Introduction to Logistic Regression

P. Boily, N. Afodjo

Contents

Simple Logistic Regression
Preliminaries

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$Y_i = 0,1$$

The expected response $E(Y_i)$ has a special meaning in this case:

- Since $E(\varepsilon_i) = 0$,

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

- Also, as Y_i is a discrete variable with two possible values, we can state its probability distribution as

$$E(Y_i) = \pi_i = P(Y_i = 1)$$

Y_i	Probability
1	$\pi_i = P(Y_i = 1)$
0	$1 - \pi_i = P(Y_i = 0)$

Simple Logistic Regression
Preliminaries

In this case, special problems arise

- Non-normal error terms
- Non-constant error variance

The response function will represent a probability so the mean responses should be constrained: $0 \leq E(Y_i) = \pi_i \leq 1$

Goal: Find an invertible function $f: [0,1] \to \mathbb{R}$, $f(Y_i) = Y_i^*$ and use **ordinary least square** (OLS) to regress Y_i^* on the X_i:

$$Y_i^* = f(Y_i) = \beta_0^* + \beta_1^* X_i$$

Simple Logistic Regression
Response Functions

Two main candidates – probit and logit responses

- The **probit response function** is

$$\pi_i' = \Phi^{-1}(\pi_i) = \beta_0^* + \beta_1^* X_i$$

where Φ represents the cdf of the standard normal distribution $N(0,1)$

- The **logit response function** is

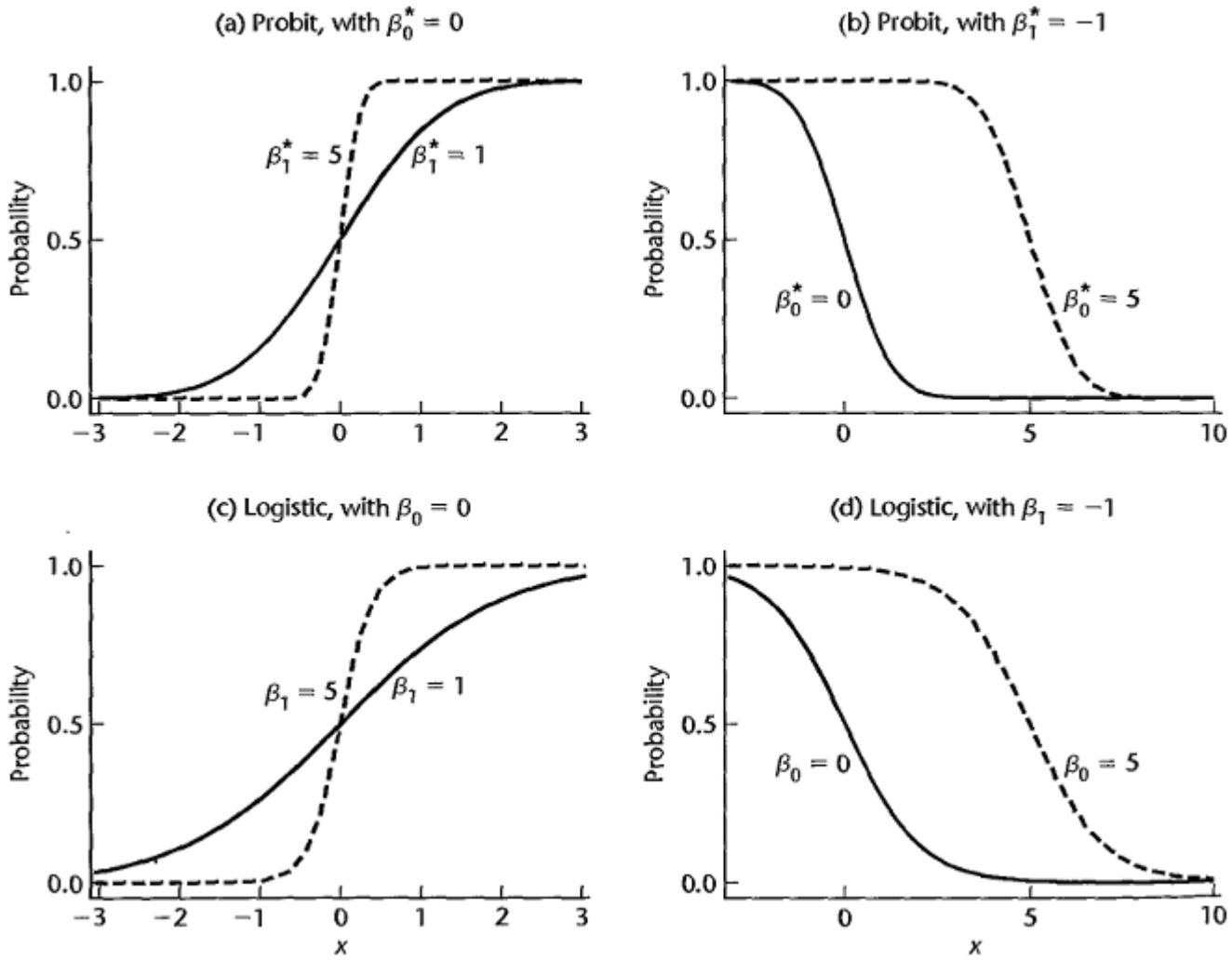
$$\pi_i' = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0^* + \beta_1^* X_i$$

also known as log-odds:

$$\pi_i = \frac{\exp(\pi_i')}{1+\exp(\pi_i')} = \frac{\exp(\beta_0^* + \beta_1^* X_i)}{1+\exp(\beta_0^* + \beta_1^* X_i)}$$

Simple Logistic Regression

Response Functions



[Kutner et. al., *Applied Linear Statistical Models*, 3rd ed.]

Simple Logistic Regression

Maximum Likelihood Expectation

The **log-likelihood** for the logistic response function is

$$LL(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i \ln \pi_i + (1 - Y_i) \ln(1 - \pi_i)],$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i)}$$

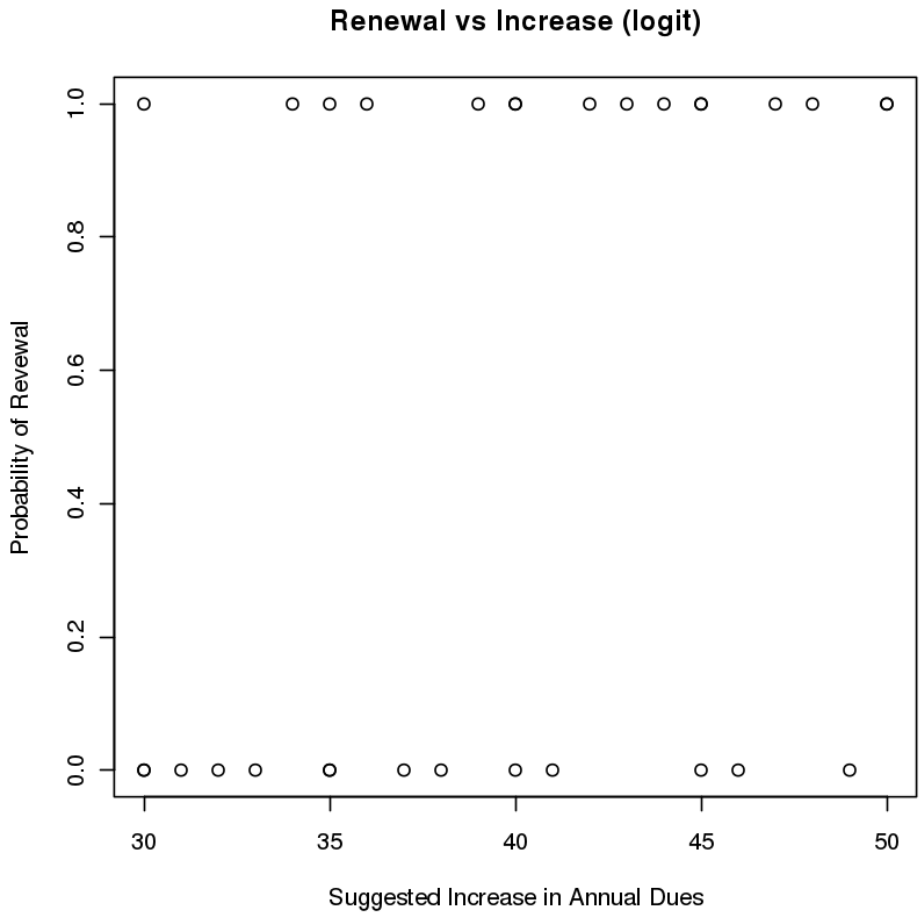
The **maximum likelihood estimates (MLE)** b_i of the β_i in the simple logistic regression model are those values of β_0 and β_1 that maximize $LL(\beta_0, \beta_1)$

No closed-form solution exists for the MLE of the β_i

The use of a computer is required

Simple Logistic Regression

Example – Annual Dues



Logistic Regression Notebook (ex. 2)

30 observations

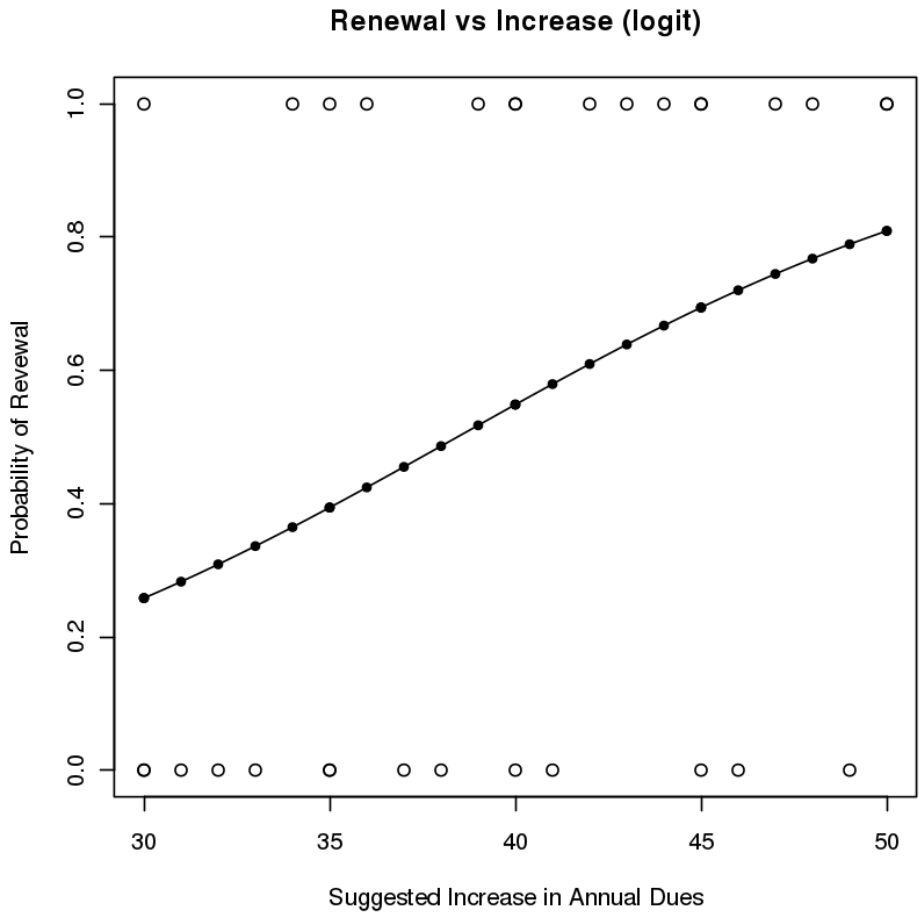
X is the increase in annual dues;
Y is the non-renewal.

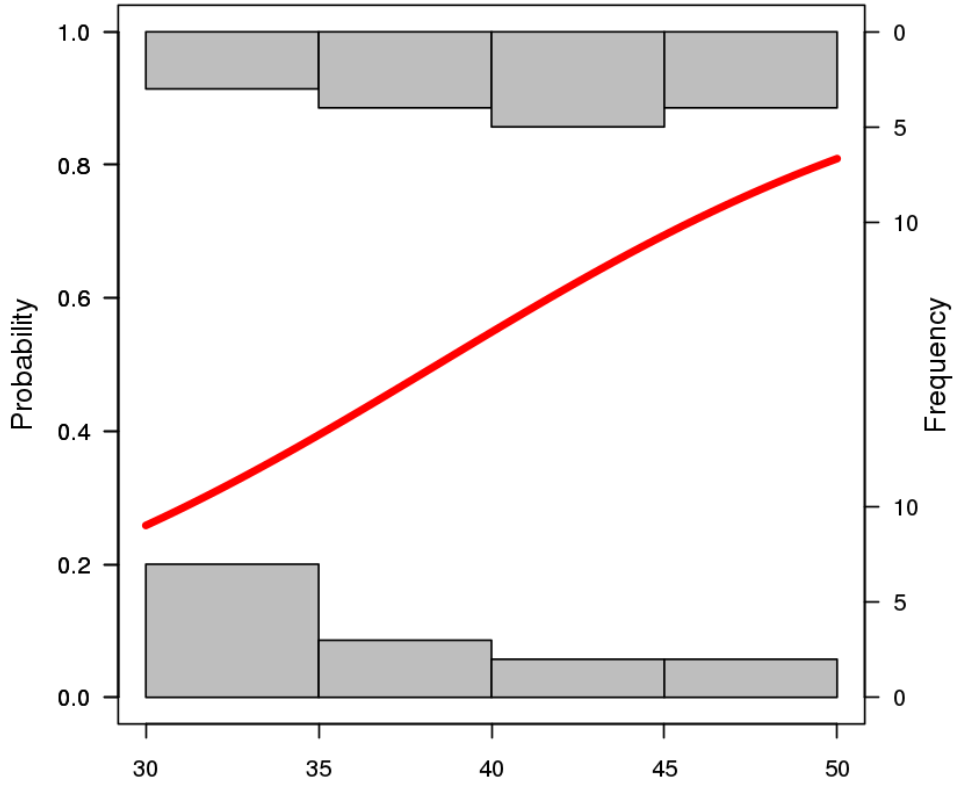
Logistic regression MLE coefficients:

$$\boldsymbol{b_0 = -4.81} \text{ and } \boldsymbol{b_1 = 0.125}$$

Simple Logistic Regression

Example – Annual Dues





Simple Logistic Regression

Example – Annual Dues - Interpreting the Coefficients

From simple linear regression, we have that

- A unit increase in X_i will increase the log(odds) by b_1 units
- The **odds ratio** (OR) will therefore increase by e^{b_1} units. We write OR $= e^{b_1}$

In this example, OR $= e^{b_1} = 1.275$, which implies that for every \$1 increase in annual dues, the odds of non-renewal increase by 27.5%.

NOT the same as saying that non-renewal increases by 27.5%.

If the initial probability of non-renewal was $\pi = 0.5$, for instance, then OR=1. An increase of \$1 in annual dues would then bring the odds ratio of non-renewal to

$$\text{OR}=\frac{\pi'}{1-\pi'}=1.275$$

which translates to $\pi' = 0.56$, an increase of 6.5% in the non-renewal probability.

Multiple Logistic Regression

Basics

The simple logistic regression model is easily extended to more than one predictor variable

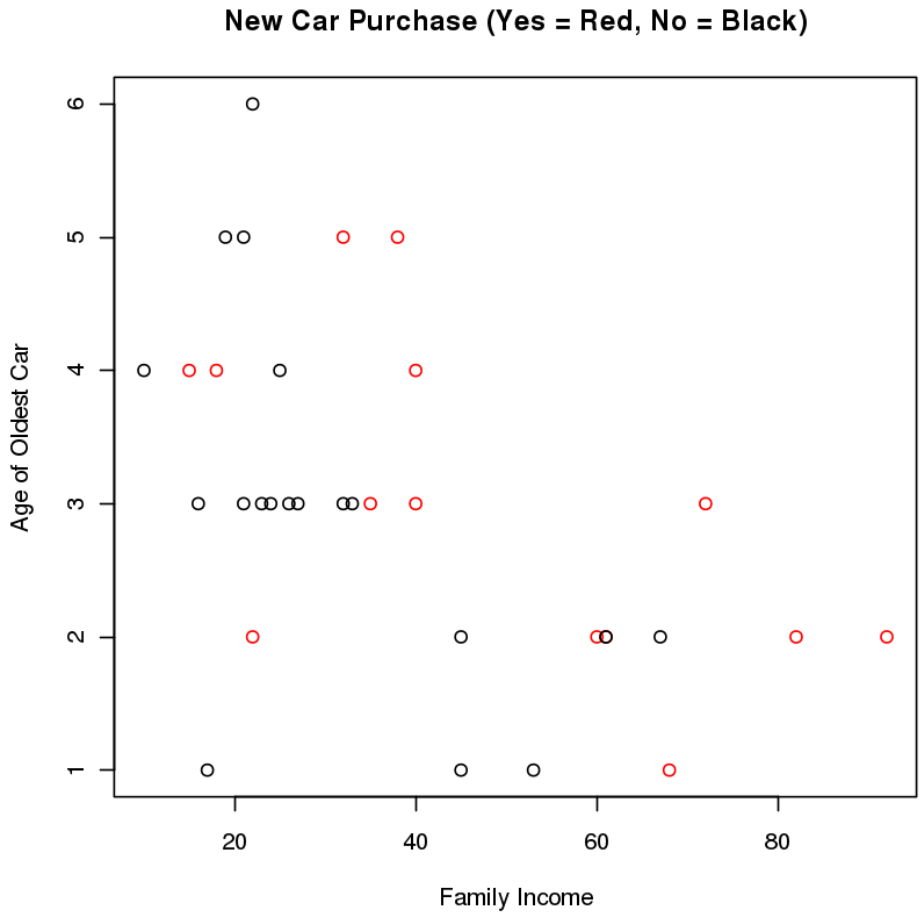
In fact, several predictor variables are usually required with logistic regression to obtain adequate description and useful predictions.

In extending the simple logistic regression model, we simply replace $\beta_0 + \beta_1 X$ by

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

Simple Logistic Regression

Example – Car Purchase



See *Logistic Regression* Notebook (ex. 3)

33 observations

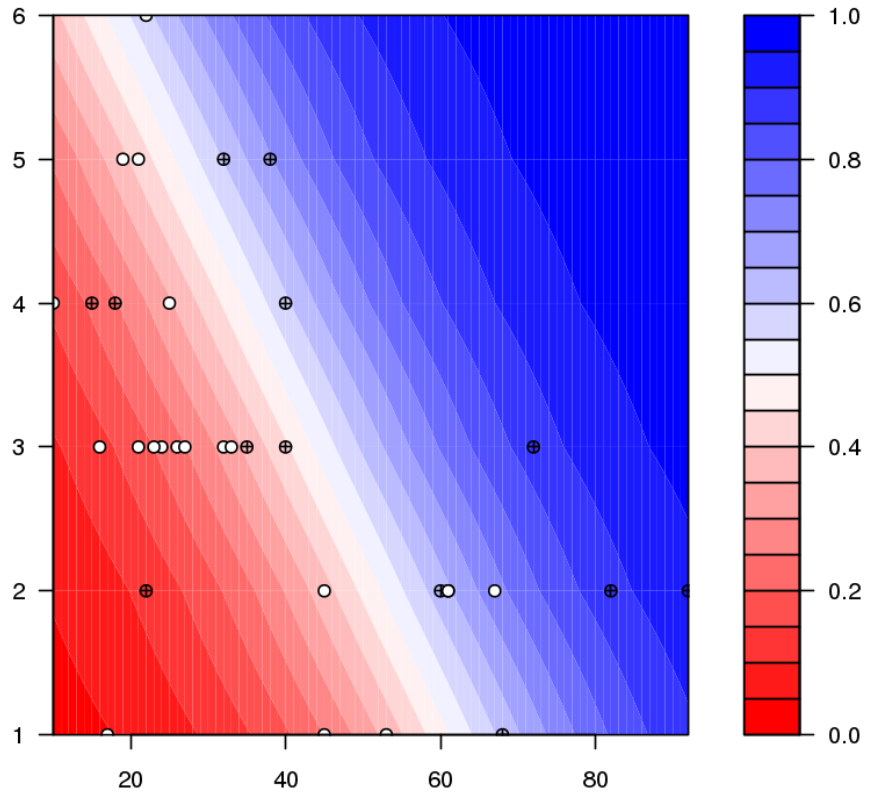
X_1 is the annual family income; X_2 is the age of the oldest family car; Y is the purchase of a car during the year.

Logistic regression MLE coeffs:

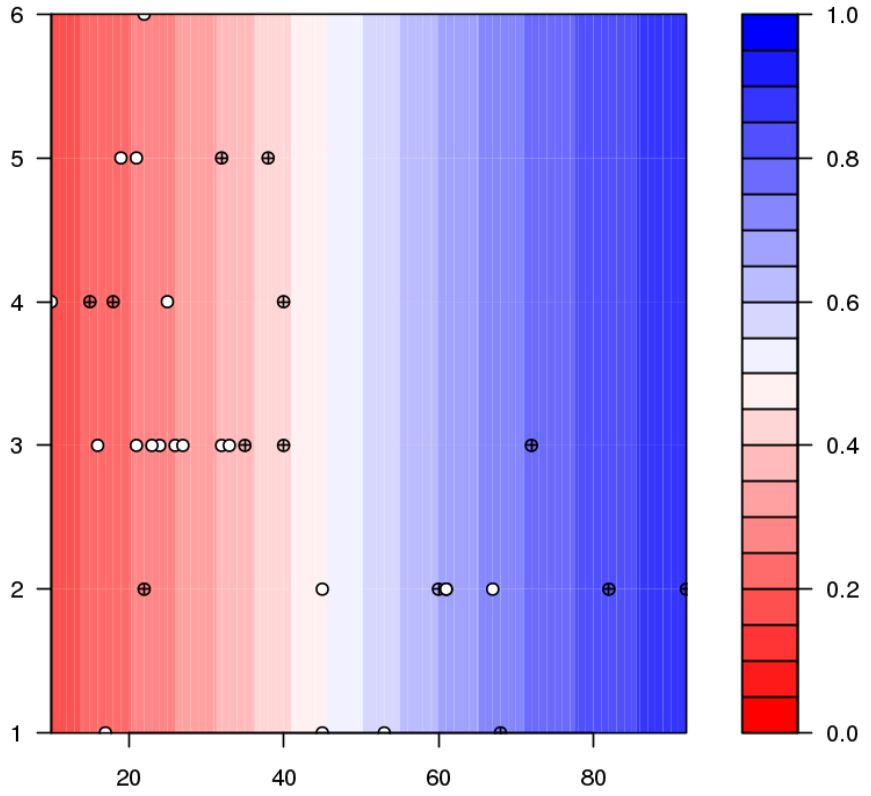
$$\boldsymbol{b_0 = -4.739, b_1 = 0.0678} \text{ and } \boldsymbol{b_2 = 0.599}$$

Simple Logistic Regression

Example – Car Purchase – Decision Boundaries



2 variables: Income, Age of Oldest Car



1 variable: Income

Multiple Logistic Regression

Example – Car Purchase – Interpretation of Coefficients

For every 10,000$ increase in annual family income, the odds of buying a new car almost double ($e^{10b_1} = 1.97$), all else being equal.

From one year to the next, everything else being equal, it appears that the odds of buying a new car increase by $e^{b_2} - 1 = 82\%$.

The 2-predictor model only has one statistically significant predictor (Income). This does not necessarily mean that the Age of the Oldest Car does not also play a role, or that Income is the sole factor that does play a role, in the purchase of a new car during the year, but what about the simpler model?

Logistic regression MLE coefficients of the simpler model are:

$$\boldsymbol{b_0 = -1.981} \text{ and } \boldsymbol{b_1 = 0.0434}$$

Other Types of Logistic Regression

Logistic regression is most frequently used to model the relationship between a dichotomous response variable and a set of predictor variables.

On occasion, however, the response variable may have more than two levels. Logistic regression can still be employed by means of a **polytomous** – or multicategory – logistic regression model

2 cases:

- The response variable is **nominal**; e.g. a market researcher may wish to relate a consumer's choice of product (product A, product B, product C) to the consumer's age, gender, location, etc.
- The response variable is **ordinal**; e.g. research linking age, alcohol use, smoking history to pregnancy duration:

$$\leq 36 \text{ wks} = \text{preterm}, 36 \text{ to } 37 \text{ wks} = \text{intermediate}, \geq 38 \text{ wks} = \text{full term}$$

Model Validation

Goodness-of-Fit Tests

The appropriateness of the fitted logistic regression model needs to be examined before it is accepted for use, as is the case for all regression models.

In particular, we need to examine whether the estimated response function for the data is monotonic and has an S-shape

Goodness-of-fit tests provide an overall measure of the fit of the model, and are usually not sensitive when the fit is poor for just a few cases. Common tests are:

- Pearson Chi-Square: uses the estimated probabilities to compute expected numbers of successes and then compares theses to the observed numbers
- Deviation: as in the Pearson case, requires replicated data
- Hosmer-Lemeshow: preferred for data with little or no replicates in the predictor levels

Model Validation

Data Reduction

When the number of potential explanatory variables is large (more than 20, say), it is productive to chose the "best" subset of these variables in order to minimize computational efforts.

Given p variables and a sample of size n, the "best" logistic model will be the one combining the highest ML value with the fewest number of variables.

Therefore, the "best" model should minimize

$$AIC_k = -2\,LL(\boldsymbol{\beta}) + 2k \quad \text{or} \quad SBC_k = -2\,LL(\boldsymbol{\beta}) + k\ln(n)$$

where k is the size of a subset of the original p coeff. (the constant term is always included).

This **best subset** procedure is available in some statistical packages

In the machine learning framework, data reduction procedures based on stepwise selection are discouraged; consider using appropriate **LASSO** (regularization) procedures instead.

Predictions

Binary Response Variables

Logistic models (after goodness-of-fit verification) may be used to make inferences about the mean response, given a new set of predictor values.

Given $\boldsymbol{X}_h' = [1 \quad X_{h1} \quad \cdots \quad X_{h,p-1}]$, we find $\hat{\pi}_h = E(Y = 1|\boldsymbol{X}_h')$

In the binary case, given a cutoff value c, we will predict

$$\begin{cases} \hat{Y} = 1 \text{ if } \hat{\pi}_h > c \\ \hat{Y} = 0 \text{ if } \hat{\pi}_h \leq c \end{cases}$$

The **cutoff point** is usually chosen as 0.5 especially if:

- Outcomes 0 and 1 are equally likely to occur
- The costs of incorrectly predicting 0 and 1 are the same

Predictions

Binary Response Variables

One can also calculate the correct prediction rate for various cutoff values and thus choose c as to minimize the proportion of incorrect predictions.

As stated earlier, this approach is reasonable if:

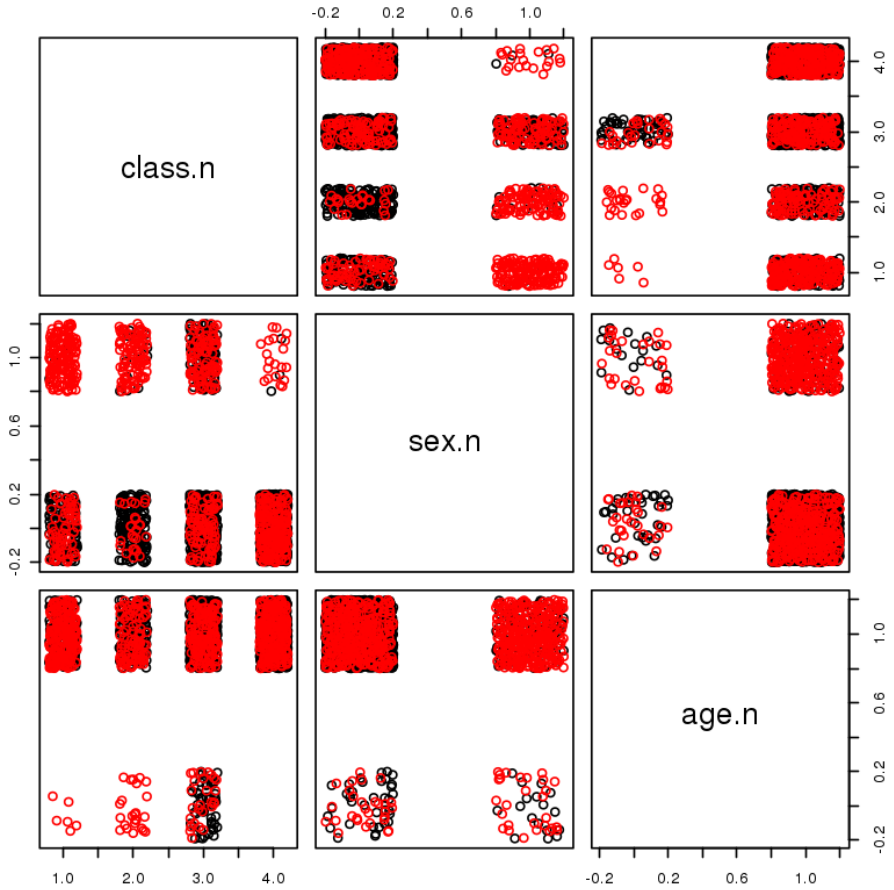
- Outcomes 0 and 1 are roughly equally likely to occur
- The costs of incorrectly predicting 0 and 1 are the same

Otherwise, when available, and in case the data set is not a random sample of the population, prior information on the likelihood of 1 and 0 may be used to assign a cutoff rate.

When logistic regression is used as a classifier, this can also become a goodness-of-fit test: the "best" regression is the one that **makes the fewest prediction errors**.

Predictions

Example – Titanic Dataset



See *Logistic Regression* Notebook (ex. 1)

2201 observations; training on 75% of data

X_1: passenger class; X_2: passenger sex; X_3: passenger age; Y: passenger survival.

Logistic regression MLE coefficients

$$b_0 = 2.0580,\ b_{2\text{Class}} = -1.0166,$$
$$b_{3\text{Class}} = -1.8570,\ b_{\text{Crew}} = -0.8243,$$
$$b_{\text{Male}} = -2.4902,\ b_{\text{Child}} = 1.2491$$

Predictions

Example – Titanic Dataset

Accuracy is fairly high, but that is mitigated by the fact that most passengers did not survive: it would be a fairly safe bet to predict "No" on the survival front.

Matthew's Correlation Coefficient (MCC) is fairly high at 43.6% (completely random is 0%).

MCC: 43.6%
Accuracy: 76.0%
Pearson: 0.00276
Hist: 0.8%

		Predicted: Did not survive	Predicted: Survived	*Total*	
Actuals	**Did not survive**	340	35	*375*	*2.3%*
	Survived	102	93	*195*	*1.2%*
	Total	*442*	*128*	***570***	
		2.7%	*0.8%*		

Validation of Prediction Error Rate

The reliability of the prediction error rate observed in the model-building data set is examined by applying the chosen prediction rule to a validation data set.

If the new prediction error rate is about the same as that for the model-building data set, then the latter gives a reliable indication of the predictive ability of both the fitted logistic regression model and the chosen prediction rule.

If the new data leads to a considerably higher prediction error rate, then the fitted logistic regression model and the chosen prediction rule do not predict new observations as well as originally indicated.