Take Home Exam 2

MATH 441

Due Tuesday, November 18, 2008, by the beginning of class.

Your responses to the questions in this exam must be **typed**, **numbered** and **lettered** with respect to the questions and parts of the questions below. A part of your grade will be based on the formatting and organization of your write-up. In this exam, you will perform the scientific method to test a hypothesis that you have about "the world" (i.e. something about your friends, your family, your dorm-mates, an engine, elk, something from your graduate thesis, etc ...). Your hypothesis needs to be about something that you can collect data on to try to either verify or refute your hypothesis.

1. **Make a hypothesis**. Give a hypothesis that you are interested in collecting data about. Part of your grade will be based on how innovative a hypothesis you come up with and how interesting a model you use to test this hypothesis (see questions #2 and #3).

2. **Collect Data**. Collecting data can be asking questions of other human beings, or using instruments to measure some object of interest, like using a thermometer to get the operating temperature of an internal combustion engine. **You must collect at least 20 data points!!** Your data can be categorical or quantitative. In your write-up, you must include a table of your data.

3. **Analyze the data**. Use MATLAB. All code and output must be included in your write-up.

   (a) Specify some model that describes the data you collected. You may be fitting a line (simple linear regression) or a parabola or exponentials or multiple linear regression. Write out your model as $y = Az$.

   (b) Explain why your model is appropriate. Is there some physics or intuition you have?

   (c) Find z_{ls}, a least squares solution to $y = Az$.

   (d) Interpret each of the components of the least squares solution in terms of the problem.

   (e) Is the least squares solution you found unique? Why or why not?

   (f) Is your least squares problem well conditioned? How large can the relative error $\frac{||\delta z||_2}{||z||_2}$ get? *Hint*: Recall that the residual is $r = -\delta b$ and

   $$\cos\theta = \frac{||\text{Projection of } y \text{ onto } \mathcal{R}(A)||_2}{||y||_2}.$$

   (g) Do you believe that the least squares solution you found is accurate? Why or why not?

4. **Test your hypothesis using the data**. To perform this step of the scientific method, you'll need to calculate how variable your least squares solution is, in the sense that if you collected many data sets from many different individuals, how much will the least squares solutions vary? This variability is quantified by $SE_{z_{ls}}$, the *standard error* of the least squares solution.

   (a) Calculate $||r||_2$, the Euclidean norm of the residual of the least squares solution (In statistics, the *sum squared error* is defined as $SSE = ||r||_2^2$).

   (b) Let p be the number of parameters in your model, and n be the number of data you collected (i.e. so the matrix A in the model $y = Az$ is $n \times p$). Then the Mean Squared Error is
   $$MSE = \frac{||r||_2^2}{n-p}$$
   Calculate the MSE for your model.

   (c) The covariance matrix of the least squares solution is given by the $p \times p$ matrix
   $$Var_{z_{ls}} = MSE \cdot (A^T A)^{-1}.$$
   Calculate $Var_{z_{ls}}$ for your model.

   (d) The standard error $SE_{z_{ls}}$ is the square root of the diagonal elements of $Var_{z_{ls}}$. Calculate the standard error (this is standard output from any statistics software).

   (e) It is common to report the least squares solution plus or minus two times the standard error:
   $$z_{ls} \pm 2SE_{z_{ls}}.$$
   Give this interval for each least squares estimate.

   (f) Does this range of values verify or refute your hypothesis in #1?

5. **Conclusion**. Write a conclusion about what you have found. Is further research warranted? Should a different model be investigated? Does more data or a different kind of data set need to be collected?

NOTE: A statistical linear models approach to the least squares problem is well studied. See for example Neter, Kutner, Nachtsheim and Wasserman, Applied Linear Statistical Models, 4th ed, 1996, p229 and 232.