UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Advanced Exam Version I - Linear Models
Thursday, August 20, 2020

Work all problems. Seventy points are required to pass.

1. The linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{y}$ is a vector of length n, is said to be saturated if the error degrees of freedom $(n - rank(\mathbf{X}))$ is equal to zero. Assume the errors are iid and mean zero.
   (a) Define BLUE. (Please do not just use four words.)
   (b) Give an example of an estimator of $\boldsymbol{\beta}$ that is not BLUE.
   (c) Show that in a saturated model, every linear unbiased estimator is the corresponding BLUE.

2. What affects the selling price of a house? Table below shows analysis results based on 100 recent home sales in Gainesville, Florida.

```
> summary(lm(price ~ size + new + taxes))
              Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   -21.3538     13.3115   -1.604   0.11196
size            0.0617      0.0125    4.937   3.35e-06
new            46.3737     16.4590    2.818   0.00588
taxes           0.0372      0.0067    5.528   2.78e-07
---
Residual standard error: 47.17 on 96 degrees of freedom
Multiple R-squared: 0.7896,   Adjusted R-squared: 0.783
F-statistic: 120.1 on 3 and 96 DF, p-value: < 2.2e-16
> anova(lm(price ~ size + new + taxes)) # sequential SS, size first
Analysis of Variance Table
Response: price
          Df  Sum Sq  Mean Sq  F value     Pr(>F)

size       1  705729   705729  317.165  < 2.2e-16
new        1   27814    27814   12.500  0.0006283
taxes      1   67995    67995   30.558  2.782e-07
Residuals 96  213611     2225
```

   (a) Report and interpret results of the global test of the hypothesis that none of the explanatory variables has an effect.
   (b) Report and interpret significance tests for the individual partial effects, adjusting for the other variables in the model.

(c) What is the conceptual difference between the tests of the effects of the independent variables in the coefficients table and in the ANOVA table?

3. Given independent random samples of sizes n_1 and n_2, the goal of this question is to inferentially compare μ_1 and μ_2 from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ populations.

   (a) Put the analysis in a normal linear model context, showing a model matrix and explaining how to interpret the model parameters.

   (b) Find the projection matrix for the model space, and find SSR (regression sum of squares) and SSE (error sum of squares).

   (c) Construct an F test statistic for testing $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$. Specify a null distribution for this statistic.

   (d) Relate the F test statistics in (c) to the t statistic for this test,

   $$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ with } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

   where $\bar{y}_i$ is the sample mean of the ith sample, and s_i^2 is sample variance of the ith sample.

   (e) Suppose the two independent samples had different variances. Could you use the F and t tests you used in parts (c) and (d)? Why or why not?

4. Suppose that $\mathbf{y}$ is $N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ and that X is an $n \times p$ matrix of constants with rank $p < n$.

   (a) Show that $\mathbf{H} = X(X'X)^{-1}X'$ and $I - H$ are idempotent, and find the rank of each.

   (b) Assuming $\boldsymbol{\mu}$ is a linear combination of the columns of X, that is, $\boldsymbol{\mu} = X\mathbf{b}$ for some $\mathbf{b}$, find $E(\mathbf{y}'\mathbf{H}\mathbf{y})$ and $E(\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y})$.

   (c) Find the distributions of $\mathbf{y}'\mathbf{H}\mathbf{y}/\sigma^2$ and $\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/\sigma^2$.

   (d) Show that $\mathbf{y}'\mathbf{H}\mathbf{y}$ and $\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$ are independent.

   (e) Find the distribution of

   $$\frac{\mathbf{y}'\mathbf{H}\mathbf{y}/p}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/(n-p)}.$$