Political Science 150B/350B
Winter 2008
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: Consider the following statistical model:

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

where $\mathbf{y}$ is a n by 1 vector of observations on a dependent variable, $\mathbf{X}$ is a n by k matrix of predictors, $\boldsymbol{\varepsilon}$ is a n by 1 vector of unobserved stochastic disturbances, and $\boldsymbol{\beta}$ is a k by 1 vector of parameters to be estimated.

Q1.1 (5 points) What is the least squares estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$?

Q1.2 (5 points) Under what conditions is the least squares estimator **undefined**?

Q1.3 (5 points) Under what conditions is the least squares estimator **unbiased**?

Q1.4 (5 points) In addition to the assumptions required to show that $\hat{\boldsymbol{\beta}}$ is unbiased, what else must be assumed in order to establish that $\hat{\boldsymbol{\beta}}$ is **BLUE**?

Question 2: A colleague seeks your advice on a methodological matter. She wants to estimate the effect of X on Y, but worries about the consequences of not being able to measure Z, another variable that is thought to be a determinant of Y.

- (3 points) Under what conditions will your colleague's estimate of the effect of X on Y be unaffected by the availability (or unavailability) of Z?
- (3 points) Suppose your colleague did have data on Z. Is there any reason *not* to include Z in the regression analysis?

Question 3: Political scientists use statistical models to provide quantitative characterizations of different electoral systems. Electoral systems take vote shares for a party in election t, $V_t \in [0,1]$ and generate seat shares, $S_t \in [0,1]$. In particular, interest centers on two features of an electoral system:

(a): *unbiasedness* or the property that $E(S_t|V_t = .5) = .5$

(b): *responsiveness*: how efficiently do changes in V_t generate changes in S_t?

The following statistical model is often used to estimate these features of electoral systems:

$$\ln\left(\frac{S_t}{1 - S_t}\right) = \alpha + \beta \ln\left(\frac{V_t}{1 - V_t}\right) + \varepsilon_t$$

Q3.1 (4 points) Interpret α.

Q3.2 (4 points) In two-party systems it was conjectured that "if the votes are divided in the ratio A:B then the seats will be divided in the ratio A^3:B^3, and the winning majority of the votes will be magnified in the proportion of seats won" (M.G. Kendall and A. Stuart, "The Law of Cubic Proportion in Election Results", *British Journal of Sociology*, 1 (1950), 183--97). An earlier statement of this so-called "cube-law" was made in 1898 by one of the founders of modern statistics, F. Y. Edgeworth. The cube law thus implies $\alpha = 0$ and $\beta = 3$.

How would you test the cube law given data on S and V?

Question 4: On November 24, 2007, elections were held in the 150 electoral divisions that comprise the Australian House of Representatives. Voting in Australia is via the alternative vote, in which a valid vote requires a complete numerical rank ordering of candidates on the ballot. Rates of invalid voting (InformalPercent) vary considerably across the 150 House seats. We also have data on the following variables, thought to be good predictors of rates of invalid voting:

- nes: the percentage of households in an electoral division where English is not spoken.
- BallotLength: the number of candidates on the ballot
- OpPref: a binary indicator variable, coded 1 (or TRUE) if the division is in a state which does not require a full enumeration of preferences when voting in state-level legislative elections (so-called optional preferential voting), and 0 otherwise.

The following table presents summary statistics for these variables:

InformalPercent		nes		BallotLength		OpPref	
Min. :	1.910	Min. :	1.60	Min. :	4.000	Mode :	logical
1st Qu.:	3.060	1st Qu.:	4.00	1st Qu.:	6.000	FALSE:	72
Median :	3.820	Median :	10.85	Median :	7.000	TRUE :	78
Mean :	3.947	Mean :	15.42	Mean :	7.027		
3rd Qu.:	4.397	3rd Qu.:	21.57	3rd Qu.:	8.000		
Max. :	9.490	Max. :	63.00	Max. :	13.000		

Least squares regression analysis produces the following results:

	Estimate	Std. Error
Intercept	1.61	0.33
nes	-0.026	0.014
nes^2	0.0016	0.00027
BallotLength	0.23	0.04
OpPref	0.81	0.13
r^2	.64	
$\hat{\sigma}$	.78	

n.b., both nes and the square of nes are included as predictors.

Q4.1 (3 points) In a sentence or two, discuss how well the model appears to fit the data.

Q4.2 (3 points) Interpret the estimated coefficient on BallotLength.

Q4.3 (3 points) Calculate and report a 95% confidence interval around the estimated coefficient for OpPref.

Q4.4 (6 points) What do the parameter estimates imply about the relationship between the preponderence of non-English speaking households and rates of invalid balloting?

Q4.5 (5 points) Consider the seat of Bennelong (formerly held by the Australian Prime Minister, John Howard). This division had nes = 36.3, BallotLength = 13, and OpPref = 1. The actual rate of invalid balloting in Bennelong was 6.22%. How well does the model predict that outcome in this seat?

Question 5: An important assumption for the optimality of the least squares regression estimator is that the disturbances be "iid".

Q5.1 (5 points) Give a formal definition of "iid".

Q5.2 (6 points) List two ways in which the "iid" assumption can be broken. Be precise.

Question 6: (3 points) A regression analysis that yields a high r^2 but few statistically significant regression coefficients is almost surely suffering from

(a): omitted variable bias

(b): multicollinearity

(c): serially correlated disturbances

(d): the "dummy variable" trap

Question 7: Consider the regression model

$$E(y_i|X_i, D_i) = \beta_0 + \beta_1 X_i + D_i(\beta_2 + \beta_3 X_i)$$

where D_i is a binary indicator variable, indicating membership of one of two mutually exclusive and exhuastive categories

Q7.1 (4 points) Suppose we can not reject the null hypothesis $H_0 : \beta_2 = 0$ but we do find that $\beta_3 > 0$. Graph the implied relationships between y and X for $D = 0$ and $D = 1$.

Q7.2 (4 points) Suppose we can not reject the null hypothesis $H_0 : \beta_3 = 0$ but we do find that $\beta_2 > 0$. Graph the implied relationships between y and X for $D = 0$ and $D = 1$.

Q7.3 (4 points) How would you test the restriction that $E(y_i|X_i) = E(y_i|X_i, D_i)$?

Q7.4 (4 points) Suppose $D_i = 1$ for one and only 1 data point, and is otherwise 0. Can the model given at the start of this question be estimated? Why or why not?

END OF EXAM

Total Number of Points: 84

	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								
df	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
∞	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.