CHAPTER 11

PITFALLS OF STATISTICAL INFERENCE

Exercise 11.1 (Selecting Predictors Using Screening)**.** A forecaster wants to predict hurricanes for the season. To do this, the forecaster evaluates the correlation between hurricane counts and 50 variables around the globe that may be relevant. The correlations are computed using 30 years of data. Based on this analysis, the forecaster selects the predictor with the maximum correlation, which has a correlation of 0.6 with hurricane counts, and uses that predictor to make a forecast. Is the predictor statistically significant at the 5% level? For your null hypothesis, you may assume the 50 variables are independent and normally distributed. Does your conclusion change if the variables are not independent? □

Exercise 11.2 (Selecting the Best Forecasts)**.** A forecaster has nine forecasts and would like to combine them to make a single, superior forecast. The forecaster decides to select the four forecasts that are most correlated with observations from a 20-year historical record, and then average those forecasts to make a prediction. When the four best forecasts are averaged, the correlation between the mean 4-member forecast and observations is 0.55. For reference, the 5% significance level for a 20-year forecast is about 0.45. Since the actual correlation exceeds the threshold, the forecaster claims that the 4-member mean forecast has a statistically significant correlation. What is "wrong" with this argument? What is the experimentwise error rate for this procedure? Plot a histogram of the correlation skill under the null hypothesis of independent variables.

Hint: generate random numbers to simulate the above scenario and perform the same steps as the forecaster. Repeat this procedure numerous times to estimate an empirical distribution of the correlation under the null hypothesis of no-skill (all variables are in-

dependent). You can use the `cor` function to determine the correlation between 1 variable and 9 others (e.g., if `obs` is 20 elements long, and `frcs` is a 20×9 matrix, then `as.numeric(cor(frcs,obs))` gives 9 correlations). Also, the `order` command allows you to identify the forecasts with the four largest correlations.

□

Exercise 11.3 (Discovery of a new Predictor)**.** An investigator believes that a particular index, called *DOOM*, is useful for predicting JFM temperature. The 50-year correlation between JFM temperature and *DOOM* is 0.2 in December and 0.1 in November, but 0.3 in October. Since the 5% significance threshold for a correlation coefficient based on 50 independent samples is about 0.28, the investigator promptly publishes a paper proclaiming that the October value of *DOOM* is a good predictor of JFM temperature. Has the significance test been applied properly? If not, why not? Is the predictor significant when screening is taken into account? If not, how large would the correlation need to be to ensure a true 5% significance? □