STOR 664: FALL 2023
Final Exam, December 8–9, 2023

The exam will be posted on gradescope and available from 6:00 pm Thursday, December 8 through 11:59 pm Friday, December 9.

This is an open-book, take-home exam. Access to course materials and standard computational tools (in particular, R) is allowed; communication with other students or with anybody via the internet, other than the instructor, is not. The university Honor Code is in effect at all times.

Answers may be given in any of the following formats (including combinations of more than one): R Markdown, Word, Latex or handwritten pages scanned or photographed; if handwritten, it is recommended you use blue or black ink on plain sheets of white paper. Hadwritted script on an iPad or tablet will also be accepted if uploaded in machine-readable (e.g. pdf) format. You are requested to submit your final solutions in gradescope but if for some reason that doesn't work, you may email tem to the instructor.

Each question will be graded out of 50 points for a total score of 100 for the whole exam. An error in one part-question will not prevent you gaining full credit in other parts of the same question even if you carry over the error. You may answer the questions (or parts within a question) in any order; answers out of sequence will not be penalized so long as you make clear which part of your answer refers to which part of the question sheet.

There is no official time limit for the exam but it is strongly recommended that you self-limit to 6 hours total; continuing to work beyond that time is unlikely to improve your grade. Any queries about the exam may be addressed directly to the instructor by email, text message or cellphone.

1. **Three-dimensional response surface design**. Consider an experiment with three variables $X_1,\ X_2,\ X_3$ laid out as follows:

```
X1: + + + + + + + + + 0 0 0 0 0 0 0 0 0 - - - - - - - - -
X2: + + + 0 0 0 - - - + + + 0 0 0 - - - + + + 0 0 0 - - -
X3: + 0 - + 0 - + 0 - + 0 - + 0 - + 0 - + 0 - + 0 - + 0 -
```

   Here, the symbols +, 0 and - represent the numerical values $1,\ 0,\ -1$ and the order is intended to represent the actual order of covariates in the 27 observations in the experiment. (For example, the variable X_1 is represented through observations $x_{1,1} = x_{2,1} = \ldots = x_{9,1}, x_{10,1} = \ldots = x_{18,1} = 0, x_{19,1} = \ldots = x_{27,1} = -1$ with corresponing notation for $x_{i,2}$ and $x_{i,3},\ i = \ldots, 27$.) Also assume that each of the 27 responses is given by a formula of the form

$$y_i \quad = \quad \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,1}^2 + \beta_5 x_{i,2}^2 + \beta_6 x_{i,3}^2 + \epsilon_i$$

   where $x_{i,1},\ x_{i,2},\ x_{i,3}$ represent the values of $X_1,\ X_2,\ X_3$ in the ith row of data and $\{\epsilon_i,\ i = 1, \ldots, 27$ represent independent normally distributed errors with mean 0 and comon variance σ^2.

   (a) Write down the X matrix for this regression problem and calculate $X^T X$ and $(X^T X)^{-1}$. (*Note: Since this is an open-book computer-aided exam, you are allowed and encouraged to use R, matlab or other appropriate software to complete the calculations, but your final answer should be an explicit statement of the result.*) **[10 points]**

(b) Define:

$$\begin{aligned} T_0 &= \sum y_i, \\ T_1 &= \sum y_i x_{i,1}, \\ T_2 &= \sum y_i x_{i,2}, \\ T_3 &= \sum y_i x_{i,3}, \\ T_4 &= \sum y_i x_{i,1}^2, \\ T_5 &= \sum y_i x_{i,2}^2, \\ T_6 &= \sum y_i x_{i,3}^2 \end{aligned}$$

Show that

$$\hat{\beta}_0 = \frac{7}{27} T_0 - \frac{1}{9}(T_4 + T_5 + T_6)$$

and derive similar expressions for $\hat{\beta}_j,\ j = 1, \ldots, 6$. What are the variances of these estimators? **[10 points]**

(c) Now suppose the objective is to find the values $X_1^*,\ X_2^*,\ X_3^*$ (not necessarily integers) to maximize the expected response $\beta_0 + \beta_1 X_1^* + \beta_2 X_2^* + \beta_3 X_3^* + \beta_4 X_1^{*2} + \beta_5 X_2^{*2} + \beta_6 X_3^{*2}$. Assume $\beta_4,\ \beta_5,\ \beta_6$ are all < 0 so that the surface has a finite maximum. Derive theoretical expressions for $X_1^*,\ X_2^*,\ X_3^*$ and show how to calculate estimators $\widehat{X_1^*},\ \widehat{X_2^*},\ \widehat{X_3^*}$ in terms of the estimators calculated in part (b). **[5 points]**

(d) Are the estimators $\widehat{X_1^*},\ \widehat{X_2^*},\ \widehat{X_3^*}$ independent? — explain why or why not. Using whatever approximations you consider appropriate, state formulas for the standard errors $\widehat{X_1^*},\ \widehat{X_2^*},\ \widehat{X_3^*}$ in terms of the quantities $\hat{\beta}_0, \ldots, \hat{\beta}_6$ and s^2. (Assume s^2 is the usual unbiased sample estimator of σ^2; you don't need to give a separate formula for that.) **[10 points]**

(e) Suppose you want to find a 99% confidence interval for X_1^*. Show how to do this using (i) delta method, (ii) Fieller method. In particular, show that the endpoints of the Fieller confidence set are the solutions of the quadratic equation

$$(4\hat{\beta}_4^2 - 5.397305 s^2)x^2 + 4\hat{\beta}_1\hat{\beta}_4 x + \hat{\beta}_1^2 - 0.4497754 s^2 = 0,$$

and state the conditions under which the confidence set is an interval.

(It is acceptable that you may get slightly different values for the numerical quantities 5.397305 and 0.4497754, but you should show where they come from. Since very similar formulas will apply for X_2^* and X_3^*, you are not asked to find those.) **[15 points]**

Question 2 on the next page.

2. **Computational Exercise (This part is expected to be completed using R)**. You may use any R packages but your final answer should include your R code including any packages that you use.

   Consider the `diabetes` dataset in Faraway's package. This may be loaded into R directly through `library(faraway)` followed by `data(diabetes)`. The dataset includes a response variable `glyhb` (Glycosolated Hemoglobin) and numerous predictors. Glycosolated hemoglobin greater than 7.0 is usually taken as a positive diagnosis of diabetes.

   Remove the variables `id` (not relevant for predictions), `bp.2s` and `bp.2d` (too many missing datapoints) and use `na.omit` to reduce the dataset to one where all observations are complete. (These operations are considered part of the initial data manipulation and not awarded formal credit. You should end up with a matrix with 366 observations and 16 variables including `glyhb`.)

   (a) Fit a linear regression model for `glyhb` on the other 15 variables. Use stepwise regression to reduce the number of variables and comment briefly on your results. [**5 points**]

   (b) Would you consider any transformation of `glynb`? Use plots and any numerical diagnostics you consider appropriate, choose a suitable transformation, and repeat the analysis of part (a). (For the rest of this question, you should use the transformed values, unless your decision is to stick with the original values of `glynb`.) [**5 points**]

   The next few parts should be based on your final model from parts (a) and (b).

   (c) Calculate the studentized residuals and comment on whether there appear to be outliers, justifying your answer with appropriate tests or plots. [**5 points**]

   (d) Plot the residuals against each of the covariate and the fitted values; hence comment on whether there is an evidence that the model dies not fit the data. [**5 points**]

   (e) Do the residuals appear to follow a normal distribution? Briefly justify your answer with suitable tests or plots. [**5 points**]

   (f) Calculate appropriate diagnostics for leverage and influence, and comment on the results. [**5 points**]

   (g) One observation appears to have very large leverage and another seems to have large influence. Identify these variables and repeat the main parts of steps (c) through (f) without those observations; do any of your conclusions substantially change? [**5 points**]

   (h) Returning to the original dataset with 366 observations and 16 variables, but still using any transformation you decided to use on `glynb`, split the data into a training dataset and a test dataset where the test dataset consists of every tenth observations (i.e. rows 10, 20, ..., 360) and the training dataset is everything else. Based on the training dataset, repeat your model selection from part (b) and find alternative models using (i) principal components regression, (ii) PLS regression, (iii) ridge regression and (iv) LASSO. You may use any appropriate package(s) but please be explicit about your method of calculation.

   Then, for all five methods, use the best model you fitted to predict the values of `glynb` on the test dataset. Use root mean squared error (on the *original* scale — this way we can compare estimators using different transformations) to decide which of the five methods is best on this example. Summarize your conclusions. [**15 points**]