

Final Exam Stat 505 Fall 2011

December 14, 2011  
100 pts total

Name: Key

1. State the Gauss-Markov theorem for linear model: (10 pts)

$$y = X\beta + \epsilon; E(\epsilon) = 0; \text{Var}(\epsilon) = \sigma^2 V$$

The best (minimum variance) unbiased linear estimator of any estimable  $\lambda^T \beta$  is  $\lambda^T \hat{\beta}$  where  $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$

2. Why is it so much easier to make causal inference in a study where treatments are randomly allocated to the units than in a study where the treatment variable is simply observed? (10 pts)

When treatments are randomly allocated, the treatment and control groups act as counterfactuals to each other. Let  $\bar{y}^1$  be mean of treated group,  $\bar{y}^0$  mean of controls. Then  $\bar{y}^1 - \bar{y}^0$  is an unbiased estimator of the mean difference in counterfactuals,  $E(\bar{y}^1 - \bar{y}^0)$ , which is the real quantity of interest.

- (b) For each school we have three available predictors:  
 size = number of students enrolled, sector (public or parochial), and  
 meanSES (average socio-economic status of all students).

- i. Add these into the model in part (a) and rewrite any distributions from above which have changed. (7 pts)

$y_i$ : description is the same.

$$\alpha_j \sim N(\gamma_0 + \gamma_1 \text{size}_j + \gamma_2 I_{\text{public } j} + \gamma_3 \text{Avg SES}_j, \sigma_a^2)$$

$j=1, \dots, 160$

- ii. What parameter(s) is(are) expected to change with the addition of these group level predictors? In what direction? (3 pts)

$\sigma_a^2$  should get smaller because the 3 school level predictors should explain some school-to-school variation.

- (c) Sample sizes vary from school to school. Results are shown for predicted intercepts from the above model for the schools with the largest and the smallest sample sizes.

School	Sample size	Estimate	SE
2305	67	14.5	0.64
8367	14	9.6	1.01
overall	7185	12.4	0.36

Explain what partial pooling does in the context of this situation. (12 pts)

$\hat{\alpha}_j$  is a weighted average of the non-pooled estimate for school  $j$  and the overall intercept from pooling all schools together. Weights are precisions = inverse variances of the nonpooled & pooled estimates. School 8367 has small sample size & small precision, so its 9.6 was more strongly influenced by the pooled estimate than school 2305 and its SE is larger.

- (c) Under what assumptions can we estimate a causal effect of the frequency of watching the TV show on the math score? (5 pts)

We could use "asking" as an Instrumental Variable on frequency, because asking was randomized and is therefore ignorable. Then we need to assume a strong connection between "asking" and frequency and <sup>assume</sup> a monotonicity of the asking  $\rightarrow$  frequency relationship. Also assume exclusion restriction & with all that baggage we should get to the point of calling a relationship between frequency & math scores causal.

6. Suppose that we make some changes in Stat 216 in Spring 2012 which are intended to help students get a better grade in the course (measured as overall percentage). There is no direct control group because we apply the "treatment" to all students in Stat 216 Spring 2012. However we have the following information about each student who took the course in the last two years:  
history of math courses, ACT, SAT, GPA, major, prior attempts at Stat 216, and number of credits toward BS degree.

- (a) Explain how to use these scores to artificially create a "control" group of individuals similar to the treatment group. (5 pts)

Build a logistic regression on response

$$\begin{cases} 1 & \text{in Stat 216 Sp 2012} \\ 0 & \text{in " " prior semester} \end{cases}$$

using all pre-treat Stat 216 variables listed.

Compute  $\hat{p}_i$  for each individual & call it a propensity score. For each student in Stat 216 Sp 2012 match one prior Stat 216 "control" student with closest  $\hat{p}$

- (b) How would you use the two groups to estimate the treatment effect? (5 pts)

Build a regression on overall % using all pre-treatment variables & an indicator for Stat 216 in Sp 2012. The coefficient on this indicator is our estimate of treatment effect.

[I'm not advocating this as I don't trust it]