# Quiz

- What are the three components of GLM?
- What is the expression of the canonical link function (in terms of $b(\theta)$)?

  $g(\mu) = b'^{-1}(\mu)$

- What is the diagram among $y_i, \mu_i, \theta_i, \eta_i$?

$$
\begin{tikzcd}
 & y_i \arrow[dl, dash] \arrow[d, dash] & \\
\theta_i \arrow[r, dash, "b'(\theta_i)=\mu_i"] \arrow[rr, bend right, dash, "g(\mu)=b'^{-1}(\mu)"'] & \mu_i \arrow[r, dash, dotted, "g(\mu_i)=\eta_i"] & \eta_i
\end{tikzcd}
$$

# Model Diagnostics

Check how well a model fits data

- Goodness-of-fit statistics
- Residuals

Compare different candidate models

- nested models
- hypothesis testing

Data: $n$ independent observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$,
$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$.
Model: GLM

- Two extreme models:
  - Null Model: Common $\mu$ for $y_1, \ldots, y_n$; only 1 parameter.
  - Full (Saturated) Model: $\mu_i = y_i$ for $i = 1, \cdots, n$; $n$ parameters.
  - the null model is too simple,
  - the full model is uninformative and not generalizable.
- We need something in between: an intermediate $p$-parameter model $(1 < p < n)$

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is $p$-dimensional.


degree of freedom

- Assume the following log-likelihood (dispersion $\phi = 1$),

$$l(y, \mu) = y\theta - b(\theta) + c(y).$$

- Let $l(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ denote the maximized log-likelihood over $\boldsymbol{\beta}$, where

$$\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$$

- The maximum possible value of the log-likelihood is $l(\boldsymbol{y}, \boldsymbol{y})$, i.e. the full (saturated) model.

- The full model fits each data point exactly.

# Deviance

- (scaled) **Deviance** measures the discrepancy between the two fits, which is twice the difference between $l(\boldsymbol{y}, \boldsymbol{y})$ and $l(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\{\underbrace{l(\boldsymbol{y}, \boldsymbol{y})}_{df=n} - \underbrace{l(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}_{df=p}\}.$$

- Deviance can be interpreted as the likelihood ratio between the full model and the $p$-parameter model.
- When the $p$-parameter model is true, the deviance *may* be approximately distributed as $\chi^2_{n-p}$.
- Deviance is commonly used to check the goodness of fit. A large value (compared to the quantile of $\chi^2_{n-p}$) means lack of fit.

# Example

- Normal linear regression: $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$ $(\epsilon_i \sim N(0, 1))$

$$l(\boldsymbol{y}, \boldsymbol{\mu}) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2}$$

$l(y, y) = -\frac{n}{2}\log(2\pi)$

$l(y, \hat{\mu}) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(y_i - x_i^T\hat{\beta})^2}{2}$

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = RSS = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2$$

- Poisson log linear regression: $\theta_i = \log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$

$\hat{\mu} = e^{x_i^T\hat{\beta}}$

$l(y, y) = -\sum_{i=1}^{n}(\log y_i! - y_i \log y_i + y_i)$

$l(y, \hat{\mu}) = -\sum_{i=1}^{n}(\log y_i! - y_i \cancel{\log} e^{x_i^T\hat{\beta}} + e^{x_i^T\hat{\beta}})$

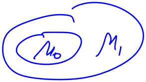
$$l(\boldsymbol{y}, \boldsymbol{\mu}) = -\sum_{i=1}^{n}(\log y_i! - y_i \log \mu_i + \mu_i)$$

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left(y_i \log y_i - y_i - y_i \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})\right)$$

# Analysis of Deviance

- Deviance can be used for model selection (comparing nested models)
- Suppose we want to compare model $M_0$ (smaller model) to model $M_1$ (larger model)

$q$    $p$




- The difference in deviances between $M_0$ and $M_1$ is

$$\overbrace{D_{M_0}}^{2(l_{full}-l_q)} - \overbrace{D_{M_1}}^{2(l_{full}-l_p)} \overset{d}{\approx} \chi^2_{p-q}, \text{ under } M_0$$

$2(l_p - l_q)$

where $\overset{d}{\approx}$ denotes "approx. distributed as".

  - Related to likelihood ratio test
  - Reject the smaller model $M_0$ is the difference in deviances is large

# Generalized Pearson's $\chi^2$ statistic

- This is another important measure of discrepancy, which takes the following form,

$$G = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i),$$

where $V(\cdot)$ is the variance function, and $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$.

- If the $p$-parameter model is true, $G$ may have an approximate distribution of $\chi^2(n-p)$.
- Both the deviance and the generalized Pearson $\chi^2$ statistic have exact $\chi^2$ distributions for normal linear models.

# Residuals

- Normal residuals: $\epsilon_i = y_i - \hat{\mu}_i$; important diagnostic tool: normality, dependence, homoscedastic.
- For GLM, we define two forms of generalized residuals:
  - **Pearson residual**
  - **Deviance residual**

# Pearson Residual

- Define Pearson residual as:

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- The raw residual scaled by the estimated sd.
- Relation to the Generalized Pearson $\chi^2$ statistic $G$:

$$G = \sum_i r_{P_i}^2.$$

- For normal dist., this reduces to the ordinary residual.

# Deviance Residual

- Define Deviance residual as:

$$r_{D_i} = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

$2\left(l(y_i, y_i) - l(y_i, \hat{\mu}_i)\right)$

where $d_i = D(y_i, \hat{\mu}_i)$.

- The deviance is $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_i d_i = \sum_i r_{D_i}^2$.
- $r_D$ is generally preferred

# GLM Model Inference

$I(\beta) = -E\left(\frac{\partial^2 \log f}{\partial \beta^2}\right) = E\left(\left(\frac{\partial \log f}{\partial \beta}\right)^2\right)$

score : $S(\beta) = \frac{\partial \log f}{\partial \beta}$

- According to general likelihood theory,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \overset{asy}{\sim} N(0, \mathcal{I}(\boldsymbol{\beta})^{-1}),$$

$I(\beta) = [\mathrm{var}(\hat{\beta})]^{-1}$

$I(\beta) = \mathrm{var}(S(\beta))$

where $\mathcal{I}(\boldsymbol{\beta})$ is the Fisher information.

- We can obtain asymptotic $100(1-\alpha)\%$ confidence intervals for $\beta_j$ using

$$\hat{\beta}_j \pm Z_{1-\alpha/2} se(\hat{\beta}_j),$$

where $Z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-th percentile of the $N(0,1)$ density.

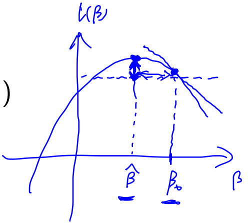
- Standard packages usually provide the estimate of $\mathcal{I}(\boldsymbol{\beta})$

# Hypothesis Tests

- Interested in testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ vs $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.
- Recall log likelihood function $l(\boldsymbol{y}, \boldsymbol{\beta})$, score vector $s(\boldsymbol{\beta})$, Fisher Information matrix $\mathcal{I}(\boldsymbol{\beta})$, and MLE $\hat{\boldsymbol{\beta}}$.
- We will introduce three asymptotically equivalent tests.
  - **Wald Test**
  - **Score Test**
  - **Likelihood Ratio Test**

- Wald test statistic:

$$TS_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$




- Score test statistic (preferred):

$$TS_S = s(\boldsymbol{\beta}_0)^T \mathcal{I}^{-1}(\boldsymbol{\beta}_0) s(\boldsymbol{\beta}_0)$$

  or sometimes replace $\mathcal{I}^{-1}(\boldsymbol{\beta}_0)$ with $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$

- Likelihood ratio test statistic (preferred):

$$TS_{LR} = 2[l(\boldsymbol{y}, \hat{\boldsymbol{\beta}}) - l(\boldsymbol{y}, \boldsymbol{\beta}_0)]$$

Under the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ and some regularity conditions, all three test statistics have asymptotic $\chi^2(p)$ distributions.

# Poisson Example

$(y_1, \cdots, y_n) \sim_{iid} Poisson(\lambda)$. We are interested in testing $H_0 : \lambda = \lambda_0$.
Questions:

- What are the expressions for different statistics?

# Poisson Example

$(y_1, \cdots, y_n) \sim_{iid} Poisson(\lambda)$. We are interested in testing $H_0 : \lambda = \lambda_0$. What are the expressions of different statistics?

**Answer:**

The problem can be viewed as a hypothesis testing problem of a null Poisson regression model. In order to obtain different test statistics, we need to calculate the key quantities first. Assume $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

$$\begin{aligned}
l(\boldsymbol{y}, \lambda) &= \sum_{i=1}^{n} [y_i \log \lambda - \lambda - \log(y_i!)] \\
s(\lambda) &= \frac{\partial l(\boldsymbol{y}, \lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{n} y_i}{\lambda} - n = \frac{n(\bar{y} - \lambda)}{\lambda} \\
\mathcal{I}(\lambda) &= \mathbb{E}\left(-\frac{\partial^2 l(\boldsymbol{y}, \lambda)}{\partial \lambda^2}\right) = \frac{\mathbb{E}(\sum_{i=1}^{n} y_i)}{\lambda^2} = \frac{n}{\lambda} \\
\hat{\lambda}_{MLE} &= \bar{y}
\end{aligned}$$

Now we can derive the expressions of different test statistics:

- **Wald:**

$$TS_W = (\hat{\lambda}_{MLE} - \lambda_0) * \mathcal{I}(\hat{\lambda}_{MLE}) * (\hat{\lambda}_{MLE} - \lambda_0) = \frac{n(\bar{y} - \lambda_0)^2}{\bar{y}}$$

- **Score:**

$$TS_S = s(\lambda_0) * \mathcal{I}^{-1}(\lambda_0) * s(\lambda_0) = \frac{n(\bar{y} - \lambda_0)^2}{\lambda_0}$$

- **LR:**

$$TS_{LR} = 2[l(\boldsymbol{y}, \hat{\lambda}_{MLE}) - l(\boldsymbol{y}, \lambda_0)] = 2n\left[\bar{y}\log\frac{\bar{y}}{\lambda_0} - (\bar{y} - \lambda_0)\right]$$

They all asymptotically follow $\chi^2_1$! We reject the null hypothesis (i.e., $\lambda = \lambda_0$) if a test statistic is too large (recall the graphical representation of different statistics).