Logistic Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multilevel Regression Modeling, 2009

Logistic Regression

Introduction

In this lecture we discuss the logistic regression model, generalized linear models, and some applications.

Probability Theory Background

Before beginning our discussion of logistic regression, it will help us to recall and have close at hand a couple of fundamental results in probability theory.

A Binary 0,1 (Bernoulli) Random Variable I

Suppose a random variable Y takes on values 1,0 with probabilities p and $1 - p$, respectively.

Then Y has a mean of

$$E(Y) = p$$

and a variance of

$$\sigma_y^2 = p(1 - p)$$

Proof I

Proof.

1. Recall from Psychology 310 that the expected value of a discrete random variable Y is given by

$$E(Y) = \sum_{i=1}^{K} y_i \Pr(y_i)$$

That is, to compute the expected value, you simply take the sum of cross-products of the outcomes and their probabilities. There is only one nonzero outcome, 1, and it has a probability of p.

Proof II

2. When a variable Y takes on only the values 0 and 1, then $Y = Y^2$. So $E(Y) = E(Y^2)$. But one formula for the variance of a random variable is $\sigma_y^2 = E(Y^2) - (E(Y))^2$, which is equal in this case to

$$\sigma_y^2 = p - p^2 = p(1-p)$$

Conditional Distributions in the Bivariate Normal Case

If two variables W and X are bivariate normal with regression line $\hat{W} = \beta_1 X + \beta_0$, and correlation ρ, the conditional distribution of W given $X = a$ has mean $\beta_1 a + \beta_0$ and standard deviation $\sigma_\epsilon = \sqrt{1 - \rho^2}\sigma_w$.

If we assume X and W are in standard score form, then the conditional mean is

$$\mu_{w|x=a} = \rho a$$

and the conditional standard deviation is

$$\sigma_\epsilon = \sqrt{1 - \rho^2}$$

An Underlying Normal Variable

It is easy to imagine a continuous normal random variable W underlying a discrete observed Bernoulli random variable Y. Life is full of situations where an underlying continuum is scored "pass-fail."

Let's examine the statistics of this situation.

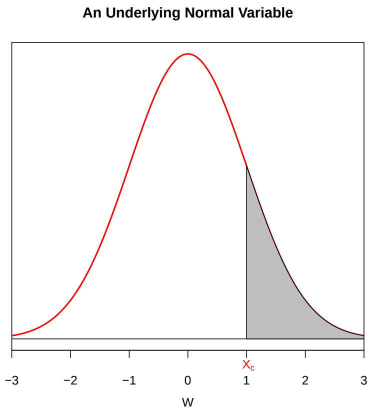
An Underlying Normal Variable

As a simple example, imagine that:

1. The distribution of scores on variable W has a standard deviation of 1, but varies in its mean depending on some other circumstance
2. There is a *cutoff score* X_c, and that to succeed, an individual needs to exceed that cutoff score. That cutoff score is $+1$.
3. What percentage of people will succeed if $\mu_w = 0$?

An Underlying Normal Variable

Here is the picture: What percentage of people will succeed?



An Underlying Normal Variable

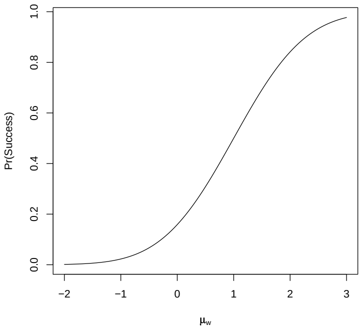
Suppose we wished to plot the probability of success as a function of μ_w, the mean of the underlying variable.

Assuming that σ stays constant at 1, and that W_c stays constant at $+1$, can you give me an R expression to compute the probability of success as a function of μ_w? (C.P.)

Plotting the Probability of Success

The plot will look like this:

```
> curve(1-pnorm(1,x,1),-2,3,
+ xlab=expression(mu[w]),ylab="Pr(Success)")
```



Plotting the Probability of Success

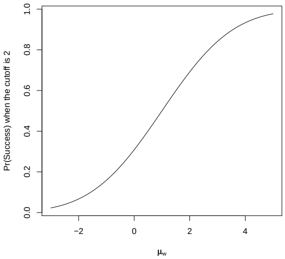
Note that the plot is non-linear. Linear regression will not work well as a model for the variables plotted here.

In fact, a linear regression line will, in general, predict probabilities less than 0 and greater than 1!

Plotting the Probability of Success

We can generalize the function we used to plot the previous figure for the general case where W_c is any value, and μ_w and σ_w are also free to vary.

```
> Pr.Success ← function(mu_w,sigma_w,cutoff)
+ {1-pnorm(cutoff,mu_w,sigma_w)}
> curve(Pr.Success(x,2,1),-3,5,
+ xlab=expression(mu[w]),
+ ylab="Pr(Success)␣when␣the␣cutoff␣is␣2")
```



Extending to the Bivariate Case

Suppose that we have a continuous predictor X, and a binary outcome variable Y that in fact has an underlying normal variable W generating it through a threshold values W_c. Assume that X and W have a bivariate normal distribution, are in standard score form, and have a correlation of ρ.

We wish to plot the probability of success as a function of X, the predictor variable.

Predicting Pr(Success) from X

We have everything we need to solve the problem. We can write

$$
\begin{aligned}
\pi(x) &= \Pr(Y = 1 | X = x) \\
&= \Pr(W > W_c | X = x) \\
&= 1 - \Phi\left(\frac{W_c - \mu_{W|X=x}}{\sigma_{W|X=x}}\right) \\
&= 1 - \Phi\left(\frac{W_c - \rho x}{\sqrt{1-\rho^2}}\right) \qquad (1) \\
&= \Phi\left(\frac{\rho x - W_c}{\sqrt{1-\rho^2}}\right) \qquad (2)
\end{aligned}
$$

Predicting Pr(Success) from X

Note that the previous equation can be written in the form

$$\pi(x) = \Phi(\beta_1 x + \beta_0) \tag{3}$$

Not only is the regression line nonlinear, but the variable Y is a Bernoulli variable with a mean that changes as a function of x, and so its variance also varies as a function of x, thus violating the equal variances assumption.

Predicting Pr(Success) from X

However, since $\Phi(\ )$ is invertible, we can write

$$\begin{aligned}\Phi^{-1}(\Pr(Y=1|X=x)) &= \Phi^{-1}(\mu_{Y|X=x}) \\ &= \beta_1 x + \beta_0 \\ &= \beta' x\end{aligned}$$

This is known as a *probit* model, but it is also our first example of a *Generalized Linear Model*, or GLM. A GLM is a linear model for a transformed mean of a variable that has a distribution in the natural exponential family. Since x might contain several predictors and very little would change, the extension to multiple predictors is immediate.

Binary Logistic Regression

Suppose we simply assume that the response variable has a binary distribution, with probabilities π and $1 - \pi$ for 1 and 0, respectively. Then the probability density can be written in the form

$$\begin{aligned} f(y) &= \pi^y (1-\pi)^{1-y} \\ &= (1-\pi)\left(\frac{\pi}{1-\pi}\right)^y \\ &= (1-\pi)\exp\left(y \log \frac{\pi}{1-\pi}\right) \end{aligned} \tag{4}$$

Binary Logistic Regression

The *logit* of Y is the logarithm of the odds that $Y = 1$.
Suppose we believe we can model the logit as a linear function of X, specifically,

$$\begin{aligned} \text{logit}(\pi(x)) &= \log \frac{\Pr(Y = 1|X = x)}{1 - \Pr(Y = 1|X = x)} \quad (5) \\ &= \beta_1 x + \beta_0 \quad (6) \end{aligned}$$

Binary Logistic Regression

The logit function is invertible, and exponentiating both sides, we get

$$
\begin{aligned}
\pi(x) &= \Pr(Y = 1|x) \\
&= \frac{\exp(\beta_1 x + \beta_0)}{1 + \exp(\beta_1 x + \beta_0)} \\
&= \frac{1}{1 + \exp(-(\beta_1 x + \beta_0))} \\
&= \frac{1}{1 + \exp(-\beta' x)} \qquad (7) \\
&= \mu_{Y|X=x}
\end{aligned}
$$

Once again, we find that a transformed conditional mean of the response variable is a linear function of X.

Extension to Several Predictors

Note that we wrote $\beta_1 x + \beta_0$ as $\beta' x$ in the preceding equation.

Since X could contain one or several predictors, the extension to the multivariate case is immediate.

Binomial Logistic Regression

In binomial logistic regression, instead of predicting the Bernoulli outcomes on a set of cases as a function of their X values, we predict a sequence of binomial proportions on I occasions as a function of the X values for each occasion.

Binomial Logistic Regression

The mathematics changes very little. The ith occasion has a probability of success $\pi(x_i)$, which now gives rise to a *sample proportion* Y based on m_i observations, via the binomial distribution.

The model is

$$\pi(x_i) = \mu_{Y|X=x_i} = \frac{1}{1 + \exp -\beta' x_i} \tag{8}$$

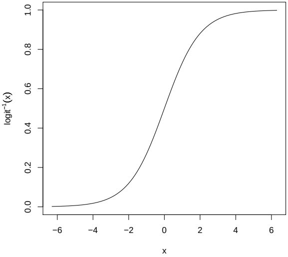
Interpreting Logistic Regression Parameters

How would we interpret the estimates of the model parameters in simple binary logistic regression?

Exponentiating both sides of Equation 5 shows that the odds are an exponential function of x. The odds increase multiplicatively by $\exp(\beta_1)$ for every unit increase in x. So, for example, if $\beta_1 = .5$, the odds are multiplied by 1.64 for every unit increase in x.

Characteristics of Logistic Regression

- Logistic regression predicts the probability of a positive response, given values on one or more predictors
- The plot of $y = \text{logit}^{-1}(x)$ is shaped very much like the normal distribution cdf
- It is S-shaped, and you can see that the slope of the curve is steepest at the midway point, and that the curve is quite linear in this region, but very nonlinear in its outer range



Interpreting Logistic Regression Parameters

If we take the derivative of $\pi(x)$ with respect to x, we find that it is equal to $\beta\pi(x)(1 - \pi(x))$.

This in turn implies that the steepest slope is at $\pi(x) = 1/2$, at which $x = -\beta_0/\beta_1$, and the slope is $\beta_1/4$.

In toxicology, this is called LD_{50}, because it is the dose at which the probability of death is 1/2.

Interpreting Logistic Regression Coefficients

1. Because of the nonlinearity of logit^{-1}, regression coefficients do not correspond to a fixed change in probability
2. In the center of its range, the logit^{-1} function is close to linear, with a slope equal to $\beta/4$
3. Consequently, when X is near its mean, a unit change in X corresponds to approximately a $\beta/4$ change in probability
4. In regions further from the center of the range, one can employ R in several ways to calculate the meaning of the slope

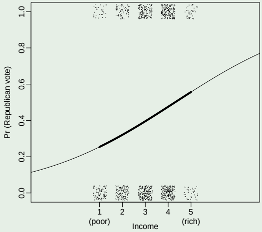
Interpreting Logistic Regression Coefficients
An Example

Example (Interpreting a Logistic Regression Coefficient)

Gelman and Hill (p. 81) discuss an example where the fitted logistic regression is

$$\Pr(\text{Bush Support}) = \text{logit}^{-1}(.33\ \text{Income} - 1.40)$$

Here is their figure 5.1a.



Interpreting Logistic Regression Coeffients

An Example

Example (Interpreting a Logistic Regression Coefficient)

The mean value of income is

```
> mean(income,na.rm=TRUE)
[1] 3.075488
```

Around the value $X = .31$, the probability is increasing at a rate of approximately $\beta/4 = .33/4 = .0825$. So we can estimate that on average the probability that a person with income level 4 will support Bush is about 8% higher than the probability that a person with income level 3 will support Bush.

Interpreting Logistic Regression Coeffients
An Example

Example (Interpreting a Logistic Regression Coefficient)

We can also employ the inverse logit function to obtain a more refined estimate. If I fit the logistic model, and save the fit in a `fit.1` object, I can perform the calculations on the full precision coefficients using the `invlogit()` function, as follows

```
> invlogit(coef(fit.1)[1] +coef(fit.1)[2]*3)

(Intercept)
  0.3955251

> invlogit(coef(fit.1)[1] +coef(fit.1)[2]*4)

(Intercept)
  0.4754819

> invlogit(coef(fit.1)[1] +coef(fit.1)[2]*4)-
+ invlogit(coef(fit.1)[1] +coef(fit.1)[2]*3)

(Intercept)
 0.07995678
```

Interpreting Logistic Regression Coefficients
The Odds Scale

Interpreting Logistic Regression Coefficients

- We can also interpret a logistic regression coefficient in terms of odds
- Since the coefficient β is linear in the *log* odds, e^{β} functions multiplicatively on odds
- That is, around the mean value of 3.1, a unit increase in income should correspond to an $e^{.326}$ increase in odds
- Let's check out how that works by doing the calculations

Interpreting Logistic Regression Coefficients

Example (Interpreting Logistic Regression Coefficients)

We saw in the preceding example that, at a mean income of 3, the predicted probability of supporting Bush is 0.3955251, which is an odds value of

```
> odds.3 = .3955251/(1-.3955251)
> odds.3
```

```
[1] 0.6543284
```

At an income level of 4, the predicted probability of supporting Bush is 0.4754819, which is an odds value of

```
> odds.4 = 0.4754819/(1-0.4754819)
> odds.4
```

```
[1] 0.9065119
```

The ratio of the odds is the same as e^{β}.

```
> odds.4/odds.3
```

```
[1] 1.385408
```

```
> exp(.3259947)
```

```
[1] 1.385408
```

Crabs and Their Satellites

- Agresti (2002, p. 126) introduces an example based on a study in *Ethology*
- Each female horseshoe crab has a male crab resident in her nest
- The study investigated factors associated with whether the fameale crab had any other males, called *satellites*, residing nearby
- Potential predictors include the female's color, spine condition, weight, and carapace width

Predicting a Satellite

- The crab data has information on the number of satellites
- Suppose we reduce these data to binary form, i.e., $Y = 1$ if the female has a satellite, and $Y = 0$ if she does not
- Suppose further that we use logistic regression to form a model predicting Y from a single predictor X, carapace width

Entering the Data

Entering the Data

- The raw data are in a text file called *Crab.txt*.
- We can read them in and attach them using the command

```
> crab.data←read.table("Crab.txt",header=TRUE)
> attach(crab.data)
```

Setting Up the Data

Next, we create a binary variable corresponding to whether or not the female has at least one satellite.

```
> has.satellite←ifelse(Sa > 0,1,0)
```

Fitting the Model with R

We now fit the logistic model using R's GLM module, then display the results

```
> fit.logit <- glm(has.satellite ~ W,
+ family=binomial)
```

Fitting the Model with R

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−12.3508	2.6287	−4.70	0.0000
W	0.4972	0.1017	4.89	0.0000

Interpreting the Results

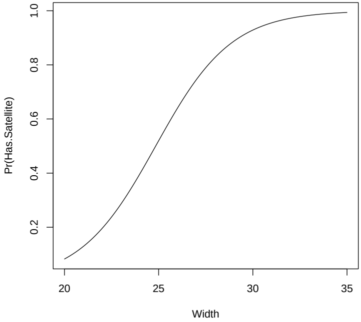
Interpreting the Results

- Note that the slope parameter $b_1 = 0.4972$ is significant
- From our $\beta/4$ rule, this indicates that 1 additional unit of carapace width around the mean value of the latter will increase the probability of a satellite by about $0.4972/4 = 0.1243$
- Alternatively, one additional unit of carapace width is associated with a log-odds multiple of $e^{0.4972} = 1.6441$
- This corresponds to a 64.41% increase in the odds

Interpreting the Results

Here is a plot of predicted probability of a satellite vs. width of the carapace.

```
> curve(invlogit(b1 * x + b0), 20,35,xlab="Width",ylab="Pr(Has.Satel
```



Including Color as Predictor

Dichotomizing Color

- The crab data also include data on color, and use it as an additional (categorical) predictor
- In this example, we shall dichotomize this variable, scoring crabs who are dark 0, those that are not dark 1 with the following command:

```
> is.not.dark ← ifelse(C == 5,0,1)
```

Specifying the Model(s)

The Additive Two-Variable Model

- The additive model states that

$$\text{logit}(p_i) = b_0 + b_1 W + b_2 C$$

- Let's fit the original model that includes only width(W), then fit the model with width(W) and the dichotomized color(is.not.dark)

```
> fit.null ← glm(has.satellite ~ 1, family = binomial)
> fit.W ← glm(has.satellite ~ W ,
+ family=binomial)
> fit.WC ← glm(has.satellite ~ W + is.not.dark,
+ family=binomial)
```

Results

Results for the null model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.5824	0.1585	3.67	0.0002

Results for the simple model with only W:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−12.3508	2.6287	−4.70	0.0000
W	0.4972	0.1017	4.89	0.0000

Results for the additive model with W and C:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−12.9795	2.7272	−4.76	0.0000
W	0.4782	0.1041	4.59	0.0000
is.not.dark	1.3005	0.5259	2.47	0.0134

Comparing Models

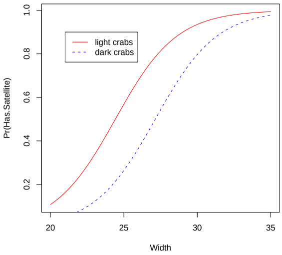
We can compare models with the `anova()` function

```
> anova(fit.null,fit.W,fit.WC,test="Chisq")
```

	Resid. Df	Resid. Dev	Df	Deviance	P(>|Chi|)
1	172	225.76			
2	171	194.45	1	31.31	0.0000
3	170	187.96	1	6.49	0.0108

Plotting the W + C Model

```
> b0 ← coef(fit.WC)[1]
> b1 ← coef(fit.WC)[2]
> b2 ← coef(fit.WC)[3]
> curve(invlogit(b1 * x + b0 + b2), 20,35,xlab="Width",
+   ylab="Pr(Has.Satellite)",col="red")
> curve(invlogit(b1 * x + b0), 20,35,lty=2,col="blue",add=TRUE)
> legend(21,0.9,legend=c("light␣crabs","dark␣crabs"),
+   lty = c(1,2), col=c("red","blue"))
```



Specifying the Model(s)

The additive model states that

$$\text{logit}(p_i) = b_0 + b_1 W + b_2 C$$

Let's add an interaction effect.

```
> fit.WCi ← glm(has.satellite ~ W + is.not.dark
+ + W:is.not.dark,
+ family=binomial)
```

The result is not significant.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−5.8538	6.6939	−0.87	0.3818
W	0.2004	0.2617	0.77	0.4437
is.not.dark	−6.9578	7.3182	−0.95	0.3417
W:is.not.dark	0.3217	0.2857	1.13	0.2600

An Important Application — Case Control Studies

An important application of logistic regression is the *case control study*, in which people are sampled from "case" and "control" categories and then analyzed (often through their recollections) for their status on potential predictors.

For example, samples of patients with or without lung cancer can be sampled, then asked about their smoking behavior.

Relative Risk

With binary outcomes, there are several kinds of effects we can assess. Two of the most important are *relative risk* and the *odds ratio*.

Consider a situation where middle aged men either smoke ($X = 1$) or do not ($X = 0$) and either get lung cancer ($Y = 1$) or do not ($Y = 0$). Often the effect we would like to estimate in epidemiological studies is the *relative risk*,

$$\frac{\Pr(Y = 1|X = 1)}{\Pr(Y = 1|X = 0)} \tag{9}$$

Retrospective Studies

In *retrospective studies* we ask people in various criterion groups to "look back" and indicate whether or not they engaged in various behaviors.

For example, we can take a sample of lung cancer patients and ask them if they ever smoked, then take a matched sample of patients without lung cancer and ask them if they smoked.

After gathering the data, we would then have estimates of $\Pr(X = 1 | Y = 1)$, $\Pr(X = 0 | Y = 1)$ $\Pr(X = 1 | Y = 0)$,and $\Pr(X = 1 | Y = 0)$.

Notice that these are not the conditional probabilities we need to estimate relative risk!

The Odds Ratio

An alternative way of expressing the impact of smoking is the *odds ratio*, the ratio of the odds of cancer for smokers and nonsmokers. This is given by

$$\frac{\Pr(Y = 1|X = 1)/1 - \Pr(Y = 1|X = 1)}{\Pr(Y = 1|X = 0)/1 - \Pr(Y = 1|X = 0)} \tag{10}$$

Some Key Identities

By repeatedly employing

1. The definition of conditional probability, i.e., $\Pr(A|B) = \Pr(A \cap B)\Pr(B) = \Pr(B \cap A)\Pr(B)$
2. The fact that $A \cap B = B \cap A$

it is easy to show that

$$\frac{\Pr(Y = 1|X = 1)/(1 - \Pr(Y = 1|X = 1))}{\Pr(Y = 1|X = 0)/(1 - \Pr(Y = 1|X = 0))} = \frac{\Pr(X = 1|Y = 1)/(1 - \Pr(X = 1|Y = 1))}{\Pr(X = 1|Y = 0)/(1 - \Pr(X = 1|Y = 0))} \tag{11}$$

Some Key Identities

Equation 11 demonstrates that the information about odds ratios is available in retrospective studies with representative sampling.

Furthermore, suppose that an outcome variable Y fits a logistic regression model $\text{logit}(Y) = \beta_1 X + \beta_0$. As Agresti (2002, p. 170–171) demonstrates, it is possible to correctly estimate β_1 in a retrospective case-control study where Y is fixed and X is random. The resulting fit will have a modified intercept $\beta_0^* = \log(p_1/p_0) + \beta_0$, where p_1 and p_0 are the respective sampling probabilities for $Y = 1$ cases and $Y = 0$ controls.