SSC 325H: Midterm Exam: Model Answers

March 2013

Scott, MW 8:30 – 10:00

Think of these as model answers rather than a key—that is, rigid adherence was neither expected nor required for full credit, as long as the essential ideas are there.

(1) (5 points) *Define confounder in a single phrase or sentence.*

A confounder is some third variable that is associated with both the predictor and response in a statistical model.

(2) (15 points) Identify the following statements as true or false. For any falsehoods, propose a modification that would make the statement true.

(a) *To describe the behavior of a continuous random variable X, we specify its probability mass function $M(k) = Pr(X = k)$.*

False. Change "continuous" to "discrete." You could also change "probability mass function" to "cumulative distribution function" (preferred) or "probability density function" (still acceptable).

(b) *The p-value associated with a hypothesis test is the probability of having seen an experimental result at least as extreme as the one actually observed, given that the null hypothesis is true.*

True.

(c) *If the Pearson product-moment correlation coefficient between two variables is near zero, we may conclude that there is a weak or nonexistent relationship between those two variables.*

False. Insert "linear" before relationship, or "not" before conclude.

(3) (30 points) *We have talked at length about the idea of statistical adjustment, as distinct from experimental manipulation. Describe your present understanding of statistical adjustment, explaining the various ways in which you have learned to operationalize this concept using models.*

The point of statistical adjustment is to hold some variable constant using a statistical model, rather than an experiment. It is usually done to remove the influence of confounders.

A simple form of statistical adjustment is to match like with like, by breaking a data set into groups. For example, in looking at mammals' dreaming patterns, we adjusted for the danger from predators faced by the various mammals by splitting them into five groups, matched by a "predation danger" index.

If both the outcome (y) and the thing we want to adjust for (x) are continuous, we can fit a regression model for y versus x and take the residuals. If the regression model is good, then the residuals can be thought of as "y adjusted for x," in the sense that they have no systematic dependence on the predictor.

Sometimes we have both continuous predictors and grouping information, as when we knew the age, finishing time, and sex of many different runners in a 10-mile race. If we are interested in the relationship between finishing time and age, then we might need to adjust for a runner's sex by fitting different models for men and women. This will be important if a runner's sex is correlated with both his or her age (e.g. if the men are systematically older) and finishing time. The general principle here is that ignoring a grouping variable associated with both x and y can lead to an aggregation paradox in estimating the x–y relationship.

Finally, another form of statistical adjustment is when we want to construct a "pure" predictor, free of association with some confounder. For example, we may want to assess the relationship between the elevation of an island and the number of plant species on that island, holding the island's area constant. Regressing species count on elevation will not work here: in doing so, we are also implicitly regressing on area, because elevation and area are correlated. To fix this, we can first regress elevation on area and take the residuals. Then we can regress species count versus these residuals, which (as argued above) are a measure of elevation with area held constant.

(4) (50 points) *We have used the idea of resampling, or sampling from your sample, in at least three different ways:*

- *in bootstrapping;*
- *in permutation tests under the Neyman–Pearson framework; and*
- *in cross-validation.*

Explain how and why resampling is used in each of these contexts. That is: what kinds of questions are we trying to answer in each case, and how does resampling help us answer them? Cite at least one example for each case, drawn from the class materials or your own imagination/experience. (By "example," I mean a substantive scientific question that could be answered using the method in question.)

The point of resampling is to simulate the forces of randomness that played a role in generating our data set.

In bootstrapping, the goal is to quantify the precision of an estimator—for example, the least-squares estimate of a truck's price versus its mileage, or the maximum-likelihood estimate of the rate of a Poisson model for photon counts. The issue is sampling variability. A different sample would have led to a different estimate. We want to know how different the estimate might have been. A more formal way of saying this is that our estimator has a sampling distribution across all possible samples that might have been drawn from the same population. We can estimate this sampling distribution by repeatedly drawing bootstrapped samples, which are samples taken with replacement from our original sample, and of the same size as the original sample. For each bootstrapped sample, we compute the statistic of interest. The standard deviation of these statistics across all the bootstrapped samples is called the standard error, and (usually) gives a decent estimate of the spread of the estimator's sampling distribution.

In permutation testing, the goal is to assess whether some observed association in a data set is plausibly due to chance (often called the "null hypothesis" or H_0). Suppose we have some statistic t that measures association between two outcomes x and y—perhaps the odds ratio in a two-by-two contingency table of treatment versus clinical outcome, or R^2 in a regression model. We wish to simulate $p(t \mid H_0)$, the sampling distribution of t under the null hypothesis that x and y are unrelated. To do this, we repeatedly "reshuffle the cards" by randomly re-assigning each person a different value of x (drawn from the actual data), and calculating t for the reshuffled data. This explicitly breaks the connection between x and y. Under the Neyman–Pearson framework, we inspect $p(t \mid H_0)$, choose a rejection region and calculate its alpha level (or vice versa), and report whether the observed statistic fell into the rejection region.

In cross-validation, the goal is to assess how well a statistical model could be used to predict new data, without actually collecting any new data. Instead, we split our data into "training" and "testing" subsets. We use the training data to fit the model, and the testing data to check our predictions. For example, we might want to choose the order of a polynomial regression model: higher-order models will provide a tighter in-sample fit, but might generalize poorly if we are unable to estimate their coefficients precisely enough using the available data.