

Political Science 150B/350B
 Final Exam
 Winter 2007

- 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 after the last question.
- A table of critical quantiles of the χ^2 distribution appears after the exam questions.

Question 1: Many large corporations and government agencies administer a preemployment test in an attempt to screen job applicants. The test is supposed to measure an applicant's aptitude for the job and the results are used as part of the information for making a hiring decision. Data were collected on twenty job applicants, each of whom were hired on a trial basis for six weeks. One week was spent in a training class. The remaining five weeks were spent on the job. The participants were selected from a pool of applicants by a method that was not related to the preemployment test scores. A test was given at the end of the training period and a work performance evaluation was developed at the end of the six-week period. These two scores were combined to form an index of job performance, denoted y_i . Let X_i be the score on the preemployment test. Applicants were also classified into two racial groups, $Z_i = 1$ for minority applicants, and $Z_i = 0$ for white applicants. Regression analysis yielded the following results:

	Model 1		Model 2	
	Estimate	Std. Error	Estimate	Std. Error
Constant	2.01	1.05	1.03	.87
X_i	1.31	.67	2.36	.54
Z_i	-1.91	1.54		
$X_i \times Z_i$	2.00	0.95		
r^2	.664		.52	
$\hat{\sigma}$	1.41		1.59	

The data have the following descriptive statistics:

	Mean	Min	Max
X_i	1.47	0.28	2.51
Z_i	0.50	0.00	1.00
y_i	4.51	1.39	8.14

- (a): (3 points) How many of the job applicants are white?
- (b): (4 points) The model implies two relationships between preemployment test score and job performance, one for whites and one for minorities. What are the slope and intercept parameters of these separate relationships?
- (c): (5 points) How would you test whether the relationship between pre-employment test score and job performance differs across the two racial groups? Can you test that hypothesis with the information provided above? n.b., you don't actually have to compute the test.
- (d): (4 points) For this particular job performance assessment protocol, $y^* = 4$ will be used as a cut-off on hiring decisions (i.e., applicants with $y < 4$ are not hired). For both racial groups, compute the preemployment test score that yields the minimum acceptable job performance score. Comment briefly on the result.

Question 2: (7 points) What does it mean for an estimator to be consistent? If you can give a formal definition, be sure to also give an explanation in words that could be understood by a colleague who has not taken a class in statistics.

Question 3: (5 points) A researcher estimates a regression with $n = 100$ iid observations. The researcher comes to you with a methodological question: if she were to go out and collect 3 times as much data (which will also be iid) how much smaller would the standard errors of her regression be? What is your answer?

Question 4: Consider the data on y and x represented as a scatterplot in Figure 1.

- (a): (5 points) Consider estimating the regression model $E(y) = \beta_0 + \beta_1 x$. Why are the estimates of β_0 and β_1 unlikely to be BLUE?
- (b): (5 points) How might you recover BLUE estimates with these data?
- (c): (5 points) Now consider some additional information about these data. Specifically, Figure 2 shows the data from Figure 1 plotted by a third, binary variable z (solid squares indicate $z = 0$, and open squares $z = 1$). In light of this additional information, how would you re-analyze these data?

Question 5: (4 points) A model with a high r^2 but none or relatively few statistically significant coefficients is an indication of

- (a): omitted variable bias
- (b): heteroskedasticity
- (c): multicollinearity
- (d): non-normal disturbances

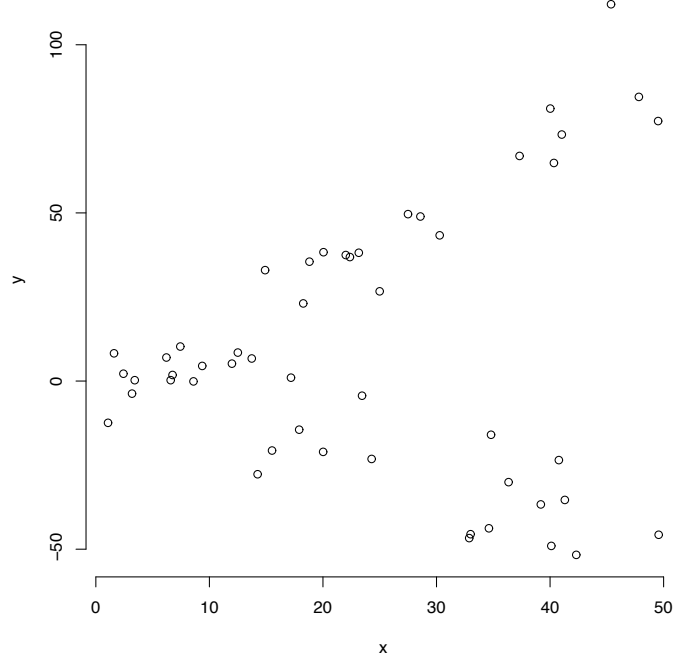


Figure 1: Scatterplot of Hypothetical Data.

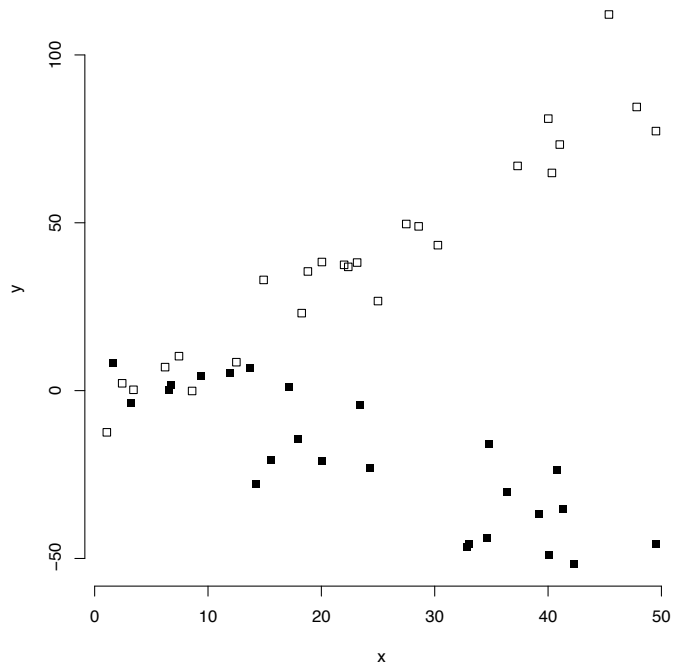


Figure 2: Scatterplot of Hypothetical Data.

Question 6: Consider the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$.

- (a): (5 points) What are the *consequences* of violation of the assumption $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}_n$?
- (b): (10 points) Heteroskedasticity and autocorrelation are two different ways that this assumption can be violated. Describe how each of these phenomena can arise in social-science data.
- (c): (5 points) Sketch a diagnostic residual plot typical of the pattern one would see with a moderate to high level of residual serial autocorrelation. Title and label your graph neatly and accurately.
- (d): (8 points) How would you test the suspicion that $\sigma_i^2 = \sigma^2 X_i^2$? What would you do if your suspicions were confirmed?
- (e): (6 points) A researcher says that he used heteroskedasticity-robust standard errors in presenting his results and in making inferences about $\boldsymbol{\beta}$. Explain how what the researcher did differs from OLS and EGLS (4 points each).

Question 7: Consider the pattern of data shown in Figure 3. Each data point belongs to one of 10 groups. The plotted symbols on the graph, “1”, “2”, etc indicate the group membership of the corresponding data point.

- (a): (2 points) Is $\hat{\beta}_1$ positive or negative?
- (b): (4 points) If we estimated a 2nd regression model that included fixed effects for the 10 groups, what would happen to the estimate of β_1 ?
- (c): (6 points) Based on your answer to the previous question, is the estimate of β_1 from the regression of y on x (shown in Figure 3) biased or unbiased?

Question 8: (5 points) In time series data, what is the “spurious regression” problem and how does it arise?

Question 9: Suppose that, for a given state in the United States, you wish to use annual time series data to estimate the effect of the state-level minimum wage on the employment of those 18 to 25 years old (*EMP*). A simple model is

$$gEMP_t = \beta_0 + \beta_1 gMIN_t + \beta_2 gPOP_t + \beta_3 gGSP_t + \beta_4 gGDP_t + \varepsilon_t$$

where MIN_t is the minimum wage in real dollars, POP_t is the population from 18-25 years old, GSP_t is gross state produce, and GDP_t is U.S. gross domestic product. The g prefix indicates the growth rate from year $t - 1$ to t .

- (a): (5 points) If we are worried that the state chooses its minimum wage partly based on factors that affect youth employment, but that are unobserved (unobserved to us), what is the problem with OLS estimation?

- (b): (5 points) Let $USMIN_t$ be the U.S. minimum wage, which is also measured in real terms. Do you think $gUSMIN_t$ is uncorrelated with u_t ?
- (c): (5 points) By law, any state's minimum wage must be at least as large as the U.S. minimum wage. Explain why this makes $gUSMIN_t$ a potential IV candidate for $gMIN_t$.

Question 10: Maximum likelihood.

- (a): (3 points) What is a likelihood function? Provide as precise a definition as you can.
- (b): (5 points) Given the regression model $E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, under certain conditions it can be shown that the maximum likelihood estimator of $\boldsymbol{\beta}$ is the least squares estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. What are those conditions? Be precise.
- (c): (5 points) Provide an example where least squares and maximum likelihood yield different estimates. Under what conditions is this difference inconsequential?
- (d): (8 points) In a study of public opinion in France, Simon Jackman and Paul Sniderman measured survey respondents' general ideological position on a zero-to-one right-to-left scale (LibIssue), their level of knowledge and awareness about politics (PollInfo) on a zero-to-one scale (low to high) and whether respondents agree ($y = 1$) or disagree ($y = 0$) with the proposition that "more must be done to help the unemployed". The responses were analyzed using logit, yielding the following parameter estimates (obtained via maximum likelihood, standard errors in parentheses):

	Model 1	Model 2
Intercept	-1.62 (.17)	.29 (.49)
LibIssue	2.50 (.28)	-.52 (.81)
PollInfo		-3.74 (.93)
LibIssue \times PollInfo		5.85 (1.48)
Log Likelihood	-1295.90	-1287.50

Test the restrictions of Model 1 vis-a-vis Model 2. Clearly state the null hypothesis, the test statistic, and the conclusion of the statistical test.

- Question 11:** (6 points) A researcher estimates a logit (or probit) model, $\Pr(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i\boldsymbol{\beta})$ where \mathbf{x}_i is a vector of independent variables, $i = 1, \dots, n$. However, the

researcher does not include a “1” in the \mathbf{x}_i , suppressing the intercept term. What strong assumption is the researcher making in estimating a logit/probit model with this restriction? [Hint: what is the interpretation of the intercept in a regular regression model; translate that interpretation into the logit/probit context.]

END OF EXAM

Total Number of Points: 140

df	Upper Tail Area				
	.25	.10	.05	.01	.001
2	2.77	4.61	5.99	9.21	13.8
3	4.11	6.25	7.81	11.3	16.3
4	5.39	7.78	9.49	13.3	18.5
5	6.63	9.24	11.1	15.1	20.5
6	7.84	10.6	12.6	16.8	22.5
7	9.04	12.0	14.1	18.5	24.3
8	10.2	13.4	15.5	20.1	26.1
9	11.4	14.7	16.9	21.7	27.9
10	12.5	16	18.3	23.2	29.6
11	13.7	17.3	19.7	24.7	31.3
12	14.8	18.5	21.0	26.2	32.9
13	16.0	19.8	22.4	27.7	34.5
14	17.1	21.1	23.7	29.1	36.1
15	18.2	22.3	25	30.6	37.7
20	23.8	28.4	31.4	37.6	45.3
30	34.8	40.3	43.8	50.9	59.7
50	56.3	63.2	67.5	76.2	86.7
100	109	118	124	136	149
200	213	226	234	249	268
300	316	332	341	360	381
500	521	541	553	576	603
1000	1030	1058	1075	1107	1144
3000	3052	3100	3129	3183	3245

Table 1: Critical values of the χ^2 distribution.