Lesson 06 Logistic Regression

Lusine Zilfimian

March 23 (Monday), 2020

Contents

- Binominal Logistic Regression

Contents

Contents

Last Lecture ReCap

- Why Not Linear Regression and OLS?

Last Lecture ReCap

- Why Not Linear Regression and OLS?
- What is the difference between binomial and multinomial LogReg?

Last Lecture ReCap

- Why Not Linear Regression and OLS?
- What is the difference between binomial and multinomial LogReg?
- Which type of dependent/independent variable is used in LogReg?

Binary simple case

- Binary response using one predictor

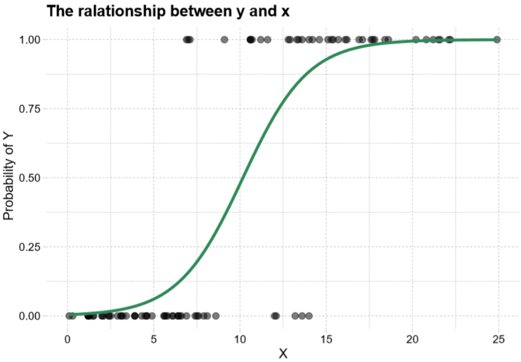
Binary simple case

- Binary response using one predictor
- When $p = 2$, there is only a single linear function to estimate.

Binary simple case

- Binary response using one predictor
- When p $= 2$, there is only a single linear function to estimate.
- The probability: $\mathbb{P}(y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$

S-shaped curve



Terminology

- Odds: $\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = e^{\beta_0+\beta_1 x} \in [0; +\infty)$

Terminology

- Odds: $\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = e^{\beta_0+\beta_1 x} \in [0;+\infty)$
- Log-odds or logit: $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0+\beta_1 x \in (-\infty;+\infty)$

Terminology

- Odds: $\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = e^{\beta_0+\beta_1 x} \in [0; +\infty)$
- Log-odds or logit: $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x \in (-\infty; +\infty)$
- Odds ratio: $\frac{\frac{\mathbb{P}(y=1|x=1)}{1-\mathbb{P}(y=1|x=1)}}{\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)}}$

Estimating the Coefficients

- The most common approach – maximum likelihood: choose that value of parameter, under which it is most likely to get the (your) data.

Estimating the Coefficients

- The most common approach – maximum likelihood: choose that value of parameter, under which it is most likely to get the (your) data.
- In our case, find/choose **estimates** for coefficients such that the **predicted probability** of Y corresponds as closely as possible to Y.

Estimating the Coefficients

- The most common approach – maximum likelihood: choose that value of parameter, under which it is most likely to get the (your) data.
- In our case, find/choose **estimates** for coefficients such that the **predicted probability** of Y corresponds as closely as possible to Y.
- Likelihood Function is the joint PD(M)F, considered as a function of the parameter

Estimating the Coefficients

- The most common approach – maximum likelihood: choose that value of parameter, under which it is most likely to get the (your) data.
- In our case, find/choose **estimates** for coefficients such that the **predicted probability** of Y corresponds as closely as possible to Y.
- Likelihood Function is the joint PD(M)F, considered as a function of the parameter
- Likelihood function:

Estimating the Coefficients

- The most common approach – maximum likelihood: choose that value of parameter, under which it is most likely to get the (your) data.
- In our case, find/choose **estimates** for coefficients such that the **predicted probability** of Y corresponds as closely as possible to Y.
- Likelihood Function is the joint PD(M)F, considered as a function of the parameter
- Likelihood function:
- $$L(\beta_0, \beta_1) = \prod_{y_i=1} F(\beta_0 + \beta_1 x_i) \prod_{y_i=0} (1 - F(\beta_0 + \beta_1 x_i)) \to max$$

Estimating the Coefficients

- The most common approach – maximum likelihood: choose that value of parameter, under which it is most likely to get the (your) data.
- In our case, find/choose **estimates** for coefficients such that the **predicted probability** of Y corresponds as closely as possible to Y.
- Likelihood Function is the joint PD(M)F, considered as a function of the parameter
- Likelihood function:
- $$L(\beta_0, \beta_1) = \prod_{y_i=1} F(\beta_0 + \beta_1 x_i) \prod_{y_i=0} (1 - F(\beta_0 + \beta_1 x_i)) \to max$$
- Likelihood is not a Probability - it can be larger than 1. It is not a PDF either, it is a function of the parameter.

Estimating the Coefficients

- The Maximum Likelihood Method suggests to find a point that makes our Likelihood Maximal:

Estimating the Coefficients

- The Maximum Likelihood Method suggests to find a point that makes our Likelihood Maximal:
- Log-Likelihood function:

Estimating the Coefficients

- The Maximum Likelihood Method suggests to find a point that makes our Likelihood Maximal:
- Log-Likelihood function:
- $l(\beta_0, \beta_1) = \sum_{i=1}^{N}(y_i ln\mathbb{P}(x_i) + (1 - y_i)ln(1 - \mathbb{P}(x_i))$

Estimating the Coefficients

- The Maximum Likelihood Method suggests to find a point that makes our Likelihood Maximal:
- Log-Likelihood function:
- $l(\beta_0, \beta_1) = \sum_{i=1}^{N}(y_i ln\mathbb{P}(x_i) + (1 - y_i)ln(1 - \mathbb{P}(x_i))$
- The points of maximum of $L(\beta_0, \beta_1)$ and $lnL(\beta_0, \beta_1)$ coincide:

Estimating the Coefficients

- The Maximum Likelihood Method suggests to find a point that makes our Likelihood Maximal:
- Log-Likelihood function:
- $l(\beta_0, \beta_1) = \sum_{i=1}^{N}(y_i ln\mathbb{P}(x_i) + (1 - y_i)ln(1 - \mathbb{P}(x_i))$
- The points of maximum of $L(\beta_0, \beta_1)$ and $lnL(\beta_0, \beta_1)$ coincide:
- $\underset{\beta}{argmax} L(\beta_0, \beta_1) = \underset{\beta}{argmax} lnL(\beta_0, \beta_1)$

Estimating the Coefficients

- The Maximum Likelihood Method suggests to find a point that makes our Likelihood Maximal:
- Log-Likelihood function:
- $l(\beta_0, \beta_1) = \sum_{i=1}^{N}(y_i ln\mathbb{P}(x_i) + (1 - y_i) ln(1 - \mathbb{P}(x_i))$
- The points of maximum of $L(\beta_0, \beta_1)$ and $lnL(\beta_0, \beta_1)$ coincide:
- $\underset{\beta}{argmax} L(\beta_0, \beta_1) = \underset{\beta}{argmax} lnL(\beta_0, \beta_1)$
- The estimates for coefficients are calculated using iterative procedure.

Binomial, Multiple logistic regression

- Binary response using multiple predictors

Binomial, Multiple logistic regression

- Binary response using multiple predictors
- Model specification:

Binomial, Multiple logistic regression

- Binary response using multiple predictors
- Model specification:
- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$

Binomial, Multiple logistic regression

- Binary response using multiple predictors
- Model specification:
- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
- $\mathbb{P}(y=1) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}$

Interpretations (Simple LogReg case)

Numeric predictors (Simple LogReg case)

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x$, where X is numeric

The rate of change in $\mathbb{P}(x)$ per unit change in X depends on the current value of X.

Interpretations (Simple LogReg case)

Numeric predictors (Simple LogReg case)

- $ln \frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x$, where X is numeric
- Increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}.

The rate of change in $\mathbb{P}(x)$ per unit change in X depends on the current value of X.

Interpretations (Simple LogReg case)

Numeric predictors (Simple LogReg case)

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x$, where X is numeric
- Increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}.
- If $\beta_1 > 0$, $\uparrow x \Rightarrow \uparrow \mathbb{P}(x)$

The rate of change in $\mathbb{P}(x)$ per unit change in X depends on the current value of X.

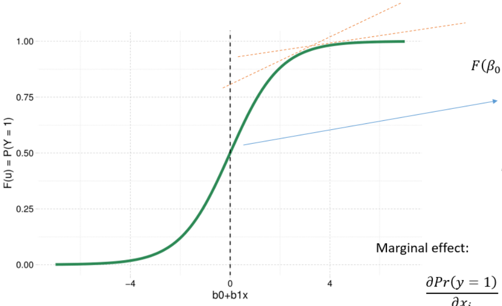
Interpretations (Simple LogReg case)

Numeric predictors (Simple LogReg case)

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 x$, where X is numeric
- Increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}.
- If $\beta_1 > 0$, $\uparrow x \Rightarrow \uparrow \mathbb{P}(x)$
- If $\beta_1 < 0$, $\uparrow x \Rightarrow \downarrow \mathbb{P}(x)$

The rate of change in $\mathbb{P}(x)$ per unit change in X depends on the current value of X.

Numeric predictors (Simple LogReg case)



$$F(\beta_0+\beta_1 x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = $$
$$= \frac{1}{1+e^0} = 0.5$$

$$\beta_0+\beta_1 x = 0 \Rightarrow \mathrm{x} = -\beta_0/\beta_1$$

Marginal effect:

$$\frac{\partial Pr(y=1)}{\partial x_j} = F(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k) * \beta_j$$

Categorical predictors

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 X,$

Categorical predictors

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 X,$
- X is binary variable

Categorical predictors

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 X,$
- X is binary variable
- $ln\frac{\mathbb{P}(y=1|x=1)}{1-\mathbb{P}(y=1|x=1)} - ln\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)} = \beta_1$

Categorical predictors

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 X,$
- X is binary variable
- $ln\frac{\mathbb{P}(y=1|x=1)}{1-\mathbb{P}(y=1|x=1)} - ln\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)} = \beta_1$
- $e^{\beta_1} = \frac{\frac{\mathbb{P}(y=1|x=1)}{1-\mathbb{P}(y=1|x=1)}}{\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)}} = Odds\ ratio$

Categorical predictors

- $ln\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)} = \beta_0 + \beta_1 X,$
- X is binary variable
- $ln\frac{\mathbb{P}(y=1|x=1)}{1-\mathbb{P}(y=1|x=1)} - ln\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)} = \beta_1$
- $e^{\beta_1} = \frac{\frac{\mathbb{P}(y=1|x=1)}{1-\mathbb{P}(y=1|x=1)}}{\frac{\mathbb{P}(y=1|x=0)}{1-\mathbb{P}(y=1|x=0)}} = Odds\ ratio$
- By changing x from 0 to 1, the odds ratio of $y = 1$ will be changed by $\frac{1}{e^{\beta_1}}$ times

Multinomial logistic regression

- Response variable has more than two classes:

Multinomial logistic regression

- Response variable has more than two classes:
- $p = 1, 2, ..., P - 1$

Multinomial logistic regression

- Response variable has more than two classes:
- $p = 1, 2, ..., P - 1$
- $ln\frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=P|X=x)} = \beta_{10} + \beta_{11}X_1 + ... + \beta_{1k}X_k$

Multinomial logistic regression

- Response variable has more than two classes:
- $p = 1, 2, ..., P - 1$
- $ln\frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=P|X=x)} = \beta_{10} + \beta_{11}X_1 + ... + \beta_{1k}X_k$
- $ln\frac{\mathbb{P}(Y=2|X=x)}{1-\mathbb{P}(y=P|x=x)} = \beta_{20} + \beta_{21}X_1 + ... + \beta_{2k}X_k$

Multinomial logistic regression

- Response variable has more than two classes:
- $p = 1, 2, ..., P - 1$
- $ln\frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=P|X=x)} = \beta_{10} + \beta_{11}X_1 + ... + \beta_{1k}X_k$
- $ln\frac{\mathbb{P}(Y=2|X=x)}{1-\mathbb{P}(y=P|x=x)} = \beta_{20} + \beta_{21}X_1 + ... + \beta_{2k}X_k$
- ...

Multinomial logistic regression

- Response variable has more than two classes:
- $p = 1, 2, ..., P - 1$
- $ln \frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=P|X=x)} = \beta_{10} + \beta_{11}X_1 + ... + \beta_{1k}X_k$
- $ln \frac{\mathbb{P}(Y=2|X=x)}{1-\mathbb{P}(y=P|x=x)} = \beta_{20} + \beta_{21}X_1 + ... + \beta_{2k}X_k$
- ...
- $ln \frac{\mathbb{P}(Y=P-1|X=x)}{1-\mathbb{P}(Y=P|X=x)} = \beta_{(P-1)0} + \beta_{(P-1)1}X_1 + ... + \beta_{(P-1)k}X_k$

Multinomial logistic regression

- $\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_{10}+\beta_{11}X_1+\ldots+\beta_{1k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$

Multinomial logistic regression

- $\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_{10}+\beta_{11}X_1+\ldots+\beta_{1k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\mathbb{P}(Y = 2|X = x) = \frac{e^{\beta_{20}+\beta_{21}X_1+\ldots+\beta_{2k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$

Multinomial logistic regression

- $\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_{10}+\beta_{11}X_1+\ldots+\beta_{1k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\mathbb{P}(Y = 2|X = x) = \frac{e^{\beta_{20}+\beta_{21}X_1+\ldots+\beta_{2k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\ldots$

Multinomial logistic regression

- $\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_{10}+\beta_{11}X_1+\ldots+\beta_{1k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\mathbb{P}(Y = 2|X = x) = \frac{e^{\beta_{20}+\beta_{21}X_1+\ldots+\beta_{2k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\ldots$
- $\mathbb{P}(Y = P|X = x) = \frac{1}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$

Multinomial logistic regression

- $\mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_{10}+\beta_{11}X_1+\ldots+\beta_{1k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\mathbb{P}(Y = 2|X = x) = \frac{e^{\beta_{20}+\beta_{21}X_1+\ldots+\beta_{2k}X_k}}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- $\ldots$
- $\mathbb{P}(Y = P|X = x) = \frac{1}{1+\sum_{l=1}^{P-1} e^{\beta_{l0}+\beta_{l1}X_1+\ldots+\beta_{lk}X_k}}$
- **Derive these formulas for p=3 case.**

Goodness of fit

- Confusion matrix

		Predicted	
		Negative (0)	Positive (1)
Actual	Negative (0)	TN	FP
	Positive (1)	FN	TP

Goodness of fit

- Confusion matrix

		Predicted	
		Negative (0)	Positive (1)
Actual	Negative (0)	TN	FP
	Positive (1)	FN	TP

- $Accuracy = \frac{TP+TN}{Total}$

Goodness of fit

- Confusion matrix

		Predicted	
		Negative (0)	Positive (1)
Actual	Negative (0)	TN	FP
	Positive (1)	FN	TP

- $Accuracy = \frac{TP+TN}{Total}$
- $Sensitivity = \frac{TP}{TP+FN}$

Goodness of fit

- Confusion matrix

		Predicted	
		Negative (0)	Positive (1)
Actual	Negative (0)	TN	FP
	Positive (1)	FN	TP

- $Accuracy = \frac{TP+TN}{Total}$
- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$

Goodness of fit

- How to choose the threshold?

Predicted	Actual1	Actual2	Actual3	Actual4	Actual5
0.1000000	0	1	0	0	0
0.1888889	0	1	0	0	1
0.2777778	0	1	0	0	0
0.3666667	0	1	0	0	1
0.4555556	0	1	0	1	0
0.5444444	0	1	1	1	1
0.6333333	0	1	1	1	0
0.7222222	0	1	1	1	1
0.8111111	0	1	1	1	0
0.9000000	0	1	1	1	1

Goodness of fit

- How to choose the threshold?

Predicted	Actual1	Actual2	Actual3	Actual4	Actual5
0.1000000	0	1	0	0	0
0.1888889	0	1	0	0	1
0.2777778	0	1	0	0	0
0.3666667	0	1	0	0	1
0.4555556	0	1	0	1	0
0.5444444	0	1	1	1	1
0.6333333	0	1	1	1	0
0.7222222	0	1	1	1	1
0.8111111	0	1	1	1	0
0.9000000	0	1	1	1	1

- Which one is the worst case?

Goodness of fit

- How to choose the threshold?

Predicted	Actual1	Actual2	Actual3	Actual4	Actual5
0.1000000	0	1	0	0	0
0.1888889	0	1	0	0	1
0.2777778	0	1	0	0	0
0.3666667	0	1	0	0	1
0.4555556	0	1	0	1	0
0.5444444	0	1	1	1	1
0.6333333	0	1	1	1	0
0.7222222	0	1	1	1	1
0.8111111	0	1	1	1	0
0.9000000	0	1	1	1	1

- Which one is the worst case?
- Maximum cutoff value $\Rightarrow$ all records will be classified as 0.

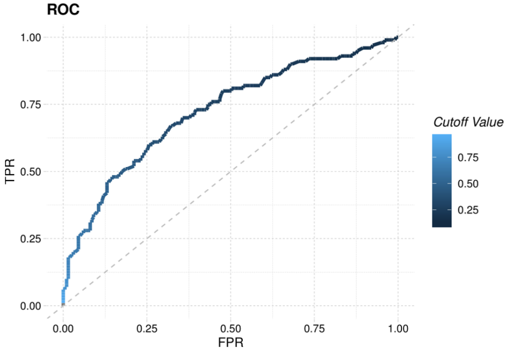
Goodness of fit

- How to choose the threshold?

Predicted	Actual1	Actual2	Actual3	Actual4	Actual5
0.1000000	0	1	0	0	0
0.1888889	0	1	0	0	1
0.2777778	0	1	0	0	0
0.3666667	0	1	0	0	1
0.4555556	0	1	0	1	0
0.5444444	0	1	1	1	1
0.6333333	0	1	1	1	0
0.7222222	0	1	1	1	1
0.8111111	0	1	1	1	0
0.9000000	0	1	1	1	1

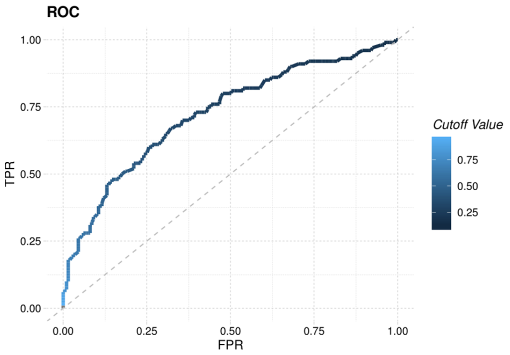
- Which one is the worst case?
- Maximum cutoff value $\Rightarrow$ all records will be classified as 0.
- Minimum cutoff value $\Rightarrow$ all records will be classified as 1.

Goodness of fit: ROC and AUC



- ROC - the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity)

Goodness of fit: ROC and AUC



- ROC - the trade-off between True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity)
- AUC – Area under ROC

Goodness of fit: Tests

- Likelihood ratio Test

Goodness of fit: Tests

- Likelihood ratio Test
- $LR = 2(loglik(m_1) - loglik(m_0)) \sim \chi^2$

Goodness of fit: Tests

- Likelihood ratio Test
- $LR = 2(loglik(m_1) - loglik(m_0)) \sim \chi^2$
- $H_0 : L_{reduced} > L_{current}$

Goodness of fit: Tests

- Likelihood ratio Test
- $LR = 2(loglik(m_1) - loglik(m_0)) \sim \chi^2$
- $H_0 : L_{reduced} > L_{current}$
- $H_1 : L_{reduced} < L_{current}$

Goodness of fit: Tests

- Likelihood ratio Test
- $LR = 2(loglik(m_1) - loglik(m_0)) \sim \chi^2$
- $H_0 : L_{reduced} > L_{current}$
- $H_1 : L_{reduced} < L_{current}$
- Hosmer-Lemeshow Test

Goodness of fit: Tests

- Likelihood ratio Test
- $LR = 2(loglik(m_1) - loglik(m_0)) \sim \chi^2$
- $H_0 : L_{reduced} > L_{current}$
- $H_1 : L_{reduced} < L_{current}$
- Hosmer-Lemeshow Test
- $H_0 : Observed\ proportion = Expected\ Proportion$

Goodness of fit: Tests

- Likelihood ratio Test
- $LR = 2(loglik(m_1) - loglik(m_0)) \sim \chi^2$
- $H_0 : L_{reduced} > L_{current}$
- $H_1 : L_{reduced} < L_{current}$
- Hosmer-Lemeshow Test
- $H_0 : Observed\ proportion = Expected\ Proportion$
- $H_1 : Observed\ proportion \neq Expected\ Proportion$

Ideas for final Project

- Play with Goodness of fit

Ideas for final Project

- Play with Goodness of fit
- LogReg for ordinal response dependent variable.

Coding examples in R

- See in Lab 06