SDS 325H Midterm Model Answers, Spring 2015

Question 1: Short answers (30 points)

A) *Under what general circumstances would you expect to need an interaction term in a statistical model? Give a specific example, either from class or your own imagination, of where this concept is important.* Suppose we want to build a model for y versus x_1 and x_2. An interaction term will be necessary anytime the joint effect of x_1 and x_2 on y is different than the sum of the individual effects: y = (effect of x_1) + (effect of x_2) + (joint effect of x_1 and x_2). We used this tool on the reaction-times data set, where the joint effect of a scene being Littered and Far Away was larger than the sum of the two individual effects.

B) *Suppose that we have data on a response y_i and a predictor x_i and that we want to estimate a power law of the form $\hat{y}_i = Kx_i^{\beta}$ for some constant K and exponent β. Explain how we would use ordinary least squares (OLS) to fit this nonlinear model. Relate the quantities K and β to the coefficients from the OLS fit. Give a specific example, either from class or your own imagination, of where this trick is useful.* Take the logarithm of both variables to get the equation $\log \hat{y}_i = \log K + \beta \log x_i$. This is a linear equation in log x and log y, with coefficients $\beta_0 = \log K$ and β_1 as the original exponent. We used this to fit a power law to brain weight versus body weight.

C) *Briefly describe the logic behind a permutation test and how it fits into the Neyman-Pearson framework of hypothesis testing.* If x_i and y_i are associated, we can break this association by randomly permuting the values of x_i, i.e. re-assigning them at random to cases, irrespective of y_i. This device to used to simulate the sampling distribution of some test statistic T under the null hypothesis that x and y are unrelated. T measures the association between x and y, and is recomputed for each permutation.

Question 2: Sampling distributions (40 points)

A) *Concisely define the term "sampling distribution". Why is this concept useful for quantifying uncertainty in statistical models?* A sampling distribution of an estimator is the probability distribution of values we get for that estimator when we compute it for different random samples from the population. It quantifies how stable the estimator is under random sampling, and therefore how much trust we can place it it for any particular random sample.

B) *One way of estimating a sampling distribution is via bootstrapping. Concisely describe this procedure—both what we do and why we do it.* We simulate taking samples of size n from the population by taking samples of size n with replacement from our original sample. We do this to approximate an estimator's sampling distribution so that we can compute standard errors, confidence intervals, etc.

C) *In class and the course packet, it was stated that confidence intervals generated via bootstrapping will "approximately" satisfy the frequentist coverage property. What does the frequentist coverage property entail?* The FCP is a property of a procedure used to generate confidence intervals. If the procedure produces X% confidence intervals that contain the true parameter X% of the time under repeated use, then the procedure satisfies the FCP.

D) *Recall that "pseudo-code" lays out the logical structure of a computer program in a human-readable way, without reference to the specific commands in any one language (like R). Write pseudo-code for a Monte Carlo simulation that could be used to judge the statement in part C—that is, whether bootstrapped confidence intervals satisfy the frequentist coverage property.*

1. Fix true parameters of model and sample size.

2. Repeat these steps many times:

a) Generate data set using true parameters.
b) Bootstrap data set.
c) Form confidence interval from bootstrapped sampling distribution.
d) Check whether interval covers the true value.

3. Calculate the coverage frequency from simulation.

Question 3: Statistical adjustment (30 points)

Briefly describe the logic of statistical adjustment in each of the following contexts. For each one, give an example, either from class or from your own vivid imagination.

A) *In ordinary linear regression with a single predictor. Here, we wished to look at the y (response) variable after "controlling for" or "adjusting for" the x (predictor) variable.* Here we fit a model for y versus x and compute the residuals. By definition these residuals have the "x-ness" removed or subtracted: $e_i = y_i - (\beta_0 + \beta_1 x_i)$. Example: calculate restaurant value as the price of a meal, adjusted for food quality.

B) *In regression models with both a numerical predictor (x_1) and a categorical/grouping predictor (x_2). Here, we wished to estimate the relationship between the response y and the numerical predictor x_1, after adjusting for group membership.* We added dummy variables for the categorical predictor. This allow the fitted line for y versus x_1 to move up or down depending on group membership, thereby adjusting for the grouping variable and avoiding a possible aggregation paradox. If required, we could also add an interaction term if we believed that y changed faster/slower as a function of x across the different categories. Example: estimate effect of aging on finishing time in a race, adjusting for a person's sex.

C) *In multiple regression models, with more than one numerical predictor.* Suppose we fit a multiple regression model $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. If we want the effect of x_2 on y adjusting for x_1, we have to subtract the x_1 effect. This gives $\hat{y}_i - \beta_1 x_{i1} = \beta_0 + \beta_2 x_{i2}$. This says that y adjusted for x_1 is an ordinary regression on x_2 with slope β_2, and that the coefficient in the multiple-regression model is a partial slope: how fast y changes as a function of x_2, holding x_1 constant. Example: estimate the association between species count and area of island, adjusting for differences in elevation.