Topic IX: Generalized Linear Model

Wei You

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Fall, 2023

General Linear Models

In a **multiple linear regression model**

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

the response variable $Y_i, i = 1, \ldots, n$ is modeled by a linear function of explanatory variables $x_j, j = 1, \ldots, p$ plus an error term ε_i.

- Multiple linear regression model is sometimes called a **general linear model**.

Here, "general" refer to the dependence on potentially more than one explanatory variable (so $p \geq 1$), as oppose to the **simple linear regression model**:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i.$$

General Linear Models

This model is linear in the parameters, e.g.

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_i.$$

- The error ε_i is assumed to be i.i.d. normal with mean 0 and same variance σ^2.
- Although in generalized least square approach, we allow general multivariate normal with known variance matrix.
- Normal assumption is a basis for statistical inference we see earlier.

Restrictions of General Linear Models

General linear models are not appropriate when

- the range of Y is restricted (e.g., binary, discrete)
- the error term ε is not normally distributed $\Rightarrow$ the variance of Y depends on the mean.

  Example: Poisson distribution is good for modeling the number of arrivals to a hospital or the number of misprints in a book. We know that the the variance of a Poisson RV is equal to its mean.

Generalized linear models extend the general linear model to address these issues.

- We start with two special cases, other than the linear regression model.
- **Logistic regression** and **Poisson regression**.

Logistic Regression – Motivation

- Many applications have a binary response variable Y (e.g. in medicine, patients may be healed or dead).
- The success probability $\mathbb{P}(Y = 1)$ depends on explanatory variables.
- This can also be viewed as a binary classification problem: Given the explanatory variables, which class do we expect? 0 or 1? (More on it later)

- Y_i takes value only from $\{0, 1\}$ and the mean response $\mu_i = \mathbb{E}[Y_i] \in [0, 1]$.
- Linear regression is no longer appropriate!
- How do we generalize the linear regression model we see earlier?

Logistic Regression

Recall that in a linear regression model:

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \sim N(\mu_i, \sigma^2), \text{with } \mu_i = \eta_i \triangleq \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

For binary response variable, we set

$$Y_i \sim \text{Bernoulli}(\mu_i(\eta_i)), \text{where } \; g(\mu_i) = \eta_i$$

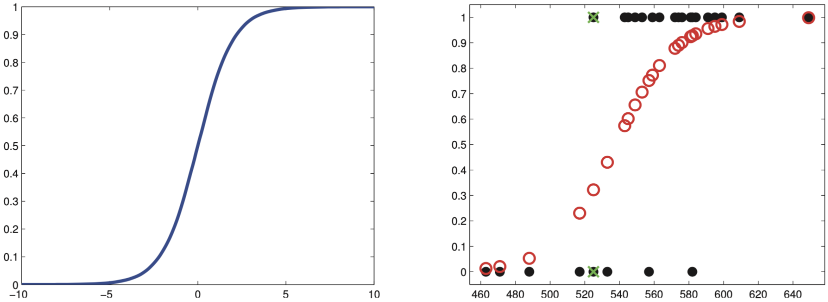
$$\text{with } \; \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \triangleq \boldsymbol{x}_i^\top \boldsymbol{\beta} \in \mathbb{R}.$$

so that the mean response $\mathbb{E}[Y_i]$ is a function of a linear predictor $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

- The linear predictor takes values in the entire real line. But $\mathbb{E}[Y] \in [0, 1]$!
- We use a link function g to remove restrictions on the range of $\mathbb{E}[Y]$:
- g is assumed to be smooth, invertible and does not depend on the sample.

We usually choose g to be the logit (log-odds) function

$$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \Rightarrow \mu = \frac{1}{1+\exp(-\eta)}.$$



Prediction of the success probability: $\mathbb{P}(Y = 1|x) = \mu = \frac{1}{1+\exp(-\eta)}$.

Classification/decision rule: $\widehat{Y} = 1 \iff \mathbb{P}(Y = 1|x) > 0.5$.

Logistic Regression – Decision Boundary

Decision boundary:

$$\mathbb{P}(Y = 1|x) = 0.5 \iff \frac{1}{1 + \exp(-\eta)} = 0.5 \iff \eta = 0 \iff \boldsymbol{\beta}^\top \boldsymbol{x} = 0$$

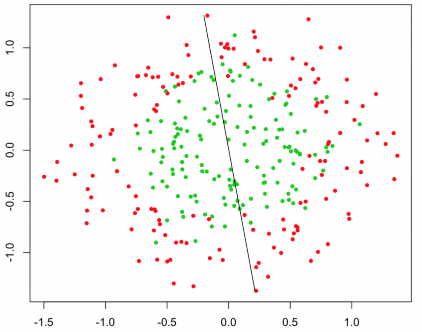
The decision boundary is a hyperplane.

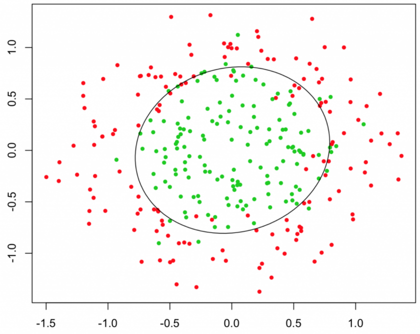
If we replace $\boldsymbol{x}$ by $\boldsymbol{\phi}(\boldsymbol{x})$, like what we did in linear regression.

Nonlinear decision boundary:

$$\mathbb{P}(Y = 1|x) = 0.5 \iff \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) = 0$$

Logistic Regression – Decision Boundary





- Use $\phi(\boldsymbol{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2) \Rightarrow$ quadratic decision boundary.

Logistic Regression – Likelihood Function

We will estimate the parameters $\boldsymbol{\beta}$ by MLE!

Recall that

$$\mu = \frac{1}{1+\exp(-\eta)}, \quad 1-\mu = \frac{1}{1+\exp(\eta)},$$

$$\frac{\mu}{1-\mu} = \exp(\eta), \quad \eta = \log(\mu/(1-\mu)).$$

Then the PDF of Y_i under our model is given by

$$f_{Y_i}(Y_i) = \mu_i^{Y_i}(1-\mu_i)^{1-Y_i} = \left(\frac{\mu_i}{1-\mu_i}\right)^{Y_i}(1-\mu_i) = \frac{(\exp(\eta_i))^{Y_i}}{1+\exp(\eta_i)} = \frac{\exp(Y_i\eta_i)}{1+\exp(\eta_i)}.$$

The joint PDF of n i.i.d. samples is

$$f_{Y_1,\ldots,Y_n}(\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{\exp(Y_i\eta_i)}{1+\exp(\eta_i)} = \prod_{i=1}^{n} \frac{\exp(Y_i\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}.$$

Logistic Regression – MLE

Log-likelihood function

$$\log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} Y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{n} \log(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})).$$

Take partial derivative

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} Y_i \boldsymbol{x}_i - \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))}.$$

Let $\mu_i(\boldsymbol{\beta}) \triangleq \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{(1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))} = \frac{1}{(1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta}))}$, (note that this is exactly the fitted value $\widehat{Y}_i$!) above can be written as

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} (Y_i - \mu_i) \boldsymbol{x}_i = \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}).$$

Logistic Regression – Estimation Equations

Setting it to $\mathbf{0}$ yields the logit-model estimation equations

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}) = \mathbf{0}.$$

- $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ is the <u>fitted value</u>.
- $Y_i - \mu_i$ is the <u>residual</u>.
- Resembles the normal equation in the multiple linear regression

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}.$$

- But the equations here are <u>nonlinear</u> in $\boldsymbol{\beta}$ and, therefore, require iterative solution, e.g., the Newton-Raphson method.

Logistic Regression – Newton-Raphson Method 1-D

Example: The Newton-Raphson method for optimization in one-dimensional problem: find x^* that maximize $f(x)$, where f is continuously differentiable.

- Start with a initial point x_0.
- Taylor's expansion around $x_0 \Rightarrow$ a quadratic function to approximate $f(x)$.

$$f(x_0 + dx) \approx f(x_0) + f'(x_0)dx + \frac{1}{2}f''(x_0)(dx)^2 =: f_{\mathsf{quad}}(dx).$$

- Find dx that maximizes the quadratic approximation f_{quad}:

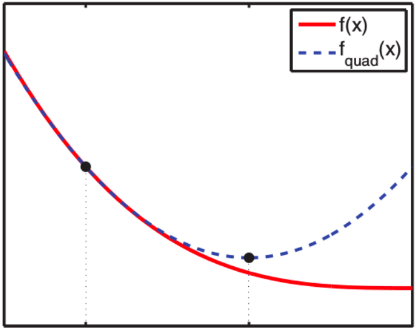
$$dx = -\frac{f'(x_0)}{f''(x_0)}.$$

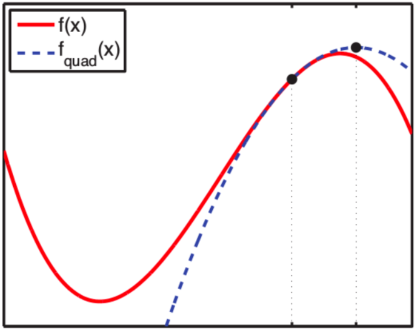
- Update

$$x_1 = x_0 + dx = x_0 - \frac{f'(x_0)}{f''(x_0)}.$$

- Repeat until $x_n - x_{n-1}$ is sufficiently small.

Logistic Regression – Newton-Raphson Method 1-D





Logistic Regression – Newton-Raphson Method N-D

Example: The Newton-Raphson method for optimization in N-dimensional problem: find $\boldsymbol{x}^*$ that maximize $f(\boldsymbol{x})$, where f is twice continuously differentiable.

- Start with a initial point $\boldsymbol{x}_0$.
- Taylor's expansion around $\boldsymbol{x}_0$

$$f(\boldsymbol{x}_0 + \boldsymbol{dx}) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)\boldsymbol{dx} + \frac{1}{2}\boldsymbol{dx}^\top \boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{dx},$$

  where $\nabla f(\boldsymbol{x}_0)$ is the gradient and $\boldsymbol{H}(\boldsymbol{x}_0)$ is the Hessian matrix at $\boldsymbol{x}_0$, i.e.,

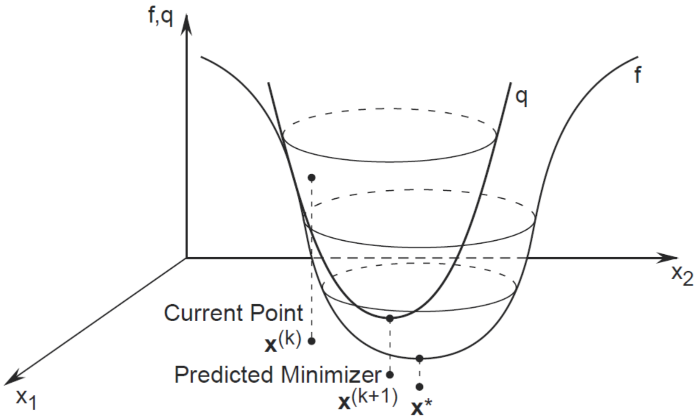
$$(\boldsymbol{H}(\boldsymbol{x}_0))_{ij} = \frac{\partial f(\boldsymbol{x}_0)}{\partial x_i \partial x_j}.$$

- Find $\boldsymbol{dx}$ that maximizes the right-hand-side. Update

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - (\boldsymbol{H}(\boldsymbol{x}_0))^{-1} \nabla f(\boldsymbol{x}_0).$$

- Repeat until $\boldsymbol{x}_n \approx \boldsymbol{x}_{n-1}$.

Logistic Regression – Newton-Raphson Method N-D



Logistic Regression – Estimation

Apply NR method to MLE for logistic regression.

Gradient of $\log(L(\boldsymbol{\beta}))$ at $\boldsymbol{\beta}$

$$\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{\mu}),$$

Hessian of $\log(L(\boldsymbol{\beta}))$ at $\boldsymbol{\beta}$

$$-\boldsymbol{X}^\top W\boldsymbol{X},$$

where

$$\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\beta}) \triangleq \left\{1/(1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})), i=1,\ldots,n\right\}$$

is the vector of fitted values and

$$W=W(\boldsymbol{\beta}) \triangleq \mathsf{diag}(\mu_i(1-\mu_i)).$$

- One can show that the Hessian $-\boldsymbol{X}^\top W\boldsymbol{X}$ is negative definite. (Why?)
- Hence the log-likelihood function has a unique global maximizer (and NR converges).

Logistic Regression – Estimation

- Start with a initial point $\boldsymbol{\beta}_0$.
- At step l, update

$$\boldsymbol{\beta}_{l+1} = \boldsymbol{\beta}_l + \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l),$$

  where $W_l = W(\boldsymbol{\beta}_l)$ and $\boldsymbol{\mu}_l = \boldsymbol{\mu}(\boldsymbol{\beta}_l)$.
- Repeat until $\boldsymbol{\beta}_n \approx \boldsymbol{\beta}_{n-1}$.

$$\begin{aligned}
\boldsymbol{\beta}_{l+1} &= \boldsymbol{\beta}_l + \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l) \\
&= \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^\top W_l \boldsymbol{X} \boldsymbol{\beta}_l + \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l)\right) \\
&= \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top W_l \left(X\boldsymbol{\beta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l)\right) \\
&= \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top W_l \boldsymbol{z}_l,
\end{aligned}$$

where z_l is the **Psudo-response**

$$\boldsymbol{z}_l \triangleq \boldsymbol{X}\boldsymbol{\beta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l).$$

- $\boldsymbol{\beta}_{l+1}$ is the minimizer of a **weighted least square problem**

$$\boldsymbol{\beta}_{l+1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} W_{lii}(z_{li} - \boldsymbol{\beta}^\top \boldsymbol{x}_i)^2, \text{where } W_{lii} = \mu_{li}(1 - \mu_{li}).$$

Logistic Regression – Iteratively Reweighted Least Square (IRLS)

This algorithm is also called **iteratively reweighted least square** (IRLS).

- Start with a initial point $\boldsymbol{\beta}_0$.
- At step l, update

$$\boldsymbol{\mu}_l = \left\{ 1/(1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_l)), i = 1, \dots, n \right\}$$

$$W_l = \text{diag}(\mu_{li}(1-\mu_{li}))$$

$$\boldsymbol{z}_l = \boldsymbol{X}\boldsymbol{\beta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l)$$

$$\boldsymbol{\beta}_{l+1} = \left(\boldsymbol{X}^\top W_l \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top W_l \boldsymbol{z}_l.$$

- Repeat until $\boldsymbol{\beta}_n \approx \boldsymbol{\beta}_{n-1}$.

Simplicity to implement and adaptability to various general settings has made IRLS popular for applications in statistics and engineering contexts.

Logistic Regression – Binomial Data

Suppose, instead of binary response variable, that we observe m groups of experiments. Within each group, we fix the same combination of explanatory-variable $\boldsymbol{x}_i$ and counts the proportion of success Y_i in group i.

The log-likelihood function is

$$\log(L(\boldsymbol{\beta})) = c + \sum_{i=1}^{m} n_i Y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{m} n_i \log(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})),$$

for some constant c that does not depend on $\boldsymbol{\beta}$.

This leads to exactly the same MLE as the logistic regression for binary data.

Poisson Regression for Counts

Count data

- Mortality studies: Number of dead in a period as a function of age, gender, lifestyle...
- Health insurance: Number of claims as a function of age, gender, profession...
- Car insurance: Number of claims as a function of car type, age, gender, previous accidents...
- Train traffic: Number of passengers as a function of time of year, weekday, time of day...
- Football: Number of goals to each team...

The response variable Y_i takes value in integer numbers. One commonly seen discrete random variable that has integer range is the Poisson random variable.

Poisson Regression for Counts

Similar to linear logistic and regression models, in a Poisson regression, we assume that

$$Y_i \sim \text{Poisson}(\mu(\eta_i)), \text{where} \quad g(\mu_i) = \eta_i$$

$$\text{with} \quad \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \triangleq \boldsymbol{x}_i^\top \boldsymbol{\beta} \in \mathbb{R}.$$

- For integer response variable, we usually choose $g(\mu) = \log(\mu)$, or equivalently $\mu = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. Note that the exponential function maps the real line to the positive real line.
- For this reason, a Poisson regression is also called a **log-linear model**.
- Note that g maps positive real line to the entire real line.

Poisson Regression for Counts – Maximum Likelihood Estimation

The log-likelihood function is

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{\beta} Y_i - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \sum_{i=1}^{n} (Y_i!)$$

Hence, the score function is

$$\frac{dl(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \sum_{i=1}^{n} (Y_i - \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))\boldsymbol{x}_i = \sum_{i=1}^{n} (Y_i - \mu_i)\boldsymbol{x}_i = \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu})$$

Likelihood equations

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}) = 0.$$

- $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ is the fitted value.

Review: Contingency Table and Chi-Squared Test of Independence

- Observe discrete samples $(Y_1^{(1)}, Y_1^{(2)})$, ..., $(Y_1^{(n)}, Y_1^{(n)})$, want to test H_0: $Y^{(1)}$ is independent of $Y^{(2)}$.
- Build a frequency table of $Y^{(1)}$ and $Y^{(2)}$. Say $\#\{Y^{(1)} = i, Y^{(2)} = j\} = y_{ij}$, and the marginal frequency $y_{\cdot j}$ and $y_{i\cdot}$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.
- Traditionally, we use Pearson's chi-squared test statistic:

$$V = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(y_{ij} - y_{i\cdot} y_{\cdot j}/n)^2}{y_{i\cdot} y_{\cdot j}/n}.$$

The limiting distribution of V is $\chi^2_{(I-1)(J-1)}$.

An Alternative Test of Independence

Given the total number n of observations, the frequencies in the contingency table follows a multinomial distribution with probabilities π_{ij} so that

$$\pi_{ij} = \mathbb{P}(Y^{(1)} = i, Y^{(2)} = j).$$

- This set of probabilities completely describes the *joint distribution* of $(Y^{(1)}, Y^{(2)})$.
- Define marginal pmf $\pi_{i\cdot} = \mathbb{P}(Y^{(1)} = i)$ and $\pi_{\cdot j} = \mathbb{P}(Y^{(2)} = j)$.
- Test independence $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$.
- Alternatively, we rely on MLE and likelihood ratio test.

Likelihood-ratio Test of Independence

Likelihood function

$$L(\boldsymbol{Y}) = \frac{n!}{\prod_i \prod_j y_{ij}!} \prod_i \prod_j \pi_{ij}^{y_{ij}}.$$

Log-likelihood function

$$l(\boldsymbol{Y}) = C + \sum_i \sum_j y_{ij} \log(\pi_{ij}).$$

- Under the alternative hypothesis, we have a constraint $\sum_i \sum_j \pi_{ij} = 1$. Use Lagrange multiplier to obtain the MLE:

$$\widehat{\pi}_{ij} = \frac{y_{ij}}{n}$$

- Under the null hypothesis, we have constraints $\sum_i \pi_{i\cdot} = 1$ and $\sum_j \pi_{\cdot j} = 1$.

$$\widehat{\pi}_{i\cdot} = \frac{y_{i\cdot}}{n}, \quad \widehat{\pi}_{\cdot j} = \frac{y_{\cdot j}}{n} \quad \Rightarrow \quad \widehat{\mu}_{ij} = n\widehat{\pi}_{i\cdot}\widehat{\pi}_{\cdot j} \quad \text{(MLE of the expected counts)}$$

Likelihood-ratio Test of Independence

For likelihood-ratio test, we compare the ratio between likelihoods. Equivalently, we consider 2 times the difference between the log-likelihoods:

$$D = 2 \sum_i \sum_j y_{ij} \log(y_{ij}/\widehat{\mu}_{ij}) \approx \chi^2_{(I-1)(J-1)}.$$

- This test statistic is also called a deviance.
- The deviance is always positive (why?), and we reject H_0 if D is large.

Test of Independence as Poisson Regression

We now show that the likelihood-ratio test of independence under the multinomial model is the same as that under a special nested model test for Poisson regression.

- Think of the frequencies in the contingency table as independent Poisson random variables with mean μ_{ij}, so that $Y_{ij} \sim \text{Poisson}(\mu_{ij})$.
  - Imagine that there are IJ groups of observations arrive randomly over time, and that the mean number of arrival is μ_{ij}.
  - Unlike the multinomial case, the total number of observations is not fixed to n but is random.
  - (Thinning of Poisson process.) One can check that conditioning on the number of observations n, the vector of the Poisson random counts follows a multinomial distribution with probabilities $\pi_{ij} = \mu_{ij}/n$.

Test of Independence as Poisson Regression

- If model Y_{ij} by a Poisson regression

$$\log(\mu_{ij}) = \eta + \alpha_i + \beta_j$$

  - Note that the above implies that $\mu_{ij} = e^{\eta} e^{\alpha_i} e^{\beta_j}$. This is the independence model!
- If model Y_{ij} by a Poisson regression

$$\log(\mu_{ij}) = \eta + \gamma_{ij}$$

  - This is the satruated (full) model!
- The MLE in this Poisson regression (conditioning on n) under both models are the same as that obtained in the corresponding multinomial model.
- Although we will not prove it rigorously, the multinomial model is equivalent to a Poisson regression. As a result, the test statistic for Poisson regression is the same

$$D = 2 \sum_i \sum_j Y_{ij} \log(Y_{ij}/\widehat{\mu}_{ij}).$$

Poisson Regression for Three-Way Tables

We have seen that Poisson regression can be used to model two-dimensional contingency table, and perform test of independence.

- Poisson regression handles contingency table with more than two dimensions.
- Poisson regression yields much richer models.

Example: Denote S as social status ($I = 4$ states), E as parental encouragement ($J = 2$ states) and P as college plans ($K = 2$ states).

Social Stratum	Parental Encouragement	College Plans No	College Plans Yes	Total
Lower	Low	749	35	784
	High	233	133	366
Lower Middle	Low	627	38	665
	High	330	303	633
Upper Middle	Low	420	37	457
	High	374	467	841
Higher	Low	153	26	179
	High	266	800	1066
Total		3152	1938	4991

Poisson Regression for Three-Way Tables

- The $S+E+P$ model: the three factors are mutually independent.
- The $SE+P$ model: S and E are associated, but are jointly independent of P;
- The $SE+EP$ model: S and E are associated, E and P are associated, but conditioning on E, S and P are independent.
- The $SE+SP+EP$ model: the three factors are pairwise associated, but there is no three-factor interactions. This implies that the association between any two of the factors is the same regardless of the level of the third factor.

Model	Hypothesis	Poisson Regression
$S+E+P$	$H_0: \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k$
$SE+P$	$H_0: \pi_{ijk} = \pi_{ij.}\pi_{..k}$	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$
$SE+EP$	$H_0: \pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk}$
$SE+SP+EP$	???	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$

Poisson Regression for Three-Way Tables

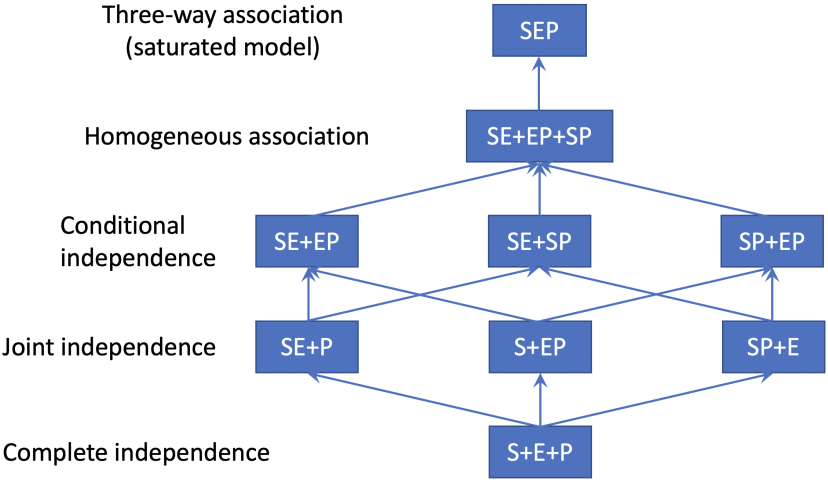
$$D = 2\sum_i\sum_j\sum_k Y_{ijk}\log(Y_{ijk}/\widehat{\mu}_{ijk}).$$

The MLE and deviances for log-linear models fitted to the education data.

Model	MLE $\widehat{\mu}_{ijk}$	Deviance	d.f.	Significance level 0.05
$S+E+P$	$Y_{i\cdot\cdot}Y_{\cdot j\cdot}Y_{\cdot\cdot k}/n^2$	2714.0	10	reject
$SE+P$	$Y_{ij\cdot}Y_{\cdot\cdot k}/n$	1877.4	7	reject
$SE+EP$	$Y_{ij\cdot}Y_{\cdot jk}/Y_{\cdot j\cdot}$	255.5	6	reject
$SE+SP+EP$	???	1.575	3	fail to reject

- For the $SE+SP+EP$ model, no explicit MLE can be written, and the poisson regression most be solved by an iterative algorithm.

Nested Models



Motivation: Exponential Family and Generalized Linear Model

In linear regression:

$$Y_i \sim N(\mu_i, \sigma^2), \text{with } \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

In logistic regression:

$$Y_i \sim \text{Bernoulli}(\mu_i), \text{with } \mu_i = \frac{1}{1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

In Poisson regression:

$$Y_i \sim \text{Poisson}(\mu_i), \text{with } \mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

- Normal, Bernoulli and Poisson belong to **exponential family**.
- In fact, we can carry out the regression analysis for Y that follows a exponential family!
- This general version of regression will be called a **Generalized Linear Model**.

Exponential Family

Exponenital Family

$$f(x) = h(x)c(\tilde{\boldsymbol{\theta}}) \exp\left(\sum_{i=1}^{k} w_i(\tilde{\boldsymbol{\theta}}) t_i(x)\right), \quad \text{for some parameter } \tilde{\boldsymbol{\theta}}.$$
$$= h(x) \exp\left(\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})\right), \quad \text{for } \theta_i = w_i(\tilde{\boldsymbol{\theta}})$$

Note: $c(\tilde{\boldsymbol{\theta}})$ can always be written as functions of $\boldsymbol{\theta}$, even when $\boldsymbol{\theta}(\tilde{\boldsymbol{\theta}})$ is not a one-to-one function, in which case all values of $\tilde{\boldsymbol{\theta}}$ that maps to the same $\boldsymbol{\theta}$ will have the same value of $c(\tilde{\boldsymbol{\theta}})$ and $A(\boldsymbol{\theta})$.

- $\boldsymbol{\theta}$ is the **natural (canonical) parameters**.
- $\boldsymbol{t}(x)$ is the vector of **sufficient statistics**.
- $A(\boldsymbol{\theta})$ is called the **log partition function**, or **cumulant function**.

Exponential Family – Cumulants

Family	$\boldsymbol{\theta}$	$\boldsymbol{t}(x)$	$A(\boldsymbol{\theta})$
Gussian (Normal)	$(\mu/\sigma^2, -1/2\sigma^2)$	(x, x^2)	$-\theta_1/4\theta_2 - \log(-2\theta_2)/2 - \log(2\pi)/2$
Bernoulli	$\log(\mu/(1-\mu))$	x	$\log(1+\exp(\theta))$
Poisson	$\log(\lambda)$	x	e^θ

Cumulants

Cumulant generating function of $\boldsymbol{Z}$ is defined as $K(\boldsymbol{s}) \triangleq \log \mathbb{E}[\exp(\boldsymbol{s}^\top \boldsymbol{Z})]$. Taking derivatives and setting $\boldsymbol{s} = \boldsymbol{0}$ gives the cumulants of $\boldsymbol{Z}$.

- Derivatives of $A(\boldsymbol{\theta})$ can be used to generate the cumulants of the sufficient statistics.

$$\frac{\partial}{\partial \boldsymbol{\theta}} A(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{t}(X)], \quad \frac{\partial^2}{\partial \boldsymbol{\theta}^2} A(\boldsymbol{\theta}) = \mathrm{cov}(\boldsymbol{t}(X)), \ldots$$

Let's check the first two cumulants. Note that

$$A(\boldsymbol{\theta}) = \log \int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x.$$

The first moment (cumulant) is then given by

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} A(\boldsymbol{\theta}) &= \frac{\int h(x) \frac{\partial}{\partial \boldsymbol{\theta}} \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x}{\int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x} \quad \text{[Interchange of limits justified by DCT.]} \\
&= \frac{\int \boldsymbol{t}(x) h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x}{\int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x} \\
&= \int \boldsymbol{t}(x) h(x) \exp\left(\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})\right) \mathrm{d}x \\
&= \mathbb{E}[\boldsymbol{t}(x)].
\end{aligned}$$

The second cumulant is then given by

$$
\begin{aligned}
\frac{\partial^2}{\partial\theta_i\partial\theta_j}A(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\int \boldsymbol{t}(x)h(x)\exp\left(\boldsymbol{\theta}^\top\boldsymbol{t}(x)-A(\boldsymbol{\theta})\right)\mathrm{d}x \\
&= \int \boldsymbol{t}(x)\left(\boldsymbol{\theta}^\top\boldsymbol{t}(x)-\frac{\partial}{\partial\boldsymbol{\theta}}A(\boldsymbol{\theta})\right)^\top h(x)\exp(\boldsymbol{\theta}^\top\boldsymbol{t}(x)-A(\boldsymbol{\theta}))\mathrm{d}x \\
&= \mathbb{E}[\boldsymbol{t}(X)\boldsymbol{t}(X)^\top]-\mathbb{E}[\boldsymbol{t}(X)]\mathbb{E}[\boldsymbol{t}(X)]^\top \\
&= \mathrm{cov}(\boldsymbol{t}(X)).
\end{aligned}
$$

Remark

The cumulant generating function of $\boldsymbol{t}(X)$ is in fact given by

$$K(\boldsymbol{s}) \triangleq \log\mathbb{E}[\exp(\boldsymbol{s}^\top\boldsymbol{t}(X))] = A(\boldsymbol{s}+\boldsymbol{\theta})-A(\boldsymbol{\theta}).$$

But taking derivatives of $C(\boldsymbol{s})$ and setting $\boldsymbol{s}=\mathbf{0}$ is the same as that of $A(\boldsymbol{\theta})$.

Exponential Family – Cumulants

Example: Bernoulli.

$$A(\theta) = \log(1 + \exp(\theta)) \Rightarrow \mathbb{E}[x] = \frac{1}{1 + e^{-\theta}} = \mu$$

$$\Rightarrow \mathrm{Var}(x) = \frac{d}{d\theta} \frac{1}{1 + e^{-\theta}} = \frac{e^{-\theta}}{1 + e^{-\theta}} \frac{1}{1 + e^{-\theta}} = (1 - \mu)\mu.$$

Example: Poisson. $A(\theta) = e^{\theta}, \mathbb{E}[x] = \lambda, \mathrm{Var}(x) = \lambda.$

Exponential Family – MLE

The MLE for $\boldsymbol{\theta}$ in the exponential family can be obtained by method of moments.

Log-likelihood function

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(h(\boldsymbol{x}_i)) + \boldsymbol{\theta}^\top \sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i) - nA(\boldsymbol{\theta}).$$

$$\Rightarrow \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i) - n\frac{\partial}{\partial \boldsymbol{\theta}} A(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i) - n\mathbb{E}[\boldsymbol{t}(X)].$$

Thus the MLE $\widehat{\boldsymbol{\theta}}$ satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i) = \mathbb{E}[\boldsymbol{t}(X)]$$

Generalized Linear Models – Introduction

Generalized linear models are models in which

- the output density is in the exponential family $\Rightarrow$ variance depends on mean.
- the mean parameters are a linear combination of the inputs, passed through a possibly nonlinear "link" function;

They are used in scenarios such that

- the range of Y is restricted (e.g., binary, discrete)
- the response Y is not normally distributed.

Model	Family	Link function	Range of Y_i	$\mathrm{Var}(Y_i)$
Linear regression	Gussian	Identity	$(-\infty, \infty)$	$\phi = \sigma^2$
Logistic regression	Bernoulli	Logit	$0, 1$	$\mu_i(1 - \mu_i)$
Poisson regression	Poisson	Log	$0, 1, 2, \ldots$	μ_i

Generalized Linear Models (GLMs)

A **Generalized linear model** has the following elements:

- A **linear predictor**

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

- A **random component**, specifying the conditional distribution of the response variable, given the values of the explanatory variables.
  - We will only look at exponential families, so Y_i may also take binomial, Poisson, gamma, etc.
- A **link function** that "links" the mean response $\mathbb{E}[Y_i] = \mu_i$ to the predictor η_i

$$g(\mu_i) = \eta_i.$$

  - The link function is assumed to be smooth and invertible.

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}).$$

Exponential Family

Let's focus on a simple case

$$Y \sim f(y|\theta, \phi) = \exp\left[\frac{y\theta - A(\theta)}{\phi} + c(y, \phi)\right], \quad \text{for a fixed } \phi$$

where ϕ is the dispersion parameter and θ is the natural parameter, A is the log partition function.

- For normal distribution, $\phi = \sigma^2$.
- For Binomial and Poisson distributions, $\phi = 1$.

Exponential Family

- We can show that

$$\mu_i \triangleq \mathbb{E}[Y_i|\theta_i, \phi] = A'(\theta_i),$$

$$\mathrm{Var}(Y_i|\theta_i, \phi) = A''(\theta_i)\phi.$$

- There is a one-to-one correspondence between θ and μ, i.e.,

$$\theta_i = (A')^{-1}(\mu_i) \triangleq \psi(\mu_i),$$

because $A''(\theta_i) = \mathrm{Var}(Y_i|\theta_i, \phi)/\phi > 0$. Hence, we can define

$$V(\mu_i) = A''((A')^{-1}(\mu_i)) = A''(\theta_i) = \mathrm{Var}(Y_i|\theta_i, \phi)/\phi.$$

Overview of GLM Notations

linear predictor — mean response — exp. family parameter

$$\left.\begin{array}{l} \boldsymbol{x}_i \\ \boldsymbol{\beta} \end{array}\right\} \rightarrow \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \underset{g}{\overset{g^{-1}}{\rightleftarrows}} \mu_i \underset{\psi^{-1}}{\overset{\psi}{\rightleftarrows}} \theta_i$$

Maximum Likelihood Estimation

GLM can be fit using MLE.

Log-likelihood function

$$l(\boldsymbol{\beta}) \triangleq \log(L(\boldsymbol{\beta})) = \frac{1}{\phi}\sum_{i=1}^{n}(\theta_i Y_i - A(\theta_i)) + \sum_{i=1}^{n} c(Y_i, \phi) \triangleq \sum_{i=1}^{n} l_i + \sum_{i=1}^{n} c(Y_i, \phi).$$

To compute the gradient

$$\begin{aligned}\frac{dl_i}{d\beta_j} = \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_j} &= \frac{1}{\phi}(Y_i - A'(\theta_i))\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}x_{ij} \\ &= \frac{1}{\phi}(Y_i - \mu_i)\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}x_{ij}.\end{aligned}$$

Exponential Family – Canonical Link Function

To simplify the log-likelihood function, we define the **canonical link function**

$$g(\mu_i) = \psi(\mu_i),$$

so that

$$\underbrace{\theta_i = \psi(\mu_i)}_{\text{def of } \psi} = \underbrace{g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}}_{\text{GLM assumption}}.$$

Model	Canonical link	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = A'(\theta)$
Linear regression	Identity	$\theta = \mu$	$\mu = \theta$
Logistic regression	Logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \frac{1}{1+e^{-\theta}}$
Poisson regression	Log	$\theta = \log(\mu)$	$\mu = e^\theta$

- Recall that $A''(\theta) = \mathrm{Var}(Y|\theta, \phi)/\phi > 0$. Then A is a convex function in θ.
- Under canonical link function $\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, hence A is a convex function in $\boldsymbol{\beta}$.
- The log-likelihood function is a concave function in $\boldsymbol{\beta}$, and Newton-Raphson method can be used to find the MLE.

Maximum Likelihood Estimation – Canonical Link Function

With canonical link, we have $\theta_i = \eta_i$.

- Hence, $\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i} = 1$ and

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - \mu_i)\boldsymbol{x}_i = \frac{1}{\phi}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}).$$

- The likelihood equation takes the exact same form as in the linear, logistic and Poisson regression models

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}) = \mathbf{0}.$$

- It is a non-linear equation $\Rightarrow$ Newton-Raphson method.

Maximum Likelihood Estimation – Canonical Link Function

- The gradient $s(\boldsymbol{\beta}) \triangleq \frac{dl(\boldsymbol{\beta})}{d\boldsymbol{\beta}}$ is often referred to as the **score function**, so that

$$s_j(\boldsymbol{\beta}) \triangleq \frac{dl(\boldsymbol{\beta})}{d\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - \mu_i)x_{ij}$$

- Similarly, we have (recall Fisher information!)

$$\boldsymbol{H} = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right]_{1\le j,k\le p} = -\frac{1}{\phi}\sum_{i=1}^{n}\frac{d\mu_i}{d\theta_i}\boldsymbol{x}_i\boldsymbol{x}_i^\top = -\frac{1}{\phi}\boldsymbol{X}^\top W\boldsymbol{X},$$

where

$$W = \text{diag}\left(\frac{d\mu_1}{d\theta_1}, \frac{d\mu_2}{d\theta_2}, \ldots, \frac{d\mu_n}{d\theta_n}\right), \quad \frac{d\mu_i}{d\theta_i} = (\psi^{-1})'(\theta_i) = A''(\theta_i) = V(\mu_i).$$

Note that W here (under canonical link) depend on β through $\boldsymbol{\theta} = \boldsymbol{\eta} = \boldsymbol{X}^\top\beta$.

Maximum Likelihood Estimation – Canonical Link Function

The Newton update is exactly the same as that in a Logistic regression

$$\boldsymbol{\beta}_{l+1} = \boldsymbol{\beta}_l + \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l).$$

It can also be reformulated as iteratively re-weighted least square update:

$$\begin{aligned}
\boldsymbol{\beta}_{l+1} &= (\boldsymbol{X}^\top W_l \boldsymbol{X})^{-1} \boldsymbol{X}^\top W_l \boldsymbol{z}_l \\
\boldsymbol{z}_l &= \boldsymbol{\theta}_l + W_l^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_l) \\
\boldsymbol{\theta}_l &= \boldsymbol{X}\boldsymbol{\beta}_l = \boldsymbol{\eta}_l \\
\boldsymbol{\mu}_l &= g^{-1}(\boldsymbol{\eta}_l)
\end{aligned}$$

General Link Function

For general link function g, we need to find out $\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}$.

- Link function
$$g(\mu) = \eta \Rightarrow \frac{d\mu_i}{d\eta_i} = \frac{1}{\frac{d\eta_i}{d\mu_i}} = \frac{1}{g'(\mu_i)}.$$
- The mean μ and the natural parameter η are linked by
$$\theta = \psi(\mu), \quad \mu = \psi^{-1}(\theta) = A'(\theta) \Rightarrow \quad \frac{d\theta_i}{d\mu_i} = \frac{1}{\frac{d\mu_i}{d\theta_i}} = \frac{1}{A''(\theta_i)} \triangleq \frac{1}{V(\mu_i)}.$$

The gradient (score function) is computed by

$$s_j = \frac{dl}{d\beta_j} = \sum_{i=1}^{n} \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - A'(\theta_i))\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}x_{ij}$$
$$= \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}\frac{Y_i - \mu_i}{g'(\mu_i)V(\mu_i)}.$$

General Link Function

The Hessian is computed by

$$
\begin{aligned}
H_{jk} = \frac{ds_j}{d\beta_k} &= \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}\frac{d}{d\beta_k}\left(\frac{Y_i-\mu_i}{g'(\mu_i)V(\mu_i)}\right) \\
&= \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}\frac{d\eta_i}{d\beta_k}\frac{d\mu_i}{d\eta_i}\frac{d}{d\mu_i}\left(\frac{Y_i-\mu_i}{g'(\mu_i)V(\mu_i)}\right) \\
&= \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}x_{ik}\frac{1}{g'(\mu_i)}\frac{d}{d\mu_i}\left(\frac{Y_i-\mu_i}{g'(\mu_i)V(\mu_i)}\right)
\end{aligned}
$$

where

$$
\frac{d}{d\mu_i}\left(\frac{Y_i-\mu_i}{g'(\mu_i)V(\mu_i)}\right) = -\frac{1}{g'(\mu_i)V(\mu_i)} + (Y_i-\mu_i)\frac{d}{d\mu_i}\left(\frac{1}{g'(\mu_i)V(\mu_i)}\right)
$$

General Link Function – Newton-Raphson

With the gradient and the Hessian, we can apply Newton-Raphson just as before.

Note that the Hessian is stochastic (depends on $\boldsymbol{Y}$). We have

$$\mathbb{E}\left[(Y_i - \mu_i)\frac{d}{d\mu_i}\left(\frac{1}{g'(\mu_i)V(\mu_i)}\right)\right] = 0.$$

General Link Function – Fisher Scoring

Hence, we have

$$\mathbb{E}\left[H\right] = -\frac{1}{\phi}\boldsymbol{X}^\top W \boldsymbol{X},$$

where

$$W = \text{diag}\left(\frac{1}{g'(\mu_1)^2 V(\mu_1)}, \frac{1}{g'(\mu_2)^2 V(\mu_2)}, \ldots, \frac{1}{g'(\mu_n)^2 V(\mu_n)}\right).$$

- W depends on $\boldsymbol{\beta}$ through

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

- **Fisher information matrix** is defined as

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}\left[-H\right] = \frac{1}{\phi}\boldsymbol{X}^\top W \boldsymbol{X}.$$

It does not depend on the observed response Y_i.

Maximum Likelihood Estimation – Fisher Scoring

If we use the expected Hessian (or, equivalently, the information matrix) in the updates, then it takes the same form (except that W_l are different!) as we see in previous cases:

$$\boldsymbol{\beta}_{l+1} = \boldsymbol{\beta}_l + \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l),$$

Same iteratively re-weighted least square formulation.

- This procedure is called **Fisher Scoring**.
- The difference between Newton-Raphson and Fisher Scoring is that the former one use (stochastic) observed Hessian, whereas the later one use the expected Hessian.
- Observed and expected information is equivalent under canonical links.

Estimation of the Dispersion Parameter

MLE for GLM does not depend on the dispersion parameter ϕ!

- We shall need an estimation for ϕ when we perform inference on the GLM.
- It is usually estimated using the **Pearson chi-squared statistic**

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

- The *scaled* Pearson chi-squared statistic is

$$\mathcal{X}_s^2 = \mathcal{X}^2/\phi. \quad \text{(Recall that } \mathrm{Var}(Y_i) = V(\mu_i)\phi.\text{)}$$

- As $n \to \infty$, we have $\mathcal{X}_s^2 \approx \chi^2_{n-p}$.
- To estimate ϕ, we use the asymptotically unbiased estimator

$$\widehat{\phi} = \frac{\mathcal{X}^2}{n-p}.$$

Estimation of the Dispersion Parameter

Example: Linear regression. $V(\mu_i) = 1$ and $\widehat{\phi}$ is a unbiased estimator of σ^2.

$$\widehat{\phi} = \widehat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-p}$$

Example: Logistic regression. $V(\mu_i) = \mu_i(1-\mu_i)/\phi$ and $\phi = 1$. The estimator is

$$\widehat{\phi} = \widehat{\sigma}^2 = \frac{1}{n-p} \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\mu_i(1-\mu_i)}.$$

- Note that there is actually no need to estimate ϕ in logistic regression. If the model is correct, then we know exactly that $\phi = 1$.
- But is the model correct? Later we will see tests based on the estimator above. (Basically, you check if $\widehat{\phi} \approx 1$.)

Statistical Inference

Statistical inference on GLM

- Large-sample theories.
- Hypothesis testing.
- Confidence interval.

General Linear Hypothesis

- $H_0 : \beta_j = \beta_j^*$
- $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$
- $H_0 : \beta_i = \beta_j$
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- $H_0 : C\boldsymbol{\beta} = \boldsymbol{r}$, where C is a $q \times p$ matrix.

Large Sample Results – MLE

The MLE $\widehat{\boldsymbol{\beta}}$ is a p-dimensional random vector.

As $n \to \infty$, we have

$$\widehat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta})),$$

where $\mathcal{I}^{-1}(\boldsymbol{\beta})$ is the inverse of the Fisher information matrix.

- Then plug in the MLE $\widehat{\boldsymbol{\beta}}$

$$\widehat{\mathcal{I}} = \mathcal{I}(\widehat{\boldsymbol{\beta}}) = \frac{1}{\phi} X^\top \widehat{W} X.$$

When testing **one** parameter, the marginal distribution can be used

- in z-test, if the dispersion parameter ϕ is known; and
- in t-test, if the dispersion parameter ϕ is unknown and is estimated.

When testing **multiple** parameter, we will need the **Wald test**.

Wald Test

Suppose we want to test $H_0 : C\boldsymbol{\beta} = \boldsymbol{r}$, where C is a full rank $q \times p$ matrix with $q < p$.

Fact by eigenvalue decomposition

For a q-dimensional normal random vector $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, V)$, we have

$$(\boldsymbol{Y} - \boldsymbol{\mu})^\top V^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_q.$$

Define the **Wald test statistic** as, (note that $\widehat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta}))$)

$$(C\widehat{\boldsymbol{\beta}} - \boldsymbol{r})^\top [C\widehat{\mathcal{I}}^{-1}C^\top]^{-1} (C\widehat{\boldsymbol{\beta}} - \boldsymbol{r}) \approx \chi^2_{q}, \quad \text{under } H_0.$$

The hypothesis is rejected if the test statistics is large.

- In the case of logistic regression and Poisson regression, $\phi = 1$.
- If the dispersion parameter is unknown, e.g., in linear regression, plug in a consistent estimate $\widehat{\phi}$ of ϕ.

Wald Test

Example: Linear regression. Consider linear regression and $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, then

- W and C are identity matrices.
- $\widehat{\mathcal{I}}(\boldsymbol{\beta}) = \frac{1}{\phi}\boldsymbol{X}^\top W \boldsymbol{X} = \frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X}$.

If the variance is unknown, we estimate it using $\widehat{\sigma}^2 = \sum_{i=1}^n \frac{(Y_i - \widehat{Y}_i)^2}{n-p}$, then

$$\frac{1}{\sigma^2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim \chi^2_p, \quad \text{and} \quad \frac{1}{\widehat{\sigma}^2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim F_{p,n-p}$$

What's the difference?

- More precisely, when we plug in $\widehat{\phi} \sim \chi^2_{n-p}$, we should use $F_{p,n-p}$ instead of χ^2_p. But they are asymptotically equivalent when n is large.

Large Sample Results – Likelihood Ratio Function

Consider the general linear hypothesis $H_0 : C\boldsymbol{\beta} = \boldsymbol{r}$, where C is a $q \times p$ matrix.

- Let $\widehat{\boldsymbol{\beta}}$ denote the MLE of the GLM.
- Let $\widehat{\boldsymbol{\beta}}_0$ denote the MLE in the restricted model, where $C\boldsymbol{\beta} = \boldsymbol{r}$.

Then the **likelihood ratio statistics** is

$$2\log[L(\widehat{\boldsymbol{\beta}})/L(\widehat{\boldsymbol{\beta}}_0)] = 2\left[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_0)\right] \approx \chi_q^2$$

- If the ϕ is unknown, one has to use the **same** consistent estimate of ϕ in the two log likelihoods.

Large Sample Results – The Score Function

The score function $s(\boldsymbol{\beta}) = \frac{dl(\boldsymbol{\beta})}{d\boldsymbol{\beta}}$ under general link function is given by

$$s_j(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{i=1}^{n} x_{ij} \frac{Y_i - \mu_i}{g'(\mu_i)V(\mu_i)}.$$

For any n, we have

$$\mathbb{E}[s(\boldsymbol{\beta})] = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}(s(\boldsymbol{\beta})) = \mathcal{I}(\boldsymbol{\beta}).$$

Furthermore, when $n \to \infty$, we have

$$s(\boldsymbol{\beta}) \approx N(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}))$$

$$s(\boldsymbol{\beta})^\top \mathcal{I}^{-1}(\boldsymbol{\beta}) s(\boldsymbol{\beta}) \approx \chi^2_p$$

- Can be used to test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, in which case $\widehat{\boldsymbol{\beta}}$ is not needed. (plug in $\boldsymbol{\beta}_0$)

Remarks on the Three Tests

- When n is large, the Wald test, likelihood ratio test (LRT) and score test are asymptotically equivalent.
- For small sample size, the LRT and score test usually work better. (Think of the profile confidence interval, we are using the full shape of the likelihood function, instead its asymptotic properties.)
- The LRT can be computed directly from the likelihood and does not need an estimate of the Fisher information. It is simple to use (if the likelihood can be easily expressed).

Testing Goodness-of-Fit

Data: $(Y_i, \boldsymbol{x}_i), i = 1, 2, \ldots, n$.

- Saturated model: the number of parameters equals the number of observations so that each observation have a dedicated parameter μ_i.
- Null model: one parameter $\Rightarrow \mu_i = \mu$.
- GLM model: p parameters $\Rightarrow \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.

We want a model that *describes data well* and has *as few parameters as possible*.

To measure the discrepancy between the data $\boldsymbol{Y}$ and the fitted values $\widehat{\boldsymbol{\mu}}$, there are two commonly used measures: **Pearson's chi-squared (goodness-of-fit) statistic** and **deviance**.

Pearson's Chi-Squared

Pearson's chi-square statistics:

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

It is a generalization of residual sum of squares (RSS) in linear regression.

Example: Normal. $\mathcal{X}^2 = RSS$.

Example: Poisson. $\mathcal{X}^2 = \sum (Y_i - \widehat{\mu}_i)^2 / \widehat{\mu}_i$.

Example: Binomial. $\mathcal{X}^2 = \sum (Y_i - \widehat{\mu}_i)^2 / (\widehat{\mu}_i (1 - \widehat{\mu}_i))$.

If the model is correct, then $\mathcal{X}^2/(n-p) \approx \phi$.

- Normal: $RSS/(n-p) \approx \sigma^2 = \phi$.
- Binomial: $\mathcal{X}^2/(n-p) \approx 1 = \phi$.
- Poisson: $\mathcal{X}^2/(n-p) \approx 1 = \phi$.

Deviance

Recall that in the exponential family for GLM, the natural parameter η is connected to the mean parameter via

$$\mu_i = A'(\theta_i), \quad \theta_i = \psi(\mu_i).$$

Hence, the log-likelihood function can also be viewed as a function of the mean parameter μ_i

$$l(\boldsymbol{\mu}) = \frac{1}{\phi}\sum_{i=1}^{n}(\theta_i Y_i - A(\theta_i)) = \frac{1}{\phi}\sum_{i=1}^{n}(\psi(\mu_i)Y_i - A(\psi(\mu_i)))$$

- Saturated model: MLE of μ_i is given by Y_i. (Why?)
- GLM: MLE of μ_i is given by $\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$.

Deviance

The deviance is defined as

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}) = 2[l(\boldsymbol{Y}) - l(\widehat{\boldsymbol{\mu}})] = \frac{2}{\phi} \sum_{i=1}^{n} ((\tilde{\theta}_i - \widehat{\theta}_i) Y_i - A(\tilde{\theta}_i) + A(\widehat{\theta}_i)),$$

where $\tilde{\theta}_i = \psi(Y_i)$ and $\widehat{\theta}_i = \psi(\widehat{\mu}_i)$.

- Deviance is always non-negative. (Why?)
- Deviance is finite in usual cases, but NOT always.

Deviance and The Likelihood Ratio Statistic

Difference between the likelihood ratio (LR) statistics and the deviance

- The LR statistic can be viewed as the difference between the deviances for two models: full GLM model and the restricted GLM model with $C\boldsymbol{\beta} = \boldsymbol{r}$.
- The LR statistic is asymptotically χ^2_q.
- Under some restrictive conditions, the deviance is asymptotically χ^2_{n-p}. It is NOT always true, mainly because the degrees of freedom (DF) is growing as fast as the number of observations.

Testing Goodness-of-Fit

When the deviance can be approximated by χ^2_{n-p}, we use it to test goodness-of-fit. A GLM model with too large a deviance does not fit the data well.

Nested Models

Recall that in linear regression, we looked at nested model tests using the difference of RSS's. For GLM, we can perform the same test.

For a model $\mathcal{M}$, let $\mathcal{B}$ be the restricted set of parameters in model $\mathcal{M}$.

Example: For the model with $\beta_1 = 0$, we have $\mathcal{B} = \{\boldsymbol{\beta} : \beta_1 = 0\}$.

Example: For the model with $C\boldsymbol{\beta} = \boldsymbol{r}$, we have $\mathcal{B} = \{\boldsymbol{\beta} : C\boldsymbol{\beta} = \boldsymbol{r}\}$.

We write $\mathcal{M}_1 \subset \mathcal{M}_2$ if $\mathcal{B}_1 \subset \mathcal{B}_2$. Equivalently, $\mathcal{M}_1$ is a special case of $\mathcal{M}_2$.

Definition (Nested Model)

We say that a sequence of models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is nested, if $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B}_m$.

Let D_i be the deviance of the ith model.

The deviance is monotone, i.e., if $\mathcal{M}_1 \subset \mathcal{M}_2$ then $D_1 \geq D_2$.

Nested Models

We now consider only the case where $\mathcal{B}_i$ is a q_i dimensional subspace in $\mathbb{R}^p$ and $q_1 < q_2 < \cdots < q_m$.

Example: A special case. In model i, let $\mathcal{B}_i = \{\boldsymbol{\beta} : \beta_{i+1} = \beta_{i+2} = \cdots = \beta_p = 0\}$.

For nested models, let D_i be the deviance of the ith model.

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}_{(i)}) = 2[l(\boldsymbol{Y}) - l(\widehat{\boldsymbol{\mu}}_{(i)})],$$

where $\widehat{\boldsymbol{\mu}}_{(i)}$ is the MLE for the ith model.

For all $i < j$, the difference of deviance $D_i - D_j \approx \chi^2_{q_j - q_i}$ for large samples.

- Deviance measures the goodness-of-fit of a model with respect to the saturated model.
- The difference in deviance help us to understand the improvement in goodness-of-fit due to additional parameters.

Nested Model Tests

For hypothesis test $H_0 : \mathcal{M}_i$ vs $H_1 : \mathcal{M}_j$, we use

$$\Delta D = D_i - D_j = 2[l(\widehat{\boldsymbol{\mu}}_{(j)}) - l(\widehat{\boldsymbol{\mu}}_{(i)})] \approx \chi^2_{p_j - p_i}.$$

- ΔD is the log of the likelihood ratio of M_i and M_j.
- The distribution of ΔD is usually better approximated by the chi-squared distribution than a single deviance because the DF is bounded.

Hypothesis testing

- If ϕ is unknown, plug in a consistent estimate of it.
- Reject model $\mathcal{M}_i$ in favor of model $\mathcal{M}_j$ if the observed ΔD is in the critical region, i.e., if $\Delta D > \chi^2_{p_j - p_i, 1-\alpha}$.
- Rejecting model $\mathcal{M}_i$ does not mean that model $\mathcal{M}_j$ fit the data well.

Nested Model Tests – Linear Regression

Example: Linear regression $Y_i \sim N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$. One can check that the deviance is

$$D = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})^2 = \frac{1}{\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{X}^\top \widehat{\boldsymbol{\beta}})$$

$$= \frac{1}{\sigma^2}(\boldsymbol{Y}^\top \boldsymbol{Y} - 2\widehat{\boldsymbol{\beta}}\boldsymbol{X}^\top \boldsymbol{Y} + \widehat{\boldsymbol{\beta}}\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

$$= \frac{1}{\sigma^2}\left(\boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y}\right)$$

$$= \frac{1}{\sigma^2}\left(\boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^\top \widehat{\boldsymbol{Y}}\right)$$

because $X^\top X\widehat{\boldsymbol{\beta}} = X^\top \boldsymbol{Y}$ (normal equation for linear regression).

Nested Model Tests – Linear Regression

For the testing of two nested models $\mathcal{M}_0$ and $\mathcal{M}_1$ with $q_0 < q_1$ parameters, respectively. If the null hypothesis (model $\mathcal{M}_0$) is true, then

$$\Delta D = D_0 - D_1 = \frac{1}{\sigma^2}\left(\widehat{\boldsymbol{\beta}}_{(1)}^T \boldsymbol{X}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}_{(0)}^T \boldsymbol{X}^\top \boldsymbol{Y}\right) \sim \chi^2_{q_1 - q_0}.$$

Because $\mathcal{M}_0 \subset \mathcal{M}_1$, model $\mathcal{M}_1$ is also true, then

$$D_1 = \frac{1}{\sigma^2}\left(\boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}_{(1)}^\top \boldsymbol{X}^\top \boldsymbol{Y}\right) \sim \chi^2_{n - q_1}.$$

We have seen that D_1 is independent of $D_0 - D_1$. Hence,

$$\frac{D_0 - D_1}{q_1 - q_0} \bigg/ \frac{D_1}{n - q_1} \sim F_{q_1 - q_0, n - q_1},$$

where the unknown σ^2 is cancelled out.

Nested Model Tests – An Example

Example: GLM with two groups of factors.

Let Model 1 be

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1, \quad \boldsymbol{x}_i \in \mathbb{R}^{p_1}$$

Let Model 2 be

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2, \quad \boldsymbol{x}_i \in \mathbb{R}^{p_1}, \boldsymbol{z}_i \in \mathbb{R}^{p_2}$$

Then $\mathcal{M}_1 \subset \mathcal{M}_2$.

Predictor	Model	#parameters	Deviance	DF
Intercept only	$g(\mu) = \alpha_0$	1	D_0	$n-1$
Single factor $\boldsymbol{x}$	$g(\mu) = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}$	p_1	D_1	$n-p_1$
Single factor $\boldsymbol{z}$	$g(\mu) = \lambda_0 + \boldsymbol{z}^\top \boldsymbol{\lambda}$	p_2	D_2	$n-p_2$
Two factors	$g(\mu) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}_1 + \boldsymbol{z}^\top \boldsymbol{\beta}_2$	$p_3 = p_1 + p_2 - 1$	D_3	$n-p_3$

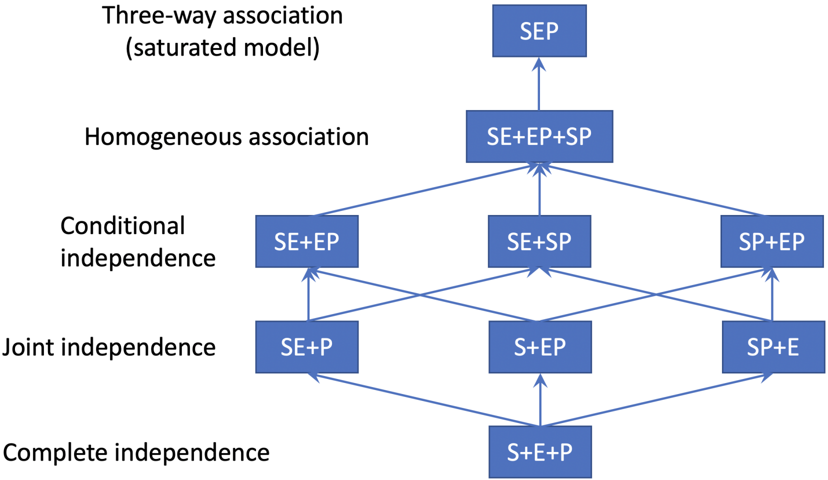
Nested Model Tests – An Example Cont.

Predictor	Model	#parameters	Deviance	DF
Intercept only	$g(\mu) = \alpha_0$	1	D_0	$n-1$
Single factor $\boldsymbol{x}$	$g(\mu) = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}$	p_1	D_1	$n-p_1$
Single factor $\boldsymbol{z}$	$g(\mu) = \lambda_0 + \boldsymbol{z}^\top \boldsymbol{\lambda}$	p_2	D_2	$n-p_2$
Two factors	$g(\mu) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}_1 + \boldsymbol{z}^\top \boldsymbol{\beta}_2$	$p_3 = p_1 + p_2 - 1$	D_3	$n-p_3$

Goodness-of-fit tests

Hypothesis	Effect to be detected	Test statistic	DF
$H_0 : \boldsymbol{\alpha} = 0$ vs. $H_1 : \boldsymbol{\alpha} \neq 0$	Effect of $\boldsymbol{x}$ ignoring $\boldsymbol{z}$	$D_0 - D_1$	$p_1 - 1$
$H_0 : \boldsymbol{\beta}_1 = 0$ vs. $H_1 : \boldsymbol{\beta}_1 \neq 0$	Effect of $\boldsymbol{x}$ with $\boldsymbol{z}$ in the model	$D_2 - D_3$	$p_1 - 1$
$H_0 : \boldsymbol{\lambda} = 0$ vs. $H_1 : \boldsymbol{\lambda} \neq 0$	Effect of $\boldsymbol{z}$ ignoring $\boldsymbol{x}$	$D_0 - D_2$	$p_2 - 1$
$H_0 : \boldsymbol{\beta}_2 = 0$ vs. $H_1 : \boldsymbol{\beta}_2 \neq 0$	Effect of $\boldsymbol{z}$ with $\boldsymbol{x}$ in the model	$D_1 - D_3$	$p_2 - 1$

Nested Model Tests – Poisson Regression for Three-Way Contingency Table



Reference

Kevin Murphy, *Machine Learning: A Probabilistic Perspective*

Annette Dobson, *An Introduction to Generalized Linear Models*, second edition

Germán Rodríguez, `https://data.princeton.edu/wws509`