Generalized Linear Models

Instructor: Dr. Sharandeep Singh Pandher*

1 Introduction

The term `generalized linear model (GLM)` was first introduced in a landmark paper by Nelder and Wedderburn (1972). Basically, Generalized linear Models (GLMs) extend linear models by allowing the response variable to follow distributions in the exponential family, which includes a wide range of commonly used distributions such as normal, binomial, and Poisson distributions etc. In other words, in a GLM, the response variable can be continuous, binary, discrete or count. The covariates or predictors still enter the model in a linear fashion, but the response and predictors are linked by a nonlinear link function. That is, if the response y_i is assumed to follow an exponential family distribution then the mean μ of the distribution is a function of $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ (often nonlinear).

In this chapter, we first describe GLMs in a general form. Then, we discuss the most popular GLMs, the logistic regression models, the Poisson regression models, Negative binomial regression models and Gamma regression models in greater details as their applications can be found in medicine, social sciences, biology and many other areas

2 Exponential family

A distribution in the exponential family has the following general form for its probability density function (pdf)

$$f\left(y|\boldsymbol{\theta},\phi\right)=\exp\left(\frac{y\boldsymbol{\theta}-b(\boldsymbol{\theta})}{a(\phi)}+c(y,\phi)\right),$$

where where $a(.)$, $b(.)$ and $c(.)$ are known functions, $\boldsymbol{\theta}$ is called the canonical parameter representing the location, and $\boldsymbol{\phi}$ is called the dispersion parameter representing the scale.The exponential family distributions have mean and variance as $E(y)=\mu=b^{'}(\boldsymbol{\theta})$ and $\text{Var}\,(y)=\sigma^2=a(\boldsymbol{\phi})b^{''}(\boldsymbol{\theta})$.

The following distributions are in the exponential family: normal distribution, binomial distribution, multinomial distribution, Poisson distribution, gamma distribution, Dirichlet distribution, Beta distribution, Chi-squared distribution, Geometric distribution, Negative Binomial distribution,and some other distributions. However, some common distributions are not in the exponential family, such as the t-distribution and uniform distribution.

*Address: Department of Mathematics and Statistics, University of Alberta, Edomoton, AB,T6G 2G1, Canada, e-mail: sharand1@ualberta.ca;

We focus on the most important ones: **binomial**, **Poisson**, **negative binomial**, and **Gamma** distributions in this Chapter. For instance,

1. if y follows Poisson distribution instead of exponential family distribution then probability distribution

$$P(y=k) = (k!)^{-1} \exp^{-\mu} \mu^k, \quad , k = 0, 1, 2, ...,$$

where $\mu > 0$, $\boldsymbol{\phi} = 1$, $a(\boldsymbol{\phi}) = 1$, $\boldsymbol{\theta} = \log \mu$, $b(\boldsymbol{\theta}) = \exp(\boldsymbol{\theta})$, $c(y; \boldsymbol{\phi}) = -\ln(y!)$ with $E(y) = Var(y) = \mu$

2. if y follows binomial distribution, then

$$P(y=k) = \binom{n}{k} \mu^k (1-\mu)^{n-k}, \quad k = 0, 1, 2, .., n,$$

where $0 < \mu < 1$, $\boldsymbol{\phi} = 1$, $\boldsymbol{\theta} = \ln(\mu/(1-\mu))$, $b(\boldsymbol{\theta}) = -n\ln(1-\mu)$, $c(y; \boldsymbol{\phi}) = \ln(\binom{n}{y})$, with $E(y) = n\mu$, $Var(y) = n\mu(1-\mu)$.

3. **Negative binomial distribution**: Given a series of independent trials, each with probability of success p, let n be the number of trials until the m'th success, then

$$P(z=n) = \binom{n-1}{m-1} p^m (1-p)^{n-m}, \quad n = m, m+1, m+2, ..$$

More convenient parameterization if we let $y = n - m$ and $p = (1+\alpha)^{-1}$

$$P(y=n) = \binom{n+m-1}{m-1} \frac{\alpha^n}{(1+\alpha)^{n+m}}, \quad n = 0, 1, 2, ..$$

$E(y) = \mu = m\alpha$ and $Var(y) = m\alpha + m\alpha^2 = \mu + \mu^2/m$

4. **Gamma distribution**: The density of the gamma distribution in the context of a GLM is

$$f(y) = \frac{1}{\Gamma(v)} (\frac{v}{\mu})^v y^{v-1} \exp^{-\frac{yv}{\mu}}, y > 0$$

where v is a shape parameters, λ is scale parameters of the distribution, $\mu = \frac{v}{\lambda}$, $E(y) = \mu$, $Var(y) = \frac{\mu^2}{v}$, and dispersion parameter is $\tau = v^{-1}$.

- **Note**: The binomial distribution reduces to the Bernoulli distribution when $n = 1$.
- **Note**: In the Poisson and binomial distribution, $\boldsymbol{\phi}$ is fixed at one. Hence, Poisson and binomial are one parameter families.
- **Note**: The Gamma distribution is two parameter families.
- **Note**: **binomial** distribution is for **binary** response, **Poisson** and **negative** binomail is for **count** response while **Gamma** distribution for **continuous** and **skewed** response.

3 The General Form of a GLM

Let us suppose we may express the effect of the predictors on the response through a linear predictor. let

$$\mu_i \quad = \quad E(y_i)$$

be the mean response. The linear combination

$$\eta_i = \boldsymbol{X_i}^\top \boldsymbol{\beta} = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

the linear predictor. The link function, g, describes how the mean response, $E(y_i) = \mu_i$, is linked to the covariates through the linear predictor: $\eta_i = g(\mu_i)$. The identity link function $g(x) = x$ for linear regressions, however, may not be appropppriate for other types of response. For example, if y_i is a binary variable, then the mean response $\mu_i = P(y_i = 1)$, which is a number between 0 and 1, while the value of the linear predictor η_i can take any value from $-\infty$ to ∞, so we cannot use the identity link to link the mean response to the linear predictor. In other words, for binary responses we should use other link functions to link the mean response to the linear predictor in regression modelling. For example, we may consider the following link function for binary response

$$g(x) = log\frac{x}{1-x}$$

Then, a regression model for the binary response y can be written as $g(\mu_i) = \eta_i$ i.e.

$$log(\frac{P(y_i = 1)}{1 - P(y_i = 1)}) = \boldsymbol{X_i}^\top \boldsymbol{\beta}, i = 1, 2, ..., n,$$

or

$$odds = \exp(\boldsymbol{X_i}^\top \boldsymbol{\beta}),$$

or

$$\mu_i = P(y_i = 1) = \frac{\exp(\boldsymbol{X_i}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{X_i}^\top \boldsymbol{\beta})}, i = 1, 2, ..., n,$$

which is the well known logistic regression model the most popular generalized linear model. As an example, suppose that $y_i = 1$ if a person gets a cancer and $y_i = 0$ otherwise. Let x_i be the smoking status. Then, the above logistic regression model can be used to study the relationship between the probability (or odds) of getting a cancer if a person smokes.
A generalized linear model (GLM) can be written as

$$g(E(y_i)) = \boldsymbol{X_i}^\top \boldsymbol{\beta}, i = 1, 2, ..., n,$$

where $g(.)$is a monotone and differentiable function, called the link function. Thus, a GLM has two components:

- the response y_i follows a distribution in the exponential family;
- the link function $g(.)$ describes how the mean response $E(y_i) = \mu_i$ is related to the linear predictor η_i, ie $g(\mu_i) = \eta_i$.

The following table provides a good summary of GLM

Family	Link	Variance Function
Normal	$\eta = \mu$	1
Poisson	$\eta = log\mu$	μ
Binomial	$\eta = log\frac{\mu}{1-\mu}$	$\mu(1-\mu)$
Negative Binomial	$\eta = log\frac{\mu}{\mu+m}$	$\mu + (\mu)^2/m$
Gamma	$\eta = \mu^{-1}$	μ^2
Inverse Gaussian	$\eta = \mu^{-2}$	μ^3

4 Inference for GLM (i.e. Estimation and hypothesis testing)

The likelihood method is used for parameter estimation and inference of GLMs as below:

$$l(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \left(\frac{y_i \boldsymbol{\theta_i} - b(\boldsymbol{\theta_i})}{a(\phi)} + c(y_i, \phi) \right) \quad (1)$$

Note that the regression parameters $\boldsymbol{\beta}$ is implicit in the loglikelihood function $l(\boldsymbol{\beta}, \phi)$ since

$$\begin{aligned} g(E(y_i)) &= \boldsymbol{X_i}^\top \boldsymbol{\beta}, \\ E(y_i) &= b'(\boldsymbol{\theta_i}) \\ &= \partial b(\boldsymbol{\theta_i}) / \partial \boldsymbol{\theta_i} \end{aligned}$$

The likelihood equation for $\boldsymbol{\beta}$ is given by

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

The resulting solution is a candidate for the MLE of $\boldsymbol{\beta}$. Since the loglikelihood $l(\boldsymbol{\beta}, \phi)$ is nonlinear in the parameters $\boldsymbol{\beta}$ and ϕ, MLEs are obtained using an iterative algorithm such as the Newton-Raphson method or the iteratively reweighted least squares method described in McCullagh and Nelder (1989). Since an iterative algorithm is used, sometimes convergence of the algorithm can be an issue, such as non-convergence, especially when the observed data are poor or the model is too complex.

Statistical inference for GLMs is often based on the **deviance**, which can be defined as the difference between the log-likelihoods for the **full model** and for the **fitted model(i.e. Restricted model)**

Firtly, we define Restricted model under the null and alternative hypotheses:

$$H_0 : \boldsymbol{R\beta} = \mathbf{0} \quad \text{vs.} H_1 : \boldsymbol{R\beta} \neq \mathbf{0}, \quad (2)$$

where $\boldsymbol{R}$ is an appropriate known matrix with full rank $p_2 \leq p$ and $\mathbf{0}$ is a $p_2 \times 1$ vector . For example, we consider the partition of the full parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, are $p_1 \times 1$, $p_2 \times 1$, vectors, respectively. Note that by letting $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, one can test whether $\boldsymbol{\beta}_2 = \mathbf{0}$.

We now form a modified log-likelihood function (1) under restriction $\boldsymbol{R\beta} = \mathbf{0}$ in the null hypothesis (2) as $l(\boldsymbol{\beta}, \phi) + \boldsymbol{\lambda}^\top (\boldsymbol{R\beta} - \mathbf{0})$, where $\boldsymbol{\lambda}$ is a $p_2 \times 1$ vector of Lagrange multipliers.

This likelihood function can be maximized by using the same methodology method discussed as above for full model with an appropriate choice of matrix $\boldsymbol{R}$. We called this fitted model or restricted model and it is denoted by $l(\tilde{\boldsymbol{\beta}}, \boldsymbol{\phi})$. Therefore, **deviance** for fitted model $l(\tilde{\boldsymbol{\beta}}, \boldsymbol{\phi})$ is defined as

$$D = 2[l(\boldsymbol{\beta}, \boldsymbol{\phi}) - l(\tilde{\boldsymbol{\beta}}, \boldsymbol{\phi})]$$

- **Note;** The **deviance** is simply 2 times the log-likelihood ratio statistic of fitted model compared to the full model, so the **deviance** measures how close fitted model is to the full model (the perfect model). For linear regression models, the deviance is simply the residual sum of squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)$, which usually measures the goodness-of-fit of the model or the discrepancy between observed data and fitted values. An alternative measure of discrepancy is the so-called Pearsons χ^2 statistic defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)}{Var(\hat{\mu}_i)}$$

5 Model Selection and Model Diagnostics

Model Selection:
In this chapter, only Akaike information criterion (AIC) will use for model selection of GLMs. The value of AIC is given by

$$AIC = -2logl(\boldsymbol{\beta}, \boldsymbol{\phi}) + 2p,$$

where the first term measures the goodness of fit and the second term is a penalty for the number of parameters in the model. Thus, AIC describes a tradeoff between accuracy and complexity of the model or between bias and variance. Given a set of candidate models, the model with the smallest AIC value is preferred but AIC do not test the significance of a model over an alternative. Hypothesis tests can be used to compare models with significance. Two types of hypothesis tests are often considered:

- **Goodness of fit test:** test if the current model fits the observed data well by comparing the current model with the full model. i.e. for goodness of fit tests, under some regularity conditions, the scaled deviance $D(y)/\boldsymbol{\phi}$ and the Pearsons χ^2 statistics are both asymptotically distributed as the $\chi^2(d)$ distribution, where d is the number of parameters in fitted model .

- **Compare two nested models**: When comparing two nested models, under the null hypothesis of no difference between the two models, the difference in the deviances of the two models asymptotically follows the χ^2 distribution, with degrees of freedom "d" being the difference of the number of parameters in the two models being compared, i.e.,

$$D_R - D_F \to \chi^2(d), n \to \infty$$

where D_R and D_F are the deviances for the restricted model and the full model respectively. This result can be used for model comparison and model selection. For example, if the p-value is large (say larger than 0.05), we prefer the restricted model since there is no significant difference between the two models. When the p-value is small (say, less than 0.05), we prefer the larger model since the larger model fits significantly better

than the smaller model. Note that the above χ^2 approximations are more accurate when comparing nested models than for the goodness of fit statistic.

- **Wald-type test**: It use for testing the significance of individual predictors in a model. Suppose that we wish to test the significance of predictor x_j i.e. testing the hypotheses $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. The test statistic is given by

$$z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1)$$

which is asymptotically under H_0.
So H_0 is rejected (i.e., predictor x_j is significant) if the p-value is small (say, less than 0.05).

Model diagnostics: For GLM model diagnostics, we can use the following two approaches:

- **Pearson residuals**:

$$r_{ip} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}$$

which are usual residuals scaled by the standard deviations. Note that $\sum_{i=1}^{n} {r_{ip}}^2$ is simply the familiar Pearson statistic.

- **Deviance residuals** r_{iD},

$$Deviance = \sum_{i=1}^{n} {r_{iD}}^2 \equiv \sum_{i=1}^{n} d_i$$

Thus, we can write as the deviance residual is the signed squared root of the unit deviance

$$r_{iD} = sign(y_i - \hat{\mu}_i)\sqrt{d_i}, i = 1, 2, .., n$$

- **Note**: When fitting a GLM to a dataset, model diagnostics often include the following:

1. Check if there are any outliers or influential observations, and if so compare analysis without these observations to analysis with these observations.
2. check whether the structure form of the model is reasonable, which includes choices of predictors and possible transformations of the predictors.
3. check whether the stochastic part of the model is reasonable, such as the distributional assumptions or the nature of the variance

.

- **Note**: Common diagnostic methods include

1. **residual plots**: we can plot the deviance residuals against the estimated linear predictor η_i. This may help the choice or transformations of the predictors, but for a binary response the residual plot is not very useful due to the nature of the response data.

2. **Transformation of the predictors**: polynomial terms or interaction terms or other transformations may be considered.
3. **Cooks distance plot**: check influential observations

6 Over-Dispersion Problem

Over-Dispersion Problem does not exist for linear regression models as the response is assumed to a normal distribution in which the variance is unrelated to the mean. That is, if y_i follows a normal distribution $N(\mu, \sigma^2)$, the mean μ and the variance σ^2 are independent and both can vary freely, which allows great flexibility in modelling real data. On the other hand, Over-Dispersion is common for GLMs due to possible relationship between mean and variance i.e. in a GLM, we assume a mean structure $g(E(y_i)) = \boldsymbol{X_i}^\top \boldsymbol{\beta}$ and assume that y_i follows a distribution in the exponential distribution, but the observed variation in the data may be different from the theoretical variance obtained from the assumed distribution. For example, for the Poisson regression model, the theoretical variance is the same as the mean, which is $E(y_i) = g(\boldsymbol{X_i}^\top \boldsymbol{\beta})^{-1}$ but the variation in the data may be much larger or much smaller than the theoretical variance, so the assumed model is inappropriate.
If the variation in the data is larger (or smaller) than the theoretical variance determined by the assumed distribution, the problem is called an over-dispersion (or a under-dispersion) problem. When over-dispersion or under-dispersion problem arises in data analysis, the assumed distribution for the GLM does not hold, so this problem must be addressed for correct inference. Usually over-dispersion problems are more common than under-dispersion problems. Overdispersion problems can arise in longitudinal or clustered data if the correlation within clusters are not incorporated in the models.
One way to address the over-dispersion problem is to specify the mean and variance functions separately in a GLM, without a distributional assumption. This approach is called the quasi-likelihood method. For example, for a logistic regression model, we may assume that

$$Var(y_i) \quad = \quad \tau E(y_i)(1 - E(y_i))$$

where τ is called a dispersion parameter. If $\tau = 1$, then y_i follows a binomial distribution. If $\tau \neq 1$, then y_i does not follows a binomial distribution. In this case, the model is called **quasibinomial**, which is not a parametric distribution, so the modified likelihood is called a **quasi-likelihood**.
Similarly, for Poisson model, we may assume that

$$Var(y_i) \quad = \quad \tau E(y_i)$$

When $\tau \neq 1$, y_i does not follow a Poisson distribution, and the resulting **quasi-likelihood** model is called **quasi-Poisson**.

7 Logistic Regression Models

A logistic regression model is used when the response y is a binary variable taking only two possible values (say, 0 or 1), which is very common in practice such as success/failure, death/alive, and cancer/health, etc. In this case, it may be reasonable to assume that y follows a binomial or Bernoulli distribution. In logistic regression model, we have several choices of link functions

such as logit link, probit link and complementary log-log link but logit link is most popular choice due to attractive interpretation as compared to others. The logit link can be written as

$$g(\mu) = log(\frac{\mu}{1-\mu})$$

where $\mu = E(y) = P(y = 1)$. The resulting GLM is the following logistic regression model

$$log(\frac{\mu_i}{1-\mu_i}) = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p, i = 1, 2, ..., n$$

where $\mu_i = E(y_i) = P(y_i = 1)$ and $y_i = 1$ is assumed to follow a Binomial distribution with mean μ_i.

Important features of Logistic Regression Models described below based on the model selection, model diagnostics, and parameters estimation:

- Residuals are not well defined for logistic regression models due to the value of response variable either 1 or 0. i.e. **residual plots** are not very useful for checking the **goodness-of-fit** of logistic regression models.
- Model selection or variable selection for logistic regression models can be based on the χ^2 test of deviances, which is similar to the likelihood ratio test. Note that, for binary data, the deviance does not assess goodness of fit and it is not approximately χ^2 distributed. Therefore, comparing two nested model of χ^2 test based on the difference of the two deviances is reasonable for binary data.
- **Wald-type z-test** is useful to test the significance of a single continuous predictor.
- **AIC criteria** is important for model or variable selections.
- For logistic regression models, the **MLEs** of model parameters are usually obtained based on an iterative algorithm such as the **Newtons method**. In some cases, the algorithm may not convergence. The reasons of non-convergence of a model may be as follows:

  1. Too many 0's or too many 1's in the response values.
  2. Multicollinearity problem

  .

Example 1. Low birth weight (less than 2500 grams) is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. Moreover, low birth babies usually suffer from many chronic conditions in their adulthood such as obesity, diabetes, and cardiovascular disease. The obstetrical literature provides evidence that a woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. In this exercise, we will use a 1986 study at the Baystate Medical Center in Springfield, MA in which data were collected from 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. The goal of the study was to identify risk factors associated with giving birth to a low birth weight baby. See: Hosmer and Lemeshow, Applied Logistic Regression: Second Edition, 2000.

1. **ID**: Mothers identification number (1-189)
2. **LOW**: 1 if birth weight less than 2.5kg (low birth weight), 0 otherwise
3. **AGE**: Mother's age in years
4. **LWT**: Mother's weight in pounds at last menstrual period
5. **RACE**: Mothers race (1=white, 2=black, 3=other)
6. **SMOKE**: Smoking status during pregnancy (1 if yes, 0 if no)
7. **PTL**: History of premature labor (0 = None 1 = One, etc.)
8. **HT**: History of hypertension (1 = Yes, 0 = No)
9. **UI**: Presence of uterine irritability (1 = Yes, 0 = No)
10. **FTV**: Number of physician visits during the first trimester (0 = None, 1 = One, 2 = Two, etc.)
11. **BWT**: birth weight (in grams).
    Note: This data set is available in R-package **"MASS"** by the name of **"birthwt"**.

We will use the binary logistic regression to develop a model that can estimate the probability of low birth weight (defined as a baby weighing less than 2500 grams) given the mothers age, the weight during her last menstrual period, race, whether she smoked during the pregnancy, number of previous premature labours, whether she had any hypertension, presence of uterine irritability, and number of physician visits during the first trimester.

Solution:

- Firstly, Categorical variables such as race" and others need to be declared in **R** using the **factor()** function so that they will not be treated as numerical variables by computer. see the below code:

```
rm(list=ls())
library(faraway)
library(MASS)

############## Data of Low Infant Birth Weight #######
data(birthwt)

#### Use factor() function ##############

bwt <- with(birthwt, {
  race <- factor(race, labels = c("white", "black", "other"))
  ptd <- factor(ptl > 0)
  ftv <- factor(ftv)
  levels(ftv)[-(1:2)] <- "2+"
  data.frame(low = factor(low), age, lwt, race, smoke = (smoke > 0),
             ptd, ht = (ht > 0), ui = (ui > 0), ftv)
})
options(contrasts = c("contr.treatment", "contr.poly"))
```

- Fit binary logistic regression as a Full model

```
fit1 = glm(low ~ .,family = binomial, data = bwt)

> summary(fit1)
........
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.82302    1.24471   0.661  0.50848
age         -0.03723    0.03870  -0.962  0.33602
lwt         -0.01565    0.00708  -2.211  0.02705 *
raceblack    1.19241    0.53597   2.225  0.02609 *
raceother    0.74069    0.46174   1.604  0.10869
smokeTRUE    0.75553    0.42502   1.778  0.07546 .
ptdTRUE      1.34376    0.48062   2.796  0.00518 **
htTRUE       1.91317    0.72074   2.654  0.00794 **
uiTRUE       0.68019    0.46434   1.465  0.14296
ftv1        -0.43638    0.47939  -0.910  0.36268
ftv2+        0.17901    0.45638   0.392  0.69488
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 195.48  on 178  degrees of freedom
AIC: 217.48
Number of Fisher Scoring iterations: 4
```

- Next, Stepwise method for variable selection using **stepAIC()** command in R.

```
fit2= stepAIC(fit1,direction = c("both"),k = 2)

summary(fit2)
..........
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.125326   0.967561  -0.130  0.89694
lwt         -0.015918   0.006954  -2.289  0.02207 *
raceblack    1.300856   0.528484   2.461  0.01384 *
raceother    0.854414   0.440907   1.938  0.05264 .
smokeTRUE    0.866582   0.404469   2.143  0.03215 *
ptdTRUE      1.128857   0.450388   2.506  0.01220 *
htTRUE       1.866895   0.707373   2.639  0.00831 **
uiTRUE       0.750649   0.458815   1.636  0.10183
---
```

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

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 197.85  on 181  degrees of freedom
AIC: 213.85
Number of Fisher Scoring iterations: 4
```

We see that the stepwise method based on AIC values only removes variables **age, "ftv1" and "ftv2+"** from the original model.

- Compare the two nested models using the deviance test to see if they are significantly different using **anova()** in R.ver

```
> ################## Compare the two nested models using deviance test####
> anova(fit2, fit1)
Analysis of Deviance Table

Model 1: low ~ lwt + race + smoke + ptd + ht + ui
Model 2: low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
  Resid. Df Resid. Dev Df Deviance
1       181     197.85
2       178     195.48  3   2.3761
```

We see that the difference in deviances from the two models is 2.3761, and the two models differ by 3 parameters (3 degrees of freedom). Compared with a $\chi^2(3)$ -distribution 5th percentile critical value, we see that models 1 and 2 are not significantly different (at 5 percentage level), and the p-value of the χ^2 test is 0.498. Thus, we should choose the smaller model(**Restricted model or fitted model**), which is **fit2** i.e. model 1.

- Model selections or variable selections often involve comparing nested models. R function **anova()** can also be used to test each covariate sequentially, i.e., it can compare models by dropping/adding one covariate at a time using a χ^2 test. This approach is especially helpful for **categorical covariates** since the **Wald-type z-tests** may not be good choice for categorical variables.

```
anova(fit2, test="Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: low
```

```
Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    188     234.67
lwt    1   5.9813       187     228.69 0.014458 *
race   2   5.4316       185     223.26 0.066153 .
smoke  1   8.2444       184     215.01 0.004088 **
ptd    1   7.9752       183     207.04 0.004742 **
ht     1   6.5572       182     200.48 0.010446 *
ui     1   2.6307       181     197.85 0.104817
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- To check **over-dispersion** using **quasibinomial model**

```
 ################### To check over-dispersion ###

> fit3 = glm(low ~ .,family = quasibinomial, data = bwt)

> summary(fit3)
..........
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.823019   1.249114   0.659  0.51082
age         -0.037234   0.038839  -0.959  0.33902
lwt         -0.015653   0.007105  -2.203  0.02888 *
raceblack    1.192413   0.537859   2.217  0.02789 *
raceother    0.740685   0.463376   1.598  0.11172
smokeTRUE    0.755528   0.426519   1.771  0.07821 .
ptdTRUE      1.343763   0.482319   2.786  0.00591 **
htTRUE       1.913166   0.723284   2.645  0.00890 **
uiTRUE       0.680195   0.465982   1.460  0.14613
ftv1        -0.436380   0.481088  -0.907  0.36560
ftv2+        0.179009   0.457991   0.391  0.69637
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasibinomial family taken to be 1.007082)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 195.48  on 178  degrees of freedom
```

The estimated dispersion parameter is $\hat{\tau} = 1.007082$, very close to 1. Thus, the assumed **binomial distribution** may hold or there may be **no over-dispersion** problem, suggesting that the results from **fit2** may be reliable.

8 Poisson Regression Models

If the response y is a count, it may be reasonable to assume that y follows a Poisson distribution. Then, the Poisson GLM is a natural choice. For the Poisson GLM, the standard link function is the following log-link

$$g(\mu) = log(\mu)$$

where $\mu = E(y)$. The resulting GLM is called a Poisson GLM and it can be written as follows

$$log(\mu_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p, i = 1, 2, ..., n$$

where $\mu_i = E(y_i)$ and y_i is assumed to follow a Poisson distribution with mean μ_i.

- **Note**:If an over-dispersion problem exists, the parameter estimates based on an assumed Poisson GLM will still be consistent, but the standard errors will be wrong. Thus, the overdispersion problem must be addressed in order for the statistical inference to be valid.

- Where there is an over-dispersion exist in count data, an **alternative model** of choice is the **negative binomial GLM**, rather than Poisson GLM.

- To check the goodness of fit of a Poisson model, we can check the deviance against a χ^2 distribution. For Poisson models, an alternative goodness of fit measure is the Pearsonsχ^2 statistic

  $$\chi^2 = \sum_1^n \frac{(y_i - \hat{\mu_i})}{\hat{\mu_i}}$$

  which is like scaled (squared) differences between observed counts and estimated counts based on the assumed model.

- For comparing two nested models, we can also use a χ^2 distribution based on the difference of the deviances of the two models.

Example.2(Species diversity on the Galapagos Islands) Gala data is available in the "faraway" r-package that contain 30 Galapagos islands and 7 variables. The relationship between the number of plant species and several geographic variables is of interest. Faraway suggested that log transformation of all the predictors is helpful to see relationship with number of plant species through Poisson Regression Models. Authors used Poisson Regression Models as below:

$$\begin{aligned} mod1 &= glm(Species \sim log(Area) + log(Elevation) + log(Nearest) + log(Scruz + 0.1) \\ &+ log(Adjacent), family = poisson, gala) \end{aligned}$$

Solution:

- Firstly, Fit Possion GLM Model by taking a log transformation of all the predictors and find the summary as below:

```
> summary(mod1)

Call:
glm(formula = Species ~ log(Area) + log(Elevation) + log(Nearest) +
    log(Scruz + 0.1) + log(Adjacent), family = poisson, data = gala)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.287941   0.284661  11.550  < 2e-16 ***
log(Area)       0.348445   0.018029  19.327  < 2e-16 ***
log(Elevation)  0.036421   0.056983   0.639  0.52272
log(Nearest)   -0.040644   0.013781  -2.949  0.00318 **
log(Scruz + 0.1)-0.030045  0.010492  -2.864  0.00419 **
log(Adjacent)  -0.089014   0.006948 -12.812  < 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: 3510.73  on 29  degrees of freedom
Residual deviance:  359.12  on 24  degrees of freedom
AIC: 531.96
```

- For a **Poisson GLM**, if the assumed Poisson distribution holds, the **mean** should be equal to the **variance**. Lets check to see if this is consistent with the observed data by plotting the **estimated variance** in the data against the **estimated mean**.

Figure 1 shows that the **estimated variance** is much larger than the **estimated mean** (both in log-scale), so there is an **over-dispersion**.

- We can estimate the **dispersion parameter** as below

```
DIS = sum(residuals(mod1 ,type="pearson")^2)/mod1$df.res
> DIS
[1] 16.55919
```

The estimated dispersion parameter is **16.55919**, which is much larger than 1, so it confirms the **over-dispersion** problem. Therefore, the assumed Poisson distribution for the Poisson GLM does not hold.

- The over-dispersion problem will only affect the **standard errors** of the parameter estimates in the GLM, so for correct inference of the parameters we should adjust the standard errors only as below.

```
summary (mod1, dispersion = DIS)
glm(formula = Species ~ log(Area) + log(Elevation) + log(Nearest) +
```

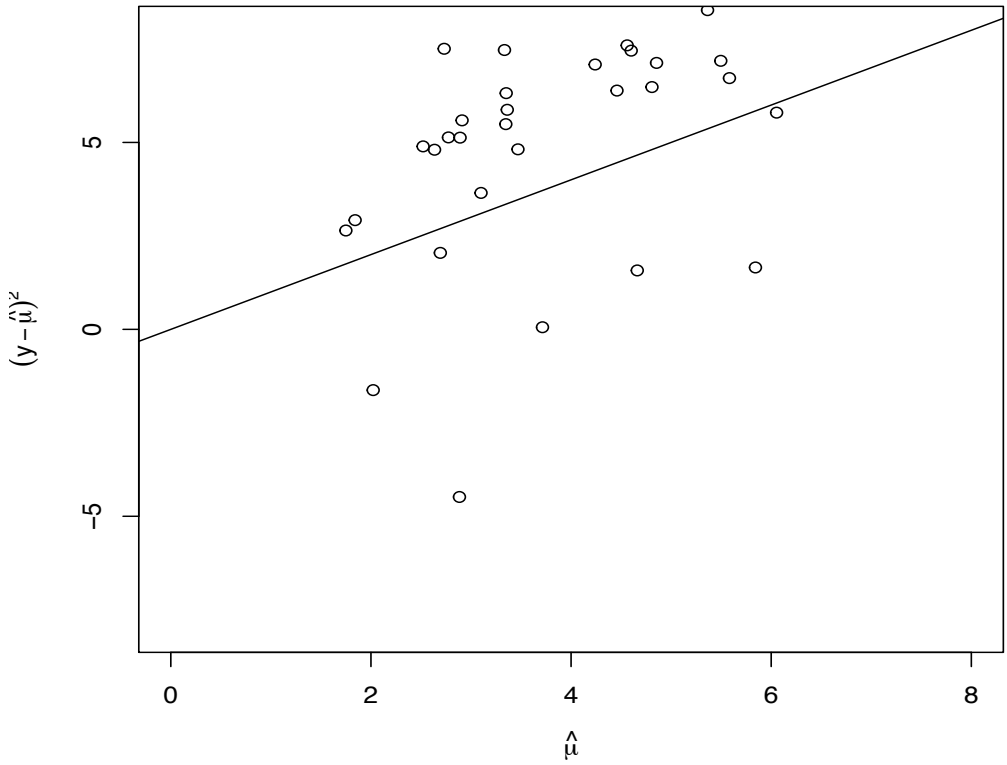

Figure 1: Estimated mean (in log-scale) versus estimated variance (in log-scale). Over-dispersion is obvious.

```
    log(Scruz + 0.1) + log(Adjacent), family = poisson, data = gala)

..........
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.28794    1.15837   2.838  0.00453 **
log(Area)         0.34844    0.07337   4.749 2.04e-06 ***
log(Elevation)    0.03642    0.23188   0.157  0.87519
log(Nearest)     -0.04064    0.05608  -0.725  0.46859
log(Scruz + 0.1) -0.03005    0.04270  -0.704  0.48162
log(Adjacent)    -0.08901    0.02827  -3.149  0.00164 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 16.55919)

    Null deviance: 3510.73  on 29  degrees of freedom
```

```
Residual deviance:  359.12  on 24  degrees of freedom
AIC: 531.96
```

The above **standard errors**, and thus the corresponding **z-values and p-values**, are more reliable than the original ones since the dispersion problem is addressed by introducing a dispersion problem. Instead of estimating the dispersion parameter and updating the fitting, we can simply use the quasi-likelihood or **quasi-poisson method:**

```
 ######################## quasipoisson ######
> modp2 <-glm(Species ~ log(Area) + log(Elevation) + log(Nearest) +
        log(Scruz + 0.1) + log(Adjacent), family = quasipoisson, data = gala)
> summary(modp2)
..........
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.28794    1.15837   2.838  0.00908 **
log(Area)          0.34844    0.07337   4.749 7.85e-05 ***
log(Elevation)     0.03642    0.23188   0.157  0.87650
log(Nearest)      -0.04064    0.05608  -0.725  0.47559
log(Scruz + 0.1) -0.03005    0.04270  -0.704  0.48839
log(Adjacent)     -0.08901    0.02827  -3.149  0.00435 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for quasipoisson family taken to be 16.55919)
```

- We obtain the same results. The dispersion parameter is estimated to be **16.55919** , larger than 1, indicating a possible **over-dispersion** problem in **gala data**. **So, an alternative model of choice is the negative binomial GLM, rather than Poisson GLM for this data.**

9 Negative binomial GLM

The **negative binomial** requires the use of the **glm.nb()** function in the **MASS package**. The call to **glm.nb()** is similar to that of **glm()**, except **no family** is given. The following script requred for the analysis of negative binomial GLM.

- The following script requred for the analysis of negative binomial GLM

```
 ######################## Use Negative Binomial due to Over-dispersion Problem ##########
> mod3 <- glm.nb(Species ~ log(Area) + log(Elevation) + log(Nearest) +
    log(Scruz + 0.1) + log(Adjacent), data = gala)
> summary(mod3)

Call:
glm.nb(formula = Species ~ log(Area) + log(Elevation) + log(Nearest) +
    log(Scruz + 0.1) + log(Adjacent), data = gala, init.theta = 2.946482723,
    link = log)
```

Compare SE of log(nearest) for Poisson & negative Binomial G.L.M.

Poisson N.B
0.05608 | 0.35582

Standard procedure $\alpha = 0.05$

$\alpha = 0.05$ $p < \alpha$ significant $p > \alpha$ Insignificant

```
.....
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.59575    1.26579   3.631 0.000283 ***
log(Area)         0.42258    0.07683   5.500 3.79e-08 ***
log(Elevation)   -0.23268    0.24798  -0.938 0.348105
log(Nearest)     -0.08893    0.08744  -1.017 0.309118
log(Scruz + 0.1) -0.02797    0.07375  -0.379 0.704474
log(Adjacent)    -0.03720    0.03591  -1.036 0.300267
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Insignificant Covariates.

p = 0.309118 > 0.05 Insignificant

α

```
(Dispersion parameter for Negative Binomial(2.9465) family taken to be 1)

    Null deviance: 149.531  on 29  degrees of freedom
Residual deviance:  32.802  on 24  degrees of freedom
AIC: 287.84

Number of Fisher Scoring iterations: 1


              Theta:  2.946
          Std. Err.:  0.868

 2 x log-likelihood:  -273.842
```

- We know that over-dispersion problem will only affect the **standard errors** of the parameter estimates in the GLM, so for correct inference of the parameters we should adjust the standard errors only as below for Negative Binomial GLM's

Signficant

```
> summary (mod3, dispersion = DIS )

Call:
glm.nb(formula = Species ~ log(Area) + log(Elevation) + log(Nearest) +
    log(Scruz + 0.1) + log(Adjacent), data = gala, init.theta = 2.946482723,
    link = log)
.....
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.59575    5.15088   0.892    0.372
log(Area)         0.42258    0.31263   1.352    0.176
log(Elevation)   -0.23268    1.00912  -0.231    0.818
log(Nearest)     -0.08893    0.35582  -0.250    0.803
log(Scruz + 0.1) -0.02797    0.30011  -0.093    0.926
log(Adjacent)    -0.03720    0.14613  -0.255    0.799
```

$p > \alpha$ Insignificant

> 0.05

> 0.05

> 0.05

> 0.05

> 0.05

```
(Dispersion parameter for Negative Binomial(2.9465) family taken to be 16.55919)

    Null deviance: 149.531  on 29  degrees of freedom
Residual deviance:  32.802  on 24  degrees of freedom
AIC: 287.84

Number of Fisher Scoring iterations: 1


              Theta:  2.946
          Std. Err.:  0.868

 2 x log-likelihood:  -273.842
```

10 Gamma GLM

In the previous sections, we discussed binomial GLM for binary response while possion GLM and negative binomail GLM for count response but the **Gamma GLM** is suitable for **continuous** and **skewed response**. Basically, The use of the gamma GLM depends on two scenario. Firstly, we assure that the response to follow a gamma distribution. Secondly, when we may be willing to speculate on the relationship between the mean and the variance of the response but we are not sure about the distribution. so we will use gamma distribution in the context of GLM model. The density of the gamma distribution for the purposes of a GLM can be written as:

$$f(y) = \frac{1}{\Gamma(v)}(\frac{v}{\mu})^v y^{v-1} \exp^{-\frac{yv}{\mu}}, y > 0$$

where v is a shape parameters, λ is scale parameters of the distribution, $\mu = \frac{v}{\lambda}$, $E(y) = \mu$, $Var(y) = \frac{\mu^2}{v}$, and dispersion parameter is $\tau = v^{-1}$.

- **Note:** The possible choices of **link functions** for Gamma GLM are **inverse, identity or log**
- The canonical link is $\eta = \mu^{-1}$. Since $-\infty < \eta < \infty$, the link does not guarantee $\mu > 0$ which could cause problems and might require **restrictions on** β or on the range of possible predictor values. On the other hand the reciprocal link has some advantages. The Michaelis-Menten model has:

$$E(y) \quad = \quad \mu = \frac{\alpha_0 x}{1 + \alpha_1 x}$$

which can be represented after some reexpression as:

$$\eta \quad = \quad \frac{\alpha_1}{\alpha_0} + \frac{1}{(\alpha_0 x)} = \mu^{-1}$$

As x increases, $\eta \to \alpha_1/\alpha_0$, which means that the mean μ will be bounded. The **inverse link** can be **useful** in such situations where we know the **mean response** to be **bounded**.

- . The **linear link**, $\eta = \mu$, is useful for modeling sums of squares or variance components which are χ^2. This is a special case of the **gamma**.

- The **log link**, $\eta = log\mu$, should be used when the effect of the predictors is suspected to be **multiplicative** on the **mean**. When the **variance** is **small**, this approach is similar to a **Gaussian model** with a logged response.

- **Note:** If we wanted to apply a **Gaussian linear model**, the **log transform** is indicated. This would imply a **lognormal distribution** for the original response. Alternatively, if $Var(y) \propto E(y)^2$ so a **gamma GLM** is also appropriate in this situation. We are comparing both models in the following example.
  Example.1 Myers and Montgomery (1997) present data from a step in the manufacturing process for semiconductors. Four factors are believed to influence the resistivity of the wafer and so a full factorial experiment with two levels of each factor was run. Previous experience led to the expectation that **resistivity** would have a **skewed distribution** and so the need for transformation was anticipated. Therefore, the application of the Box-Cox method or past experience suggests the use of a log transformation(**lognormal distribution**) on the response. We fit the **full model** and then reduce it using **AIC-based model selection**:

```
rm(list=ls())
library(faraway)
library(graphics)

data(wafer)
################## lognormal distribution ########
Model1 <-lm(log(resist)~ .^2, wafer)

Res.= step(Model1)

summary(Res.)
........
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.31111    0.04762 111.525 4.67e-14 ***
x1+          0.20088    0.04762   4.218  0.00292 **
x2+         -0.21073    0.04762  -4.425  0.00221 **
x3+          0.43718    0.06735   6.491  0.00019 ***
x4+          0.03537    0.04762   0.743  0.47892
x1+:x3+     -0.15621    0.06735  -2.319  0.04896 *
x2+:x3+     -0.17824    0.06735  -2.647  0.02941 *
x3+:x4+     -0.18303    0.06735  -2.718  0.02635 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.06735 on 8 degrees of freedom
Multiple R-squared:  0.9471,Adjusted R-squared:  0.9008
F-statistic: 20.46 on 7 and 8 DF,  p-value: 0.000165
```

- Now we fit the corresponding **gamma GLM** and again select the model using the **AIC criterion**.
  Note: The **family** must be specified as **Gamma** rather than **gamma** to avoid confusion with the Γ **function**. We use the **log link** to be consistent with the **linear model**. This must be specified as the default is the **inverse link**.

```
######## Gamma with log link ############

gModel2 <- glm(resist ~ . ^2, family=Gamma(link=log), wafer)

GammaRes. = step(gModel2)

summary(GammaRes.)
........
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.31195    0.04757 111.677 4.62e-14 ***
x1+          0.20029    0.04757   4.211  0.00295 **
x2+         -0.21101    0.04757  -4.436  0.00218 **
x3+          0.43674    0.06727   6.493  0.00019 ***
x4+          0.03537    0.04757   0.744  0.47836
x1+:x3+     -0.15549    0.06727  -2.312  0.04957 *
x2+:x3+     -0.17626    0.06727  -2.620  0.03064 *
x3+:x4+     -0.18195    0.06727  -2.705  0.02687 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for Gamma family taken to be 0.004524942)
   Null deviance: 0.697837  on 15  degrees of freedom
Residual deviance: 0.036266  on  8  degrees of freedom
AIC: 139.2
Number of Fisher Scoring iterations: 4
```

- The comparison of **lognormal distribution vs gamma GLM** model as below:

  1. The **coefficients** and **standard errors** are almost identical of both Models.
  2. The **square root of the dispersion** corresponds to the residual standard error of the linear model(**log-normal distribution**) is **sqrt(0.0045249)= 0.067267**. The **maximum likelihood** estimate of ϕ may be computed using the **MASS package** a as below

     ```
     library(MASS)
     gamma.dispersion(GammaRes.)
     [1] 0.0022657
     ```

  3. Which gives a substantially smaller estimate, which would suggest smaller standard errors. However, it is not consistent with our experience with the Gaussian linear

model as the value of $v = \frac{1}{\phi}$ is large ($\frac{1}{0.0045249} = 221$), the **gamma distribution** is well approximated by a normal. Similarly, for the logged response linear model, a **lognormal distribution** with a small variance ($\sigma = 0.0673$) is also very well approximated by a normal. For **this reason**, there is not much to distinguish these two models. The **gamma GLM** has the advantage of modeling the **response directly** while the **lognormal** has the added convenience of working with a standard linear model.