Karel Zvára

Confidence band for regression line with exponential distribution of errors

Kybernetika, Vol. 31 (1995), No. 1, 31--44

Persistent URL: http://dml.cz/dmlcz/125187

KYBERNETIKA — VOLUME 31 (1995), NUMBER 1, PAGES 31–44

CONFIDENCE BAND FOR REGRESSION LINE WITH EXPONENTIAL DISTRIBUTION OF ERRORS

KAREL ZVÁRA

This paper deals with the maximum likelihood estimation of a regression line parameters in the model with an exponential distribution of errors. A testing procedure of simple hypothesis is used for standard derivation of a confidence band for this regression line. Some quality features are given by means of a simulation experiment.

1. INTRODUCTION

In some biological applications statisticians should look for a "boundary line" that sepa ates real from non-real situations. For example Nátrová and Nátr [2], Walworth [5] search for the dependence of maximal possible grain yield on the size of a chosen growth factor. A model for this situation can be given by

$$y_i = \beta_0 + \beta_1 x_i + \theta e_i, \quad 1 \leq i \leq n, \tag{1.1}$$

where β_0, β_1 and θ are unknown parameters. Random errors $e_1, \ldots, e_n$ are independent, identically and continuously distributed with $\mathsf{P}[e_1 \geq 0] = 1$. Real points (x_i, y_i) lie only in one of the half-planes determined by a true regression line. When they lie over this line, we suppose that $\theta > 0$. In the opposite case ($\theta < 0$) we can use the transformations

$$\tilde{y}_i = -y_i, \ \tilde{x}_i = x_i, \quad 1 \leq i \leq n,$$
$$\tilde{\beta}_0 = -\beta_0, \ \tilde{\beta}_1 = -\beta_1, \ \tilde{\theta} = -\theta,$$

which convert this case to the first one. Therefore, we will suppose that $\theta > 0$. Moreover, we will suppose that random variable e_1 follows the exponential distribution with expectation one.

The last assumption is

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0. \tag{1.2}$$

2. MAXIMUM LIKELIHOOD ESTIMATE OF PARAMETERS

Let us denote

$$M_n = \left\{ \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathsf{R}^2 : \beta_0 + \beta_1 x_i \leq y_i, \quad 1 \leq i \leq n \right\}.$$

The joint density of the random vector $\boldsymbol{y} = (y_1, \ldots, y_n)'$ can be written in the form

$$f(\boldsymbol{y}) = \begin{cases} \exp\left(-n\left(\ln\theta + \dfrac{\bar{y} - \beta_0 - \beta_1 \bar{x}}{\theta}\right)\right) & \text{if } \binom{\beta_0}{\beta_1} \in M_n, \theta > 0, \\ 0 & \text{otherwise.} \end{cases}$$

To find the maximum likelihood estimate of β_0 and β_1 we try to solve the linear programming problem

$$\text{maximize } (\beta_0 + \beta_1 \bar{x}) \tag{2.1}$$

over β_0, β_1 satisfying conditions

$$\beta_0 + \beta_1 x_i \leq y_i, \quad 1 \leq i \leq n. \tag{2.2}$$

Let us denote by b_0^A, b_1^A any solution of the linear programming problem (2.1),(2.2). Moreover, let us denote

$$t^A = \bar{y} - b_0^A - b_1^A \bar{x}$$

and

$$\ell(\beta_0, \beta_1, \theta) \equiv \ln f(\boldsymbol{y}).$$

Logarithm of the likelihood function $\ell(\beta_0, \beta_1, \theta)$ can be written as

$$\ell(\beta_0, \beta_1, \theta) = \begin{cases} -n\left(\ln\theta + \dfrac{\bar{y} - \beta_0 - \beta_1 \bar{x}}{\theta}\right) & \text{for } \binom{\beta_0}{\beta_1} \in M_n, \theta > 0, \\ -\infty & \text{otherwise.} \end{cases}$$

Theorem 1. The estimates b_0^A, b_1^A, t^A are maximum likelihood estimates of β_0, β_1, θ.

Proof. We should prove that for all $(\beta_0, \beta_1)' \in M_n, \theta > 0$ it is true that $\ell(\beta_0, \beta_1, \theta) \leq \ell(b_0^A, b_1^A, t^A)$, or equivalently

$$-\ln\theta - \frac{1}{\theta}(\bar{y} - \beta_0 - \beta_1 \bar{x}) \leq -\ln t^A - 1.$$

The last inequality is the same as

$$-\ln\left(\frac{\bar{y} - b_0^A - b_1^A \bar{x}}{\theta}\right) - 1 + \frac{\bar{y} - \beta_0 - \beta_1 \bar{x}}{\theta} \geq 0. \tag{2.3}$$

Because of $(b_0^A, b_1^A)'$ is a solution of linear programming problem (2.1),(2.2) and $\theta > 0$ for all $(\beta_0, \beta_1)' \in M_n$, it is true that

$$\frac{\bar{y} - \beta_0 - \beta_1 \bar{x}}{\theta} \geq \frac{\bar{y} - b_0^A - b_1^A \bar{x}}{\theta}.$$

Therefore, the left side of (2.3) can be estimated from below by the term

$$-\ln\left(\frac{\bar{y}-b_0^A-b_1^A\bar{x}}{\theta}\right)-1+\frac{\bar{y}-b_0^A-b_1^A\bar{x}}{\theta}.$$

This term is nonnegative for every $u=(\bar{y}-b_0^A-b_1^A\bar{x})/\theta>0$ (e. g. (8a.5.9) in [3]). □

It can be seen that the maximal value of $\ell(\beta_0,\beta_1,\theta)$ on $M_n\times\mathsf{R}^+$ is equal to

$$\ell(b_0^A,b_1^A,t^A)=-n\left(\ln t^A+1\right). \tag{2.4}$$

Now we will study some properties of the maximum likelihood estimate (b_0^A,b_1^A,t^A). We cannot use standard asymptotic properties of this estimate, because some of assumptions are not fulfilled. Especially, the set

$$\{\boldsymbol{y}\in\mathsf{R}^n : f(\boldsymbol{y})>0\}$$

depends on the unknown parameter $(\beta_0,\beta_1)'$. The consistency of (b_0^A,b_1^A) was proved in [8] under the mild assumptions on asymptotic properties of $\{x_i\}$:

- the existence of a finite limit point of $\{x_i\}_{i=1}^{\infty}$;
- $x_L=\liminf x_i<\limsup x_i=x_U$;
- at least one of x_L,x_U is finite;
- the existence of $\lim\limits_{n\to\infty}\overline{x}_n=x_0$ and $x_0\in(x_L,x_U)$.

3. THE SIMPLE HYPOTHESIS TESTING

Let us consider testing of the hypothesis

$$\mathsf{H}_0:\beta_0=\beta_0^0,\ \beta_1=\beta_1^0,$$

where β_0^0,β_1^0 are given real numbers. A critical region of the classical likelihood ratio test is given by inequality

$$-2\left(\ell(\beta_0^0,\beta_1^0,t^0)-\ell(b_0^A,b_1^A,t^A)\right)>c,$$

where

$$t^0=\bar{y}-\beta_0^0-\beta_1^0\bar{x}.$$

Common requirement for the first kind error to be bounded by a predetermined significance level α determines value of the constant c. After the substitution by (2.4) we will come to an equivalent inequality for the critical region

$$\frac{t^A}{t^0}<\exp\left(-\frac{c}{2n}\right)\equiv c^A. \tag{3.1}$$

In order to find the constant c^A (therefore the constant c, too), we should be able to compute the probability of (3.1) in case when the hypothesis H_0 is true. May be,

it would be sufficient to bound this probability from above. After the substitution for t^A, t^0 in (3.1) we can sequentially modify the left side:

$$\begin{aligned}\frac{t^A}{t^0} &= \frac{\bar{y} - b_0^A - b_1^A\bar{x}}{\bar{y} - \beta_0^0 - \beta_1^0\bar{x}} \\ &= \frac{\beta_0^0 + \beta_1^0\bar{x} + \theta\bar{e} - b_0^A - b_1^A\bar{x}}{\beta_0^0 + \beta_1^0\bar{x} + \theta\bar{e} - \beta_0^0 - \beta_1^0\bar{x}} \\ &= \frac{\sum e_\nu - n(b_0^A - \beta_0^0)/\theta - n((b_1^A - \beta_1^0)/\theta)\bar{x}}{\sum e_\nu}.\end{aligned}$$

Therefore, the inequality (3.1) is equivalent to

$$\frac{b_0^A - \beta_0^0}{\theta\sum e_\nu} + \frac{b_1^A - \beta_1^0}{\theta\sum e_\nu}\bar{x} > \frac{1}{n}\left(1 - c^A\right) \equiv g. \tag{3.2}$$

If the hypothesis H_0 is true then original inequalities (2.2) are equivalent to

$$\beta_0 + \beta_1 x_i \le y_i = \beta_0^0 + \beta_1^0 x_i + \theta e_i, \quad 1 \le i \le n,$$

which are the same as

$$\frac{\beta_0 - \beta_0^0}{\theta\sum e_\nu} + \frac{\beta_1 - \beta_1^0}{\theta\sum e_\nu}x_i \le \frac{e_i}{\sum e_\nu}, \quad 1 \le i \le n.$$

Since the function maximized in (2.1) can be rewritten as

$$\beta_0 + \beta_1\bar{x} = \left(\frac{\beta_0 - \beta_0^0}{\theta\sum e_\nu} + \frac{\beta_1 - \beta_1^0}{\theta\sum e_\nu}\bar{x}\right)\theta\sum e_\nu + \left(\beta_0^0 + \beta_1^0\bar{x}\right),$$

the maximization of (2.1) over β_0, β_1 is equivalent to

$$\text{maximize } \left(\frac{\beta_0 - \beta_0^0}{\theta\sum e_\nu} + \frac{\beta_1 - \beta_1^0}{\theta\sum e_\nu}\bar{x}\right).$$

If we define

$$\delta_0 = \frac{\beta_0 - \beta_0^0}{\theta\sum e_\nu}, \ \delta_1 = \frac{\beta_1 - \beta_1^0}{\theta\sum e_\nu},$$

we can rewrite the original linear programming problem to a new problem

$$\text{maximize } (\delta_0 + \delta_1\bar{x}) \tag{3.3}$$

for δ_0, δ_1 satisfying the conditions

$$\delta_0 + \delta_1 x_i \le \frac{e_i}{\sum e_\nu} \equiv \varepsilon_i, \quad 1 \le i \le n. \tag{3.4}$$

These considerations lead to the next theorem.

Theorem 2. The expressions

$$d_0 = \frac{b_0^A - \beta_0^0}{\theta \sum e_\nu}, \; d_1 = \frac{b_1^A - \beta_1^0}{\theta \sum e_\nu}$$

are solutions of the linear programming problem (3.3),(3.4) if and only if b_0^A, b_1^A are solutions of the linear programming problem (2.1),(2.2).

The main advantage of the (3.3), (3.4) over (2.1), (2.2) is independence of its solution on the concrete values of β_0^0, β_1^0. Instead of probability of (3.2) it suffices to compute probability of the equivalent random event

$$d_0^A + d_1^A \bar{x} > g. \tag{3.5}$$

Let us consider the geometric meaning of the modified problem. The inequality (3.4) is fulfilled if and only if any of the points $(x_i, \varepsilon_i), 1 \leq i \leq n$, lies above the line $z = \delta_0 + \delta_1 x$. The maximization of (3.3) means to select the line, which value $\delta_0 + \delta_1 \bar{x}$ is maximal. It is the same situation as in Introduction, but with random variables $\varepsilon_1, \ldots, \varepsilon_n$ instead of $y_1, \ldots, y_n$.

Distribution of the random vector $\varepsilon_1, \ldots, \varepsilon_{n-1}$ is investigated in the Appendix. This vector has the Dirichlet distribution $\mathrm{D}(1, \ldots, 1; 1)$ with a constant density on the simplex $\left\{ \boldsymbol{z} \in \mathsf{R}^{n-1} : z_1 \geq 0, \ldots, z_{n-1} \geq 0, \sum z_i \leq 1 \right\}$. Moreover, the next theorem is proved in A5 of the Appendix.

Theorem 3. Suppose that for given integer p, $(1 \leq p < n)$ and for given real numbers $z_1^0 \geq 0, \ldots, z_p^0 \geq 0$ it is true that

$$z^0 = 1 - \sum_{j=1}^{p} z_j^0 \geq 0.$$

Then for every $z_{p+1} \geq 0, \ldots, z_n \geq 0$ such that $\sum_{i=p+1}^{n} z_i \leq z^0$ it is true that

$$\mathsf{P}\left[\varepsilon_{p+1} \geq z_{p+1}, \ldots, \varepsilon_n \geq z_n | \varepsilon_1 = z_1^0, \ldots, \varepsilon_p = z_p^0\right] = \left(1 - \frac{1}{z^0} \sum_{i=p+1}^{n} z_i\right)^{n-p-1}.$$

We can suppose that

$$x_1 \leq x_2 \leq \ldots \leq x_p \leq \bar{x} < x_{p+1} \leq \ldots \leq x_n.$$

Now, we will show that probability of inequality (3.5) can be bounded from above by probability of a similar inequality with only two distinct values of x_i.

Let us consider conditional probability

$$\mathsf{P}\left[d_0^A + d_1^A \bar{x} \geq g | \varepsilon_1 = z_1^0, \ldots, \varepsilon_p = z_p^0\right], \tag{3.6}$$

where $z_1^0,\ldots,z_p^0$ fulfil the conditions of Theorem 3. Considered random event occurs if and only if any of the points $(x_{p+1},\varepsilon_{p+1}),\ldots,(x_n,\varepsilon_n)$ does not lie under the line

$$\varepsilon = g + \gamma(x - \bar{x}) \equiv g' + \gamma x,$$

which is the closest to the points $(x_1,\varepsilon_1),\ldots,(x_p,\varepsilon_p)$ from below and which is going through the point $(\bar{x},g)$. Its slope γ is a function of $g, z_1^0,\ldots,z_p^0,\bar{x},x_1,\ldots,x_p$. Considering that for some of values $g' + \gamma x_i, p+1 \le i \le n$ can be negative, probability (3.6) is the same as

$$\mathsf{P}\left[\varepsilon_{p+1} \ge (g' + \gamma x_{p+1})^+,\ldots,\varepsilon_n \ge (g' + \gamma x_n)^+ | \varepsilon_1 = z_1^0,\ldots,\varepsilon_p = z_p^0\right],$$

where $(x)^+ = \max(0,x)$. By the Theorem 3 this probability is equal to

$$\left(1 - \frac{1}{z^0}\sum_{i=p+1}^{n}(g' + \gamma x_i)^+\right)^{n-p-1}.$$

Symbol z^0 has the same meaning as in the Theorem 3. When we apply Jensen inequality to the convex function $(x)^+$, we get

$$\frac{1}{n-p}\sum_{i=p+1}^{n}(g' + \gamma x_i)^+ \ge \left(g' + \gamma\frac{1}{n-p}\sum_{i=p+1}^{n}x_i\right)^+, \tag{3.7}$$

therefore

$$\mathsf{P}_0\left[d_0^A + d_1^A\bar{x} \ge g | \varepsilon_1 = z_1^0,\ldots,\varepsilon_p = z_p^0\right] \le \left(1 - \frac{p}{z^0}\left(g' + \gamma\frac{1}{n-p}\sum_{i=p+1}^{n}x_i\right)^+\right)^{n-p-1} \tag{3.8}$$

The right side of inequality (3. ...pends on the values $x_{p+1},\ldots,x_n$ only through their average. But average is t... ...e when we replace all of values $x_{p+1},\ldots,x_n$ by their average. In this last ca... ...) is fulfilled as equality and on the right side of (3.8) is the maximally possibl... ...ue of this probability.

This assertion holds for every ...nstants $z_1^0,\ldots,z_p^0$, which fulfilled the conditions of the Theorem 3, therefore it ho... unconditionally. The same consideration we can do for the values $x_i \le \bar{x}$, too.

The probability of (3.5) is maximal, if it is true that

$$x_1 = \ldots = x_p = x_L, \quad x_{p+1} = \ldots = x_n = x_U.$$

Let us define

$$\varepsilon_L = \min_{1\le i\le p}\varepsilon_i,\ \varepsilon_U = \min_{p+1\le i\le n}\varepsilon_i.$$

Points (x_L,ε_L), (x_U,ε_U) lie on the line $\varepsilon = d_0^A + d_1^A x$. For $x = \bar{x}$ we get on this line the value

$$\varepsilon_g = \left(1 - \frac{\bar{x} - x_L}{x_U - x_L}\right)\varepsilon_L + \frac{\bar{x} - x_L}{x_U - x_L}\varepsilon_U.$$

Because of

$$\bar{x} = \frac{p}{n} x_L + \frac{n-p}{n} x_U,$$

we can write

$$\varepsilon_g = \frac{p}{n} \varepsilon_L + \frac{n-p}{n} \varepsilon_U.$$

In our special case (only two distinct values of x_i, see Appendix A6) we get

$$\mathsf{P}\left[\varepsilon_g \geq g\right] = \mathsf{P}\left[T \geq ng\right],$$

where random variable T has beta distribution with parameters $2, n-2$. It follows that

$$c^A = 1 - \mathrm{B}(2, n-2; 1-\alpha),$$

where $\mathrm{B}(f_1, f_2; q)$ is a q-quantile of beta distribution with parameters f_1, f_2. We are able to formulate our main assertion:

Theorem 4. The maximal level of significance of the critical region defined by the inequality

$$t^0 > t^A \left(1 - \mathrm{B}(2, n-2; 1-\alpha)\right)^{-1}$$

is α.

The standard asymptotic value of the constant c is the quantile $\chi^2(2, 1-\alpha)$. We mentioned that the standard properties of the ML-estimate are not fulfilled in our case. Let us compare for $\alpha = 0.05$ our value of c^A with analogous value

$$c^{A\text{class.}} = \exp\left(-\frac{\chi^2(2, 1-\alpha)}{2n}\right) \tag{3.9}$$

of the standard asymptotic test.

Table 1. Comparison of critical values c^A and $c^{A\text{class.}}$.

n	5	10	20	50	100
$c^{A\text{class.}}$	0.54928	0.74113	0.86089	0.94184	0.97049
c^A	0.24860	0.57086	0.77363	0.90681	0.95298

In Table 1 it can be seen that an erroneous application of $c^{A\text{class.}}$ in (3.1) gives a critical region with a first kind error probability greater than α. For $n \to \infty$ both of the critical values c^A and $c^{A\text{class.}}$ tend to 1. Asymptotic behavior of $c^{A\text{class.}}$ can be found from (3.9). Asymptotic behavior of c^A follows from the fact that the mean of the random variable T with beta distribution with parameters 2, $n-2$ tends to 0 and its variance tends to 0.

4. A CONFIDENCE BAND FOR THE REGRESSION LINE

Inequality (3.1) introduces a confidence set for β_0, β_1. The hypothesis H_0 is not rejected in case when

$$t^0 = \bar{y} - \beta_0^0 - \beta_1^0 \bar{x} \leq t^A / c^A.$$

Let us denote

$$\begin{aligned} y^A &= \bar{y} - t^A / c^A \\ &= \bar{y} - t^A / (1 - \mathrm{B}(2, n-2; 1-\alpha)). \end{aligned}$$

The set

$$K = \{(\beta_0, \beta_1)' : \beta_0 + \beta_1 \bar{x} \geq y^A\} \cap M_n \tag{4.1}$$

constitutes a $(1-\alpha)$-confidence set for $(\beta_0, \beta_1)'$.

A classical method of derivation confidence band for regression line from the confidence set K consists in maximization and minimization of $\beta_0 + \beta_1 x$ for given x over the $(\beta_0, \beta_1)' \in K$ (e. g. Chapter 11 of [7]). It can be proved [6]) that when with probability 1 the set K is convex, closed and bounded, then the confidence band has the same confidence coefficient as original confidence set K. Set K is an intersection of $n+1$ half-planes

$$\begin{aligned} &\beta_0 + \beta_1 x_i \leq y_i, \quad 1 \leq i \leq n, \\ &\beta_0 + \beta_1 \bar{x} \geq \bar{y} - t^A / c^A. \end{aligned}$$

This fact induces convexity and closeness of K. Now we prove its boundedness, too.

The p r o o f will be done by contradiction. Let us suppose that the set K is not bounded. It follows from linear programming theory that in this case there exist real numbers b_0, b_1, c_0, c_1 such that $c_0^2 + c_1^2 > 0$ and

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \lambda \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} \in K$$

for every $\lambda \geq 0$. Especially it means that

$$(b_0 + b_1 x_i) + \lambda(c_0 + c_1 x_i) \leq y_i, \quad 1 \leq i \leq n, \lambda \geq 0,$$

which is possible only when

$$c_0 + c_1 x_i \leq 0, \quad 1 \leq i \leq n. \tag{4.2}$$

The last inequality implies that

$$c_0 + c_1 \bar{x} \leq 0. \tag{4.3}$$

On the other hand it must be true that

$$(b_0 + b_1 \bar{x}) + \lambda(c_0 + c_1 \bar{x}) \geq y^A$$

for every $\lambda \geq 0$, which implies

$$c_0 + c_1 \bar{x} \geq 0. \tag{4.4}$$

Inequalities (4.3) and (4.4) together give $c_0 + c_1 \bar{x} = 0$. The last equation together with (4.2) gives

$$c_0 + c_1 x_i = 0, \quad 1 \leq i \leq n.$$

With exception $x_1 = \ldots = x_n$, which was excluded by (1.2), it follows $c_0 = c_1 = 0$. This is contradiction with $c_0^2 + c_1^2 > 0$, therefore set K is bounded.

Now we will derive the confidence band. Let us define two functions

$$\begin{aligned} L(x) &= \inf_{(\beta_0,\beta_1)' \in K} (\beta_0 + \beta_1 x), \\ U(x) &= \sup_{(\beta_0,\beta_1)' \in K} (\beta_0 + \beta_1 x). \end{aligned}$$

Because of properties of K the confidence band defined by the inequalities

$$L(x) \leq y \leq U(x), \; x \in \mathsf{R}$$

has the same confidence coefficient as the set K.

For a chosen x the lower bound $L(x)$ is given by the solution of linear programming problem

$$\text{minimize } (\beta_0 + \beta_1 x)$$

over the solutions of

$$\begin{aligned} &\beta_0 + \beta_1 x_i \leq y_i, \quad 1 \leq i \leq n, \\ &\beta_0 + \beta_1 \bar{x} \geq y^A. \end{aligned}$$

From linear programming theory it follows that any solution lies in some of the finite number of extreme points of K. It follows that $L(x)$ is piecewise linear function.

Similarly, the function $U(x)$ is piecewise linear function, which is in a surrounding of $\bar{x}$ equal to the estimated regression function $b_0 + b_1 x$.

5. SIMULATION EXPERIMENT

To verify properties of the proposed test and the proposed confidence intervals a simulation experiment similar to the experiment in [8] was made. We used the regression line with parameters $\beta_0 = 0, \beta_1 = 1$. To compare behavior of the statistics in different situations we used eight designs: seven of them was similar to those in cited paper, the last one (signed by "R") was given by random selection of n points in interval $(-2, 2)$. For simulation of error term we used the exponential distribution with expectation 1. Table 2 gives list of these designs. For each design the table gives estimated confidence level of confidence set K given by (4.1). The designs are sorted by their variances. It can be seen that the estimated confidence levels are very close to the nominal levels. The designs with a greater number of distinct points (with a greater variability) have its estimated confidence levels a little bit greater than their nominal counterparts.

Table 2. The estimated confidence levels for $n = 20$ estimated by 5000 simulations.

	Design	90.0%	95.0%
A	$-1.0(10\times), 1.0(10\times)$	90.6%	95.1%
F	$-1.0(10\times), 0.5(3\times), 1.0(4\times), 1.5(3\times)$	89.8%	94.6%
E	$-1.0(10\times), 0.5(5\times), 1.5(5\times)$	90.0%	95.2%
D	$-1.5(3\times), -1.0(4\times), -0.5(3\times), 0.5(3\times), 1.0(4\times), 1.5(3\times)$	90.6%	95.3%
C	$-1.0(10\times), 0.1, 0.3, \ldots, 1.9$	91.2%	95.5%
G	$-1.5(5\times), -0.5(5\times), 0.5(5\times), 1.5(5\times)$	91.9%	96.1%
B	$-1.9, -1.7, \ldots, -0.1, 0.1, 0.3, \ldots, 1.9$	92.3%	96.4%
R	randomly selected on $(-2.0, 2.0)$	91.4%	95.9%

6. EXAMPLE

Nátrová and Nátr [2] among others studied dependency of kernel dry weight on the phloem cross-sectional area. They used 51 measurements which are given in the Table 3. The best limiting line was found to be given by

$$y = 0.401 + 0.0390x$$

with the estimate $t^A = 0.214$ of parameter θ. This line with a 95 % confidence region can be seen in Figure 1. A simple graphical test of the exponential distribution of errors is done by a probability plot for exponential distribution in Figure 2. It can be seen that the assumption of exponential distribution is a quite realistic. All figures and computations are given by the FamStat programs [1].

Table 3. Kernel dry weight and total phloem cross-sectional area of wheat.

Area	26.96	33.87	28.97	31.21	28.45	36.18	32.35	37.26	31.56
Weight	1.45	1.67	1.45	1.46	1.47	1.41	1.52	1.67	1.58
Area	45.17	28.66	27.20	35.60	36.23	34.31	37.71	38.05	34.11
Weight	2.10	1.48	1.30	1.54	1.66	1.58	1.77	1.55	1.46
Area	28.99	39.11	31.30	34.15	29.14	41.32	20.34	35.90	37.09
Weight	1.53	1.80	1.50	1.70	1.45	1.41	1.19	1.19	1.44
Area	26.21	35.33	32.23	34.51	36.40	47.35	36.89	43.55	32.16
Weight	1.27	1.60	1.56	1.32	1.58	1.73	1.36	1.69	1.32
Area	26.76	39.02	36.51	48.31	35.41	41.95	39.19	35.15	38.98
Weight	1.41	1.55	1.64	1.71	1.75	1.56	1.86	1.77	1.71
Area	32.33	34.58	28.63	32.93	45.92	25.05			
Weight	1.61	1.57	1.48	1.37	1.46	1.11			

APPENDIX. DIRICHLET DISTRIBUTION

A1. Let $X_1, \ldots, X_n$ be independent random variables with gamma distribution $X_i \sim \text{Gamma}(f_i), 1 \leq i \leq n$. Let us introduce new random variables

$$Z_i = \frac{X_i}{\sum_{j=1}^{n} X_j}, \quad 1 \leq i \leq n.$$

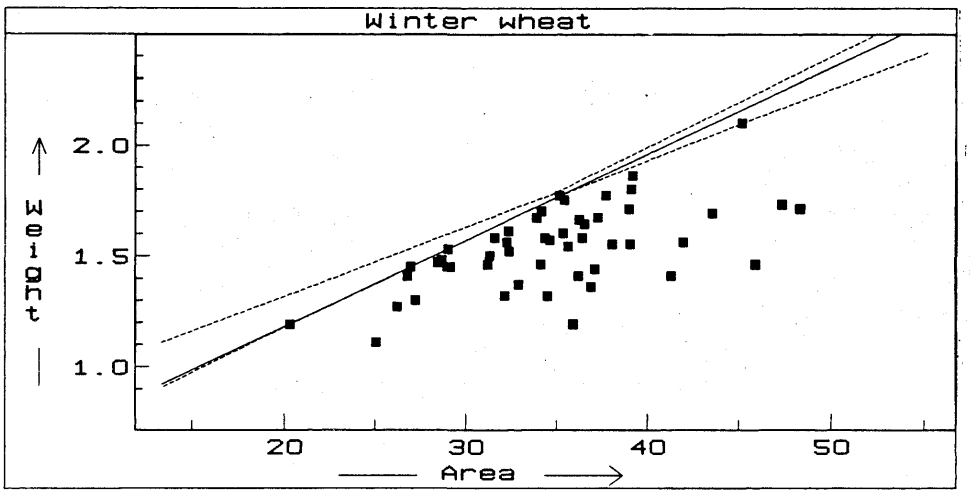

Fig. 1. Dependence of kernel weight on phloem area.

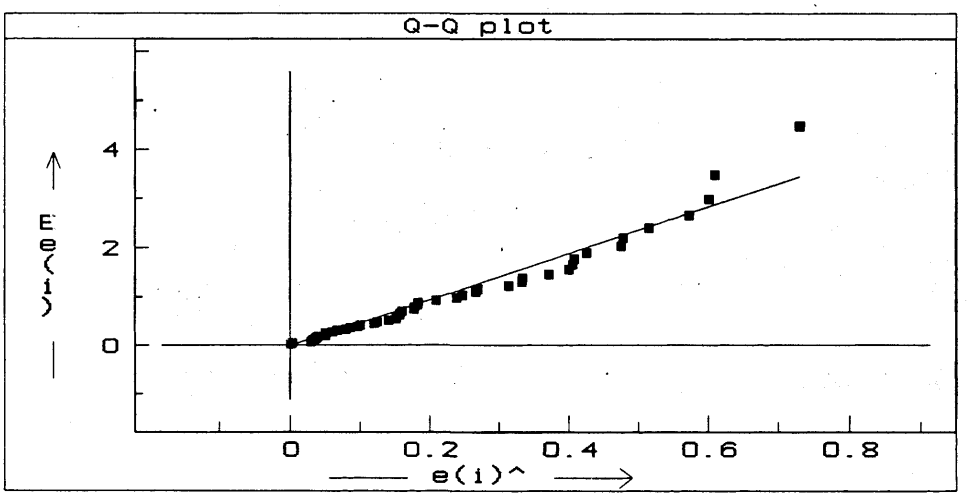

Fig. 2. Probability plot of residuals for exponential distribution.

It can be seen that $\sum_{i=1}^{n} Z_i = 1$. By a standard calculus (see 7.7.1 of [4]) we get that the random vector $(Z_1, \ldots, Z_{n-1})'$ has the Dirichlet distribution $\mathrm{D}(f_1, \ldots, f_{n-1}; f_n)$

with density function (we denote $z_n = 1 - \sum_{i=1}^{n-1} z_i$)

$$h(z_1, \ldots, z_{n-1}) = \begin{cases} \dfrac{\Gamma(\sum_{i=1}^{n} f_i)}{\prod_{i=1}^{n} \Gamma(f_i)} \prod_{i=1}^{n} z_i^{f_i - 1} & \text{for } z_1 \geq 0, \ldots, z_{n-1} \geq 0, \sum_{i=1}^{n-1} z_i \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Exponential distribution Ex(1) is gamma distribution Gamma(1), therefore the random variables $\varepsilon_1, \ldots, \varepsilon_{n-1}$ defined in (3.4) have the Dirichlet distribution D(1,..., 1; 1). We can see that in this case the density of the random vector $(\varepsilon_1, \ldots, \varepsilon_{n-1})'$ is equal to the constant $(n-1)!$ on the simplex $\{z \in \mathsf{R}^{n-1} : z_1 \geq 0, \ldots, z_{n-1} \geq 0, \sum_{i=1}^{n-1} z_i \leq 1\}$.

A2. Let $1 \leq p < n$. By 7.7.2 of [4] the marginal distribution of the random variables $Z_1, \ldots, Z_p$ is the Dirichlet distribution $\mathrm{D}(f_1, \ldots, f_p; f_{p+1} + \ldots + f_n)$.

A3. Let us have real numbers $z_1^0 \geq 0, \ldots, z_p^0 \geq 0$, let $z^0 = 1 - \sum_{i=1}^{p} z_i^0 \geq 0$. From A1 and A2 it follows that the conditional density of

$$Z_{p+1}, \ldots, Z_{n-1} | Z_1 = z_1^0, \ldots, Z_p = z_p^0$$

for $z_{p+1} \geq 0, \ldots, z_{n-1} \geq 0, z_n = z^0 - \sum_{i=p+1}^{n-1} z_i \geq 0$ is given by

$$h(z_{p+1}, \ldots, z_{n-1} | Z_1 = z_1^0, \ldots, Z_p = z_p^0) = \frac{\Gamma(\sum_{i=p+1}^{n} f_i))}{\prod_{i=p+1}^{n} \Gamma(f_i)} z_{p+1}^{f_{p+1}-1} \ldots z_n^{f_n - 1} (z^0)^{\sum_{i=p+1}^{n} f_i - 1},$$

otherwise the conditional density is equal to zero. It follows that the conditional distribution

$$Z_{p+1}/z^0, \ldots, Z_{n-1}/z^0 | Z_1 = z_1^0, \ldots, Z_p^0 = z_p^0$$

is the Dirichlet distribution $\mathrm{D}(f_{p+1}, \ldots, f_{n-1}; f_n)$.

A4. Suppose that $f_i = 1$, $1 \leq i \leq n$. Let us have real numbers $z_1 \geq 0, \ldots, z_n \geq 0$, so that $\sum_{i=1}^{n} z_i \leq 1$. To compute the probability $\mathsf{P}[\varepsilon_1 \geq z_1, \ldots, \varepsilon_n \geq z_n]$ we have to integrate the constant density of $\mathrm{D}(1, \ldots, 1; 1)$ distribution over the simplex

$$\left\{ t_1 \geq z_1, \ldots, t_{n-1} \geq z_{n-1}, \sum_{i=1}^{n-1} t_i \leq 1 - z_n \right\}.$$

The volume of this $(n-1)$-dimensional simplex is equal to

$$\frac{1}{(n-1)!} \left(1 - \sum_{i=1}^{n} z_i\right)^{n-1},$$

therefore

$$\mathsf{P}\left[\varepsilon_1 \geq z_1, \ldots, \varepsilon_n \geq z_n\right] = \left(1 - \sum_{i=1}^{n} z_i\right)^{n-1}.$$

A5. If follows from A3 that the conditional distribution of $(1/z^0)$-multiple of the random variables $\varepsilon_{p+1}, \ldots, \varepsilon_{n-1}$ can be identify with the $\mathrm{D}(1, \ldots, 1; 1)$ distribution. Therefore,

$$\mathsf{P}\left[\varepsilon_{p+1} \geq z_{p+1}, \ldots, \varepsilon_n \geq z_n | \varepsilon_1 = z_1^0, \ldots, \varepsilon_p = z_p^0\right] = \left(1 - \sum_{i=p+1}^{n} \left(\frac{z_i}{z^0}\right)\right)^{n-p-1}.$$

A6. Let us have an integer p $(2 \leq p \leq n-1)$, let $n \geq 3$. Let us denote

$$\varepsilon_L = \min_{1 \leq i \leq p} \varepsilon_i, \quad \varepsilon_U = \min_{p+1 \leq i \leq n} \varepsilon_i.$$

To find the density of the random vector $(\varepsilon_L, \varepsilon_U)$ we will compute for $z_L \geq 0$, $z_U \geq 0$, $pz_L + (n-p)z_U \leq 1$ probability

$$\begin{aligned} \mathsf{P}\left[\varepsilon_L \geq z_L, \varepsilon_U \geq z_U\right] &= \mathsf{P}\left[\varepsilon_1 \geq z_L, \ldots, \varepsilon_p \geq z_L, \varepsilon_{p+1} \geq z_U, \ldots, \varepsilon_n \geq z_U\right] \\ &= \left(1 - pz_L - (n-p)z_U\right)^{n-1}, \end{aligned}$$

where we used A4. The joint density of $(\varepsilon_L, \varepsilon_U)$ is given for $p\varepsilon_L + (n-p)z_U \leq 1$, $z_L \geq 0$, $z_U \geq 0$ by

$$\begin{aligned} f_{L,U}(z_L, z_U) &= \frac{\partial^2}{\partial z_L \partial z_U} \mathsf{P}\left[\varepsilon_L \leq z_L, \varepsilon_U \leq z_U\right] \\ &= \frac{\partial^2}{\partial z_L \partial z_U} \mathsf{P}\left[\varepsilon_L \geq z_L, \varepsilon_U \geq z_U\right] \\ &= (n-1)(n-2)p(n-p)\left(1 - pz_L - (n-p)z_U\right)^{n-3}. \end{aligned}$$

Otherwise $f_{L,U}(z_L, z_U) = 0$.

Let us define a new random vector (T, S) by

$$\begin{aligned} T &= p\varepsilon_L + (n-p)\varepsilon_U, \\ S &= -(n-p)\varepsilon_L + p\varepsilon_U. \end{aligned}$$

Standard calculus gives the joint density of (T, S)

$$f_{T,S}(t, s) = \frac{p(n-p)}{p^2 + (n-p)^2}(n-1)(n-2)(1-t)^{n-3}$$

for $-\frac{n-p}{p}t < s < \frac{p}{n-p}t$, $0 \leq t \leq 1$, otherwise the density is equal to zero. By integration we get the marginal density of the random variable T for $0 \leq t \leq 1$ as

$$\begin{aligned} f_T(t) &= \frac{(n-1)!}{(n-3)!} t(1-t)^{n-3} \\ &= \frac{1}{\mathrm{B}(2, n-2)} t(1-t)^{n-3}, \end{aligned}$$

$f_T(t) = 0$ otherwise. Therefore, the variable T has B distribution with 2 and $n-2$ degrees of freedom.

ACKNOWLEDGEMENT

The author is grateful to the referees for their suggestions.

This research has been supported by the grant No. 2169 from the Grant Agency of the Czech Republic.

(Received November 29, 1993.)

REFERENCES

[1] J. Hanousek, P. Charamza, M. Malý and K. Zvára: FamStat – Statistické knihovny (Famulus 3.5). FamStat 1993.
[2] Z. Nátrová and L. Nátr: Limitation of kernel yield by the size of conducting tissue in winter wheat varieties. Field Crops Research *XX* (1993), 121–130.
[3] C. R. Rao: Linear Statistical Inference and its Applications. Wiley, New York 1973.
[4] S. Wilks: Mathematical Statistics. Wiley, New York 1962.
[5] J. L. Walworth, W. S. Letzsch and M. E. Sumner: Use of boundary lines in establishing diagnostic norms. Soil Science Society of America Journal *50* (1986), 123–127.
[6] K. Zvára: On exact confidence regions for linear regression functions. Math. Operationsforsch. u. Statist., Ser. Statist. *10* (1979), 55–62.
[7] K. Zvára: Regression Analysis (in Czech). Academia, Prague 1989.
[8] K. Zvára: Consistency of an estimate in linear regression with non-negative errors. Kybernetika *28* (1992), 129–139.

RNDr. Karel Zvára, CSc., Matematicko-fyzikální fakulta Univerzity Karlovy (Faculty of Mathematics and Physics – Charles University), Sokolovská 83, 186 00 Praha 8. Czech Republic.