

Political Science 150B/350B  
Winter 2007  
Midterm Examination

This is a closed book, in-class examination. You may use a calculator. Attempt all questions. Show working for partial credit. The total number of points appears at the end of the exam.

**Question 1:** To show that the least squares estimator is BLUE, we require the assumption  $E(\epsilon|\mathbf{X}) = \mathbf{0}$ .

- (a): (4 points) Explain what this assumption means in words that could be understood by a colleague who has not taken a statistics class.
- (b): (4 points) What important property does the least squares estimator *not* possess if this assumption does *not* hold? Again, your answer should make sense to a colleague who has not taken a statistics class.
- (c): (4 points) How can this assumption be violated when working with data? Be specific.

**Question 2:** (4 points) Choose the statement that best completes the following proposition:

Given the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with  $\mathbf{X}$  full rank, then the assumption

$$\boldsymbol{\epsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Psi}), \boldsymbol{\Psi} \neq \mathbf{I}$$

- (a): implies that the least squares estimator of  $\boldsymbol{\beta}$  has smallest sampling variance in the class of linear unbiased estimators
- (b): implies that the least squares estimator of  $\boldsymbol{\beta}$  is unbiased, but no longer has smallest sampling variance in the class of linear unbiased estimator
- (c): implies that the least squares estimate of  $\boldsymbol{\beta}$  is unbiased
- (d): implies that the least squares estimator of  $\boldsymbol{\beta}$  suffers from omitted variable bias

**Question 3:** Let  $U_i$  be a categorical variable (a “factor”) with levels indexed by  $j \in \{1, \dots, M\}$ . Consider the regression

$$E(Y_i|U_i) = \sum_{j=1}^M \beta_j D_{ij}$$

where  $D_{ij} = 1$  if  $U_i = j$  and 0 otherwise.

- (a): (3 points) Interpret the  $\beta_j$  parameters.
- (b): (3 points) Interpret the  $r^2$  of this model.
- (c): (3 points) As written, this model does not have an intercept term. Why not?

**Question 4:** A researcher runs an experiment, designed to assess the conditions under which people will engage in various acts of political participation. The researcher randomly assigning a large number of observations to a treatment group (the treatment consists of a face-to-face, in-house visit from a member of the research team, during which a standardized speech about the importance of political participation is delivered), and a control group. Randomization is successful, in the sense that (a) all units randomly assigned to treatment are treated; (b) all units randomly assigned to the control group are not treated; (c) treated and control cases do not communicate with each other. Assume that after the field work, the researcher is able to accurately measure the political participation of both treatment and control subjects, and this measure constitutes  $y$ , the researcher’s dependent variable.

- (a): (5 points) The researcher asks your for your help in coming up with an unbiased estimate of the effect of the treatment. The researcher is concerned that aside from the treatment, other factors make people more or less likely to be politically active (e.g., socio-economic status, education level, ideological extremism). What statistical procedure do you recommend to the researcher and why?
- (b): (3 points) The researcher is particularly concerned that some members of her research team are better suited to the task of delivering the treatment than others (i.e., the researcher recruited a dozen of her political science PhD students to do the field work for this experiment). Does this fact alter the advice you gave above in answering the previous question?

**Question 5:** (3 points) Which statement best completes the following sentence: “If we add a predictor to a regression model then...”

- (a): the  $r^2$  will decrease
- (b): the  $r^2$  will decrease if the new predictor is negatively correlated with the predictors already in the model
- (c): the  $r^2$  will not decrease
- (d): the  $r^2$  will increase

**Question 6:** True or false, and why?

- (a): (4 points) An important assumption of the regression model is  $E(\epsilon|\mathbf{X}) = \mathbf{0}$ . One way to test if this assumption is met in practice is to check if the estimated residuals have zero mean.
- (b): (4 points) The least squares estimator of  $\beta$  is a random variable.

**Question 7:** Using the notation introduced above, state the dimension and the substantive interpretation of the following quantities:

- (a): (3 points)  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- (b): (4 points)  $\mathbf{H}\mathbf{y}$ , where  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ .
- (c): (5 points)  $(n - k)^{-1}\mathbf{y}'\mathbf{M}\mathbf{y}(\mathbf{X}'\mathbf{X})^{-1}$ , where  $\mathbf{M} = \mathbf{I} - \mathbf{H}$ .

**Question 8:** A researcher has cross-national data on the number of deaths in prison in a country as a proportion of all deaths in a country in a given time period ( $y$ ) and a predictor,  $X$  (GDP per capita of each country, measured contemporaneously with  $y$ ), and the data can be classified into two mutually exclusive and exhaustive groups, democracies and non-democracies. For simplicity, suppose there are  $n$  observations in each group, for a total of  $2n$  observations.

- (a): (3 points) Specify a linear regression model that will let the researcher test the null hypothesis that  $y$  varies across democracies and non-democracies. Carefully define all terms and/or the null hypothesis, test-statistic etc.
- (b): (5 points) Carefully specify how you would test whether the linear relationship between GDP per capita and  $y$  is the same across democracies and non-democracies. Again, carefully define all variables in the regression models you specify and/or the null hypothesis, test-statistic etc.
- (c): (3 points) What assumption is being made when the researcher “pools” the data from democracies and non-democracies into one regression model, as in the previous two questions?

**(d):** (6 points) Suppose the researcher find that the relationship between GDP per capita and deaths from prisons is the same across democracies and non-democracies. The researcher decides to estimate two separate regressions, one for democracies and one for non-democracies. Will the parameter estimates from these two separate regressions be the same as the estimates from a single pooled regression in which interaction terms are used to estimate intercepts and slopes for the two groups in the data? If so, then what advantages are there to pooling the data into one regression? Be specific in your answer.

**Question 9:** Economists have conjectured that cigarette smokers may be penalized with lower wages in the job market: for instance, smokers may receive lower levels of pay because they are less productive (if the act of smoking takes a worker away from his or her job); because they may be absent from work more often (due to their greater risk of respiratory infections); because it is more expensive to provide them with health insurance; or simply because firms discriminate against them. A regression analysis of log wages on a dummy variable for smoking and other predictors of wages is summarized in the following table (standard errors in parentheses):<sup>1</sup>

	Model 1	Model 2	Model 3
Smoking	-.176 (.021)	-.080 (.021)	-.069 (.019)
Education		.070 (.004)	.045 (.005)
Other factors included	no	no	yes

**(a):** (8 points) In two or three sentences, explain why the estimated impact of smoking on wages decreases in magnitude across the three models.

**(b):** (5 points) Who incurs the higher wage penalty for smoking in absolute terms, relatively higher-paid workers or relatively lower-paid workers?

<sup>1</sup>See Philip B. Levine, Tara A. Gustafson and Ann D. Velenchik, “More Bad News for Smokers? The Effects of Cigarette Smoking on Wages,” *Industrial and Labor Relations Review*, Vol. 50, No. 3 (April 1997), pp. 493-509; my source is Orley Ashenfelter, Phillip B. Levine and David J. Zimmerman (2003), *Statistics and Econometrics: Methods and Applications*, Wiley, New York, p190.

**Question 10:** Consider the following set of regression estimates, obtained from a random sample of 200 observations:

<i>Variable</i>	<i>Parameter</i>	<i>Estimate</i>	<i>t-value</i>
Intercept	$\beta_0$	324	12.2
$X_1$	$\beta_1$	8.8	3.1
$X_2$	$\beta_2$	.5	6.3
$X_3$	$\beta_3$	.7	4.2
$X_4$	$\beta_4$	1.2	.5
$D$	$\beta_5$	-506.3	-3.8
$D \times X_3$	$\beta_6$	.65	6.9
$r^2$		.69	

The independent variables have the following ranges:

- $X_1$  ranges between 0 and 10
- $X_2$  ranges between 1000 and 8000
- $X_3$  ranges between 3000 and 10,000
- $X_4$  ranges between 1000 and 8000
- $D$  is a binary indicator, taking the value zero for all observations in group A and 1 for all observations in group B.

Assume that the population and sample distributions of the  $X$  variables are approximately uniform over the ranges used in the estimation, and do not differ between groups A and B.

- (a):** (4 points) Test  $H_0 : \beta_1 = 1$  against the one-sided alternative  $H_A : \beta_1 > 1$ .
- (b):** (5 points) Discuss the relative importance of  $X_1$  and  $X_2$  for explaining variation in  $y$ .
- (c):** (5 points) Does the model predict higher or lower values of  $y$  for members of group B relative to members of group A?

**END OF EXAM**

Total Number of Points: 100

df	One-Tailed Significance Level								
	0.001	0.005	0.01	0.025	0.05	0.1	0.15	0.2	0.25
	Two-Tailed Significance Level								
	0.002	0.010	0.02	0.05	0.1	0.2	0.3	0.4	0.5
1	318.309	63.657	31.821	12.706	6.314	3.078	1.963	1.376	1.000
2	22.327	9.925	6.965	4.303	2.920	1.886	1.386	1.061	0.816
3	10.215	5.841	4.541	3.182	2.353	1.638	1.250	0.978	0.765
4	7.173	4.604	3.747	2.776	2.132	1.533	1.190	0.941	0.741
5	5.893	4.032	3.365	2.571	2.015	1.476	1.156	0.920	0.727
6	5.208	3.707	3.143	2.447	1.943	1.440	1.134	0.906	0.718
7	4.785	3.499	2.998	2.365	1.895	1.415	1.119	0.896	0.711
8	4.501	3.355	2.896	2.306	1.860	1.397	1.108	0.889	0.706
9	4.297	3.250	2.821	2.262	1.833	1.383	1.100	0.883	0.703
10	4.144	3.169	2.764	2.228	1.812	1.372	1.093	0.879	0.700
11	4.025	3.106	2.718	2.201	1.796	1.363	1.088	0.876	0.697
12	3.930	3.055	2.681	2.179	1.782	1.356	1.083	0.873	0.695
13	3.852	3.012	2.650	2.160	1.771	1.350	1.079	0.870	0.694
14	3.787	2.977	2.624	2.145	1.761	1.345	1.076	0.868	0.692
15	3.733	2.947	2.602	2.131	1.753	1.341	1.074	0.866	0.691
16	3.686	2.921	2.583	2.120	1.746	1.337	1.071	0.865	0.690
17	3.646	2.898	2.567	2.110	1.740	1.333	1.069	0.863	0.689
18	3.610	2.878	2.552	2.101	1.734	1.330	1.067	0.862	0.688
19	3.579	2.861	2.539	2.093	1.729	1.328	1.066	0.861	0.688
20	3.552	2.845	2.528	2.086	1.725	1.325	1.064	0.860	0.687
21	3.527	2.831	2.518	2.080	1.721	1.323	1.063	0.859	0.686
22	3.505	2.819	2.508	2.074	1.717	1.321	1.061	0.858	0.686
23	3.485	2.807	2.500	2.069	1.714	1.319	1.060	0.858	0.685
24	3.467	2.797	2.492	2.064	1.711	1.318	1.059	0.857	0.685
25	3.450	2.787	2.485	2.060	1.708	1.316	1.058	0.856	0.684
26	3.435	2.779	2.479	2.056	1.706	1.315	1.058	0.856	0.684
27	3.421	2.771	2.473	2.052	1.703	1.314	1.057	0.855	0.684
28	3.408	2.763	2.467	2.048	1.701	1.313	1.056	0.855	0.683
29	3.396	2.756	2.462	2.045	1.699	1.311	1.055	0.854	0.683
30	3.385	2.750	2.457	2.042	1.697	1.310	1.055	0.854	0.683
50	3.261	2.678	2.403	2.009	1.676	1.299	1.047	0.849	0.679
100	3.174	2.626	2.364	1.984	1.660	1.290	1.042	0.845	0.677
200	3.131	2.601	2.345	1.972	1.653	1.286	1.039	0.843	0.676
500	3.107	2.586	2.334	1.965	1.648	1.283	1.038	0.842	0.675
1000	3.098	2.581	2.330	1.962	1.646	1.282	1.037	0.842	0.675
3000	3.093	2.577	2.328	1.961	1.645	1.282	1.037	0.842	0.675
10000	3.091	2.576	2.327	1.960	1.645	1.282	1.036	0.842	0.675
$\infty$	3.090	2.576	2.326	1.960	1.645	1.282	1.036	0.842	0.674

Table 1: Critical values of the  $t$  distribution.