

Political Science 150B/350B
Final Exam
Winter 2005

Answer all questions. Show work for partial credit where appropriate. You will be rewarded for answers that are clear, correct, insightful, and brief. You will not be rewarded for answers that are rambling, incorrect, or confused. The maximum score is printed at the end of the exam.

Question 1: Needless to say, running a marathon is hard work, and doing so in hot weather is likely to reduce runners' performance. Data from the New York City marathon confirms this fact. The male winners' time from the 1978-1998 NYC marathons (in minutes) is regressed on temperature and temperature squared (temperature is measured in Fahrenheit, F), yielding the following results:

$$E(\text{Time}_t|F_t) = 148.51 - .71F_t + .0065F_t^2$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)' = (148.51, -.71, .0065)'$ and are all more than twice the size of their standard errors, and $t = 1978, \dots, 1998$ indexes the marathons.¹

- (a): (5 points) What is the optimal temperature for running the NYC marathon? [Hint: this is a math problem, not a statistics problem]
- (b): (5 points) How would your answer change if $\hat{\beta}_2$ was only half the size of its standard error?
- (c): (5 points) How would you augment the estimated regression model to test for the possibility that (net of year-to-year fluctuations in winning times due to temperature) winning times have been improving over the years?
- (d): (5 points) How would you test the possibility that over time, winning times have become less sensitive to race-day temperature?

Question 2: 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

¹See David E. Martin and John F. Buoncristiani, "The Effect of Temperature on Marathon Runners' Performance", *Chance*, Vol. 12, No. 4 (Fall 1999), pp. 20-24; my source is Orley Ashenfelter, Phillip B. Levine and David J. Zimmerman (2003), *Statistics and Econometrics: Methods and Applications*, Wiley, New York, p193.

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):²

	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): (10 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 or lower-paid workers?

Question 3: (4 points) Let $\mathbf{y} = \mathbf{\Xi}\boldsymbol{\beta} + \mathbf{u}$ be a statistical model of substantive interest. A researcher analyzes the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{w}$, using least squares to estimate $\boldsymbol{\beta}^*$, where $\mathbf{X} = \mathbf{\Xi} + \mathbf{v}$, and

$$\begin{aligned}
 E(\mathbf{v}'\mathbf{w}) &= \mathbf{0} \\
 E(\mathbf{\Xi}'\mathbf{v}) &= \mathbf{0} \\
 E(\mathbf{\Xi}'\mathbf{w}) &= \mathbf{0} \\
 E(\mathbf{v}) &= \mathbf{0} \\
 \text{var}(\mathbf{w}) &= \sigma_w^2 \mathbf{I} \\
 \text{var}(\mathbf{v}) &= \sigma_v^2 \mathbf{I}
 \end{aligned}$$

Is $\hat{\boldsymbol{\beta}}^*$ an unbiased estimator of $\boldsymbol{\beta}^*$?

²See 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 4: A colleague is interested in the effects of campaigns on U.S. presidential elections. She has data on Clinton’s vote margin in 1992 and in 1996 (Clinton vote share minus the Republican candidate vote share) for each state, denoted CVM_{it} , where i indexes states, and $t \in \{92, 96\}$. She also has data on states targeted in a media campaign in 1996 by Clinton’s 1996 political consultant Dick Morris, coded as a binary indicator $D_i = 1$ if state i was a target state, and 0 otherwise.

(a): (10 points) What *substantive* assumptions are being made with the model

$$CVM_{i96} = CVM_{i92} + \varepsilon_i$$

(b): (7 points) Offer a *substantive* interpretation of the intercept in the following model

$$CVM_{i96} = \alpha + \beta CVM_{i92} + \varepsilon_i$$

(c): (7 points) Assume CVM_{i92} is a reasonable measure of a baseline level of support for Clinton in state i . How would you augment the model in the previous question so as to test for the effects of Dick Morris’ media campaign?

Question 5: (5 points) Multicollinearity means

(a): Regression analysis can not proceed because the matrix $\mathbf{X}'\mathbf{X}$ can not be inverted.

(b): Conditional on the predictors \mathbf{X} , the disturbances \mathbf{u} are not iid.

(c): Because the predictors \mathbf{X} are correlated with one another, estimates of their partial impact on \mathbf{y} (i.e., “controlling for” one another) are less precise than if the predictors were uncorrelated with one another.

(d): A critical predictor of \mathbf{y} has been omitted from the regression model.

Question 6: (5 points) When the disturbances are not iid, it is well known that OLS estimates of the regression coefficients $\hat{\beta}_{OLS}$ that are unbiased but not BLUE. GLS can be used for this situation, yielding estimated coefficients $\hat{\beta}_{GLS}$ that are also unbiased. Thus, if OLS and GLS are applied to a data set with disturbances that are not iid:

(a): $\hat{\beta}_{OLS} = \hat{\beta}_{GLS}$

(b): $\hat{\beta}_{OLS} \neq \hat{\beta}_{GLS}$

- (c): each element of the vector $\hat{\beta}_{OLS}$ is smaller than the corresponding element of $\hat{\beta}_{GLS}$
- (d): each element of the vector $\hat{\beta}_{OLS}$ is larger than the corresponding element of $\hat{\beta}_{GLS}$

Question 7: (5 points) A regression's residuals are likely to be highly autocorrelated if

- (a): the estimated t -statistics are smaller than 2 in absolute value
- (b): the X variables have distinct time trends
- (c): the dependent variable y rises and falls over time but none of the X variables have a similar pattern
- (d): we include a linear time trend as one of the predictors in the model

Question 8: Figure 1 presents the results of a Monte Carlo experiment, which assessed the repeated sampling properties of two linear estimators of θ , $\hat{\theta}_A$ and $\hat{\theta}_B$, at a fixed sample size n . θ is known to be 1.0.

- (a): (10 points) Based on Figure 1, what can you say about the statistical properties of estimators A and B ?
- (b): (5 points) Do you have sufficient information to conclude that either estimator is BLUE?

Question 9: Consider the following regression analysis, evaluating a remedial reading project in a large school district. The data are a random sample of 125 students from across the school district who completed the fifth grade and who had been identified as "slow readers" in the previous year (more than a year behind the average 4th grade students). The project evaluation measure (the dependent variable of the regression analysis, *RCHANGE*) is the number of reading points gained in the sample year. Average students gain 100 reading points a year.

Two treatments were administered to the students: remedial group reading sessions, and individual tutoring sessions. These two treatments are measured in hours per week, for each student, with the variables *GROUP* and *TUTOR*, respectively. The variable *COHORT* measures a "baseline growth" or "maturation effect": i.e., $COHORT_i$ is defined as the average gain in reading points in the classroom of slow reader i .

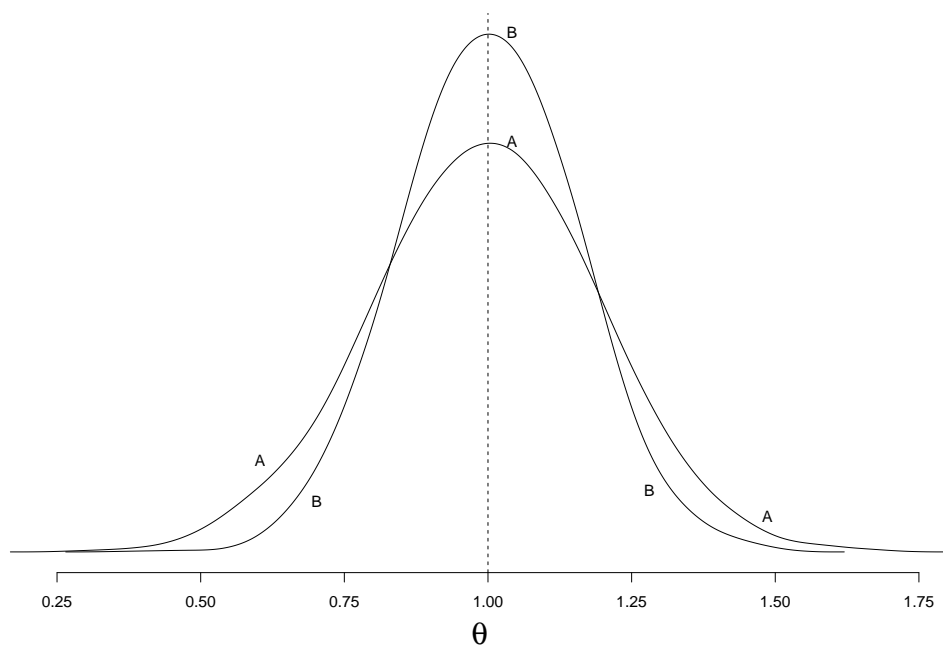


Figure 1: Results of Monte Carlo Experiment, Linear Estimators of θ .

<i>Parameter</i>	<i>Estimate</i>	<i>Std Err</i>
<i>COHORT</i>	.81	.19
<i>GROUP</i>	5.66	.68
<i>TUTOR</i>	12.52	.90
<i>Intercept</i>	-33.32	19.11
<i>Adjusted r²</i>	.63	
$\hat{\sigma}$	12.05	
<i>Mean of dep. var</i>	87.91	

- (a): (18 points) Fully interpret each of the three slope parameters. Think carefully about the effect of the *COHORT* variable (8 points for the interpretation of this coefficient; 5 each for the other slope coefficients).
- (b): (5 points) Briefly describe how to test the null hypothesis that both treatments are ineffective.
- (c): (7 points) Test the null hypothesis that the effects of each tutoring hour per week is greater than 11.5.
- (d): (10 points) Each hour of tutoring costs about \$9 per student, and each hour of group sessions costs about \$3 per student. Test the null hypothesis that the two programs are equally cost effective. The following information will be helpful:

	Lower-triangle of $var(\hat{\beta})$.			
	<i>COHORT</i>	<i>GROUP</i>	<i>TUTOR</i>	<i>Intercept</i>
<i>COHORT</i>	.034783			
<i>GROUP</i>	.0074	.456218		
<i>TUTOR</i>	.003512	.419009	.806414	
<i>Intercept</i>	-3.50749	-2.66247	-3.05492	365.024

Question 10: (10 points) A researcher estimate a regression model with panel data. She finds that without fixed effects for the observational units, the sign of the regression coefficient for an important predictor is negative, and the r^2 is about .15. However, when the researcher includes fixed effects for each unit, the regression coefficient on the important predictor is positive, and the r^2 is about .85. The researcher comes to your office hours and seeking methodological guidance. What is going on in her data?

Question 11: Suppose a researcher is interested in whether attending class is

useful or not and estimates the following model:

$$\text{GRADGPA} = \beta_0 + \beta_1 \text{COLGPA} + \beta_2 \text{GRE} + \beta_3 \text{SKIPPED} + u$$

where GRADGPA is GPA in graduate school, COLGPA is college GPA, GRE is the student's GRE score, and SKIPPED is the average number of classes skipped per week. We believe that u includes, among other things, how lazy the student is.

- (a):** (7 points) Discuss whether OLS estimates of this equation will be biased or not. If biased, can you say anything about the direction of bias?
- (b):** (10 points) You also have data on the distance (in miles) that each student lives from campus, denoted DIST. Describe how you might use this information to obtain consistent estimates of the effects of class attendance.

END OF EXAM

Total Number of Points: 160

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
∞	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.