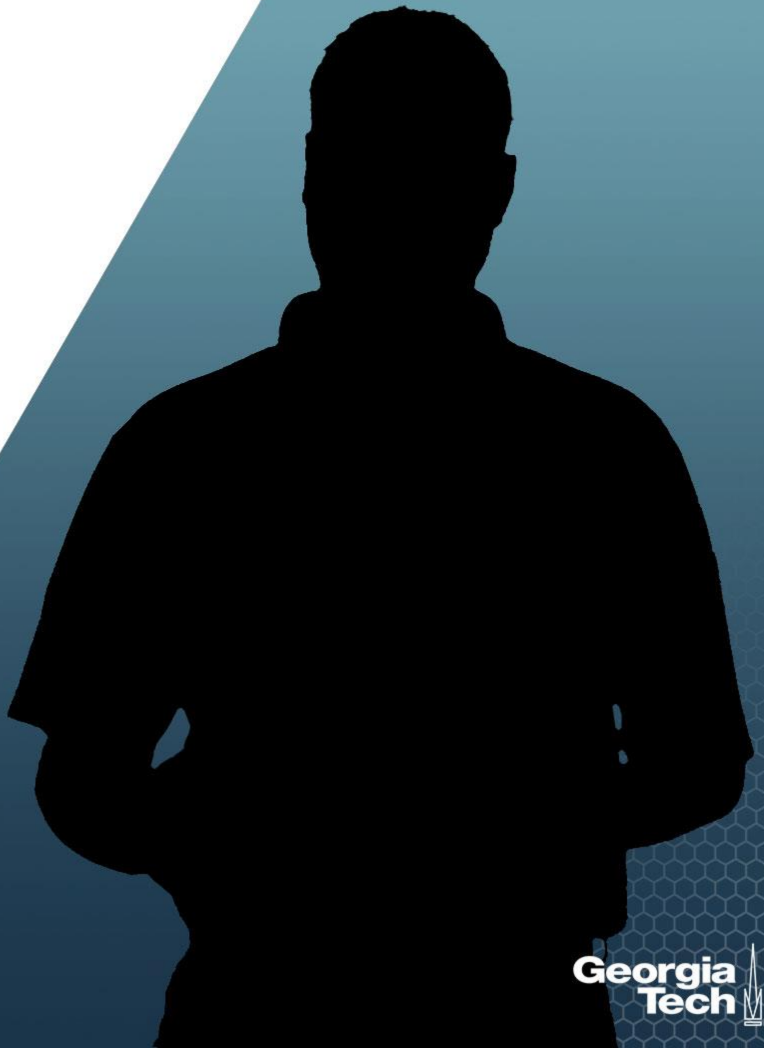
Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

Stewart School of Industrial and Systems Engineering



Statistical Inference

About This Lesson

Georgia Tech

Model Estimation

Model the probability of success given predictor(s):

$$\text{Logit}\Big(\Pr\big(Y = 1 \mid X_1, \cdots, X_p\big)\Big) = \text{Logit}\Big(\mathrm{p}\big(X_1, \cdots, X_p\big)\Big) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Parameters: $\beta_0, \beta_1, \cdots, \beta_p$

Approach: Maximum Likelihood Estimation

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \mathcal{L}(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i=1}^{n} \mathrm{p}\big(\boldsymbol{X}_{i,1}, \boldsymbol{X}_{i,2}, \cdots, \boldsymbol{X}_{i,p}\big)^{\boldsymbol{Y}_i} \Big(1 - \mathrm{p}\big(\boldsymbol{X}_{i,1}, \boldsymbol{X}_{i,2}, \cdots, \boldsymbol{X}_{i,p}\big)\Big)^{1-\boldsymbol{Y}_i}$$

or

$$\max_{\beta_0,\beta_1,\cdots,\beta_p} \ell\big(\beta_0, \beta_1, \cdots, \beta_p\big) = \max_{\beta_0,\beta_1,\cdots,\beta_p} \log\big(\mathcal{L}(\beta_0, \beta_1, \cdots, \beta_p)\big)$$

$$= \max_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} \left(\boldsymbol{Y}_i \log\left(\frac{\mathrm{e}^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + \mathrm{e}^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}\right) + (1 - \boldsymbol{Y}_i)\log\left(\frac{1}{1 + \mathrm{e}^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}\right)\right)$$

Statistical Inference

Maximum Likelihood Estimators (MLEs):

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_\mathrm{p})$$

Statistical Properties of MLEs:

- Approximate Sampling Distribution: $\hat{\boldsymbol{\beta}} \approx \boldsymbol{N}(\boldsymbol{\beta}, \boldsymbol{V})$
- The normal approximation relies on the assumption of large sample size
- Statistical inference is not reliable for small sample data

1-*α* Approximate Confidence interval $\left\{ \hat{\beta}_j \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta}_j)} \right.$

Statistical Inference (cont'd)

- Hypothesis testing and Confidence Intervals rely on the approximately normal distribution of large sample sizes

- Use the z-test (**Wald test**)
  - Test is for the statistical significance of $\hat{\beta}_j$ *given all other predicting variables in the model*
  - Null hypothesis is that β_j is not significant

    $$\mathrm{H_0}: \beta_j = 0 \;\text{ vs. }\; \mathrm{H_a}: \beta_j \neq 0$$

  - $\mathrm{z{-}value} = \dfrac{\hat{\beta}_j - 0}{\mathrm{se}(\hat{\beta}_j)} = \dfrac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$
  - Reject $\mathrm{H_0}$ if $|\mathrm{z{-}value}|$ is too large
    - Implies that β_j is statistically significant

Statistical Inference (cont'd)

For $\mathbf{H_0}$: $\boldsymbol{\beta_j} = \boldsymbol{b}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} \neq \boldsymbol{b}$ (to test if the coefficient equals constant b)

- $\text{z−value} = \frac{\hat{\beta}_j - b}{\text{se}(\hat{\beta}_j)}$
- Reject H_0 if $|\text{z−value}| > \text{z}_{\alpha/2}$ for significance level α
- Alternatively, compute P-value
  - P−value = 2Pr$(Z > |\text{z−value}|)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j} \leq \mathbf{0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} > 0$ (to test for a significantly positive coefficient)

  - P−value = Pr$(Z > z - value)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j} \geq \mathbf{0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} < \mathbf{0}$ (to test for a significantly negative coefficient)

  - P−value = Pr$(Z < z - value)$

Georgia Tech

Statistical Inference (cont'd)

For $\mathbf{H_0}$: $\boldsymbol{\beta_j} = \boldsymbol{b}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} \neq \boldsymbol{b}$ (to test if the coefficient equals constant b)

- $\text{z−value} = \frac{\hat{\beta}_j - b}{\text{se}(\hat{\beta}_j)}$
- Reject H_0 if $|\text{z−value}| > \mathrm{z}_{\alpha/2}$ for significance level α
- Alternatively, compute P-value
  - P−value = 2Pr$(Z > |\text{z−value}|)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j} \leq \mathbf{0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} > 0$ (to test for a significantly positive coefficient)

- P−value = Pr$(Z > z - value)$

For $\mathbf{H_0}$: $\boldsymbol{\beta_j} \geq \mathbf{0}$ vs. $\mathbf{H_a}$: $\boldsymbol{\beta_j} < \mathbf{0}$ (to test for a significantly negative coefficient)

- P−value = Pr$(Z < z - value)$

- Because the approximation of the normal distribution relies on large sample size, so do the hypothesis testing procedures.
- What if *n* is small?
  - The hypothesis testing procedure will have a probability of type I error larger than the significance level.
  - In other words, there will likely be more type I errors than expected.

Testing for Subsets of Coefficients

Full model:

$$\text{Logit}\left(\mathrm{p}(X_1, \cdots, X_p, Z_1, \cdots, Z_q)\right)$$
$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q$$

Reduced model:

$$\text{Logit}\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The hypothesis test:

$$\mathbf{H_0}: \boldsymbol{\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0}$$

vs.

$$\mathbf{H_a}: \boldsymbol{\alpha_i \neq 0 \text{ for at least one } \alpha_i, \mathrm{i} = 1, \cdots, q}$$

- Maximize the likelihood function under reduced model: $\mathcal{L}\,(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)\right) \approx \chi_q^2$
  - P-value = Pr($\chi_q^2 >$ Deviance)

Georgia Tech

Testing for Subsets of Coefficients

Full model:

$$\text{Logit}\left(\text{p}(X_1, \cdots, X_p, Z_1, \cdots, Z_q)\right)$$
$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q$$

Reduced model:

$$\text{Logit}\left(\text{p}(X_1, \cdots, X_p\ )\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The hypothesis test:

$$\mathbf{H_0}:\ \boldsymbol{\alpha_1 = \alpha_2 = \cdots = \alpha_q = 0}$$

vs.

$$\mathbf{H_a}:\ \boldsymbol{\alpha_i \neq 0 \text{ for at least one } \alpha_i, \mathrm{i} = 1, \cdots, q}$$

- The hypothesis test for subsets of coefficients is approximate
  - It relies on large sample size
- This is not a test for goodness of fit!
  - It only compares two models

- Maximize the likelihood function under reduced model: $\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1,\ldots, \bar{\beta}_p))$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1,\ldots, \hat{\beta}_p, \hat{\alpha}_1,\ldots, \hat{\alpha}_q)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}(\bar{\beta}_0, \bar{\beta}_1,\ldots, \bar{\beta}_p)\right) - \log\left(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1,\ldots, \hat{\beta}_p, \hat{\alpha}_1,\ldots, \hat{\alpha}_q)\right) \approx \chi^2_q$
  - P-value = $\Pr(\chi^2_q > \text{Deviance})$

Georgia Tech

Testing for Overall Regression

Full model:

Logit $\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Reduced model:

Logit $\left(\mathrm{p}(X_1, \cdots, X_p)\right) = \beta_0$

The hypothesis test:

$\mathbf{H_0}$: $\boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0}$

vs.

$\mathbf{H_a}$: $\boldsymbol{\beta_i \neq 0}$ **for at least one** $\boldsymbol{\beta_i}$, $\mathbf{i = 1, \cdots, \boldsymbol{p}}$

- Maximize the likelihood function under reduced model: $\mathcal{L}\,(\bar{\beta}_0)$
- Maximize the likelihood function under full model: $\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- Test Statistics
  - $\text{Deviance} = \log\left(\mathcal{L}\,(\bar{\beta}_0)\,\right) - \log(\mathcal{L}(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)) \approx \chi_p^2$
  - P-value = Pr($\chi_p^2 >$ Deviance)

Summary