Alternative Goodness-of-Fit Measure

know any 2 columns, know all the response data ($\because O_{i1}+O_{i2}=n_i$)

covariates \ response	$j=1$, sucess (1)	$j=2$, failure (0)	total
$i=1, \mathbf{X}=\mathbf{x}_1$	y_1 (O_{11}) $\hat{y}_1$ (E_{11})	n_1-y_1 (O_{12}) $n_1-\hat{y}_1$ (E_{12})	n_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=k, \mathbf{X}=\mathbf{x}_k$	y_k (O_{k1}) $\hat{y}_k$ (E_{k1})	n_k-y_k (O_{k2}) $n_k-\hat{y}_k$ (E_{k2})	n_k

a cell — a 2-way contingency table

- Recall: deviance is one measure of how well the model fits the data. (check in LNp.3-8&9) through y_i, $\hat{y}_i$; $\hat{y}_i$'s; y_i's
  Q: are there others?
- Pearson's X^2 statistic (for any contingency table) for a model S (i.e., $\eta_i=\sum_{l=1}^{p}\beta_l h_l(\boldsymbol{x}_i)$)
  - ➢ General definition (for J, ≥ 2, categories)

$$X_S^2=\sum_{i=1}^{k}\sum_{j=1}^{J}\left[(O_{ij}-E_{ij})^2/E_{ij}\right]$$

use difference "−" to compare O & E

related to weighting ($\because$ non-constant var. in O)

diagnal(E): $(O-E)^T\Sigma^{-1}(O-E)$

$$X_S^2=\sum_{i=1}^{k}\left\{\frac{[(n_i-y_i)-(n_i-n_i\hat{p}_i)]^2}{n_i(1-\hat{p}_i)}+\frac{(y_i-n_i\hat{p}_i)^2}{n_i\hat{p}_i}\right\}$$

where O_{ij}: observed counts for ij^{th} cell (for each cells $\overset{cf.}{\longleftrightarrow}$ deviance (LNp.3-9))
E_{ij}: expected counts for ij^{th} cell under S

- ➢ For a binomial response (J=2): $\hat{E}_S(O_{ij})$
  - ▪ number of 1's: $O_{i1}=y_i$ while $E_{i1}=n_i\hat{p}_{i,S}$ ($\hat{\eta}_i=\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, $\hat{p}_i=g^{-1}(\hat{\eta}_i)$)
  - X^2 same if change
  - ▪ number of 0's: $O_{i2}=n_i-y_i$ while $E_{i2}=n_i(1-\hat{p}_{i,S})$

for each covariate class

which results in $X_S^2=\sum_{i=1}^{k}\frac{(y_i-n_i\hat{p}_{i,S})^2}{n_i\hat{p}_{i,S}(1-\hat{p}_{i,S})}\overset{a}{\sim}\chi^2_{k-p}$ (under S)

$\hat{E}_S(y_i)$; $\widehat{Var}_S(y_i)$; same as deviance; check in LNp 3-33

RSS under WLS in LM $=\sum_i(\sqrt{w_i}y_i-\sqrt{w_i}\hat{y}_i)^2=\sum_i\frac{(y_i-\hat{y}_i)^2}{1/w_i}$ where $1/w_i\propto Var(y_i)$ (cf.)

- Some properties of the Pearson's X^2 statistic:
  - ➢ X^2 typically close in size to the deviance. Why?

check LNp.3-9

$$D=2\sum_{i=1}^{k}\{y_i\log(y_i/\hat{y}_i)+(n_i-y_i)\log[(n_i-y_i)/(n_i-\hat{y}_i)]\}$$

$f(x)=x\log\left(\frac{x}{x_0}\right)=0+(x-x_0)+\frac{1}{2x_0}(x-x_0)^2+\cdots$

$$=\cancel{2}\sum_{i=1}^{k}\{[(\cancel{y_i}-\cancel{\hat{y}_i})+(y_i-\hat{y}_i)^2/(\cancel{2}\hat{y}_i)+\cdots]$$
$$+\left[(\cancel{n_i}-\cancel{y_i}-\cancel{n_i}+\cancel{\hat{y}_i})+\frac{[(n_i-y_i)-(n_i-\hat{y}_i)]^2}{\cancel{2}(n_i-\hat{y}_i)}+\cdots\right]\}$$

O_{i1}, E_{i1}, O_{i2}, E_{i2}; ≈ 0 under S; ≈ 0 under S

$$\overset{\star}{\approx}\sum_{i=1}^{k}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}=X^2$$

Recall. In LM, $\underset{\beta}{\text{argmin}}\ RSS(\beta)$ = LS est'or = $\hat{\beta}_{MLE}$

goodness-of-fit test; deviance-based test for ω vs. Ω

  - ➢ X^2 can often be used in the same manner as the deviance e.g.
  - ➢ Alternative versions of hypothesis tests described above: use X^2 in place of the D (with same asymptotic null distribution)
    - ▪ However, some care is necessary, e.g., the model is fit (i.e., $\boldsymbol{\beta}$ is estimated) to minimize the deviance and not the Pearson's X^2 — $\hat{\beta}_{MLE}$
    - Why? $\because\star$ equivalent to maximizing log-likelihood (LNp.3-8)
    - Deviance always decreases when adding effects
    - ▪ It is possible, although unlikely, that the X^2 could increase as an effect is added to the model
    - possible that $X^2(\hat{\beta}_{MLE,\omega})<X^2(\hat{\beta}_{MLE,\Omega})$

Define $D(\beta)$. Define $X^2(\beta)$ and $\hat{\beta}_p=\underset{\beta}{\text{argmin}}\ X^2(\beta)$. Then, $X^2(\hat{\beta}_{p,\omega})\geq X^2(\hat{\beta}_{p,\Omega})$ (cf)

- We can define *Pearson residuals* under model S as:

Deviance residuals: defined from ★ in LNp.3-9

defined for each i, i.e., covariate class, not each cell

$$r^P_{i,S} = (y_i - \hat{E}_S(y_i))/\sqrt{\hat{\text{var}}_S(y_i)}$$
$$= (y_i - n_i\hat{p}_{i,S})/\sqrt{n_i\hat{p}_{i,S}(1-\hat{p}_{i,S})}$$

Recall residuals and diagnostics in LM

which can be viewed as a type of standardized residual

can be used to do diagnostic, e.g., outlier detection, residual plot, ...

preferred in residual plot

check LNp.3-32

  - $X^2_S = \sum_{i=1}^k (r^P_{i,S})^2$
  - So, Pearson's X^2 is analogous to the residual sum of squares used in Normal linear model

$D \approx X^2 \longleftrightarrow$ RSS in LM

- generalized R^2

cf. $\hat{\sigma}^2 = \frac{RSS}{n-p}$ in LM $\longleftrightarrow D, X^2$ in GLM

  - Recall: R^2 for Normal linear model is a popular goodness-of-fit measure, which represents the *proportion of variance explained*.

unit free

Q: how to generalize this concept to binomial GLM?

  - For a data with K Bernoulli ungrouped data z_j's and k binomial grouped data y_i's, the parameters p's (and η's) in the models of z_j's and y_i's are: $z_j \leftrightarrow p_j \leftrightarrow \eta_j,\ j=1,\ldots,K,$ $y_i \leftrightarrow p_i \leftrightarrow \eta_i,\ i=1,\ldots,k.$

use covariate class

more parameters

cf. $\because$ z_j's in same covariate class have same p & η

$k \le K$

Consider the following models:

irrelevant to the covariates

- null model: *all* η's are equal (intercept-only model) ← dim = 1
- full model: *all* the η's are free to vary

• For z_j's, η_j's $\in \mathbb{R}^K$, dim = K

• For y_i's, η_i's $\in \mathbb{R}^k$, dim = k

depend on covariates

- A model S ($\eta_i = \sum_{l=1}^p \beta_l h_l(\boldsymbol{x}_i)$) in which *all* η's are restricted to fall on a linear function of effect codings and their coefficients β

dim = $p \le k$

Deviance ↓ when adding effects

$D_S = 2\log\left(\frac{\hat{\mathcal{L}}_{full}}{\hat{\mathcal{L}}_S}\right)$

Then, null model $\subseteq$ model S $\subseteq$ full model

D_{full} D_S D_{null}

cf.

  - Approach 1: proportion of deviance explained

Q ① $z_j = \hat{z}_j$, $y_i = \hat{y}_i$ ② check its equation in LNp.3-9

always between 0 & 1

$$\frac{D_S - D_{null}}{D_{full} - D_{null}} = \frac{D_{null} - D_S}{D_{null}} \overset{★}{=} 1 - \frac{D_S}{D_{null}}$$

cf. $1 - \frac{RSS}{TSS} = R^2$ in LM

RSS_S, RSS_{null}

  - Approach 2: interpretation of R^2 from likelihood viewpoint

$\mathcal{L}$ → can be generalized to any distributions

(exercise)

$\hat{\sigma}^2_{MLE} = \frac{RSS}{K}$

$\hat{\mathcal{L}} \propto \hat{\sigma}_{MLE}^{-K}$

$\hat{\mathcal{L}}^{-2/K} \propto \hat{\sigma}^2_{MLE}$

$\hat{\mathcal{L}}^{-2/K} \propto RSS$

  - Recall. In Normal Linear model,

geometric mean

$$R^2_S = 1 - (\hat{\mathcal{L}}_0/\hat{\mathcal{L}}_S)^{2/K},$$ where

likelihood-ratio test ←→ overall F-test

its F-statistic is a function of R^2

- K: number of observations
- $\hat{\mathcal{L}}_0$: maximized likelihood of K obs. under *null* model
- $\hat{\mathcal{L}}_S$: maximized likelihood of K obs. under model S

- Generalization to binomial case: For y_i's, $\mathcal{L}=\prod_{i=1}^{B}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}$
  - $K=n_1+\cdots+n_B$ → K: number of Bernoulli observations z_j's → $\mathcal{L}=\prod_{j=1}^{K}p_j^{z_j}(1-p_j)^{1-z_j}$
  - $\hat{\mathcal{L}}_0$: the maximized likelihood of K obs. under *null* model
  - $\hat{\mathcal{L}}_S$: the maximized likelihood of K obs. under model S

  Why not use the likelihood of y_i's? → z_j's, not y_i's

- However, Nagelkerke (1991) pointed out that for *discrete models*, i.e., models whose likelihood is a product of probabilities, instead of densities,

  $0\le$ prob. ≤ 1; $0\le$ pdf $<\infty$

$$1\ \ge\ \max_S R_S^2\ =\ 1-(\hat{\mathcal{L}}_0/\hat{\mathcal{L}}_{\text{full}})^{2/K}$$
$$=\ 1-(\hat{\mathcal{L}}_0)^{2/K}\ \neq 1$$

Under full, for z_j's, $\hat{p}_j=\begin{cases}1, & \text{if } z_j=1\\ 0, & \text{if } z_j=0\end{cases}$ $\mathcal{L}_{\text{full}}=\prod_{j=1}^{K}0^0 1^1=1$

For y_i's, $\hat{p}_i=y_i/n_i$, $\mathcal{L}_{\text{full}}<1$ in most cases

For normal data, $\mathcal{L}_{\text{full}}=\infty$ ∵ $y_i=\hat{y}_i\Rightarrow RSS=0$; >0

Recall LM data with replicates. ⇒ R^2 cannot reach 1. z_j's in same covariate class is like replicates; cannot reach 1

- Nagelkerke (1991) suggested (for model S)

$$\bar{R}_S^2\ =\ R_S^2/(\max R^2)=[1-\overbrace{(\hat{\mathcal{L}}_0/\hat{\mathcal{L}}_S)^{2/K}}^{\exp(\log())}]/[1-(\hat{\mathcal{L}}_0)^{2/K}]$$
$$=\ \frac{1-\exp\big((D_S-D_{\text{null}})/K\big)}{1-\exp\big(-D_{\text{null}}/K\big)}\ \Rightarrow 0\le\bar{R}_S^2\le 1$$

can reach 1

cf. ★ in LN p. 3-35

➢ For sparse data (i.e., n_i is small), it is quite common to see low value of R^2 even when the model is a good fit.

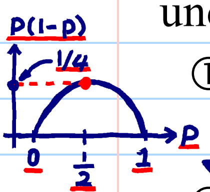
e.g. n_i's $=1$, y_i's $=0$ or 1, But, $0\le\hat{p}_i\le1$

❖ **Reading**: Faraway (1st ed.), 2.9

Over-dispersion

only offer a structure for mean, no statement about variance of y_x

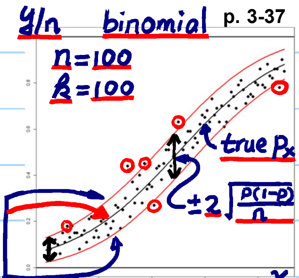
- **Q**: What is overdispersion?
  - ➢ Recall 1: If data really follows a binomial GLM model, i.e., $y_{\mathbf{x}}\sim B(n_{\mathbf{x}},p_{\mathbf{x}}=g^{-1}(\eta_{\mathbf{x}}))$, where under model S, $\eta_{\mathbf{x}}=\sum_{l=1}^{p}\beta_l\,h_l(\mathbf{x})$, then
    - ① $E(y_{\mathbf{x}}/n_{\mathbf{x}})=p_{\mathbf{x}}$, and → know mean ⇒ know variance
      $Var(y_{\mathbf{x}}/n_{\mathbf{x}})=p_{\mathbf{x}}(1-p_{\mathbf{x}})/n_{\mathbf{x}}$
    - ② D_S (deviance) $\overset{a}{\sim}\chi^2_{k-p}$ (under S)
      ⇒ can perform goodness-of-fit test

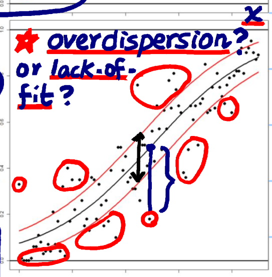


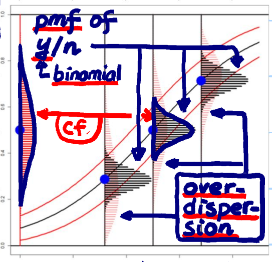
GofF, H_0: S is correct. cf.

mean structure of y carry no information about var of y

Recall lack-of-fit test (LM, LN p6-6~10)



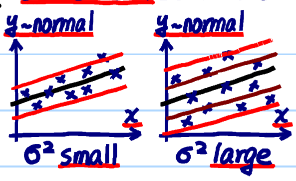




- ➢ **Q**: Why no goodness-of-fit test for Normal linear model if no further information about σ^2 is available?

  e.g., σ^2 known or data with replicates

  same mean structure



- ➢ Recall 2: $D_S\approx X_S^2=\sum_i\big(r^P_{i,S}\big)^2$ standardized residuals

  In LM: $E(y)=X\beta$, $Var(y)=\sigma^2$ ← different parameters

  say, GofF rejected

- ➢ **Q**: what cause large D_S? ⇒ may suspect $Var(y_{\mathbf{x}})\gg n_{\mathbf{x}}p_{\mathbf{x}}(1-p_{\mathbf{x}})$ which is referred to as over-dispersion compared to variance of binomial with mean n_xp_x
  - ⇒ if true, $y_{\mathbf{x}}$'s no more ~ binomial even $E(y_{\mathbf{x}})=n_{\mathbf{x}}p_{\mathbf{x}}$ (S) holds

y more possible to generate value far away from its mean