

STAT 40001/STAT 50001 Statistical Computing

Lecture 18

Department of Mathematics and Statistics
Purdue University Northwest



PURDUE
UNIVERSITY
NORTHWEST

Logistic Regression

Logistic regression is a part of a category of statistical models called generalized linear models (GLM). The GLM is a unification of both linear and nonlinear regression models that also allows the incorporation of nonnormal response distributions. In GLM, the response variable distribution must be a member of exponential family which includes normal, Poisson, binomial, exponential, gamma etc. This broad class of models includes ordinary regression and ANOVA, as well as multivariate statistics such as ANCOVA and loglinear regression.

Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure. Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables are a categorical, or a mix of continuous and categorical, logistic regression is preferred.

Logistic Regression

The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success π , or the value 0 with probability of failure $1 - \pi$. This type of variable is called a Bernoulli (or binary) variable. Although not as common, applications of logistic regression have also been extended to cases where the dependent variable is of more than two cases, known as multinomial or polytomous [Tabachnick and Fidell (1996) use the term polychotomous].

As mentioned previously, the independent or predictor variables in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of π :

Sigmoidal Response Functions for Binary Responses

The three response functions which can be used for modeling binary responses from zero-one coding are the following:

- Probit mean Response Function
- Logistic mean response function
- Complementary log log response function.

Note that all of these response functions have the following properties:

- Bounded between 0 and 1.
- They are sigmoidal means S-shaped
- Approaches asymptotically to 0 and 1.

Simple Logistic Regression

When the response variable is binary, taking 1 and 0 with probabilities π and $1 - \pi$, respectively, y is a Bernoulli random variable with parameter $E(y) = \pi$.

The simple logistic regression model is given by

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

We know that $E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \implies \pi_i = E(y_i)$

Straightforward algebra yields

$$E(y_i) = \pi_i = [1 + \exp(-\beta_0 - \beta_1 x_i)]^{-1}$$

The x observations are assumed to be known constants. If x observations are random, $E(y_i)$ is viewed as a conditional mean given the value of x_i .

The parameter π defines the mean of the distribution: $E(y_i) = \pi$. The logistic regression models the success probability as a function of the severity(x). That is, $\pi = \pi(x)$ so that for case i with severity x_i , $\pi_i = \pi(x_i)$.

Simple Logistic Regression

Once the MLE b_0 and b_1 are found, we substitute these values in the response function

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

to obtain the response function

$$\hat{\pi}_i = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$$

Hence, the logistic response function will be

$$\hat{\pi} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

Considering the logit transformation we can rewrite the response function as

$$\hat{\pi}' = b_0 + b_1 x$$

where $\hat{\pi}' = \ln \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right)$.

Remark: The logarithm of the odds $\ln \left[\frac{\pi}{1 - \pi} \right]$, is referred to as the log odds or the logit. It is clear that the logistic regression model assumes a linear model for the logit: that is

What is b_1 in a logistic regression?

In the case of a linear regression the estimated regression coefficient b_1 represent the slope of the fitted line but in case of the logistic regression it represents the the logarithm of the ratio of odds of two consecutive observations. This means

$$b_1 = \log \left(\frac{odds_2}{odds_1} \right)$$

where $\log(odds_1)$ is the logarithm of the estimated odds when $x = x_j$ and $\log(odds_2)$ is the logarithm of the estimated odds when $x = x_{j+1}$. Hence on taking the antilog in the above expression we have the estimated ratio of odds, called the odds ratio denoted by \widehat{OR} is equal to $\exp(b_1)$.

Therefore

$$\widehat{OR} = \frac{odds_2}{odds_1} = \exp(b_1)$$

For example, a regression coefficient $b_1 = -0.2$ with $\exp(b_1) = 0.82$ indicates that a change from x to $x + 1$ reduces the odds of occurrence by the multiplicative factor 0.82; it reduces the odds of occurrence by 18%. A value of $b_1 = 0$ and $\exp(b_1) = 1$ a change in the explanatory variable has no effect on the odds of occurrence.

Example

The board of directors of professionals association conducted a random sample survey of 30 members to assess the effects of several possible amount os dues increase. The table below is the result of the survey. X denotes the dollar increase in annual dues posited in the survey interview and $Y = 1$ if the interviewee indicated the membership will not be renewed at the amount of the dues increase and 0 denotes if the membership will be renewed.

y	x	y	x	y	x
0.0	30.0	1.0	30.0	0.0	30.0
0.0	31.0	0.0	32.0	0.0	33.0
1.0	34.0	0.0	35.0	0.0	35.0
1.0	35.0	1.0	36.0	0.0	37.0
0.0	38.0	1.0	39.0	0.0	40.0
1.0	40.0	1.0	40.0	0.0	41.0
1.0	42.0	1.0	43.0	1.0	44.0
0.0	45.0	1.0	45.0	1.0	45.0
0.0	46.0	1.0	47.0	1.0	48.0
0.0	49.0	1.0	50.0	1.0	50.0

Assuming logistic regression is appropriate we have the following maximum likelihood parameter estimates: $\hat{\beta}_0 = -4.8075$ and $\hat{\beta}_1 = 0.1251$. Hence the fitted response function is given by

$$\hat{\pi} = [1 + \exp(4.8075 - 0.1251x_i)]^{-1}$$

Example

```
> data=read.table("C://STAT4001//membership.txt", header=T)
> y<-data$y
> x<-data$x
> model=glm(y~x,family=binomial(logit))
> model
Call:  glm(formula = y ~ x, family = binomial(logit))
Coefficients:
(Intercept)          x
    -4.8075      0.1251
Degrees of Freedom: 29 Total (i.e. Null);  28 Residual
Null Deviance:      41.46
Residual Deviance: 37.46      AIC: 41.46
```

We can display the fitted model using R code below

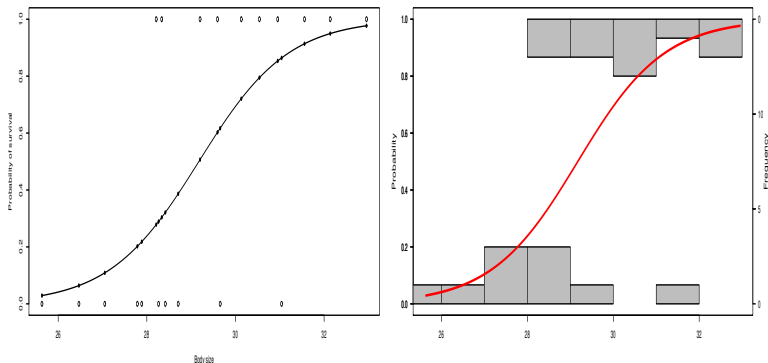
```
> plot(x,y)
> curve(predict(model,data.frame(x=x),type="resp"),add=TRUE)
> points(x,fitted(model),pch=20)
```

Example- Ecological Data

The survival response and body size conducted in a ecological study

```
> data
  bodysize survive
1 25.63320      0
2 26.46724      0
3 27.05107      0
4 27.78609      0
5 27.88550      0
6 28.21167      1
7 28.26452      0
8 28.33589      1
9 28.41717      0
10 28.70792      0
11 29.20096      1
12 29.59447      1
13 29.65526      0
14 30.13143      1
15 30.54080      1
16 30.95309      1
17 31.04085      0
18 31.55717      1
19 32.13806      1
20 32.95669      1
> bodysize=data$bodysize
> survive=data$survive
> plot(bodysize,survive,xlab="Body size",ylab="Probability of survival")
> g=glm(survive~bodysize,family=binomial)
> curve(predict(g,data.frame(bodysize=x),type="resp"),add=TRUE)
> points(bodysize,fitted(g),pch=20)
```

Example- Ecological Data



We can draw the logistic curve along with the histogram (displayed in the right above) with R code

```
>library(popbio)
>logi.hist.plot(bodysize,survive,boxp=FALSE,type="hist",col="gray")
```

Logistic Regression - Goodness of Fit Measures

In ordinary least square (OLS) regression, the R^2 statistics measures the amount of variance explained by the regression model. The value of R^2 ranges in $[0,1]$ with a larger value indicating more variance is explained by the model (higher value is better). For OLS regression, is defined as following.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Pseudo- R^2 measures proportion reduction in log-likelihood over null model. This is a useful measure, but more like another F-test than a measure of model explanatory power. For logistic regression, there have been many proposed pseudo- R^2 including

- Efron's R^2
- McFadden's R^2
- McFadden's Adjusted R^2
- Cox & Snell R^2
- Nagelkerke/Cragg & Uhler's R^2
among others

Example

We can use “PseudoR2” function in DescTools package by choosing the named pseudo R^2 . For the membership data

```
> model=glm(y~x, family=binomial)
> model
```

```
Call:  glm(formula = y ~ x, family = binomial)
```

Coefficients:

(Intercept)	x
-4.8075	0.1251

Degrees of Freedom: 29 Total (i.e. Null); 28 Residual

Null Deviance: 41.46

Residual Deviance: 37.46 AIC: 41.46

```
> PseudoR2(model, which="McFadden")
McFadden
0.09626268
```

TRY: `PseudoR2(model, which="all")`

Example-Prostate Cancer

Treatment for prostate cancer changes depending on whether or not the lymphatic nodes surrounding the prostate are affected. In order to avoid an invasive investigation procedure (opening the abdominal cavity), a certain number of variables are considered as explanatory variables for the binary variable y : if $y=0$ the cancer has not reached the lymphatic system and if $y=1$ the cancer has reached the lymphatic system. A sample of 53 observations is available in <http://www.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/prostate.txt>. The data are posted in the Brightspace

The variables include

age: Age of the patient at the time of diagnosis

acid: Serum acid phosphate level

Xray: X-ray analysis, 0=negative, 1=positive

size: Tumor size

grade: State of the tumor as determined by biopsy, 0=medium, 1=serious

log.acid: Logarithm of the acidity level

```
> prostate=read.table("C:\\Users\\arya1\\Purdue\\STAT 40001\\prostate.txt", header=T)
> head(data,5)
  age acid Xray size grade Y   log.acid
1  66 0.48    0    0     0  0 -0.7339692
2  68 0.56    0    0     0  0 -0.5798185
3  66 0.50    0    0     0  0 -0.6931472
4  56 0.52    0    0     0  0 -0.6539265
5  58 0.50    0    0     0  0 -0.6931472
```

Example-Prostate Cancer

Fitting logistic regression with different regressor variable

```
> prostate=read.table("C:\\Users\\aryal\\Purdue\\STAT 40001\\prostate.txt", header=T)
> model1=glm(Y~log.acid,data=prostate,family=binomial)
> summary(model1)
Call:
glm(formula = Y ~ log.acid, family = binomial, data = prostate)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9802  -0.9095  -0.7266   1.1951   1.7302
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.404      0.509   0.794  0.4274
log.acid       2.245      1.040   2.159  0.0309 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 64.813  on 51  degrees of freedom
AIC: 68.813
```

Hence, the fitted model is

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = 0.404 + 2.245x$$

Equivalently,

$$\pi(x) = \frac{\exp(0.404 + 2.245x)}{1 + \exp(0.404 + 2.245x)}.$$

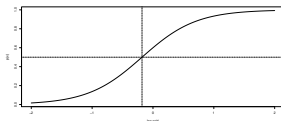
Note that the p-value for testing the hypothesis $\beta_1 = 0$ Vs $\beta_1 \neq 0$ is 0.0309.

Therefore the variable log.acid is retained in the model.

Example-Prostate Cancer

Fitting logistic regression with different regressor variable

```
> prostate=read.table("C:\\Users\\aryal\\Purdue\\STAT 40001\\prostate.txt", header=T)
> model1=glm(Y~log.acid,data=prostate,family=binomial)
> beta<-coef(model1)
> x<-seq(-2,2,0.01)
> y<- exp(beta[1]+beta[2]*x)/(1+exp(beta[1]+beta[2]*x))
> plot(x,y, type="l", xlab="log.acid", ylab="p(x)")
> abline(h=0.5, lty=2)
> xlim<--beta[1]/beta[2]
> abline(v=xlim, lty=2)
```



Observe that

$$\hat{p}(x) = \begin{cases} \leq 0.5 & \text{for } x \leq -0.18 \\ > 0.5 & \text{for } x > -0.18 \end{cases}$$

Multiple Logistic Regression

The simple logistic regression model

$$\pi_i = [1 + \exp(-\beta_0 - \beta_1 x_i)]^{-1}$$

can be extended to more than one predictor variable. The multiple logistic regression model for x_1, x_2, \dots, x_{p-1} predictor variables can be expressed as

$$E(y) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}$$

equivalently

$$E(y) = [1 + \exp(-\mathbf{x}'\boldsymbol{\beta})]^{-1}$$

where,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_{p-1} \end{pmatrix}, \mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{i,p-1} \end{pmatrix}$$

The fitted logistic response function and the fitted values can be expressed as

$$\hat{\pi} = \frac{\exp(\mathbf{x}'\mathbf{b})}{1 + \exp(\mathbf{x}'\mathbf{b})} = [1 + \exp(-\mathbf{x}'\mathbf{b})]^{-1}$$

Example-Market Data

A marketing research firm was investigating whether a family will purchase a new car during next year using logistic regression. A random sample of 33 suburban families was selected. Data on annual family income (x_1 in thousand dollars) and the current age of the oldest family automobile (x_2 , in years) was collected. A follow up interview conducted 12 months later was to determine whether the family actually purchased a new car $y = 1$ or didn't purchase a new car $y = 0$ during the year. The data is as below:

y	x_1	x_2	y	x_1	x_2
0.0	32.0	3.0	0.0	45.0	2.0
1.0	60.0	2.0	0.0	53.0	1.0
0.0	25.0	4.0	1.0	68.0	1.0
1.0	82.0	2.0	1.0	38.0	5.0
0.0	67.0	2.0	1.0	92.0	2.0
1.0	72.0	3.0	0.0	21.0	5.0
0.0	26.0	3.0	1.0	40.0	4.0
0.0	33.0	3.0	0.0	45.0	1.0
1.0	61.0	2.0	0.0	16.0	3.0
1.0	18.0	4.0	0.0	22.0	6.0
0.0	27.0	3.0	1.0	35.0	3.0
1.0	40.0	3.0	0.0	10.0	4.0
0.0	24.0	3.0	1.0	15.0	4.0
0.0	23.0	3.0	0.0	19.0	5.0
1.0	22.0	2.0	0.0	61.0	2.0
0.0	21.0	3.0	1.0	32.0	5.0
0.0	17.0	1.0			

Example- Market Data

Considering that a multiple logistic regression model fits the data we have the maximum likelihood estimates of the parameters β_0, β_1 and β_2 are

$$b_0 = \hat{\beta}_0 = -4.7393$$

$$b_1 = \hat{\beta}_1 = 0.0677$$

$$b_2 = \hat{\beta}_2 = 0.5986$$

Therefore the estimated regression model is

$$\hat{\pi} = [1 + \exp(4.7393 - 0.0677x_1 - 0.5986x_2)]^{-1}$$

```
> data=read.table("C://STAT4001//market.txt", header=T)
> y<-data$y
> x1<-data$x1
> x2<-data$x2
> x1
[1] 32 45 60 53 25 68 82 38 67 92 72 21 26 40 33 45 61 16 18 22 27 35 40 10 24 15 23 19 22 61 21 32 17
> x2
[1] 3 2 2 1 4 1 2 5 2 2 3 5 3 4 3 1 2 3 4 6 3 3 3 4 3 4 3 5 2 2 3 5 1
> y
[1] 0 0 1 0 0 1 1 1 0 1 1 0 0 1 0 0 1 0 1 0 0 1 1 0 0 1 1 0 0 1 0 0 1 0
> model=glm(y~x1+x2, family=binomial(logit))
> summary(model)
Call:
glm(formula = y ~ x1 + x2, family = binomial(logit))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73931    2.10195  -2.255   0.0242 *
x1           0.06773    0.02806   2.414   0.0158 *
x2           0.59863    0.39007   1.535   0.1249
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example-Prostate Cancer

Treatment for prostate cancer changes depending on whether or not the lymphatic nodes surrounding the prostate are affected. In order to avoid an invasive investigation procedure (opening the abdominal cavity), a certain number of variables are considered as explanatory variables for the binary variable y : if $y=0$ the cancer has not reached the lymphatic system and if $y=1$ the cancer has reached the lymphatic system. A sample of 53 observations is available in <http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/prostate.txt>

The variables include

age: Age of the patient at the time of diagnosis

acid: Serum acid phosphate level

Xray: X-ray analysis, 0=negative, 1=positive

size: Tumor size

grade: State of the tumor as determined by biopsy, 0=medium, 1=serious

log.acid: Logarithm of the acidity level

```
> data=read.table("http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/prostate.txt", header=T)
> head(data,5)
  age acid Xray size grade Y   log.acid
1  66 0.48    0    0     0 0 -0.7339692
2  68 0.56    0    0     0 0 -0.5798185
3  66 0.50    0    0     0 0 -0.6931472
4  56 0.52    0    0     0 0 -0.6539265
5  58 0.50    0    0     0 0 -0.6931472
```

Example-Prostate Cancer

Logistic regression with all explanatory variables can be expressed as

```
> prostate=read.table("http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/prostate.txt", header=T)
> model3=glm(Y~.,data=prostate,family=binomial)
> summary(model3)
```

Call:

```
glm(formula = Y ~ ., family = binomial, data = prostate)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0960	-0.6102	-0.2863	0.4834	2.2000

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	10.08672	7.83450	1.287	0.1979
age	-0.04289	0.06166	-0.696	0.4867
acid	-8.48006	7.63305	-1.111	0.2666
Xray	2.06673	0.85469	2.418	0.0156 *
size	1.38415	0.79546	1.740	0.0819 .
grade	0.85376	0.81247	1.051	0.2933
log.acid	9.60912	6.21652	1.546	0.1222

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 70.252 on 52 degrees of freedom
Residual deviance: 44.768 on 46 degrees of freedom
AIC: 58.768

Number of Fisher Scoring iterations: 5

Choosing the Model

The `anova` function can be used to compare two models using the deviance statistic to test the nullity of the coefficients of the bigger model. For example in the prostate cancer data

```
> prostate=read.table("http://www.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/prostate.txt", header=T)
> model1=glm(Y~log.acid,data=prostate,family=binomial)
> model3=glm(Y~,data=prostate,family=binomial)
> anova(model1,model3, test="Chisq")
Analysis of Deviance Table

Model 1: Y ~ log.acid
Model 2: Y ~ age + acid + Xray + size + grade + log.acid
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         51      64.813
2         46      44.768  5    20.044 0.001226 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the p-value of the test is 0.001226, the null hypothesis is rejected. Thus the model 1 is rejected and at least one of the supplementary variable in model 3 (full model) is considered to be relevant. Like in the multiple linear regression model we can use `step` function to choose the best model.

```
> choose<-step(model, direction="backward")
```

Making a Prediction

Once a best logistic regression has been selected it can be used for prediction using the function `predict`.

```
> prostate=read.table("http://math.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/prostate.txt", header=T)

> model=glm(Y~., data=prostate,family="binomial")
> bestmodel=step(model, direction="backward")
> xnew<-matrix(c(61,0.60,1,0,1,-0.51),nrow=1)
> colnames(xnew)<-c("age","acid","Xray","size","grade","log.acid")
> xnew=as.data.frame(xnew)
> predict(bestmodel,xnew, type="response")
      1
0.4835626
```

Since the predicted probability is less than 0.5 we therefore predict $\hat{Y} = 0$, that is to say the cancer has not reached the lymphatic system.

Poisson Regression

Data is often collected in counts (e.g. the number of heads in 12 flips of a coin or the number of car thefts in a city during a year). Many discrete response variables have counts as possible outcomes. Binomial counts are the number of successes in a fixed number of trials, n . Poisson counts are the number of occurrences of some event in a certain interval of time (or space). While Binomial counts only take values between 0 and n , Poisson counts have no upper bound. We now consider a nonlinear regression model where the response outcomes are discrete counts that follow a Poisson distribution. Poisson regression provides a model that describes how the mean response, changes as a function of one or more explanatory variables. The logarithm of the mean can be modeled as a linear function of observed covariates. The result is a generalized linear model with Poisson response and link log. Poisson regression is a natural choice when the response variable is a small integer. The explanatory variables model the mean of the response variable.

Poisson Regression

We have n observations of a response random variable y , and $p > 1$ fixed explanatory variables, x_1, x_2, \dots, x_p . We assume that for observation $i = 1, 2, \dots, n$,

$$y_i \sim \text{Poi}(\mu_i)$$

where

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$$

Thus, our generalized linear model has a Poisson random component, a linear systematic component $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ and a log link

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

A consequence of using the log link function with a linear predictor is that the explanatory variables affect the response mean multiplicatively. Consider a Poisson regression model with one explanatory variable: $\mu(x) = \exp(\beta_0 + \beta_1 x)$, where our notation emphasizes that μ changes as a function of x . When we increase the explanatory variable by c units, the result is $\mu(x + c) = \exp(\beta_0 + \beta_1(x + c)) = \mu(x) \exp(c\beta_1)$. Thus, the ratio of the means at $x + c$ and at x is

$$\frac{\mu(x + c)}{\mu(x)} = \frac{\exp(\beta_0 + \beta_1(x + c))}{\exp(\beta_0 + \beta_1 x)} = \exp(c\beta_1).$$

Example

School administrators study the attendance behavior of high school juniors at two schools. Predictors of the number of days of absence include gender of the student and standardized test scores in math and language arts.

VARIABLES:

Daysabs: days absent

Math: standardized test scores

Langarts: standardized test scores

Male: gender (1=male, 0=female)

```
absence=read.csv("H://STAT 40001//absence.csv", header=T)
head(absence)
tail(absence)
attach(absence)
model=glm(daysabs~math+langarts+male, family=poisson)
```