# Interval Estimate for Specific Points in Polynomial Regression

Katarina Košmelj[1], Andrej Blejec[2] and Anton Cedilnik[1]

[1]Biotechnical Faculty, University of Ljubljana, Slovenia
[2]National Institute of Biology, Ljubljana, Slovenia

This paper presents the interval estimate for specific points in polynomial regression: zero of a linear regression, abscissa of the extreme of a quadratic regression, abscissa of the inflection point of a cubic regression. Two different approaches are under study. An application of these two approaches based on quadratic regression in presented: interval estimate for the plant density giving optimal yield of maize is under consideration.

*Keywords:* zero of a linear regression, extreme of a quadratic regression, density of the ratio of two normal variables.

## 1. Introduction

Consider a general polynomial regression $E(Y|x) = \beta_0 + \beta_1 x + \ldots + \beta_m x^m$, $m \geq 1$. Let us define $Z = -B_{m-1}/(mB_m)$. For $m = 1$, $Z$ is an estimator for the zero of a linear regression. For $m = 2$, $Z$ is an estimator for the abscissa of the extreme of a quadratic regression, and for $m = 3$, $Z$ is an estimator for the abscissa of the inflection point of a cubic regression.

These points are often of research interest, in particular in biological and agricultural setting. Experimenters are interested in the point and interval estimate of $Z$. In this context, we consider these points as *specific points* of polynomial regression and focus our attention on the distribution of $Z$. In the last section, we consider the data from an agricultural experiment on maize. The experimenters were interested in the point and interval estimate of the plant density giving optimal yield.

## 2. Distribution of the Ratio of Jointly Normal Variables

Under standard regression assumptions, $Z$ is expressed as the ratio of two normally distributed and dependent variables. From the standard probability literature [3], it is known that the ratio of two centred normal variables:

$$Z = X/Y,$$
$$[X\ Y]^{\mathrm{T}} : N\,(\mu_X = \mu_Y = 0,\ \ \sigma_X, \sigma_Y,\ \ \rho \neq \pm 1)$$

is a non-centered Cauchy variable,

$$Z : C\left(a = \rho\frac{\sigma_X}{\sigma_Y},\ b = \frac{\sigma_X}{\sigma_Y}\sqrt{1-\rho^2}\right).$$

In [4] and [2], authors discussed the general situation. Independently of Hinkley, we [1] followed the same procedure as he did, but we expressed the density differently, as a product of two parts:

$$p_Z(z) = \frac{\sigma_X\sigma_Y\sqrt{1-\rho^2}}{\pi(\sigma_Y^2 z^2 - 2\rho\sigma_X\sigma_Y z + \sigma_X^2)} \cdot \left[\exp\left(-\frac{1}{2}\cdot \sup R^2\right) + \sqrt{2\pi}\cdot R \cdot \Phi(R)\cdot \exp\left(-\frac{1}{2}\cdot\left[\sup R^2 - R^2\right]\right)\right] \quad (1)$$

where:

$$R = R(z) = \frac{\left(\frac{\mu_X}{\sigma_X} - \rho\frac{\mu_Y}{\sigma_Y}\right)z - \left(\rho\frac{\mu_X}{\sigma_X} - \frac{\mu_Y}{\sigma_Y}\right)\frac{\sigma_X}{\sigma_Y}}{\sqrt{1-\rho^2}\cdot\sqrt{z^2 - 2\rho\frac{\sigma_X}{\sigma_Y}z + \left(\frac{\sigma_X}{\sigma_Y}\right)^2}}$$

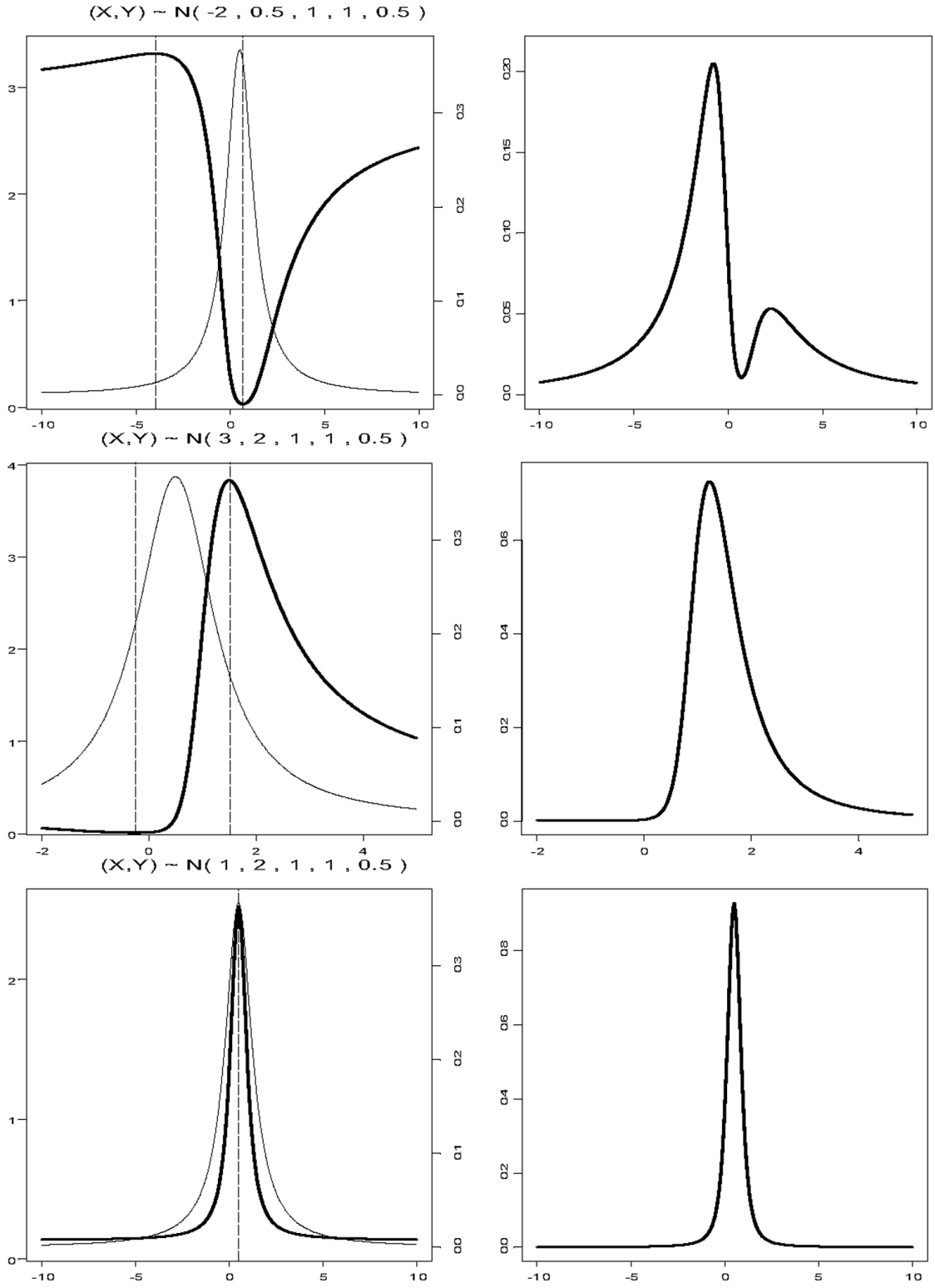


*Fig. 1.* On the left, the standard Cauchy part (thick line) and the deviant part (thin line) are presented; the scale for the deviant part is on the left, for the Cauchy part on the right y-axes. The vertical dashed lines indicate the abscissas of the local extremes of the deviant part. The right plot presents the density $p_Z(z)$.

$$\sup R^2 = const = \frac{\left(\frac{\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{\mu_X}{\sigma_X}\frac{\mu_Y}{\sigma_Y} + \left(\frac{\mu_Y}{\sigma_Y}\right)^2}{1-\rho^2}$$

$$\sup R^2 - R^2 = \frac{\left(\frac{\mu_X}{\sigma_X}\frac{\sigma_X}{\sigma_Y} - \frac{\mu_Y}{\sigma_Y}z\right)^2}{z^2 - 2\rho\frac{\sigma_X}{\sigma_Y}z + \left(\frac{\sigma_X}{\sigma_Y}\right)^2}.$$

The first factor in (1), the *standard part*, is the density for a non-centred Cauchy,

$$C\left(a = \rho\frac{\sigma_X}{\sigma_Y},\ b = \frac{\sigma_X}{\sigma_Y}\sqrt{1-\rho^2}\right);$$

it is independent of the expected values $\mu_X$ and $\mu_Y$. The second factor, the *deviant part*, is a rather complicated function of $z$. The asymptotic behaviour of $p_Z(z)$ is the same as that of the Cauchy density, consequently $E(Z)$ and other moments do not exist. We need four parameters to describe the distribution: $\rho$, $\frac{\mu_X}{\sigma_X}$, $\frac{\mu_Y}{\sigma_Y}$ and $\frac{\sigma_X}{\sigma_Y}$. In general, $p_Z(z)$ is bimodal.

We studied the density originating from the distribution N ($\mu_X$,$\mu_Y$, 1, 1, 0.5) for different values of $\mu_X$ and $\mu_Y$. The standard part is $C\left(0.5;\ \sqrt{3}/2\right)$, it is symmetric around its median 0.5. Different values for $\mu_X$ and $\mu_Y$ determine different shapes of the deviant part. Figure 1 displays three possible shapes of the density: evident bimodality (above), the deviant part prevails (in the middle); the Cauchy and the deviant part are superimposed and the density is unimodal (below).

## 3. Application

In an agricultural setting in Slovenia, experimenters studied the impact of the plant density on the yield of maize. In a field experiment, 15 different plant densities were included, each in four replications. The yield was the variable of interest. The experimenters were interested in the point and interval estimate for the plant density providing optimal yield.

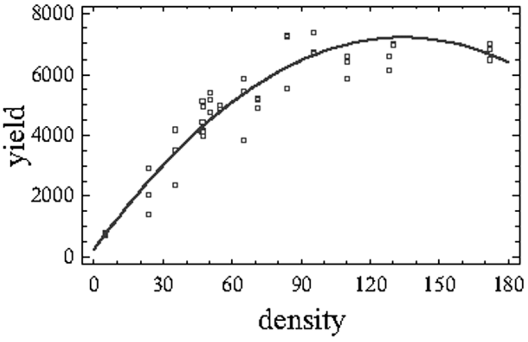


*Fig. 2.* Yield of maize [kg per ha)] depending on plant density [1000plants per ha]. The thick line represents the graph for quadratic polynomialfitted to the data.

The graph (Figure 2) and the analysis (Table 1) show that quadratic polynomial fits the data very well. The point estimate for the optimal density is $z = 103.588/(2 \cdot 0.385395) = 134.4$, however the interval estimate requires more elaborate consideration.

| Parameter | Estimate | St. Error | T-Statistic | P-Value |
|---|---|---|---|---|
| $b_0$ | 249.193 | 321.881 | 0.77418 | 0.4432 |
| $b_1$ | 103.588 | 8.34109 | 12.419 | 0.0000 |
| $b_2$ | -0.385395 | 0.046127 | -8.35504 | 0.0000 |

$R^2 = 88.5$ %.

*Table 1.* Quadratic Regression Analysis. Dependent variable: yield per ha, independent variable: plant density.

We obtained the estimates for the bivariate normal distribution from Table 1 and correlation matrix of the regression estimates:
$-b_1 = -103.6,\ 2b_2 = 0.771,\ s(-b_1) = 8.341,$
$s(2b_2) = 0.0923,\ cor(-b_1, 2b_2) = 0.961.$

These values were plugged into the computer program as if they were the true values of the bivariate normal distribution (quasi-analytical approach). Figure 3 shows the Cauchy part (first factor) and the deviant part (second factor), the probability density and the distribution function, the latter is obtained numerically. Interval estimate is obtained from the distribution function as the interval containing 95% of the values, half on each side of the point estimate.

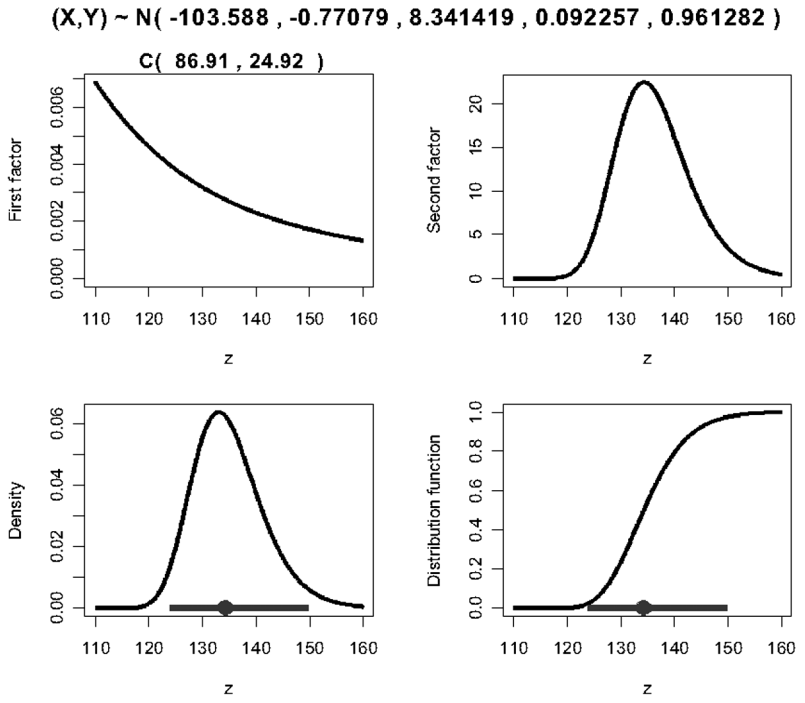


*Fig. 3.* Cauchy part (First factor, above left) and the deviant part (Second factor, above right) for the bivariate normal distribution, probability density (below left) and distribution function (below right). Interval estimate for optimal plant density is obtained numerically from the distribution function.

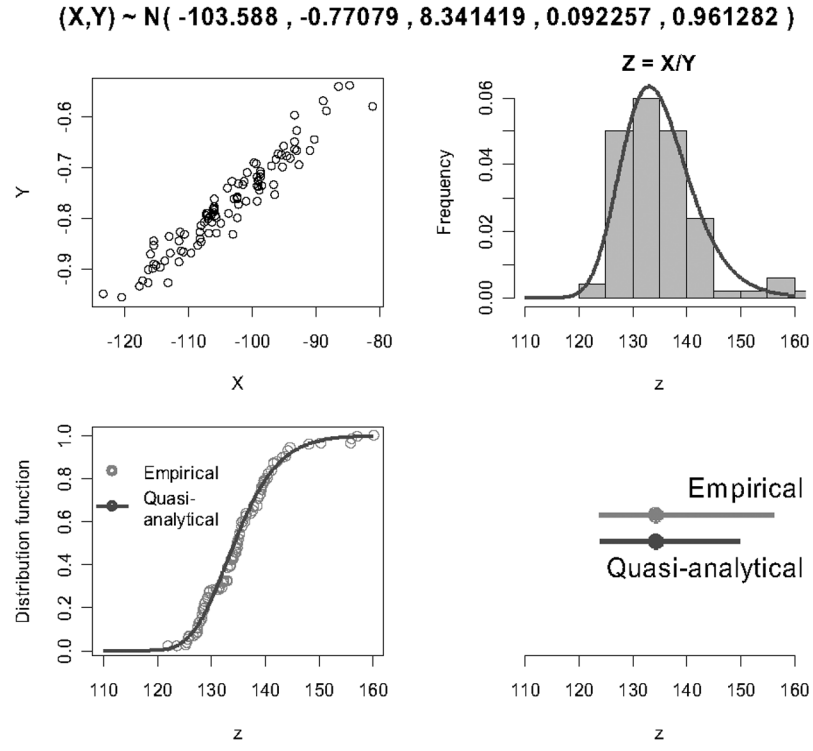


*Fig. 4.* Scatterplot of 100 values generated from the bivariate normal distribution (left above), histogram of the ratios and the density obtained from the quasi-analytical approach (right above). The two distribution functions obtained numerically (left below), the comparison of the interval estimates.

As an alternative approach we generated a random sample from the bivariate normal distribution obtained above (empirical approach). Two sample sizes were considered, 100 and 1000. Figure 4 presents the scatterplot of 100 generated values, the histogram for the ratios fitted by the density obtained from the quasi-analytical approach. The two distribution functions were obtained numerically and were used to acquire the two interval estimates.

We present the final results for the interval estimate in Table 2. The comparison of the intervals shows minor discrepancy with quasi-analytical approach, in particular for the sample size 1000.

| Approach | Lower bound | Upper bound |
|---|---|---|
| Quasi analytical | 123.9 | 149.9 |
| Empirical: 100 points | 123.3 | 151.0 |
| Empirical: 1000 points | 123.8 | 150.4 |

*Table 2.* Interval estimate for optimal plant density obtained by two different approaches (see text).

## 4. Conclusion

This paper presents two different approaches to obtain the interval estimate for a ratio of two normally distributed and dependent variables. The quasi-analytical approach is based on the derived probability density, the regression estimates are taken as true values of the parameters of the bivariate normal distribution. The second approach is empirical, random samples from the corresponding distribution are generated.

## 5. Acknowledgments

We thank Prof. dr. Anton Tajnšek, University of Ljubljana, for the data.

## References

[1] A. CEDILNIK, K. KOŠMELJ, A. BLEJEC, (2004). The Distribution of the Ratio of Jointly Normal Variables. *Metodološki zvezki*. Vol.1, No. 1, 2004, pp. 99–108.

[2] D. V. HINKLEY, (1969). On the ratio of two correlated normal random variables. *Biometrika*, **56**, 3, pp. 635–639.

[3] N. L. JOHNSON, S. KOTZ AND N. BALAKRISHNAN, (1994). Continuous Univariate Distributions. 1. John Wiley and Sons.

[4] G. MARSAGLIA, (1965). Ratios of normal variables and ratios of sums of uniforms variables. *JASA*, 60, pp. 163–204.

*Recived:* June, 2005.
*Accepted:* October, 2005.

*Contact address:*
Biotechnical Faculty
University of Ljubljana
Ljubljana
Slovenia
Katarina.Kosmelj@bf.uni-lj.si

Andrej Blejec
National Institute of Biology
Ljubljana
Slovenia
Andrej.Blejec@nib.si

Anton Cedilnik
Biotechnical Faculty
University of Ljubljana
Ljubljana
Slovenia
Anton.Cedilnik@bf.uni-lj.si

KATARINA KOŠMELJ is a professor of statistics at Biotechnical Faculty, University of Ljubljana. She is specialized in design and analysis of experiments and data analysis. Her area of work is applied biostatistics.

ANDREJ BLEJEC is an assistant professor of statistics and computer science at Biotechnical Faculty, University of Ljubljana. Currently he works as a statistician at the National Institute of Biology in Ljubljana. His main interest is computational statistics, data visualization and statistical simulation systems for statistics education.

ANTON CEDILNIK is an associate professor of mathematics at Biotechnical Faculty, University of Ljubljana. His main research interest is functional analysis and algebra. As an applied mathematician he collaborates in many interdisciplinary projects.