Political Science 150B/350B
Winter 2004
Midterm Exam

This is a closed-book, in-class exam. You may use a pocket calculator.
There are five questions:

Q1	20 points
Q2	20 points
Q3	8 points
Q4	6 points
Q5	18 points
Total	72 points
Exam Time	50 minutes

Q1 (20 points) Consider the following statistical model:

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

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, $\mathbf{u}$ 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 $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$?

Answer: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

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

Answer: If $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist; i.e., $\mathbf{X}'\mathbf{X}$ is singular, as can be the case if $k < n$ or there are linear dependencies in the columns of $\mathbf{X}$.

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

Answer: $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$.

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?

Answer: $E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma^2_\varepsilon \mathbf{I}_n$.

[Elaboration, not neccessary for full credit:] the variance-covariance matrix of the disturbances is equal to a scalar times a n-by-n identity matrix; this captures: (1) homoskedasticity or $E(u_i^2|X_1,\ldots,X_n) = \sigma^2_\varepsilon, \forall\, i$ and (2) independence or $E(u_i u_j|X_1,\ldots,X_n) = 0, \forall\, i \neq j$.

Q2 (20 points) Given the notation introduced in the previous question, answer the following questions.

Q2.1 (4 points) What is

$$(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})'(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y})$$

Answer: Sum of the squared residuals, $\hat{\mathbf{u}}'\hat{\mathbf{u}}$.

Q2.2 (4 points) True or false, and why?: "Given a set of regression models fit to the same data, a sign of a better fitting model is a smaller $\hat{\sigma}^2_\varepsilon$"

Answer: True. As $\hat{\sigma}^2_\varepsilon \rightarrow 0$, the data points are collapsing around the regression line; recall that $\hat{\sigma}^2_\varepsilon$ is the estimated variance of the disturbances.

Q2.3 (4 points) True or false, and why?: "If each of k coefficients in a regression model is statistically significant, then we will reject the null hypothesis that the k coefficients are jointly zero."

Answer: False. The question of joint significance requires that we consider the covariance terms among the $\hat{\beta}_j, j = 1, \ldots, k$. Depending on the sign of these we might come up with the unusual result of statistically significant t-ratios, but an insignificant F-statistic.

Q2.4 (4 points) True or false, and why?: "The residuals produced by least squares regression are uncorrelated with **y**"

Answer: False. The "why" requires a bit of work to verify, but not impossible:

$$\begin{aligned} \mathbf{y}'\hat{\mathbf{u}} &= \mathbf{y}'(\mathbf{y} - \mathbf{H}\mathbf{y}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y} \\ &= \mathbf{0} \iff \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n \end{aligned}$$

which will almost always be false (and will only be true if we ran a truly degenerate regression where $\mathbf{X} = \mathbf{I}_n$ i.e., $k = n$).

Q2.5 (4 points) True or false, and why?: "A useful check of omitted variable bias is to plot the residuals produced by least squares regression against the regression's predictors, **X**."

Answer: False. The least squares residuals $\hat{\mathbf{u}}$ are uncorrelated with the **X** by construction. See below. The best way to look for omitted variable bias is to plot the residuals against suspected omitted variables themselves, not against inclued variables.

$$\begin{aligned} \mathbf{X}'\hat{\mathbf{u}} &= \mathbf{X}'(\mathbf{y} - \mathbf{H}\mathbf{y}) \\ &= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{0} \end{aligned}$$

Q3 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:

1. *unbiasedness* or the property that $E(S_t | V_t = .5) = .5$
2. *responsiveness*: how much does S_t change in result to change in V_t?

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

$$\ln \frac{S_t}{1 - S_t} = \alpha + \beta \ln \frac{V_t}{1 - V_t} + \varepsilon_t$$

Q3.1 (4 points) Interpret α.

Answer: $\alpha = E\left[\ln \frac{S_t}{1\text{-}S_t} | \ln \frac{V_t}{1\text{-}V_t} = 0\right] = E\left[\ln \frac{S_t}{1\text{-}S_t} | V_t = .5\right]$ If $\alpha = 0$ then we have an unbiased electoral system, given the definition, above, since $ln \frac{S_t}{1\text{-}S_t} = 0 \iff S_t = .5$

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?

Answer: F-test. 2 degrees of freedom in the numerator. Would compare the sum of squared errors one obtains under $H_0 : \alpha = 0, \beta = 3$ with the sum of squared errors obtained from an unconstrained model.

Q4 (6 points) Imagine you are attending a seminar where the speaker presents a cross-national model of welfare expenditures. The speaker's model contains just one independent variable, GDP per capita, which has a large, positive, and statistically significant parameter estimate. But you are skeptical. You reason that welfare expenditures are also driven by unemployment levels: for example, when more people are unemployed, the demand on the welfare system goes up. What is the implication of this reasoning for the speaker's estimate of the effect of GDP per capita on welfare expenditures?

Answer:

This is a possible case of omitted variable bias. The bias term is

$$E(\hat{\beta}_{GDP}) - \beta_{GDP} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{x}_2\beta_{UNEMP}$$

where $\mathbf{x}_2$ is the vector of observations on unemployment, and $\mathbf{X}$ is the seminar speaker's data matrix, containing the data on GDP per capita (and perhaps an intercept term).

Signing the bias here requires signing the $\mathbf{X}'\mathbf{x}_2$ term, since $(\mathbf{X}'\mathbf{X})^{-1}$ is a positive definite matrix and we assume that β_{UNEMP} is positive. There are three possible cases to consider (full credit requires considering at least one):

1. if unemployment is **negatively** correlated with GDP per capita (poor countries have higher unemployment rates), then $\mathbf{X}'\mathbf{x}_2$ is negative, and the sign of the bias term is *negative* (negative times a positive). That is, the seminar speaker is *under-estimating* the true effect of GDP per capita on welfare expenditures.

2. if unemployment is **positively** correlated with GDP per capita (rich countries have higher unemployment rates?) then $\mathbf{X}'\mathbf{x}_2$ is positive, and the sign of the bias term is *positive* (positive times a positive). That is, the seminar speaker is *over-estimating* the true effect of GDP per capita on welfare expenditures.
3. if unemployment is **uncorrelated** with GDP per capita, then $\mathbf{X}'\mathbf{x}_2$ is zero, and the bias term is zero. That is, the omission of unemployment from the model does not effect the estimate of the effect of GDP per capita on welfare expenditures.

Q5 Consider the following set of regression estimates, obtained from a random sample of 300 observations:

Variable	*Parameter*	*Estimate*	*t-value*
Intercept	β_0	324	12.2
X_1	β_1	8.8	3.1
X_2	β_2	.5	6.3
X_3	β_3	.7	4.2
X_4	β_4	1.2	.5
D	β_5	-506.3	-3.8
$D \times X_3$	β_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 zero for group A and 1 for 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.

Q5.1 (4 points) Discuss the relative importance of X_1 and X_2 for explaining variation in y.

Answer: Despite differences in parameter estimates, the range of X_2 is so large relative to range of X_1 that X_2 will be more important in accounting for variation in y. Movement across the entire range of X_2 would yield a 3500 unit change in y, but movement across the entire range of X_1 would yield only a 88 unit change in y.

Q5.2 (4 points) Does the model predict higher or lower values of y for members of group B relative to members of group A?

Answer: Despite the significant intercept decrease for group B, this is more than offset by the interaction with X_3. Even at the lowest level of X_3, the contribution in terms of y is $3000 \times .65 = 1950$ units, less the intercept shift of -506, to yield "net" y values for X_3 about 1450 units greater than group A.

Q5.3 (6 points) A researcher is interested in whether the impact of a unit increase in X_2 is offset by a unit decrease in X_3? What is the null hypothesis here? How would you test it? Can you implement your proposed test with the information provided above?

Answer: The null hypothesis here is $\beta_2 = \beta_3$ or $H_0 : \beta_2 - \beta_3 = 0$. Use a *t*-test:

$$t = \frac{\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_3}{\sqrt{var(\hat{\boldsymbol{\beta}}_2) + var(\hat{\boldsymbol{\beta}}_3) - 2cov(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3)}}$$

Require covariance term, which can not be recovered from information in the table.

Q5.4 (4 points) State the form of the null hypothesis that tests for the significance of the intercept shift for group B, *and* the group B-specific effect of X_3. What kind of statistical test is this? Describe how you might implement the test and interpret the result.

Answer: *F* test for the two joint restrictions on β_5 and β_6, i.e., $H_0 : \beta_5 = \beta_6 = 0$. This restriction nests within the reported model, so simply re-run the regression with the corresponding variables dropped and comparing the explained sums of squares as per usual procedure. If the null is true, the *F* test-statistic follows the *F* distribution with two degrees of freedom in the numerator, $n - 7 = 293$ degrees of freedom in the denominator.