Regression Analysis I

Generalized Linear Models

Hossein Moradi Rekabdarkolaee

South Dakota State University

email: *hossein.moradirekabdarkolaee@sdstate.edu*.

Office: Arch, Math, and Engineering Building, *Room: 254*.

Multiple Linear Regression

- A multiple linear regression model can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Generalized Linear Model (General Concept)

- Consider a general regression model:

$$y_i = f(\mathbf{x}_i, \theta) + \epsilon_i$$

where $E(\epsilon_i) = 0$ and the variance of error is $var(\epsilon_i) = g(\mathbf{x}_i, \theta)$.

- Note that both f and g involves the regressors.
- In order to use MLE or LS, we must assume g to be fixed.
- If the model is linear, then we use WLS or GLS to minimize

$$\sum_{i=1}^{n} \frac{[y_i - f(\mathbf{x}_i, \theta)]^2}{g(\mathbf{x}_i, \theta)}$$

- This is a conditional optimization where we update $g(\mathbf{x}_i, \theta)$ with new θ's.

Generalized Linear Model (General Concept)

- If the model is nonlinear, then we use Iteratively Reweighted Nonlinear Least Squares (IRWLS).
- We choose starting values called θ_0 , form weights, and compute residuals $\mathbf{e} = \mathbf{y} - f(\mathbf{x}_i, \theta_0)$.
- IRWLS computes estimates by

$$\hat{\gamma} = (\mathbf{W}^T\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{V}^{-1}\mathbf{e}$$

  where $\mathbf{V} = diag(g(\mathbf{x}_1, \theta), \ldots, g(\mathbf{x}_n, \theta))$ is considered to be known.
- Set $\hat{\theta}_1 = \theta_0 + \hat{\gamma}$.
- Repeat until converge.

Generalized Linear Model (General Concept)

- An example of the equivalency between MLE and IRWLS is logisic regression with grouped structure.
- For any distribution, IRWLS and MLE both solve

$$\sum_{i=1}^{n}(y_i - \mu_i)\mathbf{x}_i = 0$$

or equivalently

$$\sum_{i=1}^{n} e_i\mathbf{x}_i = 0$$

or in matrix form $\mathbf{X}^T\mathbf{e} = 0$ where $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$.

- Consider linear model with $\boldsymbol{\mu} = \mathbf{X}\beta$.

Generalized Linear Model (Exponential Family)

- A density that can be written in the form of

$$f(y_i) = \exp\left\{r(\phi)[y_i\theta_i - g(\theta_i)] + h(y_i, \phi)\right\}$$

  where ϕ is a scale or nuisance parameter,

  θ_i is a natural location parameter which in some cases equals to the mean μ_i,

  and $g(\theta_i)$ is related to the mean and variance.

- is said to belong to **exponential family**.

Generalized Linear Model (Exponential Family)

- Normal Distribution.
- Poisson Distribution.
- Binomial Distribution.
- Gamma/Exponential Distribution.

Generalized Linear Model (Exponential Family)

- Let $L_i = \log f(y_i)$ denotes the log-likelihood and $L = \sum_{i=1}^{n} L_i$.

- Since
$$L_i = r(\phi)[y_i\theta_i - g(\theta_i)] + h(y_i, \phi),$$

- Then the first and second derivatives are
$$\frac{\partial L_i}{\partial \theta_i} = r(\phi)[y_i - g'(\theta_i)], \quad \text{and} \quad , \frac{\partial^2 L_i}{\partial \theta_i^2} = -r(\phi)g''(\theta_i).$$

- Apply the general likelihood results
$$E\left(\frac{\partial L}{\partial \theta}\right) = 0, \quad \text{and} \quad , -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = E\left(\frac{\partial L}{\partial \theta}\right)^2,$$

  which hold under regularity conditions satisfied by the exponential dispersion family.

Generalized Linear Model (Principles and Restrictions)

- A set of independent observation $y_1, \ldots, y_n$ with mean $E(y_i) = \mu_i$.
- A regression structure: $\mathbf{X}$ matrix.
- A density function that belongs to exponential family.
- θ_i varies from data point to data point and links the mean to data.
- The link function is $s(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$.
- The link function is not necessarily linear, but it should produce a function that mean is a monotonic and differentiable function of $\mathbf{x}_i^T \boldsymbol{\beta}$.
- $var(y_i)$ is not necessarily homogeneous but varies with regressors only through the mean function.

Generalized Linear Model (Principles and Restrictions)

- The link function links the mean to the regressors and determines the model.
- Usually $s(\mu_i) = \theta_i$, which is called cononical link.
- Thus, we let $s(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ and solve for μ_i.
- We build the model around cononical link.
- $\mu_i = \frac{\partial g(\theta_i)}{\partial \theta_i}$.
- $\sigma_i^2 = \frac{\left(\frac{\partial^2 g(\theta_i)}{\partial \theta_i^2}\right)}{r(\phi)}$
- The only homogeneous variance case is normal distribution.

Generalized Linear Model (Link function)

- Normal Distribution.
- Poisson Distribution.
- Binomial Distribution.
- Gamma/Exponential Distribution.

Generalized Linear Model (Link function)

- The Cononical link is just a special case where $s(\mu_i) = \theta_i$.
- In general, we can use other link fundtions.
- For instance, for Poisson distribution, instead of log link, we can use the square root link and the model become:

$$y_i = (\mathbf{x}_i^T \beta)^2 + \epsilon_i.$$

Table: Some Commonly Used Link Functions.

Normal	Poisson	Binomial	Exponential/Gamma
Identity	Log	Logit	Reciprocal
Log	Square root		Log
Squared root	Identity		Identity
Exponential			
Reciprocal			

Generalized Linear Model

- For n independent observations, the log likelihood is

$$\begin{aligned}\sum_{i=1}^{n} L_i &= \sum_{i=1}^{n} \{r(\phi)[y_i\theta_i - g(\theta_i)] + h(y_i,\phi)\} \\ &= \sum_{i=1}^{n} \left\{r(\phi)[y_i\mathbf{x}_i^T\boldsymbol{\beta} - g(\theta_i)] + h(y_i,\phi)\right\}.\end{aligned}$$

- The part of the log likelihood involving both the data and the model parameters is

$$\sum_{i=1}^{n} y_i \sum_{j=1}^{p} x_{ij}\beta_j = \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} y_i x_{ij}.$$

- Thus the sufficient statistics for $\{\beta_j\}_{j=1}^{p}$ is $\{\sum_{i=1}^{n} y_i x_{ij}; j = 1, 2, \ldots, p\}$.

Generalized Linear Model

- For GLM with link function $g(\theta_i)$, the derivitives of likelihood are

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial L_i}{\partial \beta_j} = 0, \quad \text{for all j.}$$

  using the chain rule:

$$\begin{aligned}
\frac{\partial L_i}{\partial \beta_j} &= \frac{\partial L(\theta_i)}{\partial g(\theta_i)} \frac{\partial g(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mathbf{x}_i^T \beta_j} \frac{\partial \mathbf{x}_i^T \beta_j}{\partial \beta_j} \\
&= r(\phi)[y_i - \mu_i] \frac{1}{r(\phi) var(y_i)} \frac{\partial \theta_i}{\partial \mathbf{x}_i^T \beta_j} x_{ij} \\
&= \frac{[y_i - \mu_i] x_{ij}}{var(y_i)} \frac{\partial \theta_i}{\partial \mathbf{x}_i^T \beta_j}
\end{aligned}$$

- Summing over the n observations yields the likelihood equations.

- Recall that both IRWLS ad MLE solve $\mathbf{X}^T\mathbf{e} = 0$.
- This holds true only if we use the canonical link.
- Suppose we use a noncanonical link.
- Then the equations that should be solve become: $\mathbf{X}^T\Delta\mathbf{e} = 0$.

where $\Delta = diag(\delta_1, \ldots, \delta_n)$, and $\delta_i = \frac{\partial\theta_i}{\partial\mathbf{x}_i^T\boldsymbol{\beta}}$.

Generalized Linear Model

- Suppose having canonical link functio, the information matrix is

$$\mathbf{I} = E\left(\frac{\partial^2 lnL}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)$$

- We showed that

$$\frac{\partial lnL}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n} r(\phi)(y_i - \mu_i)\mathbf{x}_i = r(\phi)\mathbf{X}^T(\mathbf{y} - \mu).$$

- Then

$$\mathbf{I} = E\left[r(\phi)\mathbf{X}^T(\mathbf{y} - \mu)r(\phi)(\mathbf{y} - \mu)^T\mathbf{X}\right] = r(\phi)^2\mathbf{X}^T\mathbf{D}\mathbf{X}$$

where **D** denote the diagonal matrix of variances of the observations.

Generalized Linear Model

- Thus the asymptotic variance-covariance matrix is

$$var(\hat{\beta}) = \mathbf{I}^{-1} = \frac{1}{r(\phi)^2}\left(\mathbf{X}^T\mathbf{D}\mathbf{X}\right)^{-1}.$$

- Using noncanonical link the asymptotic variance-covariance matrix is

$$var(\hat{\beta}) = \mathbf{I}^{-1} = \frac{1}{r(\phi)^2}\left(\mathbf{X}^T\Delta\mathbf{D}\Delta\mathbf{X}\right)^{-1}.$$

- This is the weighted regression.

- **Example**:

  Exponential distribution, with Identity link

  Gamma Distribution.

Generalized Linear Model

- Inference about GLMs has three standard ways to use the likelihood function.
- This is for a generic scalar model parameter β.
- We focus on test of $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$
  - Likelihood-Ratio Tests.
  - Wald Tests.
  - Score Tests.

- **Likelihood-Ratio Tests**:

- Recall that the test is

$$\lambda = \frac{L(H_0)}{L(H_1)}$$

and $-2 \log \lambda \sim \chi^2_{df}$ where $df = df(numerator) - df(denumerator)$ as $n \to \infty$.

Generalized Linear Model

- **Wald Test**:
- Standard errors obtained from the inverse of the information matrix depend on the unknown parameter values.
- When we substitute the unrestricted ML estimates (i.e., not assuming the null hypothesis), we obtain an estimated standard error (SE) of $\hat{\beta}$.
- For $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, the test statistic using this non-null estimated standard error,

$$z = \frac{\hat{\beta} - \beta_0}{SE},$$

- is called a Wald statistic which has an approximate standard normal distribution when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

- **Score Tests**:

- The score test, referred to in some literature as the Lagrange multiplier test, uses the slope (i.e., the score function) and expected curvature of the log-likelihood function, evaluated at the null value β_0. The chi-squared form of the score statistic is

$$\frac{[\partial L(\beta)/\partial \beta_0]^2}{-E[\partial^2 L(\beta)/\partial \beta_0^2]}$$

  where the notation reflects derivatives with respect to β that are evaluated at β_0.

- In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood and the inverse information matrix, both evaluated at the H_0 estimates.

Generalized Linear Model

- Consider a binomial parameter p and testing $H_0 : p = p_0$.
- With sample proportion $\hat{p} = y$ for n observations, it can be shown that the chi-squared forms of the test statistics are

$$\begin{aligned} LR: \quad -2ln(\lambda) &= -2ln\left[\frac{p_0^{ny}(1-p_0)^{n(1-y)}}{y^{ny}(1-y)^{n(1-y)}}\right]; \\ \text{Wald}: \quad z^2 &= \frac{(y-p_0)^2}{y(1-y)/n}; \\ \text{Score}: \quad z^2 &= \frac{(y-p_0)^2}{p_0(1-p_0)/n}. \end{aligned}$$

- As $n \to \infty$, the three tests have certain asymptotic equivalences.

Generalized Linear Model

Advantages of using deviance

- If $r(\phi) = 1$, it is a proper goodness of fit test and the distribution is exact.
- Deviances are additive, so they can be used to test a subset of coefficients. This is especially good for small sample size.
- For a given error distribution, deviances are a good diagnostic tool for comparing link function. The smaller the deviance, the better the fit.

Disadvantages of Using deviance:

- If $r(\phi) \neq 1$ it is not a valid good of fit test.
- The deviance cannot be used to compare different error distributions.

Wald test:

- It can be shown that the results of the Wald test depend on the scale for the parameterization.
- Also, Wald inference is useless when an estimate or H_0 value is on the boundary of the parameter space.

Score Test:

- The Score test should be used, as a proper goodness of fit test, only for cases that $r(\phi) = 1$.
- It can be shown that the score statistic divided by $n - p$ is an estimator of $\frac{1}{r(\phi)}$.

Generalized Linear Model (Over Dispersion)

- Over dispersion exists if the variance is larger than what we expected.
- Usually it is a results of a clustering experimental units into homogeneous groups.
- These cluster in turn produce a scale parameter $r(\phi) \neq 1$.
- Example: Let $r(\phi) = \frac{1}{\sigma^2}$, then
  $f(y_i) = exp\left\{\frac{1}{\sigma^2}[y_i\theta_i - g(\theta_i)] + h(y_i, \phi)\right\}$
- For binomial with $r(\phi) = 1$, we had $var(y_i) = n_i p_i(1 - p_i)$, now it is $\sigma^2 n_i p_i(1 - p_i)$.
- Posson??
- Gamma??

Generalized Linear Model

- How to determine Over Dispersion?
- We can estimate σ^2.
- We conclude to have over dispersion if $\hat{\sigma^2}$ is significantly different from 1.
- There are two cases: with replication and no replication.
- **replication**: Suppose at i-th data point, we have R replication, then compute

$$s_i^2 = \sum_{j=1}^{R} \frac{(y_{ij} - \bar{y}_i)^2}{var(y_i)(R-1)}.$$

- This is the regular variance divided by $var(y_i)$.

Generalized Linear Model

- This takes the heterogeneity of the variance in the model.
- In addition, $var(y_i)$ is scaled since we are testing $H_0 : \sigma^2 = 1$.
- Example: In Binomial case, $var(y_i) = n_i p_i (1 - p_i)$
- Once we get s_i^2, we pool them to get a final estimate of σ^2 via

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{s_i^2}{n}$$

Generalized Linear Model

- **No Replication**:
- Without replication, we cannot estimate σ^2 using previous formula.
- Thus, there is nothing to pool.
- Instead, we estimate σ^2 with

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{var(y_i)(n-p)}$$

where $\hat{\mu}_i$ are the MLE and $var(y_i)$ is scaled.

Logistic Regression

- Logistic regresion is a type of regression that involves binary reposes.
- Thus the y_i's are 0 or 1.
- Model:
$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i; \quad i = 1, 2, \ldots, n.$$
where $E(\epsilon_i) = 0$ and $E(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta} = p_i$.
- This means that each observation is a Bernoulli trial.
- As a results, ϵ_i can take only two possible values: $-p_i$ or $1 - p_i$.
- Normality of error????
- Homogeneous variance???

Logistic Regression

- There two different structures for data in the logistic regression: Group Structure and Ungrouped Structure.

  Group Structure:

- Usually come from designed experiments where we can control regressors.
- For n different Combinations of the regressor variables, we record r_i of successes in the n_i trials at that level.
- Then compute $\hat{p}_i = \frac{r_i}{n_i}$.
- The responses are these proportion.

		$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_k$
n_1	r_1	$\hat{p}_1$	x_{11}	x_{21}	$\dots$	x_{k1}
n_2	r_2	$\hat{p}_2$	x_{12}	x_{22}	$\dots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
n_n	r_n	$\hat{p}_n$	x_{1n}	x_{2n}	$\dots$	x_{kn}

Ungrouped Structure:

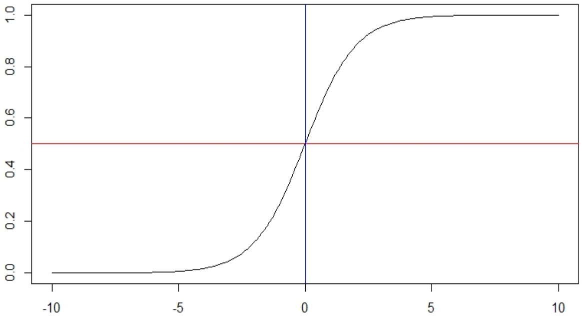
- Usually come from observational studies where y's are responses.
- There are n combinations of the regressors.

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\dots$	$\mathbf{x}_k$
y_1	x_{11}	x_{21}	$\dots$	x_{k1}
y_2	x_{12}	x_{22}	$\dots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
y_n	x_{1n}	x_{2n}	$\dots$	x_{kn}

- In either case, we need a function between 0 and 1 that is increasing function of $\mathbf{x}_i^T\beta$.
- We use the logistic function: $p(\mathbf{x}_i) = \frac{1}{1+\exp\{-\mathbf{x}_i^T\beta\}}$.

Logistic Regression

- $p(\mathbf{x}_i)$ is increasing and ranges from 0 to 1.
- In addition, when $\mathbf{x}_i^T\beta = 0$, then $p(\mathbf{x}_i) = 0.5$.



Logistic Regression

- The logistic model is

$$y_i = \frac{1}{1 + \exp\{-\mathbf{x}_i^T \beta\}} + \epsilon_i$$

- We use the MLE for the parameters.
- The goals are:

1. Estimate β using MLE.
2. Screen variables: Variable selection.
3. Confidence limits on $p(\mathbf{x}_i)$.
4. Diagnostics.

Logistic Transformation

- Let's consider a transformation to linearize the logistic function.
- Starting with $p(\mathbf{x}_i) = \frac{1}{1+\exp\{-\mathbf{x}_i^T\beta\}}$.
- Variance???
- Now one can use WLS to estimate the parameters.
- This procedure is quick, and dirty and has no optimal properties.
- It should be used only if the number of observations at each individual $\mathbf{x}_i$ is relatively large.

Logistic Regression

- Logistic regression uses MLE which depends on the structure of the data.

  Group Structure:

- The likelihood function for the i-th group is

$$\binom{n_i}{r_i} p(x_i)^{r_i}[1-p(x_i)]^{n_i-r_i} = \binom{n_i}{r_i}\left(\frac{1}{1+\exp\{-\mathbf{x}_i^T\beta\}}\right)^{r_i}\left(1-\frac{1}{1+\exp\{-\mathbf{x}_i^T\beta\}}\right)^{n_i-r_i}$$

- The likelihood:

$$L(\beta,\mathbf{x}_i) = \prod_{i=1}^{n}\binom{n_i}{r_i}\left(\frac{1}{1+\exp\{-\mathbf{x}_i^T\beta\}}\right)^{r_i}\left(1-\frac{1}{1+\exp\{-\mathbf{x}_i^T\beta\}}\right)^{n_i-r_i}$$

Logistic Regression

- How find the MLE??

$$\frac{\partial ln[L(\beta, \mathbf{x}_i)]}{\partial \beta} = 0$$

- Simplifying, the MLE for β are the solution of

$$\sum_{i=1}^{n} n_i \left(1 - \frac{\exp\{-\mathbf{x}_i^T \beta\}}{1 + \exp\{-\mathbf{x}_i^T \beta\}}\right) \mathbf{x}_i = \sum_{i=1}^{n} r_i \mathbf{x}_i$$

- We have p equations with p unknowns.
- These equations are not linear in β.
- Thus, we need to use an iterative procedure.

Ungroup Structure:

- Assume the errors are independent from each other but not identically distributed.
- Assume there are n_1 successes.
- The likelihood function is

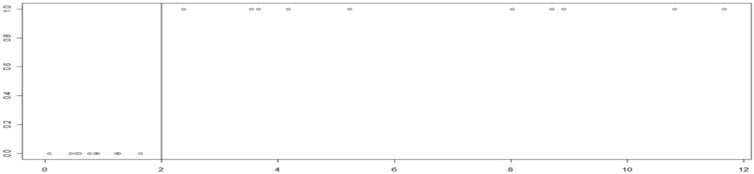
$$L(\beta, \mathbf{x}_i) = \prod_{i=1}^{n} \binom{1}{y_i} p(x_i)^{y_i} [1 - p(x_i)]^{1-y_i}$$

- Simplifying, the MLE for β are the solution of

$$\sum_{i=1}^{n} n_i \left(\frac{\exp\{\mathbf{x}_i^T \beta\}}{1 + \exp\{\mathbf{x}_i^T \beta\}} \right) \mathbf{x}_i = \sum_{i=1}^{n} \mathbf{x}_i$$

Logistic Regression

- In order to interpret the results, be sure the algorithm converges
- When might we not have convergence?
- Many possible curves work so no unique solution
- Even though R may give you a line of best fit, this is an approximation.
- It is NOT the line of best fit.
- With separation, there is no line of best fit.



Logistic Regression

- Test of the logistics equation require to use the likelihood ratio statistic.
- The goal is to find if the logistics regression model is appropriate.
- This can be done by comparing the likelihood of the logistic model with the likelihood when we have perfect fit (a saturated model).
- Likelihood of the logistic model (for ungroupe structure):

$$L(\beta, \mathbf{x}_i) = \frac{\prod_{i=1}^{n_1} \exp\{\mathbf{x}_i^T \beta\}}{\prod_{i=1}^{n} (1 + \exp\{\mathbf{x}_i^T \beta\})}$$

- The likelihood of the saturated model:

$$L(\mathbf{p}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Logistic Regression

- The test statistic to test

$$
\begin{aligned}
H_0 &: \text{Logistic model is appropriate.} \\
H_1 &: \text{Saturated model is appropriate.}
\end{aligned}
$$

- The likelihood ratio statistics:

$$
\lambda(\beta) = -2ln\left(\frac{L(\hat{\beta})}{L(\hat{p})}\right).
$$

- Thus this model deviance follows a chi-square distribution with $n - p$ degrees of freedom.

Logistic Regression

- If the likelihood is close to 1, i.e. model deviance is close to 0, then logistic model is appropriate. .
- If we reject the null hypothesis, then logistic model is not appropriate.
- Model deviance add and subtract like sum of squares.
- Thus, it can be used to test a subset of parameters.
- Suppose we split the $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$.

Logistic Regression

- Suppose we want to test the following:

$$H_0: \quad \boldsymbol{\beta}_1 = 0.$$
$$H_1: \quad \boldsymbol{\beta}_1 \neq 0.$$

- The test statistic is $\lambda(\boldsymbol{\beta}_2) - \lambda(\boldsymbol{\beta})$.

- This is equivalent to

$$-2ln\left(\frac{L(\hat{\boldsymbol{\beta}}_2)}{L(\hat{\boldsymbol{\beta}})}\right) \sim \chi_r^2$$

Logistic Regression (Standard error of the Coefficients)

- Using Fisher information matrix, we have:

$$C = [c_{ij}] = -E\left(\frac{\partial^2 lnL(\hat{\beta})}{\partial\hat{\beta}_i \partial\hat{\beta}_j}\right)$$

- The variance-covariance matrix is C^{-1}.
- The standard error of the coefficients are the square root of the diagonal elements of variance-covariance matrix.
- Note that this procedure can be used to test single coefficients.
- However, it often provides different results from test using deviance method.

Logistic Regression (Measure of performance)

- The fit of the logistic regression model can be analyzed using a R^2-like and adjusted R^2-like statistic.
- Recall that
$$R^2 = \frac{SS_{model}}{SS_{total}}.$$
- The SSE is the model deviance $\lambda(\boldsymbol{\beta})$.
- The SS_{total} does not depend on the regression and equals to the model deviance if one fits a logistic model containing only β_0.

Logistic Regression (Measure of performance)

- Thus

$$R^2 = 1 - \frac{\lambda(\boldsymbol{\beta})}{\lambda(\beta_0)} = \frac{\lambda(\beta_0) - \lambda(\boldsymbol{\beta})}{\lambda(\beta_0)}$$

- The R^2 is a non-decreasing function of the number of covariates.
- The adjusted R^2-like is

$$adj. - R^2 = \frac{\lambda(\beta_0) - \lambda(\boldsymbol{\beta}) - 2p}{\lambda(\beta_0)}$$

Poisson Regression

- Poisson Regression depends on the Poisson probability function

$$f(y,\lambda) = \frac{\lambda^y \exp\{\lambda\}}{y}$$

- The mean and the variance of the Poisson distribution is λ.
- One of the popular model for Poisson Regression is

$$y_i = \exp\{\mathbf{x}_i^T \beta\} + \epsilon_i$$

- Recall that this is the cononical link function for Poisson distribution.

Poisson Regression

- There are three ways we can use the log link to analyze the Poisson data
  - Use IRWLS.

    Since $var(y_i) = \lambda_i = \exp\{\mathbf{x}_i^T\beta\}$, we should weight by $\frac{1}{var(y_i)} = \frac{1}{\exp\{\mathbf{x}_i^T\beta\}}$. Thus, we want to find β that minimizes

    $$\sum_{i=1}^{n} \frac{(y_i - \exp\{\mathbf{x}_i^T\beta\})^2}{\exp\{\mathbf{x}_i^T\beta\}}$$

    The log-link is used to get starting values by regressing $ln(y_i)$ versus $\mathbf{x}_i$ using OLS.
  - Use MLE: We want to find β that maximizes $L(\mathbf{x}, \beta) = \prod_{i=1}^{n} \frac{\lambda^y \exp\{\lambda\}}{y!}$
  - Use transformation to stabilize the Variance. (Similar to HW).

Poisson Regression

- Poisson regression assumes that the variance of the data is equal to their means.
- In real world, that is rarely true and variance is usually greater than mean.
- there are two ways to tackle these types of problem:
  - Use Negative Binomial regression
  - Use zero-inflated Poisson (ZIP) regression.
- The zero-inflated Poisson (ZIP) regression is used for count data that exhibit overdispersion and excess zeros.
- One can divide the data to two groups. A group that is always zero and a group that takes non-zero values.

Zero-Inflated Poisson (ZIP) Regression

- Let assume that π_i shows the probability of being from the always zero group, then we have following probability:
  - Zero counts is always-zero group:

$$p(y_i = 0) = \pi_i \times 1 = \pi_i.$$

  - Zero counts is not from always-zero group:

$$p(y_i = 0) = (1 - \pi_i) \times \frac{\exp\{-\mu_i\}\mu_i^0}{0!} = (1 - \pi_i)\exp\{-\mu_i\}.$$

  - Non-zero counts which is from not always-zero group:

$$p(y_i = j) = (1 - \pi_i) \times \frac{\exp\{-\mu_i\}\mu_i^j}{j!}.$$

  - Put these together:

$$p(y_i = j) = \begin{cases} \pi_i + (1 - \pi_i)\exp\{-\mu_i\} & j = 0 \\ (1 - \pi_i) \times \frac{\exp\{-\mu_i\}\mu_i^j}{j!} & j \neq 0 \end{cases}$$

Zero-Inflated Poisson (ZIP) Regression

- Mean:
$$E(y_i) = 0 \times \pi_i + (1 - \pi_i)\mu_i = (1 - \pi_i)\mu_i$$

- Variance:
$$var(y_i) = (1 - \pi_i)(1 + \mu_i\pi_i)\mu_i$$

- Since $0 \leq \pi_i \leq 1$ for all i, the mean of the ZIP is always less than or equal to mean of the Poisson regression.

- From $Var(y_i) > E(y_i)$, ZIP face the overdespersion problem immediately.