

Math 362: Mathematical Statistics II

Le Chen

le.chen@emory.edu
chenle02@gmail.com

Emory University
Atlanta, GA

Last updated on Spring 2021
Last compiled on January 15, 2023

2021 Spring

Creative Commons License
(CC By-NC-SA)

Chapter 11. Regression

§ 11.1 Introduction

§ 11.4 Covariance and Correlation

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.A Appendix Multiple/Multivariate Linear Regression

§ 11.5 The Bivariate Normal Distribution

Chapter 11. Regression

§ 11.1 Introduction

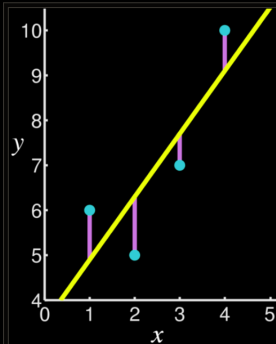
§ 11.4 Covariance and Correlation

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.A Appendix Multiple/Multivariate Linear Regression

§ 11.5 The Bivariate Normal Distribution



In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (y) and an independent variable (x). □

Goal: Find a blue line that minimizes the sum of the square of the green lines

Thm. Given n points $(x_1, y_1), \dots, (x_n, y_n)$, the straight line $y = a + bx$ minimizing

$$L(a, b) = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

when

$$b = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n (\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

and

$$a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n} = \bar{y} - b\bar{x}.$$

Proof.

$$\begin{cases} \frac{\partial}{\partial a} L(a, b) = \sum_{i=1}^n (-2) [y_i - (a + bx_i)] = 0 \\ \frac{\partial}{\partial b} L(a, b) = \sum_{i=1}^n (-2x_i) [y_i - (a + bx_i)] = 0 \end{cases} \quad \text{(Normal equations)}$$

$$\Longleftrightarrow \begin{cases} \sum_{i=1}^n y_i - na - b \sum_{i=1}^n x_i = 0 & (1) \\ \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0 & (2) \end{cases}$$

$$(1) \implies a = \bar{y} - b\bar{x}$$

$$(1) \times \sum_{i=1}^n x_i - (2) \times n \implies b = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n (\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

□

(Moore-Penrose) Pseudoinverse

1. Well determined system

$$Ax = b \implies x = A^{-1}y.$$

2. Overdetermined system

$$\begin{aligned} Ax &= y \\ A^T Ax &= A^T y \\ \underbrace{(A^T A)^{-1} A^T A}_{=I} x &= (A^T A)^{-1} A^T y \\ x &= \underbrace{(A^T A)^{-1} A^T}_{=: A^+} y \end{aligned}$$

3. Under determined system

$$Ax = y \implies x = \underbrace{A^T (AA^T)^{-1}}_{=: A^+} y.$$

Proof. (Another proof based on pseudoinverse)

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n \times 2}, \quad x = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2 \times 1}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{1 \times n}$$

$$A^T A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$(A^T A)^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

$$A^T y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = x = (A^T A)^{-1} A^T y$$

$$= \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

$$= \begin{pmatrix} \frac{(\sum_{i=1}^n x_i^2) (\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i) (\sum_{i=1}^n x_i y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \end{pmatrix}$$

$$b = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}.$$

$$\begin{aligned} a &= \frac{(\sum_{i=1}^n x_i^2) (\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i) (\sum_{i=1}^n x_i y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{(\sum_{i=1}^n x_i^2) (\sum_{i=1}^n y_i) - (\sum_{i=1}^n x_i) [(\sum_{i=1}^n x_i y_i) - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)]}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &\quad - \frac{\frac{1}{n} (\sum_{i=1}^n x_i)^2 (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{1}{n} \sum_{i=1}^n y_i - b \frac{1}{n} \sum_{i=1}^n x_i = \bar{y} - b\bar{x}. \end{aligned}$$

□

A probabilistic view ...

Def. The function $f(X)$ for which

$$\mathbb{E} [(Y - f(X))^2]$$

is minimized is called the **regression curve of Y on X** .

Thm. Let (X, Y) be two random variables such that $\text{Var}(X)$ and $\text{Var}(Y)$ both exist. Then the regression curve of Y on X is given (for all x) by

$$f(x) = \mathbb{E} [Y|X = x] .$$

Proof. Let $f(x) = \mathbb{E}[Y|X = x]$ and let $\phi(x)$ be a general function. Then

$$\begin{aligned}\mathbb{E}[(Y - \phi(X))^2] &= \mathbb{E}[(Y - f(X)) + (f(X) - \phi(X))]^2 \\ &= \mathbb{E}[(Y - f(X))^2] + \mathbb{E}[(f(X) - \phi(X))^2] \\ &\quad + \mathbb{E}[(Y - f(X))(f(X) - \phi(X))].\end{aligned}$$

Let $\psi(x)$ be either $f(x)$ or $\phi(x)$. We claim that

$$\mathbb{E}[(Y - f(X))\psi(X)] = 0.$$

Indeed,

$$\begin{aligned}\mathbb{E}[Y\psi(X)] &= \int \int_{\mathbb{R}^2} f_{X,Y}(x,y) y \psi(x) dy dx \\ &= \int_{\mathbb{R}} dx \psi(x) f_X(x) \underbrace{\int_{\mathbb{R}} dy \frac{f_{X,Y}(x,y)}{f_X(x)} y}_{= \mathbb{E}[Y|X = x]} \\ &= \mathbb{E}[f(X)\psi(X)].\end{aligned}$$

Hence,

$$\mathbb{E}[(Y - \phi(X))^2] = \mathbb{E}[(Y - f(X))^2] + \mathbb{E}[(f(X) - \phi(X))^2]$$

which is minimized when $\phi(x) = f(x)$. □

If one imposes that $f(x) = a + bx$, then

Thm. The following squared error:

$$\mathbb{E} [\{Y - (a + bX)\}^2]$$

is minimized at

$$b = \rho_{XY} \frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2} \quad \text{and} \quad a = \mathbb{E}[Y] - b\mathbb{E}[X]$$

with the mean squared error

$$\mathbb{E} [\{Y - (a + bX)\}^2] = (1 - \rho_{XY}^2) \sigma_Y^2.$$

Proof.

$$\begin{aligned}
 & \mathbb{E} [\{Y - (a + bX)\}^2] \\
 = & \mathbb{E} \left[\left\{ [Y - \mathbb{E}(Y)] - b[X - \mathbb{E}(X)] - [a - \mathbb{E}[Y] + b\mathbb{E}(X)] \right\}^2 \right] \\
 = & \mathbb{E} \left[[Y - \mathbb{E}(Y)]^2 \right] + b^2 \mathbb{E} \left[[X - \mathbb{E}(X)]^2 \right] + \left[a - \mathbb{E}[Y] + b\mathbb{E}(X) \right]^2 \\
 & - 2b \mathbb{E} \left[[Y - \mathbb{E}(Y)][X - \mathbb{E}(X)] \right] \\
 & - 2 \left[a - \mathbb{E}[Y] + b\mathbb{E}(X) \right] \mathbb{E} [Y - \mathbb{E}(Y)] \\
 & + 2b \left[a - \mathbb{E}[Y] + b\mathbb{E}(X) \right] \mathbb{E} [X - \mathbb{E}(X)] \\
 = & \text{Var}(Y) + b^2 \text{Var}(X) + \left[a - \mathbb{E}[Y] + b\mathbb{E}(X) \right]^2 - 2b \text{Cov}(X, Y) + 0 + 0
 \end{aligned}$$

$$\begin{aligned}
& \Downarrow \\
& \mathbb{E} \left[\{Y - (a + bX)\}^2 \right] \\
& \parallel \\
& \text{Var}(Y) + b^2 \text{Var}(X) + \left[a - \mathbb{E}[Y] + b\mathbb{E}(X) \right]^2 - 2b \text{Cov}(X, Y)
\end{aligned}$$

The best a , called a^* , should be such that

$$\left[a^* - \mathbb{E}[Y] + b\mathbb{E}(X) \right]^2 = 0 \quad \Longleftrightarrow \quad a^* = \mathbb{E}[Y] - b\mathbb{E}[X]$$

$$\begin{aligned}
& \Downarrow \\
& \mathbb{E} \left[\{Y - (a^* + bX)\}^2 \right] \\
& \parallel \\
& \text{Var}(Y) + b^2 \text{Var}(X) - 2b \text{Cov}(X, Y) \\
& \parallel \\
& \sigma_Y^2 + b^2 \sigma_X^2 - 2b \rho_{XY} \sigma_X \sigma_Y \\
& \parallel \\
& (1 - \rho_{XY}^2) \sigma_Y^2 + \left(b \sigma_X - \rho_{XY} \sigma_Y \right)^2
\end{aligned}$$

The best b , called b^* , should be

$$(b^* \sigma_X - \rho_{XY} \sigma_Y)^2 = 0 \quad \Longleftrightarrow \quad b^* = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

$$\begin{aligned}
& \Downarrow \\
& \mathbb{E} \left[\{Y - (a^* + b^* X)\}^2 \right] \\
& \parallel \\
& (1 - \rho_{XY}^2) \sigma_Y^2
\end{aligned}$$

with

$$b^* = \rho_{XY} \frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2} \quad \text{and} \quad a^* = \mathbb{E}[Y] - b^* \mathbb{E}[X]$$

□

Remark In practice, we have data $(x_1, y_1), \dots, (x_n, y_n)$ instead of the joint law of (X, Y)

\Downarrow

Replace

$$\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}, \sigma_{XY}$$

by their maximum likelihood estimates

$$\bar{x}, \bar{y}, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, r_{XY}, \hat{\sigma}_{XY}$$

$$1. \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$2. \hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n^2}$$

$$\hat{\sigma}_Y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2 = \frac{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}{n^2}$$

$$\begin{aligned} 3. \hat{\sigma}_{XY} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \\ &= \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n^2} \end{aligned}$$

$$4. r_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

↓

$$b = r_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}, \quad a = \bar{y} - b\bar{x}$$

Maximum likelihood estimates

$$\hat{\sigma}_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\hat{\sigma}_Y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\hat{\sigma}_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})$$

Sample (co)variances

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

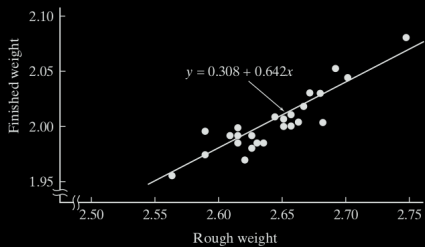
$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})$$

E.g. 1 Producing air conditioners. x = rough weight of a rod. y = finished weight. Find the best linear approximation of xy -relationship. Predict the weight when $x = 2.71$

Table 11.2.1					
Rod Number	Rough Weight, x	Finished Weight, y	Rod Number	Rough Weight, x	Finished Weight, y
1	2.745	2.080	14	2.635	1.990
2	2.700	2.045	15	2.630	1.990
3	2.690	2.050	16	2.625	1.995
4	2.680	2.005	17	2.625	1.985
5	2.675	2.035	18	2.620	1.970
6	2.670	2.035	19	2.615	1.985
7	2.665	2.020	20	2.615	1.990
8	2.660	2.005	21	2.615	1.995
9	2.655	2.010	22	2.610	1.990
10	2.655	2.000	23	2.590	1.975
11	2.650	2.000	24	2.590	1.995
12	2.650	2.005	25	2.565	1.955
13	2.645	2.015			

Sol. ...



...



Def. Let a and b be the least squares coefficients with the sample $(x_1, y_1), \dots, (x_n, y_n)$.

$\hat{y} = a + bx$: **predicted value** of y

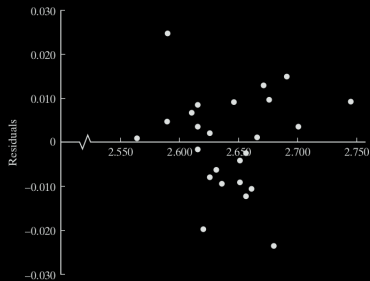
$y_i - \hat{y}_i = y_i - (a + bx_i)$: **i th residual**

Remark Use the residual plots to assessing the model.

E.g. 1' Here are the residues and their plots:

Table 11.2.2

x_i	y_i	\hat{y}_i	$y_i - \hat{y}_i$
2.745	2.080	2.070	0.010
2.700	2.045	2.041	0.004
2.690	2.050	2.035	0.015
2.680	2.005	2.029	-0.024
2.675	2.035	2.025	0.010
2.670	2.035	2.022	0.013
2.665	2.020	2.019	0.001
2.660	2.005	2.016	-0.011
2.655	2.010	2.013	-0.003
2.655	2.000	2.013	-0.013
2.650	2.000	2.009	-0.009
2.650	2.005	2.009	-0.004
2.645	2.015	2.006	0.009
2.635	1.990	2.000	-0.010
2.630	1.990	1.996	-0.006
2.625	1.995	1.993	0.002
2.625	1.985	1.993	-0.008
2.620	1.970	1.990	-0.020
2.615	1.985	1.987	-0.002
2.615	1.990	1.987	0.003
2.615	1.995	1.987	0.008
2.610	1.990	1.984	0.006
2.590	1.975	1.971	0.004
2.590	1.995	1.971	0.024
2.565	1.955	1.955	0.000



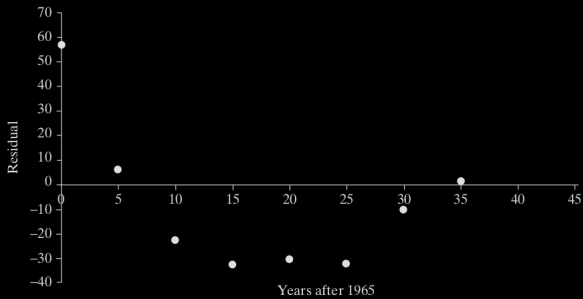
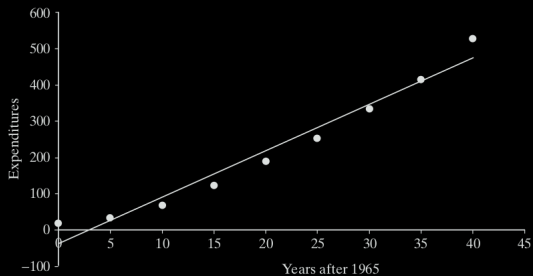
E.g. 2 Predict the Social Security expenditures.

Table 11.2.3		
Year	Years after 1965, x	Social Security Expenditures (\$ billions), y
1965	0	19.2
1970	5	33.1
1975	10	69.2
1980	15	123.6
1985	20	190.6
1990	25	253.1
1995	30	339.8
2000	35	415.1
2005	40	529.9

Source: www.socialsecurity.gov/history/trustfunds.html.

Does the the least squares line $y = -38.0 + 12.9x$ a good model to predict the cost in 2010 would be \$543, i.e., the case $x = 45$?

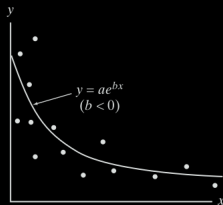
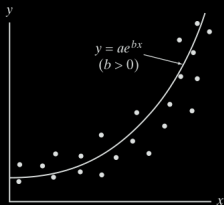
Sol.





"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Exponential Regression



$$y = ae^{bx} \iff \ln y = \ln a + bx$$

$$b = \frac{n \sum_{i=1}^n x_i \ln y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n \ln y_i)}{n (\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2}$$

$$\ln a = \frac{\sum_{i=1}^n \ln y_i - b \sum_{i=1}^n x_i}{n}$$

E.g. Moore's law:

Gordon Moore predicted in 1965 that the number of transistors per chip would double every 18 months.

Based on the real data, check:

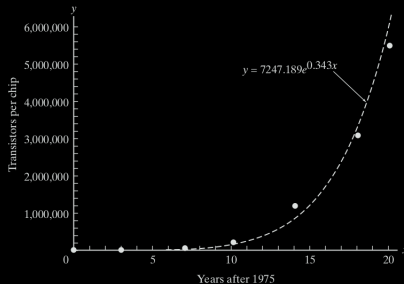
- 1) Whether is the chip capacity doubling at a fixed rate?
- 2) Find out the rate.

Table 11.2.5			
Chip	Year	Years after 1975, x	Transistors per Chip, y
8080	1975	0	4,500
8086	1978	3	29,000
80286	1982	7	90,000
80386	1985	10	229,000
80486	1989	14	1,200,000
Pentium	1993	18	3,100,000
Pentium Pro	1995	20	5,500,000
Source: en.wikipedia.org/wiki/Transistor—count.			

Sol. To check whether chip capacity doubles in a fixed rate, one needs to carry out exponential regression:

Table 11.2.6				
Years after 1975, x_i	x_i^2	Transistors per Chip, y_i	$\ln y_i$	$x_i \cdot \ln y_i$
0	0	4,500	8.41183	0
3	9	29,000	10.27505	30.82515
7	49	90,000	11.40756	79.85292
10	100	229,000	12.34148	123.41480
14	196	1,200,000	13.99783	195.96962
18	324	3,100,000	14.94691	269.04438
20	400	5,500,000	15.52026	310.40520
72	1078		86.90093	1009.51207

$$\Rightarrow b = \dots = 0.342810, \quad a = \dots = e^{\ln a} = e^{8.89} = 7247.189.$$



Finally, to find out the rate:

$$e^{0.343x} = e^{\ln 2 \times \frac{0.343}{\ln 2} x} = 2^{\frac{0.343}{\ln 2} x}$$

$$\frac{0.343}{\ln 2} x = 1 \quad \implies \quad x = \frac{\ln 2}{0.343} = 2.020837.$$



Other curvilinear models

Table 11.2.10

- a. If $y = ae^{bx}$, then $\ln y$ is linear with x .
- b. If $y = ax^b$, then $\log y$ is linear with $\log x$.
- c. If $y = L/(1 + e^{a+bx})$, then $\ln\left(\frac{L-y}{y}\right)$ is linear with x .
- d. If $y = \frac{1}{a+bx}$, then $\frac{1}{y}$ is linear with x .
- e. If $y = \frac{x}{a+bx}$, then $\frac{1}{y}$ is linear with $\frac{1}{x}$.
- f. If $y = 1 - e^{-x^b/a}$, then $\ln \ln\left(\frac{1}{1-y}\right)$ is linear with $\ln x$.