APPLIED STATISTICS

Log-Linear Regression for Poisson Counts

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics
The Australian National University

Last Updated: Wed Oct 11 20:54:25 2017

Overview

- Motivating Examples
- Log-Linear Regression for Poisson Counts
- Model Diagnostics
    1. Log response verus explanatory variable plot.
    2. Pearson residual plot.
    3. Deviance goodness-of-fit test.

References

1. **F.L. Ramsey and D.W. Schafer** (2012)
   Chapter 22 of *The Statistical Sleuth*

2. **H. Wang** (2008)
   Chapter 6 of *Applied Business Statistical Analysis*

3. The slides are made by **R Markdown**.
   http://rmarkdown.rstudio.com

Example: Elephant Data

(Taken from 22.1.1 of "The Statistical Sleuth".)

The object "case2201" in the R package "Sleuth3" contains information on the age of 41 male elephants and for each male the number of successful matings over the past eight-year period. It was of interest to study the relationship between mating success and age.

Data

```
setwd('~/Desktop/Research/AppliedStat2017/L13')
library(Sleuth3)
elephant=case2201;elephant
```

```
##    Age Matings
## 1   27       0
## 2   28       1
## 3   28       1
## 4   28       1
## 5   28       3
## 6   29       0
## 7   29       0
## 8   29       0
## 9   29       2
## 10  29       2
## 11  29       2
## 12  30       1
## 13  32       2
## 14  33       4
## 15  33       3
## 16  33       3
## 17  33       3
## 18  33       2
## 19  34       1
## 20  34       1
## 21  34       2
## 22  34       3
## 23  36       5
## 24  36       6
## 25  37       1
## 26  37       1
## 27  37       6
## 28  38       2
## 29  39       1
```

Some Terminologies to Interpret Data

"Trial": A male elephant at some specific age has a lot of trials of mating in this eight-year period.

"Success": one trial is called "success" if in this trial, the elephant has a successful mating.

Probability of "Success": the probability that one trial is a "success", **or equivalently**, the probability that the elephant has a successful mating in the trial.

Some Terminologies to Interpret Data (Con'd)

A typical row for the data:

Explanatory	Number of Successes	Total Number
X	Z	M

Z is the number of "successes" given some specific X.

M is the number of "trials" given some specific X.

Given X, the probability of one "success" in a trial is denoted by $\pi \in [0, 1]$.

In the previous lecture, the count Z is modelled by Binomial(M, π) distribution. Specifically,

$$\mathrm{P}(Z = z) = \binom{M}{z} \pi^z (1-\pi)^{M-z}, \text{ for } z = 0, \cdots, M.$$

However, in this case, the number of "trials" of mating for an elephat in this eight-year period, namely M, is unknown, but usually M is large.

Since we do not have the information of M, how to model the count Z?

Poisson Approximation

"Rare Events": the probability that an elephant has a successful mating in the trial, namely π is usually very small. For elephants, we call a successful mating a rare event.

When M is very large and π is very small, let $\mu = M\pi$. Then we have the following approximation

$$\mathrm{P}(Z = z) = \binom{M}{z} \pi^z (1-\pi)^{M-z} \approx \frac{e^{-\mu}\mu^z}{z(z-1)(z-2)\cdots 1}.$$

By making use of this approximation, M does not appear in the expression.

In this case, we no longer need to use the information of M to model the count Z, and hence we can model the elephant data now.

Poisson Count

In a lot of real data, we only have **the number of "successes"** instead of the number of "trials". Specifically in this example, we have

1. at a specific age of an elephant, we know **the number of "successes"** (the number of successful matings), which is denoted by Z;

2. at a specific age of an elephant, we do not know **the total number of "trials"** (the total number of matings for this elephant), denoted by M.

When the total number of "trials" is very large and the probability of "success" is very small (a successful mating for an elephant is a rare event), then we can call Z a Poisson count.

In this case, the total number of "trials" is usually not recorded in the data.

Log-linear regression for Poisson counts deals with this particular case, and is used to model the Poisson count response Z.

Another Example

Suppose we are interested in modelling the number of car accidents at the intersection of Barry Drive and Kingsley Street as the response variable.

1. We know **the number of "successes"** (the number of car accidents), which is denoted by Z;

2. We do not know **the total number of "trials"** (the total number of cars that pass through this intersection), denoted by M.

Obviously the total number of "trials" is very large and the probability of "success" is very small (a car accident is a rare event), then we can call Z a Poisson count.

In this case, no one will record the total number of cars that pass through this intersection. However, the number of car accidents is likely to be recorded by the police.

The number of rare events that occur over a given time, space or volume is usually a Poisson count.

Examples of Poisson Counts

1. The number of hospital admissions that occur over a given period of time;
2. The number of kangaroos found over a particular area of land;
3. In the statistical literatures, the first application of the Poisson approximation was the description of the number of deaths by horse kicking.

Count Response Variables

Example 4: Y takes value of "once", "twice", "three times" $\cdots$

We can set 1="once", 2="twice", 3="three times" $\cdots$ Obviously,

$$\text{“once”} < \text{“twice”} < \text{“three times”} < \cdots,$$

and $2 - 1 = 3 - 2$, which mean there is a numerical meaning for 1="once", 2="twice", 3="three times" $\cdots$

The above is the main difference between the count response variables and the ordinal response variables in Lecture Notes 11.

Overview of This Course

	Continuous X + Categorical X
Continuous Y	MLR + Indicator Variables
Two-Category Y	Binary Logistic Regression + Indicator Variables
Multicategory Y - Nominal	Nominal Response Regression + Indicator Variables
Multicategory Y - Ordinal	Ordinal Response Regression + Indicator Variables
Binomial Count Z	Binomial Logistic Regression + Indicator Variables
Poisson Count Z	**Poisson Regression + Indicator Variables**

Poisson Log-Linear Regression Model Assumptions

1. **Poisson distribution**: There is a Poisson distributed (sub)population of responses Z for given values of the explanatory variables $(X_1 = x_1, \cdots, X_k = x_k)$. That means if we let $X = (X_1, \cdots, X_k)$, the probability that $Z = z$ given X is

$$\mathrm{P}(Z = z) = \frac{e^{-\mu}\mu^z}{z(z-1)(z-2)\cdots 1}, \text{ where } z = 0, 1, 2, \cdots.$$

Based on the properties of the Poisson distribution, the mean of response Z is given by

$$\mu\{Z|X\} = \mu.$$

2. **Generalised Linearity**: The transformation of the mean of response μ falls on a linear function of the explanatory variables

$$g(\mu\{Z|X\}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \text{ for } X = (X_1, \cdots, X_k),$$

where $g(u) = \log(u)$, which is the log link function.

Poisson Log-Linear Regression Model Assumptions (Con'd)

Remark: $\mu\{Z|X\} = e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}$.

DISPLAY 22.5 A representation of a log-linear model in which the distribution of Y (as a function of X) is Poisson with mean μ and $\log(\mu) = -1.7 + 0.20X$; the histograms are the Poisson distributions of Y at three values of X

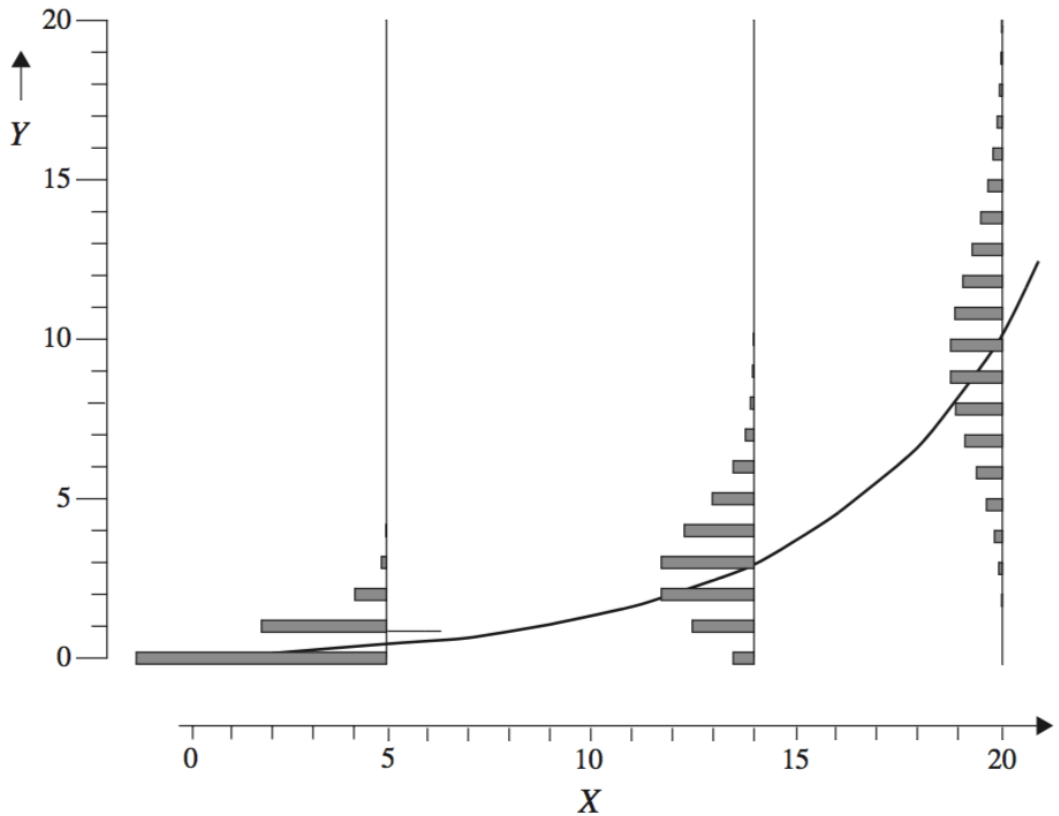


Figure from Chapter 22 of "The Statistical Sleuth"

Poisson Log-Linear Regression Model Assumptions (Con'd)

3. **Independence**: Observations

$$(X_{1,1}, \cdots X_{k,1}, Z_1),$$

$$\vdots$$

$$(X_{1,m}, \cdots X_{k,m}, Z_m),$$

are independent, where m is the sample size.

Poisson Log-Linear Regression Model and Interpretation

$$\mu\{Z|X_1 = x_1 + 1, X_2 = x_2, \cdots, X_k = x_k\} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k} e^{\beta_1}, \text{ and}$$

$$\mu\{Z|X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k\} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}, \text{ and}$$

With the other variables held constant, if X_1 is increased by 1 unit, the mean of response Z will change by a multiplicative factor of e^{β_1}.

If β_1 is small, $e^{\beta_1} \approx 1 + \beta_1$. For instance, $e^{0.05} \approx 1 + 0.0513$ and $e^{0.5} \approx 1 + 0.6482$. Hence

$$\mu\{Z|X_1 = x_1 + 1, X_2, \cdots, X_k\} \approx (1 + \beta_1)\mu\{Z|X_1 = x_1, X_2, \cdots, X_k\}.$$

So if β_1 is small, we can interpret β_1 as the percentage increase in the mean of response Z for a one unit increase in X_1.

Estimation, z-Test and CI

The likelihood function for the observations and the MLE can be obtained.
The inferential tools for generalised linear model can be used.

```
attach(elephant)
elephant.pois=glm(Matings~Age,family=poisson(link=log))
summary(elephant.pois)
```

```
##
## Call:
## glm(formula = Matings ~ Age, family = poisson(link = log))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80798  -0.86137  -0.08629   0.60087   2.17777
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.58201    0.54462  -2.905  0.00368 **
## Age          0.06869    0.01375   4.997 5.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 75.372  on 40  degrees of freedom
## Residual deviance: 51.012  on 39  degrees of freedom
## AIC: 156.46
##
## Number of Fisher Scoring iterations: 5
```

Drop-in-Deviance χ^2-Test

$$\mathrm{H}_0 : \beta_1 = 0 \leftrightarrow \mathrm{H}_a : \text{ otherwise.}$$

```
#drop in deviance test
#reduced model
elephant.poisr=glm(Matings~1,family=poisson(link=log))
anova(elephant.poisr,elephant.pois,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Matings ~ 1
## Model 2: Matings ~ Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        40     75.372
## 2        39     51.012  1    24.36 7.991e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fitted Values of Response

A typical row for the data:

Explanatory	Poisson Count
X	Z

The Poisson count Z is modelled by Poisson distribution. Based on the properties of the Poisson distribution, the mean of response Z is given by

$$\mu\{Z|X\} = \mu, \text{ where } \mu \text{ is the parameter in the Poisson distribution.}$$

Using MLE $\hat{\beta}_0, \cdots, \hat{\beta}_k$, the fitted values of response Z are given by:

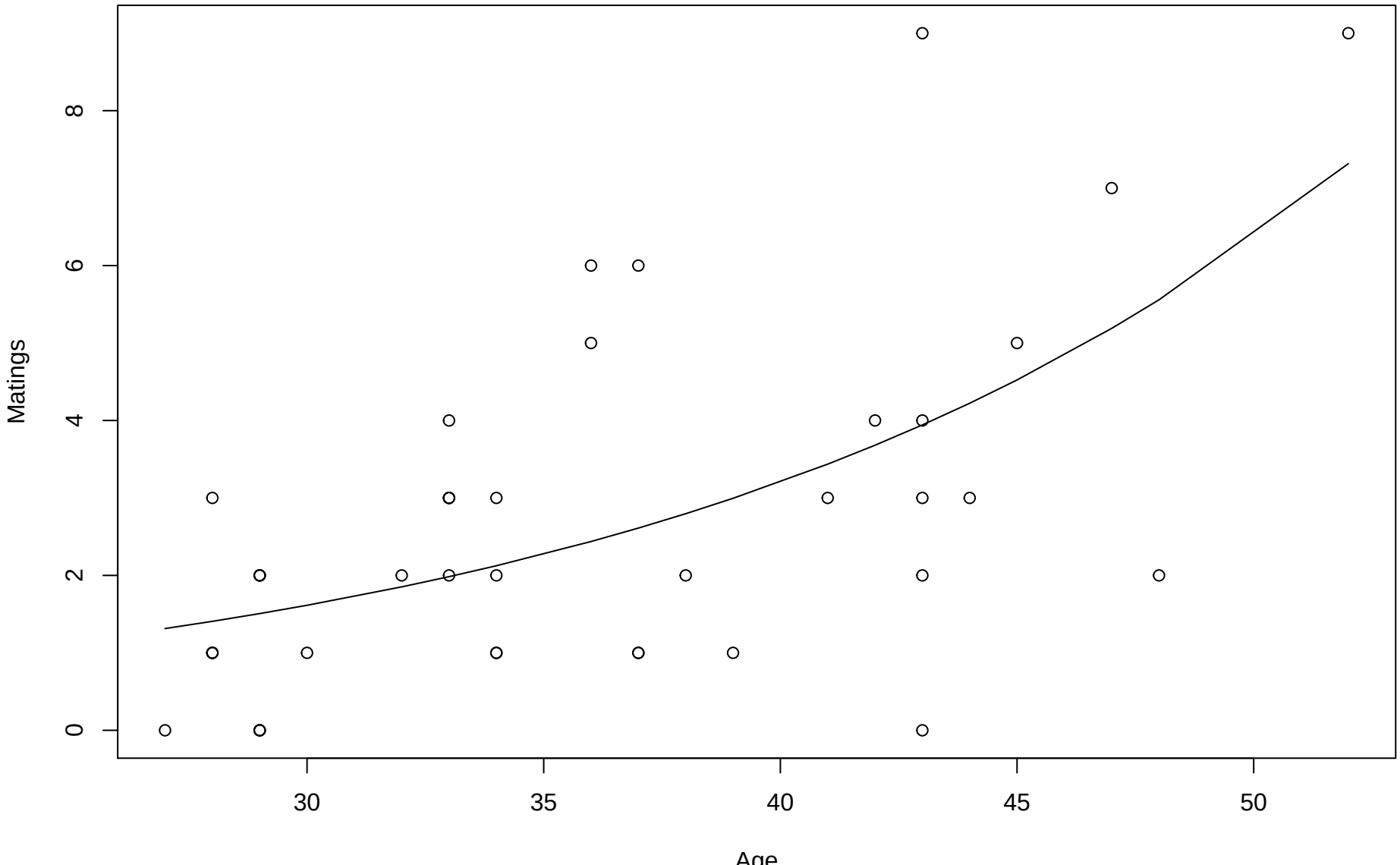
$$\hat{Z} = \hat{\mu}\{Z|X\} = e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k}.$$

When we talk about "fitted", X is usually from the training dataset (see Lecture Notes 8).

When X_{new} is from the new dataset or the test dataset, we actually talk about prediction.

Example: Elephant Data (Con'd)

```
plot(Age,Matings)
lines(Age,elephant.pois$fitted.values)
```



Example: Elephant Data (Con'd)

By using this example, we might be interested in predicting the number of successful matings if an elephant is 31. The prediction is:

```
Xnew=data.frame(Age=31)
predict(elephant.pois,Xnew,type='response')
```

```
##        1
## 1.728872
```

However, it is not an integer. So we consider

```
round(predict(elephant.pois,Xnew,type='response'),0)
```

```
## 1
## 2
```

roughly as the prediction of response Z.

1. Log Response verus Explanatory Variable Plot

Recall that in the Poisson log-linear regression model assumptions, the **log** transformation of the mean of response falls on a linear function of the explanatory variables

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \text{ for } X = (X_1, \cdots, X_k),$$

where $g(u) = \log(u)$ is the **log** link function. Here the mean of response depends on unknown parameters and hence is also unknown.

For a typical row of the data:

Explanatory	Poisson Count
X	Z

The response is Z given X, which can be used to approximate μ, since the total number of "trials" is unobservable but is usually very large for the Poisson count.

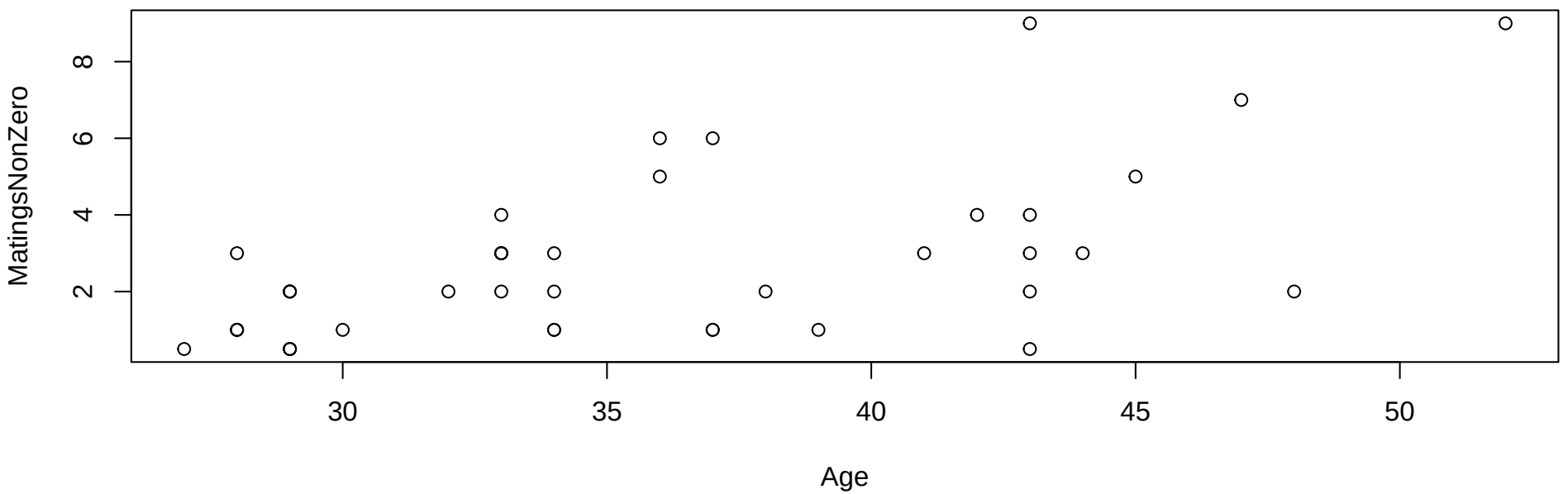
Hence it can be useful to plot the $\log(Z)$ versus explanatory variable (e.g., X_1).

1. Log Response verus Explanatory Variable Plot (Con'd)

The plot should show a straight line. Otherwise the model assumption is violated. See Lecture Notes 3 for solutions to this problem.

Note: $\log(Z)$ is undefined for $Z = 0$. We need to add a small quantity to values of 0.

```
MatingsNonZero=ifelse(Matings==0, 0.5, Matings)
plot(Age,MatingsNonZero)
```



This plot suggests that $\log(\mu)$ might be reasonably well approximated by a linear function of age.

Review: Studentized Residuals for Continuous Response Y

Recall residual: $\text{res}_i = \hat{\mathcal{E}}_i = Y_i - \hat{Y}_i$ for observation i, where $\hat{Y}_i$ is the fitted value of response.

One can obtain

$$\text{SD}(\text{res}_i) = \sigma\sqrt{1-h_i}, \text{ and}$$

$$\text{SE}(\text{res}_i) = \hat{\sigma}\sqrt{1-h_i}, \text{ where } \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}\text{res}_i^2}{n-k-1}}.$$

A studentized residual is a residual divided by its standard error, namely

$$\text{studres}_i = \frac{\text{res}_i}{\text{SE}(\text{res}_i)}$$

Using studentized residuals allows the residuals to be viewed on the same scale.

These studentized residuals are roughly standard normally distributed $[N(0,1)]$, if the observation is from the MLR model with the all the assumptions satisfied.

Pearson Residuals for Poisson Count Z

The residual for Poisson count Z is defined as: $\text{res}_i = Z_i - \hat{Z}_i$ for observation i, where $\hat{Z}_i$ is the fitted value of response.

One can obtain

$$\text{SD}(\text{res}_i) = \sqrt{\mu_i}, \text{ and}$$

$$\text{SE}(\text{res}_i) = \sqrt{\hat{Z}_i}, \text{ where } \hat{Z}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i}}.$$

A Pearson residual is a residual divided by its standard error, namely

$$\text{Peares}_i = \frac{\text{res}_i}{\text{SE}(\text{res}_i)}$$

Using Pearson residuals allows the residuals to be viewed on the same scale.

These Pearson residuals are roughly standard normally distributed $[N(0,1)]$, if the observation is from the Poisson log-linear model with the all the assumptions satisfied.

2. Pearson Residual Plot

Due to the nature of the $N(0,1)$ distribution, most values of the $N(0,1)$ distribution concentrate in the middle around 0.

Hence, if Peares_i falls into the two tails of the $N(0,1)$ distribution, namely $|\text{Peares}_i|$ is too large, then it is unlikely that observation i is from the Binomial logistic model with the all the assumptions satisfied.

$\text{Peares}_i > 1.96$ (97.5 % quantile of $N(0,1)$), or
$\text{Peares}_i <$- 1.96 (2.5 % quantile of $N(0,1)$).

$\Rightarrow$

Peares_i falls into the two tails of the $N(0,1)$ distribution.

$\Rightarrow$

$|\text{Peares}_i|$ is too large.

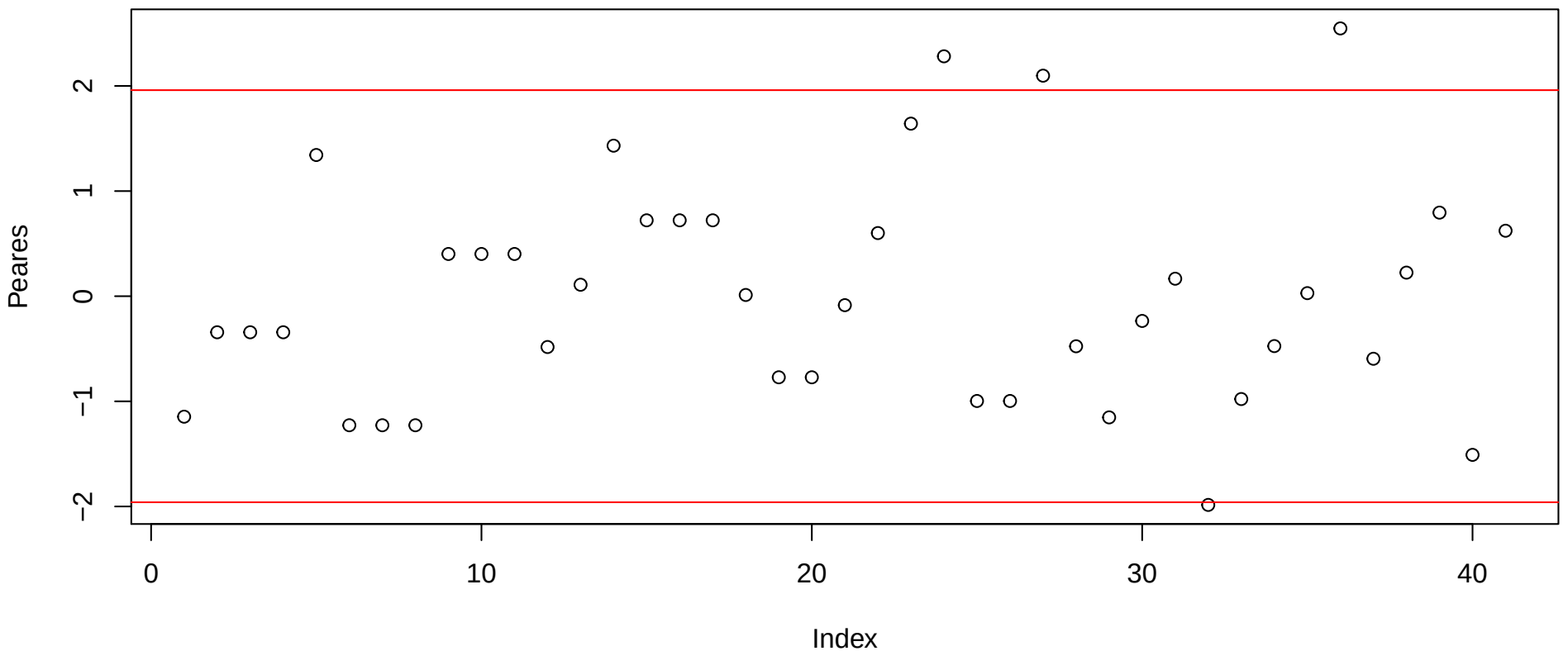
$\Rightarrow$

Observation i is **an outlier**.

Sometimes we use 2 instead of 1.96 for simplicity.

2. Pearson Residual Plot (Con'd)

```
Peares=residuals(elephant.pois,type="pearson")
plot(Peares)
abline(h=1.96,col='red')
abline(h=-1.96,col='red')
```



The Pearson residuals indicate some outliers.

3. Deviance Goodness-of-Fit Test

$$\mathrm{H}_0 : \text{Poisson model } g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \text{ is appropriate} \leftrightarrow$$

$$\mathrm{H}_a : \text{The Poisson log-linear model is not appropriate.}$$

The test statistic is

$$TS = \text{deviance of model } g(\pi) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

This test statistic should be compared to χ^2_{m-k-1} distribution approximately, where k is the number of explanatory variables.

The p-value is

$$p\text{-value} = \mathrm{P}(S > TS), \text{ where } S \sim \chi^2_{m-k-1}.$$

If p-value $< \alpha \Rightarrow$ reject H_0; p-value $\geq \alpha \Rightarrow$ not reject H_0.

3. Deviance Goodness-of-Fit Test (Con'd)

The output from the summary() output can be used to perform the goodness-of-fit test.

```
summaryfit=(summary(elephant.pois))
summaryfit
```

```
##
## Call:
## glm(formula = Matings ~ Age, family = poisson(link = log))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80798  -0.86137  -0.08629   0.60087   2.17777
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.58201    0.54462  -2.905  0.00368 **
## Age          0.06869    0.01375   4.997 5.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 75.372  on 40  degrees of freedom
## Residual deviance: 51.012  on 39  degrees of freedom
## AIC: 156.46
##
## Number of Fisher Scoring iterations: 5
```

3. Deviance Goodness-of-Fit Test (Con'd)

The test statistic and the numer of degress of freedom are

```
summaryfit$deviance
```

```
## [1] 51.01163
```

```
summaryfit$df.residual
```

```
## [1] 39
```

The p-value for the test is given by

```
1 - pchisq(summaryfit$deviance, summaryfit$df.residual)
```

```
## [1] 0.09426231
```

```
detach(elephant)
```

We cannot reject the null. There is no evidence that the model is inappropriate.