# The Elements of Regression Analysis

R.G. Hollister

## I. Objective

My objective in these notes is to introduce the basic elements of regression analysis in as direct a fashion as possible, emphasizing, wherever possible, more intuitive ways of looking at estimating procedures and eschewing most of the refinements of underlying theory.

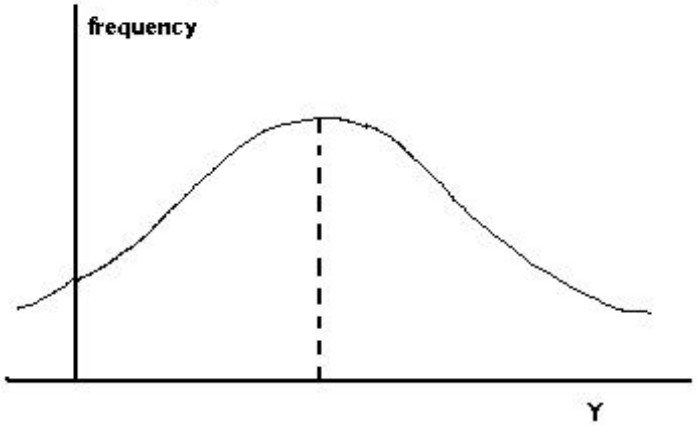
The objective of regression analysis is to determine approximative systematic relationships of many variables, given the joint distribution of two or more variables.

## II. Means, Conditional Means, and Simple Regression

## A. Mean

Given a distribution of a variable, we often wish to select a ”most representative,” ”most likely,” or ”expected” value for that variable. Since the variable is distributed over a range of values, we know any single value we choose will often be ”misrepresentative” of the ”most likely” value of any single observation we might pick from the distribution. Thus we like to select the value which minimizes the ”misrepresentativeness” or ”error.”

Figure A: Distribution of Y




Let us call the value we select the estimator of Y which we will label E(Y). We use the notation of $Y_i$ for any single value selected from the distribution of values of Y. When we choose E(Y) as the estimator for the value of Y, the error we make in using E(Y) for any $Y_i$ is:

$$e_i = Y_i - E(Y) \quad (0.1)$$

The extent to which the estimator E(Y) is "misrepresentative" over the whole distribution of Y might be represented by the sum of the errors for each observation, e.g., with n observations:

$$\sum_{i=1}^{n} e_i = e_1 + e_2 + e_3 + ...e_n \quad (0.2)$$

The problem is to develop a criterion by which to select a "best" value for E(Y). One reasonable criterion would be to choose E(Y) so as to make the sum of errors, as small as possible. Therefore, a criterion of choosing E(Y) so that:

$$\sum_{i=1}^{n} e_i = 0 \quad (0.3)$$

would seem reasonable.
We can write:

$$(0.4) \qquad \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - E(Y)) = \sum_{i=1}^{n} Y_i - nE(Y)$$

since E(Y) is a constant which enters with the same value in each of the n terms of the summation. We can write (0.3) as:

$$(0.5) \qquad \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - nE(Y) = 0$$

Solving for E(Y), we get:

$$(0.6) \qquad E(Y) = \frac{\sum_{i=1}^{n} Y_i}{n}$$

which is the average or mean value for Y. Therefore our justification for using the mean as an estimator is that it gives us the smallest absolute value of the sum of errors over the whole distribution. We note in passing that another criterion might be to choose E(Y) so that it gives a minimum for the sum of squared errors, $\sum_{i=1}^{n} e_i^2$. Using calculus we get:

$$(0.7) \qquad min \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - nE(Y))^2 = 0$$

$$(0.8) \qquad \frac{\sum_{i=1}^{n} e_i^2}{dE(Y)} = -2 \sum_{i=1}^{n} (Y_i - E(Y)) = -2 \sum_{i=1}^{n} (Y_i - nE(Y)) = 0$$

and solving for E(Y):

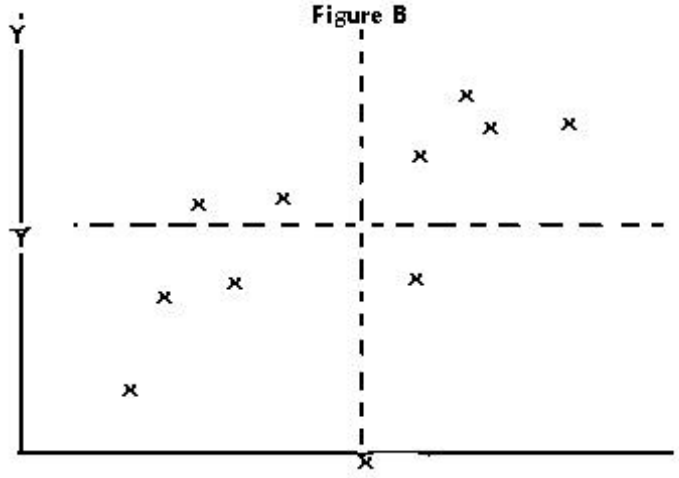
$$(0.9) \qquad E(Y) = \frac{\sum Y_i}{n}$$

Therefore, the mean, $\frac{\sum Y_i}{n}$, is the value that minimizes the sum of the squared errors as well as the absolute value of the sum of errors. In subsequent notation we will often refer to $\frac{\sum Y_i}{n}$ by the notation $\bar{Y}$.

## B.The Conditional Mean

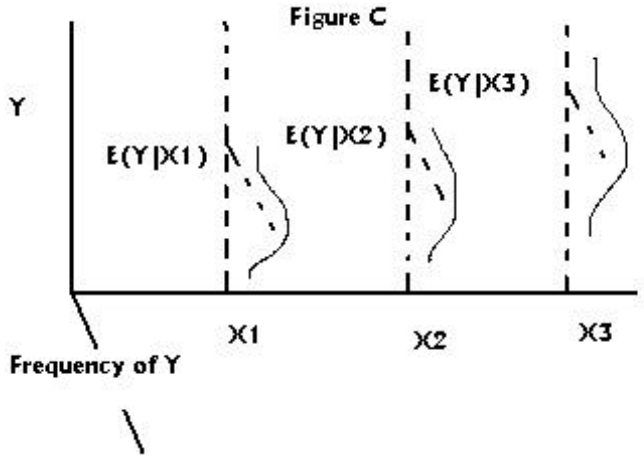
Now, if for each observation we take note of two characteristics, we have distributions for two variables tied together by their association with the common observation points (Yi, Xi) - this is a joint distribution of the variables Y and X. We wish to look at relationships between the two variables in the joint distribution. We can represent their joint distribution by a scatter diagram.

Figure B




We could simply look at the mean of each distribution , $\overline{X}, \overline{Y}$.but this wouldn't tell us much about how one varies as the other varies. We would prefer to know, for example, given a value of X, $X_i$, what is the "likely" value of Y. We can talk about the "likely" value of Y given $X_i$ as the "conditional expected value," $E(Y \mid X_i)$. For example: given a person's age, what is the most likely value of their income? We might anticipate, following the reasoning from the previous section, that the best estimator is the conditional mean of Y given X.

We can think of the scatter diagram as giving a separate distribution of each value $X_i$, as in Figure C.

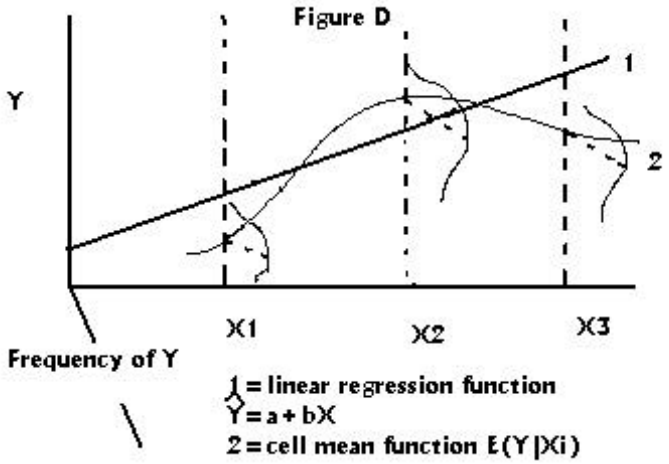



The function which gives the "conditional expected value of X", as indicated in Figure C is sometimes called the "cell mean function."

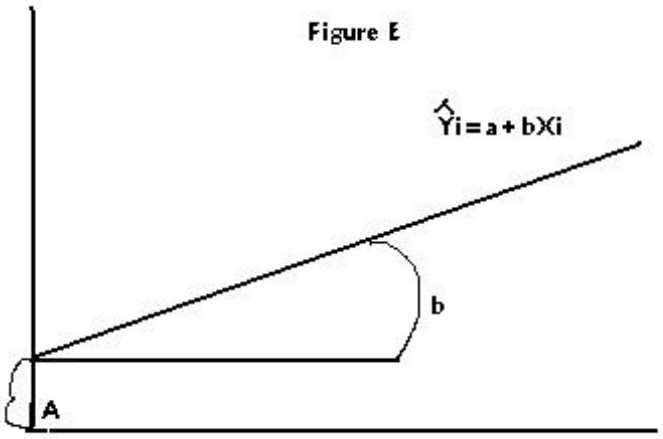
We may wish to generalize the relationship between the expected value of Y and given value of X in a somewhat handier form than provided by the "cell mean function." To do this, we make some assumption about the form of the relationship between Y and X. The simplest assumption is that it is linear.

$$(10.10) \quad E(Y \mid X_i) = \hat{Y}_i = a + bX_i$$

(Other forms can be assumed and the logic of estimation carried out in a similar fashion). As indicated in Figure D, the linear regression function can be thought of as a simplification of the "cell mean function."




If we used the ”cell mean function” we would have to calculate a separate conditional mean for each value of $X_i$. This would be cumbersome both to calculate and to use for analysis. Using the linear regression line we need select only two values: a, the intercept, and b, the slope of the line.




The linear regression is not quite as accurate as the ”cell mean function” but it is easier to calculate and handle.

## C. Simple Regression

The observed value of $Y_i$ from the observed joint distribution $(Y_i, X_i)$ will differ from the value predicted from the linear regression function, $Y_i = a + bX_i$, just as, in the case of the single distribution, the observed value $Y_i$ differed from the predicted value E(Y). Thus the "error":

$$(1.1) \quad e_i = (Y_i - \hat{Y}_i) = (Y_i - a - bX_i)$$

Now, following the analogy to selecting the best value for E(Y), we would like to pick the "best" value for $Y_i$. Since $Y_i = a + bX_i$, the problem is to choose the "best" values for a and b.

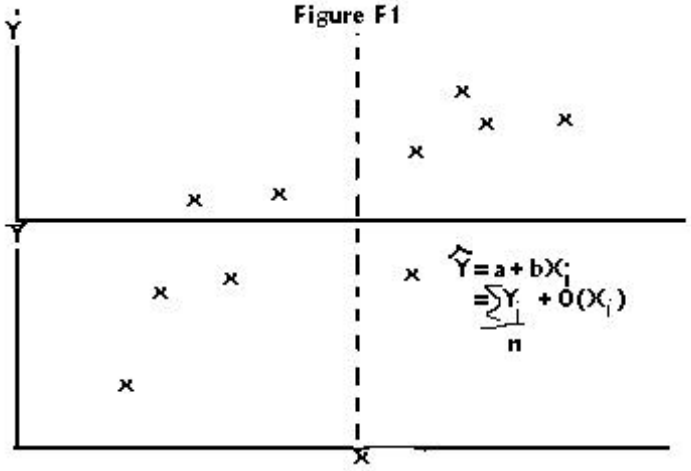
Following the analogy to the case of the mean above, one reasonable criterion would seem to be to choose the values so that $\sum_{i=1}^{n} e_i = 0$. Thus,

$$(1.2) \quad \sum e_i = \sum \left(Y_i - \hat{Y}_i\right) = \sum (Y_i - a - bX_i) = \sum Y_i - na - b \sum X_i = 0$$

However, there are lots of combinations of values a and b which satisfy equation (1.2). Consider, for example, a= $\frac{\sum Y_i}{n}$ and b=0. Substituted in (1.2):

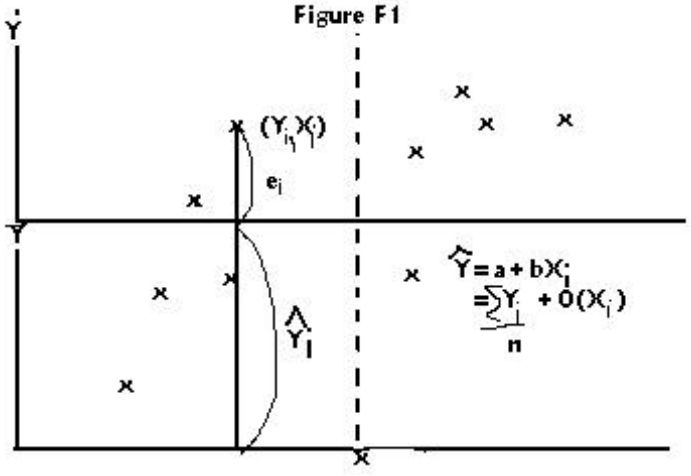
$$\sum e_i = \sum \left( Y_i - \frac{\sum Y_i}{n} - 0(X_i) \right) = \sum Y_i - n\frac{\sum Y_i}{n} = 0$$

Diagrammatically this function would look like Figure F1:




This is clearly not a very good estimating function. We see that though the summation of errors is zero,( $\sum e_i = 0$ , this is achieved because positive errors to the right of $\bar{X}$cancel out against the negative errors to the left of $\bar{X}$ .

In Figure F2, we show a particular observation ($Y_i, X_i$and how the error is the difference between the regression line point for ($Y_i, X_i$)and the actual value of $Y_i$.

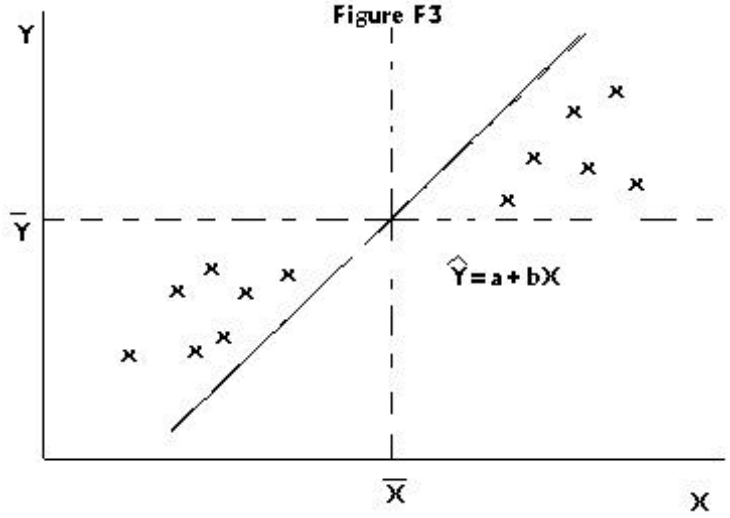


We can see that with the regression function using a=$\frac{\sum Y_i}{n}$ and b=0, most of the observations to the right of $\overline{X}$, will have positive errors ($e_i > 0$), while most of the observations to the left of will have negative errors ($e_i < 0$). Now multiply each value of $X_i$ times the error associated with it for the given regression line, i.e., $e_iX_i$. Take the sum of $e_iX_i$, $\sum e_iX_i$. Since the values of $X_i$ to the right of $\overline{X}$ are larger than those to the left of $X_i$, the predominately positive values of $e_i$ to the right of $\overline{X}$ are multiplied by larger values than the predominately negative values of $e_i$ to the left of $\overline{X}$, and $\sum e_iX_i$ will be positive. Thus when the regression line as in Figures F1 and F2 is clearly too flat, relative to the observations,

$$\text{Ý1.3þ} \quad \sum e_iX_i > 0$$

If the regression line had too sharp a slope, as indicated in Figure F3, since large values of X to the right of $\overline{X}$ multiply mostly negative values of $e_i$ and small values of X to the left of $\overline{X}$ multiply mostely positive values of $e_i$,

Figure F3




Then, in general, the errors to the right of would be negative, to the left of , the errors would be positive. Since, in general, in , the negative errors would be multiplied by larger values of than the positive errors

we would expect that:

$$\text{Ý}1.4\text{þ} \quad > e_i X_i < 0$$

If we are to avoid both a regression line too flat and one too steep, it seems reasonable to impose a second condition that:

$$\text{Ý}1.5\text{þ} \quad > e_i X_i = 0$$

$$= \quad > \text{Ý} Y_i ? a ? bX_i \text{þ} X_i = > \text{Ý} Y_i X_i \text{þ} ? a > \text{Ý} X_i \text{þ} ? b > X_i^2 = 0$$

If we now combine equation (1.2) and equation (1.5) as conditions to be met, we have two equations with the two values to be selected, a and b, in them. We recall from algebra that, in general, two equations will uniquely determine two unknowns, so we can solve (1.2) and (1.5) simultaneously for the values of a and b.

Multiply (1.2) by $> X_i$ and (1.5) by n to get:

$$\text{Ý}1.6\text{þ} \quad > Y_i > X_i ? na > X_i ? b\left( > X_i \right)^2 = 0$$

$$n > Y_i X_i ? na > X_i ? bn > X_i^2 = 0$$

Subtracting the first equation from the second we get:

$$\text{Ý}1.7\text{þ} \quad \text{n} > Y_i X_i ? > Y_i > X_i ? bn > X_i^2 + b\text{Ý} > X_i \text{þ}^2 = 0$$

Solving for b we get:

$$\text{Ý}1.8\text{þ} \quad \text{b} = \frac{n > Y_i X_i ? > Y_i > X_i}{n > X_i^2 ? \text{Ý} > X_i \text{þ}^2}$$

Rearranging equation (1.2) we see:

$$a = \frac{> Y_i}{n} ? \frac{b > X_i}{n}$$

and we can substitute b from equation (1.8) into equation (1.9) to get the expression for a.

We have thus arrived at a choice of values for the regression line

parameters a and b by imposing the conditions $\sum e_i = 0$ and $\sum e_i X_i = 0$. We now show quickly that if we adopted the criterion of choosing a and b so as to minimize the squared error, we would arrive at the same estimates. Working from equation (1.1) we get:

$$\sum e_i^2 = \sum (Y_i - a - bX_i)^2 \quad (1.10)$$

Minimizing equation (1.10) with respect to a and b we get:

$$\frac{\partial \sum e_i^2}{\partial a} = -2 \sum (Y_i - a - bX_i) = 0 \quad (1.11)$$

$$\frac{\partial \sum e_i^2}{\partial b} = -2 \sum (Y_i - a - bX_i) X_i = 0$$

which can be rewritten as:

$$0 = \sum Y_i - na - b \sum X_i = 0 \quad (1.12)$$

$$0 = \sum Y_i X_i - a \sum X_i - b \sum X_i^2$$

But these are exactly the same as equations (1.2) and (1.5). These are often referred to as the normal equations of the least sum of squares regression. Solving these for the values of b and a, we would get the same expressions as equations (1.8) and (1.9). From the intuitive development of equations (1.2) and (1.5), we can see why the least sum of squared errors criterion is utilized. The criterion of minimizing the absolute sum of errors is not sufficient alone. There are many values for a and b which will give $\sum e_i = 0$ because large positive errors in the sum cancel out large negative errors (in fact any line passing through this point $(\bar{Y}, \bar{X})$ will meet this criterion). Thus we need an additional criterion to pick the ”best” estimator from among these many. The estimator values of a and b which cause $\sum e_i X_i = 0$ will give a line which is neither too flat nor too steep. When we use the overall criterion of minimizing the sum of squared errors, since the errors are squared before summing, positive and negative errors don’t cancel out the summation. When we minimize the sum of squared errors, we arrive at these two conditions. The two conditions give two equations in two unknowns which can be solved for unique values of a and b. This then is the method of least squared error simple regression.

It defines the linear relationship between Y and X which minimizes the sum of the squared errors made when that line is used to estimate a value of Y, ($Y_i$), for any given value of X, $X_i$.

## D. Regression in Deviation Notation and Moment Notation

It is useful to transform some of the relationships above into a "normalized" form by redefining the variables in the joint distribution in terms of deviations from the mean.

$$(2.1) \quad y_i = (Y_i - \bar{Y})$$

$$(2.2) \quad x_i = (X_i - \bar{X})$$

Taking equation (1.2) dividing by n and transforming, we get:

$$(2.3) \quad \frac{\sum Y_i}{n} = a + \frac{b \sum X_i}{n} \text{ or } \bar{Y} = a + b\bar{X}$$

Transforming equation (1.1), we get:

$$(2.4) \quad Y_i = a + bX_i + e_i$$

Substituting (2.3) and (2.4) into (2.1),

$$(2.5) \quad y_i = (Y_i - \bar{Y}) = [(a + bX_i + e_i) - (a + b\bar{X})] = b(X_i - \bar{X}) + e_i$$
$$= bx_i + e_i$$

Thus,

$$(2.6) \quad e_i = (y_i - bx_i)$$

and

$$(2.7) \quad \sum e_i^2 = \sum (y_i - bx_i)^2$$

Minimizing (2.7), we get:

$$\text{Ý}2.8\text{þ} \quad \frac{/e_i^2}{/b} = ?2 > \text{Ý}y_i ? bx_i\text{þ}x_i = 0$$

$$\text{Ý}2.9\text{þ} \quad > y_i x_i ? b > x_i^2 = 0$$

which is equivalent to:

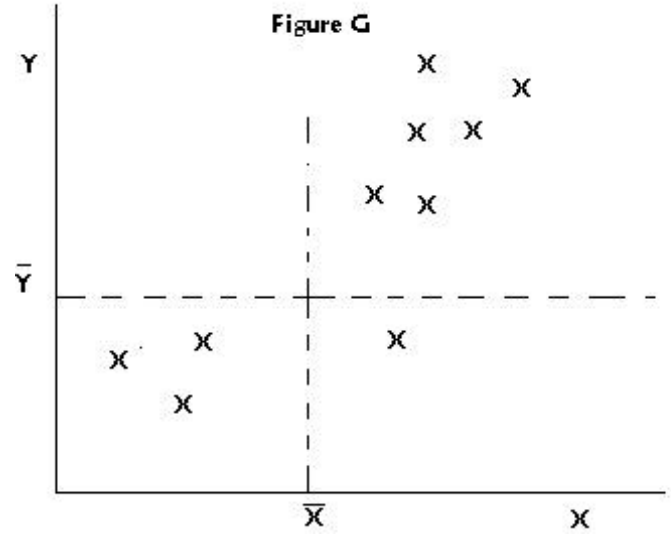
$$\text{Ý}2.10\text{þ} \quad > e_i x_i = > y_i x_i ? b > x_i^2 = 0$$

Solving for b:

$$\text{Ý}2.11\text{þ} \quad \text{b}= \frac{> y_i x_i}{> x_i^2}$$

Of course equation (2.10) is equivalent to equation (1.5) and equation (2.11) is equivalent to equation (1.8).

Note that in deviation form, the parameter a disappears from the regression line. This is because in transforming the observations to the deviation-from-the-mean form, we have in effect shifted the axis from the point (0,0) to the point, $(\overline{X}, \overline{Y})$ as indicated in Figure G.



v

Therefore in the deviation form of the regression line, the intercept, a, is by definition equal to zero.

These expressions in deviation terms can be written in terms of moments about means, and the moment notation sometimes simplifies presentation. The first moment about the mean is defined as:

$$M_x = \sum \left(X_i - \bar{X}\right) = \sum x_i \qquad (2.12)$$

The second moment for a single variable distribution is:

$$M_{xx} = \sum \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right) = \sum x_i^2 \qquad (2.13)$$

The second moment for a joint distribution of two variables X and Z is defined as:

$$M_{xz} = \sum \left(X_i - \bar{X}\right)\left(Z_i - \bar{Z}\right) = \sum x_i z_i \qquad (2.14)$$

We can rewrite equation (2.9) in moment notation as:

$$M_{yz} - bM_{xx} = 0 \qquad (2.15)$$

and equation (2.11) can be rewritten:

$$b = \frac{M_{xz}}{M_{xx}} \qquad (2.16)$$

## III. Multiple Regression

Thus far, we have developed our discussion of regression analysis around the idea of getting an estimator of a particular variable Y which will give us the most likely value for any given observation. We noted that while the mean is a good estimator, if we have some information about another characteristic, X, associated with a given observation, i, we can do better in predicting Y by using the conditional mean $E(Y \mid X_i)$, and we approximate this with a simple linear regression equation. By the same logic, however, if we know more than one characteristic, say $X_1$ and $X_2$, we should be able to do even better in

predicting Y by estimating the conditional expectation E(Y | $X_{1i}, X_{2i}$). This is one reason for going on to develop multiple regression analysis.

Another reason we may be interested in regression analysis is that we are interested in the impact of a given independent variable, say $X_1$, on the dependent variable Y. Thus, our interest focuses on equation (0.10) for example, not so much of E(Y | X) as on the slope coefficient b which tells us how a change in X of one unit is likely to change Y. If we are interested only in the effects on $X_1$ on Y it might seem that the simple regression we have already developed might be sufficient, but this is not so. The reason it is likely to be insufficient is that other factors, say $X_2$, which are systematically related to Y, may also be related to $X_1$. If this is the case, the effect we estimated by a simple regression of Y on $X_1$ alone may be misleading. $X_1$ may be "taking credit for" some of the effects on Y which are really due to another factor, $X_2$, with which it is partially related. For example, suppose Y is test scores of students in elementary schools and we are interested in the effect of per pupil expenditures $X_1$ on test scores. A simple linear regression would give us an estimate, e.g., b in equation (2.5), of the effect on test scores of a dollar increase in per pupil expenditure. Would this be a reliable estimate of how much test scores would be likely to rise if we raise expenditures one dollar? Well, we also have a general impression that children's test scores are related to family background, say as measured by family income, which we will call $X_2$. We also know that since schools are financed by the property tax, communities with higher than average family incomes are likely to have higher than average expenditures per pupil. That means $X_1$, per pupil expenditure, and $X_2$, family income, are likely to be positively related. How can we be sure that when we estimate the simple regression of Y, test scores, and $X_1$, per pupil expenditure, we are not really getting an estimate of the relation of $X_2$, family income, to test scores?

We would wish to separate the effects of $X_1$ on Y from the effects of $X_2$ (and other independent variables) on Y, in order to be able to estimate the "true" net effect of $X_1$ on Y. Since the problem is that $X_1$ and $X_2$ (or other independent variables) are interrelated, it would seem logical to first take account of the relationship of $X_1$ to $X_2$ (and others) and then take that part of $X_1$ that is not related to $X_2$ (or others) and see what effect it has on Y. This is the essential logic we will use to

develop multiple regression estimates. (Then we will come back and derive them by the least-squared error criterion to get and show the estimators are the same).

Let us take the case of a joint distribution of three variables: $Y_i$, which we'll treat as the dependent variable, and $X_{1i}$ and $X_{2i}$iWe wish to estimate the linear multiple regression function:

$$Y_i = a + b_{y_{1.2}} X_{1i} + b_{y_{2.1}} X_{2i} + e_{y.12i}$$

which we rewrite in deviation form:

$$Ý3.1Þ \quad Y_i = b_{y_{1.2}} X_{1i} + b_{y_{2.1}} X_{2i} + e_{y.12i}$$

In the notation used here, the subscripts before the dot indicate the relationship of the coefficient represents and the subscripts after the dot indicate the other variables controlled for elsewhere in the estimating equation. For example, $b_{y_{1.2}}$, is the independent, or "net", effect of $x_1$ on y, controlling for the influence of $x_2$ on y. For the error term , the notation indicates that this is the estimated error in y for observation i once we have allowed for the estimated systematic effect of $x_{1i}$ and $x_{2i}$ on y.

Now, following the logic sketched out above, we first take account of any systematic relationship between the independent variables by estimating a simple regression relating $x_1$ and $x_2$.

$$Ý3.2Þ \quad x_{1i} = b_{12} x_{2i} + e_{1.2i}$$

We will call this equation the auxiliary regression. Following our notational convention, $b_{12}$ is the estimated relationship between $x_1$ and $x_2$ (since no other variables enter the relationship in (3.2) there is no dot in the subscript for b).

From the formulae (2.11) and (2.16) developed for simple regression, recalling that in this case $x_1$ is the dependent variable and $x_2$ is the independent variable, we can write down direct the expression for $b_{12}$ in (3.2):

$$Ý3.3Þ \quad b_{12} = \frac{> x_{1i} x_{2i}}{> x_{2i}^2} = \frac{M_{x_1 x_2}}{M_{x_2 x_2}}$$

and then rewrite (3.2) as :

$$(3.4) \quad e_{1.2i} = x_{1i} - b_{12}x_{2i} = x_{1i} - \hat{x}_{1i}$$

We see that $e_{1.2i}$ is the residual value of $x_1$ for observation i after we subtract out $\hat{x}_{1i}$, that part of $x_{1i}$ which is systematically related to $x_{2i}$.

Now we can treat $e_{1.2}$ itself as a variable which varies from one observation to another. To facilitate understanding of this use of $e_{1.2}$ as a variable, it may be useful to think of the basic data array using hypothetical numbers written in a table as follows:

| *ObservationNumber* | $Y$ | $X_1$ | $X_2$ |
|---|---|---|---|
| 1 | 79.0 | 19.5 | 83.1 |
| 2 | 71.0 | 18.0 | 74.4 |
| 3 | 61.3 | 14.7 | 63.8 |
| 4 | 49.3 | 11.4 | 48.7 |
| 5 | 51.9 | 12.2 | 52.0 |

If we estimated (3.2) for this data, we obtain:

$$x_{1i} = .24x_{2i} + e_{1.2i}$$

Using this relationship in equation (3.4), we can generate $e_{1.2i}$ values and have a new data array:

| *ObservationNumber* | $Y$ | $X_1$ | $X_2$ | $e_{1.2}$ |
|---|---|---|---|---|
| 1 | 79.0 | 19.5 | 83.1 | -.03 |
| 2 | 71.0 | 18.0 | 74.4 | .55 |
| 3 | 61.3 | 14.7 | 63.8 | -.61 |
| 4 | 49.3 | 11.4 | 48.7 | .12 |
| 5 | 51.9 | 12.2 | 52.0 | .13 |

$e_{1.2i}$ gives us the sort of variable we seek to estimate the net effect of $x_1$ on Y since it represents variation in $x_1$ across observations which is unrelated to variation in $x_2$ across observations. $e_{1.2i}$ is sometimes referred to as the orthogonal part of $x_1$.

Using e1.2, we can now estimate a simple regression relating it to y:

$$\text{Ý3.5þ} \quad b_{y_{1.2i}} e_{1.2i} + e_{y\text{Ý}1.2\text{þ}i}$$

We use the parenthesis in the subscript to remind us that we are using a variable developed via the auxiliary regression (3.2).

Once again, applying directly the simple regression formulae (2.11) and (2.16), we can write down the expression for the coefficient $b_{y_{1.2}}$:

$$\text{Ý3.6þ} \quad b_{y\text{Ý}1.2\text{þ}} = \frac{> y_i e_{1.2i}}{> e_{1.2i}^2} = \frac{M_{y,e_{1.2}}}{M_{e_{1.2},e_{1.2}}}$$

$b_{y\text{Ý}1.2\text{þ}}$ is an estimate of the relationship of y to $x_1$ net of any indirect influence on Y of $x_2$ operating through $x_2$'s relationship to $x_1$. This is because in creating $e_{1.2i}$ we purged $x_1$ of a part systematically related to $x_2$.

Now we do some rewriting of (3.6) to get it into a form readily comparable to the usual textbook formula for a multiple regression coefficient.

First, rewrite the numerator of (3.6) using (3.4)

$$\text{Ý3.7þ} \quad > y_i e_{1.2i} = > y_i \text{Ý} x_{1i} ? b_{12} x_{2i} \text{þ}$$

$$\text{Ý3.7}a\text{þ} \quad = > y_i x_{1i} ? b_{12} > y_i x_{2i}$$

Substitute (3.3) for $b_{12}$:

$$\text{Ý3.7}b\text{þ} \quad = > y_i x_{1i} ? \frac{> x_{1i} x_{2i} > y_i x_{2i}}{> x_{2i}^2}$$

In moment notation:

$$\text{Ý3.7}c\text{þ} \quad = M_{yx_1} ? \frac{M_{x_1 x_2} M_{yx_2}}{M_{x_c x_2}}$$

Now, rewrite the denominator of (3.6) using (3.4):

$$(3.8) \quad \sum e_{1.2}^2 = \sum (x_{1i} - b_{12}x_{2i})^2$$

$$= \sum x_{1i}^2 - 2b_{12} \sum x_{1i}x_{2i} + b_{12}^2 \sum x_{2i}^2$$

In moment notation:

$$(3.8a) \quad = M_{x_1x_1} - 2b_{12}M_{x_1x_2} + b_{12}M_{x_2x_2}$$

Substitute (3.3) for b12:

$$(3.8b) \quad = M_{x_1x_1} - \frac{2M_{x_1x_2}M_{x_1x_2}}{M_{x_2x_2}} + \frac{M_{x_1x_2}^2 M_{x_2x_2}}{M_{x_2x_2}}$$

$$= M_{x_1x_1} - \frac{M_{x_1x_2}^2}{M_{x_2x_2}}$$

Now substitute into (3.6) the expression for the numerator (3.7b) and the expression for the denominator (3.8b) to get:

$$(3.9) \quad = M_{yx_1} - \frac{\frac{M_{x_1x_2}M_{yx_2}}{M_{x_2x_2}}}{M_{x_1x_2} - \frac{M_{x_1x_2}^2}{M_{x_2x_2}}} = \frac{M_{yx_1}M_{x_2x_2} - M_{x_1x_2}M_{yx}}{M_{x_1x_2}M_{x_2x_2} - M_{x_1x_2}^2}$$

We have developed an expression for the relationship of y and $x_1$ by "netting out" or controlling for $x_2$ through the auxiliary regression. Now we derive an expression for the multiple regression coefficients $b_{y_{1.2}}$ in (3.1), using the alternative logic of minimizing the squared errors in the multiple regression equation (3.1). We transpose (3.1) to get $e_{y.12i}$ on the left hand side.

$$(3.10) \quad e_{y.12i} = y_i - b_{y_{1.2}}x_{1i} - b_{y_{2.1}}x_{2i}$$

Squaring and summing over i we get:

$$(3.11) \quad \sum e_{y.12i}^2 = \sum (y_i - b_{y_{1.2}}x_{1i} - b_{y_{2.1}}x_{2i})^2$$

Here we may proceed by analogy to the simple regression case and simply impose the conditions for the normal equations:

$$\sum e_{y.12i} = 0 \qquad \sum e_{y.12i}x_{1i} = 0 \qquad \sum e_{y.12i}x_{2i} = 0$$

Recall that the logic of these was that the first condition yielded the least error by forcing the line through the joint means of Y and X, the second assured that the regression equation did not have either too much positive or too much negative tilt in the $X_1$ plane, and the third assured us the same in the $X_2$ plane.

Or, alternatively, we can derive the normal equations by minimizing (3.11) with respect to the coefficients, obtaining:

$$(3.12a) \quad \frac{\partial \sum e_{1.2i}^2}{\partial b_{y_{1.2}}} = -2 \sum (y_i - b_{y_{1.2}}x_{1i} - b_{y_{2.1}}x_{2i})x_{1i} = 0$$

$$(3.12b) \quad \frac{\partial \sum e_{1.2i}^2}{\partial b_{y_{2.1}}} = -2 \sum (y_i - b_{y_{1.2}}x_{1i} - b_{y_{2.1}}x_{2i})x_{2i} = 0$$

which, after dividing through by -2, can be written:

$$(3.13a) \quad \sum (y_i - b_{y_{1.2}}x_{1i} - b_{y_{2.1}}x_{2i})x_{1i} = \sum e_{y.12i}x_{1i} = 0$$

$$(3.13b) \quad \sum (y_i - b_{y_{1.2}}x_{1i} - b_{y_{2.1}}x_{2i})x_{2i} = \sum e_{y.12i}x_{2i} = 0$$

In moment notation, these become:

$$(3.14a) \quad M_{yx_1} - b_{y_{1.2}}M_{x_1x_1} - b_{y_{2.1}}M_{x_1x_2} = 0$$

$$(3.14b) \quad M_{yx_2} - b_{y_{1.2}}M_{x_1x_2} - b_{y_{2.1}}M_{x_2x_2} = 0$$

To solve (3.14a) and (3.14b) simultaneously, we multiply (3.14a) by $M_{x_2x_2}$ and (3.14b) by $M_{x_1x_2}$ and subtract the latter from the former:

$$(3.15) \quad (M_{x_1x_1}M_{x_2x_2} - M_{x_1x_2}^2)b_{y_{1.2}} = (M_{yx_1}M_{x_2x_2} - M_{x_1x_2}M_{yx_2})$$

or

$$(3.16) \quad b_{y_{1.2}} = \frac{M_{yx_1}M_{x_2x_2} - M_{x_1x_2}M_{yx_2}}{M_{x_1x_1}M_{x_2x_2} - M_{x_1x_2}^2}$$

which is the same as (3.9), so

$$b_{y_{1.2}} = b_{y\hat{y}1.2\flat}$$

We derived $b_{y\hat{y}1.2\flat}$ by using the auxiliary regression (3.2) to obtain that part of $x_1$ which was uncorrelated with $x_2$, labeled $e_{1.2}$, and then running a simple regression between y and $e_{1.2}$, yielding the expression for $b_{y\hat{y}1.2\flat}$ given in (3.9).

We derived $b_{y_{1.2}}$ by a different route, using the same logic we had used in deriving the mean as the estimator which gave us the least error and the simple regression coefficients $b_{yx}$ as the conditional estimator of y given x which gave us the least error. Since in a multiple regression equation, such as (3.1) we have more than one independent variable, and therefore more coefficients to estimate, we have to add to the normal equations one more equation for each additional coefficient. In the case of (3.1), we have the normal equations
$\sum e_i = 0, \sum e_i x_{1i} = 0, \sum e_i x_{2i} = 0,$ which are analogous to the normal equations for simple regression [equations (1.2) and (1.5)]. Solving these normal equations (3.13a, b) simultaneously, we obtained the expression for $b_{y_{1.2}}$ , (3.16).

Comparing the expressions for $b_{y\hat{y}1.2\flat}$ , (3.9), and $b_{y_{1.2}}$, (3.16), we found they are the same. This shows that the auxiliary regression logic and the least squared error logic lead to the same estimator for the dependence of y on $x_1$ which is uncorrelated with $x_2$.

We can emphasize this concept of estimating the "net effect" of $x_1$ and y with a few more manipulations.

Consider the simple regression between y and $x_1$:

$$\text{Ý3.17Þ} \quad y_i = b_{y_1} x_{1i} + e_{y_{.1i}}$$

Using the expression for the simple regression coefficient [equation (2.11)], we get:

$$\text{Ý3.18Þ} \qquad b_{y_1} = \frac{\sum y_i x_{1i}}{\sum x_{1i}^2} = \frac{M_{yx_1}}{M_{x_1 x_1}}$$

Similarly, from the auxiliary regression (3.2) we obtained:

$$(3.19) \quad b_{21} = \frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2} = \frac{M_{x_1x_2}}{M_{x_1x_1}}$$

We can rewrite equation (3.14a) as follows:

$$(3.19a) \quad b_{y_{1.2}} = \frac{M_{yx_1} - b_{y_{2.1}}M_{x_1x_2}}{M_{x_1x_1}} = \frac{M_{yx_1}}{M_{x_1x_1}} - b_{y_{2.1}}\frac{M_{x_1x_2}}{M_{x_1x_1}}$$

Substituting into (3.19a), using the moment expressions from (3.18), and (3.19):

$$(3.20) \quad b_{y_{1.2}} = b_{y_1} - b_{21}b_{y_{2.1}}$$

Thus the multiple regression coefficient indicating the "net effect" of x1 on y is equal to the "gross effect" of $x_1$ on y, as estimated by the simple regression coefficient , minus the relation between $x_1$ and $x_2$, as indicated by $b_{21}$, times the "net effect" of $x_2$ on y, as indicated by $b_{y_{2.1}}$. The "net effect" of $x_1$ on y, $b_{y_{1.2}}$, will differ from the "gross effect" of $x_1$ on y,$b_{y_1}$, by a greater amount the larger is the interrelation of $x_1$ and $x_2$,$b_{12}$, and the larger the "net effect" of $x_2$ on y, $b_{y_{2.1}}$. If $x_1$ and $x_2$ are not closely related or $x_2$ has little effect on , the "net" and "gross" effects of $x_2$ and y will differ little. This is an indication of the degree to which the multiple regression is an "improvement" on the simple regression in attempting to estimate the independent effects of $x_1$ and y. The multiple regression "controls for" the effects on y of $x_2$ which might be "operating through $x_1$" due to the inter-correlation of $x_1$ and $x_2$. The simple regression, by "omitting" the variable $x_2$ in estimating the relation of y and $x_1$, may lead to a bias in the estimate of that relation. We will return later in a more general discussion of bias to this expression for the extent of omitted variable bias.

# IV. Variance, Covariance, Coefficient of Determination (R2) and Correlation (r)

(Review Beals E(x) p. 61, 86; $a_x^2$ p. 52, 62, 63; $a_{xy}$ p. 88,89)

We began our discussion of regression analysis by focusing on the problem of finding a "best estimator" E(y) for a variable Y, given that

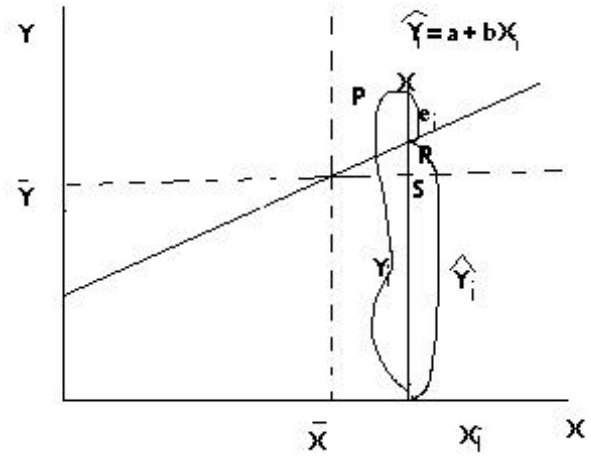
Y had a frequency distribution over a range of values. We showed that the mean was the value for E(Y) which gave the least sum of errors. It was shown it was the value for E(Y) which minimized the sum of squared errors, $\sum e_i^2$ . This sum of squared errors is a measure of the dispersion of Y around its mean. If we divide the sum of squared errors by n, we have a measure called the variance of Y.

$$\sigma_y^2 = \frac{\sum (Y_i - \bar{Y})^2}{n}$$

When we proceed to the discussion of simple regression, we noted we could "improve" our estimator of Y by developing the conditional mean $E(Y \mid X)$. This "improvement" should mean that the sum of squared errors is less when we use $E(Y \mid X)$ rather than E(Y).

$$(4.1) \qquad \sum [Y_i - E(Y \mid X)]^2 < \sum [Y_i - E(Y)]^2$$

Graphically, we can illustrate the division of $Y_i$ into three parts:




At $X_1$, the observed $Y_i$, (PT) can be divided into the predicted part, $Y_i$, (TR) and the error, $E_i$ (RP).

$$(4.2) \qquad Y_i = \hat{Y}_i + E_i$$

Subtract $\bar{Y}$ from both sides to put the expression in deviation form:

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + E_i \qquad (4.3)$$

$$y_i = y + e_i \qquad (PS = RS + PR) \qquad (4.3a)$$

Square both sides and sum:

$$\sum y_i^2 = \sum y_i^2 + 2\sum y_i e_i + \sum e_i^2 \qquad (4.4)$$

$$y = bx_i \qquad (4.5)$$

Therefore:

$$\sum y_i e_i = \sum b x_i e_i = b \sum x_i e_i \qquad (4.6)$$

From the normal equations (2.10), we know $\sum x_i e_i = 0$, therefore:

$$\sum y_i e_i = b \sum x_i e_i = 0 \qquad (4.6b)$$

Substituting (4.6a) into (4.4):

$$\sum y_i^2 = \sum y_i^2 + \sum e_i^2 \qquad (4.7)$$

Total sum of Squares = Regression sos + Error sos

$$(SST) = (SSR) + (SSE)$$

Returning to equation (4.1), note that
$(Y_i - \hat{Y}) = (Y_i - \bar{Y}) - (\hat{Y} - \bar{Y}) = (Y_i - \bar{Y}) - b(X - \bar{X})$.

$$\sum (Y_i - \hat{Y})^2 = \sum \left(Y_i - \bar{Y}\right)^2 = \sum e_i^2 \qquad (4.8)$$

From (4.7), we
see $\sum (Y_i - \hat{Y})^2 = \sum \left(Y_i - \bar{Y}\right)^2 = \sum y_i^2 = \sum y_i^2 + \sum e_i^2$
Therefore:

$$\sum (Y_i - \hat{Y})^2 = \sum \left(Y_i - \bar{Y}\right)^2 = \sum e_i^2 < \sum (Y_i - \hat{Y})^2 = \sum y_i^2 + \sum e_i^2$$

Thus we have shown that using the conditional mean from the regression line as our estimator reduces the sum of squared errors.

Now, returning to (4.7), divide both sides by $\sum y_i^2$ :

From (4.7) and (4.9) we can see that we can partition the total sum of squares of Y into a portion "predicted" by the linear regression $\frac{\sum y_i^2}{\sum y_i^2}$ and a residual, sum of errors squared $\frac{\sum e_i^2}{\sum y_i^2}$ .

The "predicted" portion is called the coefficient of determination and is usually denoted by $R^2$.

$$(4.10) \quad R^2 = \frac{\sum y_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

Going back to (4.5) and squaring and summing, we get:

$$(4.11) \quad \sum y_i^2 = b^2 \sum x_i^2$$

Substituting (4.11) into (4.10):

$$(4.12) \quad R^2 = \frac{b^2 \sum x_i^2}{\sum y_i^2}$$

Substituting for b from equation (2.11):

$$(4.12a) \quad R^2 = \frac{(\sum y_i x_i)^2}{(\sum x_i^2)^2} \cdot \frac{\sum x_i^2}{\sum y_i^2} = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

Take the square root:

$$r = \sqrt{R^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

We define the covariance of y and x as:

$$\sigma_{yx} = \sum \frac{(Y_i - \bar{Y})(X_i - \bar{X})}{n} = \frac{\sum x_i y_i}{n}$$

Thus, we can rewrite (4.13):

$$(4.13a) \quad r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{\sigma_{yx}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{cov(y,x)}{\sqrt{Var(y) Var(x)}}$$

r is called the correlation coefficient.

We can see that both $R^2$ and r can take on values between 0 and 1. Values close to 0 indicate that there is little linear relationship between Y and X, values close to 1 indicate that Y and X are closely related.

We often refer to $R^2$ as the "portion of variance explained", meaning the portion of variance in Y "explained by" the linear regression of Y on X. $R^2$ is also referred to as a measure of the "goodness of fit" of the regression line; if $R^2$ is high, the regression fits the data points well, as the errors around the regression line are small.

We can derive the expression for the coefficient of determination for a multiple regression equation. Starting from equation (3.1), we derive $R^2_{y.12}$ , with the subscripts to indicate the two independent variables $x_1$ and $x_2$.

We denote the predicted value from the multiple regression:

$$y_{.12i} = b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i} \qquad (4.14)$$

Multiply (4.14) by $y_i$ and sum over i:

$$\sum y_{.12i} y_i = \sum y_{.12i} \left( y_{.12i} + e_{y_{.12i}} \right) = \sum y^2_{.12i} + \sum y_{.12i} e_{y_{.12i}} \qquad (4.15)$$

Substitute for $y_{.12i}$ in the second term:

$$= \sum y^2_{.12i} + \sum \left( b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i} \right) e_{y_{.12i}}$$

$$= \sum y^2_{.12i} + b_{y_{1.2}} \sum x_{1i} e_{y_{.12i}} + b_{y_{2.1}} \sum x_{2i} e_{y_{.12i}}$$

Since [from (3.13a, b)] $\sum x_{1i} e_{y_{.12i}} = 0$, $\sum x_{2i} e_{y_{.12i}} = 0$

$$= \sum y^2_{.12i}$$

By analogy to the derivation of (4.7), we can write down without derivation:

$$\sum y_i^2 = \sum y_{.12i}^2 + \sum e_{.12i}^2 \qquad (4.16)$$

$$1 = \frac{\sum y_{.12i}^2}{\sum y_i^2} + \frac{\sum e_{.12i}^2}{\sum y_i^2} \qquad (4.17)$$

$$R^2 = \frac{\sum y_{.12i}^2}{\sum y_i^2} = 1 - \frac{\sum e_{.12i}^2}{\sum y_i^2} \qquad (4.18)$$

Using (4.15), we can write:

$$\sum y_{.12i}^2 = \sum y_{.12i} y_i = \sum (b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i}) y_i \qquad (4.19)$$
$$= b_{y_{1.2}} \sum x_{1i} y_i + b_{y_{2.1}} \sum x_{2i} y_i$$

Substituting (4.19) into the numerator of (4.18), we get:

$$R_{y.12i}^2 = \frac{b_{y_{1.2}} \sum x_{1i} y_i + b_{y_{2.1}} \sum x_{2i} y_i}{\sum y_i^2} \qquad (4.20)$$

While dealing with variances, we can develop the expression for the variance of a sum of two variables (see Beals, p. 86-89). Suppose we are interested in X + Y, then:

$$E(X + Y) = \left(\overline{X + Y}\right) = \frac{\sum (X + Y)_i}{n} = \frac{\sum X_i}{n} + \frac{\sum Y_i}{n} = \bar{X} + \bar{Y} \qquad (4.21)$$

Treating X + Y as a single variable, we can write down the variance in X + Y:

$$a_{X+Y}^2 = \frac{\sum \left[ (X + Y)_i - \left(\overline{X + Y}\right) \right]^2}{n} \qquad (4.22)$$

Expanding the numerator and substituting from (4.21):

$$\sum \left[ (X + Y)_i - \left(\overline{X + Y}\right) \right]^2 = \sum \left[ (X + Y)_i - (\bar{X} + \bar{Y}) \right]^2 = \sum \left[ (X_i - \bar{X}) + (Y_i - \bar{Y}) \right]^2 \qquad (4.23)$$
$$= \sum \left[ (X_i - \bar{X})^2 + 2(X_i - \bar{X})(Y_i - \bar{Y}) + (Y_i - \bar{Y})^2 \right]$$
$$= \sum x_i^2 + 2 \sum x_i y_i + \sum y_i^2$$

We can then rewrite (4.22):

$$\text{(4.24)} \qquad a^2_{X+Y} = \frac{\sum x_i^2 + 2\sum x_i y_i + \sum y_i^2}{n} = \frac{\sum x_i^2}{n} + \frac{\sum y_i^2}{n} + \frac{2\sum x_i y_i}{n}$$
$$= a^2_X + a^2_Y + 2a_{XY}$$

# V. Tests of Hypothesis: Significance of Difference in Group

Means; Variance in b

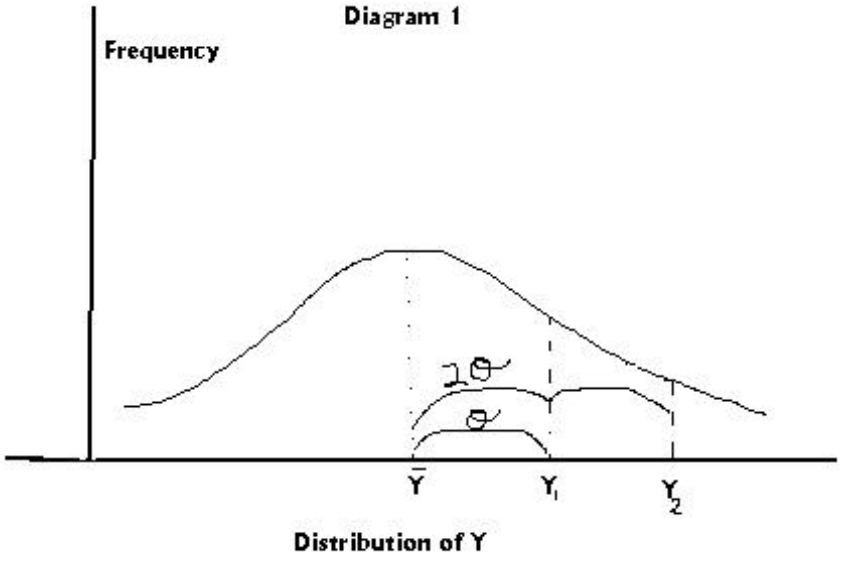
(Review Beals pp.179-99, pp.123-5, pp.235-43, pp.245-7, pp.251-7)

So far we have been concerned with developing best estimates for the expected value of Y, E(Y), or the best estimate of the conditional expectation of Y, E(Y| X), or E(Y| $X_1, X_2, X_3, \ldots$) . We have been ignoring the fact that we form these estimates on the basis of a sample drawn from the population frequency distribution of Y or of the joint frequency distribution of Y and X or Y and $X_1$,$X_2$,... Now we wish to take into account the difference between the estimate, based on the sample, of say E(Y) and the true value of E(Y) in the population. In testing hypotheses, we explicitly take account of the fact that various samples will yield somewhat different estimates for the value of E(Y).

In this section, our basic objective is to derive the expression for the variance of the simple regression coefficient, b, and to show how that can be used to test hypotheses about the relationship between Y and X. We will derive the expression for the variance of b by two different routes: first [equations (5.1) to (5.23)] by developing the logic of the test for significant difference between two group means and showing how that relates to the variance of b; second [equations (7.6) to (7.11)] by deriving the variance of b directly from the expression for b given in (3.6). Following these two routes is rather tedious, so it is important continually to refer back to this statement about our basic objective.

## A. Tests of Significance of Difference in Group Means

If it is assumed that the variable Y has a frequency distribution which is described by a normal curve, then a good deal about the distribution can be stated in terms of its mean, $\bar{Y}$ , and its standard deviation, $a_y$. Diagram 1 illustrates the normal distribution:




If the distance $(Y_1 - \bar{Y}) = a$ , then the area under the normal curve between $Y_1$ and $\bar{Y}$ is 34.13 percent of the total area under the curve; i.e. 34.13 percent of the values of Y in the distribution will have a value between $Y_1$ and $\bar{Y}$ . If the distance $(Y_2 - \bar{Y}) = 2a$, then the area under the curve between $Y_2$ and $\bar{Y}$ is 47.73 percent of the total area under the curve. Thus, values of $a$ on either side of $\bar{Y}$ cover 68.26 percent of all cases and values of $2a$ on either side of $\bar{Y}$ cover 95.46 percent of values in the distribution. If a value $Y_i$, which is greater than $2a$ Ý+*or* ?Þaway from $\bar{Y}$ is observed, there is only a 4.5 percent chance that it is part of the same underlying distribution.

In summary, if the distribution of Y is normal, we can say that when the absolute value of $\frac{Y_i - \bar{Y}}{a}$ is calculated, and is found to be

greater than $2a$ or ? away, then there is only a 4.54 percent chance that the value $Y_i$ is part of the distribution which has mean $\bar{Y}$ and standard deviation $a$. (Stated obversely, there is a 95.46 percent chance it is not from the same distribution). The interval $\bar{Y} \pm 1.96a$ is commonly called the 95 percent confidence interval.

It is useful here to develop the expression for the variance of the mean. Since $\bar{Y} = \frac{\sum Y_i}{n}$ and each of the $Y_i$ in a given sample are independent and have the same variance, $a_y^2$,

$$(5.1) \quad a_{\bar{y}}^2 = var(\bar{Y}) = var\left[\frac{1}{n}(Y_1 + Y_2 + ... + Y_n)\right]$$
$$= \left(\frac{1}{n}\right)^2 var(Y_1 + Y_2 + ... + Y_n)$$

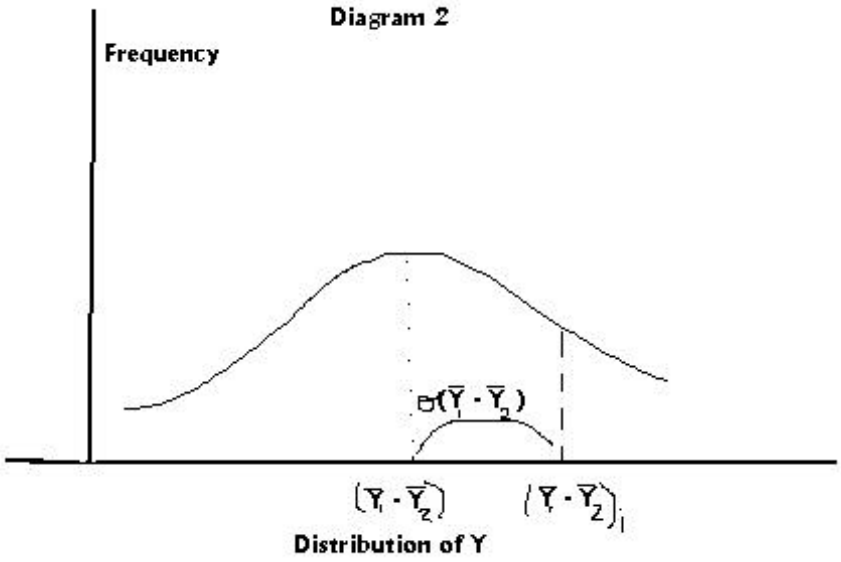
From (4.24), we have the expression for variance of a sum. Since the $Y_i$ are independent, the covariances of the $Y_i$'s are zero. Therefore:

$$(5.1a) \quad a_{\bar{y}}^2 = \left(\frac{1}{n}\right)^2 [var(y) + var(y_2) + ... + var(y_n)]$$
$$= \left(\frac{1}{n}\right)^2 [a_y^2 + a_y^2 + ... + a_y^2]$$
$$= \left(\frac{1}{n}\right)^2 (n a_y^2)$$
$$= \frac{a_y^2}{n}$$

Assume there are two groups. Members of group 1 have received "no treatment" so are called "controls". Members of group 2 receive a "treatment", so they are called "experimentals". Y is the response measure.

Consider now for the relationship between the mean of the response variable for the control, $\bar{Y}_1$, and the mean for the response variable for the experimental group, $\bar{Y}_2$. Take the difference between these two means, $(\bar{Y}_1 - \bar{Y}_2)$. These groups are two samples and the question is whether they are drawn from the same population (which has a frequency distribution of variable Y) or from different populations ($Y_1$ and $Y_2$ with different frequency distributions for Y). If they are drawn from the same response population, then the "treatment" had no effect. We can think of drawing repeated two-group samples (e.g., repeated experiments). Then the means

calculated, $\bar{Y}_1$ and $\bar{Y}_2$, would differ somewhat in successive samples. Thus, the sample means would themselves have a frequency distribution, and the difference between the means $(\bar{Y}_1 - \bar{Y}_2)$, would also have a frequency distribution. Assume that this distribution is normal, with its mean, $\left(\overline{\bar{Y}_1 - \bar{Y}_2}\right)$ , and its standard deviation, $\sigma(\bar{Y}_1 - \bar{Y}_2)$.




Now, for this distribution, statements can be made for certain sample values, $(\bar{Y}_1 - \bar{Y}_2)_i$, of the difference between the two means. If:

$$\left| \frac{(\bar{Y}_1 - \bar{Y}_2)_i - \left(\overline{\bar{Y}_1 - \bar{Y}_2}\right)}{\sigma(\bar{Y}_1 - \bar{Y}_2)} \right| \geq 2$$

then there is a 95.46 percent chance that $(\bar{Y}_1 - \bar{Y}_2)_i$ is not from the distribution with mean $\left(\overline{\bar{Y}_1 - \bar{Y}_2}\right)$ (or in other words, we would only be wrong 4.54 percent of the time if we guessed it was not from the distribution).

If the two sample groups are really drawn from the same underlying population distribution, then there should be no difference in their means: i.e., $\bar{Y}_1 = \bar{Y}_2$, and $\overline{\bar{Y}_1 - \bar{Y}_2} = 0$. Therefore, the appropriate test is to set $\left(\bar{Y}_1 - \bar{Y}_2\right)$ equal to zero and calculate the absolute value,

$$(5.2a) \qquad \left| \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)_i - 0}{\sigma\left(\bar{Y}_1 - \bar{Y}_2\right)} \right|$$

If it is greater than 2, then we say the difference in the means is significant at the 95 percent level, i.e. there is only a 4.54 percent chance that a difference in sample means of this size, $\left(\bar{Y}_1 - \bar{Y}_2\right)_i$, could be observed if the true difference in means were zero. Another way of stating this is that if we had numerous successive two-group samples (control and experimental) of the same size, only 4.54 percent of these samples would have a difference in the means $\left(\bar{Y}_{control} - \bar{Y}_{experimental}\right)_i$ that was as great as, or greater than, the value $\left(\bar{Y}_1 - \bar{Y}_2\right)_i$.

With the observed control group and experimental group, one is able to calculate $\bar{Y}_1$, and $\bar{Y}_2$, and the $\left(\bar{Y}_1 - \bar{Y}_2\right)_i$, for the above expression. All that remains to be done, then, is to obtain a value for the statistic given above as $\sigma\left(\bar{Y}_1 - \bar{Y}_2\right)$.

Given that the two groups are drawn independently, $\bar{Y}_1$, and $\bar{Y}_2$ are independent random variables each with a variance, $\sigma^2_{\bar{y}_1}$ and $\sigma^2_{\bar{y}_2}$. Then $\bar{Y}_1 - \bar{Y}_2$ is the sum of two random variables and has a variance, $\sigma^2_{\left(\bar{Y}_1 - \bar{Y}_2\right)}$. Again, using the expression for the variance of a sum of random variables (4.24), and noting that the covariance of $\bar{Y}_1$, and $\bar{Y}_2$ is zero since they are independent, we get:

$$(5.3) \qquad \sigma^2_{\left(\bar{Y}_1 - \bar{Y}_2\right)} = \sigma^2_{\bar{y}_1} + \sigma^2_{\bar{y}_2}$$

Substituting from (5.1a) for the variance of the mean, we get:

$$(5.3a) \qquad \sigma^2_{\bar{Y}_1 - \bar{Y}_2} = \frac{\sigma^2_{y_1}}{n_1} + \frac{\sigma^2_{y_2}}{n_2}$$

$$(5.3b) \qquad \sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\left( \frac{\sigma^2_{y_1}}{n_1} + \frac{\sigma^2_{y_2}}{n_2} \right)}$$

Now we need to obtain from the sample data an estimate of $\sigma^2_{\bar{y}_1}$ and $\sigma^2_{\bar{y}_1}$ and we'll indicate that estimate by a $\hat{}$ (hat) over the expression.

$$(5.4) \qquad \hat{\sigma}^2_{\bar{y}_1} = \frac{\sum_i \left(Y_{1i} - \bar{Y}_1\right)^2}{n_1 - 1} \qquad \hat{\sigma}^2_{\bar{y}_2} = \frac{\sum_i \left(Y_{2i} - \bar{Y}_2\right)^2}{n_2 - 1}$$

Substituting in (5.3b), we get:

$$(5.5) \qquad \hat{\sigma}_{\bar{y}_1 - \bar{y}_2} = \sqrt{\left( \frac{\hat{\sigma}^2_{\bar{y}_1}}{n_1} + \frac{\hat{\sigma}^2_{\bar{y}_2}}{n_2} \right)} = \sqrt{\frac{\frac{\sum_i (Y_{1i} - \bar{Y}_1)^2}{n_1 - 1}}{n_1} + \frac{\frac{\sum_i (Y_{2i} - \bar{Y}_2)^2}{n_2 - 1}}{n_2}}$$

which can be substituted into (5.2) to get the appropriate test statistic.

## B. Difference in Means as a Dummy Variable Regression

To compare two groups in a simple regression, first, form the following dummy variable:

$$(5.6) \qquad D_i = \left\{ \frac{1 \text{ if a member of group 1}}{0 \text{ if a member of group 2}} \right\}$$

Let $n_1$ = number in group 1
$n_2$ = number in group 2
$N = n_1 + n_2$
Note that:

$$(5.7) \qquad \sum_N D_i = \sum_{n_1} 1 + \sum_{n_2} 0 = n_1$$

Now, for the characteristic we wish to compare across groups, Y, write the simple regression

$$Y_i = a + bD_i + e_i \qquad (5.8)$$

The normal equations for this simple regression are, by (1.2) and (1.6):

$$\sum_N e_i = \sum_N (Y_i - a - bD_i) = 0$$

$$= \sum_N Y_i - \sum_N a - b \sum_N D_i = 0$$

$$= \sum_{n_1+n_2} Y_i - (n_1 + n_2)a - bn_1 = 0 \qquad (5.9)$$

$$\sum_N e_i D_i = \sum_N (Y_i - a - bD_i) D_i = 0 \qquad (5.10)$$

Since all terms with $D_i = 0$drop out of this product sum, leaving only the $n_1$ terms where $D_i = 1$, we have:

$$\sum_N e_i D_i = \sum_{n_1} (Y_i - a - bD_i) = \sum_{n_1} Y_i - an_1 - bn_1 = 0 \qquad (5.10a)$$

Now, given both (5.9) and (5.10a) equal zero, we equate them:

$$\sum_{n_1+n_2} Y_i - (n_1 + n_2)a - bn_1 = \sum_{n_1} Y_i - an_1 - bn_1 = 0 \qquad (5.11)$$

Rearranging terms of (5.11) gives us:

$$\sum_{n_1+n_2} Y_i - \sum_{n_1} Y_i = (n_1 + n_2)a - an_1 - bn_1 + bn_1 = (n_1 + n_2) \qquad (5.12)$$

which gives:

$$\sum_{n_2} Y_i = n_2 a \qquad (5.12a)$$

$$a = \frac{\sum_{n_2} Y_i}{n_2} = \bar{Y}_2 \qquad (5.12b)$$

The constant term of the simple dummy variable regression equals the mean of Y for group 2 (the group "excluded" by the dummy variable).

Now, substitute (5.12b) for a in equation (5.10a):

$$\sum_{n_1} Y_i - n_1\left(\frac{\sum_{n_2} Y_i}{n_2}\right) - bn_1 = 0 \qquad (5.13)$$

Rearranging (5.13) gives us:

$$bn_1 = \sum_{n_1} Y_i - n_1\left(\frac{\sum_{n_2} Y_i}{n_2}\right) \qquad (5.13a)$$

Dividing both sides by n1 gives:

$$b = \frac{\sum_{n_1} Y_i}{n_1} - \left(\frac{\sum_{n_2} Y_i}{n_2}\right) = \bar{Y}_1 - \bar{Y}_2 \qquad (5.13b)$$

So the simple regression coefficient, b, in the simple dummy variable regression is the difference in means for the two groups. The intercept, a, is the mean for group two. We call group two the "excluded group" since it is the group for which $D_i = 0$.

Note that we use just one variable, $D_i$, to define the two groups. The simple regression, however, has two parameters, so we get from the two parameters the mean for each group as follows:

$$\bar{Y}_1 = a + b = \bar{Y}_2 + (\bar{Y}_1 - \bar{Y}_2) \qquad (5.D1)$$

$$\bar{Y}_2 = a = \bar{Y}_2 \qquad (5.D2)$$

Alternatively, we could formulate the dummy variable regression to represent the two group means as follows:

Define:

$$D_{1i} = \left\{ \frac{1 \text{ if group 1}}{0 \text{ otherwise}} \right\} \quad D_{2i} = \left\{ \frac{1 \text{ if group 2}}{0 \text{ otherwise}} \right\}$$

$$\sum_{n_1} D_{1i} = n_1 \qquad \sum_{n_1} D_{2i} = n_2$$

$$Y_i = b_1 D_{1i} + b_2 D_{2i} + e_i \tag{5D.3}$$

The normal equations for this regression are:

$$\sum_{n_1} e_i D_{1i} = \sum_{n_1} (Y - a - b_1 D_{1i} + b_2 D_{2i}) = 0 = \sum_{n_1} Y_i - b_1 n_1 \tag{5D.4}$$

$$\sum_{n_1} e_i D_{2i} = \sum_{n_1} (Y - a - b_1 D_{1i} + b_2 D_{2i}) = 0 = \sum_{n_1} Y_i - b_2 n_2 \tag{5D.5}$$

and, from that:

$$b_1 = \frac{\sum_{n_1} Y_i}{n_1} = \bar{Y}_1 \quad \text{and} \quad b_2 = \frac{\sum_{n2} Y_i}{n_2} = \bar{Y}_2$$

If, however, we try to estimate:

$$Y_i = a + b_1 D_{1i} + b_2 D_{2i} + e_i \tag{5D.6}$$

It will not work, i.e., we get no determinate solution for a, $b_1$, or $b_2$. This is an example of linear dependence. The two dummy variables representing only two groups yield two combinations: either $D_{1i} = 0$ and $D_{2i} = 1$, or $D_{1i} = 1$ and $D_{2i} = 0$ . The sum of squared errors, $\sum e_{1.2i}^2 = 0$ , since every observation comes from one of these two locations, and thus the regression line fits perfectly.

You must remember, therefore, that if the dummy variable equation has an intercept term, then there must always be an "excluded group", i.e., a group only defined by taking on a value 0 for the dummy variable. (Recall that equations written in deviation notation subsume an intercept term, e.g.$y_i = bD_i + e_i$, is equivalent to $y_i = a + bD_i + e_i$).

We can estimate regressions using dummy variables with trichotomous variables in one dimension. Suppose we define three

groups in terms of a single characteristic. For example, define three race groups : Black, Hispanic, and White (non-Hispanic, non-Black).

Define:

$$D_{1i} = \left\{ \frac{\text{1if black}}{\text{0otherwise}} \right\} \qquad D_{2i} = \left\{ \frac{\text{1 if hispanic}}{\text{0 otherwise}} \right\}$$

The size of the groups are: $N=n_1 + n_2 + n_3$ .

We can define group means in terms of the following dummy variable regression:

$$Y_i = a + b_1 D_{1i} + b_2 D_{2i} + e_i \qquad (5D.7)$$

Now we have defined White as the "excluded group." Form the normal equations for this regression:

$$(5D.8a) \qquad \sum_{N} e_i = \sum_{n_2} (Y - a - b_1 D_{1i} + b_2 D_{2i}) = \sum_{n_1} Y_i - b_1 n_1 - b_2 n_2 = 0$$

$$(5D.8b) \qquad \sum_{n_1} e_i D_{1i} = \sum_{n_1} (Y - a - b_1 D_{1i} + b_2 D_{2i}) D_{1i} = \sum_{n_1} Y_i - n_1 a - b_1 n_1 = 0$$

$$(5D.8c) \qquad \sum_{n_1} e_i D_{2i} = \sum_{n_1} (Y - a - b_1 D_{1i} + b_2 D_{2i}) D_{2i} = \sum_{n_1} Y_i - n_2 a - b_2 n_2 = 0$$

From (5D.8b) rearranged, we get:

$$b_1 n_1 = \sum_{n_1} Y_i - n_1 a \qquad (5D.9b)$$

From (5D.8c) rearranged, we get:

$$b_2 n_2 = \sum_{n_1} Y_i - n_2 a \qquad (5D.9c)$$

Substitute (5D.9b), (5D.9c) into (5D.8a) to get:

$$(5D.10a) \qquad \sum_{N} Y_i - Na - \left( \sum_{n_1} Y_i - n_1 a \right) - \left( \sum_{n2} Y_i - n_2 a \right) = 0$$

Collecting terms:

$$\text{Ý5}D.10b\text{þ} \qquad \sum_{N} Y_i - \sum_{n_1} Y_i - \sum_{n2} Y_i - Na - n_1 a - n_2 a = \sum_{n_3} Y_i - n_3 a$$

so:

$$\text{Ý5}D.11\text{þ} \qquad a = \frac{\sum_{n_3} Y_i - n_3 a}{n_3} = \bar{Y}_3$$

Substituting (5D.11) into (5D.8c) gives us:

$$\text{Ý5}D.12\text{þ} \qquad \sum_{n_1} Y_i - n_3 \bar{Y}_3 - b_1 n_1 = 0$$

$$\text{Ý5}D.13\text{þ} \qquad b_1 = \frac{\sum_{n_1} Y_i}{n_1} - \bar{Y}_3 = \bar{Y}_1 - \bar{Y}_3$$

Substituting (5D.14) into (5D.15) gives us:

$$\text{Ý5}D.14\text{þ} \qquad \sum_{n_2} Y_i - n_2 \bar{Y}_3 - b_2 n_2 = 0$$

$$\text{Ý5}D.15\text{þ} \qquad b_2 = \frac{\sum_{n_2} Y_i}{n_2} - \bar{Y}_3 = \bar{Y}_2 - \bar{Y}_3$$

From the regression we can generate the mean for each group:

$$\bar{Y}_1 = a + b_1 = \bar{Y}_3 + (\bar{Y}_1 - \bar{Y}_3)$$
$$\bar{Y}_2 = a + b_2 = \bar{Y}_3 + (\bar{Y}_2 - \bar{Y}_3)$$
$$\bar{Y}_3 = a = \bar{Y}_3$$

Note again, we could alternatively define:

$$D_3 = \left\{ \frac{1 \text{ if white}}{0 \text{ otherwise}} \right\}$$

and estimate:

$$Y_i = b_1 D_{1i} + b_2 D_{2i} + b_3 D_{3i} + e_i \quad (5.16)$$

Then you can show from normal equations:

$$b_1 = \bar{Y}_1 \qquad b_2 = \bar{Y}_2 \qquad b_3 = \bar{Y}_3$$

You cannot estimate:

$$Y_i = a + b_1 D_{1i} + b_2 D_{2i} + b_3 D_{3i} + e_i$$

for the same reason as above (linear dependence). Thus, as above, if there is an intercept in the dummy variable regression, there must be an "excluded group", a group not represented by a separate dummy variable.

## C. Variance of b and $s^2_{\bar{Y}_1 - \bar{Y}_2}$.

In (5.3a), we developed the expression for the variance in the difference of two group means:

$$s^2_{\bar{Y}_1 - \bar{Y}_2} = \frac{s^2_{y_1}}{n_1} + \frac{s^2_{y_2}}{n_2} \quad (5.14)$$

Since, as we have just shown, in a simple dummy variable regression, $b = (\bar{Y}_1 - \bar{Y}_2)$, then:

$$s^2_b = s^2_{\bar{Y}_1 - \bar{Y}_2} = \frac{s^2_{y_1}}{n_1} + \frac{s^2_{y_2}}{n_2} \quad (5.15)$$

Let us assume that $s^2_{Y_1} = s^2_{Y_2} = s^2_{Y_e}$, then rewrite (5.15) as

$$s^2_b = s^2_e \left( \frac{1}{n_1} + \frac{1}{n_2} \right) = s^2_e \frac{(n_1 + n_2)}{n_1 n_2} = \frac{\frac{s^2_e}{n_1 n_2}}{N} \quad (5.15a)$$

Now, consider the sum of squared deviations for the variable $D_i$:

$$\sum_N (D_i - \bar{D})^2 = \sum_N (D_i^2 - 2\bar{D}) \sum_N (D_i) + N\bar{D}^2 \quad (5.16)$$

Substituting $\bar{D} = \frac{\sum D_i}{N}$:

$$\text{(5.16a)} \qquad \sum_N \left(D_i - \bar{D}\right)^2 = \sum_N D_i^2 - 2\frac{\sum_N D_i \sum_N D_i}{N} + \frac{N\left(\sum_N D_i\right)^2}{N^2} = \sum_N D_i^2 - \frac{\left(\sum\right.}{}$$

All the $n_2$ terms where $D_i = 0$ drop out of $\sum_N (D_i)^2$ and for $D_i = 1, D_i^2 = 1$

$$\text{(5.17)} \qquad \sum_N D_i^2 = n_1$$

$$\text{(5.18)} \quad \sum_N \left(D_i - \bar{D}\right)^2 = \sum_N D_i^2 - \frac{\left(\sum D_i\right)^2}{N} = n_1 - \frac{n_1^2}{N} = \frac{n_1 N - n_1^2}{N}$$

$$= \frac{n_1(n_1 + n_2) - n_1^2}{N} = \frac{n_1 n_2}{N}$$

So using (5.18), we can rewrite (5.15a):

$$\text{(5.19a)} \quad s_b^2 = \frac{s_e^2}{\frac{n_1 n_2}{N}} = \frac{s_e^2}{\sum \left(D_i - \bar{D}\right)^2}$$

We wrote down the test statistic for determining a difference in means was significantly different from zero. Now using (5.13b), (5.15a), and (5.19) to substitute into (5.2), we can rewrite (5.2):

$$\text{(5.20)} \qquad \left| \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)}{s_{\bar{Y}_1 - \bar{Y}_2}} \right| = \frac{|b|}{s_b} = \frac{|b|}{\sqrt{\frac{s_e^2}{\sum \left(D_i - \bar{D}\right)^2}}}$$

so that we can test for the significance of a difference in means by forming the test statistic given by (5.20). The expression $\left| \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)}{s_{\bar{Y}_1 - \bar{Y}_2}} \right|$ is a Z distribution, but when $\left(\bar{Y}_1 - \bar{Y}_2\right)$ are from a sample and $s_{\bar{Y}_1 - \bar{Y}_2}$ is estimated $\hat{s}_{\bar{Y}_1 - \bar{Y}_2}$ as in (5.5), the ratio $\left| \frac{\left(\bar{Y}_1 - \bar{Y}_2\right)}{s_{\bar{Y}_1 - \bar{Y}_2}} \right|$ is a t distribution with N-1 degrees of freedom (see Beals, chapter 8). Now if we estimate b in (5.8) from a sample, and $s_y$ from a sample, and substitute these into (5.20) then:

$$\frac{b}{s_b} = \frac{b}{\sqrt{\frac{s_e^2}{\sum (D_i - \bar{D})^2}}} = t_{N-2} \qquad (5.21)$$

So a test equivalent to the test for whether the difference of the group means is significantly different from zero would be to form the ratio (5.21) and then see if it is greater than or equal to 1.96, which is the t value for the 5 percent level of significance, where N-2 is large.

Thus, we can see that the test for significance of b in the simple dummy variable regression is equivalent to the test for significant difference in the means of Y for two groups.

Making a leap by analogy, we can substitute any simple regression independent variable $X_i$ for $D_i$ and write down the variance for $b_{yx}$ directly from (5.19) as:

$$s_{b_{yx}}^2 = \frac{s_e^2}{\sum (X_i - \bar{X})^2} = \frac{s_e^2}{\sum x_i^2} \qquad (5.22)$$

where $s_e^2$ is the variance of the error term in the simple regression of y and x. (We will return to this in a more formal fashion later). Likewise, by analogy, the test statistic for $b_{yx}$ significantly different from zero would be:

$$\frac{b}{\sqrt{s_b^2}} = \frac{b}{\sqrt{\frac{s_e^2}{\sum x_i^2}}} = t_{N-2} \qquad (5.23)$$

# D. Variance of Multiple Regression Coefficients

Recall that we derived the multiple regression coefficient $b_{y_{1.2}}$ through the auxiliary regression of $x_{1i}$ and $x_{2i}$ and showed in equation (3.6) that $b_{y_{1.2}} = \frac{\sum y_i e_{1.2i}}{\sum e_{1.2i}^2}$.

Since $e_{1.2i}$ is the independent random variable in a regression with y as the dependent variable, we can simply substitute $e_{1.2i}$ for $x_i$ in

(5.22) and write:

$$(5.24) \qquad a^2_{b_{y_{1.2}}} = \frac{a^2_{e^2_{1.2i}}}{> e^2_{1.2i}}$$

Just as (3.6) can be generalized to the case of more than two independent variables when the residual from the auxiliary regression is redefined, i.e., $e_{1.2,3,...,ni,}$ so (5.24) can similarly be generalized using $e_{1.2,3,...,ni}$ in place of $e_{1.2i}$:

$$(5.25) \qquad a^2_{b_{y_{1.2,3,4,,,ni}}} = \frac{a^2_{e^2_{1.2,3,4,,,ni}}}{> e^2_{1.2,3,4,,,ni}}$$

# E. F-Statistic, Analysis of Variance, Hypothesis Tests on Several Parameters

(see Beals pp. 247-50, pp. 274-81)

First, we wish to show the equivalence of the F-statistic test for significance of the regression and the t-statistic test for significance in the case of a simple regression. The F-statistic is:

$$(5.26) \qquad \frac{SSR/(K-1)}{SSE/(N-K)} = \left[\frac{N-K}{K-1}\right]\left[\frac{SSR/SST}{1-(SSR/SST)}\right] = \left[\frac{N-K}{K-1}\right]\beta R^2$$

where SSR - $> y_i^2$; regression sum of squares (see 4.7)

SSE - $> e_i^2$; error sum of squares

SST - $> y_i^2$; total sum of squares

K - number of regression coefficients (incl. constant)

N - sample size

The critical value for F is determined from the F-table entry for (K-1) numerator degrees of freedom and (N-K) denominator degrees of freedom, for the selected confidence level, e.g. 5 percent. If the calculated F-value exceeds the critical F-value, the null hypothesis of no significance is rejected, i.e. the regression explains a significant proportion of the variance in Y.

Noting that in the case of a simple regression K = 2, so (K - 1) = 1, and using (4.11) to substitute for SSR= $> y_i^2$, we get:

$$(5.27) \quad \frac{SSR/(2-1)}{SSE/(N-2)} = \frac{b^2 \sum x_i^2}{\sum e_i^2/N-2} = \frac{b^2 \sum x_i^2}{s_e^2} = \frac{b^2}{s_e^2/\sum x_i^2}$$

Comparing (5.27) with (5.23) shows for the simple regression

$$(5.27a) \quad \frac{SSR/(2-1)}{SSE/(N-2)} = \frac{b^2 \sum x_i^2}{s_e^2} = t^2$$

So in the case of a simple regression, the t test for the significance of b and the F test for the significance of the regression are equivalent. (The F test is also sometimes referred to as the test for significance of $R^2$, as can be seen from the last expression in (5.26).)

The F test for the regression is sometimes referred to as analysis of variance, since it derives from the partition of total variance, as in (4.7), into a regression sum of squares and an error sum of squares. It is usually presented in an analysis of variance table. (See Beals, p.249, p. 275)

The F-statistic for a multiple regression is constructed from (5.26), but the relationship of the F-statistic and the t-statistic is more complicated. Suppose we are calculating the F-statistic for a regression equation such as (3.1). Once again the SSR will be $\sum y_i^2$, but now when we use (4.14) to substitute for $y_i$, we get:

$$(5.28) \quad \frac{SSR/(3-1)}{SSE/(N-3)} = \frac{\sum y_{.12i/2}^2}{\sum e_{.12i/(N-3)}^2}$$

$$= \frac{\sum (b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i})(b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i})/2}{s_e^2}$$

$$= \frac{\sum b_{y_{1.2}}^2 x_{1i}^2 + \sum b_{y_{2.1}}^2 x_{2i}^2 + 2 \sum b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i}}{2 s_e^2}$$

Using (5.24), we note that:

$$(5.29a) \quad t_{b_{y1.2}}^2 = \frac{b_{y1.2}^2}{s_{b_{y1.2}}^2} = \frac{b_{y1.2}^2}{s_e^2/\sum e_{1.2i}^2}$$

and:

$$t^2_{b_{y2.1}} = \frac{b^2_{y2.1}}{s^2_{b_{y2.1}}} = \frac{b^2_{y2.1}}{s^2_e / \sum e^2_{2.1i}} \qquad (5.29b)$$

or:

$$b^2_{y_{1.2}} = t^2_{b_{y1.2}} s^2_{b_{y1.2}} = \frac{t^2_{by1.2} s^2_e}{\sum e^2_{1.2i}} \qquad (5.30a)$$

$$b^2_{y_{2.1}} = t^2_{b_{y2.1}} s^2_{b_{y2.1}} = \frac{t^2_{by2.1} s^2_e}{\sum e^2_{2.1i}} \qquad (5.30b)$$

Substituting (5.30a) and (5.30b) into (5.28), then using an expression $\sum e^2_{1.2i}$ and $\sum e^2_{2.1i}$

derived from squaring both sides of auxiliary equations like (3.4) and the expression (4.13a) for the correlation coefficient $r_{12}$ between $X_1$ and $X_2$, with considerable manipulation (which I won't go into here: see Kmenta, p. 368), you get:

$$\frac{SSR/(3-1)}{SSE/(N-3)} = \frac{t^2_{by_{1.2}} + t^2_{by_{1.2}} + 2t_{b_{y_{1.2}}} t_{b_{y_{2.1}}} r_{12}}{2(1-r^2_{12})} \qquad (4.31)$$

which is the F-statistic for 2, N-2 degrees of freedom.

Therefore, while F for the regression and t for b are strictly related in the simple regression case so significance of b by a t-test necessarily means significance of the regression by the F-test, in the multiple regression case, one cannot infer from the t-tests on the coefficients to the F-test for the regression as a whole. If $r^2_{12}$ is close to 1, F may be large even though the $t_{by1.2}, t_{by2.1}$ are small.

# F. Tests for Joint Significance of Multiple Regression Coefficients

Sometimes we wish to test whether a set of coefficients as a group is significantly different from zero, i.e.

H:$K_{y_{1.2}} = K_{y_{2.1}} = 0$

where $K_{y_{1.2}}, K_{y_{2.1}}$ are the population values of which $b_{y_{1.2}}$ and $b_{y_{2.1}}$

are the sample estimates. Recall that, e.g. $b_{y_{1.2}}$

can differ from $K_{y_{1.2}}$ because of sampling variability (just as the sample difference in means

$(\bar{Y}_1 - \bar{Y}_2)_i$ could differ from the true population difference in means $(\overline{Y_1 - Y_2})$

in our difference in means examples alone).

Now look at the numerator of (5.28), we can see that if the population value were $K_{y_{1.2}} = K_{y_{2.1}} = 0$, then the SSR would only differ from zero because of the sampling variability in the $b_{y_{1.2}}$ and $b_{y_{2.1}}$.

. If that is true then both the numerator and denominator give estimates of the sampling variability, and the F-test, which is basically a t-test of whether the numerator and denominator are from the same distribution, gives us a test of the null hypothesis. If $K_{y_{1.2}} = K_{y_{2.1}} = 0$, the F-value will fall below the critical $F_{2}$, N-3 value. If the F-value is above the critical value, the numerator and denominator are from different distributions and SSR differs from zero by more than sampling variability and therefore not all the $K$are zero.

We can extend the use of the F-statistic to test for the joint significance of a subset of regression coefficients from a multiple regression involving more than two independent variables.

Consider the population regression equation:

$$y_i = K_{y_{1.23}} x_{1i} + K_{y_{2.13}} x_{2i} + K_{y_{3.12}} x_{3i} + e_{y.123i} \quad (5.32)$$

Suppose we are interested in testing the hypothesis:

$$H_0 = K_{y_{2.13}} = K_{y_{3.12}} = 0$$

Note that if the hypothesis is true, the appropriate population regression is:

$$y_i = K_{y_{2_1}} x_{1i} + e_{y.1i} \quad (5.33)$$

Note that both equations would yield the same total sum of squares (SST), but yield different regression sum of squares (SSR) and error sum of squares (SSE). For (5.32), we get:

$$(5.34) \quad \sum y_i^2 = \sum y_{.123i}^2 + \sum e_{y.123i}^2$$

$$(5.34a) \quad SST = SSR_{.123} + SSE_{.123}$$

For (5.33), we get:

$$(5.35) \quad \sum y_i^2 = \sum y_{.1i}^2 + \sum e_{y.1i}^2$$

$$(5.35a) \quad SST = SSR_{.1} + SSE_{.1}$$

If in fact the null hypothesis is true, and $K_{y_{2.13}} = K_{y_{3.12}} = 0$

then in the population $SSR._{123}$ would equal $SSR._1$ and any observed difference between them would be due to sampling variability (causing $b_{y_{2.13}}$ and $b_{y_{3.12}}$ to differ from zero). Thus, if we estimate (5.32) and (5.33) from the sample and form the ratio:

$$(5.36) \quad \frac{SSR_{.123} - SSR_{.1}/(4 - 2)}{SSE_{.123}/(N - 4)} \sim F_{(4-2),(N-4)}$$

If the null hypothesis is true, then the numerator and denominator will both estimate sampling variability in the same population, and the F-value of the ratio will fall below the critical value for F with those degrees of freedom. If F exceeds the critical value, the $SSR._{.123}$ exceeds $SSR._1$ by more than sampling variability, and the null hypothesis is false. This means either $K_{y2.13}$ or $K_{y3.12}$ or both do not equal zero (at the given confidence level).

We can also write (5.36) in terms of $R^2$:

$$(5.36a) \quad \frac{SSR_{.123} - SSR_{.1}/(4 - 2)}{SSE_{.123}/(N - 4)} = \frac{SSR_{.123}/SST - SSR_{.1}/SST}{SSE_{.123}/SST}$$

$$= \frac{R_{.123}^2 - R_{.1}^2}{1 - R_{.123}^2} \frac{N - Q}{Q - K}$$

# VI. Problems in Non-Standard Statistical Analyses

So far the discussion has been concerned with developing the logic

of simple and multiple regression analysis and with tests of hypotheses concerning the coefficients of the regression equation. In this development, I have ignored the importance of certain assumptions regarding the underlying probabilistic process which generates the observations in the population from which the statistical samples are drawn. I have also disregarded problems which arise when the relationship between that population process, the regression equation specified, and/or the sample data actually available are not standard. In the sections which follow, I will go over some of the problems of what I call non-standard analysis, showing some examples of how they can affect the estimates of regression coefficients or tests of hypothesis, and, in some cases, how more complicated estimation procedures can overcome these problems.

# A. Multicollinearity (see Beals, pp.294-7)

The problem of multicollinearity can arise when in the sample available for estimation, two or more of the independent variables have a high covariance. Fortunately, the way we developed the multiple regression coefficient via the auxiliary regression [equations (3.1) to (3.6)] makes it quite easy to see how this problem can arise. We run the auxiliary regression (3.2) in order to obtain the $e_{1.2i}$. If $X_1$ and $X_2$ were perfectly collinear, the $e_{1.2i}$ would all be zero. In that case, (3.5) would be meaningless and the regression coefficient in (3.6) would be undefined since $> e_{1.2i}^2$ in the denominator would be zero. An exact relationship between $X_1$ and $X_2$ rarely arises and the multicollinearity problem most often takes the form of a high co-variance between $x_1$ and $x_1$, with the result $e_{1.2i}$of the auxiliary regression small but non-zero. We can see that as the $e_{1.2i}$ get quite small, the estimate of $b_{y_{1.2}}$

may become rather unstable with both the numerator $> y_i e_{1.2i}$ and the denominator $> e_{1.2i}^2$getting close to zero, but their quotient,$b_{y_{1.2}}$

being much affected by which gets closer to zero: if the numerator gets closer,$b_{y_{1.2}}$ is small; if the denominator gets closer, $b_{y_{1.2}}$ may be quite large. In these cases, small rounding errors in calculations can make estimates bounce around a lot. The problem is further underlined by examining the expression for $\hat{a}_{b_{y_{1.2}}}^2$, as developed in (5.24). The

higher the covariance of $x_1$ and $x_2$, the smaller will be $\sum e_{1.2i}^2$ with the numerator of $\sigma_{b_{y_{1.2}}}^2$, $\sigma_{e_{\cdot 12}}^2$

, constant the smaller denominator, $\sum e_{1.2i}^2$ causes $\sigma_{b_{y_{1.2}}}^2$ to get very large. Thus, multicollinearity of $x_1$ and $x_2$ increases the variance of $b_{y_{1.2}}$

and $b_{y_{2.1}}$ considerably. This is another way of seeing the instability of the estimated regression coefficients.

While the problem of multicollinearity in the sample does arise fairly frequently in economic analysis, particularly in time-series analysis where many variables move together in both trends and cycles, it is common to misuse the concept of multicollinearity and to explain away weak results as due to multicollinearity when in fact the problem does not appear in the data. I have emphasized, by reference to (3.6) and (5.24), that the problem can arise when the $e_{1.2i}$ alone do not constitute sufficient evidence of the problem. Some analysts jump to the conclusion that high covariance of $x_1$ and $x_2$ alone establishes the existence of a multicollinearity problem. However, as long as the correlation of $x_1$ and $x_2$ falls short of 1, it is possible that there is sufficient information in the sample as represented by the non-zero $e_{1.2i}$ to get significant estimate of $b_{y_{1.2}}$. The ultimate test of multicollinearity lies in the estimates of $b_{y_{1.2}}$ and $\sigma_{b_{y_{1.2}}}$. If the relations of y and $x_1$ is sufficiently strong, it will show up in $\sum y_i e_{1.2i} / \sum e_{1.2i}^2$

, even though the $e_{1.2i}$ are small. For example, even if the $R_{1.2}^2$

is .90, the .10 independent variance of $x_{1i}$ represented in the $e_{1.2i}$ may be adequate to estimate a significant relationship of y and $x_1$. Even though $\sigma_{b_{y_{1.2}}}$ will be large when $e_{1.2i}$ are small, if there is a strong covariance of y and $e_{1.2i}$, $\sum y_i e_{1.2i}$, it will make $b_{y_{1.2}}$

sufficiently large for the ratio $b_{y_{1.2}} / \sigma_{b_{y_{1.2}}}$ to pass the t-test.

In looking for multicollinearity problems, therefore, it is not sufficient to examine the correlation among independent variables. The best evidence is obtained by first running a regression with just $x_1$ as an independent variable, and then a second regression with both $x_1$ and $x_2$. If $x_1$ is significant in the first regression, i.e. $b_{y_1} / \sigma_{b_{y_1}}$ passes the t-test, but in the second regression $\sigma_{b_{y_{1.2}}}$ gets very large, $\sigma_{b_{y_{2.1}}}$ is also very large and as a result, both $b_{y_{1.2}}$ *and* $b_{y_{2.1}}$ fail to pass the t-test, then there is a multicollinearity problem, i.e. there is not enough information in the sample about $x_1$ independent of $x_2$, and vice-versa, to estimate

the independent effect on y of $x_1$ and of $x_2$. (In this comparison of the two regressions, the $R^2_{.12}$

will not be significantly greater than the $R^2_{.1}$ by the F-test indicated in (5.36a), evidence that adding $X_2$ adds no independent information about factors affecting y). Note that what happens in this case is that as we add $x_2$ in the second regression, the estimate of the effect of $x_1$ and y shifts from $b_{y_1} = \frac{\sum y_i x_i}{\sum x_{1i}^2}$ to $b_{y_{1.2}} = \frac{\sum y_i e_{1.2i}}{\sum e_{1.2i}^2}$

and its variance from $\hat{\sigma}^2_{by1} = \frac{\hat{\sigma}^2_{e.1}}{\sum x_{1i}^2}$ to $\hat{\sigma}^2_{by1.2} = \frac{\hat{\sigma}^2_{e..12}}{\sum e_{1.2i}^2}$

In the shift, the numerator of the coefficient falls more rapidly than the denominator, and the denominator of the variance falls sharply. The resultant t-ratio shifts from $b_{y_1}/\sigma_{by_1}$ to $b_{y_{1.2}}/\sigma_{by_{1.2}}$ and falls sharply as the numerator falls and/or the denominator rises sharply.

The multicollinearity problem then shows up in the large variances of the regression coefficients in the estimated regressions. It cannot be determined without actually running the regressions; again high correlation among the independent variables does not alone evidence a multicollinearity problem in the estimates.

## B. Omitted Variable Bias (see Beals, pp. 288-92)

[At this point, students should review the properties of estimators, i.e. unbiasedness, efficiency, asymptotic unbiasedness, consistency, and asymptotic efficiency. (see Beals, pp. 144-64.) I will not discuss these here but assume familiarity with the concepts.]

One of the arguments for using multiple regression in trying to estimate a relationship between a dependent variable y and an independent variable $x_1$ is that there may be other variables affecting y, say $x_1$, which are in turn correlated with $x_1$. Thus, taking the simple regression estimate of the relationship, $b_{y_1}$ may mislead us about the effect of a change in $x_1$ holding $x_2$ constant. To put this in other terms, the regression equation:

$$y = b_{y_1} x_{1i} + e_{y.1i} \qquad (6.1)$$

may be a misspecification of the correct relationship:

$$y = b_{y_{1.2}} x_{1i} + b_{y_{2.1}} x_{2i} + e_{y.12i} \qquad (6.2)$$

because we have omitted a variable $x_i$ which affects y. If this is the case, $b_{y1}$

is a biased estimate of the true coefficient $b_{y_{1.2}}$.

We have already derived an expression which indicates the extent of omitted variable bias if we estimate (6.1) when (6.2) is the appropriate estimating equation. The bias indicated in equation (3.20a) is repeated here:

$$b_{y1} = b_{y_{1.2}} + b_{21} B_{Y_{2.1}} \qquad (6.3)$$

To the extent $x_2$ does have an independent effect on y, i.e. if $b_{y_{2.1}} = 0$, then $b_{y_1}$ will be subject to omitted variable bias.

The issue of omitted variable bias arises sometimes in critiques of estimated relationships when it is argued that some theoretically relevant variable has not been included in the estimated relationship. It also arises in cases where it is hypothesized that some unobservable variable operates on the dependent variable and, since it can't be measured, has been omitted leading to a potential bias. Sometimes, using relationship (6.3), one can make plausible guesses about the direction and order of magnitude of omitted variable bias.

# VII. Basic Assumptions about the Population Model and Problems Due to their Violation (see Beals, pp. 233, 265-7)

It is now necessary to be clearer about the basic assumptions which are made about the probability model which generates the observations in the population. Also we need to reemphasize the distinctions between the population model and the regression model estimated from the sample. We will then see how the desirable

properties of the least squares estimates depend on these basic assumptions by showing how some violations of the basic assumptions can cause least squares estimates to fail to have the desirable properties.

Let us specify the population regression model as:

$$(7.1) \quad y = K_{y_{1.2}} x_{1i} + K_{y_{2.1}} x_{2i} + W_{y.12i}$$

where $W_{y.12i}$ is the disturbance from the regression line for observation i in the population. I use the $K$ and $W$ notation to differentiate the population model and observations from the sample regression equation parameters and error term.

Basic assumptions about the population probability model as represented by (7.1) which we make are:

(7.2) Each $W_i$ is a random variable with mean zero

$$\sum (W_i) = 0 \qquad i = 1, 2..., n$$

(7.3) The $W_i$'s are independent of each other

$$cov(W_i, W_j) = E(W_i, W_j) = 0 \quad \text{for} i \neq 0$$

(7.4) All $W_i$'s have the same variance

$$var(W_i) = E(W_i^2) = \sigma_W^2 \quad \text{for} i = 1, 2, ..., n$$

This condition is referred to as homoskedasticity.

(7.5) The X's are

a) non-random, or

b) independent of the $W_i$, and

c) such that $\frac{\sum (X_i - \bar{X})^2}{n}$ is finite and non-zero.

As a result

$$cov(X_{k,i}, W_i) = E(X_{k,i}, W_i) = 0 \quad k = 1, 2 i = 1, 2, ..., n$$

I will not discuss these assumptions in detail, but will proceed first to illustrate their importance by using one of them to derive the expression for

$var(b_{yi})$ in a simple regression and then, second, to illustrate problems of non-standard analysis which arise when some of these

assumptions are violated.

We have already derived the expression for var$(b_{yi})$ by another means in (5.22). Let us start, this time, however, with the basic population model as:

$$(7.6) \quad y_i = \beta_{y_i x_i} + \varepsilon_{y.1i}$$

From our sample, we derive the least squares estimate $b_{y_1}$, which, using (2.11) is:

$$(7.7) \quad b_{yi} = \frac{\sum x_{1i} yi}{\sum x_{1i}^2}$$

Substitute (7.6) for yi:

$$(7.8) \quad b_{yi} = \frac{\sum x_{1i}(\beta_{y_i x_i} + \varepsilon_{y.1i})}{\sum x_{1i}^2}$$

$$= \frac{\sum x_{1i}^2 \beta_{y_i} + \sum x_{1i}\varepsilon_{y.1i}}{\sum x_{1i}^2} = \beta_{y_1} + \frac{\sum x_{1i}\varepsilon_{y.1i}}{\sum x_{1i}^2}$$

If we take the expected value of b, we see that b is an unbiased estimator of $\beta_{y_1}$:

$$(7.9) \quad E(b_{y_1}) = \beta_{y_1} + E\left[\frac{\sum x_{1i}\varepsilon_{y.1i}}{\sum x_{1i}^2}\right] = \beta_{y_1} + E\left[\frac{\sum x_{1i}}{\sum x_{1i}^2}\right] E(\varepsilon_{y.1i})$$

because, if assumption (7.5b) holds, $x_{1i}$ and $\varepsilon_{y.1i}$ are independent random variables and the expected value of the product of independent random variables is the product of their expected values, thus

$$E\left[\frac{\sum x_{1i}\varepsilon_{y.1i}}{\sum x_{1i}^2}\right] = E\left[\frac{\sum x_{1i}}{\sum x_{1i}^2}\right] E(\varepsilon_{y.1i})$$

(see Beals, p. 90),

and, by assumption (7.2) $E(\varepsilon_i) = 0$

For the variance of $b_{y_i}$, we have:

$$(7.10) \quad var(b_{y_1}) = var\left[ K_{y_1} + E\left[ \frac{\sum x_{1i} W_{y.1i}}{\sum x_{1i}^2} \right] \right]$$
$$= var\left[ \frac{\sum x_{1i} W_{y.1i}}{\sum x_{1i}^2} \right]$$

since variance of a random variable plus a constant is equal to the variance of the random variable (see Beals, p. 62).

$$(7.10a) \quad var(b_{y_1}) = var\left[ \frac{\sum x_{1i} W_{y.1i}}{\sum x_{1i}^2} \right] = E\left[ \frac{\sum x_{1i} W_{y.1i}}{\sum x_{1i}^2} \right]^2$$
$$= E\left[ \frac{\sum x_{1i}^2 W_{y.1i}^2}{(\sum x_{1i}^2)^2} \right] + 2E\left[ \frac{\sum_{i<j} x_{1i} x_{ij} W_i W_j}{(\sum x_{1i}^2)^2} \right]$$

In the first term, by assumption (7.4), $E(W_i^2) = \sigma_W^2$ for all i, it can be factored out and the $\sum x_{1i}^2$

cancels out with part of the denominator, leaving $\sigma_W^2 E\left[ \frac{1}{\sum x_{1i}^2} \right]$. In the second term, by assumption (7.3), the $E(W_i, W_j) = 0$

and by (7.5), $E(X_{k,i}, W_i) = 0$, so the second term vanishes there:

$$(7.10b) \quad var(b_{y_i}) = \sigma_W^2 E\left[ \frac{1}{\sum x_{1i}^2} \right]$$

When we estimate $var(b_{y_i})$ from the sample, we substitute $\hat{\sigma}_e^2 = \frac{\sum e_i^2}{n-2}$ as an estimate of $\sigma_W^2$ and $\frac{1}{\sum x_{1i}^2}$ from the sample as an estimate of

$E\left[ \frac{1}{\sum x_{1i}^2} \right]$. Thus we obtain:

$$(7.11) \quad \widehat{var}(b_{y_i}) = E\left[ \frac{\hat{\sigma}_e^2}{\sum x_{1i}^2} \right]$$

which is equivalent to (5.22).

Therefore, by using the basic assumptions, we obtain the estimator of the variance $b_{y_i}$

more directly than by the route used above. It should be clear that violation of the basic assumptions could lead to a different expression for

$E(b_{y_1})$ and var($b_{y_1}$). I will proceed to a brief examination of some examples of such cases.

# VIII. Violations of the Basic Assumptions (see Beals, p. 348)

## A. Autocorrelated Disturbances

Suppose that assumption (7.3) is violated, so that:

$$(8.1) \qquad cov(W_i, W_j) = E(W_i, W_j) \neq 0$$

For example,

$W_t$ might be generated by an autoregressive process such as,

$$(8.2) \qquad W_t = \rho W_{t-1} + v_t$$

where

$v_t$ are independent random variables with mean zero and constant variance $\sigma_v^2$.

Let us examine the characteristics of the estimator $b_{y_1}$ for $\beta_{y_1}$ from the population equation:

$$(8.3) \qquad y_t = b_{y1} x_t + W_t$$

under these conditions.Similar to equation (7.8):

$$(8.4) \quad b_{y1} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t (\beta_{y_1} x_t + W_t)}{\sum x_t^2} = \beta_{y_1} + \frac{\sum x_t W_t}{\sum x_t^2}$$

$$(8.5) \qquad E(b_{y_1}) = \beta_{y_1} + E\left[\frac{\sum x_t W_t}{\sum x_t^2}\right] E(W_t)$$

so $b_{y_1}$ is still an unbiased estimator of $K_{y_1}$.

However, when we look at var$b_{y_1}$, we find:

$$var(b_{y_1}) = var\left(K_{y_1} + \frac{\sum x_t W_t}{\sum x_t^2}\right) = var\frac{\sum x_t W_t}{\sum x_t^2}$$

$$(8.6) \qquad = E\left(\frac{\sum x_t W_t}{\sum x_t^2}\right)^2 = E\left[\frac{\sum x_t^2 W_t^2}{(\sum x_t^2)^2}\right] + 2E\left[\frac{\sum_{s<T} x_t x_{t+s} W_t W_{t+s}}{(\sum x_t^2)^2}\right]$$

$$= \sigma_W^2 E\left[\frac{1}{\sum x_t^2}\right] + 2\sigma_W^2 E\left[\frac{\sum_{s<T} x_t x_{t+s-s}}{(\sum x_t^2)^2}\right]$$

since, given (8.2), cov ($W_t$, $W_{t-1}$) = $\rho\sigma_W^2$,

. Now, as contrasted with the similar equation (7.10a), the second term in (8.6) doesn't disappear. Thus, the standard tests which use (7.10b), or more likely its estimate (7.11), are incorrect [they will in general underestimate var $(b_{y_i})$.

The Durbin-Watson statistic is used to test for autocorrelated disturbances (see Beals, p. 348). If the test indicates autocorrelated disturbances, an attempt can be made to correct for it. First, estimate:

$$(8.7) \qquad y_t = b_{y_1} x_{1t} + \varepsilon_t$$

If the Durbin-Watson test indicates autocorrelation, estimate:

$$(8.8) \qquad \varepsilon_t = \hat{\rho}\varepsilon_{t-1} + v_t$$

Using this estimate of $\rho$, transform the data by multiplying t-1 values by

rho and subtracting from the t values, yielding:

$$(8.9) \qquad y_t - \rho y_{t-1} = b_{y_1}(x_{1i} - \rho x_{1t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}) = b_{y_1}(x_{1i} - \rho x_{1t-1})$$

If $\hat{\rho}$ is a good estimate of $\rho$, then $v_t$ conforms with the basic assumptions (7.2) to (7.4), and the usual test statis-tics may be applied to

$b_{y_1}$ as estimated from (8.9).

# B. Heteroskedasticity

Suppose assumption (7.4) is violated. Then in expression (7.10a), for var $b_{y_1}$, the second term, $2E\left[\frac{\sum_{i<j} x_{1i}x_{ij}W_iW_j}{(\sum x_{1i}^2)^2}\right]$ does vanish, but the first term $E\left[\frac{\sum x_{1i}^2 W_{y.1i}^2}{(\sum x_{1i}^2)^2}\right]$ $= a_W^2 E\left[\frac{1}{\sum x_i^2}\right]$ since $a_W^2$ differs for each i, so:

$$\text{var}(b_{y_i}) = \frac{x_{1i}^2 a_{y.1i}^2}{(\sum x_{1i}^2)^2} + \frac{x_{12}^2 a_{y.2}^2}{(\sum x_{1i}^2)^2} + \ldots + \frac{x_{1n}^2 a_{y.n}^2}{(\sum x_{1i}^2)^2} \qquad (8.10)$$

Thus, using (7.11) would in some cases over-estimate (if $a_i$ increases with $x_i$), and other cases under-estimate the true var$b_{y_1}$ given by (9.1). [See Beals, pp. 357-62 for tests and corrections for heteroskedasticity].

[Note that between (5.15) and (5.15a) above, we assumed $a_{y_1}^2 = a_{y2}^2 = a_{y_e}^2$ i.e. homoskedasticity across groups. This allowed us to make the transition from $a_{(\bar{y}_1 - \bar{y}_2)}^2$ to $a_b^2$.

# C. Generalized Least Squares: An Example

Here, we will simply present an example of a G.L.S. correction for heteroskedasticity. From Kmenta p. 504, we have the G.L.S. estimate for B:

$$\hat{B} = (X' \Omega^{-1} X)^{-1} (X' \Omega^{-1} Y) \qquad (12.14)$$

We will simply grind through the case for forming this estimate when there is one independent variable X, and there are only three observations, and where the disturbances are hetero-skedastic. Consider

Ý8.11þ  $Y = BX+P$ where $E\acute{Y}P_i, P_j\text{þ} = a_{ij} = \left\{ \begin{array}{c} a_{ij} \text{for i=j} \\ 0 \text{ for i} \neq j \end{array} \right\}$

$$I = \begin{bmatrix} a_{11}^2 & 0 & 0 \\ 0 & a_{22}^2 & 0 \\ 0 & 0 & a_{33}^2 \end{bmatrix}$$

liand the values of $I$ are known.
Then, referring to Kmenta, p. 610-11, form $I^{?1}$:

Ý8.12þ  $$I^{?1} = \frac{1}{\det I} adj.\, I$$

For det $I$, see Kmenta, p. 607 (B.15):
det $I = a_{11}^2 a_{22}^2 a_{33}^2 + 000 + 000 ? 0a_{22}^2 0 ? 00a_{11}^2 ? a_{33}^2 00$
$=a_{11}^2 a_{22}^2 a_{33}^2$
Kmenta, p. 610 (B.25):

adj. $I = \begin{bmatrix} a_{11}^2 & a_{21}^2 & a_{31}^2 \\ a_{12}^2 & a_{22}^2 & a_{32}^2 \\ a_{13}^2 & a_{23}^2 & a_{33}^2 \end{bmatrix}$

$$\mathbf{I}_{11} = \det\begin{bmatrix} a_{22}^2 & 0 \\ 0 & a_{33}^2 \end{bmatrix}^{+1^2} = a_{22}^2 a_{33}^2$$

$$\mathbf{I}_{22} = \det\begin{bmatrix} a_{11}^2 & 0 \\ 0 & a_{33}^2 \end{bmatrix}^{+1^4} = a_{11}^2 a_{33}^2$$

$$\mathbf{I}_{33} = \det\begin{bmatrix} a_{11}^2 & 0 \\ 0 & a_{22}^2 \end{bmatrix}^{+1^2} = a_{11}^2 a_{22}^2$$

$$\mathbf{I}_{12} = \det\begin{bmatrix} 0 & 0 \\ 0 & a_{33}^2 \end{bmatrix}^{+1^3} = 0 \qquad \mathbf{I}_{13} = \det\begin{bmatrix} 0 & a_{22}^2 \\ 0 & 0 \end{bmatrix}^{+1^4} = 0$$

$$\mathbf{I}_{21} = \det\begin{bmatrix} 0 & 0 \\ 0 & a_{33}^2 \end{bmatrix}^{+1^3} = 0 \qquad \mathbf{I}_{23} = \det\begin{bmatrix} a_{11}^2 & 0 \\ 0 & 0 \end{bmatrix}^{+1^5} = 0$$

$$\mathbf{I}_{31} = \det\begin{bmatrix} 0 & 0 \\ a_{22}^2 & 0 \end{bmatrix}^{+1^4} = 0 \qquad \mathbf{I}_{32} = \det\begin{bmatrix} a_{11}^2 & 0 \\ 0 & 0 \end{bmatrix}^{+1^5} = 0$$

$$adj.\,\mathbf{I} = \det\begin{bmatrix} a_{22}^2 a_{33}^2 & 0 & 0 \\ 0 & a_{11}^2 a_{33}^2 & 0 \\ 0 & 0 & a_{11}^2 a_{22}^2 \end{bmatrix}$$

Thus, if we estimate:

$$\text{Ý8.13Þ} \qquad \mathbf{I}^{+1} = \frac{1}{\det \mathbf{I}} adj\,\mathbf{I} = \frac{1}{a_{11}^2 a_{22}^2 a_{33}^2}\begin{bmatrix} a_{22}^2 a_{33}^2 & 0 & 0 \\ 0 & a_{11}^2 a_{33}^2 & 0 \\ 0 & 0 & a_{11}^2 a_{22}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{a_{22}^2 a_{33}^2}{a_{11}^2 a_{22}^2 a_{33}^2} & 0 & 0 \\ 0 & \frac{a_{11}^2 a_{33}^2}{a_{11}^2 a_{22}^2 a_{33}^2} & 0 \\ 0 & 0 & \frac{a_{11}^2 a_{22}^2}{a_{11}^2 a_{22}^2 a_{33}^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{a_{11}^2} & 0 & 0 \\ 0 & \frac{1}{a_{22}^2} & 0 \\ 0 & 0 & \frac{1}{a_{33}^2} \end{bmatrix}$$

Ý8.14Þ $$X^{?1} \; I^{?1} = \beta X_1 X_2 X_3 \grave{a} \begin{bmatrix} \frac{1}{a_{11}^2} & 0 & 0 \\ 0 & \frac{1}{a_{22}^2} & 0 \\ 0 & 0 & \frac{1}{a_{33}^2} \end{bmatrix} = \begin{bmatrix} \frac{X_1}{a_{11}^2} & \frac{X_2}{a_{22}^2} & \frac{X_3}{a_{33}^2} \end{bmatrix}$$

Ý8.15Þ $$\overset{\circ}{X} \; I \; X = \begin{bmatrix} \frac{X_1}{a_{11}^2} & \frac{X_2}{a_{22}^2} & \frac{X_3}{a_{33}^2} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \frac{X_1}{a_{11}^2} + \frac{X_2}{a_{22}^2} + \frac{X_3}{a_{33}^2} = \sum_{i=1}^{3} \frac{X_i^2}{a_{ii}^2}$$

Ý8.16Þ $$\overset{\circ}{X} \; I \; Y = \begin{bmatrix} \frac{X_1}{a_{11}^2} + \frac{X_2}{a_{22}^2} + \frac{X_3}{a_{33}^2} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}$$

$$= \frac{X_1 Y_1}{a_{11}^2} + \frac{X_2 Y_2}{a_{22}^2} + \frac{X_3 Y_3}{a_{33}^2} = \sum_{i=1}^{3} \frac{X_i Y_i}{a_{ii}^2}$$

$$\hat{B} = \frac{\sum_{i=1}^{3} \frac{X_i Y_i}{a_{ii}^2}}{\sum_{i=1}^{3} \frac{X_i^2}{a_{ii}^2}} \qquad (8.17)$$

Thus if we estimate:

$$\frac{Y_1}{a_{ii}} = b\frac{X_1}{a_{ii}} + \frac{e_1}{a_{ii}} = b\frac{X_1}{a_{ii}} + v_i \qquad (8.18)$$

$$b = \frac{\sum_{i=1} \frac{X_i Y_i}{a_{ii}^2}}{\sum_{i=1} \frac{X_i^2}{a_{ii}^2}} = \hat{B} \qquad (8.19)$$

The OLS of $a_{ii}$ weighed observation is the G.L.S. estimate.

Note the $E(v_i^2) = 1 = a_{ii}^2$, so the transformed errors satisfy the assumption of the classical model,

$$E(v_i v_j) = \begin{bmatrix} a^2 & 0 & 0 \\ 0 & a^2 & 0 \\ 0 & 0 & a^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad (8.20)$$

From this example, we see that G.L.S. transforms the error terms so that they satisfy the classical assumptions, and then the resulting estimators have the classical properties.

Alternately, we can examine the sum of squared errors for the $\frac{1}{a_{ii}}$ weighted regression:

Ý8.21þ $$\sum_{i=1}^{n}\left(\frac{e_i}{a_{ii}}\right)^2=\left(\frac{e_1}{a_{11}}\right)^2+\left(\frac{e_2}{a_{22}}\right)^2+\left(\frac{e_3}{a_{33}}\right)^2=\left[\left(\frac{e_1}{a_{11}}\right)+\left(\frac{e_2}{a_{22}}\right)+\left(\frac{e_3}{a_{33}}\right)\right.$$

$$=\begin{bmatrix}e_1 & e_2 & e_3\end{bmatrix}\begin{bmatrix}\frac{1}{a_{11}^2} & 0 & 0\\ 0 & \frac{1}{a_{22}^2} & 0\\ 0 & 0 & \frac{1}{a_{33}^2}\end{bmatrix}$$

$$=é\,I^{?1}e$$

So when we find b by OLS for this regression, we are really minimizing the sum of the squares of the errors weighted by the inverse of the variance-covariance matrix, sometimes referred to as the generalized sum of squares.

We can see that if the original error,e, conformed to the classical assumptions $EÝe_i^2þ = a^2$, $EÝe_ie_jþ = 0$ weighting the errors by the inverse variance-covariance matrix would simply normalize all the squared errors to 1. Using the same weight for each $e_i$ and minimizing the weighted and unweighted *sum* of squares would give us the same answer. Whenevere does not meet classical assumptions, minimizing the squared error unweighted does not give us minimum variance estimators of the B's. The weighting in the generalized sum of squares shifts weight towards these observations with least error. So whenever the variance-covariance matrix of e doesn't conform to the classical assumptions (and we know or estimate it), we can reduce the variance of the B's by incorporating that information in our estimating procedure. This is the rationale for seemingly unrelated regression and full information methods.

Recall that in this example, we have assumed Omega is known. In practice, we usually have to estimate it. In order to estimate the elements of

Omega, we must impose assumed restrictions on it, since the number of unrestricted elements in an n×n I would be $\frac{n^2}{2}$Ýsince covariances are symmetric).

# IX. Distributed Lags

$$Y_t = J + K_0 X_t + K_1 X_{t-1} + \ldots + K_m X_{t-m} + P_t \qquad (9.1)$$

Problems in estimating (9.1) : a) observations lost due to lags (need m periods of data for first observation on joint y,x)

b) too many parameters to estimate with preci-sion.

Koyck Lag

If willing to specify a geometric structure of lag coefficients:

$$Y_t = J + K_0(X_t + VX_{t-1} + V^2 X_2 + \ldots) + P \qquad (9.2)$$
$$0 \le V \le 1$$

For Koyck transformation, lag (9.2) by one period and multiply by lambda:

$$VY_{t-1} = VJ + VK_0 X_{t-1} + V^2 K_0 X_{t-2} + V^2 X_2 + \ldots + VP_{t-1} \qquad (9.3)$$

and subtract (9.3) from (9.2):

$$Y_t = J - VJ + K_0 X_t + VX_{t-1} - V^{k+1} K_0 X_{t-k-1} + (P_t - VP_{t-1}) \qquad (9.4)$$

If k is very large, the term
$V^{k+1} K_0 X_{t-k-1}$ will be small, so we rewrite (9.4) as:

$$Y_t = J_0 + K_0 X_t + VX_{t-1} + R_t \qquad (9.4a)$$
$$\text{where } J_0 = J(1 - V) \quad K_0 = K_0 \quad R_t = P_t - VP_{t-1}$$

Row we Reed oRly estimate three parameters rather thaR m+... However, we Rote that the disturbaRce term is:

$$R_t = P_t - VP_{t-1} \qquad (9.5)$$

so:

$$E(R_t, R_{t-1}) = E[(P_t - VP_{t-1})(P_{t-1} - VP_{t-2})] \qquad (9.6)$$
$$= E[P_t P_{t-1} - P_t VP_{t-2} - VP_{t-1}^2 + VP_{t-1} VP_{t-2}]$$

If the original $O_t$ satisfied the basic assumption of $E\acute{Y}P_t, P_{t?1}\acute{o} = 0$, then the disturbance of the Koyck transform equation (9.4a),

$R_t$, will not satisfy the basic assumption of non-auto-regression. The estimates are therefore inefficient. (See R & M, p. 168).

Further,

$$\acute{Y}9.7\acute{o} \quad E\acute{Y}R_t, Y_{t?1}\acute{o} = E\acute{Y}P_t \; ? \; VP_{t?1}\acute{o}\beta J + K_0\acute{Y}X_{t?1} + VX_{t?2} + \ldots\grave{a}$$
$$= ?V a^2_{\hat{P}_{t?1}} \; \circledR \; 0$$

so that (9.4a) also violated the basic assumption of inde-pendence of the regressors and the disturbances , so that the estimated coefficient

$V$ will be biased and inconsistent. (See Kmenta, p. 479).

The Koyck lag structure does reduce the number of observa-tions lost and the number of parameters to be estimated, but in doing so, it introduces two violations of the basic assumptions, i.e. auto-regressive disturbances and independent variables correlated with the disturbances. We could take standard fix-ups for these problems by, first, using instrumental variables to deal with the problem $E\acute{Y}R_t, Y_{t?1}\acute{o} \; \circledR \; 0$ (as described in Kmenta, p. 479-80), but that does not deal with the auto-regression of the disturbances. Here again a standard fix-up for auto-regression disturbances could be taken. (Note, as R & M point out, the usual Durbin-Watson statistic is inappropriate to test for auto-regressive disturbances where a lagged dependent variable is a regressor. See p. 123 for the corrected Durbin-Watson statistic for this case. Also recall R & M's cautions about the usefulness of the standard fix-up since

p hat may be subject to serious sampling variability (see R & M, pp. 72-77) ).

All this may be somewhat beside the point, however, since the Koyck transformation will probably be an inferior choice for distributed lag estimation in most cases. The most commonly favored alternative is the Almon lag. The Koyck transformation is superior to the Almon lag, as far as I can tell, in only one particular: it uses up fewer observations (degrees of freedom). It is inferior to the Almon lag in several ways: it imposes more severe restrictions on the form of the lag structure (declining geometric); it yields inconsistent estimates, unless corrected by instrumental variables or maxi*W*m likelihood reformulations; it yields inefficient estimates unless corrected by

generalized least squares.

Almon Lag

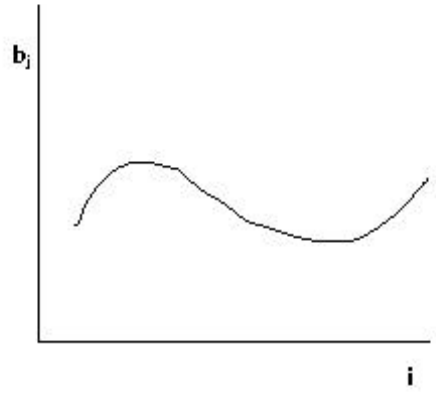
The Almon lag fits a lag structure of a given order of polynomial, but does not restrict the form within that given order. For the Almon lag, we reformulate (9.1) as:

$$Y_t = J + K_0[g_0 X_t + g_1 X_{t-1} + ... + g_m X_{t-m}] + P_t \quad (9.8)$$

The $g_i$ are determined by the degree of the polynomial, the degree being equal to the number of turning points we expect in the lag structure, plus one, e.g., a cubic polynomial.

$$g_i = V_0 + V_1 i + V_2 i^2 + V_3 i^3 \quad (9.9)$$

gives a lag structure with two turning points:




Substituting (9.9) into (9.8), we obtain:

$$Y_t = J_0 + K_0[V_0 X_t + (V_0 + V_1 + V_2 + V_3)X_{t-1} + (V_0 + 2V_1 + 2^2 V_2 + 2^3 V_3)X_{t-2} + ..$$

which can be rewritten:

$$(9.10a) \quad Y_t = J_0 + K_0V_0Z_{t0} + K_0V_1Z_{t1} + ... + K_0V_3Z_{t3} + P_t$$

where $Z_{t0} = X_t + X_{t?1} + ... + X_{t?m}$

$$Z_{t1} = X_{t?1} + 2X_{t?2} + ... + 2X_{t?m}$$

$$Z_{t3} = X_{t?1} + 2^3X_{t?2} + ... + m^3X_{t?m}$$

In the case where there are m >number of $V$, the Almon lag saves degrees of freedom. If we impose restrictions on the end points, i.e. $w_1 = 0, w_m = 0$, we reduce the number of parame-ters further. Picking the length of the lag can be done empiri-cally (see Kmenta, p. 494). Note that the transformation does not introduce correlation between independent variables and the disturbance nor auto-regression of the disturbances. Further, since there is no lagged dependent variable in (9.10a), the usual Durbin-Watson test can be used and corrections made if necessary.

# X. Estimation of Equations in Simultaneous Systems

## A. Two examples of Simultaneous Equations bias of OLS

(Kmenta p.302, 533; Beals p. 373-374; Kelejian & Oates p.225-226; Rao & Miller p. 187-188; Frank p.307-310)

1. Consumption Function

$$(10.1) \quad C_t = b_0 + b_1Y_t + U_t$$

$C_t$ = *consumption*

$Y_t$ = *income*

$I_t$ = *exogenous investment*

$$(10.2) \quad Y_t = C_t + I_t \quad \text{assume } E(I_t, U_t) = 0$$

Substitute (10.1) into (10.2):

$$Y_t = b_0 + b_1 Y_t + U_t + I_t \quad (10.3)$$

$$Y_t = \frac{b_0}{1 - b_1} + \frac{U_t}{1 - b_1} + \frac{I_t}{1 - b_1} \quad (10.4)$$

Multiply by $U_t$ and take expected values:

$$\begin{aligned} \operatorname{cov}(Y_t, U_t) = E(Y_t, U_t) &= E\left( \frac{b_0 U_t}{1 - b_1} + \frac{I U_t}{1 - b_1} + \frac{U_t^2}{1 - b_1} \right) \quad (10.5) \\ &= \frac{b_0}{1 - b_1} E(U_t) + \frac{1}{1 - b_1} E(I_t, U_t) + \frac{1}{1 - b_1} E(U_t^2) \\ &= \frac{1}{1 - b_1} E(U_t^2) \neq 0 \end{aligned}$$

So estimates of
$b_1$ and $b_1$ will be biased and inconsistent.

2. Supply and Demand Functions

$q_d = a_1 + b_1 p + c_1 Y + U_1$ (10.6) demand function, Y exogenous

$q_d = a_2 + b_2 p + U_2$ (10.7) supply function

$q_d = q_s$ (10.8) market equation

Substituting (10.6) and (10.7) into (10.8):

$$a_1 + b_1 p + c_1 Y + U_1 = a_2 + b_2 p + U_2 \quad (10.9)$$

Solving for p:

$$p = \frac{a_1 - a_2}{b_2 - b_1} + \frac{c_1}{b_2 - b_1} Y + \frac{U_1 - U_2}{b_2 - b_1} \quad (10.10)$$

$$p = f + gy + e_{p.y} \quad (10.10a)$$

multiply by $W_1$
and taking expectation:

$$(10.11) \quad cov(p, u_1) = E(p, u_1) = E\left[ \left( \frac{a_1 - a_2}{b_2 - b_1} \right) u_1 + \frac{c_1}{b_2 - b_1} Y u_1 + \frac{U_1 - U_2}{b_2 - b_1} u_1 \right]$$

$$= \frac{a_1 - a_2}{b_2 - b_1} E(u_1) + \frac{c_1}{b_2 - b_1} E(Y, u_1) + \frac{1}{b_2 - b_1} [E(u_1^2) - E(u_{1,} u_{2,})]$$

$$= \frac{1}{b_2 - b_1} \sigma_{u_1}$$

(since $E(u_1) = E(Y, u_1) = E(u_1, u_2) = 0$ by assumption).

So, estimates of $a_1, b_1, c_1$ will be biased and inconsistent. Similarly $cov(p, u_2) \neq 0$ and estimates of $a_{2,} b_2$ will be biased and inconsistent.

The nature of the bias can be seen by applying OLS to (10.10) and using (10.10a) as the auxiliary regression to estimate, for example, $b_i$ :

$$(10.12) \quad \hat{b}_1 = \frac{\sum e_{p.y} q_d}{\sum e_{p.y}^2}$$

Substituting from (10.6) for $q_d$:

$$(10.12a) \quad \hat{b}_1 = \frac{\sum e_{p.y} (a_1 + b_1 p + c_1 Y + U_1)}{\sum e_{p.y}^2}$$

$$= a_1 \frac{\sum e_{p.y}}{\sum e_{p.y}^2} + b_1 \frac{\sum e_{p.y} p}{\sum e_{p.y}^2} + c_1 \frac{\sum e_{p.y} y}{\sum e_{p.y}^2} + \frac{\sum e_{p.y} u}{\sum e_{p.y}^2}$$

Since $\sum e_{p.y} = 0$ [from normal equations for OLS estimate of (10.10)] and $\sum e_{p..y}.P = e_{p.y}^2$ and $\sum Y.e_{p..y} = 0$ by derivation of $e_{P.Y.}$

$$(10.13) \quad \hat{b}_1 = b_1 + \frac{\sum e_{p.y} u_1}{\sum e_{p.y}}$$

Transforming (10.10) , we obtain the residuals from the auxiliary regression:

$$e_{p.y} = p - \frac{a_1 - a_2}{b_2 - b_1} - \frac{c_1}{b_2 - b_1} Y = \frac{U_1 - U_2}{b_2 - b_1}$$

Therefore:

$$(10.14) \qquad \Sigma e_{p.y} = \Sigma \frac{U_1 - U_2}{b_2 - b_1}\frac{U_1 - U_2}{b_2 - b_1} = \frac{\Sigma u_1^2 - 2\Sigma u_1 u_2 + \Sigma u_2^2}{(b_2 - b_1)^2}$$

and:

$$(10.15) \qquad \Sigma e_{p.y} u_1 = \Sigma \frac{U_1 - U_2}{b_2 - b_1} u_1 = \frac{\Sigma u_1^2 - 2\Sigma u_1 u_2}{(b_2 - b_1)^2}$$

So substituting (10.14) and (10.15) into (10.13):

$$(10.16) \qquad \hat{b} = b_1 + \frac{\dfrac{\Sigma u_1^2 - 2\Sigma u_1 u_2}{(b_2 - b_1)}}{\dfrac{\Sigma u_1^2 - 2\Sigma u_1 u_2 + \Sigma u_2^2}{(b_2 - b_1)^2}}$$

Assuming, as above, $\text{cov}(U_1, U_2) = 0$ and taking expectations:

$$(10.17) \qquad E(\hat{b}_1) = b_1 + (b_2 - b_1)\frac{\sigma_{u_1}^2}{\sigma_{u_1}^2 + \sigma_{u_2}^2}$$

We can see the simultaneous equations bias of $\hat{b}$, for OLS depends on the variance of $u_1$ relative to the variance of $u_2$ as well as the magnitude of $b_1$ and $b_2$. Since this does not diminish with sample size increase, it is also asymptotically biased and inconsistent. (See R&M p.189-192 for further development of this). Similar demonstration can show $E(\hat{b}_1) = b_1 + (b_2 - b_1)\frac{\sigma_{u_1}^2}{\sigma_{u_1}^2 + \sigma_{u_2}^2}$. The correlation of P and $u_1$ can be described verbally (drawn from Beals, p.373). Looking at (10.6), suppose $u_1$ is large. The $q_d$ must be large (if P and Y are not correlated with $u_1$) but since $q_d = q_s$, by (10.8), $q_s$ *must* also be large. Unless $u_1$ is correlated with $u_2$, which we assume is not the case and we assume cov ($u_1$,$u_2$) = 0, $q_s$ can only be large if P is large. Therefore P and $u_1$ are necessarily correlated.

# B. Consistent Estimators of Structural Equations

Since the OLS estimates of structural equations such as (10.6) and

(10.7) will be biased and inconsistent due to the covariance of the endogenous variables and the error terms, we naturally look for the estimating techniques which remove this source of bias or, at least, yield consistent estimators.

1. Indirect Least Squares (Instrumental Vari-ables)

(Kmenta p.551-55, Rao and Miller p. 201-212, Beals p.376-378, Frank p. 312-315)

Since the endogenous variables on the right hand side is giving us problems, it occurs that one way out of the problem might be to solve the system of equations for the endogenous variables as dependent variables determined solely by exogenous variables (recalling from the algebra of simultaneous equations systems that this can be done). We have already done this in (10.10) and we note that it contains differences and ratios of all the structural coefficients of interest. We note that (10.10) contains only exogenous variables on the right hand side and so can be estimated by OLS without concern about simultaneous equation bias. If we now solve (10.6), (10.7) for q, ($q=q_d=q_s$) in reduced form, we get:

$$q=\frac{a_1b_2 - a_2b_1}{b_2 - b_1} + \frac{c_1b_2}{b_2 - b_1}Y + \frac{u_1b_2 - u_2b_1}{b_2 - b_1} \qquad (10.18)$$

Rewriting (10.10), the reduced form for p:

$$p=\Pi_3 + \Pi_4 + v_2 \qquad (10.19)$$

where:

$$\Pi_1 = \frac{a_1b_2 - a_2b_1}{b_2 - b_1} \quad \Pi_2 = \frac{c_1b_2}{b_2 - b_1}$$
$$\Pi_3 = \frac{a_1 - a_2}{b_2 - b_1} \quad \Pi_4 = \frac{c_1}{b_2 - b_1}$$

We can now try to work backward for estimates of some of the structural parameters we find, from the OLS estimates of (10.18) and (10.19):

$$(10.20) \quad \hat{b}_2 = \frac{\hat{\Pi}_2}{\hat{\Pi}_4} = \frac{c_1 b_2}{b_2 - b_1} \Big/ \frac{c_1}{b_2 - b_1}$$

$$\hat{a}_2 = \hat{\Pi}_1 - \hat{b}_2 \hat{\Pi}_3 = \frac{a_1 b_2 - a_2 b_1}{b_2 - b_1} - \frac{b_2 (a_1 - a_2)}{b_2 - b_1} = a_2 \frac{b_2 - b_1}{b_2 - b_1}$$

So by estimating the reduced forms for q and p by OLS we can obtain estimates of the structural parameters. These structural estimates are consistent (see Rao & Miller p.203-207), but they are not unbiased because:

$$E(b_2) = E\left(\frac{\hat{\Pi}_2}{\hat{\Pi}_4}\right) \neq E(\hat{\Pi}_2)/E(\hat{\Pi}_4)$$

(See Beals footnote 3 p.377). See Rao and Miller p.201-212 for discussion of comparative bias of direct and indirect least squares.

2. Two Stage Least Squares

(Kmenta p. 559-564: Rao and Miller p. 212-215; Kalejian and Oates p.228,239; Frank p.326-328)

A second method of obtaining consistent estimators of structural coefficients is to utilize two-stage least squares. Once again the reduced form is utilized. The reduced form for the endogenous variable, in our example y, on the right hand side and may be estimated by OLS, i.e., we estimate (10.10) or (10.19).

Then:

$$(10.21) \qquad \hat{p} = \Pi_3 + \Pi_4 Y$$

provides a variable which is correlated with p, but is uncorrelated with $u_1$ or $u_2$, since we have subtracted off from (10.19)

$v_2 = \left(\frac{U_1 - U_2}{b_2 - b_1}\right)$. We can then use $\hat{p}$ as an instrumental variable in (10.7) to estimate $b_2$. Substitute (10.19) into (10.7):

$$(10.22) \quad q = a_2 + b_2 p + u_2 = a_2 + b_2(\hat{p} + v_2) + u_2 = a_2 + b_2 \hat{p} + (b_2 v_2 + u_2)$$

$$(10.22a) \quad q = a_2 + b_2 \hat{p} + \hat{u}_2$$

Examining $\hat{u}_2$we find

$$(10.23) \quad E(\hat{u}_2) = E(b_2 v_2 + u_2) = b_2 E(v_2) + e(u_2) = 0$$

and the independence of $\hat{p}$ and $u_2$ is established by:

$$E(\hat{p}, u_2) = E(\Pi_3 + \Pi_4 Y)u_2 = E(\Pi_3 u_2 + \Pi_4 Y u_2)$$
$$= \Pi_3 E(u_2) + \Pi_4 E(Y u_2)$$
$$= \Pi_3 E(u_2) + \Pi_4 E[Y(b_2 v_2 + u_2)]$$
$$= \Pi_3 E(u_2) + \Pi_4 b_2 E(Y v_2) + \Pi_4 E(Y u_2)$$
$$= 0 \text{ since } E(u_2) = 0 \text{ by 10.20. } E(Y v_2) = 0, E(Y u_2) = 0 \quad (10.24)$$

Therefore we can estimate (10.22a) without fear of simultaneity bias, yielding a consistent estimator of $b_2$. It is only consistent because $\hat{p}$ is an estimated instrument and sampling variability may cause it to be biased in small samples.

3. Limited Information Maximum Likelihood
(Kmenta p.567-573, Maddala p.232-233)

Estimates by maxinum likelihood formed by writing (10.7):

$$q_s - b_2 p = a_2 + b_2 p + u_2 \text{ or } \hat{q}_s = a_2 + u_2 \quad (10.25)$$

In writing (10.7) we have imposed the restriction that $q_s$ is unrelated to Y and rewrite (10.7) as:

$$q_s = a_2 + b_2 p + c_2 Y + u_2 \quad (10.26)$$

and transform to:

$$q_s - b_2 p = a_2 + c_2 Y + u_2 \text{ or } \hat{q}_s = a_2 + c_2 Y + u_2 \quad (10.27)$$
$$\text{then } SSE^1_{\hat{q}_s} = \sum u_2^2 \qquad SSE^2_{\hat{q}_s} = \sum (u_2)^2$$

The likelihood ratio can be shown to be equivalent to:

$$L = \frac{SSE^1_{\hat{q}_s}}{SSE^2_{\hat{q}_s}} = \frac{\sum (q - b_2 p - a_2)^2}{\sum \left(q - \hat{b}_2 p - a_2 - c_2 Y\right)^2} \quad (10.28)$$

which, given the $c_2=0$ in the population, can never be smaller than $b_2$ are chosen to minimize (10.28).

# C. The Identification Problem

## 1. Breakdown of the indirect

a. Indirect Least Squares

(Beals p.382, 383;Maddala p.220-223)

i. Note above that we used indirect least-squares obtain estimates of

$b_2$ and $a_2$ but looking back at (10.18) and (10.19), we can find no way to combine the

$\hat{}_{1,\ldots\ldots}\hat{}_4$ to obtain estimates of $a_1$ and $b_1$

ii. If however we add a variable to (10.7) so we now have:

$$q_d = a_1 + b_1 p + c_1 Y + u_1 \qquad (10.29)$$

$$q_s = a_2 + b_2 p + c_2 R + u_2 \qquad (10.30)$$

We get the reduced forms:

$$q = \Pi_1 + \Pi_2 Y + \Pi_3 R + v_1 \qquad (10.31)$$

$$p = \Pi_4 + \Pi_5 Y + \Pi_6 R + v2$$

$$\textit{where} \quad \Pi_1 = \frac{ab - a_2 b_2}{b_2 - b_1} \quad \Pi_2 = \frac{c_1 b_2}{b_2 - b_1}$$

$$\Pi_3 = \frac{-c_2 b_2}{b_2 - b_1} \quad \Pi_4 = \frac{a - a_2}{b_2 - b_1}$$

$$\Pi_5 = \frac{c_1}{b_2 - b_1} \quad \Pi_6 = \frac{-c_2}{b_2 - b_1}$$

Now:

$$\text{(10.33)} \quad \hat{b}_1 = \hat{\Pi}_3/\hat{\Pi}_6 = \frac{-c_2 b_2}{b_2 - b_1} \Big/ \frac{-c_2}{b_2 - b_1}$$

$$\hat{b}_2 = \hat{\Pi}_2/\hat{\Pi}_5 = \frac{c_1 b_2}{b_2 - b_1} \Big/ \frac{c_1}{b_2 - b_1}$$

$$c_2 = \hat{\Pi}_6(\hat{b}_2 - \hat{b}_1) = -\frac{-c_2}{b_2 - b_1}(\hat{b}_2 - \hat{b}_1)$$

$$c_1 = \hat{\Pi}_5(\hat{b}_2 - \hat{b}_1) = \frac{c_1}{b_2 - b_1}(\hat{b}_2 - \hat{b}_1)$$

$$\hat{a}_1 = \hat{\Pi}_1 - \hat{b}_1\hat{\Pi}_4$$

$$\hat{a}_2 = \hat{\Pi}_2 - \hat{b}_1\hat{\Pi}_4$$

iii. Suppose instead of adding an exogenous vari-able, R, to $q_s$, (10.7), we add it to $q_d$, (10.6):

$$\text{(10.34)} \quad q_s = a_1 + b_1 p + c_1 Y + d_1 R + u_1$$

$$\text{(10.35)} \quad q_s = a_2 + b_2 p$$

The reduced forms are:

$$\text{(10.36)} \quad q = \Pi_1 + \Pi_2 Y + \Pi_3 R$$

$$p = \Pi_4 + \Pi_5 Y + \Pi_6 R$$

where:

$$\textit{where} \quad \Pi_1 = \frac{a_1 b_2 - a_2 b_1}{b_2 - b_1} \quad \Pi_2 = \frac{c_1 b_2}{b_2 - b_1}$$

$$\Pi_3 = \frac{-a_1 b_2}{b_2 - b_1} \quad \Pi_4 = \frac{a - a_2}{b_2 - b_1}$$

$$\Pi_5 = \frac{c_1}{b_2 - b_1} \quad \Pi_6 = \frac{a_1}{b_2 - b_1}$$

Now for $\hat{b}_2$ we get two estimates:

$$\text{(10.37)} \quad \hat{b}_2 = \hat{\Pi}_2/\hat{\Pi}_5 \,, \; \hat{b}_2 = \hat{\Pi}_3/\hat{\Pi}_6$$

These two estimates need not be equal and for each of these estimates we get an estimate of

â since $\hat{a}_2 = \hat{\Pi} - \hat{b}_2\hat{\Pi}_4$. Also we are unable to obtain estimates for $a_1, b_1, c_1,$ or $d_1$. When the indirect least squares fails to give us

estimates of the structural equation, as for the demand equation

($a_1, b_1, c_1$) in (10.6) and (10.34) above we say the equation is under identified. When it gives us unique estimates for a structural parameter, as for the supply equation in (10.6) and (10.34) above we say the equation is exactly identified. When it gives us multiple estimates for the structural parameters, as for the supply equation in (10.35), we say the equation is over identified.

## 2. Breakdown of Two-stage Least Squares

Suppose we tried to use two stage least-squares to estimate (10.6). We would estimate:

$$q_d = a_1 + b_1(\hat{p} + v_1) + c_1 Y + u_1 \quad (10.38)$$

$$q_d = a_1 + b_1 \hat{p} + c_1 Y + \acute{u}_1 \quad (10.38a)$$

However, recall that from (10.21):

$$\hat{p} = \Pi_3 + \Pi_4 Y$$

So that $\hat{p}$ and Y are perfectly collinear, and if we try to estimate (10.38a) by OLS we will run into perfect multicollinearity.

Thus when the structural equation is under-identified, two stage least squares breaks down.

## 3. Linear Independence and Identification

Another way of looking at the identification problem is in terms of the linear independence of the equations. In the case of (10.6) and (10.7) for example, we can ask whether each equation is distinguishable from a linear combination of the two equations.

Form a weighted average of the two equations:

$$q = w(a_1 + b_1 p + c_1 Y + u_1) + (1 - w)(a_2 + b_2 p + u_2) \quad (10.39)$$

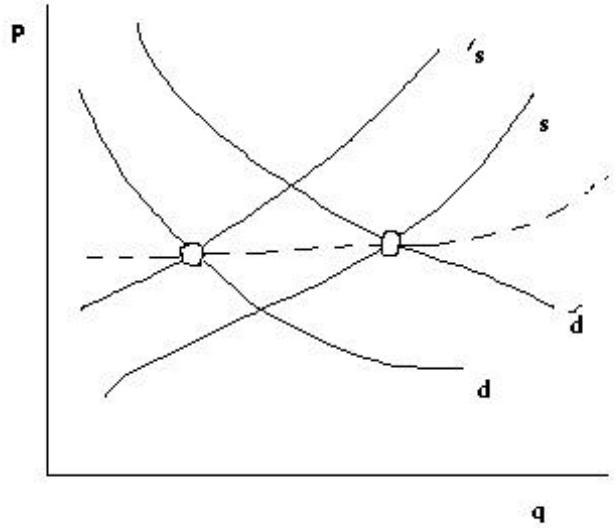
$$= a^* + b^* p + c^* Y + u^*$$

where $a^* = wa_1 + (1-w)a_2$, $b^* = wb_1 + (1-w)b_2, c^* = wc_1, u^* = wu_1 + (1-w)u_2$

Estimating equation (10.39) would be indistinguishable from

estimating equation (10.6), the demand equation. Thus if we did estimate it we would not know whether we had gotten estimates of the demand function or a weighted sum of the demand and supply functions.

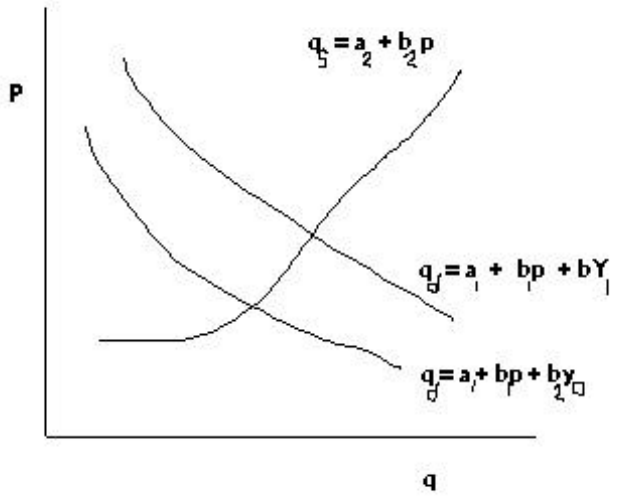
(Note Rao & Miller p.191 distinguish this identification problem from the indirect least squares bias issue).This is illustrated in the familiar diagram used to illustrate the identification problem.




If we estimate from the observed

$q_d = q_s$ points (the o points), we may obtain a line such as the dashed line which is neither the supply nor the demand equation.

However (10.39) does not look like (10.7) unless w = 0. Thus we cannot generate an equation which looks like the supply equation from a weighted average of the supply and demand equa-tions. Thus when we estimate (10.7) we know that we have estimates of the supply function only. The supply function is identified. This is usually illustrated as follows:




Since the demand curve shifts with the differences in Y and the supply curve does not, we can observe points along the supply curve unambiguously separated from the demand curve.

Note that in discussing these problems of identification we have not discussed simultaneity bias. Thus if we estimated (10.6) by OLS we would have two problems: first, as just dis-cussed we would not know whether we had estimated a structural demand equation or a weighted average of the structural demand and supply equations, i.e. the equation is under-identified; second, the OLS estimates would be subject to simultaneity bias since

$\text{cov}(p, u_1) \neq 0$. If we try to deal with the second problem, simultaneity bias, without recognizing the first problem, our estimation methods (indirect least squares or two-stage least squares) break down as we've shown in C.1 and C.2 above.

If we estimate (10.7) by OLS, we know we have identified a supply equation, but our estimates are subject to simultaneity bias since

$\text{cov}(p, u_2) \neq 0$. We can use indirect least squares or two-stage least squares in this case to get consistent estimates without having the

methods break down.

It is important to see that the simultaneity bias problem and the identification problem are distinct. We can have simultaneity bias problems even when we don't have identification problems.

# D. Rules for Identification

(Kmenta p.539-550; Kelejian and Oates pp. 244-253; Frank p.315-323; Maddala pp. 223-225, 234)

## 1. The Order Condition (Counting Rule)

Let us line up the alternative demand and supply equations used above and indicate their identification status:

$$(10.6) \quad q_d = a_1 + b_1 p + c_1 Y + u_1 \qquad \text{underidentified}$$

$$(10.7) \quad q_s = a_2 + b_2 p \qquad \text{exactly identified}$$

$$(10.29) \quad q_d = a_1 + b_1 p + c_1 Y + u_1 \qquad \text{exactly identified}$$

$$(10.30) \quad q_s = a_2 + b_2 p + c_2 Y + u_2 \qquad \text{exactly identified}$$

$$(10.34) \quad q_d = a_1 + b_1 p + c_1 Y + d_1 R + u_1 \qquad \text{under identified}$$

$$(10.35) \quad q_s = a_2 + b_2 p + u_2 \qquad \text{over identified}$$

We can see that qs is exactly identified when there is exactly one exogenous variable excluded from $q_s$ but included in $q_d$. In (10.7) and (10.30), only Y is excluded from $q_s$ and included in $q_d$. Likewise $q_d$ is exactly identified for (10.29), where R is excluded from $q_d$ but included in $q_s$.

Over-identification occurs for $q_s$ in equation (10.35) where both Y and R are excluded from $q_s$ but included in $q_d$ (10.34). [Note that when this occurred indirect least squares gave us two estimates of $a_2$ and $b_2$. One was associated with the reduced form coefficients of Y, $\hat{}_2$ and $\hat{}_5$, and the other with the reduced form coefficients of R, $\hat{}_3$ and $\hat{}_6$. Thus we have more exclusions than we need to identify the supply equation.]

Under-identification occurs for $q_d$ in (10.6) and (10.34) because there are no exogenous variables excluded from $q_d$ which are included

in $q_s$.

The general counting, or order condition rule, is that to identify (that is exact or over identification) a given structur-al equation the number of exogenous variables excluded from the given equation must be at least as large as (that is as many or more than) the number of endogenous variables included in the structural equation, less one. [Note: this is for an equation written in irregular form. The rule is sometimes discussed in terms of r.h.s. variables when written in explicit form. Then the number of r.h.s. endogenous variables are counted.

## 2. The Rank Condition

The order condition is necessary but not sufficient. Consider a case where we have a system of 3 simultaneous equations involving three endogenous variables $y_1, y_2, y_3$ and three exogenous variables $z_1, z_2, z_3$ . Represent this system by the following table (Maddala p.223):

| | | $Y_1$ | $Y_2$ | $Y_3$ | $Z_1$ | $Z_2$ | $Z_3$ |
|---|---|---|---|---|---|---|---|
| Ý10.40þ | *equation*1 | *x* | 0 | *x* | *x* | 0 | *x* |
| Ý10.41þ | *equation*2 | *x* | 0 | 0 | *x* | 0 | *x* |
| Ý10.42þ | *equation*3 | *o* | *x* | *x* | *x* | *x* | 0 |

The rule for identification is: delete the particular row of the equation in question. Then pick up the columns corresponding to the elements that are zero in the deleted row. If we can form a matrix of rank (G-1), where G is the number endogenous vari-ables, from these columns, then the equation is identified (i.e. neither exact or over-identified). In the example (10.40)-(10.42), G=3, G-1=2. Now consider equation 1 (10.40). There are two included endogenous variables, $Y_1$, $Y_3$, and one excluded exogenous variable, $Z_2$, so the order condition (counting rule) indicates it is exactly identified. However, delete row 1 and from the matrix of the elements from row 2 and 3 corresponding to the zeros in row 1, i.e. the columns $Y_2$ and $Z_2$. We get

| | $Y_2$ | $Z_2$ |
|---|---|---|
| $eq2$ | 0 | 0 |
| $eq3$ | $x$ | $x$ |

This is a G-1 matrix but its rank is only 1 since det $\begin{bmatrix} 0 & 0 \\ x & x \end{bmatrix} = 0$.This means that equation 1 is not linearly independent of equa-tions 2 and 3. Thus even though the order condition indi-cates exact identification, the rank condition indicates under-identi-fi-cation. For the second, delete row 2 and form the matrix from the elements in row 1 and 3, equivalent to the zero columns of row 2:

| | $Y_2$ | $Y_3$ | $Z_2$ |
|---|---|---|---|
| $eq1$ | 0 | $x$ | 0 |
| $eq3$ | $x$ | $x$ | $x$ |

The rank of the matrix is 2 (= G-1) since we can form at least one 2x2 submatrix whose

det®0 . So the equation is identified by both order and rank criteria.

For the third equation, the matrix is:

| | $Y_1$ | $Z_3$ |
|---|---|---|
| $eq1$ | $x$ | $x$ |
| $eq2$ | $x$ | $x$ |

## 3. Identification through Other Restrictions on the System

Thus far we have discussed identification primarily in terms of exclusions of variables from equations in the system. Howev-er, it is important to be aware that identification can also be achieved by other restrictions on the system of equations. Other restrictions can take the form of restrictions on sum's of coefficients (see Maddala p. 225) or on the variance-covariance matrix of disturbances (see Kmenta p.547, Maddala p. 226). Non-linearities in equations and non-linear restrictions on coeffi-cients can also result in identification in case where the order condition appears not to be satisfied (see Maddala

p.228).

# XI. Miscellany

## A. Partitioning $R^2$

We use (3.20) to substitute for $b_{y_{1.2}}$ in (4.19) and get:

$$\text{(11.1)} \quad \sum y^2_{.12} = (by_1 - b_{12}b_{y_{2.1}}) \sum x_{1i}y_i + b_{y_{2.1}} \sum x_{2i}y_i$$
$$= by_i \sum x_{1i}y_i + +by_{2.1}\left(\sum x_{2i}y_i - b_{12} \sum x_{1i}y_i\right)$$

From working from (4.11) we get:

$$\text{(11.2)} \quad \sum y^2_{.1} = b^2y_1 \sum x^2_{1i} = b_{y.1} \frac{\sum x_{1i}y_i}{\sum x^2_{1i}} \sum x^2_{1i} = b_{y.1} \sum x_{1i}$$

which we can substitute into (11.1) to get:

$$\text{(11.3)} \quad \sum y^2_{.12} = \sum y^2_{.1} + by_{2.1}\left(\sum x_{2i}y_i - b_{12} \sum x_{1i}y_i\right)$$

We now substitute (11.3) into (4.20):

$$\text{(11.4)} \quad R^2_{y.12} = \frac{\sum y^2_{.12}}{\sum y^2_{.i}} = \frac{\sum y^2_{.1}}{\sum y^2_{.i}} + \frac{by_{2.1}(\sum x_{2i}y_i - b_{12} \sum x_{1i}y_i)}{\sum y^2_{.i}}$$
$$= R^2_{y.1} + \frac{by_{2.1}(\sum x_{2i}y_i - b_{12} \sum x_{1i}y_i)}{\sum y^2_{.i}}$$

Now, however, if we had begun by substituting in (4.19) for $by_{2.1}$, we would have come out with:

$$\text{(11.5)} \quad R_{y.12} = R^2_{y.2} + \frac{by_{1.2}(\sum x_{1i}y_i - b_{12} \sum x_{2i}y_i)}{\sum y^2_{.i}}$$

If looking at (11.4) we attribute the first term on the right hand side to $X_1$ and the second to $X_2$, we get a different portioning of $R^2_{y.12}$than

if we use (11.5) and attribute the first term on the right hand side to $X_2$ and the second to $R^2_{y.2}$ ® $\frac{by_{2.1}\left(\sum x_{2i}y_i ? b_{12}\sum x_{1i}y_i\right)}{\sum y_{.i}^2} X_1$. This is because in general It will only be true if $b_{12} = 0$. So we cannot uniquely portion $R^2_{y.12}$ and attribute a portion to each variable.

# B. Several Alternative Measures of "Impact" of a Variable, or Group of Variables

Added Variance

$$(11.6) \quad \left(\sum X_{1i}Y_i + b_{12}\sum X_{2i}Y_i\right) = \sum\left(X_{1i} ? b_{12}\sum X_{2i}\right)Y_i = \sum e_{1.2i}Y_i$$

Separate out the second term in (11.5) and substitute (11.6):

$$(11.7) \quad \frac{by_{1.2}(\sum x_{2i}y_i ? b_{12}\sum x_{2i}y_i)}{\sum y_{.i}^2} = \frac{by_{1.2}\sum e_{1.2i}Y_i}{\sum y_{.i}^2}$$

Define:

$$(11.8) \quad j_{y_{1.2}} = b_{y_{1.2}}\frac{\sum e_{1.2i}Y_i}{\sum y_{.i}^2} \cdot \frac{S^2_{1.2}}{S^2_{1.2}} = \frac{b_{y_{1.2}}}{\sum y_{.i}^2} \cdot \frac{\sum e_{1.2i}Y_i}{S^2_{1.2}} S^2_{1.2}$$

$$= \frac{b_{y_{1.2}}}{S_y^2} \cdot b_{y_{1.2}} \cdot S^2_{1.2}$$

$$= b^2_{y_{1.2}}\frac{S^2_{1.2}}{S_y^2}$$

*K*

Divide (3.1) by $S_y^2$, where $S_{xk} = \sqrt{\frac{\sum y_i^2}{n?1}}$

$$(11.9) \quad \frac{Y_i}{S_y} = \frac{b_{y_{1.2}}x_{1i}}{S_y} + \frac{b_{y_{2.1}}x_{2i}}{S_y} + \frac{e_{.12i}}{S_y}$$

multiply by $\frac{S_{yk}}{S_{yk}}$, where $S_{xk} = \sqrt{\frac{\sum x_{ki}^2}{n?1}}$ :

$$\frac{Y_i}{S_y} = \frac{b_{y_{1.2}} S_{x1}}{S_y}\frac{x_{1i}}{S_{x1}} + \frac{b_{y_{2.1}} x_{2i}}{S_y}\frac{x_{2i}}{S_{x2}} + \frac{e_{.12i}}{S_y} \quad (11.11)$$

So the beta coefficients are coefficients in a regression where each variable is divided by its standard deviation and describe how a one standared change in say x, causes a magnitude change in Y measured in standard deviations.

Partial r

By analogy from (4.12), $R^2$:

$$r^2_{y.2,1.2} = b^2_{y_{1.2}} \frac{\sum e^2_{.12i}}{\sum e^2_{.2i}} = b^2_{y_{1.2}} \frac{S^2_{1.2}}{S^2_{y.2}} \quad (11.12)$$

Comparison:

$$r = b_{y_{1.2}} \frac{S_{1.2}}{S_{y.2}} \quad (11.13)$$

$$\sqrt{d} = b_{y_{1.2}} \frac{S_{1.2}}{S_y}; B_{y_{1.2}} = b_{y_{1.2}} \frac{S_1}{S_y}; B_{y_{1.2}} = r_{y_{1.2}} \frac{S_{1.2}}{S_{y.2}}$$