Homework 4

Due Date: 11:59pm (Eastern time), Friday Nov 6, 2020

Questions 5.6, 5.7, 5.17, and 8.3 in Applied Linear Regression book. See below for descriptions of these questions.

5.6 The coding of factors into dummy variables described in the text is used by default in most regression software. Older sources, and sources that are primarily concerned with designed experiments, may use *effects coding* for the dummy variables. For a factor X with d levels $\{1, 2, \ldots, d\}$ define $V_j, j = 1, \ldots, d-1$ with elements v_{ji} are given by:

$$v_{ji} = \begin{cases} 1 & i = j \\ -1 & i = d \\ 0 & \text{otherwise} \end{cases}$$

The mean function for the one-factor model is then

$$\mathrm{E}(Y|V_1, \ldots, V_{d-1}) = \eta_0 + \eta_1 V_1 + \cdots + \eta_{d-1} V_{d-1} \tag{5.18}$$

5.6.1 Show that the mean for the jth level of the factor is $\eta_0 + \alpha_j$, where

$$\alpha_j = \begin{cases} \eta_j & j \neq d \\ -(\eta_1 + \eta_2 + \cdots + \eta_{d-1}) & j = d \end{cases}$$

By taking the mean of the level means show that η_0 is the mean of the response ignoring the factor. Thus, we can interpret α_j, the difference between the overall mean and the level mean, as the effect of level j, and $\sum \alpha_j = 0$.

5.7 Suppose X_1 were a continuous predictor, and F is a factor with three levels, represented by two dummy variables X_2 with values equal to 1 for the second level of F and X_3 with values equal to 1 for the third level of F. The response is Y. Consider three mean functions:

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{5.19}$$

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 \tag{5.20}$$

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 (x_1 - \delta) + \beta_{12} (x_1 - \delta) x_2 + \beta_{13} (x_1 - \delta) x_3 \tag{5.21}$$

Equation (5.21) includes an additional unknown parameter δ that may need to be estimated.

All of these mean functions specify that for a given level of F the plot of $\mathrm{E}(Y|X_1, F)$ is a straight line, but in each the slope and the intercept changes. For each of these three mean functions, determine the slope(s) and intercept(s), and on a plot of Y on the vertical axis and X_1 on the horizontal axis, sketch the three fitted lines.

The model (5.21) is a generalization of (5.20). Because of the extra parameter δ that multiplies some of the βs, this is a nonlinear model; see Saw (1966) for a discussion.

Homework 4
Due Date: 11:59pm (Eastern time), Friday Nov 6, 2020

5.17 Sex discrimination (Data file: `salary`) The data file concerns salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The variables include `degree`, a factor with levels PhD and MS; `rank`, a factor with levels Asst, Assoc, and Prof; `sex`, a factor with levels Male and Female; `Year`, years in current rank; `ysdeg`, years since highest degree, and `salary`, academic year salary in dollars.

5.17.1 Get appropriate graphical summaries of the data and discuss the graphs.

5.17.2 Test the hypothesis that the mean salary for men and women is the same. What alternative hypothesis do you think is appropriate?

5.17.3 Assuming no interactions between `sex` and the other predictors, obtain a 95% confidence interval for the difference in salary between males and females.

5.17.4 Finkelstein (1980), in a discussion of the use of regression in discrimination cases, wrote, "[a] variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be 'tainted.'" Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may not be acceptable to the courts.

Exclude the variable `rank`, refit, and summarize.

Data link: http://users.stat.umn.edu/~sandy/alr4ed/data/

8.3 (Data file: `water`) A major source of water in Southern California is the Owens Valley. This water supply is in turn replenished by spring runoff from the Sierra Nevada mountains. If runoff could be predicted, engineers, planners, and policy makers could do their jobs more efficiently. The data file contains snowfall depth measurements over 43 years taken at six sites in the mountains, in inches, and stream runoff volume at a site near Bishop, California. The three sites with names starting with "O" are fairly close to

each other, and the three sites starting with "A" are also fairly close to each other. `Year` is also given in the data file, but should not be used as a predictor.

8.3.1 Construct the scatterplot matrix of the data, and provide general comments about relationships among the variables.

8.3.2 Using the methodology for automatic choice of transformations outlined in Section 8.2.2, find transformations to make the transformed predictors as close to linearly related as possible. Obtain a test of the hypothesis that all $\lambda_j = 0$ against a general alternative, and summarize your results. Do the transformations you found appear to achieve linearity? How do you know?

8.3.3 Given log transformations of the predictors, show that a log transformation of the response is reasonable.

8.3.4 Consider the multiple linear regression model with mean function given by

```
log(BSAAM) ~ log(APMAM)+log(APSAB)+log(APSLAKE)+
             log(OPBPC)+log(OPRC)+log(OPSLAKE)
```

with constant variance function. Estimate the regression coefficients using OLS. You will find that two of the estimates are negative; Which are they? Does a negative coefficient make any sense? Why are the coefficients negative?

8.3.5 Test the hypothesis that the coefficients for the three "O" log predictors are equal against the alternative that they are not all equal. Repeat for the "A" predictors. Explain why these might be interesting hypotheses. (*Hint:* The geometric mean of the regressors `OPBPC`, `OPRC`, `OPSLAKE` is equal to exp[(log(`OPBPC`) + log(`OPRC`) + log(`OPSLAKE`))/3], and so the sum [log(`OPBPC`) + log(`OPRC`) + log(`OPSLAKE`)] is proportional to the logarithm of the geometric mean of these regressors. If the three coefficients are equal, then we are essentially replacing the three predictors by one regressor equivalent to the logarithm of their geometric mean.)