Coefficient Interpretation and Hypothesis Tests for Multiple Logistic Regression

Model

$$P(Y_i = 1|X_{i1}, \ldots, X_{ip}) = \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}$$

Note that this means that

$$\begin{aligned} P(Y_i = 0|X_{i1}, \ldots, X_{ip}) = 1 - P(Y_i = 1|X_{i1}, \ldots, X_{ip}) = 1 - \frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}} \\ = \frac{1}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}} \end{aligned}$$

Odds

The odds that $Y = 1$ are given by:

$$Odds(Y_i = 1) = \frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)}$$

Examples:

- If $P(Y_i = 1|X_i) = 0.75$, $Odds(Y_i = 1) = \frac{0.75}{0.25} = 3$
- If $P(Y_i = 1|X_i) = 0.5$, $Odds(Y_i = 1) = \frac{0.5}{0.5} = 1$
- If $P(Y_i = 1|X_i) = 0.1$, $Odds(Y_i = 1) = \frac{0.1}{0.9} = \frac{1}{9}$

Odds in a Logistic Regression Model

$$\begin{aligned} Odds(Y_i = 1) = \frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)} = \frac{\frac{e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}}{\frac{1}{1 + e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}}} \\ = e^{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}} \end{aligned}$$

Coefficient Interpretation

On Odds Scale

Increasing X_{ij} by 1 unit while holding all other explanatory variables fixed leads to a multiplicative change in the predicted odds by e_j^β.

On Log-Odds Scale

Increasing X_{ij} by 1 unit while holding all other explanatory variables fixed leads to an additive change in predicted log-odds of β_j units.

Credit Card Default Example

```r
library(ggplot2)
library(gridExtra)
library(dplyr)
library(ISLR)

fit <- glm(default ~ student + balance + income, data = Default, family = binomial)
summary(fit)
```

```
##
## Call:
## glm(formula = default ~ student + balance + income, family = binomial,
##     data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

What is the interpretation of the coefficient for `studentYes`? Note that $e^{-0.6468} \approx 0.523719$.

What is the interpretation of the coefficient for `balance`? Note that $e^{0.005737} \approx 1.005753$. Is it helpful to consider that $e^{(0.005737*100)} \approx 1.775$?

Hypothesis Tests

Caution: standard approaches to hypothesis testing in logistic regression models are highly dependent on having a fairly large sample size. See Stat 343 for bootstrap-based alternatives.

Tests about one coefficient

P-values for tests about one coefficient can be read from the summary output as with `lm`. Note that these are not t tests, but are large-sample z tests (based on an approximate normal distribution from the Central Limit Theorem).

Example: Conduct a test of the claim that an individual's `income` is not useful in predicting whether or not they will default on their credit card debt.

Tests about more than one coefficient

This is a little artificial, but to demonstrate the code let's consider a test of the hypotheses that we can drop both the `student` and the `income` variables from the model. We will:

- fit a reduced model (similar to what we would do in a `lm` context)
- call `anova` to compare the reduced and full model
    - Unlike `anova` comparisons with linear models, we need to specify a `test` argument to `anova`. A common option is `test = "LRT"` (for likelihood ratio test). Again, this is a large-sample approximate test procedure.

```
fit_reduced <- glm(default ~ balance, data = Default, family = binomial)
anova(fit_reduced, fit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: default ~ balance
## Model 2: default ~ student + balance + income
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9998     1596.5
## 2      9996     1571.5  2   24.907 3.904e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```