TESTING HYPOTHESIS ABOUT REGRESSION COEFFICIENT IN SAMPLES FROM A BIVARIATE NORMAL POPULATION

J. E. Grimes, California Polytechnic State University & NASA-Ames Research Center

B. V. Sukhatme, Deceased

1. INTRODUCTION

Let Y be the characteristic under study and consider the problem of estimating the finite population mean μ_y based on a random sample of size n drawn from the population. If data on an auxiliary variate X correlated with Y are available or can be obtained for all the units in the population and the finite population mean μ_X of the variate X is known, it is generally possible to estimate the population mean μ_y with greater precision than by using the sample mean $\bar{y}$. If the relationship between Y and X is linear, regression-type estimators are generally used to estimate μ_y.

A frequently used estimator (see, e.g., Cochran [1]) is the linear regression estimator

$$\bar{y}_\ell = \bar{y} + \hat{\beta}(\mu_X - \bar{x}) \tag{1.1}$$

where $\bar{y}$ and $\bar{x}$ are the sample means and $\hat{\beta}$ is the estimated regression coefficient of Y and X given by

$$\hat{\beta} = s_{xy}/s_x^2 \tag{1.2}$$

with

$$s_{xy} = \sum_i^n (x_i - \bar{x})(y_i - \bar{y})/(n-1) \tag{1.3}$$

and

$$s_x^2 = \sum_i^n (x_i - \bar{x})^2/(n-1). \tag{1.4}$$

As an example, consider the problem of estimating the total area of the leaves on a plant. Since the area of a leaf and its weight are correlated, Watson [10] used weight of the leaf as an auxiliary variate to estimate total area of the leaves of a plant. Yates [11] gives another application where cultivated area of a farm is used as an auxiliary variate to estimatè acreage under wheat.

The regression estimator $\bar{y}_\ell$ is generally a biased estimator of the population mean μ_y, the bias vanishing when the relationship between Y and X is linear. Further, assuming x and y are bivariate normal, its variance to terms of order n-2 is given by

$$V(\bar{y}_\ell) = \frac{\sigma_y^2(1 - \rho^2)}{n}[1 + \frac{1}{n-3}] \tag{1.5}$$

where σ_y^2 and σ_x^2 are the variances of Y and X and ρ is the correlation coefficient between Y and X, (see e.g., Sukhatme and Sukhatme [9]).

Another regression-type estimator used frequently in the type of situation considered here is the so called difference estimator suggested by Hansen, Hurwitz and Madow [5], defined as

$$\bar{y}_d = \bar{y} + \beta_0(\mu_X - \bar{x}) \tag{1.6}$$

where β_0 is a preassigned constant assumed to be known. It can be shown that $\bar{y}d$ is an unbiased estimator of μ_y and its variance is given by

$$V(\bar{y}_d) = \frac{\sigma_y^2(1 -\rho^2)}{n}[1 +\delta^2] \tag{1.7}$$

where

$$\delta = \frac{\rho}{\{1-\rho^2\}^{\frac{1}{2}}}(1 - \frac{\beta_0}{\beta}). \tag{1.8}$$

For fixed $\rho \neq 0$ it is clear that $V(\bar{y}_d)$ is minimum when $\delta = 0$, i.e., when β_0 is equal to β. In practice β is rarely known. If the prior distribution of β is available, then Bayesian techniques can be used to estimate μ_y. For this, the reader is referred to Han [4] and Mehta and Swamy [8]. Usually only partial information concerning the nature of the prior distribution is available.

Based on past experience, it may be possible to make a guess β_0 of the true value β. When this is the case $\bar{y}_d$ would be preferred as an estimator of μ_y. Otherwise, we would prefer $\bar{y}_\ell$ as an estimator of μ_y. In other words, based upon the relative closeness of β_0 to β, we would prefer to consider a pooled estimator of the type

$$\bar{y}_w = \omega(t)\bar{y}_d + [1 - \omega(t)]\bar{y}_\ell \tag{1.9}$$

where $\omega(t)$ is a function of statistic t used to test the hypothesis $\beta = \beta_0$ against the alternative $\beta \neq \beta_0$. Such estimators were first proposed by Huntsberger [6] and later by Mehta and Gurland [7] who discuss problems concerning the choice of the weighting function $\omega(t)$.

In this paper, we shall restrict our choice of $\omega(t)$ to a function of the type

$$\begin{aligned} \omega(t) &= 1 \quad \text{if} \quad |t| \leq t_0 \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{1.10}$$

For the above choice of the function $\omega(t)$, the estimator $\bar{y}_w$ reduces to what is known as an estimator based upon a preliminary test of significance. We shall call this estimator the sometimes regression estimator and denote it by $\bar{y}_s$.

2. Properties of Sometimes Regression Estimator

In this section the sometimes regression estimator is defined and its preliminary test is discussed.

The estimator $\bar{y}_s$ may now be defined as

$$\begin{aligned} \bar{y}_s &= \bar{y} + \beta_0(\mu_X - \bar{x}) \text{ if } |t| \leq t_0 \\ &= \bar{y} + \hat{\beta}(\mu_X - \bar{x}) \text{ if } |t| > t_0 \end{aligned} \tag{2.1}$$

where

$$t = \frac{\{n-2\}^{\frac{1}{2}}(\hat{\beta} - \beta_0)s_x}{s_y\{1-r^2\}^{\frac{1}{2}}}, \tag{2.2}$$

$$s_y^2 = \sum_i^n (y_i - \bar{y})^2/(n-1), \tag{2.3}$$

and

$$r = s_{xy}/s_x s_y. \tag{2.4}$$

t_0 being a fixed positive constant.

This estimator is based on a preliminary test using the test statistic of 2.2. It is possible to define in an analogous manner a sometimes estimator of the regression coefficient as

$$\hat{\beta}_s = \beta_0 \quad \text{if } |t| \leq t_0 \qquad\qquad (2.5)$$
$$= \hat{\beta} \quad \text{if } |t| > t_0.$$

It can be noted that $\hat{\beta}_s$ is a biased estimator of β since

$$E(\hat{\beta}_s) = \beta_0 + E[(\hat{\beta} - \beta_0)|A^c]\ P(A^c).$$

3. Power Function of Test Statistic

In this section the power function for the test statistic is derived. Usually, the power function of the test of a hypothesis is used to compare it to other tests. In this case the power function can be used to determine the appropriate level of the test.

Theorem 3.1. The power function for the test statistic of (2.2) is given by

$$P(|t| > t_0) = K\sum_{i=0}^{\infty}\frac{(2\theta)^{2i}\ \Gamma(\frac{2i+1}{2})\ \Gamma(\frac{n+2i-1}{2})}{\Gamma(2i+1)} \qquad (3.1)$$
$$I_{m_o}\left(\frac{n-2}{2}, \frac{2i+1}{2}\right)$$

where

$$m_0 = (n-2)/(t_0^2 + n-2)$$
$$K = \{\Pi\}^{\frac{1}{2}}\ \Gamma(\frac{n-1}{2})(1-\delta^2)^{\frac{n-1}{2}}$$
$$\Theta = \delta/(1+\delta^2)^{\frac{1}{2}}$$

and $I_.$ (.,.) is the incomplete beta function.

Proof: Taking the joint density function for s_x, s_y, and r, Cramer [2], and making the transformation

$$u = (n-1)s_x^2/2\sigma_x^2(1-\rho^2),$$
$$v = (n-1)rs_xs_y/2\sigma_x\sigma_y(1-\rho^2),$$

and

$$t' = t/\{n-2\}^{\frac{1}{2}} = (\hat{\beta} - \beta_0)s_x/s_y\{1-r2\}^{\frac{1}{2}}$$

it can be seen that the joint density of u, v and t' is

$$f(u,v,t') = \frac{h(u,v,t')}{t'^{n-1}}\sum_{i=0}^{\infty}\frac{(2\delta)^i(1-\rho^2)^{i/2}}{\Gamma(i+1)}\left(v - \frac{\beta_0u\sigma_x}{\sigma_y}\right)^{n+i-2} \quad \text{in } R_1$$
$$= \frac{h(u,v,t')}{|t'|^{n-1}}\sum_{i=0}^{\infty}\frac{(-2\delta)^i(1-\rho^2)^{i/2}}{\Gamma(i+1)}\left|v - \frac{\beta_0u\sigma_x}{\sigma_y}\right|^{n+i-2} \quad \text{in } R_2$$
$$= 0 \qquad \text{otherwise}$$

where

$$h(u,v,t') = \frac{2^{n-2}(1-\rho^2)^{\frac{n-1}{2}}}{\Pi\Gamma(n-2)u}\exp[-u(1-\rho^2)(1+\delta^2) - \frac{1+t'^2}{ut'^2}(v - \frac{\beta_0u\sigma_x}{\sigma y})^2],$$

$$R_1 = \{0 \leq u < \infty,\ 0 \leq t' < \infty,\ v \geq \frac{\beta_0u\sigma_x}{\sigma y}\},$$

and $$R_2 = \{0 \leq u < \infty,\ -\infty \leq t' < 0,\ v < \frac{\beta_0u\sigma_x}{\sigma_y}\}.$$

Hence

$$P(|t'| > t_0') = I_1 + I_2$$

where

$$I_j = \int_{R_{j0}} f(u,v,t')\ dv\ du\ dt',\ j = 1,2$$

$$R_{10} = \{0 \leq u < \infty,\ -\infty < t' < -t_0',\ v < \frac{\beta_0u\sigma_x}{\sigma_y}\},$$

and

$$R_{20} = \{0 \leq u < \infty,\ t_0' < t' < \infty,\ v \geq \frac{\beta_0u\sigma_x}{\sigma_y}\}.$$

Evaluating the two integrals in the order indicated and using the fact that

$$\int_{-\infty}^{\infty}|x|^n\ \exp(\frac{-x^2}{2})\ dx = 2^{\frac{n-1}{2}}\ \Gamma(\frac{n+1}{2})$$

and $$\Gamma(\frac{j-2}{2})\ \Gamma(\frac{j-1}{2})\ 2^{j-3} = \Gamma(j-2)\ \{\Pi\}^{\frac{1}{2}}$$

we obtain the desired result.

4. Estimation of β

We have noted three methods for estimating the regression coefficient

$$\hat{\beta}_d = \beta_0, \qquad (4.1)$$
$$\hat{\beta}_\ell = \hat{\beta} \qquad (4.2)$$

and $$\beta_s = \beta_0 \quad \text{if } |t| < t_0 \qquad (4.3)$$
$$= \hat{\beta} \quad \text{if } |t| > t_0$$

If conditions are such that the estimate of the regression coefficient is desired, the question arises as to when the sometimes estimator of the regression coefficient would be most appropriate. Actually the sometimes estimator of the regression coefficient includes both β_d and β_ℓ as special cases. Hence, the sometimes estimator of the regression coefficient may be used whenever it is appropriate to estimate the regression coefficient.

There are essentially four situations that can exist:

a. We have no guessed value of β_0. In this case it would be appropriate to choose $t_0 = 0$ and thus $\hat{\beta}_s = \beta_\ell$.

b. We have a guessed value β_0 and are highly confident that our guessed value is close to the true value β.

c. We have a guessed value β_0 and are highly confident that our guessed value is not likely to be close to the true value β.

d. We have a guessed value β_0 and are not sure of the relationship of the value of β_0 to the true value of β.

The situation of case (a) requires no further discussion. Cases (b), (c), and (d) require further examination of how to determine an appropriate value of t_0. A valuable tool in making this determination is the power function (3.1) along with the level of the type I error. Plots of the power function for various sample sizes are given in the figures.

Consider first the case of (b) where we are confident of the guessed value of β_0 and hence are primarily concerned about the level of the type I error (α). Suppose that the sample size is $n = 35$; that $\alpha = .05$; and that $|\delta| < 1/\{n-3\}^{\frac{1}{2}} = .213$ which implies that $\Theta < .194$. From figure C we have that the power of the test is no better than .24.

Next consider the case of (c) where we are highly confident that β_0 is not close to the true value of β. In this case we are more concerned about the power of the test. Suppose that we have taken a sample of size $n = 35$; that the power is to be at least .9; and that $|\delta| > 1/\{n-3\}^{\frac{1}{2}} = .213$ which implies that $\Theta > .194$. From figure C we have that the value of t_0 is approximately 1.17. The level of α would be approximately .25. Hence, the trade-off was not too bad.

For case (d) with a sample of size $n = 35$, and intermediate value for t_0 between those used for cases (b) and (c) should be used.

In a similar manner to the procedure indicated above, choices of t_0 can be made when the sample size is different from $n = 35$.

In conclusion, if a guessed value, β_0, for β is available, it is best to use the sometimes estimator (2.5) of the regression coefficient. The value of t_0 should be chosen in the manner indicated above taking into account the confidence in the value of β_0 and other factors such as the cost of making an incorrect decision in the test.

REFERENCES

[1] Cochran, W. G. (1978). "Sampling Techniques," Third Edition. New York, N.Y., John Wiley and Sons, Inc.

[2] Cramer, Harold (1946). "Mathematical Methods of Statistics," Princeton, N.J., Princeton University Press.

[3] Grimes, J. E. (1973). "Regression Type Estimators Based on Preliminary Test of Significance," Ph.D. Dissertation, Iowa State University, Ames, Iowa.

[4] Han, C. P. (1973). "Double Sampling with Partial Information on Auxiliary Variables," Journal of the American Statistical Association, 68, 914-918.

[5] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). "Sample Survey Methods and Theory," Vol. I, New York, N.Y., John Wiley and Sons, Inc.

[6] Huntsberger, D.V. (1955). "A Generalization of a Preliminary Testing Procedure for Pooling Data," Annals of Mathematical Statistics, 26, 734-43.

[7] Mehta, J.S. and Gurland, John (1969). "On Utilizing Information from a Second Sample in Estimating Variance," Biometrika, 56, 527-532.

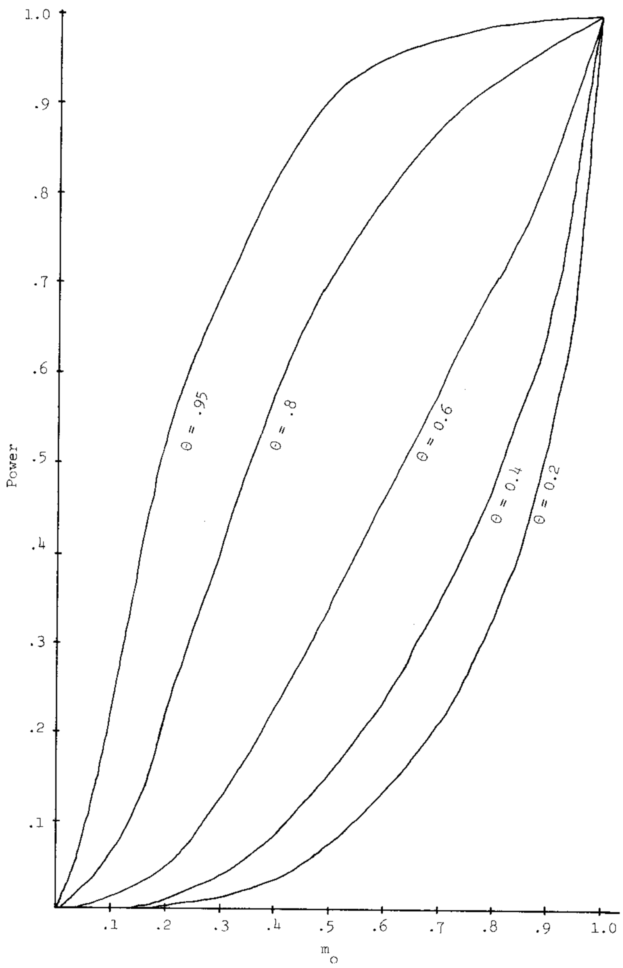
[8] Mehta, J.S. and Swamy, P.A.V.B. (1973). "Bayesian Analysis of a Bivariate Normal Distribution with Incomplete Observations," Journal of the American Statistical Association, 68, 922-27.

[9] Sukhatme, P.V. and Sukhatme, B.V. (1970). "Sampling Theory of Surveys with Applications," Ames, Iowa, Iowa State University Press.

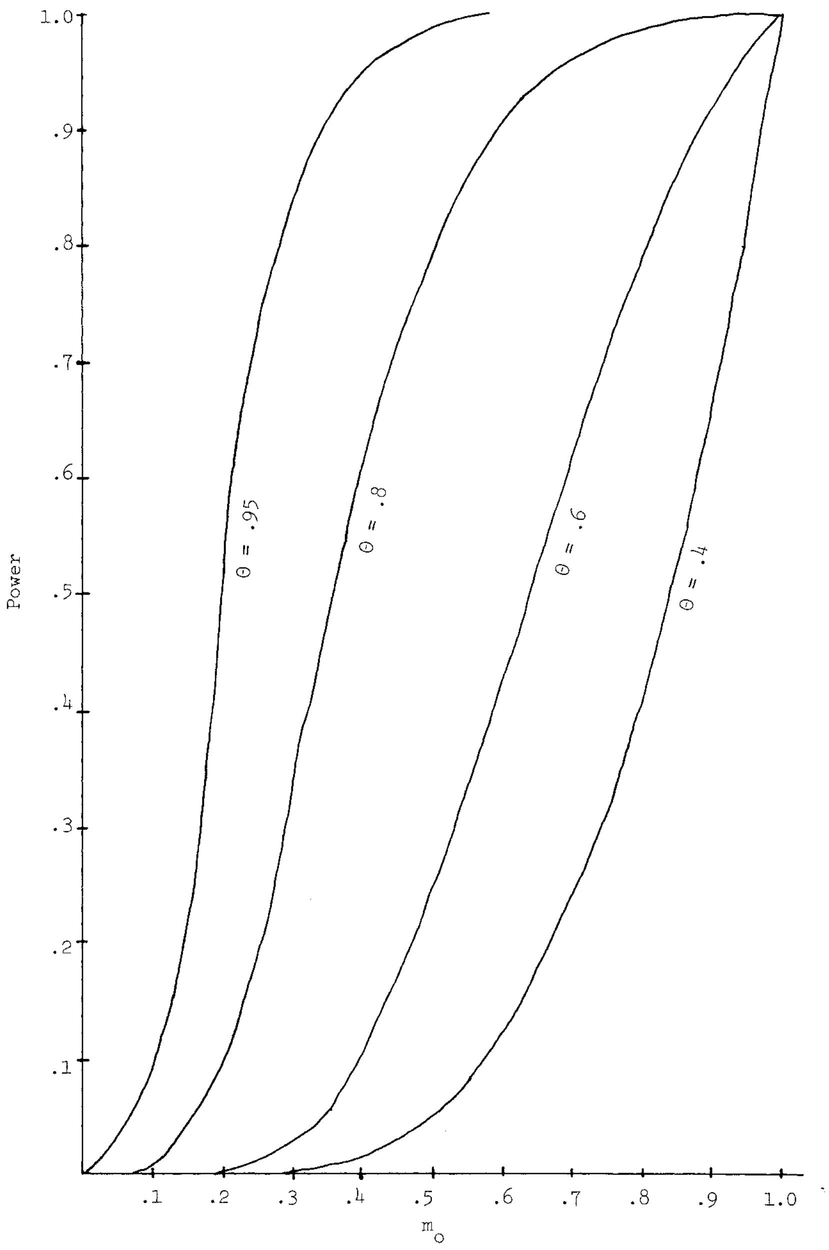
[10] Watson, D.J. (1937). "The Estimation of Leaf Areas," Journal of Agricultural Science, 27, 474.

[11] Yates, E. (1960). "Sampling Methods for Censuses and Surveys." Third Edition, Charles Griffin and Co., London.

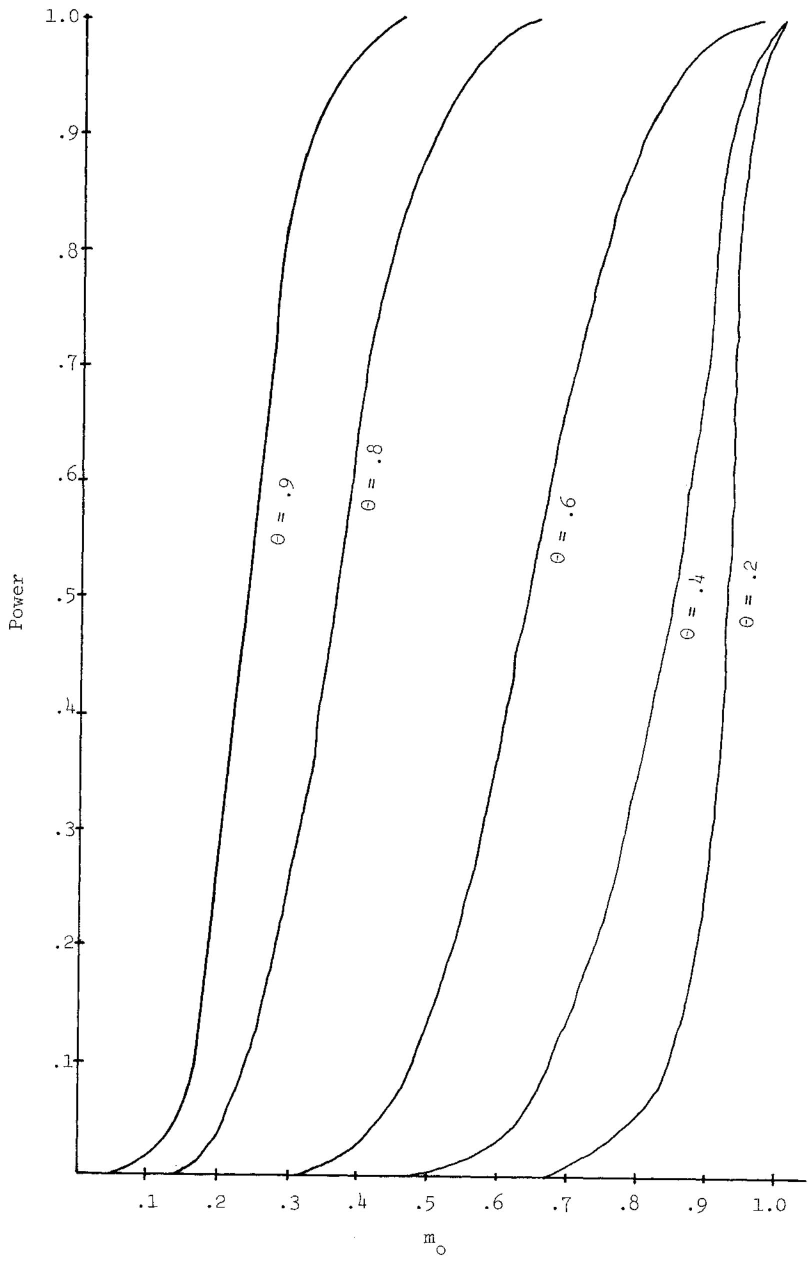
A. POWER FUNCTION FOR N = 7



B. POWER FUNCTION FOR N = 15



C. POWER FUNCTION FOR N = 35


Power
θ = .9
θ = .8
θ = .6
θ = .4
θ = .2
m_o