Logistic Regression Models

Babak Shahbaba, Ph.D.

Associate Professor, Department of Statistics
University of California, Irvine

Irvine, CA

Overview

- Generalized linear models
- Logistic regression models with one explanatory variable
- Statistical inference using logistic regression models
- Multiple logistic regression models
- Model assessment and selection

Introduction

- For linear regression models, the response variable, Y, is assumed to be a real-valued continuous random variable.
- Y is modeled as a [stochastic] function of a set of explanatory variables (predictors) as follows:

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- The right-hand side is comprised of two parts
  - Systematic component: $\alpha + \beta_1 X_1 + \cdots + \beta_p X_p$
  - Normally distributed (with mean zero and constant variance σ^2) random error: ϵ

Introduction

- The systematic component gives the expected value (mean) of the response variable,

$$\mu_X = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Note that the mean is modeled as a function of predictors so it changes as the values of predictors change. (We usually drop the index X for simplicity.)
- The values of response variable are normally distributed around this mean with a constant variance (i.e., the variance does not change with X).

Introduction

- Using a sample of n observations from the population, we can estimate the regression parameters and find the estimated value of μ (i.e., mean of the response variable) as follows:

$$\hat{\mu} \quad = \quad a + b_1 x_1 + \cdots + b_p x_p$$

- We interpret $\hat{\mu}$ as our estimate for the expected (mean) of the response variable for subjects with values $x_1, \ldots, x_p$.

Introduction

- The regression model has therefore two components:
  - Systematic component: $\alpha + \beta_1 X_1 + \cdots + \beta_p X_p$
  - Random component: $Y \sim N(\mu_X, \sigma^2)$
- And the regression model connects the mean of the random component to the systematic component,

$$\mu_X = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p$$

Generalized Linear Models

- Now consider situations where the response variable is a binary random variable (e.g., disease status), count variable (e.g., number of accidents), or it is continuous, but it can only take positive values (e.g., tumor size).
- For such problems, we can still define the systematic component as before.
- However, it would not make sense to assume that the random component Y is normally distributed.
- Also, it would not make sense to simply set the mean of the random component equal to the systematic component since the systematic component can theoretically take any positive or negative real values while that would not be the case for the mean of such response variables.

Generalized Linear Models

- To address these issues, we use a more general class of linear models that avoid the restrictive assumptions of linear regression models.
- We call these models collectively **generalized linear models**, or GLM for short.
- Using GLM, we can use other family of distributions other than normal for the random component.
- For example, we use the Bernoulli (or Binomial) distribution family when the response variable is a binary variable.
- For count response variables, we typically use the Poisson distribution family.

Generalized Linear Models

- Also, instead of simply setting the mean of the response variable to the systematic component, we set some appropriate transformations of the mean equal to the systematic component.
- For count response variables with a Poisson distribution, we usually use the log transformation,

$$\log(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p$$

- The log function in this case is called the **link function**.
- Note that although μ here can take positive values only, $\log(\mu)$ can be both positive and negative.
- The resulting models are called **Poisson regression model**.

Generalized Linear Models

- For binary response variables with a Bernoulli distribution, it is common to use the **logit** (i.e., log-odds) transformation as the link function,

$$\log(\frac{\mu}{1-\mu}) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Note that binary variable $\mu = P(Y = 1|X)$; that is, μ is the probability of the outcome of interest (denoted as 1) given the explanatory variables.
- Although μ is a real number between 0 and 1, its logit transformation can be any real number between $-\infty$ to $+\infty$.
- The resulting models are called **logistic regression models**.

Generalized Linear Models

- For linear regression models, we used the least-squares method to estimate model parameters.
- For generalized linear models, we use an alternative method called **maximum likelihood estimation** (MLE).
- Informally, this is an optimization approach to find the parameter that make the observed data most probable according to our [probabilistic] model.
- We denote the resulting parameter estimates as $\hat{\alpha} = a$, $\hat{\beta}_1 = b_1, \ldots, \hat{\beta}_p = b_p$.

Logistic Regression Models

- In what follows, we discuss logistic regression models in details.
- For these models, we use the MLE and estimate μ as follows:

$$\log\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right) = a + b_1x_1 + \ldots + b_px_p$$

- We can exponentiate both sides,

$$\frac{\hat{\mu}}{1-\hat{\mu}} = \exp(a + b_1x_1 + \ldots + b_px_p)$$

- and find $\hat{\mu}$ using the **logistic function**,

$$\hat{\mu} = \frac{\exp(a + b_1x_1 + \ldots + b_px_p)}{1 + \exp(a + b_1x_1 + \ldots + b_px_p)}$$

Logistic Regression Models

- We can use R-Commander to find the maximum likelihood estimates of model parameters.
- In R-Commander, click `Statistics` → `Fit models` → `Generalized linear model`.
- Similar to linear regression models, we specify the model formula by entering the response variable on the left side of the "$\sim$" symbol and the explanatory (predictor) variables (separated by "$+$" sings) on the right side.

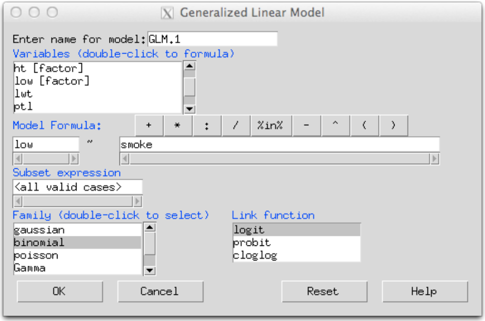
Logistic Regression Models

- Under `Family`, we specify the probability distribution of the response variable.
- In this case, the response variable is binary so we assume it has a Bernoulli distribution. It is of course more common to specify the probability distribution as binomial, which includes the Bernoulli distribution as a special case.
- After specifying the probability distribution of the response variable, R-Commander provides a list of possible link functions. The default for binary random variables is the logit link.

Logistic Regression with One Binary Predictor

- As an example, we use the `birthwt` data set to model the relationship between having low birthweight babies (a binary variable), Y, and smoking during pregnancy, X.
- The binary variable `low` identifies low birthweight babies (`low` $=1$ for low birthweight babies, and 0 otherwise).
- The binary variable `smoke` identifies mothers who were smoking during pregnancy (`smoke`$=1$ for smoking during pregnancy, and 0 otherwise).

Logistic Regression with One Binary Predictor



Enter name for model: GLM.1

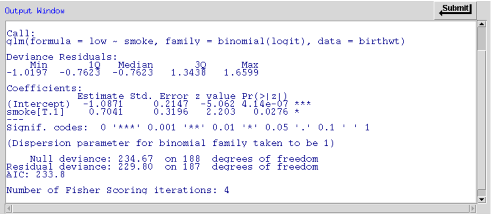
Variables (double-click to formula)
ht [factor]
low [factor]
lwt
ptl

Model Formula: low ~ smoke

Subset expression: <all valid cases>

Family (double-click to select): gaussian, binomial, poisson, Gamma — selected: binomial

Link function: logit, probit, cloglog — selected: logit



```
Call:
glm(formula = low ~ smoke, family = binomial(logit), data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.0197  -0.7623  -0.7623   1.3438   1.6599  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.0871     0.2147  -5.062 4.14e-07 ***
smoke[T.1]    0.7041     0.3196   2.203   0.0276 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 229.80  on 187  degrees of freedom
AIC: 233.8

Number of Fisher Scoring iterations: 4
```

Logistic Regression with One Binary Predictor

- For the above example, the estimated values of the intercept α and the regression coefficient β are $a = -1.09$ and $b = 0.70$ respectively.
- Therefore,

$$\frac{\hat{\mu}_x}{1-\hat{\mu}_x} \quad = \quad \exp(-1.09 + 0.70x)$$

- Here, $\hat{\mu}_x$ is the estimated probability of having a low birthweight baby for a given x.
- The left-hand side of the above equation is the estimated odds of having a low birthweight baby.

Logistic Regression with One Binary Predictor

- For non-smoking mother, $x = 0$, the odds of having low birthweight baby is

$$\begin{aligned} \frac{\hat{\mu}_0}{1 - \hat{\mu}_0} &= \exp(-1.09) \\ &= 0.34 \end{aligned}$$

- That is, the exponential of the intercept is the odds when $x = 0$, which is sometimes referred to as the **baseline odds**.

Logistic Regression with One Binary Predictor

- For mothers who smoke during pregnancy, $x = 1$, and

$$
\begin{aligned}
\frac{\hat{\mu}_1}{1 - \hat{\mu}_1} &= \exp(-1.09 + 0.7) \\
&= \exp(-1.09)\exp(0.7)
\end{aligned}
$$

- As we can see, corresponding to one unit increase in x from $x = 0$ (non-smoking) to $x = 1$ (smoking), the odds multiplicatively increases by the exponential of the regression coefficient.

Logistic Regression with One Binary Predictor

- Note that

$$\frac{\frac{\hat{\mu}_1}{1-\hat{\mu}_1}}{\frac{\hat{\mu}_0}{1-\hat{\mu}_0}} = \frac{\exp(-1.09)\exp(0.7)}{\exp(-1.09)}$$
$$= \exp(0.7)$$

- We can interpret the exponential of the regression coefficient as the odds ratio of having low birthweight babies for smoking mothers compared to non-smoking mothers.
- Here, the estimated odds ratio is $\exp(0.7) = 2.01$ so the odds of having a low birthweight baby almost doubles for smoking mothers compared to non-smoking mothers.

Logistic Regression with One Binary Predictor

- In general,
  - if $\beta > 0$, then $\exp(\beta) > 1$ so the odds increases as X increases;
  - if $\beta < 0$, then $0 < \exp(\beta) < 1$ so the odds decreases as X increases;
  - if $\beta = 0$, the odds ratio is 1 so the odds does not change with X according to the assumed model.

Confidence Interval

- We saw how $\exp(a)$ and $\exp(b)$ can be interpreted as our point estimates for the baseline odds and odds ratio respectively.
- However, it would be more informative if we present our estimates in terms of confidence intervals, which reflect the extent of our uncertainty.
- Given the point estimates a and b, we find the confidence intervals for the regression parameters as follows:

$$[a - z_{\mathrm{crit}} \times SE_a, a + z_{\mathrm{crit}} \times SE_a]$$
$$[b - z_{\mathrm{crit}} \times SE_b, b + z_{\mathrm{crit}} \times SE_b]$$

- Here, SE_a and SE_b are the standard errors of our estimators, and z_{crit} is the z-critical value for the required confidence level.

Confidence Interval

- For our example, the 95% confidence intervals of α is

$$[-1.09 - 2 \times 0.21, -1.09 + 2 \times 0.21] \quad = \quad [-1.51, -0.67].$$

- Because the baseline odds (i.e., the odds when $x = 0$) is $\exp(\alpha)$, the 95% confidence interval for the baseline odds is obtained as follows:

$$[\exp(-1.51), \exp(-0.67)] \quad = \quad [0.22, 0.51].$$

- The 95% confidence interval of β is

$$[0.70 - 2 \times 0.32, 0.70 + 2 \times 0.32] \quad = \quad [0.06, 1.34].$$

- Because the odds ratio is $\exp(\beta)$, the 95% confidence interval of the odds ratio is obtained as follows:

$$[\exp(0.06), \exp(1.34)] \quad = \quad [1.06, 3.82].$$

Confidence Interval

- Alternatively, we can use R-Commander to obtain the confidence intervals for regression parameters, baseline odds, and odds ratio.
- To do this, after you follow the above steps to fit the logistic regression model to the data, click `Models` → `Confidence intervals`, specify the confidence level, and select the `Wald statistic` option.

Hypothesis Testing

- The regression coefficient β captures the relationship between the response variable Y and the explanatory variable X.
- In logistic regression model, the assumption is that if the two variables are related, log odds for the response variable changes linearly with X.
- Using logistic regression models, we can specify our hypothesis regarding the relationship between X and Y in terms of β as $H_A : \beta > 0$, or $H_A : \beta < 0$, or $H_A : \beta \neq 0$; It is common to use two-sided alternative hypotheses.
- In contrast the null hypothesis is specified as $H_0 : \beta = 0$.

Hypothesis Testing

- As before, we use the observed significance level, i.e., p-value, to decide whether we should reject the null hypothesis.
- To do this, we first need to find the z-score by dividing b (i.e., the point estimate of β) by its standard error.
- Then, we find the p-value by calculating the probability of as or more extreme values than the z-score under the null.
- To assess the null hypothesis $H_0 : \beta = 0$, we first calculate the $z = b/SE_b$ and find the corresponding p-value as follows:

$$\begin{array}{ll} \text{if } H_A : \beta < 0, & p_{\text{obs}} = P(Z \le z), \\ \text{if } H_A : \beta > 0, & p_{\text{obs}} = P(Z \ge z), \\ \text{if } H_A : \beta \ne 0, & p_{\text{obs}} = 2 \times P\big(Z \ge |z|\big), \end{array}$$

where Z has the standard normal distribution.

Hypothesis Testing

- For the above example

$$z = \frac{0.70}{0.32} = 2.2.$$

  This is of course what R-Commander provides under the "z value" column.
- The corresponding p-value is 0.028, which is also provided by R-Commander under "Pr(>|z|)".
- In this example, we can reject the null hypothesis at 0.05 significance level.

Prediction

- We can use logistic regression models for predicting the unknown values of the response variable Y given the value of the predictor value X.

$$\hat{\mu}_x = \frac{\exp(a + bx)}{1 + \exp(a + bx)}$$

- For the above example,

$$\hat{\mu}_x = \frac{\exp(-1.09 + 0.70x)}{1 + \exp(-1.09 + 0.70x)}$$

Prediction

- Therefore, the estimated probability of having a low birthweight baby for non-smoking mothers, $x = 0$, is

$$\hat{\mu}_x = \frac{\exp(-1.09)}{1 + \exp(-1.09)} = 0.25$$

- This probability increases for mothers who smoke during pregnancy,

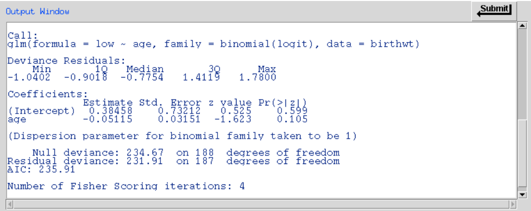
$$\hat{\mu}_x = \frac{\exp(-1.09 + 0.7)}{1 + \exp(-1.09 + 0.7)} = 0.40$$

- That is, the risk of having a low birthweight baby increases by 60% if a mother smokes during her pregnancy.

Activity 1

Logistic Regression with One Numerical Predictor

- For the most part, we follow similar steps to fit the model, estimate regression parameters, perform hypothesis testing, and predict unknown values of the response variable.
- As an example, we want to investigate the relationship between having a low birthweight baby, Y, and mother's age at the time of pregnancy, X.



```
Call:
glm(formula = low ~ age, family = binomial(logit), data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0402  -0.9018  -0.7754   1.4119   1.7800

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38458    0.73212   0.525    0.599
age         -0.05115    0.03151  -1.623    0.105

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 231.91  on 187  degrees of freedom
AIC: 235.91

Number of Fisher Scoring iterations: 4
```

Logistic Regression with One Numerical Predictor

- Finding confidence intervals and performing hypothesis testing remain as before, so we focus on prediction and interpreting the point estimates.
- For the above example, the point estimates for the regression parameters are $a = 0.38$ and $b = -0.05$.
- While the intercept is the log odds when $x = 0$, it is not reasonable to interpret its exponential as the baseline odds since mother's age cannot be zero.

Logistic Regression with One Numerical Predictor

- To interpret b, consider mothers whose age is 20 years old at the time of pregnancy,

$$\begin{aligned}\log\left(\frac{\hat{\mu}_{20}}{1-\hat{\mu}_{20}}\right) &= 0.38 - 0.05 \times 20 \\ \frac{\hat{\mu}_{20}}{1-\hat{\mu}_{20}} &= \exp(0.38 - 0.05 \times 20) \\ &= \exp(0.38)\exp(-0.05 \times 20)\end{aligned}$$

- For mothers who are one year older (i.e., one unit increase in age), we have

$$\begin{aligned}\log\left(\frac{\hat{\mu}_{21}}{1-\hat{\mu}_{21}}\right) &= 0.38 - 0.05 \times 21 \\ \frac{\hat{\mu}_{21}}{1-\hat{\mu}_{21}} &= \exp(0.38 - 0.05 \times 21) \\ &= \exp(0.38)\exp(-0.05 \times 21)\end{aligned}$$

Logistic Regression with One Numerical Predictor

- The odds ratio for comparing 21 year old mothers to 20 year old mothers is

$$\begin{aligned}\frac{\frac{\hat{\mu}_{21}}{1-\hat{\mu}_{21}}}{\frac{\hat{\mu}_{20}}{1-\hat{\mu}_{20}}} &= \frac{\exp(0.38)\exp(-0.05\times 21))}{\exp(0.38)\exp(-0.05\times 20)}\\ &= \exp(-0.05\times 21+0.05\times 20)\\ &= \exp(-0.05)\end{aligned}$$

- Therefore, $\exp(b)$ is the estimated odds ratio comparing 21 year old mothers to 20 year old mothers.
- In general, $\exp(b)$ is the estimated odds ratio for comparing two subpopulations, whose predictor values are $x+1$ and x,

$$\frac{\frac{\hat{\mu}_{x+1}}{1-\hat{\mu}_{x+1}}}{\frac{\hat{\mu}_{x}}{1-\hat{\mu}_{x}}} = \exp(b)$$

Logistic Regression with One Numerical Predictor

- As before, we can use the estimated regression parameters to find $\hat{\mu}_x$ and predict the unknown value of the response variable.

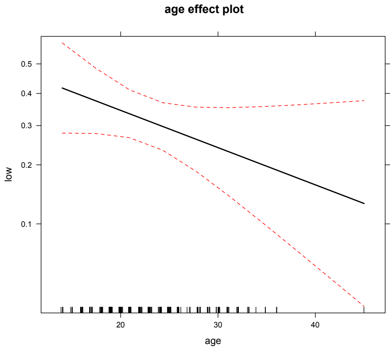
$$\begin{aligned}\hat{\mu}_x &= \frac{\exp(a + bx)}{1 + \exp(a + bx)} \\ &= \frac{\exp(0.38 - 0.05x)}{1 + \exp(0.38 - 0.05x)}.\end{aligned}$$

- For example, for mother who are 20 years old at the time of pregnancy, the estimated probability of having a low birthweight baby is

$$\hat{\mu}_x = \frac{\exp(0.38 - 0.05 \times 20)}{1 + \exp(0.38 - 0.05 \times 20)} = 0.35.$$

Logistic Regression with One Numerical Predictor

- We can use R-Commander to visualize how the estimated probability changes for different values of X.
- After you fit the logistic regression model to the data, click `Models` → `Graphs` → `Effect plots`.



Activity 2

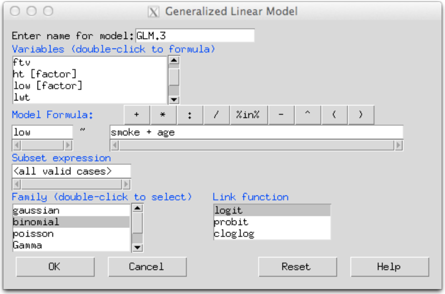
Multiple Logistic Regression

- We now discuss logistic regression models with multiple explanatory (predictor) variables.
- For example, suppose we want to investigate the relationship between smoking during pregnancy and having a low birthweight baby, but we suspect that mother's age at the time of pregnancy might influence this relationship.

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta_1 \mathrm{smoke} + \beta_2 \mathrm{age}$$

where μ is the probability of having a low birthweight baby for given values of `smoke` and age.

Multiple Logistic Regression



Generalized Linear Model

Enter name for model: GLM.3

Variables (double-click to formula)

- ftv
- ht [factor]
- low [factor]
- lwt

Model Formula:

low ~ smoke + age

Subset expression

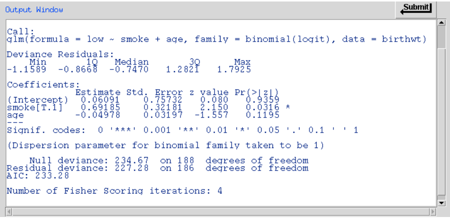
<all valid cases>

Family (double-click to select)

- gaussian
- binomial
- poisson
- Gamma

Link function

- logit
- probit
- cloglog



Output Window

```
Call:
glm(formula = low ~ smoke + age, family = binomial(logit), data = birthwt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1589  -0.8668  -0.7470   1.2821   1.7925

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.06091    0.75732   0.080   0.9359
smoke[T.1]   0.69185    0.32181   2.150   0.0316 *
age         -0.04978    0.03197  -1.557   0.1195
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 227.28  on 186  degrees of freedom
AIC: 233.28

Number of Fisher Scoring iterations: 4
```

Multiple Logistic Regression

- As before, finding confidence intervals and performing hypothesis testing remain similar to what we discussed for models with a single predictor, so we focus on prediction and interpretation of point estimates.
- For the above example, the estimated value of α is $a = 0.06$.
- In general, this could be interpreted as the odds when all predictors are set to zero. In this case, however, mother's age cannot be zero.

Multiple Logistic Regression

- The estimated value of β_1, i.e., the regression coefficient of smoke, is $b_1 = 0.69$.
- In general, this is interpreted in terms of the odds ratio corresponding to one unit increase in its related predictor *when all other predictors are kept constant*.
- Here, $\exp(0.69) = 1.99$ is the estimated odds ratio of having a low birthweight baby comparing smoking to non-smoking mothers with the same age (i.e., keeping age constant).
- Therefore, the odds of having a low birthweight baby almost doubles for smoking mothers compared to non-smoking mothers of the same age.

Multiple Logistic Regression

- The estimated value of β_2, i.e., the regression coefficient of age, is $b_2 = -0.05$.
- The estimated odds ratio for comparing two groups of mothers with 1 year age difference (e.g., comparing 20 year old mothers to 19 year old mothers) with the same smoking status (i.e., keeping `smoke` constant) is $\exp(-0.05) = 0.95$.
- Of course, this result is not statistically signifiant since p-value=0.12 is above the commonly used significance levels.

Multiple Logistic Regression

- We can use our model to estimate the risk of having a low birthweight baby,

$$\hat{\mu}_x = \frac{\exp(a + b_1 \mathrm{smoke} + b_2 \mathrm{age})}{1 + \exp(a + b_1 \mathrm{smoke} + b_2 \mathrm{age})}$$

- For a 20 year old mother who smokes during her pregnancy, the estimated probability of having a low birthweight baby is

$$\begin{aligned} \hat{\mu}_x &= \frac{\exp(0.06 + 0.69 \times 1 - 0.05 \times 20)}{1 + \exp(0.06 + 0.69 \times 1 - 0.05 \times 20)} \\ &= 0.44 \end{aligned}$$

Activity 3

Model Assessment

- We typically build regression models either for hypothesis testing or prediction.
- When our objective is to perform hypothesis testing, the set of variables we include in our model is dictated by our hypothesis and our domain knowledge.

Exploratory Analysis

- Very often, we do not have a pre-specified hypothesis in mind when analyzing the data; that is, our study is exploratory.
- In such cases, we usually want to conduct a preliminary analysis in order to generate a set of hypotheses, which can be evaluated in followup studies.
- A common approach is to try models with different sets of predictors (i.e., different systematic components) and choose the one with the best fit, or the lowest model-data discrepancy.
- The **deviance** is a common measure of discrepancy (i.e., lack of fit): the lower the deviance, the better the fit.
- The value of deviance is provided by R and R-Commander.

Exploratory Analysis

- However, the fit (and deviance) improves as we increase the number of predictors.
- Akaike proposed to penalize against model complexity by adding $2k$, where k is the number of model parameters, to the deviance.
- The resulting measure is called Akaike Information Criterion (AIC),

$$AIC = Deviance + 2k$$

- AIC is provided by R and R-Commander.
- Among different models, we choose the one with the lowest AIC.

Predictive Modeling

- When our objective is to predict the unknown values of the response variable, we should include in the systematic component any predictor that would improve the predictive power of our model.
- A common measure for predictive power is *accuracy rate*, which is defined as the proportion of the times the correct category (0 or 1 in this case) is predicted for future observations (or observations in the test set).

Predictive Modeling

- Note that the outputs of logistic regression models are in fact between 0 and 1, which are interpreted as probabilities.
- Therefore, we need to set an appropriate cutoff to obtain binary predictions, $\hat{y}$.
- In general, the cutoff depends on the loss function; that is, the cost of predicting the class as 0, when the true class is 1, and vice versa.
- For simplicity, it is common to assign a test case to the class with the highest probability; that is, we set the cutoff at 0.5.

Predictive Modeling

- Instead of averaging over all predictions, it might be more informative to separate the types of error.
- One common approach for doing this is to present the results in a *classification table* as follows:

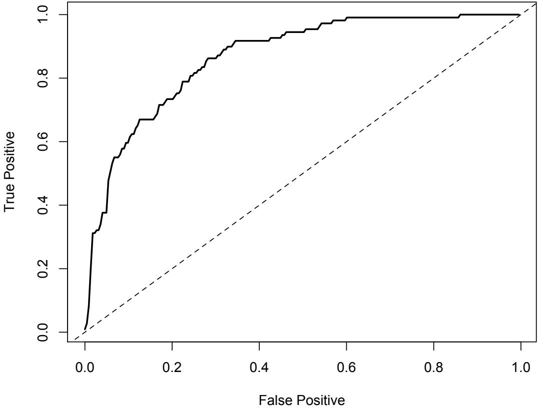
		Predicted class	
		0	1
True class	0	True Negative	False Positive
	1	False Negative	True Positive

- True positive rate is also called **sensitivity**, and true negative rate is called **specificity**.

Predictive Modeling

- Instead of setting an arbitrary cutoff (e.g., 0.5), we can use Receiver Operating Characteristic (ROC) curves, which allow for simultaneous consideration of sensitivity and specificity without setting an arbitrary cut-off.
- The curve plots true positive rate (sensitivity) as a function of false positive rate (1-specificity) for all possible cutoffs from 0 to 1.

Predictive Modeling



Predictive Modeling

- The area under the curve (AUC) is a measure of overall performance: higher AUC corresponds to higher predictive power.
- If the AUC is one, then the value of the response variable is always predicted correctly.
- If it is 0.5, the predictive model is equivalent to tossing a fair coin, e.g. useless.
- R has several packages (e.g., pROC, ROCR) for plotting ROC curves and calculating their AUC.