STAT 213: Regression With a Binary Response

Colin Reimer Dawson

Last Revised April 26, 2016

We have seen how to handle various combinations of categorical and quantitative predictor variables, but so far our response has always been quantitative. What happens if our response is categorical?

Consider the binary response variable, `ObamaWin`, which is equal to 1 for a state if Obama carried the state in the 2008 election, and 0 otherwise. A question of interest is what predictors matter in forecasting the outcome of an election in a state.

Here is a linear regression model:

$$\texttt{ObamaWin} = \beta_0 + \beta_1 \cdot \%\texttt{BA} + \varepsilon$$

where `% BA` represents the percentage of state residents with at least a college degree.

1. Under ordinary regression, the predicted response value corresponds to the *mean* response at particular values of predictors. What does the mean of a *binary* response (coded as 0 or 1) correspond to?

2. In a sample, the mean of a binary variable is the proportion of 1s. In the population, we can think of the predicted value as the *probability* of a 1.

   The relevant data for the model above is in `Election08` (in `Stat2Data`). Fit the above simple linear regression model. What is the predicted probability of an Obama win in a state where 10% of residents have a BA? 37%?

3. As you can see, if we use ordinary linear regression with a binary response, we will sometimes get nonsensical predictions. We would like a model for which the predicted probability of "success" asymptotically heads toward 0 or 1, instead of going negative or exceeding 1. Intuitively, we'd like the prediction function to have an S-shape, flattening out on the extremes.

   Represent success probabilities with π. One S-shaped function has the form

   $$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

   Use R to plot this function for different values of β_0 and β_1. Type the following, filling in BETA0 and BETA1 with a few different numbers (try values between -3 and 3 for BETA0 and between -1 and 1 for BETA1). The `x` must be lowercase.

```
BETA0 <- 0; BETA1 <- 0.5
curve(exp(BETA0 + BETA1 * x) / (1 + exp(BETA0 + BETA1 * x)),
      from = -10, to = 10, ylim = c(0,1))
```

4. Summarize the effect of β_0 on the prediction curve. Do the same for β_1.

5. The transformation applied to $\beta_0 + \beta_1 \cdot X$ to get the curve above is called the **logistic function** (hence the term "logistic regression"). Its inverse is called the **logit**. We have

$$\text{logistic}(\eta) = \frac{e^\eta}{1 + e^\eta} \qquad \text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right)$$

The prediction curve for π is therefore equivalent to

$$\text{logit}(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X$$

It seems that we could just reduce logistic regression to linear regression by applying this logit transformation to the response variable. However, this does not work in general. Why? (Hint: What happens when the logit is applied to binary values?) If this doesn't make sense, try it and see what you get for the transformed variable:

```r
library("mosaic"); library("Stat2Data"); data("Election08")
Election08.logit <-
    mutate(Election08, logitObamaWin = log(ObamaWin / (1-ObamaWin)))
```

6. Instead, we need a particular method to fit the parameters. We will get more into the details and interpretation of the coefficients, etc., later, but for now, do the following to fit and plot the model:

```r
election.logistic.model <-
    glm(ObamaWin ~ BA, family = "binomial", data = Election08)
plotModel(election.logistic.model)
```

At roughly what share of BAs is the predicted probability of an Obama win about 50%? (Just eyeball it from the plot)