PBHS 32410 / STAT 22401 Problem Set 4 Winter 2025

Possible points 84

1. [30 pts] From SPRM Chapter 7

At the end of the chapter is a dataset examining predictors of terrestrial level ozone. The data can be found on the website. Carry out the following analyses:

(a) Fit a model using all predictors and briefly describe, Compute the predicted value and raw residuals. Plot the raw residuals vs. actual ozone values to check general randomness.

(b) Calculate the standardized residuals. Plot against the predicted values. to identify any extremes

(c) Identify any was how those observations with absolute value of standardized residual greater than 2.0 are different from others.

(d) Use the QQ-plot to check normality of the standardized residuals. Also look at the mean and standard deviation of these residuals to see how they conform to the normality assumption.

(e) Check the leverage and Cook's distance values for the model. Plot against the id variable to identify the extremes.

3. [24 pts] Modified from C&H Exercise 4.7

Consider again the Cigarette Consumption data from HW 2).

(a) Predict sales with the predictor variables we used earlier.

$$Sales = \beta_0 + \beta_1 Age + \beta_2 Income + \beta_3 Price + \epsilon$$

Check the basic model assumptions (randomness of residuals, normality of residuals).

(b) Plot the standardized residuals against each predictor. This will identify whether any residuals are associated with unusual (i.e., at extremes of range) predictor variables

(c) Check for any extreme values and influential observations using residuals and Cook's distance. For any values identified as not fitting well or having high influence, examine the cases compared to the rest of the dataset and comment on anything you find.

3. [26 pts total] Brain vs. Body Weight Data

For the brain vs. body weight data we examined in class, it would be interesting to examine the hypothesis by some breakdown of type of animal. We could separate mammals, but what might be more interesting would be to evaluate primates vs. others. The data are in

`brain_12.dta`

(a) [6 pts] Create an indicator variable for primates (monkeys, apes, us, etc). Test whether this indicator improves the model and is a statistically significant predictor. You should model on the natural logarithm scale for both response and predictor (body weight).

(b) [6pts] Create and evaluate an interaction term between log body weight and primate. Interpret the result, and whether to include this term (and/or the primate indicator) in the model. What does the model say about brain vs. body weight across these species classes?

(c) [6 pts] Run on the arithmetic scale (just use brain and body weight directly without transformation) a model on a reduced dataset omitting all dinosaurs. How does it look? Why does this work better now than for the whole dataset with dinosaurs?

(d) [12 pts] Beginning with the predictor (body weight) already (natural) log transformed and the response (brain weight) in original form, run a regression model and comment on the results. Then try the Box-Cox model to see what transform on the response (brainweight) is suggested. What is the conclusion? Compare the estimated coefficient to that we obtained in class using the log transform.