Economics 551 — Professor Rust
Spring 1999

Midterm Exam
(Due at start of class, February 24, 1999)

Part I: Regression Questions (400 points: use of computers not required to do this part)
Do Question 1 and *2 out of 3* of the remaining Part I questions below.

Question 1 (200 points).

a. Compute the OLS estimates $\hat{\beta}$ for the following 2-variable linear regression problem:

$$y = \begin{bmatrix} 1 \\ -1 \\ 3 \\ 2 \\ 5 \\ 4 \\ 3 \\ -1 \\ 2 \end{bmatrix} \quad X_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \\ -2 \\ -1 \\ 0 \\ 3 \\ -2 \\ -1 \end{bmatrix} \quad X_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \end{bmatrix}$$

b. Unfortunately, there was one student who didn't know how to invert a (2×2) matrix. Thinking that it would be unnecessary to estimate $\beta = (\beta_1, \beta_2)$ as a whole, he proposed the following estimation formulas:

$\widehat{\beta}_1 = X_1'Y \ / \ X_1'X_1$, and
$\widehat{\beta}_2 = X_2'Y \ / \ X_2'X_2$.

Calculate these "naive" estimators of β_1 and β_2 and compare them with those obtained in b.

c. What do you think about his approach? Will it work generally? Will the naive estimators be unbiased and consistent? If not, specify the conditions needed to justify his approach. What's the intuition behind those conditions?

d. How can you generalize your argument in d to the case of a k-variable linear regression model?

e. Show that if a regression contains a set of K mutually exclusive *dummy variables* $\{D_1, \ldots, D_K\}$ where the variables are mutually exclusive in the sense that if $D_{ij} = 1$ (observation i is 1 for the j^{th} dummy variable), then $D_{ik} = 0$ for all $k \neq j$ (i.e. all of the other dummy variables dummy take the value 0 for observation i), then the K OLS regression estimates $(\hat{\beta}_1, \ldots, \hat{\beta}_K)$ in the regression

$$y = \beta_1 D_1 + \cdots \beta_K D_K + \epsilon$$

are given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^{N} y_i D_{ij}}{\sum_{i=1}^{N} D_{ij}} \quad j = 1, \ldots, K$$

Show that $\hat{\beta}_j$ is just the mean of y over the subpopulation of individuals i with $D_{ij} = 1$.

Question 2. (100 points) Consider the general multivariate regression model

$$y = X\beta + \epsilon$$

a. Suppose you estimate the OLS estimate of β, $\hat{\beta}$, and then compute $\hat{y} = X\hat{\beta}$, the $(N \times 1)$ vector of predicted values of y, and $\hat{\epsilon}$, the $(N \times 1)$ vector of error terms, $\hat{\epsilon} = y - \hat{y}$.

   1. What is the value of the inner product of $\hat{y}$ and $\hat{\epsilon}$?
   2. Justify your answer in part a-1 above. You can either use a geometric argument or an algebraic derivation. Can you give an intuitive explanation for your result?
   3. What are the implications of your answer for regression analysis ?

b. Consider now the quantity $c = \hat{y}'\hat{y}/y'y$.

   1. Show that c has to be between zero and one by using your answer to part a above. (**Hint:** use the pythagorean theorem.)
   2. Does c provide any sort of measure of "goodness of fit" of the regression model? Explain your answer for full credit. What is the interpretation of the case where $c = 0$? What is the interpretation of the case where $c = 1$?

Question 3 (100 points). This question considers a regression through the origin and the connection with the geometric notion of projection in three dimensional space. You should be able to answer this question using simple matrix algebra, without the use of a computer. Consider the following model:

$$y = Z_1\beta_1 + Z_2\beta_2 + \epsilon$$

where

$$\begin{aligned} Z_1' &= (2\ 0\ 0) \\ Z_2' &= (0\ 2\ 0) \\ y' &= (3\ 2\ 3) \end{aligned}$$

a. Write the model in the form $y = X\beta + \epsilon$ using matrices. Specify these matrices and their dimensions.

b. Calculate $(X'X)$ and its inverse. Verify that $(X'X)(X'X)^{-1} = I$ where I is the (2×2) identity matrix.

c. Derive the least squares estimates $\hat{\beta}$ and the predicted value $\hat{y}$.

d. In a three dimensional diagram, display the following:

   1. the subspace S spanned by the columns of X (shade region).

2. the vectors y, Z_1 and Z_2 and $\hat{y}$.
3. the orthogonal projection of y onto the subspace S sketched in part d-1 above.

e. Derive the vector of residuals $\hat{\epsilon} = y - X\hat{\beta}$ and calculate:

1. $\vec{i}'(y - X\hat{\beta})$ where $\vec{i}' = (1\,1\,1)$. Does $\vec{i}$ lie in the subspace spanned by the columns of X? Does your result for the value of $\vec{i}'(y - X\hat{\beta})$ shed any light on this?
2. $Z_1'(y - X\hat{\beta})$. Is this equal to zero? Explain your answer in either case.

f. The orthogonal projection of y into the subspace S spanned by the columns of the matrix X is given by $Py = X(X'X)^{-1}X'y$. Calculate Py and verify that $Py = \hat{y}$.

g. A symmetric square matrix A is idempotent if $A'A = A$. Show that P given above is idempotent. Calculate the rank of P. Is P an invertible matrix?

Question 4. (100 points) The following questions concern OLS estimation of the general linear model

$$y = X\beta + \epsilon \tag{1}$$

where y is $N \times 1$, X is $N \times K$ and ϵ is an $N \times 1$ vector of error terms.

a. What happens when we try to do OLS when the $N \times K$ regressor matrix X has rank less than K? Is the $X'X$ matrix invertible in this case? If not, does the OLS estimate $\hat{\beta}$ exist?

b. Does the problem of *multicollinearity* have anything to do with the rank of X?

c. Show that when $X'X$ is not invertible there are generally *infinitely many* solutions to the normal equations for the OLS estimator.

d. Does the best fitting predicted y, $\hat{y}$, exist when $X'X$ is not invertible? If yes, can you provide a formula for $\hat{y}$ or a procedure for computing it?

e. Define what is meant by the *generalized inverse*, A^+ of a square matrix A. Is the $K \times 1$ vector $(X'X)^+X'y$ a solution to the normal equations if $X'X$ is not invertible?

f. Describe the process of *stepwise regression* and discuss whether this procedure will allow us to compute the predicted values of the dependent variable, $\hat{y}$. If $X'X$ is invertible, will the coefficients produced by stepwise regression coincide with the coefficients from the standard OLS formula $(X'X)^{-1}X'y$?

Part II: Applied Regression/IV Questions (computers required for some questions in this part)
Do Question 0 and Question 1 or 2 and Question 3 or 4

Question 0. (200 points) Given an $N \times J$ matrix of *instruments* and an $N \times K$ matrix of *endogenous regressors* we form the *instrumental variables estimator.*

a. What is the equation for the instrumental variables estimator? Consider separately the three cases, $J = K$, $J > K$ and $J < K$.

b. In the overidentified case, $J > K$, we have more instruments than endogenous regressors. Suppose we form a $N \times K$ matrix of instruments $W = Z\gamma$ for some $J \times K$ matrix γ. Derive the formula for the class of IV estimators and show how it depends on the choice of γ. Is there an "optimal" choice for γ? If so, describe what the optimal γ is and in what sense this choice is optimal.

c. Sketch the argument for showing the consistency and asymptotic normality of the IV estimator for two cases: 1) the homoscedastic case, and 2) the heteroscedastic case.

d. Justify your answer in part b above by showing that in the homoscedastic case your choice of γ results in an IV estimator that has the smallest asymptotic covariance matrix among all IV estimators.

Question 1. (100 points) Consider the vector of observations contained in the file `pop` (populations in each of the 50 states, to be found in the `I:\Spring99\econ161\crossdat\fmt` directory at the Statlab). Load it and call it X.

a. Compute $(X'RX)/49$, where R is given by $R = I - \vec{\imath}(\vec{\imath}'\vec{\imath})^{-1}\vec{\imath}'$ whre I is the (50×50) identity matrix and $\vec{\imath}$ is a (50×1) vector of ones. Could $(X'RX)/49$ ever be negative? Why or why not?

b. Compute the sample standard deviation of the population of the U.S. and the sample variance by using simple Gauss commands.

c. Compare your result in a. with your result in b. Are the answers to a. and b. the same or different? If they are the same, provide an explanation for why this is the case.

(**HINT:** you might want to examine $X'RX$. Recall the properties of the matrix R, namely $R = R'R = R * R$, and see what R does to X).

Question 2 (100 points). Consider the set of hypothetical data on the regress model below.

$$y_i = \beta_1 + \beta_2 X_{2i} + \beta_{3i} X_{3i} + \epsilon_i$$

where

$$\begin{array}{rcl} y' & = & (-10 \quad -8 \quad -6 \quad -4 \quad -2 \quad 0 \quad 2 \quad 4 \quad 6 \quad 8 \quad 10) \\ X_2' & = & (1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11) \\ X_3' & = & (1 \quad 3 \quad 5 \quad 7 \quad 9 \quad 11 \quad 13 \quad 15 \quad 17 \quad 19 \quad 21) \end{array}$$

a. Can you compute OLS estimates of the three unknowns $(\beta_1, \beta_2, \beta_3)$?

b. Is there a problem of multicollinearity in this regression? If not, show that the columns of the X matrix are linearly independent. If so, show that the columns of the X matrix are linearly dependent.

c. Throwing out any redundant columns of the X matrix if necessary, what is the R^2 of the regression?

d. Suppose that there are two students in the econ 551 class, whose names are Jim and Tom. Suppose further that they estimated the parameters β_1, β_2 and β_3 by trial and error. As a result, however, Tom and Jim got different answers, i.e., $(-6, -10, 6)$ and $(-10, -2, 2)$, respectively. And each of them argues that his answer is correct. What do you think about these two answers? Which answer fits better to the data (in the sense of having a higher R^2)?

Question 3. (200 points) One researcher wants to estimate the money demand equation by the following regression:

$$\log(m_t) = \beta_0 + \beta_1 \log y_t + \beta_2 r_t + \epsilon_t, \quad t = 1, \ldots, T$$

where: m_t = real money balances (i.e., nominal money balance deflated by the price level), y_t = real GNP(i.e., nominal GNP deflated by the price level) and r_t = nominal interest rates.

a. What signs of the β's do you expect from the economic theory? Explain why.

b. Run the above regression using data files accessible via anonymous ftp from `gemini.econ.yale.edu` in the subdirectory `pub/John_Rust/courses/econ161/stats/timedat/fmt` or in the `i:\Spring99\econ161\timedat\fmt` directory at Statlab. You will be using the following variables: `GNP`=Nominal GNP, `CPI`=Price level, `R_3MO`=Interest rates, and `M2`=Nominal money balances. What is your estimates of β's ? On the basis of this evidence, what do you conclude about the validity of the money demand equation given above?

c. Now another researcher is estimating somewhat different version of money demand equation given by

$$\log M_t = \beta_0 + \beta_1 \log Y_t + \beta_2 r_t + \beta_3 \log P_t + \epsilon_t, \quad t = 1, \ldots, T.$$

where M_t = nominal money balances, r_t = nominal interest rate, Y_t = nominal GNP, and P_t = price level. What signs of β's do you expect? Why?

d. Run the regression in part c using the data files given in b. What is your estimates of β's ? On the basis of this evidence, what do you conclude about the validity of the money demand equation given above?

e. Compare the two regression models in terms of R^2 and/or the plausibility of the estimated coefficients.

f. Run a simple regression of aggregate consumption, C on a constant and GNP. What is your estimate of the *marginal propensity to consume.*?

g. Using the residuals calculated from your regression in part f above, compute the *serial correlation coefficient* of the regression residuals (i.e. compute $\rho(\hat{e}_t, \hat{e}_{t-1})$). Are the residuals serially uncorrelated, or negatively or positively correlated? Does your finding contradict the normal equations that show that the residuals in a regression should be "unpredictable" in the sense of being uncorrelated with the independent variables in the regression?

Question 4 (200 points) Due to the fact that a large number of buyers and sellers interact in a market for a nearly homogeneous good, the market for soybeans is nearly perfectly competitive. Contracts for soybeans on the Chicago Board of Trade and the daily market or equilibrium price of soybeans is known as the *spot price.* Demand for soybeans is a function of the price of soybeans, p and personal income, y. Assume that the aggregate demand curve for soybeans is linear:

$$q_d = \beta_0 + \beta_1 p + \beta_2 y + \epsilon_d$$

where q_d is the quantity of soybeans demanded, p is the market price of soybeans, y is per capita income and ϵ_d represents other unobserved factors affecting the demand for soybeans. Assume the supply of soybeans q_s is also a linear function of price, average rainfall r, and other factors ϵ_s:

$$q_s = \alpha_0 + \alpha_1 p + \alpha_2 r + \epsilon_s$$

a. What does economic theory (or common sense) tell us about the signs of the coefficients $\beta' = (\beta_0, \beta_1, \beta_2)$ of the demand curve? That is, do we expect the β_k coefficients to be negative positive or zero? (Explain your reasoning for full credit).

b. What does economic theory (or common sense) tell us about the signs of the coefficients $\alpha' = (\alpha_0, \alpha_1, \alpha_2)$ of the supply curve? That is, do we expect the α_k coefficients to be negative positive or zero? (Explain your reasoning for full credit).

c. The file `soy.asc` available via anonymous ftp at `gemini.econ.yale.edu` in the subdirectory `pub/John_Rust/courses/econ161/soy.asc` (the easiest way to get the data is simply to click on the `soy.asc` hyperlink on the version of this problem set on the Econ 551 web page). This data set contains 200 monthly observations of soybean market prices, quantities traded, per capita income y, and average rainfall, r. Retrieve these data and estimate the parameters α and β by running OLS on the demand and supply side equation separately. Report standard errors.

d. Do the results from OLS confirm or disconfirm the hypotheses you have made in part a and b? Explain why you are not getting the expected results.

e. Propose an estimator other than OLS that can improve your results. Explain the theory behind the improvement.

f. Provide estimates and standard errors of estimates using the method proposed in part e. State clearly how the method proposed in part e is implemented for this particular problem, and with this particular data. Summarize your estimation results. Do the new results confirm the hypotheses?