STATS 551 Homework 5

Missing Data & Finite Mixture Models

Due date: 6:00 pm (EST) Apr. 9, 2018

Practical missing-data imputation (5×10 points). Create a miniature version of the 2010 General Social Survey (publicly available on the Internet), including the following variables: sex, age, ethnicity (use four categories), urban/suburban/rural, education (use five categories), political ideology (on a 7-point scale from extremely liberal to extremely conservative), and general happiness.

1. Using just the complete cases, fit a logistic regression on whether respondents feel 'not too happy.'
2. Impute the missing values using $mi(\ )$ in the mi package in R. Then take one of the completed datasets and fit a logistic regression as above.
3. Repeat, this time imputing using $aregImpute(\ )$ in the Hmisc package.
4. Repeat, this time imputing using $mice(\ )$ in the mice package.
5. Briefly discuss the differences between the four inferences above.

Mixture models (50 points). Football experts provide a point spread for every football game as a measure of the difference in ability between the two teams. For example, team A might be a 3.5-point favorite to defeat team B. The implication of this point spread is that the proposition that team A, the favorite, defeats team B, the underdog, by 4 or more points is considered a fair bet; in other words, the probability that A wins by more than 3.5 points is 1 . If the point spread 2 is an integer, then the implication is that team A is as likely to win by more points than the point spread as it is to win by fewer points than the point spread (or to lose); there is positive probability that A will win by exactly the point spread, in which case neither side is paid off. The assignment of point spreads is itself an interesting

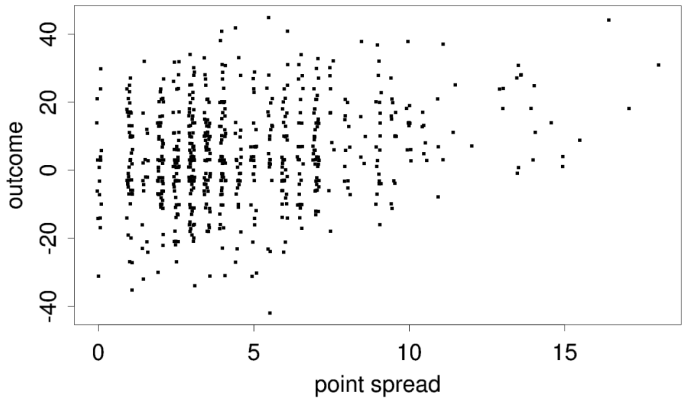

Figure 1.1 *Scatterplot of actual outcome vs. point spread for each of 672 professional football games. The x and y coordinates are jittered by adding uniform random numbers to each point's coordinates (between -0.1 and 0.1 for the x coordinate; between -0.2 and 0.2 for the y coordinate) in order to display multiple values but preserve the discrete-valued nature of each.*

exercise in probabilistic reasoning; one interpretation is that the point spread is the median of the distribution of the gambling population?s beliefs about the possible outcomes of the game. For the rest of this example, we treat point spreads as given and do not worry about how they were derived. The point spread and actual game outcome for 672 professional football games played during the 1981, 1983, and 1984 seasons are graphed in Figure 1.1. (Much of the 1982 season was canceled due to a labor dispute.) Each point in the scatter-plot displays the point spread, x, and the actual outcome (favorite?s score minus underdog?s score), y. (In games with a point spread of zero, the labels 'favorite' and 'underdog' were assigned at random.) A small random jitter is added to the x and y coordinate of each point on the graph so that multiple points do not fall exactly on top of each other.

Figure 1.2a displays the differences $y - x$ between the observed game outcome and the point spread, plotted versus the point spread, for the games in the football dataset. (Once again, random jitter was added to both coordinates.) This plot suggests that it may be roughly reasonable to model the distribution of $y - x$ as independent of x. Figure 1.2b is a histogram of the differences $y - x$ for all the football games, with a fitted normal density superimposed. This plot suggests that it may be reasonable to approximate the marginal distribution of the random variable $d = y - x$ by a normal distribution. The sample mean of the 672 values of d is 0.07, and the sample standard deviation is 13.86, suggesting that the results of football games are approximately normal with mean equal to the point spread and standard

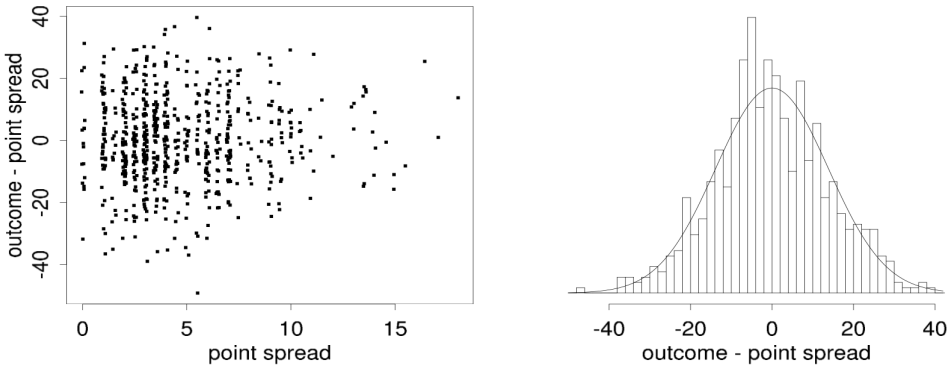

Figure 1.2 *(a) Scatterplot of (actual outcome − point spread) vs. point spread for each of 672 professional football games (with uniform random jitter added to x and y coordinates). (b) Histogram of the differences between the game outcome and the point spread, with the* $N(0, 14^2)$ *density superimposed.*

deviation nearly 14 points (two converted touchdowns). For the remainder of the discussion we take the distribution of d to be independent of x and normal with mean zero and standard deviation 14 for each x; that is, $d|x \sim N(0, 14^2)$, as displayed in Figure 1.2b. The assigned probability model is not perfect: it does not fit the data exactly, and, as is often the case with real data, neither football scores nor point spreads are continuous-valued quantities.

Instead of assuming that the differences between score differential and point spread follow a normal distribution, fit a finite mixture of Gaussians to these data using a symmetric Dirichlet prior with hyper-parameter $1/k$ where k is the number of mixture components. Run a Gibbs sampler to analyze the data, and compare the fitted distribution with that for the normal model. Comment on whether the results suggest the Gaussian density provides a good approximation.

Data: `http://www.stat.columbia.edu/~gelman/book/data/football.asc`

Guideline for Submission: Submit R markdown (or jupyter notebook) with annotated code followed by results. Discussions about the results should follow the results.