What is the Error Term in a Regression Equation?

by

David A Freedman

It is often said that the error term in a regression equation represents the effect of the variables that were omitted from the equation. This is unsatisfactory, even in simple contexts, as the following discussion should indicate. Suppose subjects are IID, and all variables are jointly normal with expectation 0. Suppose the explanatory variables have variance 1. The explanatory variables may be correlated amongst themselves, but any p of them have a non-singular p-dimensional distribution. The parameters α_j are real. Let

$$Y_i = \sum_{j=1}^{\infty} \alpha_j X_{ij} \tag{1}$$

For each $p = 1, 2, \ldots,$ consider the regression model

$$Y_i = \sum_{j=1}^{p} \alpha_j X_{ij} + \epsilon_i(p) \tag{2}$$

where

$$\epsilon_i(p) = \sum_{j=p+1}^{\infty} \alpha_j X_{ij} \tag{3}$$

The α_j are identifiable. If the X_{ij} are independent for $j = 1, 2, \ldots,$ the standard assumptions hold, and $\epsilon_i(p)$ does indeed represent the effect on Y_i of the omitted variables $\{X_{ij} : j = p+1, \ldots\}$, at least in an algebraic sense. On the other hand, if the X_{ij} are dependent, the matter is problematic. If we take (1–3) as written, then $\epsilon_i(p)$ represents the effect on Y_i of the omitted variables—but $\epsilon_i(p)$ is correlated with the explanatory variables. The standard assumptions fail, and fitting (2) to data for $i = 1, \ldots, n$ will estimate the wrong parameters. If $\epsilon_i(p)$ is replaced by $\epsilon_i(p)^{\perp}$, namely, the part of $\epsilon_i(p)$ independent of $X_{i1}, \ldots, X_{ip}$, we have a bona fide regression model, but with different α's.

There is no easy way out of the difficulty. The conventional interpretation for error terms needs to be reconsidered. At a minimum, something like this would need to be said: the error term represents the combined effect of the omitted variables, assuming that

(i) the combined effect of the omitted variables is independent of each variable included in the equation,
(ii) the combined effect of the omitted variables is independent across subjects,
(iii) the combined effect of the omitted variables has expectation 0.

This is distinctly harder to swallow. Pratt and Schlaifer have a discussion in great depth.

Some technical details

If the α_j vanish for all but finitely many j, there are no technical issues. The inferential issue remains, provided the largest j with $\alpha_j \neq 0$ is an unknown parameter. Suppose next that $\alpha_j \neq 0$ for infinitely many j. Summability and identifiability must be demonstrated. To avoid interesting but unnecessary probabilistic complications, suppose $\sum_j |\alpha_j| < \infty$. Fix i. Suppose also that part of each $X_{ij} : j = 1, 2, \ldots$ is independent of all the other X_{ik}, and has L_2 norm at least $\eta > 0$. More specifically, let $X_{ij}^{\perp}$ be X_{ij} net of $\{X_{ik} : k = 1, \ldots, p$ with $k \neq j\}$. Thus, we assume $\|X_{ij}^{\perp}\| \geq \eta$, where $\|\bullet\|$ is the L_2 norm. See below for definitions and some theory.

Now $\|\epsilon_i(p)\| \leq \sum_{j=p+1}^{\infty} |\alpha_j|$ is small, so the sum on the right hand side of (1) converges in L_2. Fix j and p with $1 \leq j \leq p$. The regression of $\epsilon_i(p)$ on $\{X_{i1}, \ldots, X_{ip}\}$ has a small coefficient on X_{ij}, because

(i) $\epsilon_i(p)$ is small,
(ii) we get the coefficient by regressing $\epsilon_i(p)$ on $X_{ij}^{\perp}$, and
(iii) $\|X_{ij}^{\perp}\| \geq \eta$.

In more formal terms, by Lemma 2 below, a regression of Y_i on $X_{i1}, \ldots, X_{ip}$ in the random-variable domain gives a coefficient on X_{ij} of $\mathrm{cov}(X_{ij}^{\perp}, Y)/\mathrm{var}(X_{ij}^{\perp})$. This coefficient is α_j, with an error that is at most

$$\text{(4)} \qquad \frac{\mathrm{cov}\big(X_{ij}^{\perp}, \epsilon_i(p)\big)}{\mathrm{var}(X_{ij}^{\perp})} \leq \frac{\|X_{ij}^{\perp}\| \|\epsilon_i(p)\|}{\|X_{ij}^{\perp}\|^2} \leq \eta^{-1} \|\epsilon_i(p)\| \leq \eta^{-1} \sum_{j=p+1}^{\infty} |\alpha_j| \to 0$$

as $p \to \infty$. That proves identifiability.

A mistake to avoid

Some may conclude from the forgoing that bigger models are better. Perhaps, but (i) eventually we run out of data, and (ii) there is always the ugly possibility of inadvertently including an endogenous variable. Also see exercise 15 on page 105 of *Statistical Models* for information on standard errors in the presence of misspecification. Kitchen-sink models have their problems too.

Regression in the domain of random variables

Changing notation, let q be a positive integer. Let $U_1, \ldots, U_q, V$ be jointly normal random variables, each having expectation 0. Let $C_{ij} = \mathrm{cov}(U_i, U_j)$. This is a symmetric $q \times q$ matrix, assumed to be positive definite. Let $D_i = \mathrm{cov}(U_i, V)$. Take $D = (D_1, \ldots, D_q)'$ as a $q \times 1$ vector. Let $B = C^{-1}D$, which is also a $q \times 1$ vector. Let $V^{\perp} = V - (U_1, \ldots, U_q) \times B$, a scalar random variable.

Lemma 1. (i) $V^{\perp}$ is normal with expectation 0, and (ii) $V^{\perp} \perp (U_1, \ldots, U_q)$ in the sense that $\mathrm{cov}(U_j, V^{\perp}) = E(U_j V^{\perp}) = 0$ for each $j = 1, \ldots, q$. In particular, (iii) $V^{\perp}$ and $(U_1, \ldots, U_q)$ are independent.

For the proof, assertion (i) is immediate. For (ii), we need only check that

$$\mathrm{cov}(U_j, V) = \mathrm{cov}\big(U_j,\ (U_1, \ldots, U_q) \times B\big) = \sum_{k=1}^{q} \mathrm{cov}(U_j, U_k)\, B_k = \sum_{i=1}^{q} C_{jk} B_k,$$

i.e., $D = CB$. But $B = C^{-1}D$ by construction, completing the proof.

In short, $(U_1, \ldots, U_q) \times B$ is the regression of V on $U_1, \ldots, U_q$; the coefficient on U_i is B_i; and $V^\perp$ is the part of V independent of $U_1, \ldots, U_q$. This is also "V net of $U_1, \ldots, U_q$." Normality is relevant only to convert orthogonality into independence. Without normality, $(U_1, \ldots, U_q) \times B$ is the linear projection of V onto $U_1, \ldots, U_q$, i.e., the linear combination of $U_1, \ldots, U_q$ closest to V in L_2—because $V^\perp$ is orthogonal to $U_1, \ldots, U_q$. The simplest special case has $q = 1$. Then the regression coefficient takes a form that may be more familiar, $\mathrm{cov}(U_1, V)/\mathrm{var}(U_1)$.

Lemma 2. The regression of V on $U = (U_1, \ldots, U_q)$ can be computed by the following stepwise procedure, with $\tilde{U} = (U_2, \ldots, U_q)$.

(i) Regress V on $U_2, \ldots, U_q$. Let α be the $(q-1) \times 1$ vector of regression coefficients. Let $\hat{V} = \tilde{U}\alpha$ and $V^\perp = V - \hat{V}$.

(ii) Regress U_1 on $U_2, \ldots, U_q$. Let β be the $(q-1) \times 1$ vector of regression coefficients. Let $\hat{U}_1 = \tilde{U}\beta$ and $U_1^\perp = U_1 - \hat{U}_1$.

(iii) Regress V on $U_1^\perp$. Let γ be the regression coefficient, a scalar.

The $q \times 1$ vector of regression coefficients of V on $U_1, \ldots, U_q$ is then

$$\begin{pmatrix} \gamma \\ \alpha - \beta\gamma \end{pmatrix}$$

Proof. Since $V = \hat{V} + V^\perp$ and $\hat{V} \perp U_1^\perp$, whether we regress V on $U_1^\perp$ or $V^\perp$ on $U_1^\perp$, the coefficient on $U_1^\perp$ will be the same, viz., γ. So $\epsilon = V^\perp - U_1^\perp\gamma \perp U_1^\perp$. Plainly, $\epsilon \perp U_2, \ldots, U_q$, because ϵ is a linear combination of $V^\perp$ and $U_1^\perp$. Thus,

$$\begin{aligned} V &= \hat{V} + V^\perp \\ &= \hat{V} + U_1^\perp\gamma + \epsilon \\ &= \tilde{U}\alpha + (U_1 - \tilde{U}\beta)\gamma + \epsilon \\ &= U_1\gamma + \tilde{U}(\alpha - \beta\gamma) \\ &= \begin{pmatrix} \gamma \\ \alpha - \beta\gamma \end{pmatrix} U + \epsilon \end{aligned} \tag{5}$$

with $\epsilon \perp U$, as required. To clarify the notation, U is $1 \times q$ and $\tilde{U}$ is $1 \times (q-1)$; both are random vectors; $\hat{V}, V^\perp, \hat{U}_1, U_1^\perp, \epsilon$ are all scalar random variables. If $U_1, \ldots, U_q, V$ are taken as jointly normal, these derived quantities are jointly normal too. The quantities α, β, γ are parameters not estimates, being computed from the joint distribution not from data. Exercise 17 on page 34 of *Statistical Models* covers regression in the data domain using a method exactly like that in Lemma 2, although the notation is little different.

References

Freedman DA (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Pratt J, Schlaifer R (1984). On the Nature and Discovery of Structure. *Journal of the American Statistical Association* 79: 9–21.

Pratt J, Schlaifer R (1988). On the Interpretation and Observation of Laws. *Journal of Econometrics* 39: 23–52.

Notes for Statistics 215
Department of Statistics
UC Berkeley, CA 94720-3860
November, 2005