Generalized Linear Models for count data:

Number of movies you watch in the last month

Many discrete response variables have counts as possible outcomes. Counts also occur in summarizing categorical variables with contingency tables.

The simplest GLMs for count data assume a Poisson distribution for the random component. Like counts, Poisson variables can take any non-negative integer value.

① Poisson Regression:

The poisson distribution has a positive mean. GLMs for the poisson mean can use the identity link, but it is more common to model the log of the mean. Like the linear predictor $\alpha+\beta x$, the log of the mean can take any real-number value. A poisson log linear model is a GLM that assumes a poisson distribution for Y and uses the log link function.

For a single explanatory variable x, the poisson log linear model has form

$$\log \mu = \alpha + \beta x$$

The mean satisfies the exponential relationship

$$\mu = \exp(\alpha + \beta x)$$

$$\mu = e^{\alpha+\beta x}$$

$$\mu = e^{\alpha} e^{\beta x}$$

$$\mu = e^{\alpha} (e^{\beta})^{x}$$

A one-unit increase in x has a multiplicative impact of e^{β} on μ.

If $\beta = 0$, then $e^{\beta} = e^0 = 1$ (multiplicative factor) then the mean of Y does not change as x changes.

If $\beta > 0$, then $e^{\beta} > 1$
then the mean of Y increase as x increase

If $\beta < 0$ then $e^{\beta} < 1$
then the mean of Y decrease as x increase.

⇒ **Goodness of Fit:**

Fit means for each observation i, would like $Y_i - \hat{Y}_i = 0$

Hypothesis for the goodness of fit tests:

H_0: model adequately fits the data

H_1: model not adequately fit the data

We generally use **Hosmer-Lemeshow test** for checking goodness of fit. The test statistic of Hosmer Lemeshow test for logistic regression is:

$$G^2_{HL} = \sum \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)}$$

Data is first organized by ordering the predicted probabilities and forming the num of groups

follows the chi-square distribution (df = (number of groups) − 2)

where

O_j = number of observation cases in the jth group

E_j = number of expected cases in the jth group

n_j = number of observations in the jth group

Another approach is used for goodness of fit that is known as **Deviance- based approach**.

This approach regards the data as representing the fit of the most complex model possible, it is called saturated model (which has a seperate parameter of each observation) ⇒ It provides a perfect fit to the data.

Let L_M denote the maximum log-likelihood value for a model M of interest. Let L_S denote the maximum log-likelihood value for the most complex model pasible. Then the deviance is the likelihood-ratio statistic for comparing model M to the saturated model.

$$\text{Deviance} = -2(L_M - L_S)$$

It is a test statistic for the hypothesis that all parameters that are in the saturated model but not in model M equal zero. Software provides the deviance, so it is not necessary to calculate L_M or L_S.

For some GLMs, the deviance has approximately a chi-square distribution with degree of freedom, number of observation minus the number of model parameters. Large test statistics and small p-values provide stoong evidence of model lack of fit.

When the predictors are categorical, the data are summarized by counts in a contigency table. In this case the deviance statistic for testing the fit of model (M).

$$\underbrace{\text{Deviance}(M)}_{\text{in book notation } G^2(M)} = 2\sum_{i=1}^{n} O_i \left(\log\left(\frac{O_i}{E_i}\right)\right)$$