CHENNAI MATHEMATICAL INSTITUTE

Predictive Analytics: Regression and Classification

Final Exam of Aug-Nov, 2022

Instructor: Ajay Shah

Total Time: 3 hours

Total Marks: 100

1) Prediction for a comparison: A linear regression is fit on high school students modeling grade point average given household income. Write Julia code to compute the 90% predictive interval for the difference in grade point average comparing two students, one with household incomes of $40, 000 and one with household income of $80, 000.
2) Correlation and explained variance: In a least squares regression with one predictor, show that R^2 equals the square of the correlation between x and y.
3) Log-log transformations: Suppose that, for a certain population of animals, we can predict log weight from log height as follows:
   - An animal that is 50 centimeters tall is predicted to weigh 10 kg.
   - Every increase of 1% in height corresponds to a predicted increase of 2% in weight.
   - The weights of approximately 95% of the animals fall within a factor of 1.1 of predicted values.

   a) Give the equation of the regression line and the residual standard deviation of the regression.

   (b) Suppose the standard deviation of log weights is 20% in this population. What, then, is the R2 of the regression model described here?
4) Logarithmic transformation and regression: Consider the following regression: log(weight) = −3.8 + 2.1 log(height) + error, with errors that have standard deviation 0.25. Weights are in pounds and heights are in inches.

(a) Fill in the blanks: Approximately 68% of the people will have weights within a factor of and of their predicted values from the regression.

(b) Using pen and paper, sketch the regression line and scatterplot of log(weight) versus log(height) that make sense and are consistent with the fitted model. Be sure to label the axes of your graph.

5) The algebra of logistic regression with one predictor: You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60, 000. Determine the logistic regression model that is consistent with this information. For simplicity, you may want to assume that income is measured in units of $10, 000.

6) Offset in a Poisson or negative binomial regression: Explain why putting the logarithm of the exposure into a Poisson or negative binomial model as an offset, is equivalent to including it as a regression predictor, but with its coefficient fixed to the value 1.

7) Decline effect: After a study is published on the effect of some treatment or intervention, it is common for the estimated effect in future studies to be lower. Give five reasons why you might expect this to happen.

8) Suppose that the zinc study described in Section 16.3 would cost $150 for each treated child and $100 for each control. Under the assumptions given in that section, determine the number of control and treated children needed to attain 80% power at minimal total cost. You will need to set up a loop of simulations as illustrated for the example in the text. Assume that the number of measurements per child is fixed at K = 7 (that is, measuring every two months for a year).

9) Experiment with pre-treatment information: An intervention is hoped to increase voter turnout in a local election from 20% to 25%.

(a) In a simple randomized experiment, how large a sample size would be needed so that the standard error of the estimated treatment effect is less than 2 percentage points?

(b) Now suppose that previous voter turnout was known for all participants in the experiment. Make a reasonable assumption about the correlation between turnout in two successive elections. Under this assumption, how much would the standard error decrease if previous voter turnout was included as a pre-treatment predictor in a regression to estimate the treatment effect?

10) Sample size calculations for main effects and interactions: In causal inference, it is often important to study varying treatment effects: for example, a treatment could be more effective for men than for women, or for healthy than for unhealthy patients. Suppose a study is designed to have 80% power to detect a main effect at a 95% confidence level. Further suppose that interactions of interest are half the size of main effects.

(a) What is its power for detecting an interaction, comparing men to women (say) in a study that is half men and half women?

(b) Suppose 1000 studies of this size are performed. How many of the studies would you expect to report a "statistically significant" interaction? Of these, what is the expectation of the ratio of estimated effect size to actual effect size?