Stat 482 Exam # 1 December 2, 2021

1. Experience with a certain type of plastic indicates that a relationship exists between the hardness (measured in Brinell units) of items molded from the plastic (response variable y) and the elapsed time (measured in hours) since the termination of the molding process (input variable x). Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels. The data is available on Blackboard as a csv file.

(a) State the simple linear regression model.
(b) Provide an interpretation for the regression parameter β_1, stated in the context of the problem.
(c) Compute an interval estimate for β_1.

(d) Compute a confidence interval for μ_h at $x_h = 40$ hours.
(e) Compute a prediction interval for $Y_{h(new)}$ at $x_h = 40$ hours.
(f) Explain the difference between a confidence interval and a prediction interval, stated in the context of the problem.

(g) State the equations for $E(MSR)$ and $E(MSE)$.
(h) Use part (g) to explain a motivation behind the F test for input effects.
(i) Compute the F^* statistic and the p-value.
Provide an interpretation of the result, stated in the context of the problem.
(j) Compute the coefficient of determination r^2.
Provide an interpretation of the result, stated in the context of the problem.

(k) State the full model (saturated model) in testing for regression model fit.
(l) Compute $SSPE$ and $SSLF$ for testing the fit of the linear regression model.
(m) Compute the F^*_{LF} statistic and the p-value.
Provide an interpretation of the result, stated in the context of the problem.
(n) Create a plot comparing the fitted values from the regression model with the fitted values from the saturated model.

(o) Explain an advantage of using the regression estimate $\widehat{\mu}_{40}$ instead of the sample mean $\overline{y}_{40}$, and explain when this advantage will lead to a more accurate estimator.

2. Measurements were made on men involved in a physical fitness course. The input variables under consideration are age (in years), weight (in kgs), time to run 1.5 miles (in minutes), heart rate while resting (in bpm), heart rate while running, and maximum heart rate. The goal is to determine which input variables are needed to model the oxygen uptake rate (ml/kg body weight per minute). The data is available on Blackboard as a csv file.

(a) Describe the goal of the discrepancy function approach to model selection. How is the best model defined? What are the two sources of model error?
(b) Plot the Cp statistic against the model dimension. Plot the Cp statistic against the candidate models. Which variables are included in the selected model?
(c) Test the selected model against its best competitor having fewer parameters. (Compute the test statistic and the p-value.) How does discrepancy based model selection compare to p-value based selection?