Homework Three

Statistics 151a (Linear Models)

Due by midnight on March 19, 2020

Instructions: Please submit with a cover sheet that has your name and student ID.

1. **Fox Exercise 7.1.** Suppose that the values -1 and 1 are used for dummy regressors (e.g. D in Equation 7.1) instead of 0 and 1. Write out the regression equations and explain how the parameters of the model are to be interpreted. Does this alternative coding of the dummy regressor adequately capture the effect of a binary variable? Is it fair to conclude that the dummy-regression model will "work" properly as long as any two distinct values of the dummy regressor are employed? Is there a reason to prefer one coding to another?

2. **Fox Exercise 9.1(a).** Solving the parametric equations in one-way and two-way ANOVA: Show that the parametric equation (Equation 9.5, page 205) in one-way ANOVA has the general solution

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix} = \begin{bmatrix} \mu_\cdot \\ \mu_1 - \mu_\cdot \\ \mu_2 - \mu_\cdot \\ \vdots \\ \mu_{m-1} - \mu_\cdot \end{bmatrix}$$

3. **Fox Exercise 9.14.**

   Prediction: one use of a fitted regression equation is to *predict* response-variable values for particular "future" combinations of explanatory-variable scores. Suppose, therefore, that we fit the model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ (with full assumptions on ε), obtaining the least squares estimate $\mathbf{b}$ of $\boldsymbol{\beta}$. Let $\mathbf{x}_0' = [1, x_{01}, ..., x_{0k}]$ represent a set of explanatory-variable scores for which a prediction is desired, and let Y_0 be the (generally unknown, or not yet known) corresponding value of Y. The explanatory variable vector $\mathbf{x}_0'$ does not necessarily correspond to an observation in the sample for which the model was fit.

a) If we use $\hat{Y}_0 = \mathbf{x}_\mathbf{0}'\mathbf{b}$ to estimate $E(Y_0)$, then the error in estimation is $\delta \equiv \hat{Y}_0 - E(Y_0)$. Show that if the model is correct, then $E(\delta) = 0$ (i.e. $\hat{Y}_0$ is an unbiased estimator of $E(Y_0)$) and that $V(\delta) = \sigma_\varepsilon^2 \mathbf{x}_\mathbf{0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

b) We may be interested not in estimating the *expected* value of Y_0 but in predicting or forecasting the *actual* value of $Y_0 = \mathbf{x}_\mathbf{0}'\boldsymbol{\beta} + \varepsilon_0$ that will be observed. The error in the forecast is then

$$D \equiv \hat{Y}_0 - Y_0 = \mathbf{x}_\mathbf{0}'\mathbf{b} - (\mathbf{x}_\mathbf{0}'\boldsymbol{\beta} + \varepsilon_0) = \mathbf{x}_\mathbf{0}'(\mathbf{b} - \boldsymbol{\beta}) - \varepsilon_0$$

Show that $E(D) = 0$ and that $V(D) = \sigma_\varepsilon^2[1 + \mathbf{x}_\mathbf{0}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]$. Why is the forecast error D greater than the variance of δ found in part (a)?

c) Use the results in part (a) and (b), along with the Canadian occupational prestige regression (see Section 5.2.2), to predict the prestige score for an occupation with an average income of \$12,000, an average education of 13 years, and 50% women. Place a 90% confidence interval around the predcition, assuming (i) that you with to estimate $E(Y_0)$ and (ii) that you wish to forecast an actual Y_0 score. (Because σ_ε^2 is not known, you will need to use S_E^2 and the t-distribution.)

d) Suppose that the methods of this problem are used to forecast a value of Y for a combination of Xs very different from the X values in the data to which the model was fit. For example, calculate the estimated variance of the forecast error for an occupation with an average income of \$50,000, an average education of 0 years, and 100% women. Is the estimated variance of the forecast error large or small? Does the variance of the forecast error adequately capture the uncertainty in using the regression equation to predict Y in this circumstance?

4. **Fox Exercise 21.3.** Random versus fixed resampling in regression (note: you can get the data for this problem at `https://socialsciences.mcmaster.ca/jfox/Books/RegressionDiagnostics/index.html`):

a) Recall (from Chapter 2) Davis's data on measured and reported weight for 101 women engaged in regular exercise. Bootstrap the least-squares regression of reported weight on measured weight, drawing r = 1000 bootstrap samples using (1) random-X resampling and (2) fixed-X resampling. In each case, plot a histogram (and, if you wish, a density estimate) of the 1000 bootstrap slopes, and calculate the bootstrap estimate of standard error for the slope. How does the influential outlier in this regression affect random resampling? How does it affect fixed resampling?

b) Randomly construct a data set of 100 observations according to the regression model $Y_i = 5 + 2x_i + \epsilon_i$, where $x_i = 1, 2, ..., 100$, and the errors are independent (but seriously heteroscedastic), with $\epsilon_i \sim N(0, x_i^2)$ As in (a), bootstrap the least-squares regression of Y on x, using (1) random resampling and (2) fixed resampling. In each case, plot the bootstrap distribution of the slope coefficient, and calculate the bootstrap estimate of standard error for this coefficient. Compare the results for random and fixed resampling. For a few of the bootstrap samples, plot the least-squares residuals against the fitted values. How do these plots differ for fixed versus random resampling?

c) Why might random resampling be preferred in these contexts, even if (as is not the case for Davis's data) the X-values are best conceived as fixed?

5. **Fox Exercise 21.4.** Bootstrap estimates of bias: The bootstrap can be used to estimate the bias of an estimator $\hat{\theta}$ of a parameter θ, simply by comparing the mean of the bootstrap distribution $\bar{\theta}*$ (which stands in for the expectation of the estimator) with the sample estimate $\hat{\theta}$ (which stands in for the parameter); that is, $\hat{bias} = \bar{\theta} * -\hat{\theta}$. (Further discussion and more sophisticated methods are described in Efron and Tibshirani, 1993, chap. 10.) Employ this approach to estimate the bias of the maximum-likelihood estimator of the variance, $\hat{\sigma}^2 = \sum(Y_i - \bar{Y})^2/n$, for a sample of n = 10 observations drawn from the normal distribution N(0, 100). Use r = 500 bootstrap replications. How close is the bootstrap bias estimate to the theoretical value $-\sigma^2/n = -100/10 = -10$.