Logistic Regression

Goals

- To introduce logistic regression for binary outcomes.
- To introduce multinomial and ordinal logistic regression.
- To describe diagnostics for logistic regression models.

Introduction

For a binary response, say y_i, there are two response categories, which we will code as 0 or 1. We consider y_i to be a realization of a random variable Y_i that can take the value of 1 with a probability, π_i, and a value of 0 with a probability of $1 - \pi_i$. The distribution of Y_i is called the Bernoulli distribution,

$$Pr(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

The expected value of Y_i is π_i and the variance is $\pi_i(1 - \pi_i)$. Note that the mean and variance depend on the underlying probability, π_i. So any factor that affects the probability will alter not just the mean but also the variance of the observations. Thus, a linear model that allows the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

Assuming $Y_1, \ldots, Y_n$ for $Y_i \sim Bern(1, \pi)$ are mutually independent, then $\sum_{i=1}^{n} Y_i \sim Bin(n, \pi)$ and so Y follows a binomial distribution with expected value $n\pi$ and variance $n\pi(1 - \pi)$.

A second problem with a linear model such as $\pi_i = \beta_0 + \beta_1 X_i$, is that the predicted values could take on any value, but since π_i is a probability, we want the predicted values to be between 0 and 1. We can transform π_i to remove the restictions by first transforming to the *odds*, $\pi_i/(1 - \pi_i)$ that $Y_i = 1$, which can take on any positive value. Taking the log of the odds removes the floor restriction so that the log-odds, also called the logit,

$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

can take on values from $-\infty$ to $+\infty$. Thus, the logit maps the probability to the real number line. If the probability is 1/2 then the odds are even and the logit is zero. The logit is symmetric around 0, and unbounded both above and below, making the logit a good candidate for the response-variable side of a linear model.

Now we can assume the *logit* of the probability, rather than the probability itself, follows a linear model.

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_2 X_{ki}$$

The transformation is one-to-one. The inverse transformation is

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

and is called the *expit* or *inverse logit* transformation.

Logistic Regression

The regression coefficients are interpreted similiarly to linear regression except that they are in terms of logits. So β_j is the change in the logit associated with a one-unit change in the jth predictor variable, holding all others constant. If we exponentiate both sides:

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}} = e^{\beta_0}(e^{\beta_1})^{X_{i1}}\ldots(e^{\beta_k})^{X_{ik}}$$

we obtain a *multiplicative* model for the odds. A one-unit change in the jth predictor is associated with multiplying the odds by a factor of e^{β_j}, holding all other variables constant. Thus, e^{β_j} is the multiplicative effect on the odds of increasing X_j by 1, holding the other X's constant. e^{β_j} is called the odds ratio because it represents the ratio of the odds of response at two different X values, where the X value in the numerator is one-unit larger than the X value in the denominator.

So, increasing X_1 by 1 changes the logit by β_1 and multiplies the odds by e^{β_1}. For example, if $\beta_1 = 2$, then increasing X by 1 increases the odds by a factor of $e^2 \approx 2.718^2 = 7.389$, holding constant the other predictors.

Finally, we can solve for the probability:

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_i+\ldots+\beta_k X_{ki}}}{1+e^{\beta_0+\beta_1 X_i+\ldots+\beta_k X_{ki}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i+\ldots+\beta_k X_{ki})}}$$

There is no simple interpretation on the probability scale. Because the slope of the relationship between π and X_1 is nonlinear, the slope is not constant; the slope is $\beta_1\pi(1-\pi)$, and hence is at a maximum when $\pi = 1/2$, where the slope is $\beta_1/4$. So, $\beta_j/4$ is the slope of the logistic regression surface in the direction of X_j at $\pi = .5$.

The slope does not change very much between $\pi = .2$ and $\pi = .8$, reflecting the near linearity of the logistic curve in this range.

Nevertheless, solving for the probability is the basis for how the predicted probabilities, or fitted values, are computed, which we will discuss later.

The X's in the linear predictor, η_i can be as general as previously, including, for example

- quantitative explanatory variables
- transformations of quantitative explanatory variables
- polynomial regressors formed from quantitative explanatory variables
- dummy regressors representing qualitative explanatory variables
- interaction regressors

The model can be fit to data by the method of maximum likelihood.

Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation.

For an individual coefficient, it is most convenient to test the hypothesis $H_0 : \beta_j = 0$ by calculating the *Wald statistic*

$$Z_0 = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$$

where $SE(\widehat{\beta}_j)$ is the asymptotic standard error of $\widehat{\beta}_j$.

The test statistic Z_0 follows an asymptotic standard normal distribution under the null hypothesis.

Similarly, an asymptotic $100(1-\alpha)\%$ CI for β_j is given by

$$\beta_j = \widehat{\beta}_j \pm z_{\alpha/2} SE(\widehat{\beta}_j)$$

where $z_{\alpha/2}$ is the value from $Z \sim N(0,1)$ with a probability of $\alpha/2$ to the right.

Wald tests for several coefficients can be formulated from the estimated asymptotic variances and covariances of the coefficients.

It is also possible to formulate a *likelihood-ratio test* for the hypothesis that several coefficients are simultaneously zero, $H_0 : \beta_1 = \ldots = \beta_q = 0$.

We proceed, as in least-squares regression, by fitting two models to the data:

The full model (model 1)

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_k X_k$$

and the null (or reduced) model (model 0)

$$\text{logit}(\pi) = \beta_0 + 0X_1 + \ldots + 0X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_k X_k = \beta_0 + \beta_{q+1} X_{q+1} + \ldots + \beta_k X_k$$

Each model produces a maximized likelihood: L_1 for the full model, L_0 for the null model.

Because the null model is a specialization of the full model, $L_1 \geq L_0$.

The generalized likelihood-ratio test statistic for the null hypothesis is

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

Under the null hypothesis, this test statistic has an asymptotic χ^2 distribution with q degrees of freedom.

A test of the omnibus null hypothesis that all the coefficients are 0, $H_0 : \beta_1 = \ldots = \beta_k = 0$, is obtained by specifying a null model that includes only the constant, $\text{logit}(\pi) = \beta_0$.

The likelihood-ratio test can be inverted to produce confidence intervals for coefficients.

The likelihood-ratio test is less prone to breaking down than the Wald test.

An analog to the multiple-correlation coefficient can also be obtained from the log-likelihoods.

By comparing $\log_e L_0$ for the model containing only the constant with $\log_e L_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y.

The quantity $G^2 \equiv -2 \log_e L$, called the residual deviance under the model, is a generalization of the residual sum of squares for a linear model.

Thus,

$$R^2 = 1 - \frac{G_1^2}{G_0^2} = 1 - \frac{\log_e L_1}{\log_e L_0}$$

is analogous to R^2 for a linear model and is referred to as McFadden's pseudo-R^2. Note that there are numerous pseudo-R^2s because an exact equivalent to the R^2 in least-squares regression does not exist.

Residuals for Logistic Regression

Unlike linear models, the observations have different variances, so two types of residuals are used.

First, the predicted values are given by

$$\widehat{\mu}_i = n_i\widehat{\pi}_i = n_i\frac{e^{\widehat{\beta}_0+\widehat{\beta}_1X_i+\ldots+\widehat{\beta}_kX_{ki}}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1X_i+\ldots+\widehat{\beta}_kX_{ki}}}$$

In the observed data, y_i is either 0 or 1 but the predicted values can take on any value between 0 and 1. We can interpret the fitted or predicted probability as the estimated population proportion of individuals sharing the ith person's characteristics for whom $Y = 1$.

Pearson Residuals

Pearson residuals are the difference between observed and fitted values and divide by an estimate of the standard deviation of the observed value,

$$e_i = \frac{y_i - n_i\widehat{\pi}_i}{\sqrt{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)}}$$

Observations with a Pearson residual exceeding 3 in absolute value may be worth a closer look.

Usually the Pearson residuals are standardized.

Standardized Pearson residuals correct for the conditional response variation and for the leverage of the observations:

$$r_i = \frac{e_i}{\sqrt{1-h_i}}$$

Deviance Residuals

An alternative residual is based on the deviance or likelihood ratio statistic. The deviance residual is defined as

$$d_i = \sqrt{2\left[y_i\log\left(\frac{y_i}{\widehat{\mu}_i}\right)+(n_i-y_i)\log\left(\frac{n_i-y_i}{n_i-\widehat{\mu}_i}\right)\right]}$$

where the residual has the same sign as $y_i - \widehat{\mu}_i$.

Observations with a deviance residual in excess of $|3|$ may indicate lack of fit.

Standardized deviance residuals are

$$s_i = \frac{d_i}{\sqrt{1-h_i}}$$

A *studentized* (i.e., jack-knifed) residual can be obtained as

$$t_i = s_i\sqrt{\frac{n-k-1}{n-k-s_i^2}}$$

Cook's Distance can be approximated by

$$D_i = s_i^2 \frac{h_i}{(1-h_i)k}$$

Checking Assumptions

There are not assumptions regarding normality or constant variance. However, there is still an assumption that the predictors are linear in the logit (i.e., log odds). Departures from linearity can be examined by plotting the predicted log odds of the outcome against each X and/ or constructing component-residual plots.

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, patients with or without lung cancer can be sampled, then asked about their past smoking behavior.

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:

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

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 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 $Pr(X = 0|Y = 0)$.

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

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

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

By repeatedly employing the definition of conditional probability, i.e.

$$Pr(A|B) = Pr(A \cap B)Pr(B) = Pr(B \cap A)Pr(B)$$

and that $Pr(A \cap B) = Pr(B \cap A)$, it is easy to show that

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

Thus, the odds ratio can be estimated from retrospective data but the relative risk and the risk difference cannot. However, the odds ratio approximates the relative risk for rare outcomes, which is typically when case-control studies are used.

Infinite Estimates

```
library(mosaic) # for the tally() function
```

Complete separation and quasi-complete separation can result in infinite estimates in logistic regression.

Complete separation occurs if there is a plane that can separate the x-values where $y = 1$ and $y = 0$ and no observations fall on that plane.

Quasi-complete separation occurs if on the plane boundary, both outcomes occur.

This is common in contingency tables.

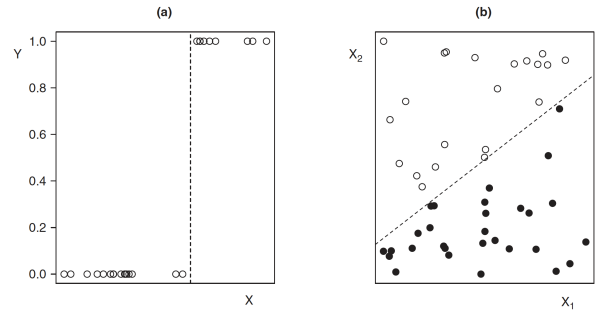

Figure 14.5 Separability in logistic regression: (a) with one explanatory variable, X; (b) with two explanatory variables, X_1 and X_2. In panel (*b*), the solid dots represent observations for which $Y = 1$ and the hollow dots observations for which $Y = 0$.

Figure 1: Separability

Here is a situation in which complete separation occurs and the estimate is actually infinite.

```
# Note, I am generating values to illustrate a point -
# this is not code that you should use
x <- c(10, 20, 30, 40, 60, 70, 80, 90)
y <- c(0, 0, 0, 0, 1, 1, 1, 1)

fit <- glm(y ~ x, family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial")
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.045e-05  -2.110e-08   0.000e+00   2.110e-08   1.045e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -118.158 296046.187       0        1
## x                2.363   5805.939       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1.1090e+01  on 7  degrees of freedom
## Residual deviance: 2.1827e-10  on 6  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 25
```

Note the enormous std. errors. Also, note that the residual deviance=0. Also, there is warning 'glm.fit: fitted probabilities numerically 0 or 1 occurred' which means that model fit gave perfect predictions, and this is also a clue that something is wrong.

But notice that estimates are given, so you, rather than the software has to detect that the source of the problem is complete or quasi-complete separation (although there is a function described below that can help).

Here is a situation in which quasi-complete separation occurs.

```
# Note, I am generating values to illustrate a point -
# this is not code that you should use
x <- c(10, 20, 30, 40, 50, 50, 60, 70, 80, 90)
y <- c(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)

fit <- glm(y ~ x, family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial")
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.177   0.000   0.000   0.000   1.177
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -98.158  39288.592  -0.002    0.998
## x               1.963    785.772   0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.8629  on 9  degrees of freedom
## Residual deviance:  2.7726  on 8  degrees of freedom
## AIC: 6.7726
##
## Number of Fisher Scoring iterations: 21
```

If you have contingency table data and one of the cell counts is 0, then you may run into this problem. Agresti (2002; Sec. 7.2.2 and 7.4.8) illustrates this with a data set on histology grade and risk factors for 79 cases of endometrial cancer. The data are available numerous places, including in an package called `brglm2`.

```
library(brglm2)
data(endometrial)
```

```
summary(endometrial)
```

```
##        NV               PI              EH              HG
##  Min.   :0.0000   Min.   : 0.00   Min.   :0.270   Min.   :0.0000
```

```
##  1st Qu.:0.0000   1st Qu.:11.00   1st Qu.:1.180   1st Qu.:0.0000
##  Median :0.0000   Median :16.00   Median :1.640   Median :0.0000
##  Mean   :0.1646   Mean   :17.38   Mean   :1.662   Mean   :0.3797
##  3rd Qu.:0.0000   3rd Qu.:21.00   3rd Qu.:2.015   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :49.00   Max.   :3.610   Max.   :1.0000
```

`HG` is the histology grade (1=high, 0=low)

`NV` is neovasculation (1=present, 0=absent)

`PI` is pulsatility index of arteria uterina (ranges 0-49)

`EH` is endometrium height (ranges 0.27 to 3.61)

We are going to fit the model

```
endo <- glm(HG ~ NV + PI + EH, family="binomial", data=endometrial)
summary(endo)
```

```
##
## Call:
## glm(formula = HG ~ NV + PI + EH, family = "binomial", data = endometrial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.50137  -0.64108  -0.29432   0.00016   2.72777
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.30452    1.63730   2.629 0.008563 **
## NV           18.18556 1715.75089   0.011 0.991543
## PI           -0.04218    0.04433  -0.952 0.341333
## EH           -2.90261    0.84555  -3.433 0.000597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 104.903  on 78  degrees of freedom
## Residual deviance:  55.393  on 75  degrees of freedom
## AIC: 63.393
##
## Number of Fisher Scoring iterations: 17
```

```
tally(~ NV | HG, margins=TRUE, data=endometrial)
```

```
##        HG
## NV       0  1
##   0     49 17
##   1      0 13
##   Total 49 30
```

For all 13 patients with NV=1, HG=1. This is quasi-complete separation.

If you try to get confidence intervals

```
confint(endo)
```

```
## Waiting for profiling to be done...
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
##                   2.5 %       97.5 %
## (Intercept)   1.4332283   7.95497588
## NV          -69.2281177 506.19375845
## PI           -0.1370875   0.03817732
## EH           -4.7859979  -1.43659412
```

it doesn't go so well.

The `detectseparation` package has a method for detecting infinite estimates. `method="detect_separation"` will work only if the `detectseparation` package is loaded.

```
library(detectseparation)
glm(HG ~ NV + PI + EH, family="binomial", data=endometrial,
    method="detect_separation")
```

```
## Implementation: ROI | Solver: lpsolve
## Separation: TRUE
## Existence of maximum likelihood estimates
## (Intercept)          NV          PI          EH
##           0         Inf           0           0
## 0: finite value, Inf: infinity, -Inf: -infinity
```

Remedies for infinite estimates include penalized likelihood (Firth, 1993) or Bayesian approaches.

Penalized Likelihood can be done using the function `logistf()` from the `logistf` package.

```
library(logistf)
endoPL <- logistf(HG ~ NV + PI + EH, family="binomial", data=endometrial)
summary(endoPL)
```

```
## logistf(formula = HG ~ NV + PI + EH, data = endometrial, family = "binomial")
##
## Model fitted by Penalized ML
## Coefficients:
##                     coef   se(coef) lower 0.95  upper 0.95      Chisq
## (Intercept)  3.77455951 1.43900672  1.0825371  7.20928050  8.1980136
## NV           2.92927330 1.46497415  0.6097244  7.85463171  6.7984572
## PI          -0.03475175 0.03789237 -0.1244587  0.04045547  0.7468285
## EH          -2.60416387 0.75362838 -4.3651832 -1.23272106 17.7593175
##                         p method
## (Intercept) 4.193628e-03      2
## NV          9.123668e-03      2
## PI          3.874822e-01      2
## EH          2.506867e-05      2
##
## Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None
##
## Likelihood ratio test=43.65582 on 3 df, p=1.78586e-09, n=79
## Wald test = 21.66965 on 3 df, p = 7.641345e-05
```

Multinomial Logistic Regression

The binary logistic regression model can be extended to an outcome with multiple categories. This approach has the advantage of treating the categories in a non-arbitrary, symmetric manner.

The response variable Y can take on any of J qualitative values, which, for convenience, we number $1, 2, \ldots, J$ (using the numbers only as category labels).

Let π_{ij} denote the probability that the ith observation falls in the jth category of the response variable; that is,

$$\pi_{ij} \equiv Pr(Y_i = j) \text{ for } j = 1, \ldots, J$$

$\sum_{j=1}^{J} \pi_{ij} = 1$ due to the assumption that the categories are mutually exclusive and exhaustive.

Y_{ij} becomes an indicator (or dummy) variable that takes on the value 1 if the ith response falls in the jth category and 0 otherwise, and $\sum_j y_{ij} = 1$ because for each case one and only one of the y_{ij} can be 1.

The probability distribution of these counts, Y_{ij} is given by the multinomial distribution

$$Pr(Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}) = \binom{n_i}{y_{i1}, \ldots, y_{iJ}} \pi_{i1}^{y_{i1}} \ldots \pi_{iJ}^{y_{iJ}}$$

We would like to model how the probabilities depend on predictors. The basic idea is that we are going to choose a baseline category, calculate the log-odds for all other categories relative to the baseline, and then let the log-odds be a linear function of the predictors.

We have k regressors, $X_1, \ldots, X_k$, on which the π_{ij} depend.

More specifically, suppose that this dependence can be modeled as

$$\pi_{ij} = \frac{e^{\gamma_{0j} + \gamma_{1j} X_{i1} + \ldots + \gamma_{kj} X_{ik}}}{1 + \sum_{j=1}^{J-1} e^{\gamma_{0j} + \gamma_{1j} X_{i1} + \ldots + \gamma_{kj} X_{ik}}}$$

for $j = 1, \ldots, J-1$.

There is one set of parameters, $\gamma_{0j}, \gamma_{1j}, \ldots, \gamma_{kj}$, for each response variable category except the last; category J functions as a type of baseline. Thus, there are $(k+1) \times (J-1)$ parameters to estimate!

The use of a baseline category is one way of avoiding redundant parameters because of the restriction that $\sum_{j=1}^{J} \pi_{ij} = 1$, thus $\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}$

Some algebraic manipulation of the model produces

$$\log_e \frac{\pi_{ij}}{\pi_{iJ}} = \gamma_{0j} + \gamma_{1j} X_{i1} + \ldots + \gamma_{kj} X_{ik} \text{ for } j = 1, \ldots, J-1$$

The regression coefficients affect the log-odds of membership in category j versus the baseline category, J.

It is also possible to form the log-odds of membership in any pair of categories j and j'

$$\log_e \frac{\pi_{ij}}{\pi_{ij'}} = \log_e \left(\frac{\pi_{ij}}{\pi_{iJ}} \Big/ \frac{\pi_{ij'}}{\pi_{iJ}} \right) = \log_e \frac{\pi_{ij}}{\pi_{iJ}} - \log_e \frac{\pi_{ij'}}{\pi_{iJ}} = (\gamma_{0j} - \gamma_{0j'}) + (\gamma_{1j} - \gamma_{1j'}) X_{i1} + \ldots + (\gamma_{kj} - \gamma_{kj'}) X_{ik}$$

The regression coefficients for the logit between any pair of categories are the differences between corresponding coefficients.

Now suppose that the model is specialized to a binary response variable. Then, $J = 2$, and

$$\log_e \frac{\pi_{i1}}{\pi_{i2}} = \log_e \frac{\pi_{i1}}{1 - \pi_{i1}} = \gamma_{01} + \gamma_{11} X_{i1} + \ldots + \gamma_{k1} X_{ik}$$

Applied to a binary outcome, the multinomial logistic model is identical to the binary logistic model.

Ordinal Logistic Regression

The multinomial logistic regression model may be applied to ordinal response variables as well, but it does not make explicit use of the fact that the response categories are ordered. We now consider a model designed specifically for the analysis of responses measured on an ordinal scale.

Let $\pi_{ij} = Pr(Y_i = j)$ denote the probability that the response for individual i falls in the jth category. The cumulative probability that the response falls in the jth category or *below* is:

$$\xi_{ij} = Pr(Y_i \leq j) = \pi_{i1} + \pi_{i2} + \ldots + \pi_{ij}$$

Next we can apply the logit transformation to the *cumulative* response probabilities, ξ_{ij}, directly, rather than to the response probabilites, π_{ij}.

$$\text{logit}(\xi_{ij}) = \log \frac{\xi_{ij}}{1 - \xi_{ij}} = \theta_j + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

where θ_j a constant representing the baseline value of the transformed cumulative probability for category j.

Exponentiating give the odds of $Y_{ij} \leq j$ that is, the odds of a response in category j or below is

$$\frac{\xi_{ij}}{1 - \xi_{ij}} = \lambda_j e^{(\beta_1 X_{i1} + \ldots + \beta_k X_{ik})}$$

where $\lambda_j = e^{\theta_j}$ which may be interpreted as the baseline odds of a response in category j or below when all X's are zero.

The effect of a covariate, for example, β_1, is to raise or lower the odds of a response in category j or below by a factor of e^{β_1} Note that the effect is a proportionate change in the odds of $Y_i \leq j$ for all response categories j. If a certain combination of covariate values doubles the odds of being in category 1, it also doubles the odds of being in category 2 or below, or in category 3 or below.

The logits in this model are for cumulative categories — at each point contrasting categories above category j with category j and below.

The slopes for each of these regression equations are identical; the equations differ only in their intercepts.

The logistic regression surfaces are therefore horizontally parallel to each other. For example, for $J = 4$ response categories and a single X:

For a fixed set of X's, any two different cumulative log-odds — say, at categories j and j' — differ only by the constant $(\theta_j = \theta_{j'})$

The odds, therefore, are proportional to one another, and for this reason, the ordered (or ordinal) logistic model is sometimes called the proportional-odds model. This model is also sometimes called the cumulative logit model because it models the cumulative logit.

There are $(k + 1) + (J - 1) = k + J$ parameters to estimate, which is fewer than the multinomial logistic model. Thus, if the response categories are ordered and the proportional odds assumption is reasonable, then the ordinal logistic model is to be prefered and is easier to interpret than the multinomial logistic model.

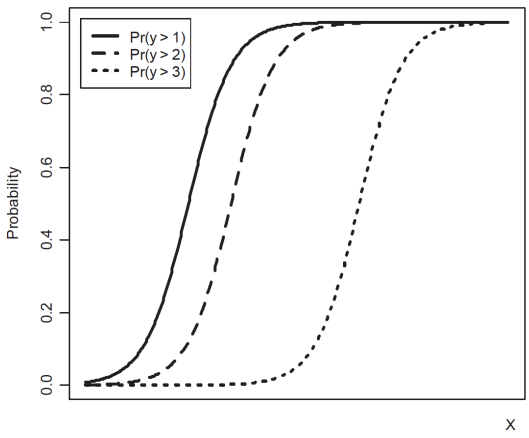

Figure 2: Proportional Odds

The multinomial logistic model does not make the proportional odds assumption, so when it does not hold, ordered response categories can be modeled using the multinomial logistic model.

As usual, likelihood-ratio tests are computed by contrasting the deviances for alternative models, with and without the terms in question.