PROPERTIES OF ORDINARY AND WEIGHTED LEAST SQUARE ESTIMATORS OF REGRESSION COEFFICIENTS FOR TWO-STAGE SAMPLES

Cathy Campbell, University of Minnesota

1. Introduction

In this paper we present a simple regression model for two-stage cluster samples. It is hypothesized that the model may be appropriate for many situations in which both the clusters and elements within the clusters are assumed to be sampled from infinite populations. As such, it is not a model for sampling from finite populations, but may also be considered as a super-population model for two-stage samples taken from finite populations.

The model of interest can be given as

$y_{ij} = \mu + \underline{x}_{ij}'\beta + u_{ij}$ $(i=1,\ldots,b;\ j=1,\ldots,n_i)$,

or in matrix notation as

$$\underline{y} = \underline{1}\mu + X\underline{\beta} + \underline{u} \qquad (1.1)$$

where $\underline{1}$ is an nx1 vector of 1's,

$\underline{y} = (y_{11},\ldots,y_{bn_b})'$ is an nx1 vector of observed dependent variables,

X is an nxp matrix of observed predictor variables,

μ and $\underline{\beta}$ (px1) are unknown parameters

$\underline{u}$ is an nx1 vector of unobserved random variables with

$E(\underline{u}|X) = 0$,

$\mathrm{Var}(\underline{u}|X) = \sigma^2 V$,

$$V = (1-\rho_y)I + \rho_y \begin{bmatrix} J_{n_1} & & \emptyset \\ & \ddots & \\ \emptyset & & J_{n_b} \end{bmatrix} \qquad (1.2)$$

$$= (1-\rho_y)I + \rho_y X_b X_b', \qquad (1.3)$$

J_{n_i} is an $n_i x n_i$ matrix of 1's,

$X_b = \begin{bmatrix} \underline{1}_{n_1} & & \emptyset \\ & \ddots & \\ \emptyset & & \underline{1}_{n_b} \end{bmatrix}$ is the matrix of indicator variables identifying the cluster from which each element was sampled,

ρ_y is the intraclass correlation of the residuals around the regression line,

$$\frac{-1}{n_b-1} < \rho_y \leq 1.$$

From (1.2) and (1.3), it is clear that

$$\mathrm{Var}(y_{ij}) = \sigma^2$$
$$\mathrm{Cov}(y_{ij},y_{ik}) = \rho_y\sigma^2 \quad (j\neq k)$$
$$\mathrm{Cov}(y_{ij},y_{k\ell}) = 0 \quad (i\neq k).$$

Since the constant σ^2 appears as a constant multiplier on all variance expressions, and will cancel from all ratios, for convenience we assume $\sigma^2=1$.

Our interest in this model lies in studying the estimation of $\underline{\beta}$; μ is considered a nuisance parameter. Conditional on X, the weighted least squares (WLS) estimator of $\underline{\beta}$ is BLU, but is rarely used because it depends on unknown parameters and is difficult to compute. More often, because of availability of computer programs and the familiarity of the technique, ordinary least squares (OLS) is used to estimate $\underline{\beta}$. In this paper we wish to consider two aspects of the estimation of $\underline{\beta}$.

(1) Sometimes cluster samples are taken for convenience or economy, sometimes from necessity. What would be the effect on the variance of the parameter estimates if a simple random sampling procedure were used instead? In sampling terminology we wish to study the design effect for the OLS estimator of $\underline{\beta}$.

(2) Is OLS an efficient estimation procedure when model (1.1) holds? If OLS is extremely inefficient, then perhaps some form of approximate WLS, using an estimate of ρ_y, should be considered as an alternative.

For convenience, we restrict our results here to models with one or two predictor variables and consider the issue of design effects first.

2. Design Effect for Simple Linear Regression

When p=1, model (1.1) becomes

$$\underline{y} = \underline{1}\mu + \underline{x}\beta + \underline{u}. \qquad (2.1)$$

We assume that $\underline{x}$ has been transformed so that $\underline{x}'\underline{1} = 0$. Then the OLS estimator of β is given by

$$\hat{\beta}_o = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}, \qquad (2.2)$$

and

$$\mathrm{Var}(\hat{\beta}_o|\underline{x}) = (\underline{x}'\underline{x})^{-2}\underline{x}'V\underline{x}. \qquad (2.3)$$

Following Frankel (1971), we define the design effect of $\hat{\beta}_o$, $\mathrm{Deff}(\hat{\beta}_o)$, as the ratio of the variance of $\hat{\beta}_o$ under model (1.1) to the variance of $\hat{\beta}_o$ under the assumption of a simple random selection of elements of the same overall sample size. As $\mathrm{Var}(y_{ij})=\sigma^2=1$, $\mathrm{Var}(\hat{\beta}_o|\underline{x})$ with simple random sampling is $(\underline{x}'\underline{x})^{-1}$. Therefore

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = \frac{(\underline{x}'\underline{x})^{-2}\underline{x}'V\underline{x}}{(\underline{x}'\underline{x})^{-1}} = (\underline{x}'\underline{x})^{-1}\underline{x}'V\underline{x}. \qquad (2.4)$$

More correctly, the expression (2.4) should be called a conditional design effect since the same $\underline{x}$ is used in both numerator and denominator.

Without loss of generality, we may assume $\underline{x}'\underline{x} = 1$ in (2.4) and obtain

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = \underline{x}'V\underline{x} \quad (\underline{x}'\underline{x} = 1,\ \underline{x}'\underline{1} = 0). \qquad (2.5)$$

Substituting (1.3) for V in (2.5) yields

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + (\underline{x}'X_bX_b'\underline{x}-1)\rho_y. \qquad (2.6)$$

To make (2.6) more easily comparable with the usual expressions for design effects, it is convenient to express $\underline{x}'X_bX_b'\underline{x}$ in terms of the intra-

class correlation, ρ_x, of the observed $\underline{x}$. We first note that $\underline{x}'X_b X_b'\underline{x}$ is the sum of squares of the b cluster totals of $\underline{x}$, and can be expressed as

$$\underline{x}'X_b X_b'\underline{x} = \sum_{i=1}^{b} n_i^2 \bar{x}_i^2 . \qquad (2.7)$$

Following Murthy (1967), we use the following definition of ρ_x that is applicable for unequal cluster sizes:

$$\rho_x = \frac{\sum_{i=1}^{b}\sum_{k\neq j}^{n_i}\sum^{n_i}(x_{ij}-\bar{x})(x_{ik}-\bar{x})}{\sum_{i=1}^{b} n_i(n_i-1)\sigma_x^2} . \qquad (2.8)$$

Using the relationships $\sigma_x^2 = \frac{1}{n}$ and $\bar{x} = 0$, (2.8) reduces to

$$\rho_x = \frac{\Sigma n_i^2 \bar{x}_i^2 - 1}{\frac{\Sigma n_i^2}{\Sigma n_i} - 1} ,$$

which gives

$$\Sigma\, n_i^2 \bar{x}_i^2 = 1 + \left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right)\rho_x . \qquad (2.9)$$

Substituting (2.9) in (2.6) gives

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + \left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right)\rho_x\rho_y \qquad (2.10)$$

$$= 1 + \left(\frac{\mathrm{Var}(n_i)}{\bar{n}} + \bar{n} - 1\right)\rho_x\rho_y \qquad (2.11)$$

where $\bar{n} = \sum_{i=1}^{b} n_i/b$ is the average sample size.

Noting that (2.9) is the design effect (see (2.13)) for estimating the mean of $\underline{x}$, we can also obtain

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + (\mathrm{Deff}(\bar{x}) - 1)\rho_y . \qquad (2.12)$$

We now wish to make the following points about $\mathrm{Deff}(\hat{\beta}_o|\underline{x})$:

(i) When the sample sizes are all equal,

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + (\bar{n}-1)\rho_x\rho_y . \qquad (2.13)$$

(ii) To obtain the design effect for estimating μ under model (1.1), we let $\underline{x}$ in (2.10) be $\frac{1}{\sqrt{n}}\underline{1}_n$ and define the intraclass correlation for a column of 1's as 1. Then (2.13) reduces to

$$\mathrm{Deff}(\hat{\mu}_o) = 1 + \left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right)\rho_y , \qquad (2.14)$$

which also shows why (2.9) is $\mathrm{Deff}(\bar{x})$.

(iii) With equal sample sizes, (2.14) becomes

$$\mathrm{Deff}(\hat{\mu}_o) = 1 + (\bar{n}-1)\rho_y , \qquad (2.15)$$

the well-known design effect for cluster samples.

(iv) If ρ_x and ρ_y have the same sign, then $\mathrm{Deff}(\hat{\beta}_o|\underline{x}) > 1$, while the converse holds if ρ_x and ρ_y have opposite signs.

(v) If either ρ_x or ρ_y is 0, then $\mathrm{Deff}(\hat{\beta}_o) = 1$.

(vi) If $\rho_x > 0$ and $\rho_y > 0$, then

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) \leq \mathrm{Deff}(\hat{\mu}_o)$$

and

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) \leq \mathrm{Deff}(\bar{x}) .$$

The last point is an important piece of theoretical evidence in support of Kish and Frankel's (1974) observation that design effects for complex statistics (including regression coefficients) tend to be less than design effects for means.

The fact that the design effects for means obtained from this model reduce to those used in practice for balanced samples is encouraging as is the fact that the empirical observation of Kish and Frankel is supported by the use of model (1.1).

Unfortunately, at the time of this writing, we do not have empirical values of the design effects for single variable regressions with which to compare (2.13) or (2.11). Therefore, it is not yet possible to verify the applicability of these results to sample survey situations.

3. Design Effects in a Two-Variable Regression

The model we use here is

$$\underline{y} = \underline{1}\mu + \underline{x}_1\beta_1 + \underline{x}_2\beta_2 + \underline{u} \qquad (3.1)$$

with $\underline{x}_1'\underline{1} = \underline{x}_2'\underline{1} = 0$, and $\underline{x}_1'\underline{x}_1 = \underline{x}_2'\underline{x}_2 = 1$. The variance of the OLS estimator of $(\beta_1,\beta_2)'$ is

$$\mathrm{Var}\begin{pmatrix}\hat{\beta}_{1o}\\ \hat{\beta}_{2o}\end{pmatrix} X = (X'X)^{-1}X'VX(X'X)^{-1} , \qquad (3.2)$$

where $X = [\underline{x}_1,\underline{x}_2]$.

With the restriction $\underline{x}_1'\underline{x}_1 = \underline{x}_2'\underline{x}_2 = 1$,

$$X'X = \begin{bmatrix}1 & r\\ r & 1\end{bmatrix} \qquad (3.3)$$

and

$$(X'X)^{-1} = \frac{1}{1-r^2}\begin{bmatrix}1 & -r\\ -r & 1\end{bmatrix} . \qquad (3.4)$$

The center matrix, X'VX, is

$$X'VX = (1-\rho_y)X'X + \rho_y\begin{bmatrix}\Sigma\, n_i^2\bar{x}_{1i}^2 & \Sigma\, n_i^2\bar{x}_{1i}\bar{x}_{2i}\\ \Sigma\, n_i^2\bar{x}_{1i}\bar{x}_{2i} & \Sigma\, n_i^2\bar{x}_{2i}^2\end{bmatrix} . \qquad (3.5)$$

Using (2.9), the diagonal elements of (3.5) can be easily represented in terms of the intraclass correlations ρ_{x_1} and ρ_{x_2}. By generalizing (2.8), we define the intraclass "co-correlation" of $\underline{x}_1$ and $\underline{x}_2$ as

$$\rho_{X_1X_2} = \frac{\sum_{i=1}^{b}\sum_{k\neq j}^{n_i}\sum^{n_i}(x_{1ij}-\bar{x}_1)(x_{2ik}-\bar{x}_2)}{\sum_{i=1}^{b} n_i(n_i-1)\sqrt{\sigma^2_{X_1}\sigma^2_{X_2}}}, \qquad (3.6)$$

which reduces to

$$\rho_{X_1X_2} = \frac{\sum n_i^2 \bar{x}_{1i}\bar{x}_{2i} - r}{\frac{\sum n_i^2}{\sum n_i} - 1}. \qquad (3.7)$$

We note that the sign of $\rho_{X_1X_2}$ does not depend on the sign of the covariance between the cluster totals, but on whether this covariance is larger or smaller than the overall correlation between $\underline{x}_1$ and $\underline{x}_2$. If the cluster totals are uncorrelated, then

$$\rho_{X_1X_2} = \frac{-r}{\frac{\sum n_i^2}{\sum n_i} - 1}.$$

To form the design effect for $\hat{\beta}_{1o}$, we perform the necessary matrix multiplication in (3.2), using (2.9) and (3.7) in (3.5), to find

$$\text{Deff}(\hat{\beta}_o|X) = 1 + \left[\frac{\sum n_i^{2-2}x_{1i} - 2r\sum n_i^{2-}\bar{x}_{1i}\bar{x}_{2i} + r^2\sum n_i^{2-2}x_{2i}}{1-r^2} - 1\right]\rho_Y \qquad (3.8)$$

$$= 1 + \left(\frac{\sum n_i^2}{\sum n_i} - 1\right)\left(\frac{\rho_{X_1} - 2r\rho_{X_1X_2} + r^2\rho_{X_2}}{1-r^2}\right)\rho_Y. \qquad (3.9)$$

Due to the number of parameters involved, it is difficult to make general statements about the value of $\text{Deff}(\hat{\beta}_{1o}|X)$. However, we can notice that:

(i) if $r = 0$, then (3.9) reduces to the single variable design effect of (2.10);

(ii) if $\rho_Y > 0$, then $\text{Deff}(\beta_{1o}|X)$ increases with ρ_{X_1} and ρ_{X_2};

(iii) $\text{Deff}(\hat{\beta}_{1o}|X)$ is larger if r and $\rho_{X_1X_2}$ have opposite signs than if they have the same sign.

(iv) $\text{Deff}(\hat{\beta}_{1o}|X)$ becomes very large if r approaches 1 or -1.

Perhaps a more intuitive parametrization of $\text{Deff}(\hat{\beta}_{1o}|X)$ occurs when it is expressed in terms of the design effects of x_1 and x_2. By letting ρ_{12} be the correlation coefficient between the block totals, we obtain

$$\text{Deff}(\hat{\beta}_{1o}|X) = \qquad (3.10)$$

$$1 + \left[\frac{\text{Deff}(\bar{x}_1) - 2r\rho_{12}\sqrt{\text{Deff}(\bar{x}_1)\text{Deff}(\bar{x}_2)} + r^2\text{Deff}(\bar{x}_2)}{1-r^2} - 1\right]\rho_Y.$$

To evaluate (3.10), we used data from Frankel's (1971) three variable regressions, considering the variables pairwise. Values of $\sqrt{\text{Deff}(\bar{x}_i)}$ were included in his appendix E. Values of r were available from Table 5.1. Values for ρ_Y were obtained by assuming $\text{Deff}(\bar{y}) = 1 + (\bar{n}-1)\rho_Y$ and solving for ρ_Y. As sufficient data were not available for evaluating ρ_{12}, we assumed it was equal to r. Only data from the six strata designs are used.

Frankel considered 2 different three variable regressions. Since we used the variables in pairs in 2 variable equations, (3.10) was evaluated twice for each regression coefficient. The results of our calculations and Frankel's empirically obtained values are given below.

Comparison of Theoretical and Empirical Design Effects

Variable	$\text{Deff}(\hat{\beta}_{io})$ from (3.10)		$\text{Deff}(\hat{\beta}_{io})$ from Frankel
6	1.067,	1.063	1.089
7	1.092,	1.088	1.134
12	1.089,	1.089	.984
8	1.058,	1.057	1.093
11	1.126,	1.128	1.080
17	1.251,	1.252	1.432

Variables 6, 7, and 12 were predictor variables in one equation while 8, 11, and 17 were included in the other equation. Except for variables 12 and 11, the results from (3.10) are somewhat smaller but ordered approximately the same as Frankel's results. Variable 12 is clearly an anomaly for which we have no explanation at this time. From Frankel's data we found variables 6 and 7 were highly correlated with each other and correlated only slightly with variable 12. Until results are obtained for three variable regressions, we do not know whether this explains the small design effect for variable 12.

In this section and the preceding one, we have presented expressions for conditional design effects for regression coefficients. These were obtained by assuming the data follow a simple linear model appropriate for two-stage sampling from infinite populations. It is hoped that these results may also shed some light on the properties of regression coefficients obtained from finite populations.

The comparisons in the above table are not totally discouraging. Further investigation is needed to determine whether the discrepancies are due to differences between two-variable and three-variable regressions or from some oversimplification in the assumed model.

4. Relative Efficiency of OLS for Cluster Samples

In this section we study the efficiency of OLS with respect to WLS when model (1.1) holds. We consider only single variable regressions and define the relative efficiency as

$$E^* = \frac{\text{Var}(\hat{\beta}_w|\underline{x})}{\text{Var}(\hat{\beta}_o|\underline{x})}, \qquad (4.1)$$

where $\hat{\beta}_w$ is the second element of

$$\begin{pmatrix} \hat{\mu}_w \\ \hat{\beta}_w \end{pmatrix} = (Z'V^{-1}Z)^{-1}Z'V^{-1}\underline{y}$$

with $Z = [\underline{1},\underline{x}]$. $\mathrm{Var}(\hat{\beta}_w)$ is the (2,2) element of $(Z'V^{-1}Z)^{-1}$. As before we assume $\underline{x}'\underline{1} = 0$ and $\underline{x}'\underline{x} = 1$. The efficiencies given here are a pessimistic reflection of the efficiency of OLS since $\mathrm{Var}(\hat{\beta}_w)$ can never be achieved.

It can be shown that

$$E^* = \frac{(1-\rho_Y)}{[1-\underline{a}_2'D\underline{a}_2][1+(\underline{a}_2'N\underline{a}_2-1)\rho_Y]}, \qquad (4.2)$$

where $\underline{a}_2' = [\sqrt{n_1}\bar{x}_1,\ldots,\sqrt{n_b}\bar{x}_b]$

$$N = \mathrm{diag}\,(n_1,\ldots,n_b)$$

$$D = \mathrm{diag}\left(\frac{n_1\rho_Y}{1+(n_1-1)\rho_Y},\ldots,\frac{n_b\rho_Y}{1+(n_b-1)\rho_Y}\right).$$

We note that

$$0 \le \underline{a}_2'\underline{a}_2 = \Sigma\, n_i\bar{x}_i^2 \le 1$$

is the between-cluster sum of squares of $\underline{x}$, since $\bar{x} = 0$. As such it is the length of the projection of $\underline{x}$ into the subspace spanned by X_b. The vector $\underline{a}_2$ contains the between-cluster information for regressing $\underline{y}$ on $\underline{x}$. If $\underline{a}_2 = \underline{0}$, all cluster means are 0 and $\underline{x}$ varies only within the clusters.

Rather than discussing the properties of E^*, we give some graphs of it in simple situations. We choose to represent $\underline{a}_2$ as

$$\underline{a}_2 = \sqrt{k}\,\underline{u}$$

where $\underline{u}'\underline{u} = 1$ and $k = \underline{a}_2'\underline{a}_2$. Using this representation, $\underline{a}_2'N\underline{a}_2 = k\underline{u}'N\underline{u}$, where the quantity $\underline{u}'N\underline{u}$ is a weighted average of the n_i and must satisfy

$$n_1 \le \underline{u}'N\underline{u} \le n_b . \qquad (4.3)$$

The restriction $\underline{x}'\underline{1} = 0$ translates to $\underline{u}'\underline{\ell} = 0$

where $\ell' = \left(\sqrt{\frac{n_1}{n}},\ldots,\sqrt{\frac{n_b}{n}}\right)$.

With this restriction on $\underline{u}$, equality on the left in (4.3) can be attained only if $n_1 = n_2$ and on the right only if $n_{b-1} = n_b$.

We also point out that k is a linear function of the intraclass correlation of $\underline{x}$ via

$$k = \frac{\rho_X\left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right) + 1}{\underline{u}'N\underline{u}} . \qquad (4.4)$$

For the illustrations, we consider only balanced samples and use

$$E^* = \frac{(1-\rho_Y)\,(1-\rho_Y+\bar{n}\rho_Y)}{(1-\rho_Y+\bar{n}\rho_Y(1-k))(1-\rho_Y+\bar{n}\rho_Y k)} \qquad (4.5)$$

which is obtained from (4.2) with $n_1 = \ldots = n_b = \bar{n}$. The relationship between k and ρ_X simplifies to

$$k = \frac{1 + (\bar{n}-1)\rho_X}{\bar{n}} \qquad (4.6)$$

for balanced samples.

In figures 1 and 2 we present graphs of E^* versus k (or ρ_X) for different values of ρ_Y. Figure 1 contains results for $\bar{n} = 100$ and Figure 2 for $\bar{n} = 50$. Small values of ρ_Y and k such as are commonly found in sample survey data were used in the calculations.

If we define "reasonable efficiency" as $E^* \ge 0.75$, then with $\bar{n} = 100$ the efficiency of OLS could be unreasonably low if $\rho_Y > 0.05$ unless ρ_X is very small. With $\bar{n} = 50$, the values of E^* remain high until $\rho_Y > 0.10$. Given the small values of ρ_Y and ρ_X commonly present in social science data, OLS should be reasonably efficient for most ρ_X and ρ_Y when $\bar{n} \le 50$. With large values of $\bar{n}$, some inefficient estimates may result if OLS is used consistently.

We also point out that with large total sample sizes, $\bar{n}$ and/or b both large, then $\mathrm{Var}(\hat{\beta}_o)$ and $\mathrm{Var}(\hat{\beta}_w)$ may both be acceptably small even though the efficiency of OLS is low - making it not worthwhile to attempt a WLS analysis.

In conclusion, with the presence of clustering as modelled in (1.1), it appears that OLS is a reasonably efficient estimator of a single regression coefficient for many parameter values commonly obtained in social science data. Therefore standard computer programs can usually be used for calculating point estimates of regression coefficients. The properties of the estimated standard error of $\hat{\beta}_o$ provided by an OLS routine when model (1.1) holds have not been investigated at this time.

REFERENCES

Frankel, Martin R. (1971) Inference from Sample Surveys: An Empirical Investigation. Ann Arbor: Institute for Social Research, The University of Michigan.

Kish, Leslie and Frankel, Martin R. (1974). "Inference from complex samples." The Journal of the Royal Statistical Society, Series B, 36. No. 1, pp. 1-37.

Murthy, M.N. (1967). Sampling: Theory and Methods. Calcutta: Statistical Publishing Society.

ACKNOWLEDGMENT

This research is based, in part, on the author's Ph.D. dissertation which was completed at Southern Methodist University.

Figure 1

E* vs. $K(\rho_x)$ for Different Values of ρ_y: $\bar{n} = 100$

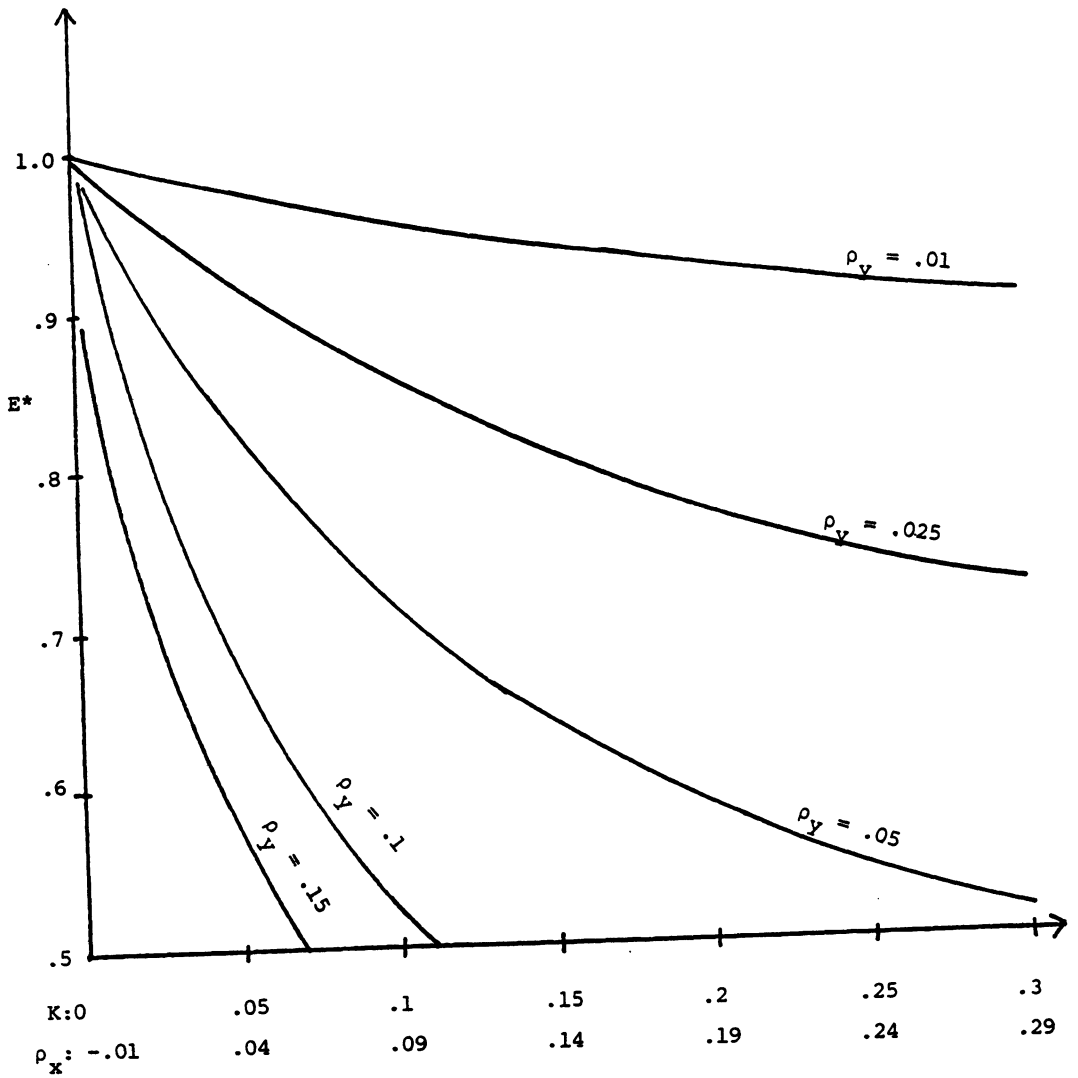

Figure 2

E* vs. ρ_x for Different Values of ρ_y: $\bar{n}$ = 50