COWLES COMMISSION DISCUSSION PAPER: STATISTICS: NO. 339

THE ACCURACY OF NAIVE MODELS

Harry Markowitz

In "A Revised Klein Econometric Model...", Carl Christ compared the predictions, for 13 variables, obtained by three methods of prediction. One method predicted from the reduced form of Christ's econometric model. The coefficients of the reduced form were estimated by least squares (L.S.). The other two methods of prediction were:

$$1) \quad (y_p)_t = y_{t-1}$$

$$2) \quad (y_p)_t = y_{t-1} + (y_{t-1} - y_{t-2}) = 2y_{t-1} - y_{t-2}$$

where $(y_p)_t$ is the prediction of y for time t. We will call 1) and 2) naive models I and II respectively (N.M.I and N.M.II).

In 7 out of the 13 cases the error $\left| y - y_p \right|$ made by N.M.I was less than that made by the least-squares, reduced form predictions. Similarly for N.M.II. (See ibid p. 47.) The accuracy of prediction obtainable by naive models, the information obtainable by a comparison of L.S. and N.M. performances, deserves consideration.

In this paper "accuracy of prediction" will mean $E(y - y_p)^2$. We will see that if the changes in Ey are gradual, N.M.I works well. N.M.II works well when σy is small and Ey does not change direction frequently. Even if the assumptions of the least squares analysis are correct a naive model (especially N.M.I) may be more accurate. But this is not typical. If some L.S. assumptions are incorrect, N.M.I and II have an added advantage. A comparison of the "goodness of fit" obtained by L.S. and N.M.I or II can serve as a test of the L.S. assumption.

Let us assume that y is <u>in fact</u> a linear combination of variables V_j and a random element u: $y = \alpha_o + \Sigma \alpha_j V_j + u$.

The statistician <u>believes</u> that E y is a linear combination of variables z_i. The accuracy of prediction of the L. S. method will depend upon the relations between the true V's and the statisticians z's. We will consider two cases (i) where the set $\{z_i\}$ is the same as the set $\{V_j\}$; (ii) where the set $\{z_i\}$ includes some but not necessarily all the V's (including "V's" whose $\alpha = 0$). Two interesting cases which should be treated in the future are (iii) the question of errors of measurement where $z_i = V_i + u_i$, u_i a random variable; and (iv) the question of aggregation where $z_i = \varphi(V_1, V_2 \ldots)$.

The assumption that E y is a linear function of V_j is not a restriction. There may be an infinite number of such V; V_k may equal $(V_1^{e'} . V_2^{e''} . V_3^{e'''} \ldots)$. If E y = $\varphi(V)$ then we write $\varphi(V)$ as an infinite polynomial. We will assume that a) E u = 0; b) the set $\{u_t\}$ are independent; c) $\sigma_{u_t} = \sigma_{u_{t+1}}$ for all t. a) and b) are not limitations of the analysis but are part of the definition of the V. c) is a limitation which should be dropped in some future study.

By accuracy of prediction we will mean $E(y - y_p)^2$. The accuracy of prediction of N. M. I at time t is

$$E(y_t - y_{t-1})^2 = E[(y_t - E\, y_t) - (y_{t-1} - E\, y_{t-1}) + (E\, y_t - E\, y_{t-1})]^2$$

Let us write E y = Y i.e., $y = Y + u$ then since u_t, u_{t-1} are independent

$$E(y_t - y_{t-1})^2 = 2\,\sigma^2 + (\Delta Y)^2$$

The accuracy of prediction of N. M. II is

$$E(y_t - 2\,y_{t-1} + y_{t-2})^2 = E[\,(y_t - Y_t) - 2(y_{t-1} - Y_{t-1}) + (y_{t-2} - Y_{t-2})$$
$$+ (Y_t - 2Y_{t-1} + Y_{t-2})]^2 = 6\,\sigma^2 + (\Delta\Delta Y)^2$$

Let $Y_p = \beta_o' + \sum_{i=1}^{k} \beta_i' (z_i - \bar{z}_i)$, where the β's are least squares estimates based on a sample of size T; suppose the set of z's is the same as the set of V's defined above; let

$$Z_i = (z_i - \bar{z}_i),$$

$$\ell_{ij} = \frac{1}{T} \sum_{t=1}^{T} (Z_i Z_j)_t$$

$$|L| = \begin{vmatrix} \ell_{11} & \cdots & \ell_{1k} \\ \vdots & & \\ \ell_{k1} & & \ell_{kk} \end{vmatrix}$$

$|L^{ij}|$ = the cofactor of ℓ_{ij}

Then the variance $V \beta_o' = V \bar{y} = \frac{\sigma^2}{T}$

$$V \beta_i' = \frac{|L^{ii}|}{|L|} \frac{\sigma^2}{T}$$

and the covariance $C \beta_i' \beta_j' = \frac{|L^{ij}|}{|L|} \frac{\sigma^2}{T}$

$$C \beta_i' \beta_o' = 0$$

$$V y_p = \frac{\sigma^2}{T} \left(1 + \frac{\sum_{i,j} |L^{ij}| (Z_i Z_j)}{|L|} T + 1\right);$$

$$E(y - y_p)^2 = \sigma^2 + V y_p$$

Now let us consider the case where the z's are some but not all the V's.

$$E(y - y_p)^2_{T+1} = E[(y - Y) - (y_p - E y_p) + (Y - E y_p)]^2$$

$$= \sigma^2 + V y_p + (Y - E y_p)^2$$

for $(y - Y)$, $(y_p - E y_p)$, $(Y - E y_p)$ are independent. The first is u_{T+1}; the second depends upon $u_1, \ldots, u_T$; the third does not depend on u.

$V(y_p)$ is the same as above; we must evaluate $(Y - E y_p)$

$E \beta_i' = \beta_i$ are the coefficients of the linear fit which minimize

$$\sum_{t=1}^{T} [Y - \beta_o - \sum_i \beta_i Z_i]^2$$

where again $Y = E\,y$ and $Z_i = (z_i - \bar{z}_i)$

$$Y = \alpha^o + \sum_j \alpha V_j = \alpha^o + \sum_i \alpha_i Z_i + \sum_k \alpha_k V_k^*$$

where the V_k^*'s are the V's which are not Z's.

Let $Y' = \alpha_o + \Sigma \alpha_i Z_i$

$Y'' = \Sigma \alpha_k V_k^*$

$Y = Y' + Y''$

Let $\mathcal{L}_{oj} = \frac{1}{T} \sum_{t=1}^{T} (Y - \bar{Y}) Z_j$

$\bar{Y} = \frac{1}{T} \Sigma Y_t$

$\mathcal{L}'_{oj} = \frac{1}{T} \sum_t (Y' - \bar{Y}') Z_j$

$\mathcal{L}''_{oj} = \frac{1}{T} \Sigma (Y'' - \bar{Y}'') (Z_j)$

$\beta_o = \bar{Y} = \bar{Y}' + \bar{Y}''$

$$\beta_i = \frac{\sum_j \mathcal{L}_{oj} \left| L^{ij} \right|}{|L|}$$

$$= \frac{\sum_j (\mathcal{L}'_{oj} + \mathcal{L}''_{oj}) \left| L^{ij} \right|}{|L|}$$

$$= \frac{\Sigma \mathcal{L}'_{oj} \left| L^{ij} \right|}{|L|} + \frac{\Sigma \mathcal{L}_{oj} \left| L^{ij} \right|}{|L|}$$

$$= \alpha_i + \gamma_i$$

$\alpha_i = \frac{\Sigma \mathcal{L}'_{oj} \left| L^{ij} \right|}{|L|}$ is the coefficient which would be obtained by fitting a linear combination of the Z's to Y'. Since $Y' = \alpha_o + \Sigma \alpha_i Z_i$, this gives the correct α_i. γ_i is the coefficient of Z_i we would got from a least squares regression of Y" on the Z's. This is a spurious coefficient. It is the basis in β'_i. The expected regression hyper-plane is the sum of the regression plane of Y' on Z and the spurious plane of Y" on Z.

$$E(y_p)_t = (Y_p)_t = \bar{Y}' + \bar{\bar{Y}}' + \sum_i \alpha_i Z_i + \sum_i \gamma_i Z_i$$

$$(Y - Y_p)^2_{T+1} = (Y'_{T+1} + Y''_{T+1} - \bar{Y}' - \bar{Y}'' - \Sigma \alpha Z - \Sigma \gamma Z)^2$$

$$= (Y''_{T+1} - \bar{Y}'' - \Sigma \alpha_i Z_i)^2$$

This is the squared deviation of Y''_{T+1} from the previously best fiting regression plane.

We may summarize our results as follows:

Source of Variation / Method	$(y - Y)^2$	$(y_p - Y_p)^2$	$(Y - Y_p)^2$
N. N. M. I.	σ^2	σ^2	$(\triangle Y)^2$
N. M. II	σ^2	$5\sigma^2$	$(\triangle \triangle Y)^2$
L. S.	σ^2	$\sigma^2\left(\frac{1}{T} + \frac{\Sigma \lvert L^{ij} \rvert Z_i Z_j}{T \lvert L \rvert}\right)$	$(Y'' - \bar{Y}'' - \Sigma \gamma_i Z_i)^2$

N. M. I will work well where the change ($\triangle Y$) is small compared to σ. If $\triangle Y < 2\sigma$, N. M. I must be more accurate than N. M. II. If the L. S. assumptions are correct, the L. S. method is more accurate than N. M. I if and only if

$$\left(\frac{I}{T} + \frac{\sum_{ij} \lvert L^{ij} \rvert (z_i - \bar{z}_i)(z_j - z_j)}{T \quad \lvert L \rvert}\right) < \left(1 + \left(\frac{\triangle Y}{\sigma}\right)^2\right)$$

The opposite of this is not impossible; I would guess that the above is more common. If the L. S. assumptions are not completely correct, the naive models

have an added advantage. We have analyzed the case where the set $\{Z\}$ included only some of the V. The case where $z = f(V_1, V_2, \ldots) + u$ would add to $(y_p - Y_p)^2$ and $(Y - Y_p)^2$ N.M.II would work well if σ were small and Y didn't change direction frequently; i.e., if Y had "inertia."

For some economic series, next year's value is usually about the same as this year's. More so, for next quarter or next month. Over a number of years Y may change considerably; but usually gradually, step by step. For such economic series, unless the assumptions of the more sophisticated analysis are almost correct, a naive model may be more accurate.

But the naive model cannot learn; it cannot predict the coming of years like 1930; it is helpless in the first years of war or revolution; it cannot tell us what effects will flow from policy decisions.

Estimates of the α_i are important for policy-decisions. But if the L.S. assumptions are false, the estimates β_i' may be biased. A check on the L.S. assumptions can be provided by comparing the goodness of fit of a naive model with that of a fitted regression plane.

We will assume u to be distributed normally. Let:

$$s_1^2 = \frac{\sum_{t=1}^{T} (y - y_f)_t^2}{T - K - 1}$$ K is the number of Z's,

$(y_f)_t$ are the values fitted by L.S.

Let $s_2^2 =$

$$\begin{cases} \dfrac{\sum_{k=1}^{T/2} (y - y_{-1})_{2k}^2}{T} & \text{if T is even} \\ \dfrac{\sum_{k=1}^{T/2-1} (y - y_{-1})_{2k+1}^2}{T - 2} & \text{if T is odd} \end{cases}$$

$$s_3^2 = \begin{cases} \dfrac{\sum_{k=1}^{T/3} (y - 2y_{-1} + y_{-2})_{3k}}{2T} & \text{if T is a multiple of 3} \\ \text{suitably modified if T is not a multiple of 3.} \end{cases}$$

If $(\triangle Y)_t = (\triangle \triangle Y)_t = 0$ for all t, then s_2^2 and s_3^2 would be distributed as $\frac{1}{n} \sum_{i=1}^{n} X_i^2$ where the X_i are distributed normally with mean of zero and variance of σ_u^2. n equals T/2 (or T/2 - 1) and T/3 for s_2^2 and s_3^2 respectively. Then

$$\frac{s_1^2}{s_2^2} = F(T - K - 1, T/2)$$

$$\frac{s_1^2}{s_2^2} = F(T - K - 1, T/3).$$

If $(\triangle Y)_t = (\triangle \triangle Y)_t = 0$ for all t, and if (e.g.) $\frac{s_1^2}{s_2^2}$ exceeded unity significantly according to the F distribution, we would reject the L.S. assumptions. But if (say) some $(\triangle Y)_t \neq 0$, then the probability that $(y - y_{-1})_t^2$ would be as low as or lower than that observed would be less than if $\triangle Y = 0$; the probability that s_2^2 would be as low or lower would be less; the probability that $\frac{s_1^2}{s_2^2}$ would be as great or greater would be less. Therefore, if the L.S. assumptions would be rejected if $\triangle Y = 0$, they would be rejected a - fortiori if $\triangle Y \neq 0$.

The power of the test depends upon the sizes of the $(\triangle Y)_t$ or $(\triangle \triangle Y)_t$.