9. Goodness-of-Fit Measures for Spatial Regression

Unlike Ordinary Least Squares, where there is a single dominant measure of goodness of fit – namely *R-squared* (and *adjusted R-squared*), no such dominant measure exists for more general linear models. So relative goodness of fit for models such as SEM and SLM is best gauged by employing a variety of candidate measures, and attempting to establish "dominance" in terms of multiple measures. Recall from Figure 7.7 that seven different measures were reported for each of these models. So the main objective of this section is to clarify the meaning and interpretation of these measures. To do so, we begin in Section 9.1 below with a detailed investigation of the classical R-squared measure. Our objective here is to show why it is appropriate for classical OLS but not for more general models. This will lead to "extended" R-squared measures that can be applied to both SEM and SLM.

9.1 The R-Squared Measure for OLS

To motivate *R-squared* (R^2) as a goodness-of-fit measure for OLS, we start with a simplest case of a single explanatory variable, *x*, and consider a scatter plot of data points, (y_i, x_i), $i = 1,..,n$, used to estimate a regression of *y* on *x*, as shown in Figure 9.1 below.

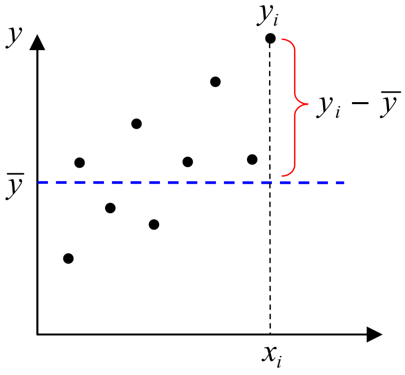

Figure 9.1. Basic Data Plot

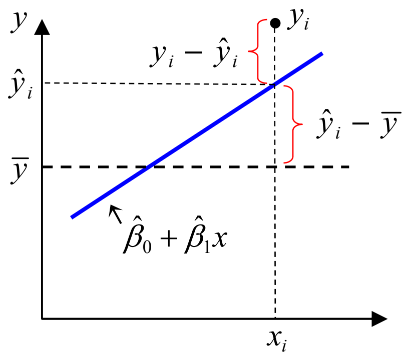

Figure 9.2. Regression Line

From an estimation viewpoint, the *regression problem* for this data is to find a linear function, $y = \beta_0 + \beta_1 x$, which best fits this data. If we let e_i denote the actual *deviation* of point (y_i, x_i) from this function (or line), so that by definition,

$$(9.1.1) \qquad y_i = \beta_0 + \beta_1 x_i + e_i \quad , \quad i = 1,..,n$$

then the *regression line* is defined to be that linear function, $y = \hat{\beta}_0 + \hat{\beta}_1 x$, which minimizes the sum of squared deviations, $\Sigma_i e_i^2$. In this case, the desired regression line is given by the blue line in Figure 9.2 [where only the single representative data point, (y_i, x_i), from Figure 9.1 is shown here].

To evaluate "goodness of fit" for this line, we first construct an appropriate benchmark for comparison. To do so, it is natural to ask how we might "fit" y-values if the explanatory variable, x, were ignored altogether. This can be accomplished by simply setting $\beta_1 = 0$, so that model (9.1.1) reduces to:

(9.1.2) $$y_i = \beta_0 + e_i \ , \ i = 1,..,n$$

In this setting the *least-squares fit*, $\hat{\beta}_0$, is now obtained by minimizing the sum of squares

(9.1.3) $$S(\beta_0) = \sum\nolimits_i (y_i - \beta_0)^2$$

By solving the first-order condition for this problem, we see that

(9.1.4) $$0 = \tfrac{d}{d\beta_0} S(\hat{\beta}_0) = 2\sum\nolimits_i (y_i - \hat{\beta}_0)(-1)$$

$$\Rightarrow \ 0 = \sum\nolimits_i (y_i - \hat{\beta}_0) = \sum\nolimits_i y_i - n\hat{\beta}_0$$

$$\Rightarrow \ \hat{\beta}_0 = \tfrac{1}{n}\sum\nolimits_i y_i = \bar{y}$$

and thus that the best least-squares fit to y in this case is precisely the *sample mean*, $\bar{y}$. [Recall also the arguments of expressions (7.1.35) and (7.1.36) in Part II]. In other words, if one ignores possible relations with other variables, then the best predictor of y values based *only* on data $(y_i : i = 1,..,n)$ is given by the sample mean of this data. So the *flat line* with value $\bar{y}$ in Figure 9.1 represents the natural *benchmark* (or *null hypothesis*) against which to compare the performance of any other possible regression model, such as (9.1.1). But for this benchmark case, it is clear that "goodness of fit" to the y-values can be measured directly in terms of their squared deviations around $\bar{y}$. This can be summarized in terms of the sum of squared deviations,

(9.1.5) $$S_y^2 = \sum\nolimits_{i=1}^{n} (y_i - \bar{y})^2$$

designated here as the *total variation* in y.[1] Note in particular that with respect to this measure, one has a perfect fit (i.e., $y_i = \bar{y}$ for all $i = 1,..,n$) if and only if $S_y^2 = 0$.

In this setting, candidate explanatory variables, x, for y only have substance in so far as they can *reduce* this benchmark level of uncertainty in y. As we shall see, it is here that

[1] Equivalently, one could take averages, and use the *sample variance*, $s_y^2 = S_y^2 / (n-1)$, of y in model (9.2). But as we shall see below, it turns out to be simpler and more direct to consider the fraction of *total variation* in y that can be accounted for by a given regression model.

the *R-squared measure* (R^2) comes into play. In short, R^2 captures the reduction in uncertainty about *y* that can be achieved by regressing *y* on any given set of explanatory variables. The key idea can be seen in an intuitive way by reconsidering the regression shown in Figures 9.1 and 9.2 above. Note first that the full deviation, $y_i - \bar{y}$, of the representative point, (y_i, x_i), from the benchmark flat line, $\bar{y}$, is shown explicitly in Figure 9.1. In the presence of the regression line in Figure 9.2, this deviation can be decomposed into two parts by using the predicted value, $\hat{y}_i$, of y_i for this regression. The lower segment, $\hat{y}_i - \bar{y}$, reflects that part of the overall deviation, $y_i - \bar{y}$, that has been "explained" by the regression line, and the upper segment, $y_i - \hat{y}_i$, reflects that part left "unexplained" by the regression. In this context, the essential purpose of R^2 is to yield a summary measure of the fractional deviations accounted for by the regression.

But notice that this example point, (y_i, x_i), has been carefully chosen so that both the deviation, $y_i - \bar{y}$, and its fractional parts are positive. To ensure positivity, it is more appropriate to ask how much of the *squared deviation*, $(y_i - \bar{y})^2$, is accounted for by the regression line. Note moreover that not all points will yield such "favorable" results for this regression. For example, data points that happen to be very close to the $\bar{y}$-line will surely be better predicted by $\bar{y}$ than by the regression, so that $(y_i - \bar{y})^2 < (y_i - \hat{y}_i)^2$. Thus the key question to be addressed how well a given regression is doing with respect to *total variation* of *y* in (9.1.5). In the context of Figure 9.2, the main result will be to show that this total variation can be decomposed into the sum of squared deviations of both $y_i - \hat{y}_i$ and $\hat{y}_i - \bar{y}$, i.e., that

$$(9.1.6) \qquad S_y^2 = \sum\nolimits_i (\hat{y}_i - \bar{y})^2 + \sum\nolimits_i (y_i - \hat{y}_i)^2 = \sum\nolimits_i (\hat{y}_i - \bar{y})^2 + \sum\nolimits_i \hat{e}_i^2$$

If these terms are designated respectively as *model variation* and *residual variation*, then this fundamental decomposition says that

$$(9.1.7) \qquad \textit{total variation} = \textit{model variation} + \textit{residual variation}$$

In this setting, the desired R^2 *measure* (also called the *Coefficient of Determination*) is taken to be the fraction of total variation accounted for by model variation, i.e.,

$$(9.1.8) \qquad \boxed{R^2 = \frac{\textit{model variation}}{\textit{total variation}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}}$$

Note from (9.1.7) that this can equivalently be written as

$$(9.1.9) \qquad \boxed{R^2 = 1 - \frac{\textit{residual variation}}{\textit{total variation}} = 1 - \frac{\sum_i \hat{e}_i^2}{\sum_i (y_i - \bar{y})^2}}$$

where this ratio can be viewed as the fraction of "unexplained" variation.

The task remaining is to demonstrate that this decomposition holds for linear regressions with any number of explanatory variables. To do so, we begin by developing a "dual" representation of the regression problem which (among other things) will yield certain key results for this construction.

9.1.1 The Regression Dual

To motivate this representation, we again begin with the simplest possible case of one explanatory variable, *x*, together with only three samples, (y_i, x_i), $i = 1,2,3$, as shown in Figure 9.3 below.

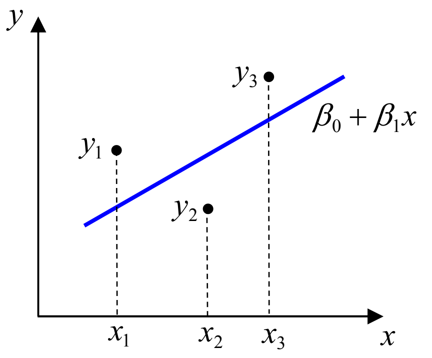

Figure 9.3 Sample Plot

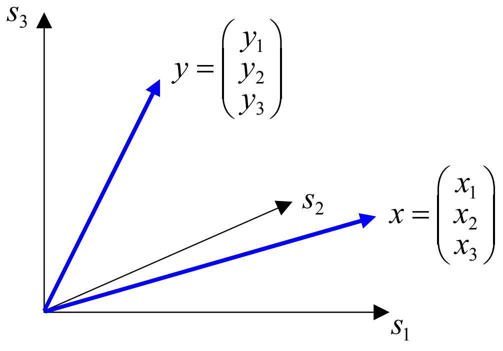
s_3
$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$
s_2
$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$
s_1

Figure 9.4. Variable Plot

This *sample plot* is simply another instance of the scatter plot in Figure 9.1, where a candidate line, $\beta_0 + \beta_1 x$, for fitting these three points is shown in blue. As in expression (9.1.1), this yields the identity,

$$(9.1.10) \qquad y_i = \beta_0 + \beta_1 x_i + e_i \quad , \quad i = 1,2,3$$

where again the desired *regression line*, $\hat{\beta}_0 + \hat{\beta}_1 x$, minimizes the sum of squared deviations, $\Sigma_i e_i^2 = e_1^2 + e_2^2 + e_3^2$. But recall that (9.1.6) can also be written in vector form as,

$$(9.1.11) \qquad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} \;\Rightarrow\; y = \beta_0 1_3 + \beta_1 x + e$$

where in particular, the vectors, $y = (y_1, y_2, y_3)'$ and $x = (x_1, x_2, x_3)'$ denote all data values of the dependent variable and explanatory variable, respectively. These two vectors are shown (in blue) in Figure 9.4, which is usually designated as the *variable plot*. Here the three axes now represent "sample dimensions", (s_1, s_2, s_3). The two representations in Figures 9.3 and 9.4 exhibit a certain *duality* property in that the roles of *samples* and *variables* are reversed. For plots such as Figure 9.3, the axes are variables and the points are samples. However, the axes in Figure 9.4 are samples and the points are variables [here drawn as vectors from the origin]. Each of these representations has its own advantages. For the present case of a single explanatory variable, x, the more standard sample plot has the advantage of allowing any number of samples to be plotted and displayed. The variable plot in Figure 9.2 is far more restrictive in this context, since the present case of a single explanatory variable with three samples is essentially the only instance in which a graphic representation is even possible.[2] Nonetheless, this dual representation, or *regression dual*, reveals key *geometric* properties of regression that simply cannot be seen in any other way. This is more apparent in Figure 9.5 below, where we have included the unit vector, $1_3 = (1,1,1)'$ from expression (9.1.11) as well.

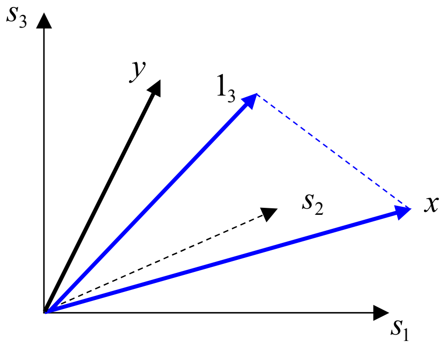

Figure 9.5. Regression Plane

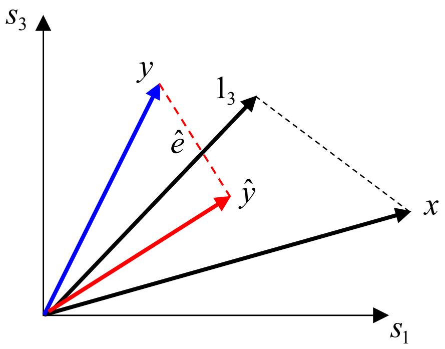

Figure 9.6. Regression as Projection

Note also that we have now colored the vectors, x and 1_3, and have connected them with a dashed line to emphasize that these two vectors define a two-dimensional *plane* called the *regression plane*. In geometric terms, the linear combinations, $\beta_0 1_3 + \beta_1 x$, in expression (9.1.10) above represent possible points on this plane (so for example, $\beta_0 = \beta_1 = 1/2$, corresponds to the point midway on dashed line joining x and 1_3). In these terms, the *regression problem* of finding a point, $\hat{\beta}_0 1_3 + \hat{\beta}_1 x$, in the regression plane that minimizes the sum of squared deviations, $\Sigma_i e_i^2$, has a very clear geometric interpretation. In particular, since the relation,

$$(9.1.12) \qquad e = y - (\beta_0 1_3 + \beta_1 x) \;\Rightarrow\; \Sigma_i e_i^2 = \| e \|^2 = \| y - (\beta_0 1_3 + \beta_1 x) \|^2$$

shows that this sum of squares is simply the *squared distance* from y to $\beta_0 1_3 + \beta_1 x$, the regression problem in this dual representation amounts geometrically to finding that point, $\hat{y} = \hat{\beta}_0 1_3 + \hat{\beta}_1 x$, in the regression plane which is *closest* to y. Without going into further details, this closest point is precisely the *orthogonal projection* of y into this

[2] Note that while more variables could in principle be included in Figure 9.4, the associated regression would be completely overdetermined. More generally, when variables outnumber sample points, there are generally infinitely many regression planes that all yield perfect fits to the data.

plane, as shown by the red arrow in Figure 9.6,[3] where the red dashed line represents the corresponding *residual vector*, $\hat{e}$, from (9.1.12), as defined by $\hat{e} = y - \hat{y}$.

This view of regression as an orthogonal projection also yields a number of insights into the algebraic structure of regression.[4] The most important of these follow from the observation that since the residual vector, $\hat{e}$, is orthogonal to the regression plane, it must necessarily be orthogonal to *every vector* in this plane. In particular, $\hat{e}$ must be orthogonal to both $\hat{y}$ and 1_3. Not surprisingly, the same is true for regressions in any dimension, n (i.e., with n samples).[5] So we can generalize these observations by first extending the present case to multiple regressions with k explanatory variables and n samples as,

$$(9.1.13) \qquad y = \hat{y} + \hat{e} = X\hat{\beta} + \hat{e} = \hat{\beta}_0 1_n + \sum\nolimits_{j=1}^{k} \hat{\beta}_j x_j + \hat{e}$$

Here $\hat{y}$ is now the orthogonal projection of y into the regression *hyperplane* spanned by the vectors $(1_n, x_1, .., x_k)$ in $\mathbb{R}^n$. Moreover (as shown in Section A2.4 of the Appendix to Part II), orthogonality between vectors can be expressed algebraically as follows: *vectors*, $a, b \in \mathbb{R}^n$, *are orthogonal if and only if their inner product is zero, i.e., if and only if* $a'b = 0$.[6] So these observations yield the following two important inner product conditions for any regression in $\mathbb{R}^n$:

$$(9.1.14) \qquad \hat{e}'\hat{y} = 0 = \hat{e}' 1_n$$

As we shall see, it is precisely these two conditions that allow the total variation of y to be decomposed as desired.

9.1.2 Decomposition of Total Variation

To develop this decomposition, we first obtain a vector representation of mean variation by employing the following notational conventions. Each sample vector, $y = (y_1, .., y_n)'$, can be transformed into *deviation form* about its about its *sample mean*,

[3] Here the s_2 axis has been hidden for visual clarity

[4] An excellent discussion of all these ideas is given in Sections 3.2.4 and 3.5 of Green (2003). In particular, his Figure 3.2 gives an alternative version of Figure 9.6. For a somewhat more advanced treatment, see Section 1.2 in Davidson and MacKinnon (1993).

[5] As an extension of footnote 2 above, it of interest to note that the present case of one explanatory variable with $n = 3$ (non-collinear) samples is in fact the unique case where *all* the relevant geometry can be seen. On the one hand, three points are just enough to yield a non-trivial regression as in Figure 9.3, while at the same time still allowing a graphical representation of variable vectors in Figure 9.4.

[6] This is perhaps the most fundamental identity linking the *algebra* of Euclidean vector spaces to their underlying *geometry*. As one simple illustrative example, note that any vectors, $a = (a_1, 0)$ and $b = (0, b_2)$, on the horizontal and vertical axes in $\mathbb{R}^2$ must be orthogonal in geometric terms, and in algebraic terms, must satisfy $a'b = a_1 \cdot 0 + 0 \cdot b_2 = 0$.

(9.1.15) $$\bar{y} = \tfrac{1}{n}\sum\nolimits_{i=1}^{n} y_i = \tfrac{1}{n}(1_n' y)$$

as follows,

(9.1.16) $$y - \bar{y}1_n = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$$

This is in fact a linear transformation on $\mathbb{R}^n$, as can be seen by defining the *n*-square *deviation matrix*,

(9.1.17) $$D = I_n - \tfrac{1}{n}(1_n 1_n')$$

and observing that for all $y \in \mathbb{R}^n$,

(9.1.18) $$Dy = (I_n - \tfrac{1}{n}1_n 1_n')y = y - \tfrac{1}{n}(1_n 1_n')y = y - \tfrac{1}{n}1_n(1_n' y) = y - \bar{y}1_n$$

Like regression, this transformation is also an orthogonal projection, where in this case D projects $\mathbb{R}^n$ onto the orthogonal complement of the unit vector, 1_n, i.e., the subspace of all vectors orthogonal to 1_n. In algebraic terms, D sends 1_n to the origin, i.e.,

(9.1.19) $$D1_n = (I_n - \tfrac{1}{n}1_n 1_n')1_n = 1_n - \tfrac{1}{n}1_n(1_n' 1_n) = 1_n - \tfrac{n}{n}1_n = 0 \quad ,$$

and leaves all vectors orthogonal to 1_n where they are. For example, the residual vector, $\hat{e}$, for any regression is orthogonal to 1_n by (9.1.10), and we see that,

(9.1.20) $$D\hat{e} = (I_n - \tfrac{1}{n}1_n 1_n')\hat{e} = \hat{e} - \tfrac{1}{n}1_n(1_n'\hat{e}) = \hat{e} - \tfrac{1}{n}1_n(0) = \hat{e}$$

More generally, as with all orthogonal projections, the matrix D is *symmetric* ($D = D'$) and *idempotent* ($DD = D$), i.e.,[7]

(9.1.21) $$DD = (I_n - \tfrac{1}{n}1_n 1_n')(I_n - \tfrac{1}{n}1_n 1_n') = I_n - \tfrac{2}{n}1_n 1_n' + \tfrac{1}{n^2}1_n(1_n' 1_n)1_n'$$

$$= I_n - \tfrac{2}{n}1_n 1_n' + \tfrac{n}{n^2}1_n 1_n' = I_n - \tfrac{1}{n}1_n 1_n' = D$$

These facts allow the total variation in (9.1.5) to be expressed directly in terms of D as,

(9.1.22) $$S_y^2 = \sum\nolimits_{i=1}^{n}(y_i - \bar{y})^2 = (y - \bar{y}1_n)'(y - \bar{y}1_n)$$

$$= (Dy)'(Dy) = y'D'Dy = y'DDy = y'Dy$$

[7] These two conditions in fact *characterize* the set of orthogonal projection matrices.

Moreover, by recalling from (9.1.13) that $y = \hat{y} + \hat{e}$, we may now employ (9.1.14), (9.1.20) and (9.1.21) to obtain the following fundamental decomposition of S_y^2:

(9.1.23)
$$\boxed{\begin{aligned} S_y^2 &= (\hat{y}+\hat{e})'D(\hat{y}+\hat{e}) = (\hat{y}'D\hat{y} + 2\hat{y}'D\hat{e} + \hat{e}'D\hat{e}) \\ &= \hat{y}'D\hat{y} + 2\hat{y}'\hat{e} + \hat{e}'\hat{e} = \hat{y}'D\hat{y} + 2(0) + \hat{e}'\hat{e} \\ &= \hat{y}'D\hat{y} + \hat{e}'\hat{e} \end{aligned}}$$

To relate this decomposition to (9.1.6), we note first that if we now denote the *residual variation* term in (9.1.6) by $S_{\hat{e}}^2$ then it follows at one that this is precisely the second term in (9.1.23), i.e, that

(9.1.24)
$$S_{\hat{e}}^2 = \sum\nolimits_{i=1}^{n} \hat{e}_i^2 = \hat{e}'\hat{e}$$

Turning next to the model variation term in (9.1.6), notice again from (9.1.14) that

(9.1.25)
$$0 = 1_n'\hat{e} = 1_n'(y - \hat{y}) = 1_n'y - 1_n'\hat{y} \Rightarrow 1_n'y = 1_n'\hat{y}$$

and thus that the mean of the regression predictions, $(\hat{y}_1,..,\hat{y}_n)$, is precisely $\bar{y}$, i.e.,

(9.1.26)
$$\tfrac{1}{n}\sum\nolimits_{i=1}^{n} \hat{y}_i = \tfrac{1}{n}(1_n'\hat{y}) = \tfrac{1}{n}(1_n'y) = \bar{y}$$

Thus if we now denote *model variation* in (9.1.6) by $S_{\hat{y}}^2$, then it follows from (9.1.17) and (9.1.26), together with the above properties of D that

(9.1.27)
$$\begin{aligned} S_{\hat{y}}^2 &= \sum\nolimits_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\hat{y} - \bar{y}1_n)'(\hat{y} - \bar{y}1_n) \\ &= (\hat{y} - [\tfrac{1}{n}1_n'\hat{y}]1_n)'(\hat{y} - [\tfrac{1}{n}1_n'\hat{y}]1_n) = (\hat{y} - \tfrac{1}{n}1_n1_n'\hat{y})'(\hat{y} - \tfrac{1}{n}1_n1_n'\hat{y}) \\ &= ([I_n - \tfrac{1}{n}1_n1_n']\hat{y})'([I_n - \tfrac{1}{n}1_n1_n']\hat{y}) = (D\hat{y})'D\hat{y} = \hat{y}'D'D\hat{y} \\ &= \hat{y}'D\hat{y} \end{aligned}$$

and thus that $S_{\hat{y}}^2$ is precisely the first term in (9.1.23). By putting these results together, we may conclude that the desired *decomposition of total variation* for y is given by

(9.1.28)
$$\boxed{S_y^2 = S_{\hat{y}}^2 + S_{\hat{e}}^2}$$

In these terms, the R-squared measure in (9.1.8) and (9.1.9) can now be re-expressed as:

$$\boxed{R^2_{OLS} = \frac{S^2_{\hat{y}}}{S^2_y} = 1 - \frac{S^2_{\hat{e}}}{S^2_y}} \tag{9.1.29}$$

where the OLS subscript is here used to emphasize that this decomposition property holds for OLS. Notice also from the nonnegativity of all terms in (9.1.28) that $0 \le R^2_{OLS} \le 1$, and thus that R^2_{OLS} can be interpreted as the *fraction* of total variation explained by a given OLS regression. For computational purposes, it is more convenient to express R-squared in vector terms as,

$$\boxed{R^2_{OLS} = \frac{\hat{y}'D\hat{y}}{y'Dy} = 1 - \frac{\hat{e}'\hat{e}}{y'Dy}} \tag{9.1.30}$$

where the latter form, in terms of *unexplained variation*, is by far the most commonly used in practice.

9.1.3 Adjusted R-Squared

While R^2_{OLS} is intuitively very appealing as a measure of goodness of fit, it suffers from certain drawbacks. Perhaps the single most important of these is that fact that the measure can *never decrease* when more explanatory variables are added to the model, and in fact it *almost always increases*. This can be most easily seen by relating residual variation to the solution of the regression problem itself. Recall that if for any given set of data, $(y_i, x_{1i},..,x_{ki})$, $i = 1,..,n$, we define the *sum-of-squares function*

$$S_k(\beta_0,\beta_1,..,\beta_k) = \sum\nolimits_i \left(y_i - \sum\nolimits_{j=0}^k \beta_j x_{ij}\right)^2 \tag{9.1.31}$$

over possible beta values $(\beta_0,\beta_1,..,\beta_k)$ [as in expression (7.1.9) of Part II], then the regression problem is to find those values $(\hat{\beta}_0,\hat{\beta}_1,..,\hat{\beta}_k)$ that minimize this function. But the *residual variation* for this regression problem, say $\hat{e}_k'\hat{e}_k$, is precisely the *value* of S_k at the *minimum*, i.e.,

$$\hat{e}_k'\hat{e}_k = \sum\nolimits_i \hat{e}^2_{ik} = \sum\nolimits_i \left(y_i - \sum\nolimits_{j=0}^k \hat{\beta}_j x_{ij}\right)^2 = S_k(\hat{\beta}_0,\hat{\beta}_1,..,\hat{\beta}_k) \tag{9.1.32}$$

$$= \min\nolimits_{(\beta_0,\beta_1,..,\beta_k)} S_k(\beta_0,\beta_1,..,\beta_k)$$

So if we add another explanatory variable, x_{k+1}, and observe that by definition $S_k(\beta_0,\beta_1,..,\beta_k)$ is just the special case of $S_{k+1}(\beta_0,\beta_1,..,\beta_k,\beta_{k+1})$ with $\beta_{k+1} = 0$, i.e., that

$$S_{k+1}(\beta_0,\beta_1,..,\beta_k,0) = \sum\nolimits_i \left(y_i - \sum\nolimits_{j=0}^k \beta_j x_{ij} - (0)x_{i,k+1}\right)^2 \tag{9.1.33}$$

$$= \sum_i \left(y_i - \sum_{j=0}^{k} \hat{\beta}_j x_{ij} \right)^2 = S_k(\beta_0, \beta_1, .., \beta_k)$$

then it follows at once from (9.1.31) through (9.1.33) that

$$(9.1.34) \qquad \hat{e}_{k+1}'\hat{e}_{k+1} = \min_{(\beta_0,..,\beta_k,\beta_{k+1})} S_{k+1}(\beta_0,..,\beta_k,\beta_{k+1})$$

$$\le \min_{(\beta_0,..,\beta_k)} S_{k+1}(\beta_0,..,\beta_k,0)$$

$$= \min_{(\beta_0,..,\beta_k)} S_k(\beta_0,..,\beta_k)$$

$$= \hat{e}_k'\hat{e}_k$$

Thus, when a new explanatory variable is added to the regression, the resulting residual variation *never increases*, and in fact must *decrease* unless the new variable, x_{k+1}, is totally unrelated to y in the sense that $\hat{\beta}_{k+1} = 0$. Finally, since $y'Dy$ is the same in both regressions, we may conclude from last term in (9.1.30) that R_{OLS}^2 never decreases, and almost always increases.[8]

This property creates serious problems when using R_{OLS}^2 as a criterion for model selection. Since R_{OLS}^2 can always be increased by adding *more* variables to a given model, this will lead inevitably to the classic problem of "overfitting the data". Indeed, for problems with n samples, it is easy to see that a perfect fit $(R_{OLS}^2 = 1)$ can be guaranteed by increasing the number of (non-collinear) explanatory variables, k, to $n-1$. For example, if there were only $n = 2$ samples, then since two points define a unique line, almost any simple regression $(k = 1)$ must yield a perfect fit.

This serves to underscore the need to modify R_{OLS}^2 to reflect the number of explanatory variables used in a given regression model. This can be accomplished by essentially "penalizing" those models with larger numbers of explanatory variables. The standard procedure for doing so is to replace R_{OLS}^2 by the following modification, $\bar{R}_{OLS}^2$, designated as *adjusted R-squared*:

$$(9.1.35) \qquad \boxed{\bar{R}_{OLS}^2 = 1 - \left(\tfrac{n-1}{n-1-k}\right)\frac{\hat{e}'\hat{e}}{y'Dy} = 1 - \left(\tfrac{n-1}{n-k}\right)(1 - R_{OLS}^2)}$$

Here the first equality is the standard definition of $\bar{R}_{OLS}^2$, and the second equality simply re-expresses this measure directly in terms of R_{OLS}^2. While this measure can be given

[8] The exact magnitude of this increase is given in Green (2003, Theorem 3.6).

some theoretical justification,[9] the popularity of $\overline{R}^2_{OLS}$ lies mainly in its simplicity and ease of interpretation as a reasonable "penalized" version of R^2_{OLS}. In particular, note that the penalty factor, $(n-1)/(n-1-k)$, must be greater than one in all cases of interest, and always increases with k. This in turn implies that $\overline{R}^2_{OLS} < R^2_{OLS}$, and that $\overline{R}^2_{OLS}$ decreases as k increases. Thus, $\overline{R}^2_{OLS}$ does indeed penalize models with larger numbers of explanatory variables. Moreover, since $\overline{R}^2_{OLS}$ approaches $-\infty$ as k approaches $n-1$, it is clear that models with numbers of variables anywhere close to the sample size will never be considered. Note however that this last property also shows that $\overline{R}^2_{OLS}$ need not be positive, and thus cannot be given any interpretation relating to the "fraction of variation explained". About all that can be said is that models with *negative* $\overline{R}^2_{OLS}$ can surely be discarded from consideration. At the other extreme, notice that penalty factor, $(n-1)/(n-1-k)$, shrinks rapidly to one as sample size, n, increases. So from a practical viewpoint, this penalty has little effect whenever sample sizes are quite large compared to the number of explanatory variables being considered. Because of this, it has been argued that $\overline{R}^2_{OLS}$ does not penalize models enough. But in any case, this measure is unquestionably preferable to R^2_{OLS} when comparing regression models of different sizes, and is far and away the most popular measure of goodness of fit in this context.

9.2 Extended R-Squared Measures for GLS

In spite of the success of R^2_{OLS} and $\overline{R}^2_{OLS}$ for OLS models, their appropriateness as goodness-of-fit measures for more general models is more problematic. Here it suffices to consider the simplest possible extension involving the GLS model in Section 7.2.2 above,

$$(9.2.1) \qquad Y = X\beta + \varepsilon\,,\;\; \varepsilon \sim N(0, \sigma^2 V)$$

with known covariance structure, V. In this modeling context, the key difficulty is that the resulting y-predictions obtained from (7.2.18) by

$$(9.2.2) \qquad \hat{y} = X\hat{\beta} = X(X'V^{-1}X)^{-1}X'V^{-1}y$$

are no longer orthogonal projections.[10] So the fundamental decomposition of total variation in (9.1.23) and (9.1.28) no longer holds, and the compelling interpretive

[9] The standard theoretical justification relies on the fact that (i) $y'Dy/(n-1)$ yields an unbiased estimate of y variance in the null model (9.1.2), (ii) $\hat{e}'\hat{e}/(n-1-k)$ yields an unbiased estimate of residual variance, σ^2, in the regression model, and (iii) the second term in (9.1.35) is precisely the ratio of these unbiased estimates. But while this argument is appealing, it does *not* imply that this ratio is an unbiased estimate of the fraction of unexplained variance. Indeed, the expectation of a ratio is almost never the same as the ratio of expectations.

[10] An excellent discussion of this issue is given in Davidson and MacKinnon (1993 ,Sections 1.2 and 9.3).

features of R^2_{OLS} now vanish. In particular, the model-oriented and error-oriented definitions of R^2_{OLS} in (9.1.30) are no longer equivalent. So there is no unambiguous way to define the "fraction of variation explained" by the given GLS model.

But as in the introductory discussion to Section 9.1 above, the residual vector, $\hat{e} = y - \hat{y}$, still captures the deviations of data, y, from their predicted values, $\hat{y}$, under any GLS model. Moreover, since $Dy = y - \bar{y}1_n$ still represents the y deviations from their least-squares prediction, $\bar{y}$, under the null model [as in (9.1.4) above], it is reasonable to gauge the goodness of fit of this model by comparing its *mean squared error*:

$$(9.2.3) \qquad MSE = \tfrac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y}_i)^2$$

with that under the null model, say

$$(9.2.4) \qquad MSE_0 = \tfrac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \bar{y})^2$$

This comparison is shown graphically in Figures 9.7 and 9.8 below:

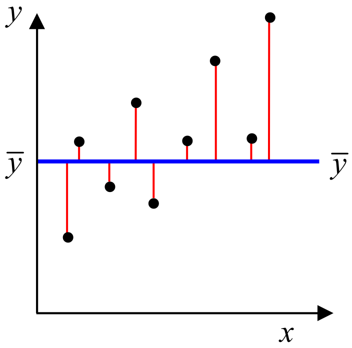

Figure 9.7. Null Deviations

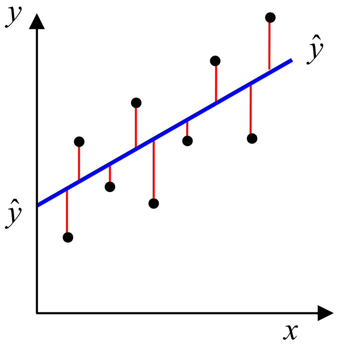

Figure 9.8. Model Deviations

In particular, the positivity (and common units) of these measures suggests that their ratio should provide an appropriate comparison, as given by

$$(9.2.5) \qquad \frac{MSE}{MSE_0} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{(y-\hat{y})'(y-\hat{y})}{(y-\bar{y}1_n)'(y-\bar{y}1_n)}$$

$$= \frac{\hat{e}'\hat{e}}{(Dy)'(Dy)} = \frac{\hat{e}'\hat{e}}{y'DDy} = \frac{\hat{e}'\hat{e}}{y'Dy}$$

which is precisely the second term in the error-oriented version of R^2_{OLS}. Finally, since smaller values of this ratio indicate better average fit relative to the null model, it follows that larger values of the difference,

$$(9.2.6) \qquad R^2_{GLS} = 1 - \frac{MSE}{MSE_0} = 1 - \frac{\hat{e}'\hat{e}}{y'Dy}$$

also indicate a better fit. To distinguish this general measure from R^2_{OLS}, it is convenient to designate (9.2.6) as *extended* R^2. This terminology also serves to emphasize that (9.2.6) *cannot* be interpreted as "explained variation" outside of the OLS case. This is made clear by the fact that extended R^2 can be *negative*. But as with adjusted R^2 for OLS, it should be clear that negative values of extended R^2 are a strong indication of poor fit. Indeed, models with higher mean squared error than y by itself can generally be ruled out on this basis alone.

Finally, as with the OLS case, it should be clear that larger numbers of explanatory variables must necessarily reduce MSE and thus increase the value of extended R^2. So goodness of fit for GLS models must be also be penalized for the addition of new variables. While the penalty ratio, $(n-1)/(n-1-k)$, in (9.1.35) is somewhat more difficult to interpret in the GLS setting,[11] it nonetheless continues to exhibit the same appealing properties discussed in Section 9.1.3 above. So in the present GLS setting, we now the designate

$$(9.2.7) \qquad \bar{R}^2_{GLS} = 1 - \left(\tfrac{n-1}{n-1-k}\right)\frac{\hat{e}'\hat{e}}{y'Dy}$$

as the appropriate *extended* form of adjusted R^2 in (9.1.35).

Before applying these extended measures to SEM and SLM, it is also of interest to note that there is an alternative approach which seeks to preserve the appealing properties of R^2_{OLS}. In particular, recall that one can convert any given GLS model to an OLS model that is equivalent in terms of parameter estimation. In the present setting, it follows from expressions (7.1.15) through (7.1.18) that if T is the Cholesky matrix for V, so that $V = TT'$, then (9.2.1) can be converted to an OLS model

$$(9.2.8) \qquad Y_o = X_o\beta + \varepsilon_o\,,\ \varepsilon_o \sim N(0,\sigma^2 I_n)$$

where these new variables are defined by

[11] While the simple "unbiasedness" argument in footnote 9 no longer holds, it can still be shown that replacing n by $n-1-k$ corrects bias in the GLS estimate of variance, $\hat{\sigma}^2$, in (7.2.20). So at least in these terms, a justification in terms of "unbiasedness" can still be made.

(9.2.9) $$Y_o = T^{-1}Y \quad , \quad X_o = T^{-1}X \quad , \quad \varepsilon_o = T^{-1}\varepsilon$$

So if goodness of fit for model (9.2.1) is now measured in terms of R^2 and $\bar{R}^2$ for model (9.2.8), then it would appear that all of the properties of these measures are preserved. In particular, if for any given y data, we set $y_o = T^{-1}y$, then the appropriate prediction, say $\hat{y}_o$, is given by

(9.2.10) $$\hat{y}_o = X_o\hat{\beta} = X_o(X_o'X_o)^{-1}X_o'y_o$$

So by setting $\hat{e}_o = y_o - \hat{y}_o$, it follows that the appropriate R-squared measure, say R_o^2, is given from (9.1.30) by

(9.2.11) $$R_o^2 = \frac{\hat{y}_o'D\hat{y}_o}{y_o'Dy_o} = 1 - \frac{\hat{e}_o'\hat{e}_o}{y_o'Dy_o}$$

Such measures are typically designated as *pseudo R-squared* measures for GLS models [see for example, Buse (1973)]. However, the most serious limitation of such measures in that they account for total variation in $y_o = T^{-1}y$ rather than in y itself. This is not only difficult to interpret, but in fact can vary depending on the factorization of covariance used. For example, the estimated SEM covariance matrix, $V_{\hat{\rho}}$ in (7.3.2) has a natural factorization in terms of the matrix, $B_{\hat{\rho}}^{-1}$, which will clearly yield different results than for the Cholesky matrix. So the essential appeal of the extended R^2 and $\bar{R}^2$ measures above is that they are *directly interpretable* in terms of y and $\hat{y}$.

9.2.1 Extended R-Squared for SEM

Turning first to SEM, recall from expression (6.1.8) that for any given spatial weights matrix, W, we can express SEM as a GLS model of the form:

(9.2.12) $$Y = X\beta + u, \quad u \sim N(0, \sigma^2 V_\rho)$$

where the *spatial covariance structure*, V_ρ, is given by

(9.2.13) $$V_\rho = (B_\rho' B_\rho)^{-1} = B_\rho^{-1}(B_\rho^{-1})'$$

with B_ρ given in terms of weight matrix, W, by

(9.2.14) $$B_\rho = I_n - \rho W$$

So for any given y data, the maximum-likelihood estimate, $\hat{y}_{SEM}$, of the conditional mean, $E(Y \mid X) = X\beta$, is given by

$$(9.2.15) \qquad \hat{y}_{SEM} = X\hat{\beta} = X(X'V_{\hat{\rho}}^{-1}X)^{-1}X'V_{\hat{\rho}}^{-1}y = X(X'B_{\hat{\rho}}'B_{\hat{\rho}}X)^{-1}X'B_{\hat{\rho}}'B_{\hat{\rho}}y$$

Finally, letting

$$(9.2.16) \qquad \hat{e}_{SEM} = y - \hat{y}_{SEM}$$

it follows from (9.2.6) that the *extended* R^2 measure for SEM is given by,

$$(9.2.17) \qquad \boxed{R^2_{SEM} = 1 - \frac{\hat{e}_{SEM}'\hat{e}_{SEM}}{y'Dy}}$$

with associated *extended* $\overline{R}^2$ measure,

$$(9.2.18) \qquad \boxed{\overline{R}^2_{SEM} = 1 - \left(\tfrac{n-1}{n-1-k}\right)(1 - R^2_{SEM})}$$

These two values are reported for the Eire data in the left panel of Figure 7.7 as

$$(9.2.19) \qquad \boxed{R^2_{SEM} = 0.3313} \quad (R^2_{OLS} = 0.5548)$$

and

$$(9.2.20) \qquad \boxed{\overline{R}^2_{SEM} = 0.3034} \quad (\overline{R}^2_{OLS} = 0.5363)$$

where the corresponding OLS values are given in parentheses. As expected, these extended measures for SEM are lower than for OLS since they incorporate more of the true error variation due to spatial dependencies among residuals.[12] So the main interest in these goodness-of-fit measures is their relative magnitudes compared to SLM, or other models which may serve to account for spatial dependencies (such as the spatial Durbin model in Section 6.3.2).

9.2.2 Extended R-Squared for SLM

Turning next to SLM, recall from (6.2.6) that this can also be expressed as a GLS model of the form:

[12] This can be seen explicitly by observing from the SEM log likelihood function in (7.3.4) that for the OLS case of $\rho = 0$, the estimate, $\hat{\beta}$, is chosen precisely to minimize mean squared error. So whenever $\hat{\rho} \neq 0$, one can expect that the associated mean squared error for SEM will be larger than this global minimum.

(9.2.21) $$Y = X_\rho \beta + u \quad , \quad u \sim N(0, \sigma^2 V_\rho)$$

where V_ρ is again given by (9.2.13) and (9.2.14) for some choice of spatial weights matrix, W, and where in this case,

(9.2.22) $$X_\rho = B_\rho^{-1} X = (I_n - \rho W)^{-1} X$$

So for any given y data, the maximum-likelihood estimate, $\hat{y}_{SLM}$, of the conditional mean, $E(Y \mid X) = X\beta$, is given in terms of (7.4.13) by

(9.2.23) $$\hat{y}_{SLM} = X_{\hat{\rho}} \hat{\beta} = X_{\hat{\rho}} (X'X)^{-1} X' B_{\hat{\rho}} y = B_{\hat{\rho}}^{-1} X (X'X)^{-1} X' B_{\hat{\rho}} y$$

Thus, by now letting

(9.2.24) $$\hat{e}_{SLM} = y - \hat{y}_{SLM}$$

it follows from (9.2.6) that the *extended* R^2 measure for SLM is given by,

(9.2.25) $$\boxed{R^2_{SLM} = 1 - \frac{\hat{e}'_{SLM} \hat{e}_{SLM}}{y'Dy}}$$

with associated *extended* $\bar{R}^2$ measure,

(9.2.26) $$\boxed{\bar{R}^2_{SLM} = 1 - \left(\tfrac{n-1}{n-1-k}\right)(1 - R^2_{SLM})}$$

These two values are reported for the Eire data in the right panel of Figure 7.7 as

(9.2.27) $$\boxed{R^2_{SLM} = 0.7335} \quad (R^2_{OLS} = 0.5548)$$

and

(9.2.28) $$\boxed{\bar{R}^2_{SEM} = 0.7224} \quad (\bar{R}^2_{OLS} = 0.5363)$$

where the corresponding OLS values are again given in parentheses. So in contrast to SEM, we see that both R^2_{SLM} and $\bar{R}^2_{SLM}$ for SLM are actually considerably higher than for OLS. The reason for this is again explained by the contrast between the "pale" effect in X and the "rippled pale" effect, $X_{\hat{\rho}}$, as illustrated in Figure 7.8 above. However, this appears to be a very exceptional case in which $\hat{y}_{SLM} (= X_{\hat{\rho}} \hat{\beta})$ happens to yield an extraordinarily good fit to y. More generally, one expects both SEM and SLM to yield extended R^2 values that are lower than R^2_{OLS}, so that the spatial components W and ρ serve mainly to capture the hidden variation arising from spatial autocorrelation effects.

9.3 The Squared Correlation Measure for GLS Models

A measure that turns out to be closely related to extended R^2 is the *squared correlation* between y and its predicted value, $\hat{y}$, under any GLS model (including OLS). Here it is again convenient to begin with the OLS case, where this measure is shown to be *identical* to R^2. We then proceed to the more general case of GLS models, including both SEM and SLM. Finally, the correlation measure itself is given a geometrical interpretation in terms of angle cosines in deviation subspaces, which helps to clarify its relevance for measuring goodness of fit.

Let us begin by recalling that the *sample correlation*, $r(x,y)$, between any pair of data vectors, $x = (x_1,..,x_n)'$ and $y = (y_1,..,y_n)'$, can be expressed in vector form by employing the properties of the deviation matrix, *D*, in (9.1.17) , (9.1.18) and (9.1.21) as follows:

$$(9.3.1) \qquad r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}}$$

$$= \frac{(x - \overline{x}_n 1_n)'(y - \overline{y}_n 1_n)}{\sqrt{(x - \overline{x}_n 1_n)'(x - \overline{x}_n 1_n)}\sqrt{(y - \overline{y}_n 1_n)'(y - \overline{y}_n 1_n)}}$$

$$= \frac{(Dx)'Dy}{\sqrt{(Dx)'Dx}\sqrt{(Dy)'Dy}}$$

$$= \frac{x'D'Dy}{\sqrt{x'D'Dx}\sqrt{y'D'Dy}}$$

$$= \frac{x'Dy}{\sqrt{x'Dx}\sqrt{y'Dy}}$$

so that *squared correlation* is always of the form

$$(9.3.2) \qquad \boxed{r^2(x,y) = \frac{(x'Dy)^2}{(x'Dx)(y'Dy)}}$$

Given this general expression, we now consider the correlation between data, *y*, and model predictions, $\hat{y}$, for the case of OLS.

9.3.1 Squared Correlation for OLS

First recall from (7.2.6) that for any given data (y, X), the predicted value, $\hat{y}$, of y is given by

$$\hat{y}_{OLS} = X\hat{\beta} = X(X'X)^{-1}X'y \tag{9.3.3}$$

In these terms, the squared correlation measure for OLS is given in terms of (9.3.2) by

$$\boxed{r^2(y, \hat{y}_{OLS}) = \frac{(y'D\hat{y}_{OLS})^2}{(y'Dy)(\hat{y}'_{OLS}D\hat{y}_{OLS})}} \tag{9.3.4}$$

With this definition, our first objective is to show that (9.3.4) is precisely the same as R^2_{OLS}. If for notational simplicity we let $\hat{y} = \hat{y}_{OLS}$ and again denote the estimated residuals for OLS by $\hat{e} = y - \hat{y}$, then it follows from expression (9.1.14) that

$$0 = \hat{y}'\hat{e} \;\Rightarrow\; \hat{y}'\hat{y} = \hat{y}'(y - \hat{e}) = \hat{y}'y - \hat{y}'\hat{e} = \hat{y}'y \tag{9.3.5}$$

and moreover that [see also (9.1.25)],

$$0 = 1_n'\hat{e} = 1_n'(y - \hat{y}) \Rightarrow 1_n'y = 1_n'\hat{y} \tag{9.3.6}$$

But given these two identities, we must have

$$\begin{aligned} \hat{y}'Dy &= \hat{y}'(I_n - \tfrac{1}{n}1_n1_n')y \\ &= \hat{y}'y - \tfrac{1}{n}1_n(1_n'y) \\ &= \hat{y}'\hat{y} - \tfrac{1}{n}1_n(1_n'\hat{y}) = \hat{y}'(I_n - \tfrac{1}{n}1_n1_n')\hat{y} = \hat{y}'D\hat{y} \end{aligned} \tag{9.3.7}$$

So it follows at once from (9.3.4) that

$$r^2(y, \hat{y}) = \frac{(y'D\hat{y})^2}{(y'Dy)(\hat{y}'D\hat{y})} = \frac{(\hat{y}'D\hat{y})^2}{(y'Dy)(\hat{y}'D\hat{y})} = \frac{\hat{y}'D\hat{y}}{y'Dy} \tag{9.3.8}$$

which together with the first (model-oriented) representation of R^2_{OLS} implies that

$$\boxed{r^2(y, \hat{y}_{OLS}) = R^2_{OLS}} \tag{9.3.9}$$

For purposes of later comparison, it follows from (9.3.9) that for the Eire case

$$\boxed{r^2(y, \hat{y}_{OLS}) = R^2_{OLS} = 0.5548} \tag{9.3.10}$$

9.3.2 Squared Correlation for SEM and SLM

By employing $\hat{y}_{SEM}$ in expression (9.2.15), it follows at once that the *squared correlation* measure for SEM is given by,

$$\boxed{r^2(y,\hat{y}_{SEM}) = \frac{(y'D\hat{y}_{SEM})^2}{(y'Dy)(\hat{y}'_{SEM}D\hat{y}_{SEM})}} \tag{9.3.11}$$

Similarly, by employing $\hat{y}_{SLM}$ in expression (9.2.23), it follows that the corresponding *squared correlation* measure for SLM is given by,

$$\boxed{r^2(y,\hat{y}_{SLM}) = \frac{(y'D\hat{y}_{SLM})^2}{(y'Dy)(\hat{y}'_{SLM}D\hat{y}_{SLM})}} \tag{9.3.12}$$

These values are reported in Figure 7.7 as

$$\boxed{r^2(y,\hat{y}_{SEM}) = 0.5548} \tag{9.3.13}$$

and

$$\boxed{r^2(y,\hat{y}_{SLM}) = 0.7512} \tag{9.3.14}$$

Notice first that the squared correlation for SEM is *identical* with that of OLS. This appears somewhat surprising, given that their estimated beta coefficients are quite different. But in fact, this is an instance of the strong scale invariance properties of correlation. To see this, we again use the simplifying notation in (9.3.8),

$$r^2(y,\hat{y}) = \frac{(y'D\hat{y})^2}{(y'Dy)(\hat{y}'D\hat{y})} \tag{9.3.15}$$

and observe that for the case of only *one* explanatory variable, the $\hat{y}$ values for both SEM and OLS, must be linear combinations of 1_n and x, i.e., must be of the form,

$$\hat{y} = a1_n + bx \tag{9.3.16}$$

for some scalars *a* and *b*. But note first from the properties of the deviation matrix, *D*, that

$$D\hat{y} = aD1_n + bDx = bDx \tag{9.3.17}$$

and thus that $D\hat{y}$ is already independent of *a*. Moreover, (9.3.17) in turn implies both that

$$y'D\hat{y} = by'Dx \quad \text{and} \quad \hat{y}'D\hat{y} = (D\hat{y})'D\hat{y} = b^2x'Dx \tag{9.3.18}$$

Thus by (9.3.15) we must have

(9.3.18) $$r^2(y,\hat{y}) = \frac{(by'Dx)^2}{(y'Dy)(b^2x'Dx)} = \frac{b^2(y'Dx)^2}{b^2(y'Dy)(x'Dx)} = r^2(y,x)$$

and may conclude that squared correlation depends *only* on y and x. So in particular, the squared correlation of OLS and SEM must always be the same for the case of *one* explanatory variable.

However, this is clearly not true for SLM, where $X = [1_n, x]$ is transformed to

(9.3.19) $$X_\rho = B_\rho^{-1}X = [B_\rho^{-1}1_n, B_\rho^{-1}x]$$

so that $\hat{y}$ is no longer of the form (9.3.16). Thus there is little relation between the squared correlations for SLM and OLS, and as we have seen before, the squared correlation fit for SLM in (9.3.14) is *much* higher than for OLS (and SEM).

9.3.3 A Geometric View of Squared Correlation

To gain further insight into the role of squared correlation as a general measure of goodness-of-fit, it is instructive to start with the correlation coefficient itself. As we shall show below, if one writes vectors, $x, y \in \mathbb{R}^n$, in *deviation form* as $Dx = x - \bar{x}1_n$ and $Dy = y - \bar{y}1_n$, then from a geometric viewpoint, the correlation coefficient, $corr(x, y)$, in (9.3.1) turns out to be precisely the *cosine* of the angle, $\theta(Dx, Dy)$, between these vectors, i.e.,

(9.3.20) $$\boxed{r(x,y) = \cos[\theta(Dx,Dy)]}$$

This is most easily seen by first considering the cosine of the angle, $\theta = \theta(x, y)$, between any pair of (nonzero) vectors, $x, y \in \mathbb{R}^n$, as shown for $n = 2$ in Figure 9.9 below:

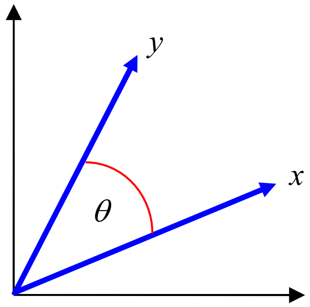

Figure 9.9. Vector Angle

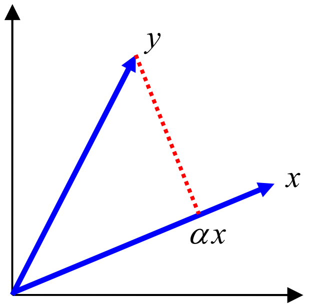

Figure 9.10. Right Triangle

To calculate the cosine of this angle, we first construct a *right triangle* by finding the point, αx, on the x-vector for which the line segment, $y-\alpha x$, is *orthogonal* to x, as shown by the red dotted line in Figure 9.10. Since vectors are orthogonal if and only if their inner product is zero, this point can be identified by solving:

$$(9.3.21) \qquad 0 = x'(y-\alpha x) = x'y - \alpha x'x \Rightarrow \boxed{\alpha = \frac{x'y}{x'x} = \frac{x'y}{\| x \|^2}}$$

Next, recall (from trigonometry) that for this right triangle, the desired cosine of $\theta(x,y)$ is given by the (signed) length of the *adjacent side*, i.e., $\alpha \| x \|$, over the length of the *hypotenuse*, $\| y \|$, so that

$$(9.3.22) \qquad \cos[\theta(x,y)] = \frac{\alpha \| x \|}{\| y \|} = \left(\frac{x'y}{\| x \|^2} \right) \frac{\| x \|}{\| y \|}$$

$$\Rightarrow \boxed{\cos[\theta(x,y)] = \frac{x'y}{\| x \| \cdot \| y \|}}$$

Before proceeding further, recall from expression (4.1.12) that this already establishes (9.3.20) for the case of "zero mean" vectors. But the more general case is now obtained by simply considering the vectors, *Dx* and *Dy*. In particular, since by definition,

$$(9.3.23) \qquad \| Dx \| = \sqrt{(Dx)'(Dx)} = \sqrt{x'DDx} = \sqrt{x'Dx}$$

and similarly, $\| Dy \| = \sqrt{y'Dy}$, it follows at once from (9.3.1) together with (9.3.22) and (9.3.23) that

$$(9.3.24) \qquad \cos[\theta(Dx,Dy)] = \frac{(Dx)'Dy}{\| Dx \| \cdot \| Dy \|} = \frac{x'Dy}{\sqrt{x'Dx}\sqrt{x'Dx}} = r(x,y)$$

and thus that (9.3.20) does indeed hold for all (nonzero) vectors, $x, y \in \mathbb{R}^n$. This in turn implies that the *squared correlation* is simply the square of this cosine:

$$(9.3.25) \qquad \boxed{r^2(x,y) = \cos^2[\theta(Dx,Dy)]}$$

So in our case, if we now let $\hat{y}$ denote the *predicted value* of data vector, y, for *any given model* (whatsoever), then it follows at once that

$$(9.3.26) \qquad \boxed{r^2(y,\hat{y}) = \cos^2[\theta(Dy,D\hat{y})]}$$

This geometric view of squared correlation helps to clarify the exact sense in which it constitutes a robust goodness-of-fit measure. In particular, it yields a measure of "similarity" between y and $\hat{y}$ which is completely independent of the measurement units employed. Indeed, this was already shown in arguments of (9.3.16) through (9.3.18) above, where shifts of measurement origins were seen to be removed by the deviation matrix, D, and where scale transformations were removed by the ratio form of squared correlation itself. Even more important is the fact that since $\cos^2(\theta)$ is close to one if and only if θ is close to 0 (or π), the identity in (9.3.26) shows that $r^2(y,\hat{y})$ is close to one if and only if the vectors, Dy and $D\hat{y}$, point in almost the same (or opposite) directions. Algebraically, this implies they are almost exact linear multiples of one another, i.e., that $D\hat{y} \approx \alpha Dy$ for some nonzero scalar, α. In practical terms, this means that the *relative sizes of all deviation components* must be approximately the same, so that if $\bar{\hat{y}}$ denotes the sample mean of $\hat{y}$, then

$$(9.3.27) \qquad \boxed{\frac{\hat{y}_i - \bar{\hat{y}}}{\hat{y}_j - \bar{\hat{y}}} \approx \frac{y_i - \bar{y}}{y_j - \bar{y}} \quad , \quad i \neq j}$$

Thus large (or small) deviations from the mean in components of y are reflected by comparable large (or small) deviations the mean in components of $\hat{y}$. The shows exactly the sense in which prediction, $\hat{y}$, is deemed to be similar to data, y, when $r^2(y,\hat{y}) \approx 1$.

Finally, a more detailed geometric investigation of squared correlation is presented in Section A3.6.3 of the Appendix. There it shown that one shortcoming of this goodness of fit measure is that, by the orthogonality property of OLS in (9.1.14), this model must almost *always* look better than GLS competitors in terms of squared correlation.[13]

9.4 Measures based on Maximum-Likelihood Values

Recall that our basic strategy for estimating model coefficients, (β, σ^2, ρ), was to find values $(\hat{\beta}, \hat{\sigma}^2, \hat{\rho})$ that maximized the likelihood of observed data, *y*, given explanatory data values, *X*. This suggests that a natural measure of fit should be provided by the maximum (log) likelihood value, $L(\hat{\beta}, \hat{\sigma}^2, \hat{\rho} \mid y, X)$, obtained. One difficulty here is that since likelihood values themselves are probability *density* values, and *not probabilities*, any direct interpretation of such values is tenuous at best. But the *ratios* of these values for different models might still provide meaningful comparisons in terms of the limiting probability-ratio arguments used in expressions (7.1.1) and (7.1.4) above.

[13] The special case of a *single* explanatory variable in (9.3.16) above is one of the few exceptions.

However, there is a second more serious difficulty with likelihood values that is reminiscent of R-squared values. Recall from the argument in expressions (9.1.31) through (9.1.34) that R-squared essentially always *increases* when new explanatory variables are added to the model. In fact, that argument really shows that the increase in R-squared results from the addition of new *beta parameters*. But this argument is far more general, and in fact shows that maximum values of functions are never decreased when more parameters are added. In particular, if we consider the case of two likelihood functions, say $L_{(k)}(\theta_1,..,\theta_k \mid y,X)$ and $L_{(k+1)}(\theta_1,..,\theta_k,\theta_{k+1} \mid y,X)$, where the first is simply a special case of the second with $\theta_{k+1}=0$, i.e., with

(9.4.1) $$L_{(k)}(\theta_1,..,\theta_k \mid y,X) \equiv L_{(k+1)}(\theta_1,..,\theta_k,0 \mid y,X)$$

then the same argument shows that

(9.4.2) $$\max_{(\theta_1,..,\theta_k)} L_{(k)}(\theta_1,..,\theta_k \mid y,X) = \max_{(\theta_1,..,\theta_k)} L_{(k+1)}(\theta_1,..,\theta_k,0 \mid y,X)$$

$$\leq \max_{(\theta_1,..,\theta_k,\theta_{k+1})} L_{(k+1)}(\theta_1,..,\theta_k,\theta_{k+1} \mid y,X)$$

with strictly inequality almost always holding. What this means for our purposes is that log likelihood functions suffer from exactly the same "inflation problem" as R-squared whenever new parameters are added. So if one attempts to compare the goodness of fit between models that are "nested" in the sense of (9.4.1), [i.e., where one is a special case of the other with certain parameters set to zero (or otherwise constrained in value)], then the larger model will always yield a better fit in terms of maximum-likelihood values.

This observation suggests that such likelihood comparisons must somehow be *penalized* in terms of the numbers of parameters in a manner analogous to adjusted R-squared. If we again let $L(\hat{\theta} \mid y)$ denote a general log likelihood function evaluated at its maximum value, then the simplest of these penalized versions is *Akaike's Information Criterion* (AIC):

(9.4.3) $$\boxed{AIC \;=\; -2L(\hat{\theta} \mid y) + 2K}$$

where K now denotes the dimension of $\hat{\theta}$, i.e., the number of parameters being estimated [and where factor "2" in AIC, as well as in the other measures to be developed, relates to the form of the log likelihood ratio statistic in expression (10.1.7) below.] For both SEM and SLM with parameters, $\hat{\theta}=(\hat{\beta}_0,\hat{\beta}_1,..,\hat{\beta}_k,\hat{\sigma}^2,\hat{\rho})$, this implies in particular that $K=(k+1)+2=k+3$. This measure is discussed in detail by Burnham and Anderson (2002), where AIC is both defined (p.61) and later derived (Section 7.2). In addition, these authors recommend a "corrected" version of AIC (p.66) for sample sizes that are small relative to the number of parameters ($n/K < 40$). This is usually designated as *corrected AIC* (AIC$_c$) and can be written in terms of (9.4.3) as

$$\boxed{AIC_c = AIC + \frac{2K(K+1)}{n-(K+1)}} \tag{9.4.4}$$

An alternative penalized version of maximum likelihood which directly incorporates sample size is the *Bayes* (or *Schwarz*) *Information Criterion* (BIC):

$$\boxed{BIC = -2L(\hat{\theta}\,|\,y) + K\log(n)} \tag{9.4.5}$$

While this measure is also developed in Burnham and Anderson (2002, Section 6.4.1), a more lucid derivation can be found in Raftery (1995, section 4.1). Given its heavier penalization term for model sizes, K [when $\log(n) > 2$], this measure is well known to favor smaller models (i.e., with fewer parameters) than AIC in terms of goodness of fit.

Finally it should be noted that when comparing SEM and SLM for a given specification of k explanatory variables, all such measures will differ only in terms of their corresponding maximum-likelihood values, $L(\hat{\theta}\,|\,y)$, for these two models. So in the present case of Eire, where Figure 7.7 shows that

$$\boxed{L_{SEM}(\hat{\theta}\,|\,y) = -49.8773} \tag{9.4.6}$$

$$\boxed{L_{SLM}(\hat{\theta}\,|\,y) = -45.6632} \tag{9.4.7}$$

it is clear that SLM must continue to yield a better fit than SEM with respect to *all* of these criteria.