Homework 6

36-401, Fall 2015, Section B

Due at the start of class on 29 October 2015

1. Refer to the **SENIC** data set in Appendix C.1 of the textbook. (The appendix is reproduced on Blackboard.) We're interested in using age, number of beds, infection risk, and available facilities/services to predict the average length of hospital stay of all patients in hospital.

   (a) (5) Create the pairs plot a correlation matrix. Describe these (briefly).

   (b) (5) Fit a model regressing the average length of hospital stay only on percent of service provided by the hospital. What is the slope, and its standard error?

   (c) (5) Fit a linear regression model for four predictor variables and state the estimated regression function.

   (d) (5) What is the difference between the coefficient on service in problem 1b and 1c? Are both of them statistically significant at the 5% level? If you see any difference, provide an explanation based on your observation in 1a.

   (e) (5) Plot the residuals versus the fitted values, and versus each of the predictor variables. Also create a QQ-plot. Describe your diagnostics; are any transformations necessary? If so, make them and then double-check the subsequent diagnostics.

   (f) (5) Give an interpretation of the coefficient of infection risk. Test, for an appropriate α, the hypothesis that the slope relating infection risk and hospital stay $\neq 0$. State your conclusion in the context of the variables in this problem.

   (g) (5) What is the root mean squared error? Is the model a good fit, over all? Is it useful? (Explain, carefully.)

   (h) (5) Obtain an interval estimate of the expected value of average length of hospital stay when average age=54, number of beds=100, infection risk=5%, and service=30%. Use 95% confidence level, and provide an interpretation.

   (i) (5) Obtain a prediction interval for the average length of stay for a new hospital with average age=58, number of beds=200, infection risk=6%, and service=40%. Use 99% confidence level, and provide an interpretation.

2. Refer to the `birthwt` data set from the `MASS` library (`library(MASS); help(birthwt)`).

   (a) (5) Fit a model predicting birth weight using mother's age, mother's weight, smoking status, self-reported race, and number of previous premature labors. For race, use Caucasian as the reference group. Include your summary output from R. Interpret your model's coefficients for self-reported race and being a smoker.

   (b) Re-design the model to find the coefficient of non-Caucasian vs. Caucasian.

      i. (5) Report and interpret the point estimate of this coefficient.

      ii. (5) Describe what effects this change had on the rest of your model (estimates, std errors, p-values). Which variables were affected? Which weren't?

   (c) (5) Someone suggests reducing the terms in the model (to increase stability) by using self-reported race as an ordered categorical variable. Good idea? Why/why not?

   (d) There are only a few unique values for number of previous premature labors; we might be better off treating this variable as categorical.

      i. (5) Keeping race as non-Caucasian vs. Caucasian, re-design the model such that each number of labors is a categorical variable, with zero labors as the reference. Interpret all adjusted effects associated with the number of premature labors. Given these effects, would you keep number of previous premature labors as an ordered categorical/continuous variable? Why/why not?

      ii. (5) Re-design your model to find the adjusted effects of 1 labor vs. 0 labors and > 1 labor vs. 0 labors. Interpret your effects. How have they changed?

      iii. (5) Re-design your model to find the adjusted effect of > 0 labors vs. 0 labors. Interpret your effect. How have things changed?

      iv. (5) Given your results in i, ii, iii), what would be your final decision about modeling the number of premature labors? Give at least one advantage and one disadvantage of your choice.

   (e) (5) You also have the variable `ftv`, the number of physician visits during the first trimester, with a similar small number of unique values. Leaving race as non-Caucasian vs. Caucasian and premature labors as > 0 vs. 0, explore different ways to include the number of physician visits in the model. Explain your final choice; present summary output from your final model and give at least one advantage and one disadvantage related to your choice of `ftv` effects.

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a

plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, please only submit the knitted file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.