Journal of Educational Statistics
Summer 1978, Volume 3, Number 2, Pp. 179-188

BAYESIAN COMPARISON OF TWO REGRESSION LINES

Robert K. Tsutakawa

University of Missouri-Columbia

Key words: Comparison of Regression Lines; Finite Interval; Bivariate t Distribution

ABSTRACT

The comparison of two regression lines is often meaningful or of interest over a finite interval I of the independent variable. When the prior distribution of the parameters is a natural conjugate, the posterior distribution of the distances between two regression lines at the end points of I is bivariate t. The posterior probability that one regression line lies above the other uniformly over I is numerically evaluated using this distribution.

INTRODUCTION

The comparison of two or more regression lines is a familiar problem to most practicing statisticians. In textbooks (e.g. Brownlee (1965) and Neter and Wasserman (1974)), it is usually discussed in terms of testing the equality of regression coefficients. In many problems where such comparisons are made, the important question is not whether the coefficients are identical, but rather whether one line lies above another over some finite interval $I = (x_*, x^*)$ of the independent variable.

In the education literature, problems of this nature are often called the Johnson-Neyman problem. The Johnson-Neyman method modified by Aitkin (1973) is based on a confidence band of the Working-Hotelling type and divides I into regions of

significant and nonsignificant differences. It may be noted that there is a certain awkwardness in the interpretation of results based on this method since a pair of lines in a plane cannot coincide over one interval while being distinct over another. Tsutakawa and Hewett (1978) present an alternative approach which tests the hypothesis that the lines intersect over I against the hypothesis that they do not and relate their method to Aitkin's in terms of the difference in the hypotheses being tested. A nonparametric test for problems related to comparing regression lines has also been proposed and is illustrated with several applications by Tsutakawa and Hewett (1977).

We note that the practicing statistician who routinely informs a client that his regression lines are significantly different is often asked, "How much assurance is there that one lies above the other?" The scene is a familiar one, often reflecting a certain amount of confusion.

The present paper departs from hypothesis testing and gives a Bayesian answer to the experimenter's question. Although the result reported here follows directly from well known Bayesian distribution theory, [see, for example, Box and Tiao (1973)], its application does not appear to be mentioned in the literature. Using the locally uniform prior, we will show that the joint posterior distribution of the (signed) distance between two regression lines at x_* and x^* is bivariate t and give a simple numerical method for evaluating the posterior probability that one regression line lies above another for all points x in I. Data from Brownlee (1965) on the effect of drugs on mental addition will be used to illustrate the technique.

2. ANALYSIS

Given two independent sets of bivariate observations $(X_{11},Y_{11}), \ldots, (X_{1n_1},Y_{1n_1})$ and $(X_{21},Y_{21}), \ldots, (X_{2n_2},Y_{2n_2})$ we wish to discuss the relative position of the regression of Y on X for one set to that of the other over an arbitrary finite interval $I = (x_*, x^*)$. Given the X values, assume that the Y values are independent and normally distributed with means and variances,

$$E(Y_{ij} \mid x_{ij}) = \alpha_i + \beta_i x_{ij} ,$$

$$\mathrm{Var}(Y_{ij} \mid x_{ij}) = \sigma^2 ,$$

where (α_i,β_i), $i = 1, 2,$ and $\sigma^2 > 0$ are unknown parameters.

The relative position of the two regression lines over I may be expressed in terms of the distance between the lines at x_* and x^*, given by

$$\underset{\sim}{\theta} = (\theta_1, \theta_2)' = \underset{\sim}{D}(\alpha_1, \beta_1, \alpha_2, \beta_2)'$$

where $'$ denotes transpose and

$$\underset{\sim}{D} = \begin{pmatrix} 1 & x_* & -1 & -x_* \\ 1 & x^* & -1 & -x^* \end{pmatrix} .$$

Inference about the relative position of the regression lines may now be made in terms of the posterior distribution of $\underset{\sim}{\theta}$. In particular, the posterior probability that line 1 is uniformly above line 2 over I is the posterior probability that $\underset{\sim}{\theta}$ belongs to the first quadrant, i.e., $\theta_1 > 0$ and $\theta_2 > 0$. Similarly the probability that line 2 is uniformly above line 1 over I is the probability associated with the third quadrant and the probability that the lines intersect over I is the sum of the probabilities associated with the second and fourth quadrants.

Let $\underset{\sim}{\hat{\mu}}' = (a_1,b_1,a_2,b_2)$ denote the least squares estimate of $\underset{\sim}{\mu}' = (\alpha_1,\beta_1,\alpha_2,\beta_2)$, s^2 the pooled variance estimate of σ^2,

$$s^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij}-a_i-b_i x_{ij})^2/\nu ,$$

where $\nu = n_1+n_2-4$, and $\underset{\sim}{X}$ the $(n_1+n_2)\times 4$ design matrix

$$\underset{\sim}{X} = \begin{bmatrix} 1 & x_{11} & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_{1n_1} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 1 & x_{2n_2} \end{bmatrix} .$$

Under the locally uniform prior, i.e.,

$$p(\underset{\sim}{\mu}', \sigma) \propto 1/\sigma, \tag{1}$$

the posterior distribution of $\underset{\sim}{\mu}$ is multivariate t with pdf

$$p(\underset{\sim}{\mu}'|\underset{\sim}{X},\underset{\sim}{Y}) \propto [1 + (\underset{\sim}{\mu}-\underset{\sim}{\hat{\mu}})'\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{\mu}-\underset{\sim}{\hat{\mu}})/\nu s^2]^{-(\nu+4)/2} . \tag{2}$$

Moreover, since $\underset{\sim}{D}$ is a 2×4 matrix of rank 2, it follows from page 118 of Box and Tiao (1973) that the posterior distribution of $\underset{\sim}{\theta}' = \underset{\sim}{D}\underset{\sim}{\mu}'$ is bivariate t with pdf,

$$p(\underset{\sim}{\theta}'|\underset{\sim}{X},\underset{\sim}{Y}) \propto [1 + (\underset{\sim}{\theta}-\underset{\sim}{\hat{\theta}})'\underset{\sim}{C}^{-1}(\underset{\sim}{\theta}-\underset{\sim}{\hat{\theta}})/\nu s^2]^{-(\nu+2)/2} , \tag{3}$$

where $\underset{\sim}{\hat{\theta}}' = \underset{\sim}{D}\underset{\sim}{\hat{\mu}}'$ and $\underset{\sim}{C} = \underset{\sim}{D}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{D}'$. (It may be noted that $\underset{\sim}{\hat{\theta}}$ is the maximum likelihood estimator of $\underset{\sim}{\theta}$ and $\sigma^2\underset{\sim}{C}$ its covariance matrix.)

Although the posterior distributions of $\underset{\sim}{\mu}$ or $\underset{\sim}{\theta}$ may not be of direct use in practice, we can derive from them a number of useful summaries depending on the user's interest. If he wishes to compare the regression lines over a particular interval I or a set of intervals, then a useful summary would be the set of posterior probabilities associated with the four quadrants of the space of $\underset{\sim}{\theta}$ for such intervals. The evaluation of the cdf of $\underset{\sim}{\theta}$ and these quadrant probabilities requires numerical integration and will be discussed in Section 4.

3. EXAMPLE

We will use the data on the first two of four drugs which appear in Brownlee (1965, p.395). For each drug, 14 subjects were scored on mental addition before and after the use of the drug. (See Fig.1,p.187). We will compare the two drugs in terms of the regression of the score after the drug (Y) on the score before the drug (X) by using I = (18,52), the range of the X values actually observed. The numerical values of the relevant statistics are given as follows.

i	n_i	a_i	b_i	$\hat{\theta}_i$
1	14	2.80	1.01	1.07
2	14	5.44	0.80	8.10

$s^2 = 15.9539,$
$\nu = 24,$

$$\underset{\sim}{D} = \begin{bmatrix} 1 & 18 & -1 & -18 \\ 1 & 52 & -1 & -52 \end{bmatrix},$$

$$\underset{\sim}{X}'\underset{\sim}{X} = \begin{bmatrix} n_1 & \sum x_{ij} & 0 & 0 \\ \sum x_{ij} & \sum x_{ij}^2 & 0 & 0 \\ 0 & 0 & n_2 & \sum x_{2j} \\ 0 & 0 & \sum x_{2j} & \sum x_{2j}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 14 & 455 & 0 & 0 \\ 455 & 16249 & 0 & 0 \\ 0 & 0 & 14 & 490 \\ 0 & 0 & 490 & 18008 \end{bmatrix},$$

$$\underset{\sim}{C} = \begin{bmatrix} .62355 & -.38744 \\ -.38744 & .73986 \end{bmatrix}.$$

Using these values and the integration method described in the next section, the probability that $\theta_1 > 0$ and $\theta_2 > 0$ turns out to be .619. This is the posterior probability that the regression line for the first drug lies above that for the second drug over the interval $I = (18,52)$. The corresponding probabilities that $\underset{\sim}{\theta}$ belongs to the 2nd, 3rd, and 4th quadrants are .368, .000, and .013, respectively. Thus there is about a 38% chance that the lines intersect over this interval.

Since I is arbitrary, extrapolations are possible. For example, for $I = (0,18)$ the posterior probabilities of the four quadrants are, in order, .326, .310, .314, and .000. However, extrapolations of this type could be misleading since there is little reason to believe that our linear model is valid over this lower region, particularly if the scores must be positive.

4. NUMERICAL METHOD

The numerical integration technique for the bivariate t distribution defined by (3) which we will now describe is a variation of the method used by Krishnaiah and Armitage (1966). When extreme accuracy is not required, the method described is simple to use for evaluating the cdf and quadrant

probabilities.

Let c_{ij}, i =1, 2, denote the ijth element of $\underset{\sim}{C}$, defined in Section 2, and $\rho = c_{12}/(c_{11}c_{22})^{\frac{1}{2}}$. Then the random vector (θ_1,θ_2) may be represented by

$$\theta_i = \hat{\theta}_i + Z_i s(\nu c_{ii})^{\frac{1}{2}}/U^{\frac{1}{2}}, \quad i = 1, 2$$

where (Z_1,Z_2) is standard bivariate normal with correlation coefficient ρ, $-1 < \rho < 1$, and U is chi-square with ν df independent of (Z_1,Z_2). Following Dunnett and Sobel (1955), let $c_1 = |\rho|^{\frac{1}{2}}$ and $c_2 = c_1$ or $-c_1$ according as $\rho \geq 0$ or $\rho < 0$. Then we may represent (Z_1,Z_2) by

$$Z_i = (1-c_i^2)^{\frac{1}{2}}Y_i - c_i Y_o \ , \ i = 1, 2 \ ,$$

where Y_o, Y_1, Y_2 are independent N(0,1) and independent of U. By conditioning on Y_o and U, the posterior cdf of $\underset{\sim}{\theta}$ may be expressed as the expectation of a conditional probability and put in the form,

$$F(k_1,k_2 \mid \underset{\sim}{X}, \underset{\sim}{Y}, I)$$

$$= \int_0^\infty e^{-w} f(w) \left\{ \int_{-\infty}^{\infty} e^{-t^2} \prod_{i=1}^{2} \Phi(\delta_i(k_i,w,t))dt \right\} dw \ , \qquad (4)$$

where

$$f(w) = w^{(\nu-2)/2}\{\pi^{\frac{1}{2}}\ \Gamma(\nu/2)\}^{-1} \ ,$$

$$\delta_i(k,w,t) = \sqrt{2}\ \{c_i t + (k-\hat{\theta}_i)w/(\nu c_{ii}s^2)^{\frac{1}{2}}\}/(1-c_i^2)^{\frac{1}{2}} \ ,$$

and Φ is the standard normal cdf. In particular, $F(0,0|\underset{\sim}{X},\underset{\sim}{Y},I)$ is the posterior probability that $\underset{\sim}{\theta}$ belongs to the 3rd quadrant, where $\theta_1 < 0$ and $\theta_2 < 0$. The probabilities for the lst, 2nd, and 4th quadrants are obtained by replacing $\prod_{i=1}^{2} \Phi(\delta_i(k_i,w,t))$ in (4) by $\prod_{i=1}^{2} \{1-\Phi(\delta_i(0,w,t))\}$, $\Phi(\delta_1(0,w,t))\{1-\Phi(\delta_2(0,w,t))\}$ and $\{1-\Phi(\delta_1(0,w,t))\}\cdot\Phi(\delta_2(0,w,t))$, respectively. (4) is now in form suitable for numerical integration combining the Gauss-Hermite with Gauss-Laguerre methods. More specifically, we may approximate (4) by a weighted average of the form

$$\sum_{i=1}^{p} \sum_{j=1}^{q} s_i u_j g(t_i, w_j)$$

where

$$g(t,w) = f(w) \prod_{i=1}^{2} \Phi(\delta_i(k_i, w, t))$$

and $\{(t_i, s_i), i = 1, \ldots, p\}$ and $\{(w_j, u_j), j = 1, \ldots, q\}$ are the sets of abscissas and weight factors for the order p Gauss-Hermite and order q Gauss-Laguerre intergrations, respectively. Readers who do not have ready access to computer subroutines for these quadrature methods may wish to use the selected sets of these constants in Abramowitz and Stegun (1964).

5. MISCELLANEOUS REMARKS

a. The reader is warned that the present method may produce misleading results when the assumptions of Section 2 are not satisfied. In particular the method is potentially inapplicable when the variances of the two populations are different.

b. The posterior probability that one regression line is at least a given distance above the other over I may be computed with a simple modification. In particular, for any k, $-\infty < k < \infty$, $P(\theta_1 > k, \theta_2 > k | \underset{\sim}{X}, \underset{\sim}{Y}, I)$ may be evaluated by replacing $\prod_{i=1}^{2} \Phi(\delta_i(k_i, w, t))$ in (4) by $\prod_{i=1}^{2} \{1-\Phi(\delta(k,w,t))\}$.

c. The locally uniform prior may be replaced by appropriate natural conjugate priors without changes in the computational techniques. In particular, suppose that the prior distribution of $(h, \underset{\sim}{\mu}')$, where $h = \sigma^{-2}$, is normal-gamma, defined by the pdf,

$$p(h, \underset{\sim}{\mu}') \propto h^{\frac{\nu'}{2}-1} \exp(-h\nu_1 v_1/2) \cdot |\underset{\sim}{\Phi}_1|^{-\frac{1}{2}} \exp\{-h(\underset{\sim}{\mu}-\underset{\sim}{\mu}_1)'\underset{\sim}{\Phi}_1^{-1}(\underset{\sim}{\mu}-\underset{\sim}{\mu})/2\},$$

where ν_1, v_1, $\underset{\sim}{\mu}_1$, $\underset{\sim}{\Phi}_1$ are prior parameters, ν_1, $v_1 > 0$ and $\underset{\sim}{\Phi}_1$ is a 2x2 nonsingular covariance matrix. Then the posterior distribution of $(h, \underset{\sim}{\mu}')$ is again a normal-gamma and the marginal distribution of $\underset{\sim}{\mu}'$ is multivariate t with pdf,

$$p(\underset{\sim}{\mu}' | \underset{\sim}{X}, \underset{\sim}{Y}) \propto \{1 + (\underset{\sim}{\mu}-\underset{\sim}{\mu}_2)'(\underset{\sim}{\Sigma}_2^{-1}/\nu_2 v_2)(\underset{\sim}{\mu}-\underset{\sim}{\mu})\}^{-\frac{(\nu_2+4)}{2}}, \quad (5)$$

where

$$\underset{\sim}{\Sigma}_2 = (\underset{\sim}{X}'\underset{\sim}{X} + \underset{\sim}{\Sigma}_1^{-1})^{-1}, \quad \nu_2 = \nu_1 + \nu + 4,$$

$$\underset{\sim}{\mu}_2 = \underset{\sim}{\Sigma}_2(\underset{\sim}{\Sigma}_1^{-1}\underset{\sim}{\mu}_1 + \underset{\sim}{X}'\underset{\sim}{X}\,\hat{\underset{\sim}{\mu}}),$$

$$v_2 = \nu_1 v_1 + \underset{\sim}{\mu}'\underset{\sim}{\Sigma}_1^{-1}\underset{\sim}{\mu}_1 + \nu v + \hat{\underset{\sim}{\mu}}'\underset{\sim}{X}'\underset{\sim}{X}\,\hat{\underset{\sim}{\mu}} - \underset{\sim}{\mu}_2'\underset{\sim}{\Sigma}_2^{-1}\underset{\sim}{\mu}_2 .$$

(For derivation see Raiffa and Schlaifer (1961, p 343).)

As in Section 3, it follows that the posterior distribution of $\underset{\sim}{\theta}'$ is bivariate t with pdf,

$$p(\underset{\sim}{\theta}' | \underset{\sim}{X}, \underset{\sim}{Y}) \propto [1 + (\underset{\sim}{\theta}-\underset{\sim}{\theta}_2)'\underset{\sim}{C}_2^{-1}(\underset{\sim}{\theta}-\underset{\sim}{\theta}_2)/\nu_2 v_2]^{-\frac{(\nu_2+2)}{2}} \quad (6)$$

where $\underset{\sim}{C}_2 = \underset{\sim}{D}(\underset{\sim}{\Sigma}_2)\underset{\sim}{D}'$ and $\underset{\sim}{\theta}_2' = \underset{\sim}{D}\underset{\sim}{\mu}_2'$. Thus the numerical evaluation may be carried out as in the case of the noninformative prior by replacing $(\underset{\sim}{C}, \hat{\underset{\sim}{\theta}}, \nu, s^2)$ in (3) with $(\underset{\sim}{C}_2, \underset{\sim}{\theta}_2, \nu_2, v_2)$.

It may be noted that, as with the use of other natural conjugate priors, the selection of a particular prior is facilitated if it is based on a previous similar experiment, in which case the likelihood function may be formally identified with a normal-gamma distribution.

d. When there are several regression lines, $\ell_1, \ldots, \ell_{k+1}$, $k \geq 2$, a variety of comparisons is possible and the numerical work becomes more involved. For example, consider the posterior probability that, say, ℓ_{k+1} lies above $\ell_1, \ldots, \ell_k$ over $I = (x_*, x^*)$. This problem may be solved by noting that the sets of distances $(\theta_{11}, \ldots, \theta_{1k})$ and $(\theta_{21}, \ldots, \theta_{2k})$ between ℓ_{k+1} and $\ell_1, \ldots \ell_k$ at x_* and x^* have a multivariate t distribution with a straightforward modification in $\underset{\sim}{X}$, $\underset{\sim}{C}$, s^2, and ν. The required orthant probability cannot be evaluated by the method used here, but may be computed by a procedure such as the one suggested by Dutt (1975).

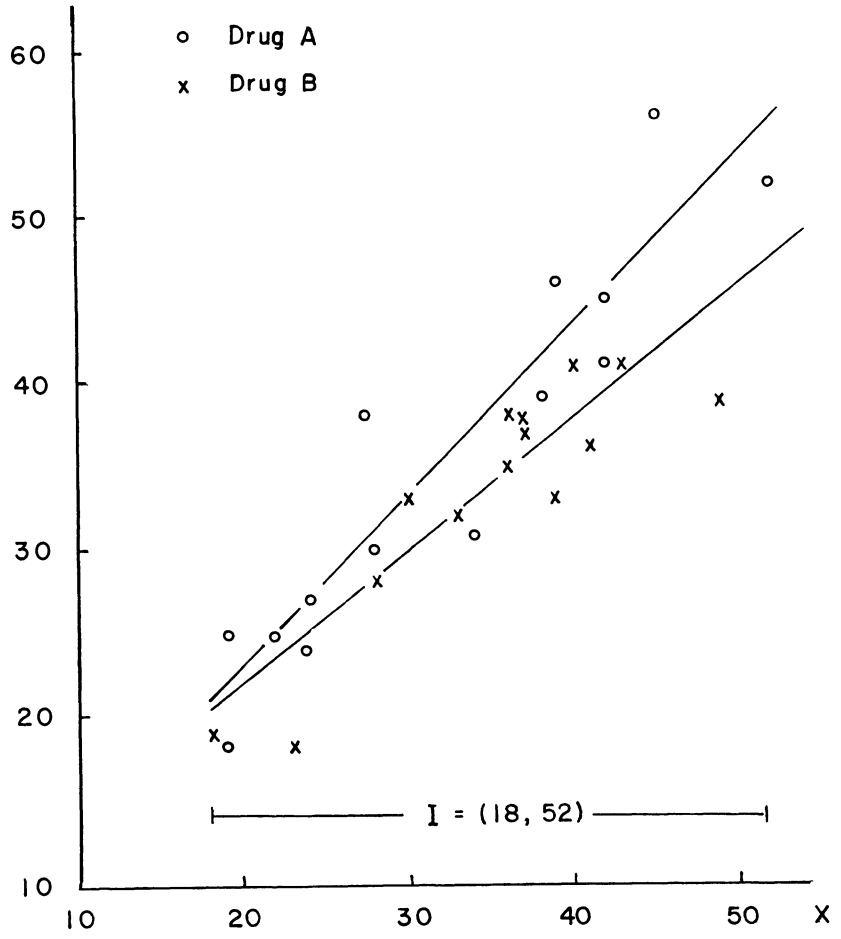

Figure 1

Mental Addition Scores Before (X) and After (Y) Drug.

ACKNOWLEDGMENTS

This research was supported by PHS Grant No. R01 GM23548 from the National Institute of General Medical Sciences. The author is indebted to John E. Hewett for discussions on comparing two regression lines.

REFERENCES

Abramowitz, M. & Setgun, I. A. Handbook of mathematical functions. National Bureau of Standards Applied Mathematics Series 55, 1964.

Box, G. E. P. & Tiao, G. C. Bayesian inference in statistical analysis. Reading, Massachusetts: Addison-Wesley Publishing Company, 1973.

Brownlee, K. A. Statistical theory and methodology in science and engineering (2nd edition). New York: John Wiley and Sons, Inc., 1965.

Dunnett, C. W. & Sobel, M. Approximations to the probability integral and certain percentage points of a multivariate analogue of Student's t-distribution. Biometrika, 1955, 42, 258-260.

Dutt, J. E. On computing the probability integral of a generalized multivariate t. Biometrika, 62, 201-205.

Krishnaiah, P. R. & Armitage, J. V. Tables for multivariate t distributions. Sankhya, Series B. 1966, 28, 31-56.

Neter, J. & Wasserman, W. Applied linear statistical models. Homewood, Illinois: Richard D. Irwin, Inc., 1974.

Raiffa, H. & Schlaifer, R. Applied statistical decision theory. Boston: Graduate School of Business Administration, Harvard, 1961.

Tsutakawa, R. K. & Hewett, J. E. Quick test for comparing two populations with bivariate data. Biometrics, 1977, 33, 215-219.

Tsutakawa, R. K. & Hewett, J. E. Comparison of two regression lines over a finite interval. In preparation, 1978.

AUTHOR

TSUTAKAWA, ROBERT K. Address: Department of Statistics, University of Missouri-Columbia, Columbia, MO 65201. Title: Associate Professor of Statistics. Degrees: B. S., M.S., Ph.D., University of Chicago. Specialization: Statistical Inference, Bioassay.

[*Manuscript received April 1977; revised February 1978.*]