Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment

Georgia Tech

About This Lesson

Georgia Tech

Logistic Regression Model

Data:

$\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

Assumptions:

- *Linearity Assumption*: $\mathrm{g}\left(\mathrm{p}\left(X_1, \cdots, X_p\right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- *Independence Assumption*: $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are independent random variables
- *Logit Link Function:* $\mathrm{g}(\mathrm{p}) = \ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)$

There is no error term!
How to check the model assumptions?

Georgia Tech

Residuals in Logistic Regression

Data:

$\{(\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{1,p}), \boldsymbol{Y}_1\}, \{(\boldsymbol{X}_{2,1}, \boldsymbol{X}_{2,2}, \cdots, \boldsymbol{X}_{2,p}), \boldsymbol{Y}_2\}, \cdots, \{(\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \cdots, \boldsymbol{X}_{n,p}), \boldsymbol{Y}_n\}$
where $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are *binary* responses

- **Logistic Regression Without Replications**
  - One separate (possibly non-unique) set of predictors $(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})$ for each individual observation $\boldsymbol{Y}_i$ (i=1,..,n)
  - $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim$ Bernoulli$\left(\mathrm{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right)$ or
    $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim$ Binomial$\left(1, \mathrm{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right)$
- **Logistic Regression With Replications**
  - Observe n_i repeated responses $\boldsymbol{Y}_i$ for each unique set of predictors $(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p})$ across all i=1,..,n
  - $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim$ Binomial$\left(n_i, \mathrm{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right)$, $n_i > 1$

Georgia Tech

Residuals in Logistic Regression

Logistic Regression With Replications

- $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim \text{Binomial}\left(n_i, \text{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right),\ n_i > 1$
- Estimated probabilities are:

$$\hat{\text{p}}_i = \hat{\text{p}}_i\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right) = \frac{e^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}$$

- Pearson residuals:

$$r_i = \frac{Y_i - n_i\hat{\text{p}}_i}{\sqrt{n_i\hat{\text{p}}_i(1-\hat{\text{p}}_i)}}$$

- Deviance residuals:

$$\text{d}_i = \text{sign}\left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}_i\right)\sqrt{2\boldsymbol{Y}_i\log\left(\boldsymbol{Y}_i/\widehat{\boldsymbol{Y}}_i\right) + 2(n_i - \boldsymbol{Y}_i)\log\left((n_i - \boldsymbol{Y}_i)/\left(n_i - \widehat{\boldsymbol{Y}}_i\right)\right)}$$

Georgia Tech

Residuals in Logistic Regression

Logistic Regression With Replications

- $\boldsymbol{Y}_i \mid \boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p} \sim \text{Binomial}\left(n_i, \text{p}\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right)\right),\ n_i > 1$
- Estimated probabilities are:

$$\hat{\text{p}}_i = \hat{\text{p}}_i\left(\boldsymbol{X}_{i,1}, \cdots, \boldsymbol{X}_{i,p}\right) = \frac{e^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1}\boldsymbol{X}_{i,1} + \cdots + \widehat{\beta_p}\boldsymbol{X}_{i,p}}}$$

- Pearson residuals:

$$r_i = \frac{Y_i - n_i\hat{\text{p}}_i}{\sqrt{n_i\hat{\text{p}}_i(1-\hat{\text{p}}_i)}}$$

- Deviance residuals:

$$\text{d}_i = \text{sign}\left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}_i\right)\sqrt{2\boldsymbol{Y}_i\log\left(\boldsymbol{Y}_i/\widehat{\boldsymbol{Y}}_i\right) + 2(n_i - \boldsymbol{Y}_i)\log\left((n_i - \boldsymbol{Y}_i)/\left(n_i - \widehat{\boldsymbol{Y}}_i\right)\right)}$$

- Pearson's residuals follow directly a normal approximation to a binomial, hence approximately *N(0,1)*.
- The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model vs. the fitted model, thus approximately *N(0,1)* if the model is a good fit.

Georgia Tech

Model Goodness of Fit

GOF Visual Analytics

- Normal probability plot & histogram of the residuals

Hypothesis Testing Procedure

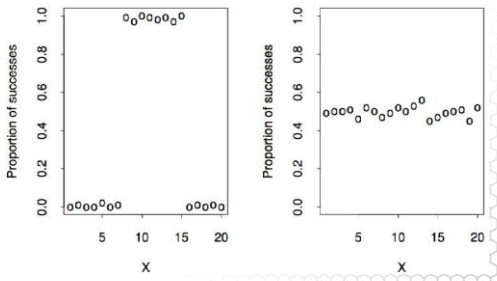
- $\mathbf{H_0}$: **the logistic model fits the data** vs.
  $\mathbf{H_a}$: **the logistic model does not fit the data**
- Deviance test statistic: $\mathrm{D} = \sum_{i=1}^{n} d_i^2$
  - Under $\mathbf{H_0}$, $\mathrm{D} \sim \chi^2_{n-p-1}$
- Reject $\mathbf{H_0}$ if $\mathrm{P-value} = \Pr(\chi^2_{df} > \mathrm{D})$ is small
- For this test we want large p-values!!!!

Georgia Tech

Goodness of Fit vs. Predictive Power

- **Goodness of fit**: Model assumptions hold
  - For example, does the S-shape logit function fit the data?
- **Predictive Power**: The predicting variables predict the data
  - Even if the one or more assumptions do not hold

While the logistic model is a sensible one for probabilities, it is not necessarily appropriate for any particular data set. That does not mean that the predicting variables are not good predictors of the probability of success.



Georgia Tech

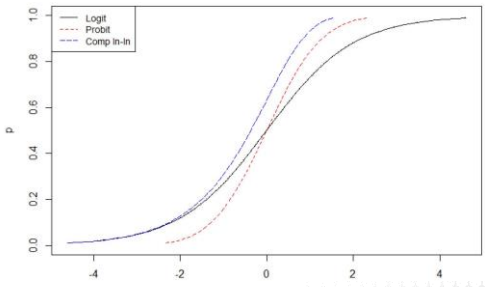
What If No Goodness of Fit?

- Add predicting variables
- Transform predicting variables to improve linearity
- Identify unusual observations (outliers, leverage points)
- The binomial distribution may not be appropriate
  - Overdispersion: The variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
- Logit function may not fit data
  - Other S-shape functions
    - probit, complementary log-log (c-log-log)

Georgia Tech

What If No Goodness of Fit?

- Add predicting variables
- Transform predicting variables to improve linearity
- Identify unusual observations (outliers, leverage points)
- The binomial distribution may not be appropriate
  - Overdispersion: The variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
- Logit function may not fit data
  - Other S-shape functions
    - probit, complementary log-log (c-log-log)



Georgia Tech

What If No Goodness of Fit?

- Add predicting variables
- Transform predicting variables to improve linearity
- Identify unusual observations (outliers, leverage points)
- The binomial distribution may not be appropriate
  - Overdispersion: The variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
- Logit function may not fit data
  - Other S-shape functions
    - probit, complementary log-log (c-log-log)

Why logistic regression?

- Logit link function is the canonical link function
- Ease of interpretation of the regression coefficients

Georgia Tech

Summary

Georgia Tech