

# How to choose a wrong model

Jiří Anděl

Charles University, Prague

Bratislava, December 15, 2008

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomic  
regression

Growth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

## 1 Introduction

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomic  
regression

Growth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomic  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model
- 7 References

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model
- 7 References

Wrong model

Jiří Anděl

Outline

**Introduction**

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

- All models are wrong but some are useful. (George Box)

Wrong model

Jiří Anděl

Outline

**Introduction**

Data

Polynomic  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

- All models are wrong but some are useful. (George Box)
- A model can be no more than a good portrait. (J. J. Faraway)

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

- **All models are wrong** but some are useful. (George Box)
- A model can be no more than a good portrait. (J. J. Faraway)
- (M. J. Crawley) There is a temptation to become personally attached to a particular model. Statisticians call this ‘falling in love with your model’. Remember:
  - ① All models are wrong.
  - ② Some models are better than others.
  - ③ The correct model can never be known with certainty.
  - ④ The simpler the model, the better it is.

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

- **All models are wrong** but some are useful. (George Box)
- A model can be no more than a good portrait. (J. J. Faraway)
- (M. J. Crawley) There is a temptation to become personally attached to a particular model. Statisticians call this ‘falling in love with your model’. Remember:
  - ① All models are wrong.
  - ② Some models are better than others.
  - ③ The correct model can never be known with certainty.
  - ④ The simpler the model, the better it is.
- So far as theories of mathematics are about reality, they are not certain; so far as they are certain, they are not about reality. (A. Einstein)

- **All models are wrong** but some are useful. (George Box)
- A model can be no more than a good portrait. (J. J. Faraway)
- (M. J. Crawley) There is a temptation to become personally attached to a particular model. Statisticians call this ‘falling in love with your model’. Remember:
  - ① All models are wrong.
  - ② Some models are better than others.
  - ③ The correct model can never be known with certainty.
  - ④ The simpler the model, the better it is.
- So far as theories of mathematics are about reality, they are not certain; so far as they are certain, they are not about reality. (A. Einstein)
- It is easy to lie with statistics. It is hard to tell truth without statistics. (A. Dunkels)

## Graphs are very important in statistics.

- Every year about  $10^{12}$  statistical graphs are printed.
- Graphs enable to find laws in data.
- Everitt (2005), p. 16, cites Chambers et al. (1983): "... there is no statistical tool that is as powerful as a well-chosen graph". Co-authors of Chambers are W. S. Cleveland and P. A. Tukey.
- Carl Sagan warns: "Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent."

# Outline

- 1 Introduction
- 2 **Data**
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model
- 7 References

Wrong model

Jiří Anděl

Outline

Introduction

**Data**Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

Here we introduce data to be analyzed. It is known how the data arised. This information will be given later. For example, if you knew that the independent variable is concentration of hexametylentetramin and dependent variable is concentration of pentaerytritol (which is not the case), most people would only understand that  $x_i$  are values of independent variable and  $y_i$  values of dependent variable.

$i$	1	2	3	4	5
$x_i$	2.28	1.03	0.19	0.49	2.52
$y_i$	2.89	3.18	0.89	3.30	2.24
$i$	6	7	8	9	10
$x_i$	0.11	0.46	0.28	1.39	0.03
$y_i$	1.00	2.42	0.17	2.53	0.01

Table: Data

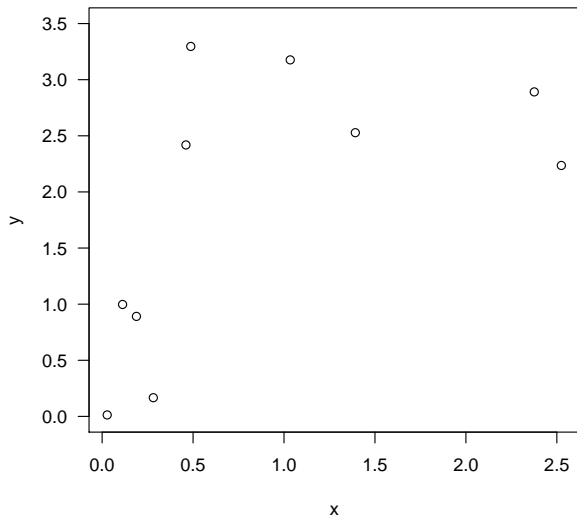


Figure: Data

## Graphical description of the data

- The boxplot (box-and-whiskers-plot): The box in the middle indicates median. The upper and the lower side of the box are hinges (nearly quartiles). It means that about 50% of the data are between hinges. The lines (whiskers) show the largest and smallest observation that falls within a distance of 1.5 times the box size from the nearest hinge. If any observations fall farther away, the additional points are considered extreme values and are shown separately.
- The bivariate boxplot (Goldberg and Iglewicz, 1992): It consists of a pair of concentric ellipses. The smaller is called hinge and includes 50% of the data. The larger is called fence and delineates potential troublesome outliers. In addition, regression lines are shown. There is a robust and a nonrobust version of the bivariate boxplot.

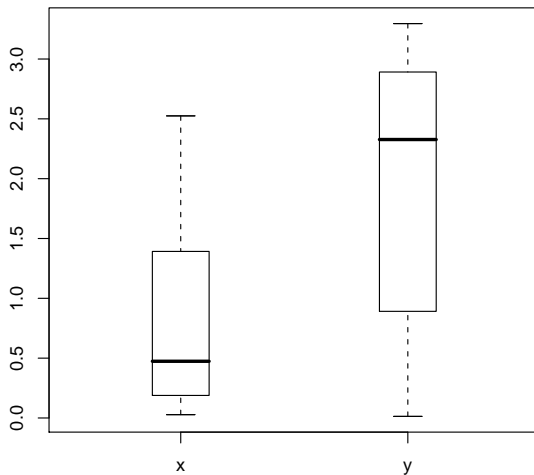


Figure: Boxplot

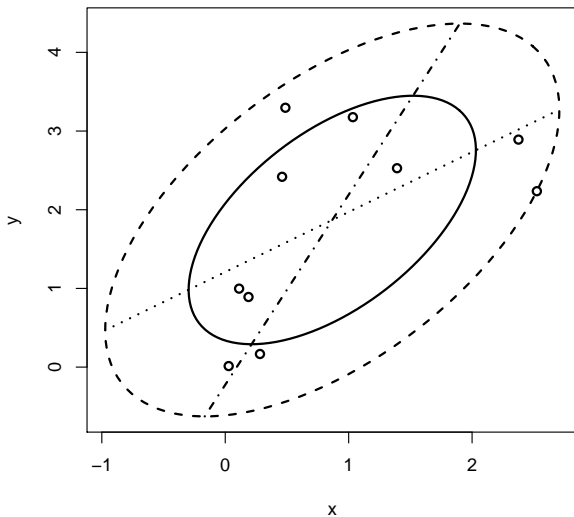


Figure: Robust bivariate boxplot

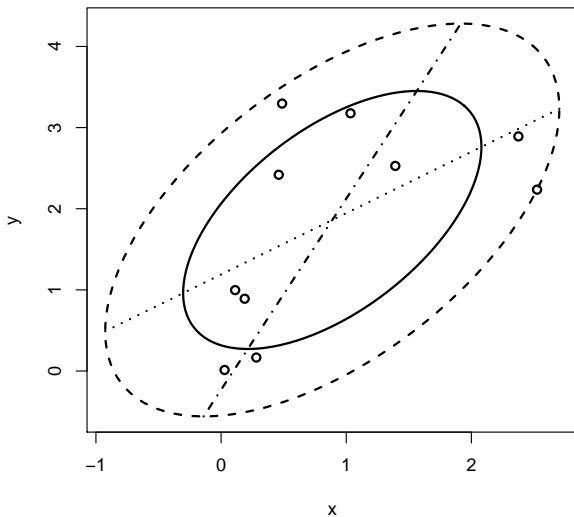


Figure: Nonrobust bivariate boxplot

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomic regression**
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model
- 7 References

Wrong model

Jiří Anděl

Outline

Introduction

Data

**Polynomic  
regression**Growth  
functions

Explanation

An application  
of a wrong  
model

References

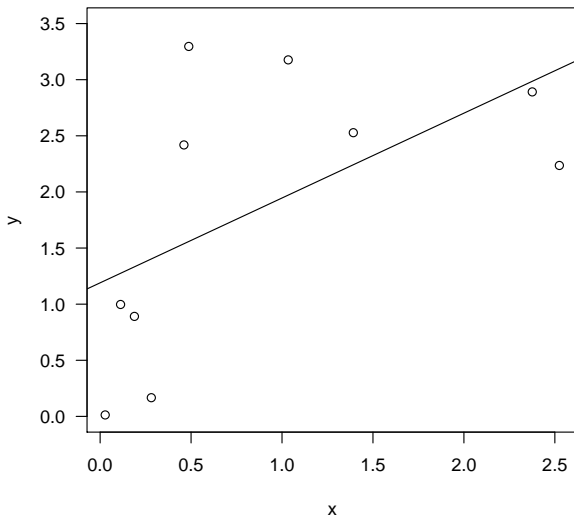


Figure: Linear regression

	Estimate	Std. Error	t value	$Pr(>  t )$
(Interc.)	1.1900	0.4855	2.451	0.0399 *
x	0.7558	0.3886	1.945	0.0877 .

Table: Linear regression

Residual standard error: 1.079 on 8 degrees of freedom

Multiple R-Squared: 0.321, Adjusted R-squared: 0.2361

F-statistic: 3.782 on 1 and 8 DF, p-value: 0.08772

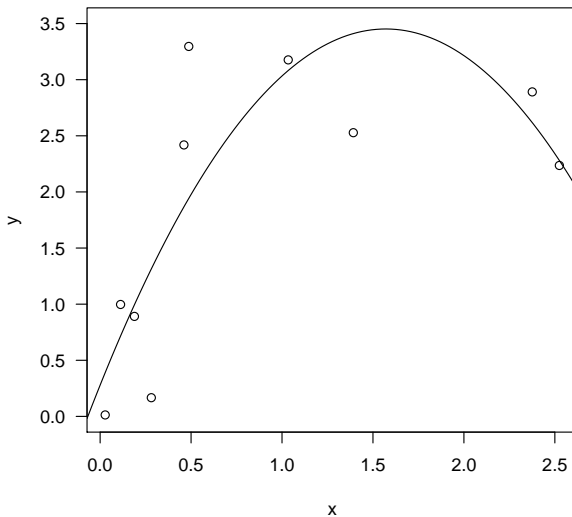


Figure: Quadratic regression

	Estimate	Std. Error	t value	$Pr(>  t )$
(Interc.)	0.2786	0.4948	0.563	0.5910
x	4.0411	1.24756	3.239	0.0143 *
$I(x^2)$	-1.2865	0.4751	-2.708	0.0303 *

Table: Quadratic regression

Residual standard error: 0.8062 on 7 degrees of freedom

Multiple R-Squared: 0.6684, Adjusted R-squared: 0.5736

F-statistic: 7.054 on 2 and 7 DF, p-value: 0.02100

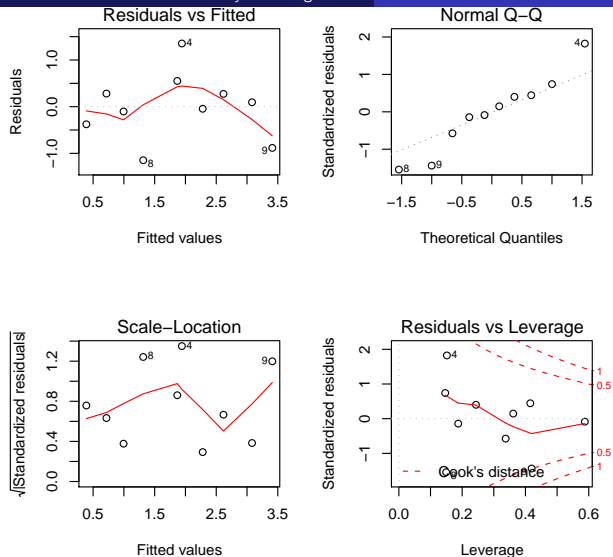


Figure: Diagnostic graphs to quadratic regression

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial regression

Growth functions

Explanation

An application of a wrong model

References

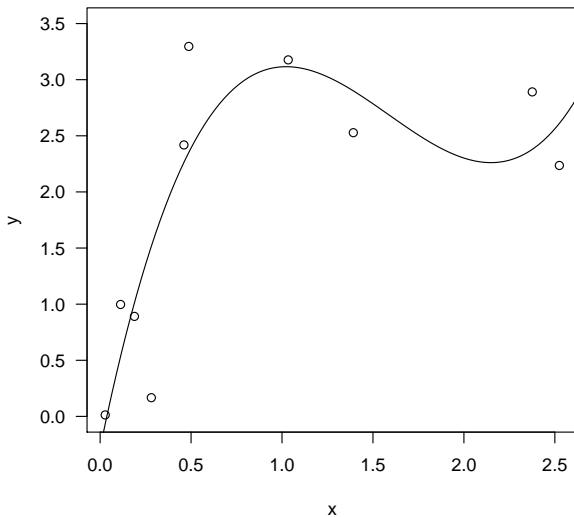


Figure: Cubic regression

In cubic regression is significant only the linear term. Neither the quadratic nor the cubic terms are significant.

Wrong model

Jiří Anděl

Outline

Introduction

Data

**Polynomic  
regression**Growth  
functions

Explanation

An application  
of a wrong  
model

References

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions**
- 5 Explanation
- 6 An application of a wrong model
- 7 References

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regression**Growth  
functions**

Explanation

An application  
of a wrong  
model

References

## Linear-plateau regression function

$$lp(x, a, b, p, c) = a - bp \ln \left[ 1 + \exp \left\{ \frac{c - x}{p} \right\} \right],$$

- a* value of dependent variable at the change point
- b* slope of the growing line
- c* value of indep. variable at the change point
- p* smoothness between both lines

[Table](#): Interpretation of parameters

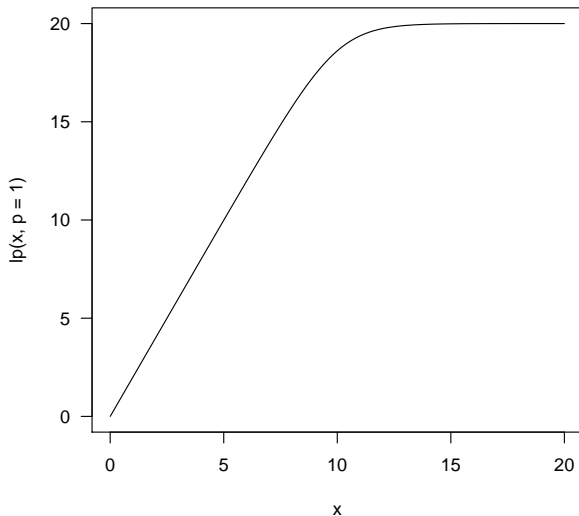


Figure: Parameters  $a = 20$ ,  $b = 2$ ,  $c = 10$ ,  $p = 1$

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomic  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

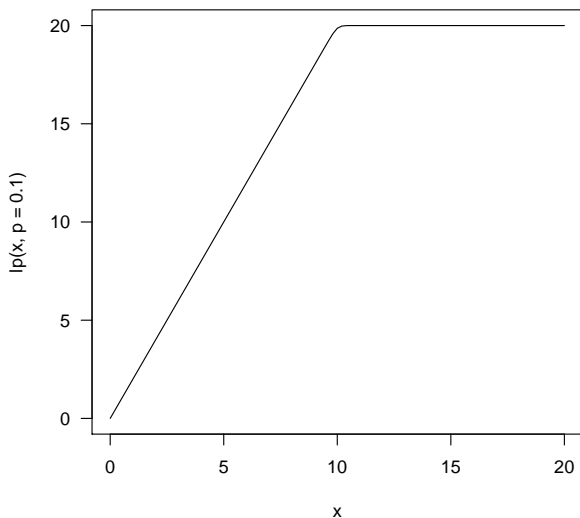


Figure: Parameters  $a = 20, b = 2, c = 10, p = 0,1$

Wrong model

Jiří Anděl

Outline

Introduction

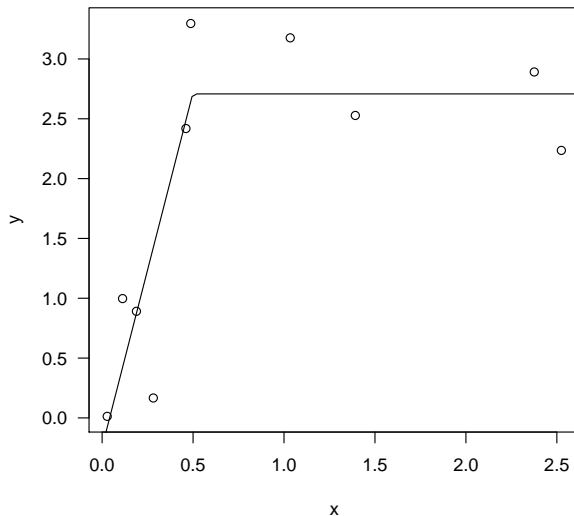
Data

Polynomic  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References



Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomic  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

Figure: Our linear-plateau regression function  
 $a = 2.708$ ,  $b = 5.928$ ,  $c = 0.498$ ,  $p = 0.0007$

## Michaelis-Menten curve

$$gmm(x) = \frac{V_m * x}{K + x}$$

- $V_m$  is a numeric parameter representing the maximum value of the response
- $K$  is a numeric parameter representing the input value at which half the maximum response is attained. In the field of enzyme kinetics this is called the Michaelis parameter.

Estimates:

$$\widehat{V_m} = 3.2750362, \widehat{K} = 0.3046473,$$

$p$ -values are 0.00172 \*\* and 0.21298, respectively, residual sum-of-squares: 4.943938

## Biexponential curve

$$gbe(x) = A1 * \exp(-\exp(lrc1) * x) + A2 * \exp(-\exp(lrc2) * x)$$

Estimates:

$$\widehat{A1} = -5.441, \widehat{lrc1} = 0.649, \widehat{A2} = 5.110, \widehat{lrc2} = -1.246,$$

all  $p$ -values are about 0.5, residual sum-of-squares: 3.59276

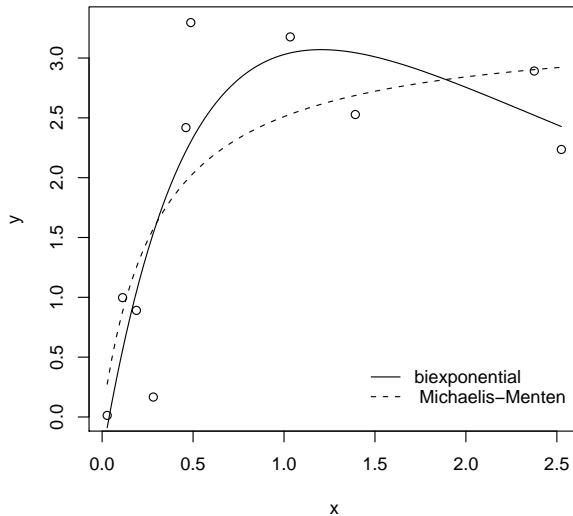


Figure: Michaelis-Menten and biexponential growth functions

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation**
- 6 An application of a wrong model
- 7 References

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions**Explanation**An application  
of a wrong  
model

References

Data were generated on computer such that  $x_i \sim N(1, 1)$ ,  $y_i \sim N(2, 1)$  and all data were independent. We used program R with `set.seed(1203)`. This constant is used by Everitt (2005).

Why we derived a wrong model? Possible reasons:

- Generator is not good enough.
- Error of the first kind occurs with probability 0.05 (this may be our case).
- Our hypothesis about quadratic regression function were formulated after seeing data.

I think that it is the third reason which lead to wrong model. The data mining and consequent statistical analysis should be used on different sets of data.

Remember: “Humans are good at discerning subtle patterns that are really there, but equally so at **imagining them when they are altogether absent.**”

- Prediction is a tricky business — perhaps the only thing worse than a prediction is no prediction at all. (J. J. Faraway)

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

- Prediction is a tricky business — perhaps the only thing worse than a prediction is no prediction at all. (J. J. Faraway)
- The good Christian should beware of mathematicians and all those who make empty prophecies. The danger already exists that mathematicians have made a covenant with the devil to darken the spirit and confine man in the bonds of Hell. (St. Augustine — *Aurelius Augustinus* 354 – 430)

- Prediction is a tricky business — perhaps the only thing worse than a prediction is no prediction at all. (J. J. Faraway)
- The good Christian should beware of mathematicians and all those who make empty prophecies. The danger already exists that mathematicians have made a covenant with the devil to darken the spirit and confine man in the bonds of Hell. (St. Augustine — *Aurelius Augustinus* 354 – 430)
- Anyone who uses arithmetic methods to produce random numbers is in a state of sin. (John von Neumann)

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model**
- 7 References

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

In the next picture we present famous data. To help to identify them, we inform that the first three observations were made in 1968 and the last two in 2008.

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomic  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

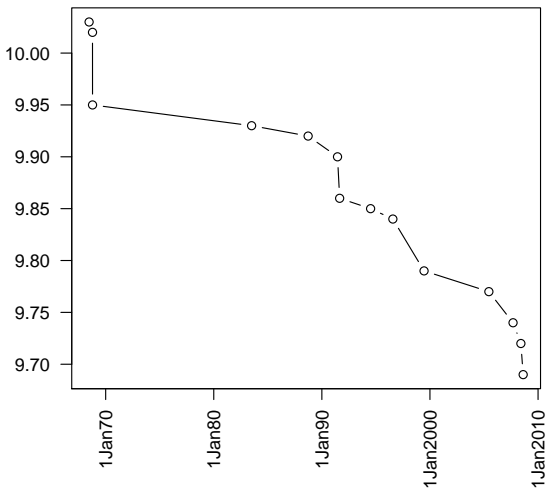


Figure: Data

I believe all of you recognized that these are records 100 m men. Here is the corresponding table.

date	record	runner
20.6.1968	10.03	J. Hines (USA)
13.10.1968	10.02	Ch. Greene (USA)
14.10.1968	9.95	J. Hines (USA)
3.7.1983	9.93	C. Smith (USA)
24.9.1988	9.92	C. Lewis (USA)
14.6.1991	9.90	L. Burell (USA)
25.8.1991	9.86	C. Lewis (USA)
6.7.1994	9.85	L. Burell (USA)
27.7.1996	9.84	D. Bailey (Can)
16.6.1999	9.79	M. Greene (USA)
14.6.2005	9.77	A. Powell (Jam)
9.9.2007	9.74	A. Powell (Jam)
31.5.2008	9.72	U. Bolt (Jam)
16.8.2008	9.69	U. Bolt (Jam)

A frequent question raised in newspapers has been: “Where are the limits of men’s speed?” Is it possible to draw a reasonable curve going through our points which should show the asymptotics? Our observations correspond to a nondecreasing curve. Since we are accustomed to use sigmoidal curves, we transform the records  $y_t$  to  $12 - y_t$ . The constant 12 is quite arbitrary and we get rid of it immediately.

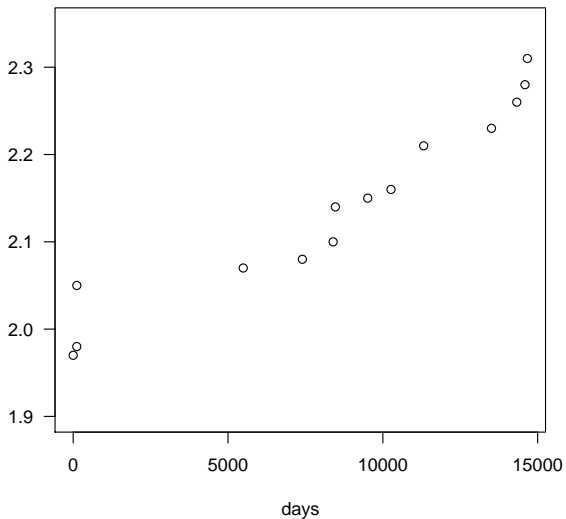


Figure: Transformed data:  $12 - y_t$

Since our constant 12 was arbitrary, we use the four-parameter logistic function

$$y = A + \frac{B - A}{1 + \exp\left(\frac{xmid - x}{scal}\right)}.$$

The estimates of the parameters are

$$A = 1.947, \quad B = 2.660, \quad xmid = 15301, \quad scal = 5998.$$

This function is drawn in the following picture.

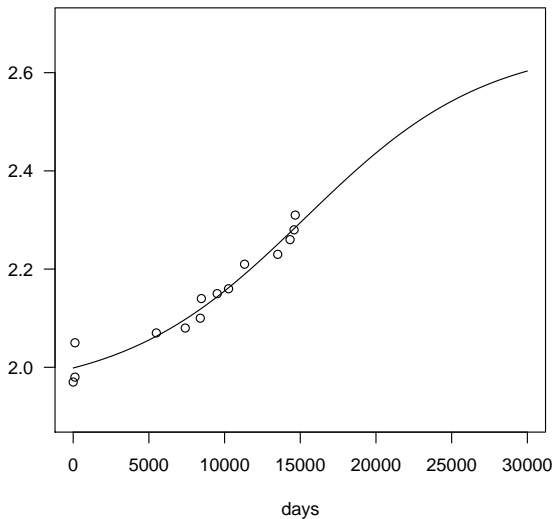


Figure: Four-parameter logistic curve

Finally, we return back to the original data. The asymptotic value is  $12 - B = 9.34$ . However, it will last some 40 years to reach record 9.4. Even if our model is completely wrong, we can easily risk that somebody will accuse us after 40 years that our extrapolation was not good.

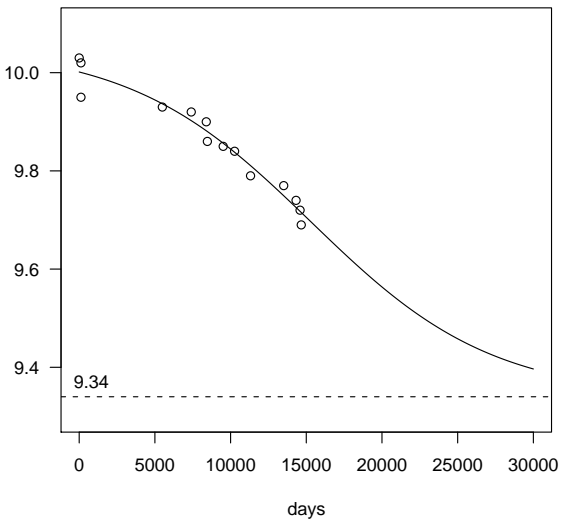


Figure: Result

# Outline

- 1 Introduction
- 2 Data
- 3 Polynomial regression
- 4 Growth functions
- 5 Explanation
- 6 An application of a wrong model
- 7 References**

Wrong model

Jiří Anděl

Outline

Introduction

Data

Polynomial  
regressionGrowth  
functions

Explanation

An application  
of a wrong  
model

References

## References

Anděl J. (2008): A choice of a regression model (in Czech - Volba regresního modelu). *Information Bulletin of the Czech Statistical Society* **19**, No. 1, 5–13.

Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A. (1983): *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Everitt B. (2005): *An R and S-PLUS Companion to Multivariate Analysis*. Springer-Verlag, London.

Faraway J. J. (2000): *Practical Regression and ANOVA using R*. (PDF file)

Goldberg K. M., Iglewicz B. (1992): Bivariate extensions of the boxplot. *Technometrics* **34**, 307–320.