April 1, 2015

NYTimes Article

- Review article posted to class website and discuss the similarities between predicted NCAA brackets and what we will be doing.
- Use of the hybrid method.
- Y_i= 0 if home team lost, 1 if home team won. Represent binary outcomes.

Bottom line: There are fancy predictive models that take into account many predictive factors. But the winning model used a logistic regression and used the Las Vegas point spread to bake in all these other factors. Shows that sometimes smaller predictive models give you better results.

bballbets.csv

- Spread refers to by how much the home team is favored to win. Need to check if point spread does have predictive value.
- When you plot the data in a scatter plot it's impossible to distinguish individual points because the outcome is 0 or 1 so they cluster together. Give an artificial jitter by adding in random noise up or down. Fit the model to the original data obviously and not the jitters.
- You see a much higher fraction of 1's for large values of the point spread
- Goal: create model $P(y_i=1 \mid x_i)$ → probability $y_i = 1$ given the data about x_i
- Discretize the data (opposite of continuous outcome like point spread and turn it into "buckets"). We want an empirical frequency so we define cut points (the buckets).
- `homewin_frequency = mean(homewin ~ spread_category, data=bballbets)`, homewin_frequency → gives us empirical frequency through groupwise means. Is in percentages so for example between -15 and -10 there is a 17.6% chance they will win.

- Best way to analyze is using a linear model in baseline-offset form. Because of the buckets we see the fitted values are flat. So the same probability exists whether 1 or 4. This is a simplification so we will modify it later. The fitted values also form a squished "S" referring the fact that frequencies can't get outside 0 and 1. That explains the bend.
- We need a model more sensitive to the two ends of a bucket. One strategy would be making the buckets smaller (1 point buckets). The big issue is that there are some buckets so small that there are few games in it. So if one home team one and the other lost, then the outcome is 50%. Estimating frequencies with precision from small sample sizes is difficult. There will be a noisy, uncertain estimate. Look at the dip in the top of the fitted values as well—there is a dip between 15 and 20 just by chance. So this problem would be exacerbated with even smaller buckets.

Bottom line: linear models are oversimplified for this example and the use of "buckets" is misleading. We can try a linear probability model but there will be a new problem, so we eventually need to solve both problems at once.

Side Calculations

1) Suppose y is a binary random variable. P(y=1) =w, P(y=0)=1-w. What is E(Y)? (expected value). Means we take the weighted average of the outcomes where the weights are the probabilities. To calculate we say the first outcome is 1 so weight it by the probability Y=1, and same for 0 as the second outcome. So 1*w + 0*(1-w)→ Becomes w. For a binary variable, the expected value is the success probability (w).
2) Suppose $y_i=b_0+b_1x_i+e_i$
   What is the expected value of y_i? $E(y_i)$?
   $E(b_0+b_1x_i+e_i)$
   $E(b_0+b_1x_i) + E(e_i)$
   Simplifies to $b_0+b_1x_i + 0$→ $y\text{-hat}_i$.
   If you plug in a zero-one outcome for y_i, you will get the fitted value for y_i. Using the side calculations together we see that when y_i is a zero-one outcome, $P(y_i=1 \mid x_i) = b_0+b_1x_i$.
   When the outcome is binary, a model by linear least squares lets us predict the probability when the outcome is one. The new problems is that the probabilities exceed 0 and 1. The straight line predicts things that exceed the values they can reasonably take. Rather than just rounding, which is unsatisfying, you should find something that respects the S shape (sigmoid).

Logistic Regression

This is where logistic regression comes in! Two-piece model:

1) Also called parametric probability model: Probability model specifying distribution of the outcome y. Need binary (or bernoulli) probability model. Binominal logistic regression—this is a type of binominal distribution.
   $Y_i \sim \text{Bernoulli}(\mu_i) = \text{Binomia}(n=1, \mu_i)$
   Speed Limits: μ_i must be between 0 and 1.
2) "Link function" that connects μ_i (parameter) to x_i (predictor).
   a) Form a linear function of x_i. Ψ_i (psi)= $b_0+b_1x_i$ (linear predictor).
   b) Run Ψ_i through "speed limit obeying" link function.
      $\mu_i = g(\Psi_i)=g(b_0+b_1x_i)$. G is another function that takes the number from Ψ_i and matches it to the limits of μ_i.

Link function in logistic regression is $\mu_i = e^{\Psi_i} / 1+ e^{\Psi_i} = e^{b0+b1xi}/1+e^{b0+b1xi}$← this will obey "speed limits"
$e^{100}/1+e^{100}$ is essentially 1 but still slightly less than 1. $e^{-100}/1+e^{-100}$ is effectively 0 because approx. 0/0+1. If you plot $e^{\Psi_i} / 1+ e^{\Psi_i}$ it will essentially form the S curve.
This is using the principle of maximum likelihood.

Use glm which stands for generalized linear model. Linear because it still involves predictor function x, but also incorporates a link function. So introducing link function obeys speed limit and therefore solves the two problems. You can also use link function with more than one predictor (b_2x_{i2} or more).

How do you quantify uncertainty in a logistic regression model and make predictions? Use ACL data to practice and continue in class next time.