SINGLE DEGREE OF FREEDOM SUMS OF SQUARES FOR TESTING THE FIT TO A LINEAR MODEL

D. S. Robson

Biometrics Unit, Cornell University, Ithaca, N. Y.

ABSTRACT

An *ad hoc* but exact test of fit to a linear model $E(Y_i|X) = X_i\beta$ which is designed to have power against alternatives of the form H_p: $E(Y_i|X) = (X_i\beta_p)^p$ may be constructed by solving the non-linear moment equations $X'Y = X'(X\tilde{\beta}_p)^p$ and testing the significance of the correlation between $e = Y - X\tilde{\beta}_1$ and $\tilde{e}_p = (X\tilde{\beta}_p)^p - X\tilde{\beta}_1$. Under the hypothesis of the linear model with NIID$(0,\sigma^2)$ errors the test statistic $\tilde{t}_p^2 = (n-r-1)r^2_{e\tilde{e}_p}/(1-r^2_{e\tilde{e}_p})$ is F-distributed, and is a test of H_p in the sense that $t_p^2 = \infty$ when $Y_i = (X_i\beta)^p$ for all i. A more robust test not requiring the specification of p is obtained by computing $\tilde{t}^2_\infty = \lim_{p\to\pm\infty} \tilde{t}_p^2$, which reduces to Tukey's test for nonadditivity in the case where $X\beta$ is the additive model for a two-way classification with one observation per cell. Greater robustness appears to be obtainable by combining $\tilde{t}^2_\infty$ with $\overset{*}{t}{}_1^2 = \lim_{p\to 1} \tilde{t}_p^2$ in the form of a test of significance of the multiple correlation coefficient $R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}$.

INTRODUCTION

We consider here an ad hoc but exact test of fit to the linear model

$$H_1 : Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, I\sigma^2)$$

against the alternative that some power transform of Y is linear in X . In particular, if the alternative is expressed in the form $E(Y_j|X) = (X_j\beta_p)^p$ then for any specified p we may estimate β_p by solving the nonlinear moment equations $X'Y = X'(X\tilde{\beta}_p)^p$, where $\tilde{\beta}_1 = \hat{\beta}$ is the linear least squares estimator. If $\hat{Y} = X\hat{\beta}$ and $e = Y - X\hat{\beta}$ then e is statistically independent of $X'Y$ and $\hat{Y}$ under H_1, so letting $\tilde{Y}^{(p)} = (X\tilde{\beta}_p)^p$ and $\tilde{e}_p = \tilde{Y}^{(p)} - \hat{Y}$ then $\tilde{e}_p$ is statistically independent of e . For a fixed value of $\tilde{e}_p$ the linear function $\tilde{e}_p'e$ is therefore normally distributed with mean zero, and since $X'\tilde{e}_p = 0$ the conditional variance of $\tilde{e}_p'e$ is simply $\tilde{e}_p'\tilde{e}_p\sigma^2$. The single d.f. sum of squares

$$\tilde{S}_p^2 = \frac{(\tilde{e}_p'e)^2}{\tilde{e}_p'\tilde{e}_p} = e'e r^2_{\tilde{e}_p e}$$

due to the regression of e on $\tilde{e}_p$ is therefore H_1-distributed as $\sigma^2\chi^2_{1\ d.f.}$, and the test statistic

$$\tilde{t}_p^2 = \frac{(n-r-1)\tilde{S}_p^2}{e'e - \tilde{S}_p^2} = \frac{(n-r-1)r^2_{\tilde{e}_p e}}{1 - r^2_{\tilde{e}_p e}}$$

has the F-distribution on 1 and n-r-1 d.f. when Y is $n \times 1$ and X is $n \times k$ with rank $r \leq k < n$. This does provide a test against the alternative hypothesis $E(Y|X) = (X\beta_p)^p$ in the sense that if $Y = (X\beta_p)^p$ then $\tilde{S}_p^2 = e'e$, or $\tilde{t}_p^2 = \infty$.

Implementation of this procedure would require specification of p; for example, the choice p=2 would test whether the square root transform of Y improves the fit to a linear model in X . In practice, however, the choice of p is likely to be arbitrary, and this raises the question of how sensitive the test is to the choice of p . If $\tilde{S}_p^2$ is a slowly changing function of p then some degree of arbitrariness in choosing p will not greatly effect the power of the test, and if $\tilde{S}_p^2$ is extremely robust then a limiting value of $\tilde{S}_p^2$ will serve almost as well as any other. With this possibility in mind we note that if the limiting form of $\tilde{e}_p$,

$$\lim_{p\to-\infty} \tilde{e}_p = \lim_{p\to\infty} \tilde{e}_p = \tilde{e}_\infty = \tilde{Y}^{(\infty)} - \hat{Y} ,$$

exists then $\tilde{Y}^{(\infty)}$ must have the form

$$\tilde{Y}_i^{(\infty)} = \tilde{B}_1^{X_{1i}} \tilde{B}_2^{X_{2i}} \cdots \tilde{B}_k^{X_{ki}}$$

where $\tilde{B}_1, \cdots, \tilde{B}_k$ is a solution to the equations

$$\sum_{j=1}^{n} X_{ij} Y_j = \sum_{j=1}^{n} X_{ij} \tilde{B}_1^{X_{1j}} \cdots \tilde{B}_k^{X_{kj}} , \qquad i=1, \cdots, k$$

when such a solution exists. Thus, with $\tilde{e}_\infty$ defined in this manner and

$$r^2_{e\tilde{e}_\infty} = \frac{(e'\tilde{e}_\infty)^2}{(e'e)(\tilde{e}_\infty'\tilde{e}_\infty)}$$

then when Y is exactly the p'th power of $X\beta$, $Y_j = \left(\sum_{i=1}^{k} \beta_i X_{ij}\right)^p$, then $r^2_{e\tilde{e}_\infty}$ approaches unity as p approaches $\pm\infty$. The test statistic $\tilde{t}^2_\infty$ might thus be expected to be robust in power against alternatives with $E(Y|X) = (X\beta)^p$, at least when p is large in absolute value.

If such a test could be combined with another which has power against small p-values the resulting test should perform reasonably well against all p . To this end we note that $r^2_{e\tilde{e}_p}$ is undefined at p=1 but does approach a limit; namely,

$$\lim_{p\to1} r^2_{e\tilde{e}_p} = r^{2*}_{ee_1}$$

where

$$\overset{*}{Y}_i^{(1)} = \hat{Y}_i \log \hat{Y}_i \qquad \overset{*}{e}_1 = \overset{*}{Y}^{(1)} - X\overset{*}{\beta}_1$$

with $X\overset{*}{\beta}_1$ defined by $X'\overset{*}{Y}^{(1)} = X'X\overset{*}{\beta}_1$, provided that $\hat{Y}_i > 0$ for $i=1,\cdots,n$. The test statistic

$$\overset{*}{t}_1^2 = \frac{(n-r-1)r^2_{e\overset{*}{e}_1}}{1-r^2_{e\overset{*}{e}_1}}$$

should thus have desirable power characteristics for p near unity, and combining this with $\tilde{t}^2_\infty$ in the form

$$F_{2,n-r-2} = \frac{(n-r-2)R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}}{2\left(1-R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}\right)}$$

should provide the desired robustness. The multiple correlation coefficient $R_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}$ is defined by

$$R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1} = \frac{r^2_{e\tilde{e}_\infty} + r^2_{e\overset{*}{e}_1} - 2r_{\tilde{e}_\infty\overset{*}{e}_1} r_{e\tilde{e}_\infty} r_{e\overset{*}{e}_1}}{1-r^2_{\tilde{e}_\infty\overset{*}{e}_1}}$$

where

$$r_{\tilde{e}_\infty\overset{*}{e}_1} = \frac{\tilde{e}_\infty'\overset{*}{e}_1}{\sqrt{\left(\tilde{e}_\infty'\tilde{e}_\infty\right)\left(\overset{*}{e}_1'\overset{*}{e}_1\right)}}$$

and the H_1-distribution of $F_{2,n-r-2}$ is then Snedecor's F-distribution with the indicated d.f. .

The power of such tests will depend upon the error structure under the alternative hypothesis as well as depending upon the parameters p and β and the design matrix X. Instead of attempting to specify error structure and evaluate power we have made a preliminary investigation of robustness by selecting some design matrices of simple form and then numerically evaluating $r^2_{e\tilde{e}_\infty}$, $r^2_{e\overset{*}{e}_1}$ and $R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}$ when Y is exactly equal to the p'th power of a specified linear function.

H_1: Simple Linear Regression

As a numerical indication of degree of robustness in the case of simple linear regression we calculated $r^2_{e\tilde{e}_\infty}$, $r^2_{ee_1^*}$ and $R^2_{e\cdot\tilde{e}_\infty e_1^*}$ when $Y_x = (\alpha + \beta X)^p$, with $\alpha + \beta X > 0$. We considered only the case of sample size n=6 with six equally spaced values of the independent variable X and, without loss of generality, we took these values to be $X=0,1,2,\cdots,5$. Also, no generality was lost by taking $\alpha=1$ and $\beta > 0$, since with this design matrix and any given pair of parameters α,β satisfying the constraints $\alpha + \beta X > 0$ for $X=0,1,\cdots,5$ the following three models

$$Y_x = (\alpha+\beta X)^p$$

$$Y_x = (1+\frac{\beta}{\alpha}X)^p$$

$$Y_x = (1-\frac{\beta}{\alpha+5\beta}X)^p$$

produce identical values of the criteria $r^2_{e\tilde{e}_\infty}$, $r^2_{ee_1^*}$ and $R^2_{e\cdot\tilde{e}_\infty e_1^*}$. Thus, the constraint $\alpha + \beta X > 0$ for $X=0,1,2,\cdots,5$ restricts β/α to the interval $-.2 < \beta/\alpha < \infty$, and $\beta/\alpha = \theta > 0$ is equivalent to $\alpha=1$, $\beta = -\theta/(1+5\theta)$ with respect to our chosen criteria.

Graphs of $r^2_{e\tilde{e}_\infty}$, $r^2_{ee_1^*}$ and $R^2_{e\cdot\tilde{e}_\infty e_1^*}$ as functions of β and p when $Y_x = (1+\beta X)^p$, $\beta > 0$, are displayed in Figures 1 - , supplemented by Table I for values of β near zero where these correlations are too near unity to permit graphing. Plotted as a family of functions of p indexed on β, these squared correlations all approach unity as $\beta \to 0$ from either direction. This and other limit points indicated by the numerical results are readily verified analytically through application of l'Hospitale's rule. Thus, the intersection at p=0 is given by

$$\lim_{p\to 0} r^2_{e\tilde{e}_\infty} = \lim_{p\to 0} r^{2*}_{ee_1} = \frac{\left(e'_{z\cdot x} e_{\hat{z}^2\cdot x}\right)^2}{\left(e'_{z\cdot x} e_{z\cdot x}\right)\left(e'_{\hat{z}^2\cdot x} e_{\hat{z}^2\cdot x}\right)}$$

where $Z_x = \log(1+\beta X)$ and $e_{v\cdot x} = V_x - \hat{V}_x$ with

$$\hat{V}_x = \bar{V} + b_{v\cdot x}(X-\bar{X}) \quad .$$

The finite domain of $r^{2*}_{ee_1}$, which conveys a somewhat synthetic appearance in the graphs, is determined by the constraint

$$\hat{Y}_x = \frac{1}{n}\sum_{X=0}^{n}(1+\beta X)^p + \frac{X-\bar{X}}{\Sigma(X-\bar{X})^2}\sum(X-\bar{X})(1+\beta X)^p > 0$$

for $X=0,1,2,\cdots,n$, and can be calculated for any given β . Results suggest that within this range the test statistic

$$F_{2,n-4} = \frac{(n-4)R^2_{e\cdot\tilde{e}_\infty \overset{*}{e}_1}}{2\left(1-R^2_{e\cdot\tilde{e}_\infty \overset{*}{e}_1}\right)}$$

might well have very desirable power characteristics. The test statistic

$$\tilde{t}^2_\infty = \frac{(n-3)r^2_{e\tilde{e}_\infty}}{1-r^2_{e\tilde{e}_\infty}}$$

which represents a linear regression analogue of Tukey's test for non-additivity, would appear to be extremely robust. As anticipated, the test statistic

$$\overset{*}{t}{}^2_1 = \frac{(n-3)r^{2*}_{ee_1}}{1-r^{2*}_{ee_1}}$$

appears to be only locally powerful in a neighborhood of p=1.

Alternative hypotheses in the close neighborhood of p=0 appear to be least favorable with respect to these test procedures, but such alternatives might also be least likely to arise in practice.

In fact, if p departs very far from unity the nonlinearity in this case of a single independent variable should become apparent from inspection of the data and not even require a statistical test; thus there may be an argument made for the test $\overset{*}{t}{}_1^2$. In the case of higher dimension design matrices X, however, nonlinearity becomes less apparent to the inspector and robustness over a wider range of p becomes definitely more desirable. As an illustration we next examine the case where X is a randomized block design matrix; i.e., the case of an additive model of a two-way classification with one observation per cell.

<u>H_1 : The Additive Two-Factor Model</u>

The additive model $EY_{ij} = \alpha_i + \beta_j$ for the rectangular array Y_{ij}, i=1,...,r and j=1,...,c, gives $\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$ and in this case $\tilde{Y}_{ij}^{(\infty)} = \bar{Y}_{i\cdot}\bar{Y}_{\cdot j}/\bar{Y}_{\cdot\cdot}$; thus,

$$\tilde{e}_{\infty ij} = \bar{Y}_{i\cdot}\bar{Y}_{\cdot j}/\bar{Y}_{\cdot\cdot} - \hat{Y}_{ij}$$

and

$$\overset{*}{e}_{1ij} = \hat{Y}_{ij}\log\hat{Y}_{ij} - \frac{1}{c}\sum_j \hat{Y}_{ij}\log\hat{Y}_{ij} - \frac{1}{r}\sum_i \hat{Y}_{ij}\log\hat{Y}_{ij} + \frac{1}{rc}\sum_{i,j}\hat{Y}_{ij}\log\hat{Y}_{ij} .$$

An $r \times c = 3 \times 3$ table with $Y_{ij} = \alpha_i + \beta_j$ was used for numerical illustration, and for graphical simplicity was constructed as a function of a single parameter θ :

i \ j	1	2	3
1	1	1+θ	3-θ
2	1+θ	1+2θ	3
3	3-θ	3	5-2θ

Taking the p'th power of these entries as our observations we calculated $r^2_{e\tilde{e}_\infty}, r^2_{e\overset{*}{e}_1}$ and $R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}$ as functions of p indexed on θ . The constraint $Y_{ij} > 0$ restricts θ to the interval $-.5 < \theta < 2.5$, and since $\theta = \theta_0$ and $\theta = 2-\theta_0$ produce permutations of the same table, the operational range of θ is $-.5 < \theta < 1$. Degeneracies occur at $\theta=0$ and 1 where e, $\tilde{e}_\infty$ and $\overset{*}{e}_1$ are perfectly correlated for all p . Again, because of the requirement $\hat{Y}_{ij}^{(p)} > 0$ the correlations $r_{e\overset{*}{e}_1}$ and $R_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}$ are defined only for p in an interval determined by θ .

The results are similar to those obtained for the simple linear regression model, suggesting that Tukey's test

$$\xi^2_\infty = \frac{[(r-1)(c-1)-1]r^2_{e\tilde{e}_\infty}}{1-r^2_{e\tilde{e}_\infty}}$$

is robust with respect to alternatives H_p: $EY_{ij} = (\alpha_i+\beta_j)^p$ and that

$$F_{2\ (r-1)(c-1)-2} = \frac{[(r-1)(c-1)-2]R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}}{2\left(1-R^2_{e\cdot\tilde{e}_\infty\overset{*}{e}_1}\right)}$$

may be even more robust when applicable.

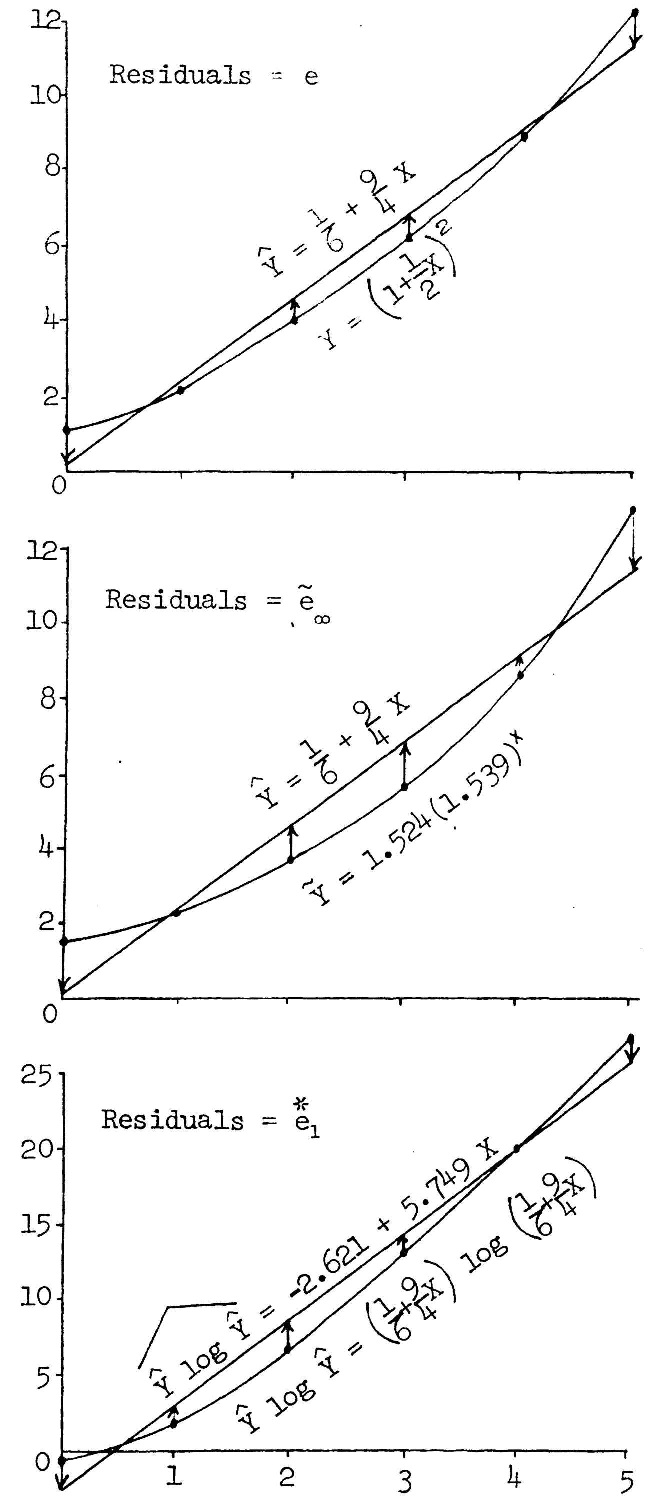

FIG. 1

An illustration of the residuals used in calculating $r^2_{e\tilde{e}_\infty}$, $r^2_{ee^*_1}$ and $R^2_{e\tilde{e}_\infty e^*_1}$ when $Y=(\alpha+\beta X)^p$ for $\alpha=1$, $\beta=.5$ and $p=2$.

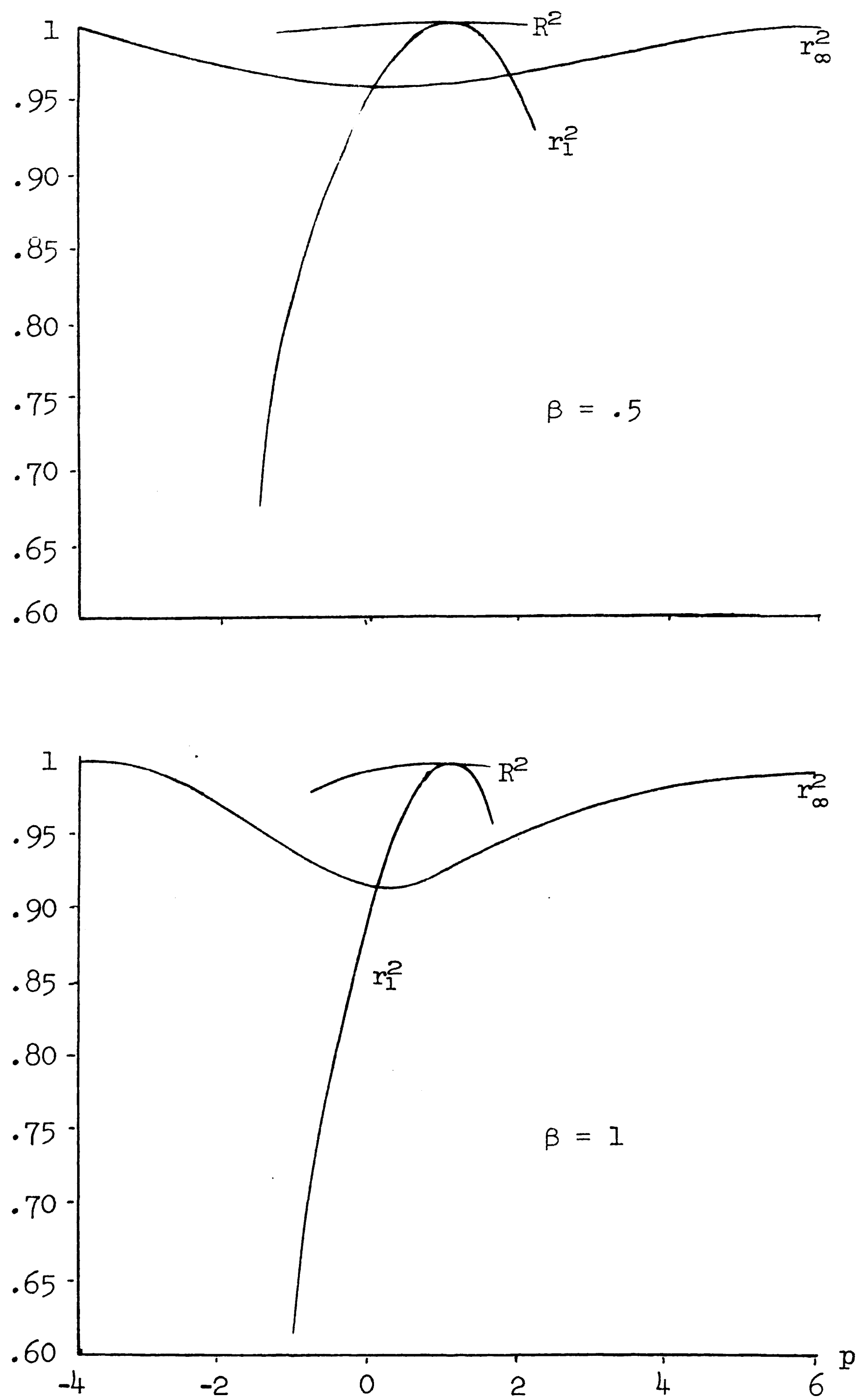

FIG. 2

Graphs of $r^{2}_{ee_1^*}$, $r^{2}_{e\tilde{e}_\infty}$ and $R^{2}_{e\cdot e_1^* \tilde{e}_\infty}$ as functions of p

when $Y = (1+\beta X)^p$ for $\beta = .5$ and 1

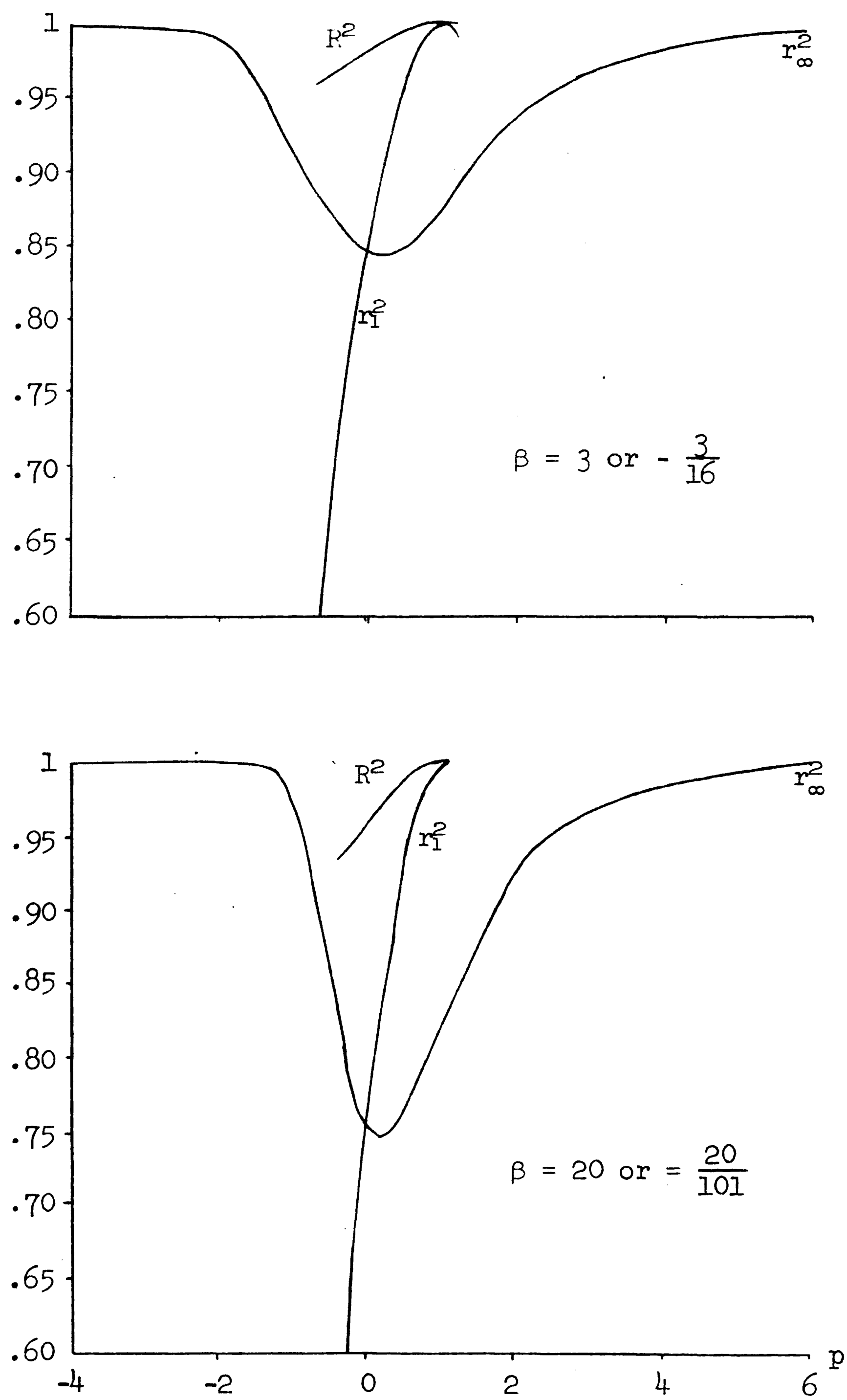

FIG. 3

Graphs of $r^{2}_{ee_1^*}$, $r^2_{e\tilde{e}_\infty}$ and $R^2_{e\cdot e_1^*\tilde{e}_\infty}$ as functions of p

when $Y = (1+\beta X)^p$ for $\beta = 3$ and 20

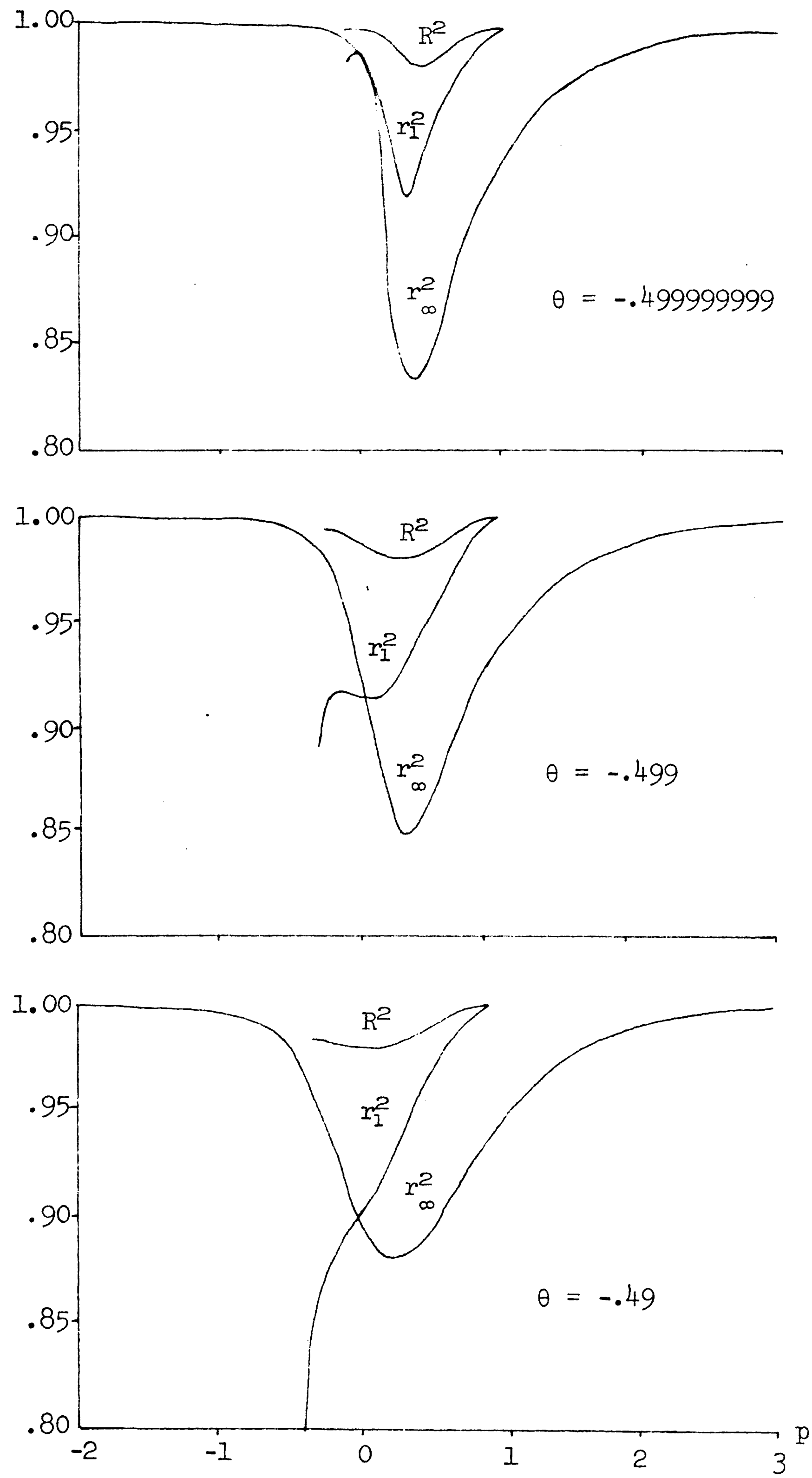

FIG. 4

Graphs of $r_1^2 = r^2_{ee_1^*}$, $r_\infty^2 = r^2_{e\tilde{e}_\infty}$ and $R^2 = R^2_{e\cdot\tilde{e}_\infty e_1^*}$ as functions of p when $Y_{ij} = (\alpha_i+\beta_j)^p$ with $\alpha_1 = \beta_1 = \frac{1}{2}$, $\alpha_2 = \beta_2 = \frac{1}{2} + \theta$, $\alpha_3 = \beta_3 = \frac{5}{2} - \theta$, for θ near $-\frac{1}{2}$.