# Logistic Regression

*Kris Gunsalus*

*11/14/2018*

## Contents



### Background Reading

- Introduction to Statistical Learning, Chapter 4: Classification, through section 4.3.
- Dalgaard, Chapter 13: Logistic Regression
- [ Aho, Chapter 9.20 ]

### Predicting qualitative responses

Last week we learned about ANOVA and linear models, which we can use to predict quantitative outcomes using categorical and quantitative explanatory variables. Sometimes instead we wish to model **qualitative outcomes** using **quantitative predictors**. Examples could be:

- Probabilty of credit card default based on income, balance, other factors
- Probability that a patient has a disease given a particular set of symptoms
- Disease prognosis given gene expression data (e.g. malignant or benign cancer)

These kinds of questions are **classification** problems, where we wish to predict the probability that $Y$ belongs to a particular category given some data.

In cases like these, when we want to predict simple binary outcomes (or multiple categorical outcomes), simple linear models are not appropriate. *Why is this?*

Consider the following "Default" dataset from *Introduction to Statistical Learning*, which contains credit card data for 10,000 individuals, including their incomes, balances, and student status. A linear model would produce something like this:

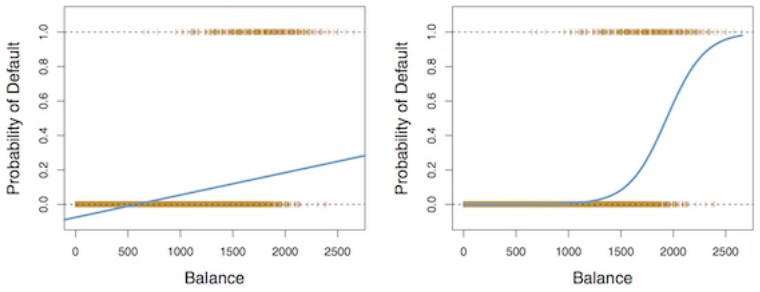


**FIGURE 4.2.** *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Figure 1: Linear vs. logistic regression

$$
\begin{aligned}
Y &= \beta_o + \sum_{i=1}^{n} \beta_i X_i + \epsilon \\
&= \beta_o + \beta_1 * income + \beta_2 * balance + \beta_3 * student + \epsilon
\end{aligned}
$$

The model may even contain some interaction terms if any of these variables are not independent from each other.

Say we to ask, "Does the chance of a credit card default depend on credit card balance? What is the probability of default given a particular balance?"

To use a linear model we could code the binary outcome with *dummy variables* as either **0 (no default)** or **1 (default)** and model a linear response. However, linear regression will produce some estimates that fall outside of the range $[0, 1]$, which doesn't make sense if we are trying to predict the *probability* of a particular binary outcome. For the Default data, low balances give a negative probability, and very high balances give a probability above 1 (Figure 1, left).

Instead, we want a model that reliably outputs values between zero and one, as a function of the explanatory variables. **Logistic regression** provides a natural way for us to do this (Figure 1, right). Now we predict a probability close to 0 for a low balance and close to 1 for a high balance. The function produces an **S-shaped** curve, and gives the same average probability of default (0.033) as the linear model.

## Logistic function

Logistic regression uses the **logistic function** to model the probability of a particular outcome $Y$, given the data $X$ – for example, the probability of default given a particular credit card balance. If we write this as $Pr(Y = 1|X)$ and use $p(X)$ as a short-hand, then the logistic function, *LOGIS(1,0)*, is:

$$
p(X) = Pr(Y = 1|X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_o + \beta_1 X)}}
$$

It can be seen that this equation ranges from 0 to 1, as desired. The probability is very close to 0 when the exponent is small, and is close to 1 when it is very large (because one of the terms in the denominator becomes neglibible). However, the logistic function is rather complicated to compute, since it is not linear.

### Odds Ratio

Rearranging the terms above gives the **odds ratio** for $p(X)$, which can take on any value between 0 and $\infty$:

$$Odds = \frac{p(X)}{1 - p(X)} = e^{\beta_o + \beta_1 X}$$

Odds close to zero give low probabilities, and odds close to infinity give high probabilities. Odds are commonly used in betting games such as horse racing, where people are interested in the odds of their favorite horse winning (the numerator) vs. losing (the denominator).

The *odds of success* are 1 (1 to 1, i.e. 1:1) when p = 1/2, meaning that one out of two outcomes will be positive. When p=0.9, the odds are 9 (9 to 1, or 9:1). For the bank default dataset, this Odds = 9 would mean that 9 out of 10 people with a particular balance will default.

Thought question: For the bank default example, what does p = 0.2 mean? What are the odds?

### Log-odds, a.k.a. logit function

With the expression above, we now have something that can easily be transformed into a linear function by taking the log of both sides. This is called the **log-odds** or **logit** function:

$$Log\ odds = log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_o + \beta_1 X$$

So now we have transformed the **non-linear** *logistic* function of $X$ into a *logit* function that is **linear** in $X$, which greatly simplifies the math required to solve for $p(X)$.

In contrast with the linear model, where the coefficient $\beta_1$ gives the change in $Y$ per unit change in $X$, for the logistic model $\beta_1$ gives the *change in the log-odds* for each increment of $X$. Note that is the same as multiplying the odds by $e^{\beta_1}$.

It is important to remember that the change in $p(X)$ itself is *not* linear in $X$; that is, **the rate of change in $p(X)$ will depend on the current value of $X$**. The function is still *monotonic*, meaning that an increase in $X$ will always be reflected by an increase in $p(X)$.

## Maximum likelihood

To estimate the coefficients $\beta_o$ and $\beta_1$, which are unknown, instead of using the method of least squares, a more general method called **maximum likelihood (ML)** is used. In fact, it turns out that the least squares approach used in linear regression is a special case of ML.

The intuition behind ML is that it seeks to find coefficients for the logistic function such that the predicted outcome for each individual, $\hat{p}(x_i)$, most closely matches the observed data. In other words, the predicted probability will be close to 1 for positive outcomes, and it will be close to 0 for negative outcomes (based on training data). Mathematically, this is achieved by finding the coefficients that maximize the **likelihood function**:

$$\ell(\beta_o, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=o} (1 - p(x_{i'}))$$

We want the first term to be big when the actual outcomes are positive ($y_i = 1|x_i$), and we want the second term to be big when they are negative ($y_i = 0|x_i$).

Once the coefficients $\beta_o$ and $\beta_1$ have been found, they can be used to solve for the predicted probability $\hat{p}(X)$ of a positive outcome (here, credit card default) for any particular value of $X$ by plugging them back into the equation above for $p(X)$. For example, for the Default dataset, the probability of default given a balance of \$1,000 is 0.6%, whereas for a balance of \$2,000 the probability is 59% (see Figure 1).

Notice that taking the log of the likelihood function will give a linear function in $p(x)$, which is called the **conditional log likelihood**:

$$LCL = \log[\ell(\beta_o, \beta_1)] = \sum_{i:y_i=1} \log p(x_i) + \sum_{i':y_{i'}=o} \log(1 - p(x_{i'}))$$

## Using binary predictors

It is also possible to use logistic regression to model a binary response to qualitative (categorical) predictors. For example, if we want to find the probability of default given the status of "student", we can code the status using a dummy variable (1 = student; 0 = non-student) and model the probability of default using logistic regression.

The model provides a *z-statistic*, which is analagous to a $t$-statistic for linear regression.

For this example, we find that *student* status is significantly associated with "default = Yes" (ISL Table 4.2: z-score = 3.52, p-value = 0.0004). The **coefficient is positive**, indicating that *students generally have a higher rate of credit card default than do non-students* (though the overall rate is still low, around 4.3% vs. 2.9% for non-students).

Extending the model to qualitative predictors with more than two states is problematic for categorical data, which have no intrinsic quantitative relationship. If ordinal data can be represented using constant intervals, these can be used as pseudo-quantitative variables.

### What if I have more than two response classes?

Logistic regression may be extended to multiple classes, but other methods are preferred for multiple-class classification, such as **linear discriminant analysis (LDA)**.

In contrast to logistic regression, which models $Pr(Y = k|X = x)$ directly, LDA first models the distribution of predictors $X$ in each response class $Y$, and then flips these around using Bayes' theorem to obtain estimates for $Pr(Y = k|X = x)$.

LDA has better statistical properties (more stable estimators) when classes are well separated, or if $n$ is small and the distribution of predictors is approximately normal in each class.

## Multiple logistic regression

Just like wtih linear regression, logistic regression can be extended to model a binary outcome using multiple predictors $X = (X_1, X_2, ...X_p)$:

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_o + \beta_1 X + ... + \beta_p X_p$$

The coefficients for the model are estimated using maximum likelihood, as above.

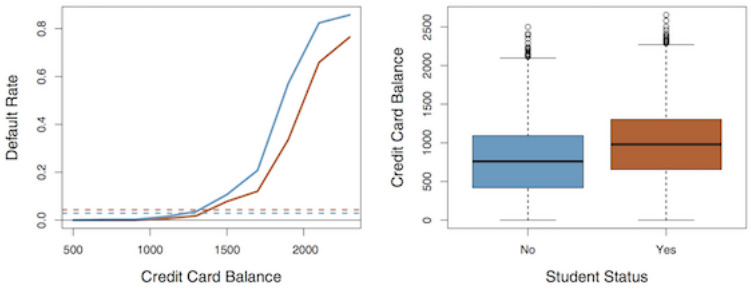


**FIGURE 4.3.** *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

Figure 2: Confounding variables

### Confounding

Multiple regression can be used to model the contributions of all the variables in the Default dataset at the same time. A funny thing happens when we use multiple regression to model the full Default dataset. Now we see that balance and student status are both significant predictors for default: $p(balance) =< 0.0001$ and $p(student) = 0.0062$ (ISL Table 4.3).

However a strange thing has happened: instead of the coefficient for student being positive, it is now negative! How can this be? As it turns out, **student** and **balance** are **correlated** (Figure 2, right panel).

In the left-hand panel, we see that students have a higher overall rate of default (dashed lines) – hence the positive z-score in the simple regression model. However, for the same amount of debt, students are *less likely* to default than non-students (solid lines) – hence the negative z-statistic in the multiple regression model.

Thus, students tend to have higher debt overall, which is associated with higher default rates, but for any given level of debt, they are less likely to default than their non-student peers. For a debt of $1500, for example, non-students are predicted to default at a rate of around 10%, whereas students would default only around 6% of the time.

## Generalized linear models

GLMs provide a framework for formulating a variety of linear and non-linear relationships between response variables $Y$ and predictors $X$. Both linear models and logistic models may be considered special cases of GLMs.

- Linear models are appropriate for continuous or interval data when there is a linear relationship variables.
- GLMs can incorporate categorical data as both predictors (like ANOVA) and response variables.
- GLMs can accommodate *bounded* response variables, which linear models cannot.

Linear models make two assumptions that GLMs do not: a *linear association* between $X$ and $Y$, and *normal* error distributions. GLMs handle cases where these assumptions do not apply. The two components of GLMs that distinguish them from simpler linear models are:

- A **link function** that transforms the response variable to allow expression of the mean function in linear terms of the predictors.
- A user-defined **error distribution** that provides ML optimization criteria and allows inferences about model parameters.

| Name | Definition | Synonyms |
|---|---|---|
| False Pos. rate | FP/N | Type I error, 1−Specificity |
| True Pos. rate | TP/P | 1−Type II error, power, sensitivity, recall |
| Pos. Pred. value | TP/P* | Precision, 1−false discovery proportion |
| Neg. Pred. value | TN/N* | |

**TABLE 4.7.** *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Figure 3: Synonyms related to classification

#### GLM Families

Does the likelihood function above look familiar? Remember the **binomial distribution**? Logistic regression belongs to a family of *binomial* GLMs where the outcomes are dichotomous (i.e. binomial). Logistic regression has a *binomial response distribution* provided by the *logit function*, $p = log[p/(1-p)]$.

In R, **logistic regression is performed using the `glm` function**, specifying `family=binomial("logit")`, or sinply `family="binomial"` (which defaults to the logit function).

GLMs allow a variety of error term distributions to be specified. The three most common are *normal*, *binomial*, and *Poisson*. Poisson GLMs are often applied when the response variable is count data. The Poisson link function is $log\lambda$, where $\lambda$ is the Poisson mean. Common family specifications are:

- Binomial: `family=binomial(link="logit")`
- Poisson: `family=poisson(link="log")`
- Linear: `family=Gaussian(link="identity")` – However, these are handled more efficiently by `lm`, which is preferred for linear models.

## ROC and AUC

A diagram called the **Receiver Operating Characteristic (ROC)** curve can be used to assess the efficacy of a GLM (**Figure 4.8**). The ROC plots the **True positive rate(sensitivity)** vs. **False-positive rate (1-specificity)** for all possible values. **Table 4.7** lists the various synonyms you will see for terms related to classification accuracy.

The **AUC**, or area under the curve, will be a diagonal on this plot if the classification is no better than random. A model where the explanatory variables are very good at predicting outcomes will push the curve toward the top left-hand corner.

## Exercise

We will see how all these concepts fit together in the class exercise, which uses characteristics of tumors to classify breast cancers as either benign or malignant.

## References

Introduction to Statistical Learning, Chapter 4

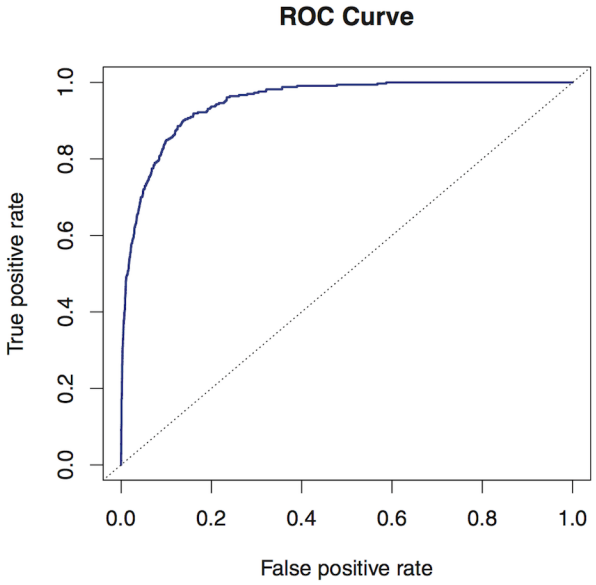


**FIGURE 4.8.** *A ROC curve for the LDA classifier on the* `Default` *data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate. The dotted line represents the "no information" classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.*

Figure 4: ROC Curve: higher area under the curve (AUC) indicates better predictive power