CSC 423/324 – Early Final Exam
May 14, 2012

Part A. Multiple Choice Problems. 3 pts. each. Answer of 20 questions. For each question give an optional for possible partial credit. CSC 423 students do all 11 questions. CSC 324 students do 10 out of 11 questions.

1. What is the definition of a random variable?
   a. A process of choosing a random number.
   b. The extreme values of a normal distribution.
   c. The extreme values of a normally distributed dataset.
   d. A function from the independent variable space to the dependent variable space.

2. The horizontal distance between the inflection points of a normal density for a population is
   a. σ b. 2σ c. σ^2 d. $\mu + \sigma$

3. Which of the following distributions has thin tails?
   a. The normal distribution N(0, 1). b. The normal distribution N(0, 0.001)
   c. The t distribution with 2 degrees of freedom. d. The uniform distribution

4. What is the most important reason why data from observational studies more difficult to interpret than data obtained from an experiment with treatments randomly assigned to subjects? It is hard to
   a. create a normal plot with data from observational studies.
   b. distinguish dependent variables from independent variables.
   c. find subjects willing to participate in a randomized study.
   d. tell if the effect being studied is due to the treatments or is due to some variable not included as an independent variable in the model.

5. For an independent two-sample t-test, if we assume that the variances of the two groups are not equal, which test is the most commonly used? The _____ test.
   a. Bertrand-Fisher b. Cauchy-Schwarz c. Kolmolgorov-Smirnov d. Welsh-Satterthwaite

6. Which of the following could be most easily tested with a paired two -sample t-test? Give a reason to support your answer.
   a. Whether automobile Model A gets better gas mileage than automobile Model B.
   b. Whether Candidate A or Candidate B is more likely to win the election next month.
   c. Whether a new drug lowers chlorestoral better than a currently popular drug.
   d. Which of two websites is easier to use.

7. What is the value of Cov($\boldsymbol{\varepsilon}$) if the residual vector is homoscedastic and uncorrelated? $\boldsymbol{H}$ is the hat matrix.
   a. 0 b. $\sigma^2\,\boldsymbol{I}$ c. $\sigma^2\,\boldsymbol{H}$ d. $\sigma^2\,(\boldsymbol{I}-\boldsymbol{H})$

8. An estimated regression parameter is unbiased if
   a. its expected value is equal to the value of the corresponding true regression parameter.
   b. its standard error is smaller than any other estimated regression parameter.
   c. only if the regression model is regression through the origin.
   d. the MSE for the regression model equals $(n-1)s_x^2$.

9. For a multiple regression model, if the sum of squares for error SSE equals the total sum of squares for error SST, then the R-squared value is
   a. 0.0 b. 0.25 c. 0.50 d. 1.00

10. If $\boldsymbol{H}$ is the hat matrix, then $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$ represents the
    a. vector of predicted values. b. vector of residuals.
    c. standard errors of the estimated parameters. d. vector of estimated parameters.

11. A large variance inflation factor for an estimated parameter indicates
    a. Heteroscedasticity b. Multicollinarity
    c. A leverage point d. An outlier

Part B: Short Answer Questions. Answer the following questions. Show your work for each question. CSC 423 students do all questions. CSC 324 students do 7 out of 9 questions.

1. A dataset having 16 observations is normally distributed with $\bar{x}$ = 3.72 and s_x = 0.245. Find a 99% confidence interval for the true value of μ.

2. What is the second normal score computed using Van der Waerden's method if n = 9?

3. When drawing a boxplot, outliers are drawn as individual points whenever the point is less than Q1 – 1.5 IQR or greater than Q3 + 1.5 IQR, where Q1 is the 25th percentile, q3 is the 75% percentile, and IQR is the interquartile range Q3 – Q1. If the dataset is exactly standard normal, what is the proportion of points that a point is marked as an outlier when drawing the boxplot? Use the standard normal table to obtain the proportions.

4. A Civil War historian is comparing two different musket powders (1 and 2). The average velocity of 8 musket balls fired with Powder 1 is 276.4 meters per second. The average velocity of 10 musket balls fired with Powder 2 is 280.6 meters per second. The t-statistic for the pooled variance independent 2-sample t-test is -2.497. The null hypotheses H_0 is that there is no difference between the firing velocities for powders 1 and 2. Should H0 be accepted or rejected at the 0.05 level (95% confidence)? Why? What does it mean to reject the null hypothesis in this case?

5. Here are descriptive statistics for height (x) and weight (y) of female models working at a modeling agency:

   $\bar{x}$ = 1.7 meters, s_x = 0.05 meters, $\bar{y}$ = 52 kilos, s_y = 3 kilos, r = 0.8

   Find the least squares linear equation for predicting y from x in slope-intercept form. If a model has height 1.6 meters, what is her predicted weight?

6. A regression model with 6 regression parameters (including the intercept) and the number of observations is 15. If SSE = 8.4 and SSM = 5.8, are any regressors significant at the 0.05 level?

7. Here is the normal equation and its solution for a regression problem with two independent variables:

$$\begin{pmatrix} 6 & 12 & 9 \\ 12 & 28 & 18 \\ 9 & 18 & 15 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 26.77 \\ 33.71 \\ 50.63 \end{pmatrix}, \qquad \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 2.6667 & -0.5000 & -1.0000 \\ -0.5000 & 0.2500 & 0.0000 \\ -1.0000 & 0.0000 & 0.6667 \end{pmatrix}\begin{pmatrix} 26.77 \\ 33.71 \\ 50.63 \end{pmatrix}.$$

Find the estimated regression coefficient $\hat{\beta}_2$.

8. If SSE = 0.339 and the number of observations is 6 in Problem B7, what is the standard error of $\hat{\beta}_2$?

9. Here is the hat matrix H and the y vector for the regression problem in Problem B7:

$$H = \begin{pmatrix} 0.58 & 0.33 & 0.08 & 0.25 & 0.00 & -0.25 \\ 0.33 & 0.33 & 0.33 & 0.00 & 0.00 & 0.00 \\ 0.08 & 0.33 & 0.58 & -0.25 & 0.00 & 0.25 \\ 0.25 & 0.00 & -0.25 & 0.58 & 0.33 & 0.08 \\ 0.00 & 0.00 & 0.00 & 0.33 & 0.33 & 0.33 \\ -0.25 & 0.00 & 0.25 & 0.08 & 0.33 & 0.58 \end{pmatrix}, \qquad y = \begin{pmatrix} 5.93 \\ 0.71 \\ -3.73 \\ 12.87 \\ 8.30 \\ 2.69 \end{pmatrix}.$$

Find the predicted value $\hat{y}_5$.

Part C: Short Essay Questions. 10 pts. each. For full credit write in compete sentences and paragraphs. All students answer only 2 out of 3 questions.

1. Explain to someone that is not familiar with statistical software some of the things you can do with SAS or R (pick one). What are the software's strong points and weak points?

2. What are influence points? How do they differ from outliers? What are some of the statistics that can be used to identify leverage points? How is information about influence points used to find a good regression model?

3. Assume a multiple regression model. Explain the difference between a confidence interval for $\hat{y}$ and a prediction interval for a new observation.