PH.D. QUALIFYING EXAMINATION-APPLIED STATISTICS, STA 682

Time: 1:00 - 4:00 PM, August 26, 2022

General Instructions

- There are 3 problems, each with subparts, in this exam. Problem #3 consists of 16 true-or-false questions. This exam has a possible total of 40 points. You are to answer all.

- Sign and print your name on this cover page only.

- Write on one side only. Begin each subpart in Problems #1 & #2 on a new sheet with the problem number noted. You must show all your work and justifications correctly and completely to receive full credits. Partial credits may be given for partially correct solutions.

- For each problem/subproblem, hand in only the answer that you want to be graded. If necessary, please make clear, e.g., by crossing out the other answer(s), which answer should be graded. Crossed-out work will be ignored. Failure to follow this instruction for a problem will result in a zero score for that problem.

- Answer T (for true) or F (for false) in the underlined spaces provided for Problem #3.

- When finished, please collate all pages according to the problem numbers and then number the pages accordingly. Your answers to Problem #3 will be the last page.

- If a theorem is applied, you must clearly state the theorem, identify its assumption(s) and conclusion(s), and explain why it is applicable. New notations must be defined before use.

By signing below, I hereby acknowledge that I have completely read and fully understand the instructions.

____

Signature

____

Printed Name

____

Date

1. Consider the following linear model with three explanatory variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \beta_3 x_{i3} + \varepsilon_i,\ i = 1, \cdots, n,\ n \geq 4.$$

Assume that $\epsilon = (\varepsilon_1, \ldots, \varepsilon_n)' \sim N(\,0_{n\times 1}, \sigma^2 I_n)$, where $0_{n\times 1}$ is a column vector of zeros of size n and I_n is the identity matrix of dimension n. Assume that the design matrix is of full rank. Derivations and answers may be in matrix form. Please take note of the difference between a statistic and a parameter.

1.a) Express the linear regression model in matrix form. Define your response, design, and parameter matrices or vectors clearly. (1.5 points)

1.b) Derive the least squares estimator of the parameter vector and the distribution of the estimator. (5 points)

1.c) Find an unbiased estimator of the model variance σ^2 and show that it is unbiased. (4 points)

1.d) Test the null hypothesis H_0: $\beta_1 = 0$ versus the alternative H_a: $\beta_1 \neq 0$.

1.d.i) Express the null hypothesis in the form of a general linear hypothesis. (0.5 points)

1.d.ii) Define the t-test statistic (T_d) and state the decision rule for the test at the significance level of α. Specify the degree(s) of freedom. (2 points)

1.d.iii) Define the t-distribution (or random variable) with degrees of freedom v in terms of the standard normal and chi-squared distributions (or random variables). (2 points)

1.d.iv) Show your test statistic in (1.d.ii) has a t-distribution under the assumption that the null hypothesis is true. (5 points)

1.e) Test the null hypothesis H_0: $\beta_1 = 0$ versus the alternative H_a: $\beta_1 \neq 0$.

1.e.i) Define the F-test statistic (F_e) and state the decision rule for the test at the significance level of α. Specify the degree(s) of freedom. (2 points)

1.e.ii) How are the chi-squared and *F* distributions related? That is, define the F distribution (or random variable) with degrees of freedom v_1 and v_2 in terms of chi-squared distributions (or random variables). (2 points)

1.e.iii) Show that the test statistic in (1.e.i) has an F distribution under the assumption that the null hypothesis is true. (4 points)

1.f) Show that $(T_d)^2 = F_e$. (1 point)

2. Let $y_i = x_i'\beta + \varepsilon_i$ with $i = 1,2,\cdots n,$ be a linear regression model where $E(\varepsilon_i) = 0$, var$(\varepsilon_i) = \sigma^2$ and cov$(\varepsilon_i, \varepsilon_i)$ =0 when $i \neq j$. Let X denote the design matrix, h_{ii} be the diagonal elements of the hat matrix $H = X(X'X)^{-1}X',$ b be the least squares estimate of β, and $e_i's$ be the resulting residuals. Let $b_{(i)}$ be the least squares estimate of β obtained after deleting the i-th case/sample point.

2.a) Show that DFBETA$_i$ = $b - b_{(i)} = \frac{(X'X)^{-1}x_i e_i}{1-h_{ii}}$. (4 points)
Hint: $(P + UV)^{-1} = P^{-1} - P^{-1}U(I + VP^{-1}U)^{-1}VP^{-1}$

2.b) Define DFFIT$_i$, the change in the predicted value for the i-th case obtained when that case is left out of the regression, in terms of e_i and h_{ii}. (1 point)

2.c) (2 points) Let $\hat{y}_{i,(i)}$ be the predicted value for the i-th case resulting from fitting the least squares without the i-th case. To evaluate the performance of a model on a data set, a leave-one-out cross-validation criterion is defined to be $CV_n = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,(i)})^2$. Show that

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2.$$

3. Answer the following questions in the context of the multiple linear regression. Write T (for true) or F (for false) in the underlined spaces provided. (0.25 points each)

____ 3.1) The coefficient of determination is the square of the sample correlation between the response values and one of the predictor values.

____ 3.2) The coefficient of determination is the square of the sample correlation between the observed response values and the fitted values.

____ 3.3) In practice, adding more predictors into a linear model decreases the R squared value.

____ 3.4) The coefficient of determination measures the proportion of variation in the response that can be explained by the set of predictors.

____ 3.5) In the case of perfect collinearity, the unique least squares estimate exists.

____ 3.6) Perfect collinearity or multicollinear occurs when one of the independent variables is a perfect linear combination of the other independent variables.

____ 3.7) When Sex (male or female) is one of the predictors, incorporating a constant term, a female indicator variable, and a male indicator variable into the model will not lead to a collinearity problem.

____ 3.8) The inverse of the matrix (X'X), where X is the design matrix, doesn't exist when there is perfect collinearity.

____ 3.9) The residual sum of squares for the centered model is the same as the one for the uncentered model.

____ 3.10) The coefficients associated with independent variables from centered and non-centered data are not identical.

____ 3.11) Changing the scale of an independent variable will lead to a change in the scale of the corresponding coefficient, but no change in the significance when testing whether the coefficient associated with the variable is significant different from zero.

____ 3.12) Changing the scale of an independent variable will not affect the fitted values/model.

____ 3.13) A leverage describes how far away an individual data point is from the centroid of all data points in the space of independent variables.

____ 3.14) DBETA's and DFFITS's are two measures for identifying influential data points.

____ 3.15) The statement that "*the plot* of *residuals versus fitted values plot is expected to show a fairly random pattern when the linear fit is reasonable"* is only true when the model error is normally distributed.

___ 3.16) A data point can be an outlier without being influential.