TR-1735 November 1986

Hypothesis Testing with Errors in the Variables

Nancy David*
G. W. Stewart†

ABSTRACT

In this paper we give reason to hope that errors in regression variables are not as harmful as one might expect. Specifically, we will show that although the errors can change the values of the quantities one computes in a regression analysis, under certain conditions they leave the distributions of the quantities approximately unchanged.

*Research Institute of Michigan, 1501 Wilson Blvd., Arlington VA 22209
†Department of Computer Science and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742. This work was supported in part by the Office of Naval Research under contract No. N00014-76-C-3091

Hypothesis Testing with Errors in the Variables

Nancy David*
G. W. Stewart†

ABSTRACT

In this paper we give reason to hope that errors in regression variables are not as harmful as one might expect. Specifically, we will show that although the errors can change the values of the quantities one computes in a regression analysis, under certain conditions they leave the distributions of the quantities approximately unchanged.

1 Introduction

Of all the practical problems associated with linear regression analysis, none is more vexing than errors in the variables. These errors are found everywhere and usually cannot be eliminated. Moreover, in many cases it is obvious that they are large enough to have important effects on quantities, such as F-statistics, calculated in the course of the analysis. Since there are no procedures for dealing with errors in the variables that do not require rather precise information about the errors themselves, the analyst usually has no choice but to ignore the errors and analyze the data as though they were not there. What is surprising is that this does not seem to result in obvious catastrophes. The purpose of this paper is to provide a partial explanation.

The nature of the explanation is that under appropriate circumstances the errors enter cooperatively into some of the procedures of regression analysis. To see what this means, we must first define our model. We will suppose that we are given the usual model

$$y = Xb + e, \tag{1.1}$$

where X is an $n \times p$ matrix of rank p and the components of e are iid $\mathrm{N}(0, \sigma^2)$. We will further assume that in place of the matrix X we observe

$$\tilde{X} = X + E,$$

where the rows of E are iid $\mathrm{N}(0, \Sigma)$ and are independent of e.

*Research Institute of Michigan, 1501 Wilson Blvd., Arlington VA 22209

†Department of Computer Science and Institute for Physical Science and Technology, University of Maryland, College Park. This work was supported in part by the Office of Naval Research under contract No. N00014-76-C-3091

Let us now consider the vector b of regression coefficients. This would ordinarily be estimated by

$$\hat{b} = X^{\dagger}y = b + X^{\dagger}e \tag{1.2}$$

where $X^{\dagger} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is the pseudo-inverse of X. Unfortunately, we are forced to compute

$$\tilde{b} = \tilde{X}^{\dagger}\tilde{y}, \tag{1.3}$$

which can differ considerably from $\hat{b}$. However, if we rewrite the model (1.1) in the form

$$y = \tilde{X}b + e - Eb,$$

then

$$\tilde{b} = b + \tilde{X}^{\dagger}(e - Eb). \tag{1.4}$$

Now it is known from perturbation theory for pseudo-inverses (Stewart, 1977) that

$$\tilde{X}^{\dagger} = X^{\dagger} + F^{\mathrm{T}},$$

where

$$\lim_{E \to 0} F = 0.$$

Consequently we have from (1.4)

$$\tilde{b} = b + X^{\dagger}(e - Eb) + F^{\mathrm{T}}(e - Eb). \tag{1.5}$$

Comparing (1.5) with (1.2) we see that if E is small, so that F is also small, the $\tilde{b}$ we compute behaves as if were the correct estimate for the model

$$\dot{y} = Xb + (e - Eb). \tag{1.6}$$

In other words, up to terms that vanish with E, the vector $\tilde{b}$ comes from a model in which the errors are iid $\mathrm{N}(0, \sigma^2 + b^{\mathrm{T}}\Sigma b)$. If these vanishing terms are small enough, the only untoward effect of the errors is that they inflate the variance of the response vector.

It is important to keep in mind that this is not simply a continuity result. Nothing is said about Eb in (1.6) being small compared to e. Indeed σ could be zero so that *all* the variability in the problem comes from E. The point is that this variability enters in a benign way.

Our approach is related to the approximation of functions of random variables by one or two terms in a Taylor series—something that is by no means new. For example, Gauss (1821) used it to linearize nonlinear least squares problems, and Brown, Kadane, and Ramage (1974) have used it in the analysis of certain econometric models. In regression analysis Hodges and Moore (1972) and Davies and Hutton (1975) have examined first and second order terms in the expansion of $\tilde{b}$. Our approach differs from this in that we actually ignore terms of the first order in E [cf. (1.5) and (1.6)]. Nontheless, a number of important quantities computed using $\tilde{X}$ approach the same quantities derived from (1.6).

In the next section, we shall show that $\tilde{b}$ converges in an appropriate sense to $b + X^{\dagger}(e + Eb)$ and establish a similar result for the residual vector. We will also show that F-statistics computed using $\tilde{X}$ converge to a true F-statistic. The paper concludes with some general observations.

2 Coefficients, residuals, and F-tests

In this section we are going to show that as $\Sigma \to 0$ a number of quantities—regression coefficients, residuals, F-statistics—computed using $\tilde{X}$ converge with probability one to the corresponding quantities from the model (1.6). As we indicated in the introduction, the results hold when $\sigma = 0$, which is in fact the most important case. Before we proceed, however, we must make sure that the result itself is formulated in such a way that it will be useful.

Let $\tilde{b}$ be defined by (1.3), and let $\dot{b} = X^{\dagger}\dot{y}$ be the corresponding vector from the model (1.6).[1] If $\sigma = 0$, it is trivial to show that as $\Sigma \to 0$ the vector $\tilde{b}$ converges to $\dot{b}$ with probability one, since they both converge to $\hat{b}$. Even when $\sigma = 0$, the result is trivial, since the two distributions are collapsing around the vector b. However, if we normalize our quantities by dividing by

$$\dot{\sigma} = \sqrt{\sigma^2 + b^{\mathrm{T}}\Sigma b}, \tag{2.1}$$

then their distributions do not collapse. Convergence will then imply that for the purpose of estimating variability the distributions of $\tilde{b}/\dot{\sigma}$ and $\dot{b}/\dot{\sigma}$ are equivalent for small Σ. This is especially gratifying, since the latter is not computable while the former does not in general have a first moment.

Turning now to the main result of this section, we must set up the underlying probability space. Let $\mathcal{E}$ denote the space of matrices $(e_0\ E_0)$, where $e_0 \in \mathcal{R}^n$ and $E_0 \in \mathcal{R}^{n\times p}$, with the elements of $(e_0\ E_0)$ iid $\mathrm{N}(0,1)$. For fixed $\sigma \geq 0$ and Σ positive semi-definite, the matrix $(\sigma e_0\ E_0\Sigma^{\frac{1}{2}})$ defines a measurable function on $\mathcal{E}$, representing our error e and perturbation E.

One final technical point. We propose to normalize our quantities by $\dot{\sigma}$ defined by (2.1). Since we have assumed only that Σ is positive semi-definite, it is possible for $\dot{\sigma}$ to be zero, even though Σ is nonzero. However, in this case $e - Eb = 0$, so that $y = \tilde{X}b$,

[1] The convention introduced here will be followed throughout the paper. A quantity with a tilde above it will refer to the quantity computed using the $\tilde{X}$. A quantity with a dot above it will refer to the model (1.6).

and it as if the model had no error at all. Consequently, we may assume that $\dot{\sigma} > 0$ as $\Sigma \to 0$.

The results on regression coefficients and residual vectors are contained in the following theorem. We will treat the F-tests separately.

Theorem 2.1 *With the definitions given in the introduction, let $\tilde{b} = \tilde{X}^{\dagger} y$ and $\dot{b} = X^{\dagger}\dot{y}$. Moreover, with $P = I - XX^{\dagger}$ and $\tilde{P} = I - \tilde{X}\tilde{X}^{\dagger}$, let $\tilde{r} = \tilde{P}y$ and $\dot{r} = P\dot{y}$ be the computed residual vector and the residual vector from the model (1.6). Then as $\Sigma \to 0$*

$$\begin{aligned} &1. \quad \frac{\tilde{b} - \dot{b}}{\dot{\sigma}} \xrightarrow{\text{wp1}} 0, \\ &2. \quad \frac{\tilde{r} - \dot{r}}{\dot{\sigma}} \xrightarrow{\text{wp1}} 0, \\ &3. \quad \frac{\tilde{r}^{\mathrm{T}}\tilde{r} - \dot{r}^{\mathrm{T}}\dot{r}}{\dot{\sigma}^2} \xrightarrow{\text{wp1}} 0. \end{aligned} \tag{2.2}$$

Proof. Since convergence with probability one is convergence almost everywhere in the underlying probability space, let $(e_0\ E_0)$ denote a fixed member of $\mathcal{E}$, and let $e = \sigma e_0$ and $E = E_0\Sigma^{\frac{1}{2}}$. As we mentioned in the introduction, $\tilde{X}^{\dagger} = X^{\dagger} + F$, where $F \to 0$ as $\Sigma \to 0$. Now from (1.4)

$$\tilde{b} = b + X^{\dagger}(e - Eb) + F^{\mathrm{T}}(e - Eb) = \dot{b} + F^{\mathrm{T}}(\sigma e_0 - E_0\Sigma^{\frac{1}{2}} b).$$

Hence

$$\frac{\tilde{b} - \dot{b}}{\dot{\sigma}} = F^{\mathrm{T}} \frac{\sigma e_0 - E_0\Sigma^{\frac{1}{2}} b}{\dot{\sigma}}.$$

When $\sigma \neq 0$, this expression clearly converges to zero, since $F^{\mathrm{T}} \to 0$ and $(\sigma e_0 - E_0\Sigma^{\frac{1}{2}} b)/\dot{\sigma} \to e_0$. When $\sigma = 0$,

$$\frac{\tilde{b} - \dot{b}}{\dot{\sigma}} = F^{\mathrm{T}} \frac{E_0\Sigma^{\frac{1}{2}} b}{\|\Sigma^{\frac{1}{2}} b\|},$$

where $\|\cdot\|$ denotes the usual Euclidean vector norm. The result then follows from the fact that $E_0\Sigma^{\frac{1}{2}} b / \|\Sigma^{\frac{1}{2}} b\|$ remains bounded as $\Sigma \to 0$.

To establish (2.2.2), observe that we can write $\tilde{P} = P + Q$, where $Q \to 0$ as $E \to 0$ (Stewart, 1977). Since

$$\tilde{r} = \dot{r} + Q(\sigma e_0 + E_0\Sigma^{\frac{1}{2}} b),$$

the result follows as above.

Finally to establish (2.2.3), write

$$\begin{aligned} \tilde{r}^{\mathrm{T}}\tilde{r} = \dot{r}^{\mathrm{T}}\dot{r} + 2\dot{r}^{\mathrm{T}} Q(\sigma e_0 + E_0\Sigma^{\frac{1}{2}} b) + \\ (\sigma e_0 + E_0\Sigma^{\frac{1}{2}} b)^{\mathrm{T}} Q(\sigma e_0 + E_0\Sigma^{\frac{1}{2}} b), \end{aligned} \tag{2.3}$$

When $\sigma \neq 0$ the ratio of the difference $\tilde{r}^{\mathrm{T}}\tilde{r} - \dot{r}^{\mathrm{T}}\dot{r}$ to $\dot{\sigma}^2$ obviously converges to zero. For the case $\sigma = 0$, consider the inequality

$$\frac{|2\dot{r}^{\mathrm{T}}QE_0\Sigma^{\frac{1}{2}}b|}{\dot{\sigma}^2} \leq \frac{\|\dot{r}\|\|Q\|\|E_0\Sigma^{\frac{1}{2}}b\|}{\|E_0\Sigma^{\frac{1}{2}}b\|^2} = \frac{\|\dot{r}\|}{\|E_0\Sigma^{\frac{1}{2}}b\|}\|Q\|$$

Since $\dot{r} = PE_0\Sigma^{\frac{1}{2}}b$, we have

$$\frac{|2\dot{r}^{\mathrm{T}}QE_0\Sigma^{\frac{1}{2}}b|}{\dot{\sigma}^2} \leq \|Q\| \to 0.$$

Thus the ratio of the second term in the right hand side of (2.3) to $\dot{\sigma}^2$ has zero for its limit. The third term is treated similarly. □

An immediate consequence of the theorem is that $\tilde{r}^{\mathrm{T}}\tilde{r}/(n-p)$ is asymptotically (small Σ) a good estimate of the inflated variance $\sigma^2 + b^{\mathrm{T}}\Sigma b$. Moreover it is asymptotically independent of $\tilde{b}$. This means that it can be used to construct confidence intervals and t-tests in the usual way.

Let us now turn to the problem of F-tests of hypotheses. Partition $X = (X_1\ X_2)$, where X_2 has k columns. Partition b conformally in the form

$$b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Then the hypothesis we will test is

$$\mathrm{H}: \quad b_2 = c.$$

Note that the most general linear hypothesis can be brought into this form by a linear (in fact orthogonal) transformation of b.

The classical F-test is computed from the residual vector of the least squares estimate when b_2 is constrained to be equal to c. In the unperturbed model, this process amounts to forming the projection $P_{\mathrm{H}} = I - X_1X_1^{\dagger}$ and computing the residual $r_{\mathrm{H}} = P_{\mathrm{H}}(y - X_2c)$. The usual F-statistic for testing Ω is

$$F = \frac{n-p}{k}\,\frac{r_{\mathrm{H}}^{\mathrm{T}}r_{\mathrm{H}} - r^{\mathrm{T}}r}{r^{\mathrm{T}}r}.$$

Following our convention, let $\dot{r}_{\mathrm{H}}$ and $\dot{F}$ be the corresponding quantities for the model (1.6) and let $\tilde{P}_{\mathrm{H}}$, $\tilde{r}_{\mathrm{H}}$ and $\tilde{F}$ be the quantities actually computed. Since

$$\dot{r} = P_{\mathrm{H}}X_2(b_2 - c) + P_{\mathrm{H}}(e - Eb),$$

we see that $\dot{F}$ has an F-distribution, which is central if and only if $b_2 = c$. Moreover we have the following theorem.

Theorem 2.2 *If $n \geq 2p$ and $b_2 = c$, then as $\Sigma \to 0$*

$$\tilde{F} - \dot{F} \xrightarrow{\text{wp1}} 0.$$

Proof. Let $\tilde{P}_{\mathrm{H}} = P_{\mathrm{H}} + Q_{\mathrm{H}}$, where $Q_{\mathrm{H}} \to 0$ as $E \to 0$. We have

$$\tilde{r}_{\mathrm{H}} = \tilde{P}_{\mathrm{H}}\tilde{X}_2(b_2 - c) + \tilde{P}_{\mathrm{H}}(e - Eb), \tag{2.4}$$

and since the hypothesis is true,

$$\tilde{r}_{\mathrm{H}} = P_{\mathrm{H}}(e - Eb) + Q_{\mathrm{H}}(e - Eb).$$

Thus as in Theorem 2.1

$$\frac{\tilde{r}_{\mathrm{H}}^{\mathrm{T}}\tilde{r}_{\mathrm{H}}}{\dot{\sigma}^2} - \frac{\dot{r}_{\mathrm{H}}^{\mathrm{T}}\dot{r}_{\mathrm{H}}}{\dot{\sigma}^2} \xrightarrow{\text{wp1}} 0. \tag{2.5}$$

Now

$$\frac{k}{n-p}(\tilde{F} - \dot{F}) = \frac{\dot{\sigma}^{-4}(\tilde{r}_{\mathrm{H}}^{\mathrm{T}}\tilde{r}_{\mathrm{H}}\dot{r}^{\mathrm{T}}\dot{r} - \dot{r}_{\mathrm{H}}^{\mathrm{T}}\dot{r}_{\mathrm{H}}\tilde{r}^{\mathrm{T}}\tilde{r})}{\dot{\sigma}^{-4}(\tilde{r}^{\mathrm{T}}\tilde{r}\dot{r}^{\mathrm{T}}\dot{r})} \tag{2.6}$$

By Theorem 2.1 and (2.5), the numerator of the right hand side of (2.6) converges to zero a.e. Thus if we can show that the denominator remains bounded below our result will be established.

As unusual the difficult case is when $\sigma = 0$, which we will now treat. For fixed E_0 in $\mathcal{E}$, we have

$$\frac{\dot{r}^{\mathrm{T}}\dot{r}}{\dot{\sigma}^2} = \frac{\|PE_0\Sigma^{\frac{1}{2}}b\|^2}{\|\Sigma^{\frac{1}{2}}b\|^2}. \tag{2.7}$$

We claim that with probability one

$$\inf(PE_0) = \inf_{\|b\|=1}\|PE_0\| > 0.$$

To see this let $V = (V_1\ V_2)$ be an orthogonal matrix such that the columns of V_2 span the column space of P. Then

$$V^{\mathrm{T}}P = \begin{pmatrix} 0 \\ V_2^{\mathrm{T}} \end{pmatrix},$$

and

$$V^{\mathrm{T}}PE_0 = \begin{pmatrix} 0 \\ V_2^{\mathrm{T}}E_0 \end{pmatrix}.$$

But since the columns of V_2 are orthonormal, the elements of the matrix $V_2^{\mathrm{T}}E_0$ are iid $\mathrm{N}(0,1)$; and since $n \geq 2p$, it has more rows that columns. Thus $V_2^{\mathrm{T}}E_0$ and hence PE_0 has full column rank except on a set of measure zero.

Since $\|PE_0\Sigma^{\frac{1}{2}}b\| \geq \inf(PE_0)\|\Sigma^{\frac{1}{2}}b\|$, it follows from (2.7) that

$$\frac{\dot{r}^{\mathrm{T}}\dot{r}}{\dot{\sigma}^2} \geq \inf(PE_0),$$

which is positive almost everywhere. Since

$$\frac{\tilde{r}^{\mathrm{T}}\tilde{r}}{\dot{\sigma}^2} \to \frac{\dot{r}^{\mathrm{T}}\dot{r}}{\dot{\sigma}^2} \quad \text{a.e.},$$

the quantity $\tilde{r}^{\mathrm{T}}\tilde{r}/\dot{\sigma}^2$ is also uniformly bounded below almost everywhere as $\Sigma \to 0$. This completes the proof of the theorem. □

When $b_2 \neq c$, the $\tilde{F}$ may still converge to $\dot{F}$. However, this case is essentially different from our previous results, since the term $\tilde{P}_{\mathrm{H}}\tilde{X}_2(b_2 - c)$ in (2.4) contributes errors of the same order of magnitude as $P_{\mathrm{H}}(e - Eb)$, errors which do not have a nice distributions. Thus, unless Σ is small compared with σ, the distribution of $\tilde{F}$ will not approximate a noncentral F-distribution. This means that we cannot use the usual procedures for relating power to sample size when there are errors in the variables. Nonetheless, equation (2.4) suggests that the errors cause a loss of power, not so much by diminishing the source $P_{\mathrm{H}}X_2(b_2 - c)$ of noncentrality as by inflating the variance with which it must be compared.

3 Concluding remarks

We have shown that if the variance of the errors in the variables is small enough, then some of the common procedures in regression analysis are unaffected by the errors. As we pointed out in the introduction, this is not a continuity result; for the errors can have a palpable effect on the numbers one calculates. But the numbers will nontheless have approximately the right distributions. What is particularly nice about these results is that they require no detailed knowledge of Σ; only that it is sufficiently small.

Now it is possible to question the value of small Σ theorems. As it has been put to one us, "You provide no statistical justification for the assumption that Σ tends to zero." In response one might note that a similar objection can be raised to large sample theorems: there is no statistical justification for assuming that $n \to \infty$. The answer to both objections is that people, not statistics, make n large or Σ small. If it turns out that the sample is too small or the errors too large for the problem at hand, no amount of mathematical analysis will alter the situation.

However, this line of argument misses the real point. The value of asymptotic theorems of all kinds is that they give hope that an otherwise intractable problem may be treated in a simple manner. In any particular problem, one must decide whether the asymptotic behavior actually obtains; but this does not detract from the value of the asymptotic result itself.

There remains the question of how small the errors must be before the results of a regression analysis can be trusted. To a large extent this is an open question. The authors (1982) attempted to obtain rigorous bounds on the terms ignored; however, the results were so conservative as to be practically useless.

Nonetheless, we can give realistic advice about when not to trust the results. Let

$$\tau = \inf{}^{-1}(\Sigma X^{\dagger}).$$

One of the authors (Stewart, 1986) has shown informally that there is a linear combination of the regression coefficients that will be biased downward by a factor of approximately $(n-p)\tau^2$. Thus unless, say,

$$(n-p)\tau^2 < 0.1 \tag{3.1}$$

the bias can be over ten percent, and $\tilde{b}$ clearly cannot be near $\dot{b}$, which is unbiased. However, at this time we cannot say that a bound like (3.1), perhaps with a smaller constant, can tell us when to trust our asymptotics.

References

[1] Brown, G. F., J. B. Kadane, and J. G. Ramage (1974), "The Asymptotic Bias and Mean-Squared Error of Double K-Class Estimators when the Disturbances are Small," *International Economic Review* **15,** 667-679.

[2] Davies, R. B. and Hutton, B. (1975), "The Effects of Errors in the Independent Variables in Linear Regression," *Biometrika* **62,** 383-391.

[3] David, N. A. and Stewart, G. W. (1982), "Significance Testing in a Functional Model," University of Maryland Computer Science Technical Report 1204.

[4] Gauss, C. F. (1821), "Theoria Combinationis Observationum Erroribus Minims Obnoxiae: Pars Prior," in *Werke* v. 4, Köiglichen Gesellschaft Der Wissenschaften zu Göttingen, 1880.

[5] Hodges, S. D. and Moore, P. G. (1972), "Data Uncertainties and Least Squares Regression," *Appl. Stat.* **21,** 185-195.

[6] Stewart, G. W. (1977), "On the Perturbation of Pseudo-Inverses, Projections, and Linear Least Squares Problems," *SIAM Review* **19,** 634-666.