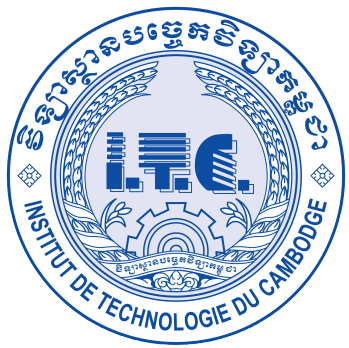
Regression Analysis

Chapter 04
Generalized Linear Models

Phauk Sokkhey
phauk.sokkhey@itc.edu.kh
Nhim Malai
nhim.malai@itc.edu.kh

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia



Introduction

Linear models = when the response variable is continuous and at least approximately normal.
In many applications the response is not continuous, but rather binary, categorical, or count variable as in the following examples:

- Patent opposition (yes/no)
- Creditworthiness of a client (yes/no)
- Benign or malignant tumor
- Person is unemployed, part-time employed, or fully employed
- Tree is very damaged, averagely or lightly damaged, or not damaged at all
- Number of cases of illness, insurance claims, or problematic credits within a certain time frame

Contents

Binary Regression Models

Assume that (ungrouped) data on n objects or individuals are given in the form $(y_i, x_{i1}, ..., x_{in})$, $i = 1, ..., n$, **with binary response** y coded by 0 and 1 and covariates denoted by $x_1, ..., x_k$.
The main goal of binary regression analysis is then to model and estimate the effects of the covariates on the (conditional) probability.

$$\pi_i = P(y_i = 1) = E(y_i),$$

Linear probability model,

$$\pi_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik},$$

The linear predictor,

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} = \boldsymbol{x}_i' \boldsymbol{\beta},$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ and $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ik})'$ **must lie in the interval** [0,1] for all vectors $\boldsymbol{x}$.

Link Functions

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}),$$

where h is a strictly monotonically increasing cumulative distribution function on the real line.

$$\eta_i = g(\pi_i),$$

with inverse function $g = h^{-1}$. h is called the *response function* and $g = h^{-1}$ is called the *link function*. Logit and probit models are the most widely used binary regression models.

Logit Model

$$\pi = h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

or (equivalently) the logit link function

$$g(\pi) = \log(\frac{\pi}{1-\pi}) = \eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k.$$

This yields a linear for the logarithmic odds (log-odds) $\log(\frac{\pi}{1-\pi})$.
Transformation with the exponential function gives

$$\frac{\pi}{1-\pi} = \exp(\beta_0)\exp(\beta_1 x_1) \cdot ... \cdot \exp(\beta_k x_k).$$

Parameter Interpretation: The parameter β determines the rate of increase or decrease of the curve presenting $\pi(x)$.

- When $\beta > 0$, $\pi(x)$ increases as x increase
- When $\beta < 0$, $\pi(x)$ decreases as x increases
- When $\beta = 0$, the curve flattens to a horizontal line

Probit Model

For h, we use the standard normal cumulative distribution function ϕ, i.e.,

$$\pi = \phi(\eta) = \phi(\boldsymbol{x}'\boldsymbol{\beta})$$

as response function, with inverse

$$\phi^{-1}(\pi) = \eta$$

A (minor) disadvantage is the required numerical evaluation of ϕ in the maximum likelihood estimation of the parameter β.

Notation: CDF of the standard normal distribution

$$\phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{u^2}{2}) du$$

Complementary Log-Log Model

The complementary log-log model uses the extreme minimum-value cumulative distribution function

$$h(\eta) = 1 - \exp(-\exp(h))$$

as response function, with inverse

$$g(\pi) = \log(-\log(1 - \pi))$$

as link function. In comparison to logit and probit models, this model is useful in more specific applications, for example, when modelling discrete duration times.

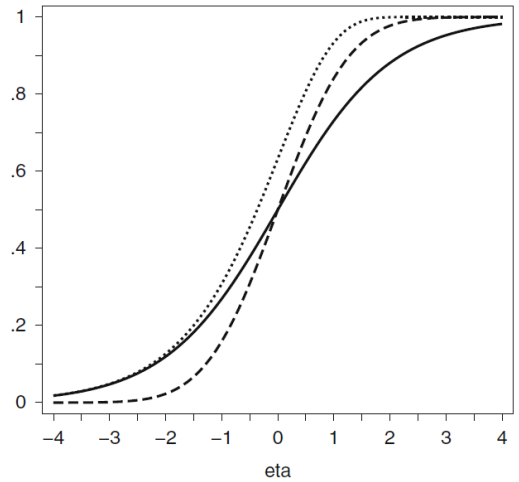


Figure 1: Response functions in binary regression models: logit model (—), probit model (- - -), and complementary log-log model ($\cdots$)

- Logit and probit models are both symmetric around 0.
- Logistic distribution function approaches 0 or 1 much slower than probit for $\eta \to -\infty$ or $\eta \to +\infty$.
- The response function of the complementary log-log is asymmetric, following a similar pattern as the logit response function for small η, but showing a faster approach towards 1 as $n \to +\infty$

Comparison of Logit and Probit Model

- For practical purposes, logistic and probit regression curves look the same.
- One seldom encounters data for which a logistic model fits well but the probit model fits poorly, or vice versa.
- Parameter estimates differ for the two models, since their links have different scales.
- The probit model was introduced in 1934 for models in toxicology.
- the logistic regression model was not studied until a decade later, but is now much more popular than the probit model.
- One **advantege** of the logistic regression model over the probit model is that the logistic regression effects can also be interpreted using odds ratios.

Example: University Admission

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

```
> ucla <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
> head(ucla)
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2
```

Models' Parameter Estimates

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{gre}_i + \beta_2 \text{gpa}_i + \sum_{j=3}^{5} \beta_j \text{rank}_{ij}$$

where $\pi_{ij} = E(Y_{ij})$ and $Y_{ij} \sim Bernoulli(\pi_{ij})$

Table 1: University Admission: Estimated Coefficients (Standard Error) of the Logit, Probit and Complementary Log-Log Models

	Logit Model Estimate (Std. Error)	**Probit Model** Estimate (Std. Error)	**Log-Log Model** Estimate (Std. Error)
(Intercept)	-3.990 (1.140)***	-2.387 (0.674)***	-3.535 (0.920)***
GRE	0.002 (0.001)*	0.001 (0.001)*	0.002 (0.01)*
GPA	0.804 (0.332)*	0.478 (0.197)*	0.641 (0.266)*
Rank (2)	-0.675 (0.316)*	-0.415 (0.195)*	-0.494 (0.228)*
Rank (3)	-1.340 (0.345)***	-0.812 (0.208)***	-1.045 (0.263)***
Rank (4)	-1.551 (0.418)***	-0.936 (0.245)***	-1.240 (0.343)***

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

Interpretation

We will give interpretation to only a few coefficient estimates as an example from the logit model since this model has odds ratio (OR) interpretation:

- $\hat{\beta}_1 = 0.02$ ($\exp(\hat{\beta}_1) = 1.02$) - an increase in GRE, resulting in an increase of 2% in the odds of getting admitted.
- $\hat{\beta}_3 = -0.675$ ($\exp(\hat{\beta}_3) = 0.509$) - the odds of getting admitted are 49.1% smaller for rank 4 compared to rank 1. (Rank 1 is the baseline in our analysis)

Example: University Admission (Grouped Data)

Suppose we got access to only the grouped data of GRE as follows:

```
> g.ucla<- ucla %>%
+   group_by(gre) %>%
+   summarise(admit=sum(admit==1), total=n())
> g.ucla<-data.frame(g.ucla)
> head(g.ucla)
  gre admit total
1 220     0     1
2 300     1     3
3 340     1     4
4 360     0     4
5 380     0     8
6 400     2    11
```

Grouped Data

- Individual data: $y_i \sim Bernoulli(\pi_i)$ with $P(y_i = 1) = \pi_i$ or $y_i \sim Binomial(1, \pi_i)$
- Assume y_i are (conditionally) independent, for grouped data:

$$n_i \bar{y}_i \sim Binomial(n_i, \pi_i)$$

$$\bar{y}_i \sim Binomial(n_i, \pi_i)/n_i$$

with the probability function

$$P(\bar{y}_i = j/n_i) = \binom{n_i}{j} \pi_i^j (1 - \pi_i)^{n_i - j} \qquad j = 0, 1, ..., n_i.$$

The mean and the variance are given by

$$E(\bar{y}_i) = \pi_i \qquad Var(\bar{y}_i = \frac{\pi_i(1 - \pi_i)}{n_i})$$

Example: University Admission (Grouped Data)

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{GRE}_i$$

where $Y_i \sim Binomial(n_i, \pi_i)$

```
> g.ucla.logit<-glm(cbind(admit, total-admit) ~ gre, family = "binomial"(link="logit"), data=g.ucla)
> summary(g.ucla.logit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.901344   0.606038  -4.787 1.69e-06 ***
gre          0.003582   0.000986   3.633  0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.337  on 25  degrees of freedom
Residual deviance: 32.417  on 24  degrees of freedom
AIC: 100.98

Number of Fisher Scoring iterations: 4>
```

Goodness of Fit Measures

In logistic regression, MLE is used rather than OLS (since OLS requires normality of the data). For more info, please refer to the reference books.

We will discuss the following four statistics for goodness-of-fit in multiple logistics regression:

1. Deviance
2. Pearson's Chi-square
3. Log-likelihood Ratio
4. AIC and BIC Statistics

Deviance

The likelihood ratio statistic for a given model M versus a "saturated" model is often called the deviance, denoted by D^2. The saturated model provides the perfect fit for the data.

$$D^2(M) = -2[\log[L(M)] - \log[L(Sat)]]$$

In a saturated model, the number of parameters equals the sample size since it contains one parameter for each observation. We have

$$L(Sat) = 1 \qquad \text{or} \qquad \log[L(Sat)] = 0$$

Then

$$D^2(M) = -2\log[L(M)]$$

- Deviance is a comparable statistic to reduce the error variance in linear regression.
- The best model tries to reduce deviance i.e. model with smaller deviance is more preferable.

Pearson's Chi-square

For a GLM with a binomial random component, the Pearson residual for the fit at setting j is:

$$e_j = \frac{y_j - n_i\hat{\pi}_j}{\sqrt{[n_j\hat{\pi}_j(1-\hat{\pi}_j)]}}$$

- If the model fits, then, asymptotically,

$$e_j \sim N(0,1)$$

  if n_j is large compared to the number of parameters!
- Hence, absolute values larger than 2 indicate a possible lack of it.

Note that the Pearson statistic for testing the model fit satisfies:

$$\chi^2 = \sum e_j^2$$

Log-Likelihood Ratio

Suppose you want to test whether model M_1 holds when the alternative is M_2, where M_1 is nested in M_2. Then the likelihood ratio statistics G^2 is as follows:

$$\begin{aligned} G^2 &= -2[\log[L(M_1)] - \log[L(M_2)]] \\ &= -2[\log[L(M_1)] - \log[L(Sat)]] + 2[\log[L(M_2)] - \log[L(Sat)]] \\ &= D^2(M_1) - D^2(M_2) \end{aligned}$$

Under the null and for large samples, $G^2 \sim \chi^2$ distribution with degree of freedom (df).

$$df = \text{ number of parameters in } M_2 - \text{ number of parameters in } M_1$$

AIC and BIC Statistics

The likelihood ratio test is used to compare models which are nested. AIC and BIC can be used to compare both nested and non-nested models. The model with a lower AIC/BIC value provides the best fit.

Akaike Information Criteria (AIC):

$$AIC = -2[\log[L(M)] - k]$$

Bayesian Information Criteria (BIC):

$$BIC = -2[\log[L(M)] - \log(n)k]$$

where

- k is the number of parameters in the model including the intercept.
- n is the sample size.

Analysis of Deviance

```
> ucla.full<-glm(admit~ gre*gpa*rank, family = "binomial", data=ucla)
> anova(ucla.full, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: admit

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         399     499.98
gre           1  13.9204       398     486.06 0.0001907 ***
gpa           1   5.7122       397     480.34 0.0168478 *
rank          3  21.8265       394     458.52 7.088e-05 ***
gre:gpa       1   2.7464       393     455.77 0.0974733 .
gre:rank      3   0.5240       390     455.25 0.9135841
gpa:rank      3   0.7067       387     454.54 0.8716287
gre:gpa:rank  3   4.1546       384     450.39 0.2452506
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Two-way and three-way interactions are not significant. We can remove them from our final model.

Likelihood Ratio Test

We can also do likelihood ratio tests to test the significance of the interaction terms one by one.

```
> ucla.logit<-glm(admit~ gre+gpa+rank, family = "binomial", data=ucla)
> ucla.2a<-glm(admit~ gre+gpa+rank+gre:gpa, family = "binomial", data=ucla)
> lrtest(ucla.2a, ucla.logit)
Likelihood ratio test

Model 1: admit ~ gre + gpa + rank + gre:gpa
Model 2: admit ~ gre + gpa + rank
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -227.89
2   6 -229.26 -1 2.7464    0.09747 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ucla.2b<-glm(admit~ gre+gpa+rank+gre:rank, family = "binomial", data=ucla)
> lrtest(ucla.2b, ucla.logit)
Likelihood ratio test

Model 1: admit ~ gre + gpa + rank + gre:rank
Model 2: admit ~ gre + gpa + rank
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -229.12
2   6 -229.26 -3 0.2824     0.9633
```

```
> ucla.2c<-glm(admit~ gre+gpa+rank+gpa:rank, family = "binomial", data=ucla)
> lrtest(ucla.2c, ucla.logit)
Likelihood ratio test

Model 1: admit ~ gre + gpa + rank + gpa:rank
Model 2: admit ~ gre + gpa + rank
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -229.06
2   6 -229.26 -3 0.4055     0.9391
```

All two-ways interaction terms are not significant at 5% level of significance. We can exclude them from our final model.

```
ucla.logit<-glm(admit~gre+gpa+rank, family="binomial"(link="logit"), data=ucla)
BIC(ucla.logit)
ucla.probit<-glm(admit~gre+gpa+rank, family="binomial"(link="probit"), data=ucla)
BIC(ucla.probit)
ucla.cloglog<-glm(admit~gre+gpa+rank, family="binomial"(link="cloglog"), data=ucla)
BIC(ucla.cloglog)
```

Model comparison

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{gre}_i + \beta_2 \text{gpa}_i + \sum_{j=3}^{5} \beta_j \text{rank}_{ij}$$

where $\pi_{ij} = E(Y_{ij})$ and $Y_{ij} \sim Bernoulli(\pi_{ij})$

Table 2: University Admission: Estimated Coefficients (Standard Error) of the Logit, Probit and Complementary Log-Log Models

	Logit Model Estimate (Std. Error)	**Probit Model** Estimate (Std. Error)	**Log-Log Model** Estimate (Std. Error)
(Intercept)	-3.990 (1.140)***	-2.387 (0.674)***	-3.535 (0.920)***
GRE	0.002 (0.001)*	0.001 (0.001)*	0.002 (0.01)*
GPA	0.804 (0.332)*	0.478 (0.197)*	0.641 (0.266)*
Rank (2)	-0.675 (0.316)*	-0.415 (0.195)*	-0.494 (0.228)*
Rank (3)	-1.340 (0.345)***	-0.812 (0.208)***	-1.045 (0.263)***
Rank (4)	-1.551 (0.418)***	-0.936 (0.245)***	-1.240 (0.343)***
Deviance	458.52	458.41	458.89
AIC	470.52	470.41	470.89
BIC	494.47	494.36	494.84

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

Conclusion

Based on Deviance, AIC and BIC among the three models, the probit model is the best-fit model. This model will be chosen for inference purposes. However, all three models lead to the same conclusion.

Overdispersion

- For grouped data, we estimate the variance within a group via $\bar{y}_i(1-\bar{y}_i)/n_i$. $\bar{y}_i$ is the MLE of π_i based on data group i.
- In application, the *empirical* variance (data) is often much larger than the variance $\hat{\pi}_i(1-\pi_i)n_i$ (theoretical) predicted by a binomial regression. This is called **overdispersion**.
- When there is overdispersion, the model doesn't fit meaning that the Deviance and the Pearson chi-square are large relative to the degree of freedom.
- Two main reasons for overdispersion:
  1. Unobserved heterogeneity.
  2. Positive correlations between the individual binary response variables.
- The easiest way to address the increased variability is through the introduction of a multiplicative overdispersion parameter $\phi > 1$ into the variance formula, i.e., we assume

$$Var(y_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$$

Overdispersion

Causes: Overdispersion can cause the analysis which assumes the logistic model **underestimates** the standard error(s) and thus, wrongly inflates the test statistics and the level of significance.

Measuring and Monitoring: Suppse that data are with replications consisting of m subgroups (with identical covariate values). Dispersion can then be measured by

- the scaled Deviance, i.e.
$$\chi^2_D/df$$
- the scaled Pearson chi-square, i.e.
$$\chi^2_P/df$$

When the values of these statistics are **much larger than one**, the assumption of binomial variability may not be valid and thus data are said to exhibit overdispersion.

Correcting Overdispersion

One way of correcting overdispersion is to multiply the covariance matrix of the parameters by the value of the overdispersion parameter ϕ, i.e.

- the scaled Deviance, or
- the scaled Pearson chi-square

In this corrections process, the parameter estimates are not changed. However, their standard errors are adjusted (increased), affecting their significance levels (reduced).

Example: Toxicity

The data investigate the toxicity of a certain chemical compound. Five groups of 20 rats each were fed for four weeks with a diet mixed with that compound at five different doses. At the end of the study, their lungs were harvested and subjected to histopathological examinations to observe for signs of toxicity (yes=1, no=0). The results were:

Group	Dose (mg)	# of Rats	# of Rats with Toxicity
1	5	20	1
2	10	20	3
3	15	20	7
4	20	20	14
5	30	20	10

Fitting a logistic model

```
> toxic.logit<-glm(cbind(n.toxic, n.rat-n.toxic)~dose, family = "binomial", data=toxicity)
> summary(toxic.logit)$coefficients
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -2.3406930 0.53804230 -4.350388 1.358966e-05
dose         0.1017379 0.02767545  3.676105 2.368224e-04
```

- Obvious sign of overdispersion
- Dose effect highly significant

```
# Scaled Deviance
> ED<-resid(toxic.logit, type = "deviance")
> Deviance<-sum(ED^2)
> Deviance/3
[1] 3.663964
> pchisq(Deviance, df = 3, lower.tail=FALSE)
[1] 0.01176981

# Scaled Pearson's chi-square
> EP<-resid(toxic.logit, type = "pearson")
> Pearson<-sum(EP^2)
> Pearson/3
[1] 3.595445
> pchisq(Pearson, df = (5 - 2), lower.tail=FALSE)
[1] 0.01293918
```

Fitting an overdispersed model, controlling for the scaled Deviance

```
> summary(toxic.logit, dispersion = Deviance/3)

Call:
glm(formula = cbind(n.toxic, n.rat - n.toxic) ~ dose, family = "binomial",
    data = toxicity)

Deviance Residuals:
      1        2        3        4        5
-1.2897  -0.6885   0.4129   2.4910  -1.5745

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.34069    1.02989  -2.273   0.0230 *
dose         0.10174    0.05297   1.920   0.0548 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 3.663964)

    Null deviance: 26.582  on 4  degrees of freedom
Residual deviance: 10.992  on 3  degrees of freedom
AIC: 29.93

Number of Fisher Scoring iterations: 4
```

- Point estimates remain the same
- Standard errors are larger
- Dose effect no longer significant at 5% level

Fitting an overdispersed model, controlling for the scaled Pearson

```
> summary(toxic.logit, dispersion = Pearson/3)

Call:
glm(formula = cbind(n.toxic, n.rat - n.toxic) ~ dose, family = "binomial",
    data = toxicity)

Deviance Residuals:
      1        2        3        4        5
-1.2897  -0.6885   0.4129   2.4910  -1.5745

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.34069    1.02022  -2.294   0.0218 *
dose         0.10174    0.05248   1.939   0.0525 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 3.595445)

    Null deviance: 26.582  on 4  degrees of freedom
Residual deviance: 10.992  on 3  degrees of freedom
AIC: 29.93

Number of Fisher Scoring iterations: 4
```

- Point estimates remain the same
- Standard errors are larger
- Dose effect no longer significant at 5% level

Quasi-Likelihood

- Overdispersion is when the empirical variance does not comply with the estimated variance.
- Quasi-likelihood models allow for a separate specification of the mean and the variance structure.

PDF of the exponential family:

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

with mean structure $E(y) = \mu = h(\boldsymbol{x}'\boldsymbol{\beta})$ and variance structure $Var(y) = \phi b''(\mu)/w$ where variance function $\nu(\mu) = b''(\mu)$.

Quasi-Likelihood

The quasi-score equation or *generalized estimating equation* (GEE) was obtained as the derivative of the log-likelihood of exponential family distribution:

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_i \frac{h'(\eta_i)}{\sigma_i^2}(y_i - \mu_i) = 0$$

QL-idea: Modify variance σ_i^2 expression in the estimating equations.
The simplest form of a (working) variance function results from overdispersion:

- Binomial model ($\mu_i = \pi_i$): $\sigma_i^2(\pi_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$
- Poisson model ($\mu_i = \lambda_i$): $\sigma_i^2(\lambda_i) = \phi\lambda_i$

where ϕ is estimated by scaled Deviance (χ^2_D/df) or scaled Pearson's chi-square (χ^2_P/df).

Quasi-Likelihood

```
> toxic.quasilogit<-glm(cbind(n.toxic, n.rat-n.toxic)~dose, family = quasibinomial, data=toxicity)
> summary(toxic.quasilogit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.34069    1.02022  -2.294    0.106
dose         0.10174    0.05248   1.939    0.148

(Dispersion parameter for quasibinomial family taken to be 3.595445)

    Null deviance: 26.582  on 4  degrees of freedom
Residual deviance: 10.992  on 3  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations
```

Contents

Count Data Regression: Poisson Regression

Many discrete response variables have counts as possible outcomes e.g

- the number of automobile thefts in a sample of cities worldwide in 1995
- the number of viruses in a solution
- the number of defective teeth in an individual
- number of suicides in New York in 1999

GLMs for these count data assume a Poisson distribution for a random component. Like counts, Poisson variates can take any non-negative integer value.
Remember, the Poisson distribution is characterized by parameter λ.

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

where

$$\lambda = E(Y)$$

Log-Linear Poisson Model

- The Poisson distribution has *positive* mean.
- Therefore, the Poisson mean in GLMs is commonly modelled using a log-link:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

- for this model, the mean satisfies the exponential relationship:

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) = \exp(\beta_0)\exp(\beta_1)^{x_{i1}} \ldots \exp(\beta_k)^{x_{ik}}$$

Interpretation

The mean of Y at $x_k + 1$ equals the mean of Y at x_k multiplied by $\exp(\beta_k)$.

- If $\beta_k = 0$, then the mean of Y does not changes as x_k changes.
- If $\beta_k > 0$, then the mean of Y increases as x_k increases.
- If $\beta_k < 0$, then the mean of Y decreases as X increases.

Overdispersion

The assumption of a Poisson distribution for the response implies

$$\lambda_i = E(y_i) = Var(y_i)$$

For similar reasons as in the case of binomial data, a significantly higher empirical variance is frequently observed in applications of Poisson regression. For this reason, it is often useful to introduce an overdispersion parameter ϕ_i by assuming

$$Var(y_i) = \phi\lambda_i$$

Dispersion can then be measured by

- the scaled Deviance, i.e.

$$\chi^2_D/df$$

- the scaled Pearson chi-square, i.e.

$$\chi^2_P/df$$

When the values of these statistics are **much larger than one**, the assumption of Poisson variability may not be valid and thus data are said to exhibit overdispersion.

Example: Species diversity on the Galapagos Islands

There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and 5 geographic variables is of interest:

- `Species`: the number of plant species found on the island
- `Area`: the area of the island (km^2)
- `Elevation`: the highest elevation of the island (m)
- `Nearest`: the distance from the nearest island (m)
- `Scruz`: the distance from the from Santa Cruz island (km)
- `Adjacent`: the area of the adjacent island (km^2)

Loading Dataset

```
> library(faraway)
> data(gala)
> gala = gala[,-2]
> glimpse(gala)
Rows: 30
Columns: 6
$ Species   <dbl> 58, 31, 3, 25, 2, 18, 24, 10, 8, 2, 97, 93, 5...
$ Area      <dbl> 25.09, 1.24, 0.21, 0.10, 0.05, 0.34, 0.08, 2....
$ Elevation <dbl> 346, 109, 114, 46, 77, 119, 93, 168, 71, 112,...
$ Nearest   <dbl> 0.6, 0.6, 2.8, 1.9, 1.9, 8.0, 6.0, 34.1, 0.4,...
$ Scruz     <dbl> 0.6, 26.3, 58.7, 47.4, 1.9, 8.0, 12.0, 290.2,...
$ Adjacent  <dbl> 1.84, 572.33, 0.78, 0.18, 903.82, 1.84, 0.34,...

> summary(gala)
    Species            Area             Elevation         Nearest          Scruz
 Min.   :  2.00   Min.   :   0.010   Min.   :  25.00   Min.   : 0.20   Min.   :  0.00
 1st Qu.: 13.00   1st Qu.:   0.258   1st Qu.:  97.75   1st Qu.: 0.80   1st Qu.: 11.03
 Median : 42.00   Median :   2.590   Median : 192.00   Median : 3.05   Median : 46.65
 Mean   : 85.23   Mean   : 261.709   Mean   : 368.03   Mean   :10.06   Mean   : 56.98
 3rd Qu.: 96.00   3rd Qu.:  59.237   3rd Qu.: 435.25   3rd Qu.:10.03   3rd Qu.: 81.08
 Max.   :444.00   Max.   :4669.320   Max.   :1707.00   Max.   :47.40   Max.   :290.20
    Adjacent
 Min.   :   0.03
 1st Qu.:   0.52
 Median :   2.59
 Mean   : 261.10
 3rd Qu.:  59.24
 Max.   :4669.32
```

Fitting Log-Linear Poisson Model

```
> gala.pois = glm(Species ~ .,family=poisson, data=gala)
> summary(gala.pois)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 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:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

Deviance/df $= 716.85/24 = 29.87$ (so much greater than 1) indicating of overdispersion exists.

Scaled Deviance

```
> #Dispersion parameter (phi = deviance / df)
> 716.85/24 # 29.86875
[1] 29.86875
> summary(gala.pois, dispersion = 29.86875)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.1548079  0.2828231  11.155  < 2e-16 ***
Area        -0.0005799  0.0001436  -4.039 5.37e-05 ***
Elevation    0.0035406  0.0004777   7.412 1.25e-13 ***
Nearest      0.0088256  0.0099536   0.887    0.375
Scruz       -0.0057094  0.0034192  -1.670    0.095 .
Adjacent    -0.0006630  0.0001603  -4.137 3.52e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 29.86875)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

Number of Fisher Scoring iterations: 5
```

Quasi-Likelihood

```
> gala.quasipois = glm(Species ~ .,family=quasipoisson, data=gala)
> summary(gala.quasipois)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.1548079  0.2915901  10.819 1.03e-10 ***
Area        -0.0005799  0.0001480  -3.918 0.000649 ***
Elevation    0.0035406  0.0004925   7.189 1.98e-07 ***
Nearest      0.0088256  0.0102622   0.860 0.398292
Scruz       -0.0057094  0.0035251  -1.620 0.118380
Adjacent    -0.0006630  0.0001653  -4.012 0.000511 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 31.74921)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```