

Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment



About This Lesson



Logistic Regression Model

Data:

$\{(X_{1,1}, X_{1,2}, \dots, X_{1,p}), Y_1\}, \{(X_{2,1}, X_{2,2}, \dots, X_{2,p}), Y_2\}, \dots, \{(X_{n,1}, X_{n,2}, \dots, X_{n,p}), Y_n\}$
 where Y_1, \dots, Y_n are **binary** responses

Assumptions:

- **Linearity Assumption:** $g(p(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$
- **Independence Assumption:** Y_1, \dots, Y_n are independent random variables
- **Logit Link Function:** $g(p) = \ln\left(\frac{p}{1-p}\right)$

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



Residuals in Logistic Regression

Data:

$\{(X_{1,1}, X_{1,2}, \dots, X_{1,p}), Y_1\}, \{(X_{2,1}, X_{2,2}, \dots, X_{2,p}), Y_2\}, \dots, \{(X_{n,1}, X_{n,2}, \dots, X_{n,p}), Y_n\}$
 where Y_1, \dots, Y_n are **binary** responses

- **Logistic Regression Without Replications**
 - One separate (possibly non-unique) set of predictors $(X_{i,1}, \dots, X_{i,p})$ for each individual observation Y_i ($i=1, \dots, n$)
 - $Y_i | X_{i,1}, \dots, X_{i,p} \sim \text{Bernoulli}(p(X_{i,1}, \dots, X_{i,p}))$ or
 $Y_i | X_{i,1}, \dots, X_{i,p} \sim \text{Binomial}(1, p(X_{i,1}, \dots, X_{i,p}))$
- **Logistic Regression With Replications**
 - Observe n_i repeated responses Y_i for each unique set of predictors $(X_{i,1}, \dots, X_{i,p})$ across all $i=1, \dots, n$
 - $Y_i | X_{i,1}, \dots, X_{i,p} \sim \text{Binomial}(n_i, p(X_{i,1}, \dots, X_{i,p}))$, $n_i > 1$



Residuals in Logistic Regression

Logistic Regression With Replications

- $Y_i | X_{i,1}, \dots, X_{i,p} \sim \text{Binomial}(n_i, p(X_{i,1}, \dots, X_{i,p}))$, $n_i > 1$
- Estimated probabilities are:

$$\hat{p}_i = \hat{p}_i(X_{i,1}, \dots, X_{i,p}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_p X_{i,p}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_p X_{i,p}}}$$

- Pearson residuals:

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

- Deviance residuals:

$$d_i = \text{sign}(Y_i - \hat{Y}_i) \sqrt{2Y_i \log(Y_i / \hat{Y}_i) + 2(n_i - Y_i) \log((n_i - Y_i) / (n_i - \hat{Y}_i))}$$



Residuals in Logistic Regression

Logistic Regression With Replications

- $Y_i | X_{i,1}, \dots, X_{i,p} \sim \text{Binomial}(n_i, p(X_{i,1}, \dots, X_{i,p}))$, $n_i > 1$
- Estimated probabilities are:

$$\hat{p}_i = \hat{p}_i(X_{i,1}, \dots, X_{i,p}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_p X_{i,p}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \dots + \hat{\beta}_p X_{i,p}}}$$

- Pearson residuals:

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

- Deviance residuals:

$$d_i = \text{sign}(Y_i - \hat{Y}_i) \sqrt{2Y_i \log(Y_i / \hat{Y}_i) + 2(n_i - Y_i) \log((n_i - Y_i) / (n_i - \hat{Y}_i))}$$

- 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.



Model Goodness of Fit

GOF Visual Analytics

- Normal probability plot & histogram of the residuals

Hypothesis Testing Procedure

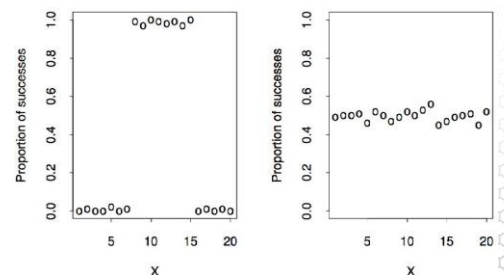
- H_0 : the logistic model fits the data vs.
 H_a : the logistic model does not fit the data
- Deviance test statistic: $D = \sum_{i=1}^n d_i^2$
 - Under H_0 , $D \sim \chi^2_{n-p-1}$
- Reject H_0 if P-value = $\Pr(\chi^2_{df} > D)$ is small
- For this test we want large p-values!!!!



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.

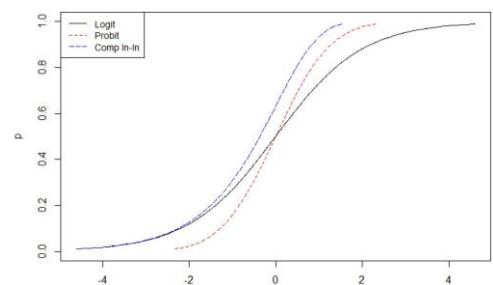


What If No Goodness of Fit?

- 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)

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)



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

Summary

