# Applied Statistical Methods II

Chapter #14

Logistic Regression, Poisson Regression, and Generalized Linear Models

Part III

# Goodness-of-Fit

- What about a global measure of model fit?
  - Goodness-of-Fit tests $H_0$ : Model fits well.
  - Linear regression: had the F-tests.
- Have three main tests of fit for logistic regression:
  - Pearson's $\chi^2$
  - Deviance test (aka Likelihood ratio)
  - Hosmer-Lemeshow
- Pearson's and the Deviance are only good if you have many repeats for covariate patterns.

# Political Protest Example

- 360 Columbia University class of 1969 alum are sampled.
- They are asked:
  1. Their current political affiliation with 1= strongly democrat and 7=strongly republican (ordered). In our example, we use that as one covariate $x$ taking 7 ordered values, not treated as categorical.
  2. Asked if they participated in any political protests in college.
- The observed data are summarized in the following table:

# Protest Data

| Party Identification | Protestors | Non Protestors |
|---|---|---|
| Strong Democrat | 10 | 18 |
| Weak Democrat | 59 | 38 |
| Leaning Democrat | 41 | 22 |
| Independent | 26 | 7 |
| Leaning Republican | 44 | 10 |
| Weak Republican | 47 | 7 |
| Strong Republican | 29 | 2 |

# Pearson's $\chi^2$

- Assume that we have repeated observations at "c" different covariate patterns.
- Let $Y_{ij}$ be the $i^{th} = 1, \ldots, n_j$ subject with $j$-th covariate pattern $X_{j1}, \ldots, X_{j(p-1)}, j = 1, \ldots, c$.
    - We will assume $Y_{ij} \mid X_{j1}, \ldots, X_{j(p-1)} \sim Ber\left(\pi_j\right)$ are **independent** across $i, j$.
- $O_{j1} = \sum_{i=1}^{n_j} Y_{ij}$ and $O_{j0} = n_j - O_{j1}$
- Note that $O_{j1} \sim Binomial(n_j, \pi_j)$ and $O_{j0} \sim Binomial(n_j, 1 - \pi_j)$. Why?
- You fit the regression model $\text{logit}(\pi_i) = \beta_0 + \cdots + \beta_{p-1} X_{i(p-1)}$ to get $\hat{\pi}_1, \ldots, \hat{\pi}_c$.
- Then the expected values of $O_{j1}$ and $O_{j0}$ are:
    - $E_{j1} = n_j \hat{\pi}_j$ and $E_{j0} = n_j - E_{j1}$

# Pearson's $\chi^2$, cont.

- Idea: compare the observed values of $O_{j1}$ and $O_{j0}$ to what you would expect under the regression model $\text{logit}(\pi_j) = \beta_0 + \cdots + \beta_{p-1} X_{j(p-1)}$.
- Pearson's statistic compares the observed numbers to the expected estimated from the regression model
$$X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{\left(O_{jk} - E_{jk}\right)^2}{E_{jk}} = n \sum_{j=1}^{c} \frac{\left(O_{j1}/n - E_{j1}/n\right)^2}{(E_{j1}/n)(1 - E_{j1}/n)}$$
- Under the null $H_0 : \text{logit}(\pi_i) = \beta_0 + \cdots + \beta_{p-1} X_{i(p-1)}$
    - $X^2$ is asymptotically $\chi^2_{c-p}$.
    - Reject the model for large values of $X^2$
    - Distribution only holds if $n_j$ is large for all j.

# Deviance Test

- Also assumes you have repeated covariate patterns.
- DEV is the likelihood ratio test of the current model to the saturated model.
- Under the saturated model
  - Observations from one group have no effect on the estimated mean of another group.
  - $\hat{\pi}_{ij} = O_{j1}/n_j$

# Deviance Test, cont.

- The likelihood ratio test first fits the proposed model $\text{logit}(\pi_{ij}) = \beta_0 + \cdots + \beta_{p-1} X_{j(p-1)}$
  - Get the log-likelihood $\ell(M)$.
- Fit the saturated model to get the log-likelihood $\ell(F)$.
- Form the statistics $Dev = -2\left[\ell(M) - \ell(F)\right]$.
- When $n_j$ are large, $Dev$ is approximately $\chi^2_{c-p}$.

# Deviance Table

- Deviance is additive and can be used to compare the fit of a sequence of nested models.
- Based on the likelihood ratio tests.
- This serves the analogous function to the F-test that is used in the normal model.

*Deviance table ("ANOVA table") :*

| Sequence of models | Deviance (e.g.) | Diff | df | $\chi^2_p$ $(p = 1)$ |
|---|---|---|---|---|
| $X_1, X_2, X_3$ | 50 | | | |
| $X_1, X_2$ | 70 | 20 | 1 | $P < 10^{-3}$ |
| $X_1$ | 100 | 30 | 1 | . |
| 0 | 200 | 100 | 1 | . |

# Protest Data

| Party Identification | Protestors | Non Protestors |
|---|---|---|
| Strong Democrat | 10 | 18 |
| Weak Democrat | 59 | 38 |
| Leaning Democrat | 41 | 22 |
| Independent | 26 | 7 |
| Leaning Republican | 44 | 10 |
| Weak Republican | 47 | 7 |
| Strong Republican | 29 | 2 |

# Political Protest Example

- $n = 360$, $c = 7$, $p = 2$, $n_1 = O_{11} + O_{10} = 10 + 18 = 28$ and so on.
- Fit a logistic model: there is a linear relationship between the logit of the probability of protesting and political leaning.
  - $Y_{ij} \mid X_{j1}, \ldots, X_{j(p-1)} \sim Ber(\pi_j)$, $j = 1, \ldots, c$.
  - $\text{logit}(\pi_j) = \beta_0 + \beta_1 X_{j1} + \cdots + \beta_{p-1} X_{j(p-1)}$
- Do a goodness-of-fit test to see if the logistic model fits well.

# Entering Aggregated Data

- Given this table, you would not want to enter a row for each subject.
- What cell subject $j$ falls into does not matter.
- Only the sufficient statistics $O_{j1}$, $O_{j0}$ are important.

# Goodness-of-Fit For the Crab Data

- We will consider the crab data set and the model:
  1. Color is linearly associated with having any satellites.
- Color=1,2,3,4 goes from light to dark.
- y=1 if satellites are present and 0 otherwise.

# Hosmer-Lemeshow Test

- The Hosmer-Lemeshow test is good for continuous variables.
- Fit a logistic regression and get $\hat{\pi}_i$ for $i = 1, \ldots, N$ and sort in increasing order.
- An algorithm is used to combine the observations into $c$ (usually 10) groups based on the closeness of $\hat{\pi}_i$.
  - Idea: under $H_0$ : logistic model is correct, $\pi(x_i) \approx \pi(x_j)$ if $\|x_i - x_j\|$ is small.
  - Choice of $c$ will depend on $p$. Ideally, $c \propto K^{(p-1)} \Rightarrow$ really only useful for small problems.
  - See *Applied Logistic Regression* by D.W. Hosmer and S. Lemeshow for details.
- Perform a Pearson's $\chi^2$ in the binned data.
- The distribution can be approximated by $\chi^2$ with df = $c - 2$.

# SAS Examples for HL test

- Hosmer-Lemeshow test is specific to logistic regression - GENMOD does not do it.
- Must use Proc LOGISTIC.
- We will consider a model that looks at the linear relationship between width and having any satellites for the Crab data, and also look at the IPO example from the book.

# IPO example from the text

- Study of 482 initial public offering companies (IPOs)
- $Y_i = 1$ if financed by venture capital funds.
- $X_i =$ face value of the company.
- Let's look at SAS examples.

# Residuals

- Since $Y_i$ is 0 or 1, the raw residuals are not that helpful.
- There are two main types of residuals that are common for identifying model fit and outliers.
  - Pearson's
  - Deviance
- Person's Residuals:
- $r_{p_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$
- Is called Pearson's Residuals since $X^2 = \sum_{i=1}^{n} r_{p_i}^2$

- The studentized residual divides by the square root of 1 - the diagonal of the hat matrix.
  $r_{s_i} = \frac{r_{p_i}}{\sqrt{1-h_{ii}}}$
  - $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$.
  - $W^{1/2}$ is the diagonal matrix of $\sqrt{\pi_i(1-\pi_i)}$.
  - Derivation (this argument applies for all GLMs using the canonical link): From HW2, there exists a function $K(\theta)$ such that $K'(\theta) = E_\theta(y)$ and $K''(\theta) = \mathrm{Var}_\theta(y)$.

$$\hat{\pi}_i = K'\left(x_i^T\hat{\beta}\right) \approx K'\left(x_i^T\beta\right) + K''\left(x_i^T\beta\right)x_i^T\left(\hat{\beta}-\beta\right)$$

$$\hat{\beta}-\beta \approx -(X^TWX)^{-1}X^T(Y-\pi)$$

$$\Rightarrow Y - \hat{\pi} \approx \left(I_n - WX(X^TWX)^{-1}X^T\right)(Y-\pi)$$

Therefore, $\mathrm{Var}(Y-\pi) \approx W^{1/2}(I_n - H)W^{1/2}$

# Deviance Residuals

- For independent data, can write
  $0 <$ deviance of the model $= \sum_{i=1}^{n} dev_i^2$ .
  - If model is correct, and since the model is a sub-model of the *saturated model* (each data point $Y_i$ has a parameter), then $\sum_{i=1}^{n} dev_i^2 \approx \chi^2_{n-p}$.
  - Akin to residual sum of squares in linear regression.
- The deviance residual is
  $dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\left[Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)\right]}$.
- Also has a standardized version: $\frac{dev_i}{\sqrt{1-h_{ii}}}$.
- Usually the deviance residuals are preferred to Pearson's.
  - Neither one of them has as nice of properties as the residuals for linear reg.
  - Pearson's can be skewed for reasons that do not relate to the fit of the model.

# Deviance Residuals

- For independent data, can write
  $0 <$ deviance of the model $= \sum_{i=1}^{n} dev_i^2$ .
  - If model is correct, and since the model is a sub-model of the *saturated model* (each data point $Y_i$ has a parameter), then $\sum_{i=1}^{n} dev_i^2 \approx \chi^2_{n-p}$.
  - Akin to residual sum of squares in linear regression.
- The deviance residual is
  $dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\left[Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)\right]}$.
- Also has a standardized version: $\frac{dev_i}{\sqrt{1-h_{ii}}}$.
- Usually the deviance residuals are preferred to Pearson's.
  - Neither one of them has as nice of properties as the residuals for linear reg.
  - Pearson's can be skewed for reasons that do not relate to the fit of the model.

# Use of the Residuals

- The residuals can be used to identify overall model fit and to find influential observations.
- Based on a plot of either the Studentized Pearson's or Deviance residual vs. the fitted value.
- For large sample sizes: $E\left(residual\right) = 0$.
- A loess plot through the residual plot should be close to a straight line.
- In linear regression, could preform outlier tests.
- In logistic regression, it is much more subjective.
    - Look for extremes and their influence on the fit.

# Generalized Additive Models

- In linear regression we have additive models $E(y) = f_1(x_1) + f_2(x_2) + \cdots + f(x_p)$, where $f$ functions can be estimated by smoothing.
- In logistic regression, $Y_i$ is 0 or 1, so direct smoothing approaches do not work well.
- GLM $\to$ GAM $\to$ GLM
  - $\text{logit}(\mu) = f_1(x_1) + f_2(x_2) + \cdots + f(x_p)$.
  - check if $f$ functions are near linear, can be used as a model building tool for GLM.
  - use R function "gam". I will post an example on Canvas.