STAT 213: Regression With a Binary Response

Colin Reimer Dawson

Last Revised April 4, 2018

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?

The dataset we will use in this handout consists of votes of 215 members of the U.S. House of Representatives in 2010 on the bill that would become the Affordable Care Act (AKA "Obamacare"). The `Vote` variable is 1 for members that voted "yes" and 0 for members that votes "no".

Several other variables about each representative are available; we will focus on just one: `ObamaMargin`, which is the margin by which then-president Obama's popular vote total in the representative's home district in the 2008 election exceeded or trailed Sen. John McCain's vote total, as a percentage of the vote in the district. For example if Obama got 55 percent of the vote and McCain got 44, we'd have `ObamaMargin = 11`. The reverse would yield `ObamaMargin = -11`.

Here is a linear regression model:

$$\texttt{Vote}_i = \beta_0 + \beta_1 \cdot \texttt{ObamaMargin}_i + \varepsilon_i$$

1. Using ordinary linear regression, the predicted response value represents to the *mean* response at particular values of predictors. What does the mean of a *binary* response (coded as 0 or 1) represent, more intuitively? (Hint: What does the sum of binary values tell us?)

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

   The relevant data for the model above is in the file `obamacare.csv` at the course website. Load `mosaic`, read in the data with `read.file()`, and fit the above simple linear regression model.

   ```
   library(mosaic)
   Obamacare <- read.file("http://colindawson.net/data/obamacare.csv")
   linear.model <- lm(Vote ~ ObamaMargin, data = Obamacare)
   ```

3. What is the predicted probability of a “yes” vote in a district where Obama and McCain tied? A “deep red” district where Obama’s total was 40 percentage points lower? A “deep blue” district where Obama’s total was 55 points higher?

4. 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 stretched S-shape, flattening out on the extremes.

   Represent the probability of a “yes” outcome (that is, the *mean* of Y at some X) with the symbol π. One S-shaped function has the form

   $$\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

   Use R to plot this function for different values of β_0 and β_1. Type the following, trying different values for β_0 and β_1 (try values between -3 and 3 for β_0 and between -1 and 1 for β_1). The `x` in the first argument to `curve` must be lowercase.

```r
beta0 <- 0; beta1 <- 0.5
prediction.function <- function(x, beta0, beta1)
{
    eta <- beta0 + beta1 * x
    pie <- exp(eta) / (1 + exp(eta))
    return(pie)
}
### We are plotting the curve of the function shown above
curve(prediction.function(x, beta0, beta1),
      from = -10, to = 10, ylim = c(0,1))
```

5. What effect does β_0 have on the prediction curve? What about β_1?

6. If we define $\eta = \beta_0 + \beta_1 \cdot X$ (η is the Greek letter “eta”), then the function that transforms η to π as above is called the **logistic function**. Its inverse (transforming π back to η) is called the **logit**. We have

$$\pi = \text{logistic}(\eta) := \frac{e^{\eta}}{1+e^{\eta}} = \frac{e^{\beta_0+\beta_1\cdot X}}{1+e^{\beta_0+\beta_1\cdot X}}$$
$$\eta = \text{logit}(\pi) := \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot X$$

   In other words, we are modeling the *logit* of π as a *linear* function of X.

   We will get more into the details and interpretation of the coefficients in logistic models later, but for now, do the following to fit and plot the model (assuming you already have the data read in):

```
### Fits a logistic regression model for a binary response
logistic.model <-
    glm(Vote ~ ObamaMargin, family = "binomial", data = Obamacare)
plotModel(logistic.model)
```

   At roughly what margin was the predicted probability of a yes vote about 50%? (Just eyeball it from the plot)