Math 5366 Notes

Logistic Regression

Jesse Crawford

Department of Mathematics
Tarleton State University

Linear Regression

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

Linear Regression

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

$$\begin{aligned}\text{Post-test}_i =& \beta_1 + \beta_2 \text{Pre-test}_i + \beta_3 \text{MathSAT}_i + \beta_4 \text{VerbSAT}_i \\ &+ \beta_5 \text{HSrank}_i + \beta_6 \text{Clickers}_i + \beta_7 \text{GroupWork}_i \\ &+ \epsilon_i, \text{ for } i = 1, \ldots, 140.\end{aligned}$$

Linear Regression

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Linear Regression

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

Logistic Regression Models

Output variable Y is dichotomous ($Y_i = 0$ or $Y_i = 1$)

Example: $Y =$ Student Retention

Logistic Regression Models

Output variable Y is dichotomous ($Y_i = 0$ or $Y_i = 1$)

Example: $Y =$ Student Retention

$$g_i = X_i\beta = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \text{ for } i = 1, \ldots, n.$$

Logistic Regression Models

Output variable Y is dichotomous ($Y_i = 0$ or $Y_i = 1$)

Example: $Y =$ Student Retention

$$g_i = X_i\beta = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \text{ for } i = 1, \ldots, n.$$

$$P(Y_i = 1) = \pi_i = \frac{1}{1 + e^{-g_i}}, \text{ for } i = 1, \ldots, n.$$

Example in R

True Model

$$g_i = -3 + 0.06X_i, \text{ for } i = 1, \ldots, 100000.$$

```
X=runif(100000,0,100)

g=-3+.06*X
Pi=(1/(1+exp(-g)))

U=runif(100000)
Y=(U<Pi)*1
```

True Model

$$g_i = -3 + 0.06X_i, \text{ for } i = 1, \ldots, 100000.$$

```
model=glm(Y~X,family=binomial)
summary(model)
```

```
Call:
glm(formula = Y ~ X, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4339  -0.6851  -0.3054   0.6772   2.5390

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.1834684  0.0205484  -154.9   <2e-16 ***
X            0.0609352  0.0003656   166.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 138436  on 99999  degrees of freedom
Residual deviance:  92361  on 99998  degrees of freedom
AIC: 92365

Number of Fisher Scoring iterations: 4
```

Maximum Likelihood Estimation

Likelihood function

$$L = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i}$$

Likelihood equations

$$\sum_{i=1}^{n} X_{ij}(Y_i - \pi_i) = 0, \text{ for } j = 1, \ldots, p.$$

Maximum Likelihood Estimator: $\hat{\beta}$

Estimating g and π

$$\hat{\beta} = \text{Maximum Likelihood Estimator for } \beta$$

$$\hat{g}_i = X_i\hat{\beta} = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}, \text{ for } i = 1, \ldots, n.$$

Estimating g and π

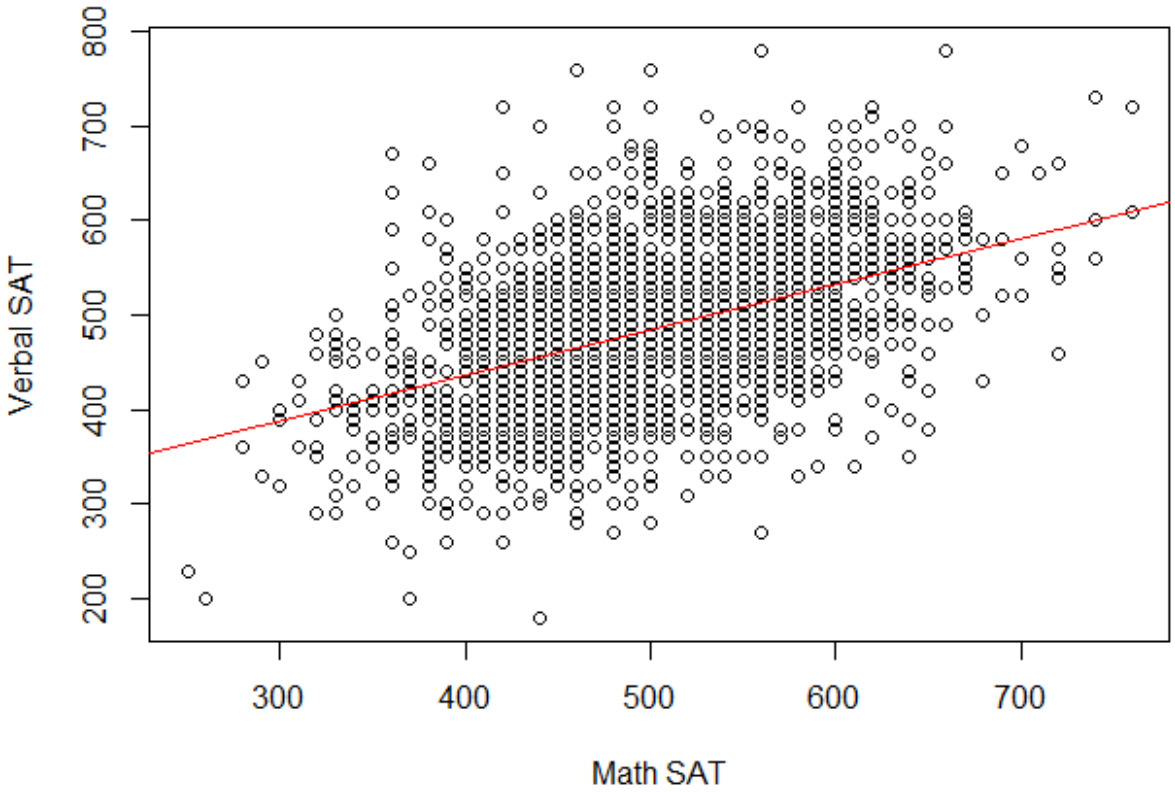
$$\hat{\beta} = \text{Maximum Likelihood Estimator for } \beta$$

$$\hat{g}_i = X_i\hat{\beta} = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}, \text{ for } i = 1, \ldots, n.$$

$$\hat{\pi}_i = \frac{1}{1 + e^{-\hat{g}_i}}, \text{ for } i = 1, \ldots, n.$$

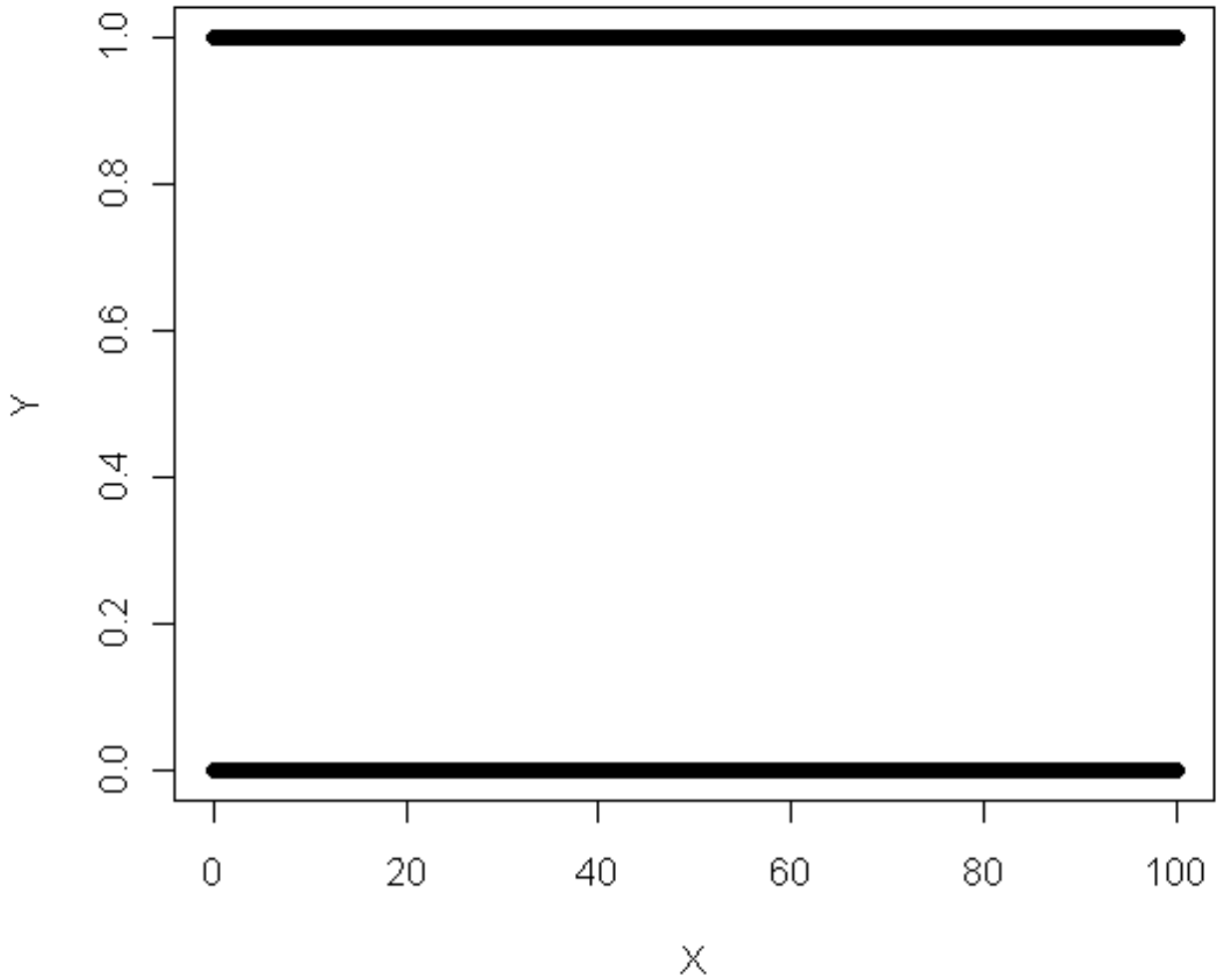
Plots

A Linear Regression Scatterplot



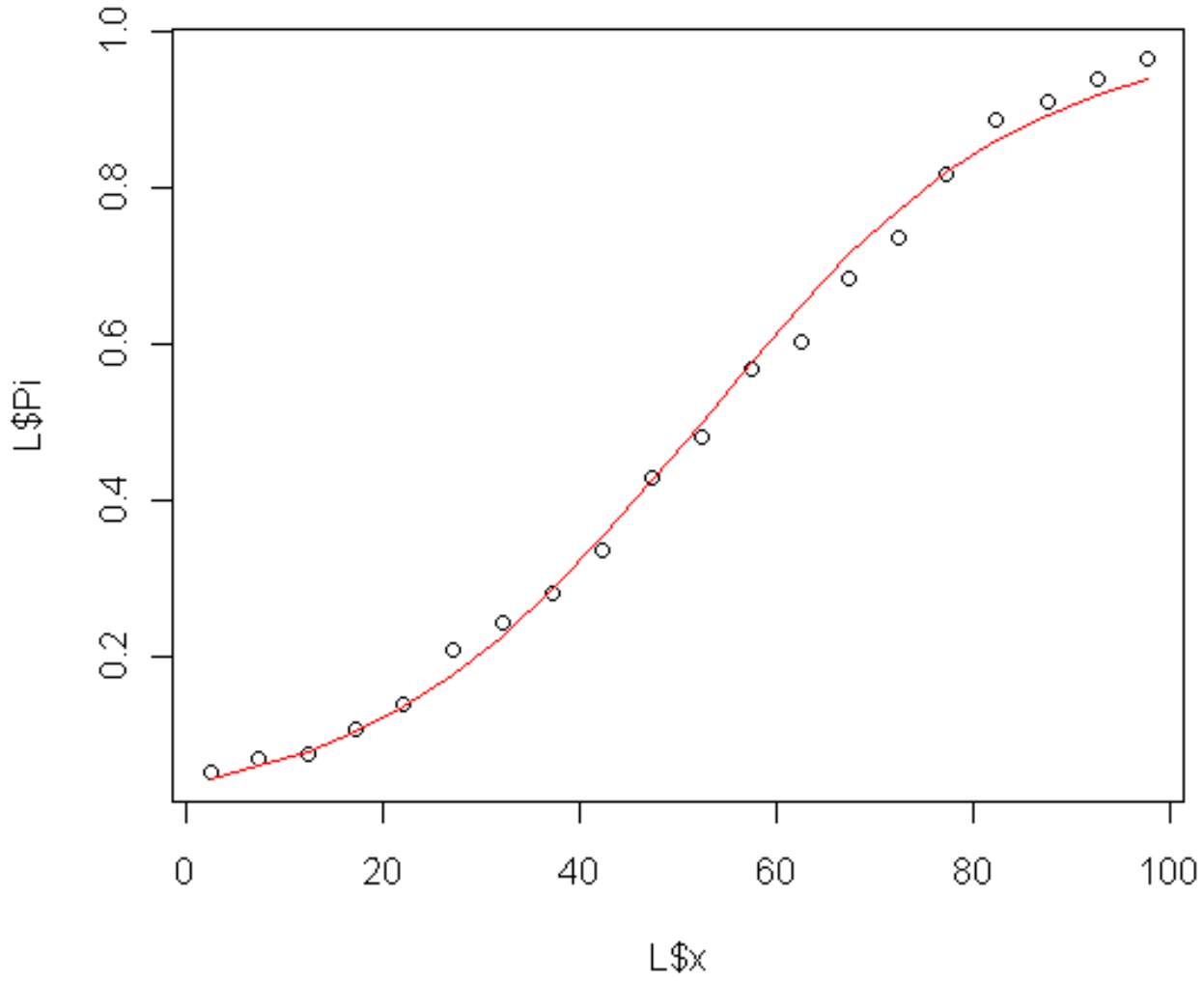
Plots

Y vs. *X* (Not very useful).



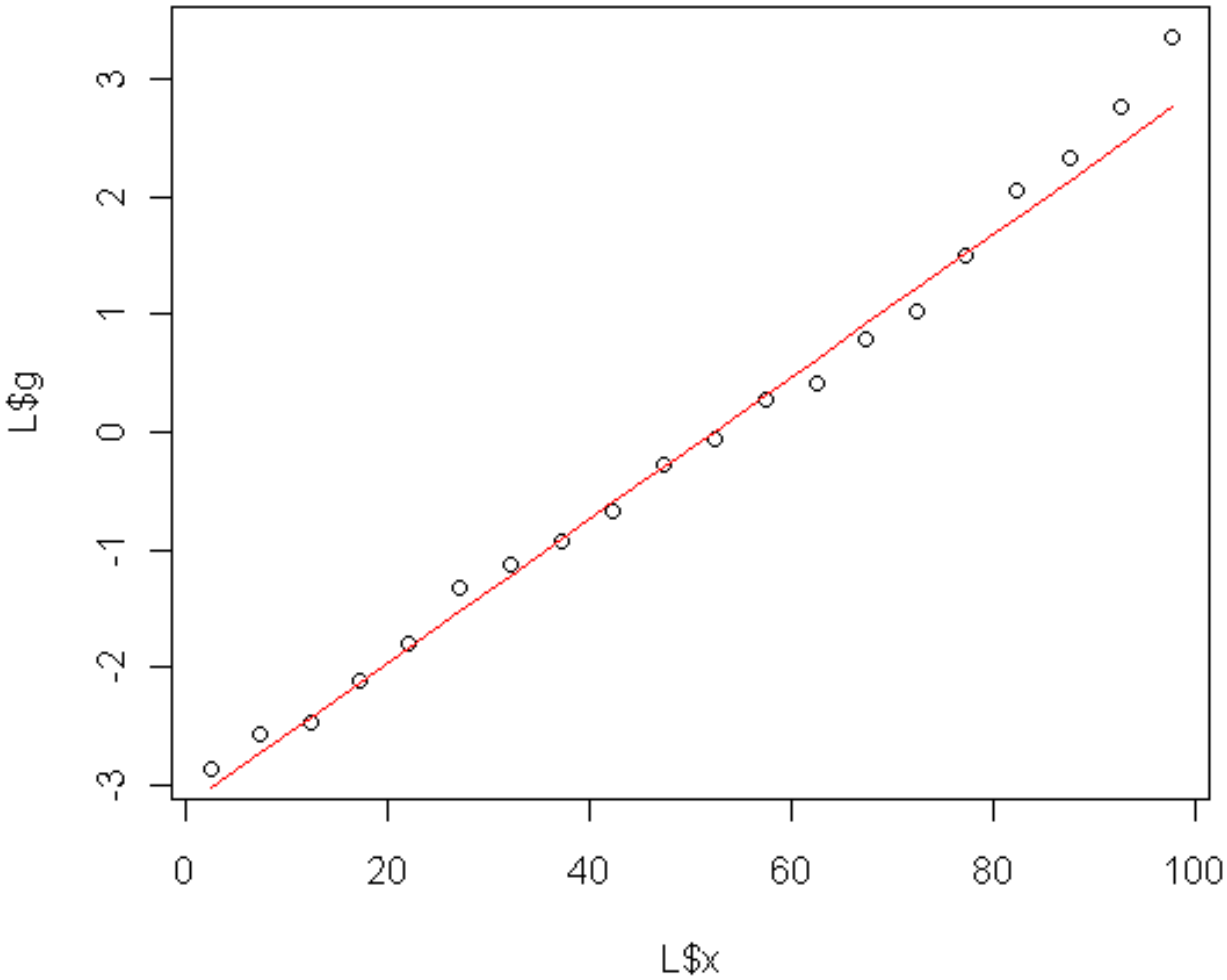
Plots

$\hat{\pi}$ vs. X



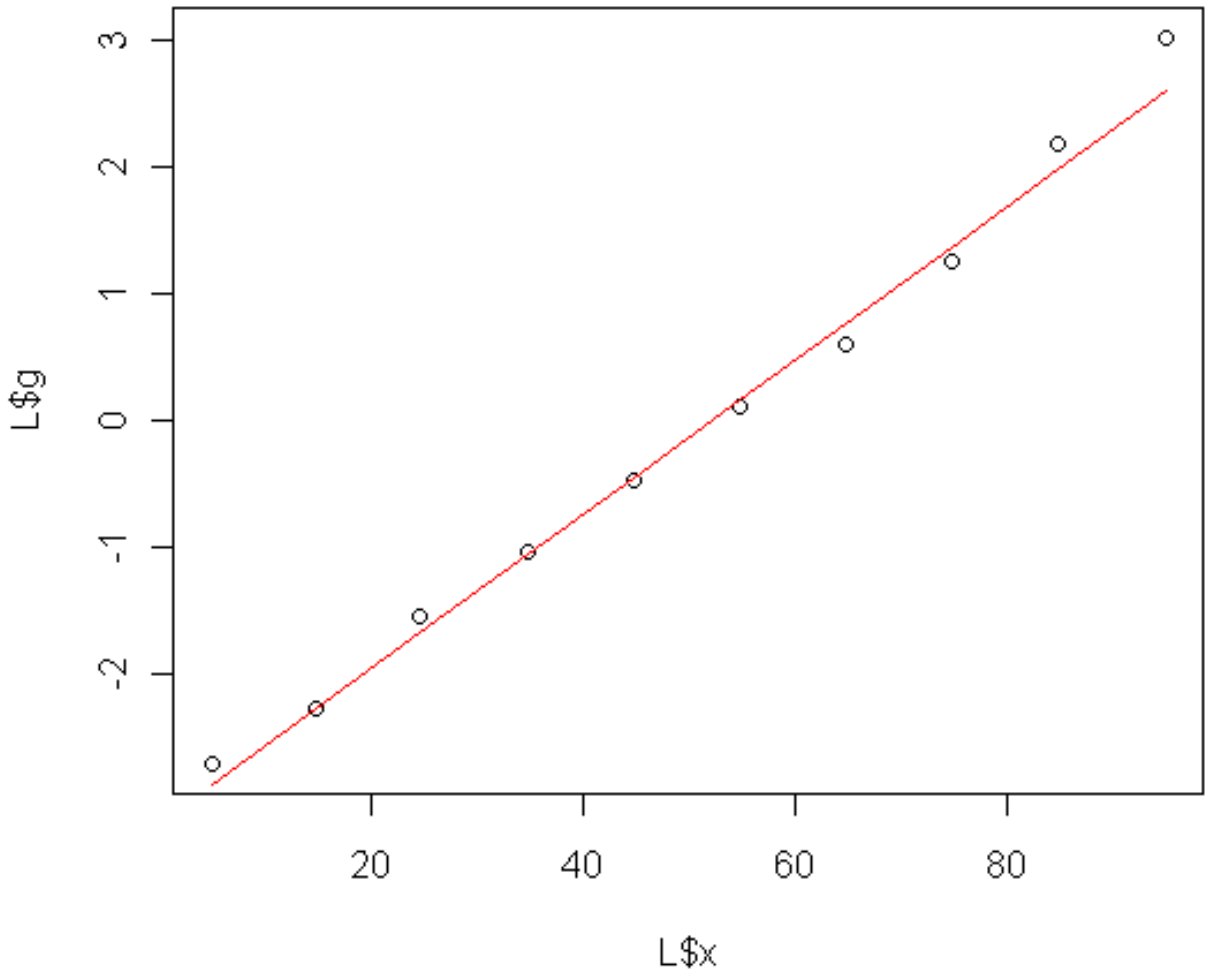
Plots

$\hat{g}$ vs. X (Best plot for assessing functional form)



Plots

$\hat{g}$ vs. X (Best plot for assessing functional form)



Model Deviance and Aikake Information Criterion

$$\text{Deviance} = -2\ln(L) = -2\sum_{i=1}^{n} Y_i \ln(\hat{\pi}_i) + (1 - Y_i)\ln(1 - \hat{\pi}_i)$$

$$\text{AIC} = 2p - 2\ln(L)$$

Hypothesis Testing

- Consider the logistic regression model

$$P(Y_i = 1) = \frac{1}{1 + e^{-g_i}}, \text{ where}$$

$$g_i = X_i\beta.$$

- Let $V_0 \leq V \leq \mathbb{R}^p$, and consider the testing problem

$$\mathrm{H}_0 : \beta \in V_0 \text{ vs. } \mathrm{H} : \beta \in V.$$

- The test statistic is $G = D_0 - D$, where D_0 and D are the deviances under H_0 and H, respectively.
- Under H_0, the approximate distribution of G is chi-square with $\dim(V) - \dim(V_0)$ degrees of freedom, so

$$\text{reject } \mathrm{H}_0 \text{ if } G > \chi^2_\alpha(\dim(V) - \dim(V_0)).$$

Variable Selection

- Manually
- Stepwise
- Best subsets

Assessing Model Performance and Fit

- Classification Accuracy
- Area under ROC Curve
- Hosmer-Lemeshow Goodness-of-fit Test

Hosmer-Lemeshow Goodness-of-fit Test

- This test is used to test the null hypothesis that a logistic regression model adequately fits the data.

Hosmer-Lemeshow Goodness-of-fit Test

- This test is used to test the null hypothesis that a logistic regression model adequately fits the data.
- Divide the data into 10 deciles based on the value of $\hat{\pi}$.
- For $k = 1, \ldots, 10$, define the following
  - n_k = number of objects (rows of data) in the kth decile
  - $\hat{\pi}_k$ = average value of $\hat{\pi}$ for objects in the kth decile
  - o_k = number of objects in the kth decile with $Y = 1$

Hosmer-Lemeshow Goodness-of-fit Test

- This test is used to test the null hypothesis that a logistic regression model adequately fits the data.
- Divide the data into 10 deciles based on the value of $\hat{\pi}$.
- For $k = 1, \ldots, 10$, define the following
  - n_k = number of objects (rows of data) in the kth decile
  - $\hat{\pi}_k$ = average value of $\hat{\pi}$ for objects in the kth decile
  - o_k = number of objects in the kth decile with $Y = 1$
- The Hosmer-Lemeshow test statistic is

$$\hat{C} = \sum_{k=1}^{10} \frac{(o_k - n_k \hat{\pi}_k)^2}{n_k \hat{\pi}_k (1 - \hat{\pi}_k)}$$

Hosmer-Lemeshow Goodness-of-fit Test

- This test is used to test the null hypothesis that a logistic regression model adequately fits the data.
- Divide the data into 10 deciles based on the value of $\hat{\pi}$.
- For $k = 1, \ldots, 10$, define the following
  - n_k = number of objects (rows of data) in the kth decile
  - $\hat{\pi}_k$ = average value of $\hat{\pi}$ for objects in the kth decile
  - o_k = number of objects in the kth decile with $Y = 1$
- The Hosmer-Lemeshow test statistic is

$$\hat{C} = \sum_{k=1}^{10} \frac{(o_k - n_k \hat{\pi}_k)^2}{n_k \hat{\pi}_k (1 - \hat{\pi}_k)}$$

- Under H_0, $\hat{C} \approx \chi^2(8)$.

Hosmer-Lemeshow Goodness-of-fit Test

- This test is used to test the null hypothesis that a logistic regression model adequately fits the data.
- Divide the data into 10 deciles based on the value of $\hat{\pi}$.
- For $k = 1, \ldots, 10$, define the following
  - n_k = number of objects (rows of data) in the kth decile
  - $\hat{\pi}_k$ = average value of $\hat{\pi}$ for objects in the kth decile
  - o_k = number of objects in the kth decile with $Y = 1$
- The Hosmer-Lemeshow test statistic is

$$\hat{C} = \sum_{k=1}^{10} \frac{(o_k - n_k\hat{\pi}_k)^2}{n_k\hat{\pi}_k(1 - \hat{\pi}_k)}$$

- Under H_0, $\hat{C} \approx \chi^2(8)$.
- Reject H_0 if $\hat{C} > \chi^2_\alpha(8)$.