0 + \beta_1 \texttt{period.adj} + \beta_2 \texttt{ATPpreventno} + \beta_3 \texttt{period.adj} \times \texttt{ATPprevent}. \tag{30}$$

Results are reported in Table 14. The interaction term is non-significant. So we now try the model with two intercepts and a single slope:

$$\log(\mu_i) = \beta_0 + \beta_1 \texttt{period.adj} + \beta_2 \texttt{ATPprevent}. \tag{31}$$

The results are reported in Table 15

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	2.241481	0.251535	8.911	< 0.000000
`period.adj`	-0.063042	0.017894	-3.523	0.000427
`ATPpreventno`	0.384092	0.325501	1.180	0.237999
`period.adj:ATPpreventno`	0.002115	0.023018	0.092	0.926776

Table 14: Coefficients of the Poison GLM model for accidents data.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	2.22866	0.21033	10.596	< 0.0000
`period.adj`	-0.06177	0.01126	-5.488	< 0.0000
`ATPpreventno`	0.40547	0.22822	1.777	0.0756

Table 15: Coefficients of the Poison GLM model for accidents data (without interaction term).

How do we interpret these results? Let's consider model (31). Then, if we consider `ATPpreventno` and `ATPpreventyes` we have

$$\log(\mu_i^{YES}) - \log(\mu_i^{NO}) = \beta_2$$

and

$$\beta_2 = \log\left(\frac{\mu_i^{YES}}{\mu_i^{NO}}\right)$$

So $\exp(\beta_2)$ is the ratio of the expected number of accidents.

Using this interpretation the results suggest that, if there is a significant difference between the level of ATP-preventable accidents and non-ATP-preventable accidents, then then average number of non-ATP-preventable accidents is $\exp(0.4055) = 1.5$ times that of ATP-preventable accidents. That is, we would predict that there are 50% more non-ATP-preventable accidents than there are ATP-preventable ones. Nevertheless the decline over time is strongly significant for both types of fatal accident, and we would predict that from one 5-year period to the next, numbers of accidents would be multiplied by $\exp(-5 \times 0.06177) = 0.734$, a reduction of 26.6%.

6.10 Modelling Positive Continuous Data

The gamma distribution can be used in a range of disciplines including financial services. Examples of events that may be modeled by gamma distribution include:

- the amount of rainfall accumulated in a reservoir
- the size of loan defaults and insurance claim cost
- the flow of items through manufacturing and distribution processes
- the load on web servers

A gamma GLM is of the form

$$y_i \sim Gamma(\mu_i, \nu)$$

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	8.2118447	0.0329095	249.528	< 0.0000
`op_time`	0.0383149	0.0006311	60.707	< 0.0000
`legrepYes`	0.4667863	0.0424613	10.993	< 0.0000

Table 16: Coefficients of the gamma GLM model for accidents data.

with

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

The canonical link for the gamma distribution is the inverse function. Since parameters from a model with inverse link are difficult to interpret, the log link is usually regarded as more useful.

For this model, we consider the personal injury insurance data. This data set contains information on 22,036 settled personal injury insurance claims. These claims arose from accidents occurring from July 1989 through to January 1999. Claims settled with zero payment are not included. The variables considered are: claim size (`total`), operational time (`op_time`) and legal representation (`legrep`).

The linear predictor for this model is

$$\log(\mu_i) = \beta_0 + \beta_1 \texttt{op\_time} + \beta_2 \texttt{legrep}.$$

The results of the fitted model are reported in Table 16.

Since, we use the same link function as in the Poisson GLM model, the interpretation of the coefficients is the same. So for example exp(0.467) = 1.59 means that claims with legal representation are 1.6 times more likely than claims with no legal representation. That is, we would predict that there are 60% more claims if there is legal representation than there are claims if there is no legal representation. Also, if we increase of one unit the operational time we would expect that the predicted claim size would be multiplied by exp(0.04) = 1.04, an increase of 4%.

6.11 Lab: Poisson GLM Model

In this section, we use the `RecreationDemand` from library `AER`. The data are cross-section data on the number of recreational boating trips to Lake Somerville, Texas, in 1980, based on a survey administered to 2,000 registered leisure boat owners in 23 counties in eastern Texas. The dependent variable is `trips`, and we want to regress it on all further variables: a (subjective) quality ranking of the facility (`quality`), a factor indicating whether the individual engaged in water-skiing at the lake (`ski`), household income (`income`), a factor indicating whether the individual paid a user's fee at the lake (`userfee`), and three cost variables (`costC`, `costS`, `costH`) representing opportunity costs. We begin with the standard model for count data, a Poisson GLM regression with canonical link (the log link).

```
library(AER)
data("RecreationDemand")
rd_pois <- glm(trips ~ ., data = RecreationDemand, family = poisson)
```

Let's look at the summary results

```
summary(rd_pois)
```

It would seem to indicate that almost all regressors are highly significant.

Let's also check the residual analysis:

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

The normal assumption and constant variance of the deviance residuals do not seem to hold and there might be problem of influential points (437, 554 and 659). These problems might indicate a problem of not good fit of the Poisson GLM which might due to overdispersion.

An alternative is to fit a quasi-Poisson model which accounts for overdispersion. We can fit this model in R

```
rd_qpois <- glm(trips ~ ., data = RecreationDemand, family = quasipoisson)
summary(rd_qpois)
```

The estimated coefficients are the same. Only the standard errors, and hence t-values and p-values are different but significance is similar. The estimated dispersion parameter is 6.3 suggesting overdispersion.

As for the interprettion, we can see that, for example, the ratio of the averages number of trips is 1.6 for one unit increase in quality ranking, holding constant the values of the other explanatory variables in the model. Also, the expected number of trips for the individual engaged in water-skiing is 1.5 times the expected number of trips for the individual not engaged in water-skiing, holding constant the values of the other explanatory variables in the model. This means that the predicted number of trips for individuals engaged in water-skiing is 50% more than that for individuals not engaged in water-skiing.

An alternative approach to tackle the issue of overdispersion is to use a more flexible distribution such as the negative binomial which allows the variance of y_i to be larger than its mean.

In R, tools for negative binomial regression are provided by the MASS package. Specifically, for estimating negative binomial GLMs, the function glm.nb(). Thus

```
library("MASS")
rd_nb <- glm.nb(trips ~ ., data = RecreationDemand)
summary(rd_nb)
```

The dispersion parameter of the fitted negative binomial distribution is 0.7293 (the closer to zero, the higher the overdispersion), suggesting a considerable amount of overdispersion.

As for the interpretation, we proceed, exactly in the same way as we did for the Poisson GLM.

6.12 Lab: Gamma GLM Model

Let's consider an example using an insurance dataset. This dataset (insurance) can be downloaded from Moodle.

```
insurance <- read.csv("insurance.csv")
```

This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. There are 4571 filed claims.

In the current section, we will focus on modeling only positive losses, we will not discuss the scenario where we want to know whether a policy has reported any claim or not (same as a policy has caused any loss or not). this scenario needs to consider other models (zero-inflated and hurdle models) which are beyond the scope of this course.

The variables we consider are: `veh_value` (vehicle value in $10,000s), `numclaims` (number of claims), `claimcst` (claim amount), `veh_body` (Vehicle body, coded as `COUPE`, `HBACK`, `HDTOP`, `MIBUS`, `other`, `PANVN`, `SEDAN`, `STNWG`, `TRUCK` and `UTE`), `veh_age` (age of vehicle, 1 = youngest,2,3,4), `gender` (a factor with levels `F` and `M`), `area` (a factor with levels `A`, `B`, `C`, `D`, `E` and `F`), and `agecat` (policyholder's age, 1=youngest, 2, 3, 4, 5, 6).

The dependent variable is `claimcst0`, whereas the remaing variables are independent variables. Number of claims (`numclaims`) will be used as an offset term.

Let's fit a GLM with Gaussian distribution and log link function and GLM with Gamma distribution and log link function, respectively:

```
model_gauss <- glm(claimcst0 ~ veh_value + veh_body + veh_age + gender + area + agecat,
                   data = insurance, offset = log(numclaims), family = gaussian(link="log"))

model_gamma <- glm(claimcst0 ~ veh_value + veh_body + veh_age + gender + area + agecat,
                   data = insurance, offset = log(numclaims), family=Gamma(link="log"))
```

After modeling, we can extract the results using a summary function and observe the goodness of fit for both the cases.

```
summary(model_gauss)
plot(model_gauss)

summary(model_gamma)
plot(model_gamma)
```

We can clearly see either of the models are not perfect, but still, gamma with log link is slightly better than the Gaussian. The pattern of spread in residuals against fit is better in gamma and as per Q-Q plot residuals are very close to the normal distribution. This is a GLM, not the linear regression where normality of residuals and homogeneity of variance is a strict condition, some deviation is expected but substantial deviation from normality indicates an issue with the assumed distribution. Finally, AIC is another indicator of model comparison, the value is lower in the case of gamma which suggests a better fit.

The estimated coefficients of the two models are comparable because both are on the same scale (log scale). For example, if we consider `gender`, the rate of claim amount for a male is $\exp(0.081) = 1.084$ times that for a female, under the Gaussian model with log link. Under the Gamma model with log link, the rate is $\exp(0.097) = 1.102$ times that for female.