Logistic Regression - Statistical Inference

Ryan Miller

Introduction

This week our focus is on **logistic regression**, a type of generalized linear model (GLM):

$$logit(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Recall that all GLMs have the following components:

- *Systematic component* - a linear combination of predictor variables (ie: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$)
- *Random component* - a probability distribution for Y, the outcome variable
- *Link function* - a function that links $E(Y)$ to the model's systematic component

The Binomial Distribution

- Logistic Regression is a statistical model due its use of the *binomial probability distribution*
- Consider a binary random variable, Y, with an underlying probability of "success" denoted by π
  - In mathematical shorthand, $Y \sim \text{binom}(1, \pi)$
  - In this framework, notice $E(Y) = \pi$

The Binomial Distribution

- Logistic Regression is a statistical model due its use of the *binomial probability distribution*
- Consider a binary random variable, Y, with an underlying probability of "success" denoted by π
  - In mathematical shorthand, $Y \sim \text{binom}(1, \pi)$
  - In this framework, notice $E(Y) = \pi$
- In logistic regression, we model $g(\pi)$ as a linear combination of predictors

Statistical Inference

- Theoretically, statistical inference could be done on the outcome, y, using a binomial distribution
  - Practically, it's more useful to apply inferential methods to the model coefficient estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_p$

Statistical Inference

- Theoretically, statistical inference could be done on the outcome, y, using a binomial distribution
  - Practically, it's more useful to apply inferential methods to the model coefficient estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_p$
- Without getting in to the details, maximum likelihood theory provides a Normal approximation for these estimates
  - $\hat{\beta}_p \sim N(\beta_p, SE)$
  - Similar to what we saw in linear regression, the `summary()` function provides a default test of $H_0 : \beta_p = 0$

Example

```r
xub$to_diff <- xub$TOV - xub$TOV.1
m <- glm(win ~ to_diff, data = xub, family = "binomial")
summary(m)
```

```
##
## Call:
## glm(formula = win ~ to_diff, family = "binomial", data = xub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5671  -1.3653   0.7045   0.9024   1.2404
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8814     0.5403   1.631    0.103
## to_diff       0.1286     0.1285   1.000    0.317
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23.699  on 18  degrees of freedom
## Residual deviance: 22.450  on 17  degrees of freedom
## AIC: 26.45
##
## Number of Fisher Scoring iterations: 4
```

Inference and Odds Ratios

- In the previous model, the estimated intercept was $\hat{\beta}_0 = 0.8814$
  - Is it meaningful to interpret this coefficient? What does it tell us?

Inference and Odds Ratios

- In the previous model, the estimated intercept was $\hat{\beta}_0 = 0.8814$
  - Is it meaningful to interpret this coefficient? What does it tell us?
- $exp(0.8814) = 2.414$
  - This is the *odds ratio* of the odds of XU winning relative to the odds of their opponent winning *when both teams have an equal number of turnovers* is 2.414
- However, notice the p-value testing $H_0 : \beta_0 = 0$ is 0.103, so we might not be statistically convinced XU is really more likely to win in this situation
  - Further, recognize $exp(0) = 1$, which implies an odds ratio of 1 indicates an equal likelihood of XU winning and their opponent winning

The Inverse-Logit Transformation

- One of the most useful aspects of logistic regression is the model's ability to generate *predicted probabilities* for various combinations of predictors
  - Determining these probabilities requires the use of the *inverse-logit* transformation on the model's *linear predictor* (often denoted η)

$$logit(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots = \eta$$

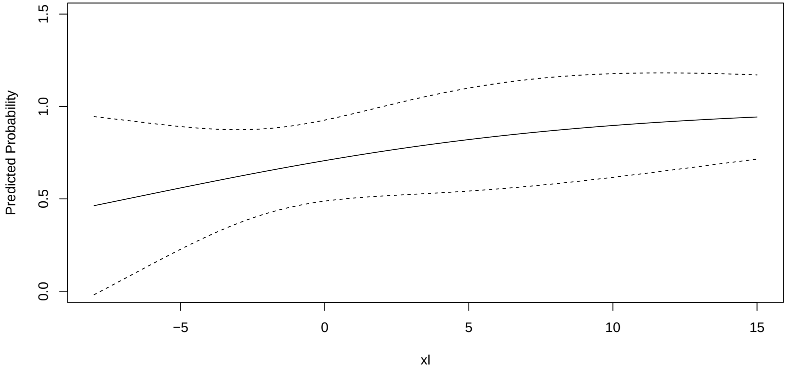
$$\pi = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)} = \frac{exp(\eta)}{1 + exp(\eta)}$$

Confidence Intervals and the Inverse-Logit Transformation

- Often, we'd like to reported predicted probabilities alongside confidence intervals
  - However, confidence intervals for predicted probabilities should be calculated on the logit scale, then the end-points should be transformed
  - Otherwise, the intervals run into Normality/boundary problems

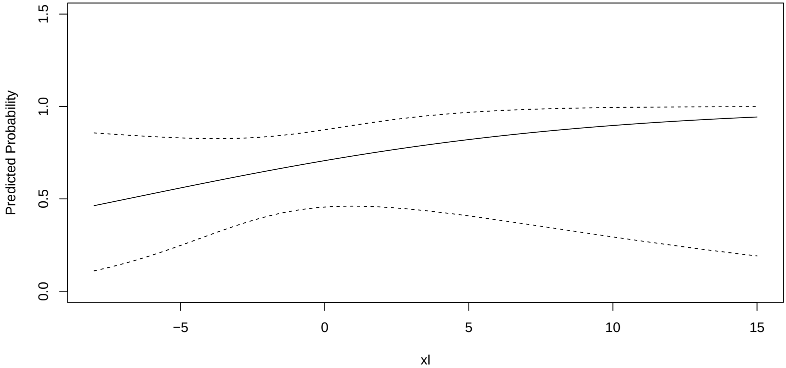
Example (Don't do this)

```
xl <- seq(min(xub$to_diff), max(xub$to_diff), by = 0.1)
preds <- predict(m, newdata = data.frame(to_diff = xl), type = "response", se = TRUE)
plot(xl, preds$fit, type = "l", ylim = c(0,1.5), ylab = "Predicted Probability")
lines(xl, preds$fit + 1.96*preds$se.fit, lty = 2)
lines(xl, preds$fit - 1.96*preds$se.fit, lty = 2)
```



Example (Do this instead)

```r
inverse_logit = function(x){exp(x)/(1+exp(x))}
preds <- predict(m, newdata = data.frame(to_diff = xl), type = "link", se = TRUE)
plot(xl, inverse_logit(preds$fit), type = "l", ylim = c(0,1.5), ylab = "Predicted Probability")
lines(xl, inverse_logit(preds$fit + 1.96*preds$se.fit), lty = 2)
lines(xl, inverse_logit(preds$fit - 1.96*preds$se.fit), lty = 2)
```



The Likelihood Ratio Test for Nested Models

- For linear regression, the F-test (ANOVA) allowed us to statistically compare two nested models
  - This allowed us to assess the overall impact of a categorical predictor that was being represented by multiple dummy variables
  - It also allowed us to compare a model of interest to an intercept-only model as an overall evaluation

The Likelihood Ratio Test for Nested Models

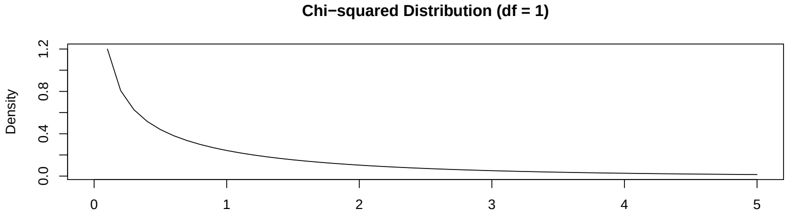
- For linear regression, the F-test (ANOVA) allowed us to statistically compare two nested models
  - This allowed us to assess the overall impact of a categorical predictor that was being represented by multiple dummy variables
  - It also allowed us to compare a model of interest to an intercept-only model as an overall evaluation
- For logistic regression, the analogous statistical test is the *Likelihood ratio test*
  - ANOVA/the F-test compares a standardized ratio of *sums of squares*, while the likelihood ratio test compares a ratio of *likelihoods*

The Likelihood Ratio Test for Nested Models

- Without getting too detailed, a larger likelihood indicates a better fit to the sample data
  - Thus, a likelihood ratio that sufficiently exceeds 1 will indicate superiority of the larger model

The Likelihood Ratio Test for Nested Models

- Without getting too detailed, a larger likelihood indicates a better fit to the sample data
  - Thus, a likelihood ratio that sufficiently exceeds 1 will indicate superiority of the larger model
- It can be shown that the distribution of the likelihood ratio, under the null hypothesis that the models have equal likelihoods, follows a Chi-squared distribution with degrees of freedom equal to the difference in model parameters



Example

```r
library(lmtest)
m1 <- glm(win ~ to_diff, data = xub, family = "binomial")
m2 <- glm(win ~ to_diff + Location, data = xub, family = "binomial")
lrtest(m1, m2)
```

```
## Likelihood ratio test
##
## Model 1: win ~ to_diff
## Model 2: win ~ to_diff + Location
##   #Df   LogLik Df  Chisq Pr(>Chisq)
## 1   2 -11.2250
## 2   3  -7.6852  1 7.0796   0.007797 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Closing Remarks

- We've now introduced logistic regression and covered a few important modes of statistical inference
- The main concept you need to be aware of is the role of the logit link function
  - In order to interpret model coefficients, you can use exponentiation
  - In order to calculate predicted probabilities (and associated confidence intervals) you must use the inverse logit transformation