Základy Regrese | (NMFM 334)

Letný semester 2024 | Prednáška 1

Matúš Maciak
Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University, Prague

Why? What is (linear) regression?

"When a numerical criterion variable is to be predicted from other numerical predictor variables, proper (linear/regression) models outperform (human) intuition."

Paul Meehl (1954)

Clinical versus statistical prediction: A theoretical analysis and a Review of the Evidence

Why? What is (linear) regression?

"When a numerical criterion variable is to be predicted from other numerical predictor variables, proper (linear/regression) models outperform (human) intuition."

Paul Meehl (1954)

Clinical versus statistical prediction: A theoretical analysis and a Review of the Evidence

dynamic, global, sensible, advanced, delicate, holistic, nice, rich, pure, configural, organized, sophiticated, natural, realistic, understandable, exemplary, vital;

mechanical, local, dashed, too simple, unreal, artificial, random, incomplete, trivial, pedant, trivial, static, forced, shallow, academic, scientific, blind;

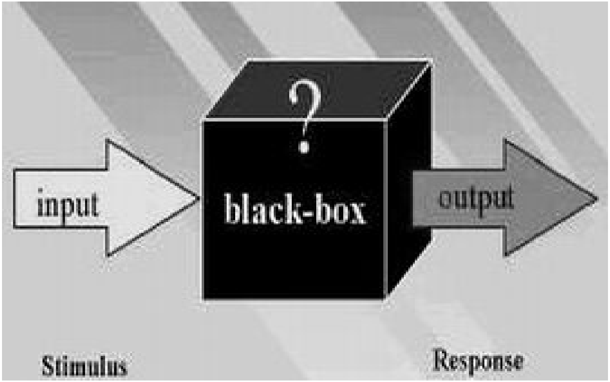
Regression (models) applied all around us ...

Regression models applied in practice

- **Black boxes:** `lm()`, `PROC REG`, `XLSTAT`, `LinearModel.fit()`;
  https://en.wikipedia.org/wiki/List_of_statistical_packages

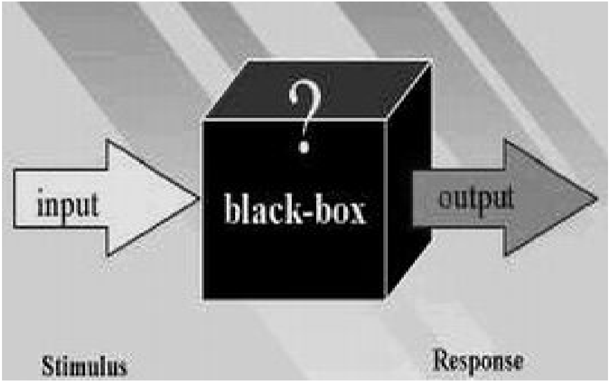
Regression models applied in practice

- **Black boxes:** `lm()`, `PROC REG`, `XLSTAT`, `LinearModel.fit()`; `https://en.wikipedia.org/wiki/List_of_statistical_packages`



Regression models applied in practice

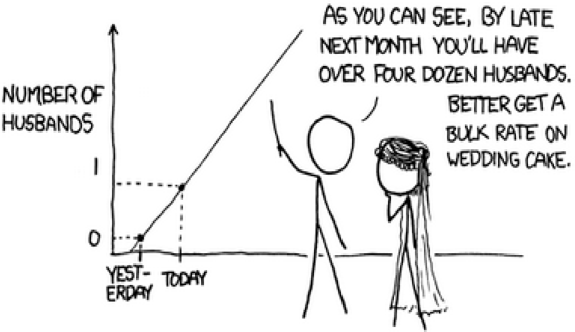
- **Black boxes:** `lm()`, `PROC REG`, `XLSTAT`, `LinearModel.fit()`;
  https://en.wikipedia.org/wiki/List_of_statistical_packages



- **Inside of the black box** there is a complex and quite sophisticated mathematical and statistical theory which makes the output reliable and useful **if and only if the input data suits the theory** in the box.

Regression models applied in practice

- **Black boxes:** `lm()`, `PROC REG`, `XLSTAT`, `LinearModel.fit()`;
  https://en.wikipedia.org/wiki/List_of_statistical_packages



- **Inside of the black box** there is a complex and quite sophisticated mathematical and statistical theory which makes the output reliable and useful **if and only if the input data suits the theory** in the box.

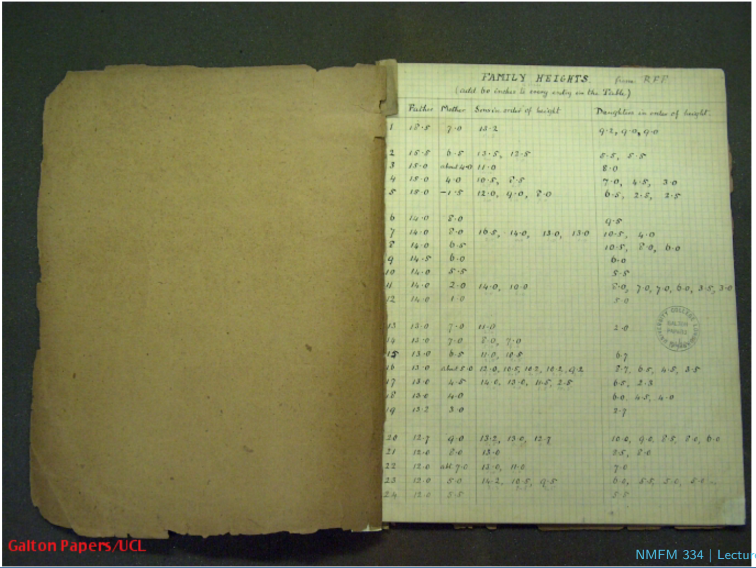
Outline

1. **Motivation & some historical background**
   Somehow, it was all a little bit different at the beginning...
   A brief look into the historical backgrounds of the regression.

2. **Basic principles of the theoretical background**
   All we need in regression is conveniently concentrated in three main pivots: cognition, calibration, and prediction.

3. **Common problems when fitting a regression model**
   What can actually go wrong at the end? A few examples of incorrect applications of the linear regression framework.

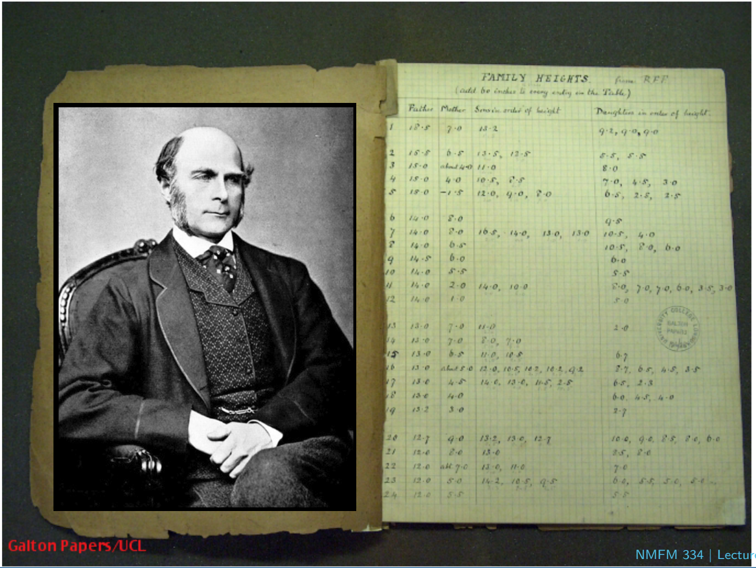
Regression: At the very beginning ...



FAMILY HEIGHTS from RFF
(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height
1	18.5	7.0	13.2	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 12.5	5.5, 5.5
3	15.0	about 4.0	11.0	8.0
4	15.0	4.0	10.5, 8.5	7.0, 4.5, 3.0
5	15.0	−1.5	12.0, 9.0, 8.0	6.5, 2.5, 2.5
6	14.0	8.0		9.5
7	14.0	8.0	16.5, 14.0, 13.0, 13.0	10.5, 4.0
8	14.0	6.5		10.5, 8.0, 6.0
9	14.5	6.0		6.0
10	14.0	5.5		5.5
11	14.0	2.0	14.0, 10.0	8.0, 7.0, 7.0, 6.0, 3.5, 3.0
12	14.0	1.0		5.0
13	13.0	7.0	11.0	2.0
14	13.0	7.0	8.0, 7.0	
15	13.0	6.5	11.0, 10.5	6.7
16	13.0	about 5.0	12.0, 10.5, 10.2, 10.2, 9.2	8.7, 6.5, 4.5, 3.5
17	13.0	4.5	14.0, 13.0, 11.5, 2.5	6.5, 2.3
18	13.0	4.0		6.0, 4.5, 4.0
19	13.2	3.0		2.7
20	12.7	9.0	13.2, 13.0, 12.7	10.0, 9.0, 8.5, 8.0, 6.0
21	12.0	8.0	13.0	8.5, 8.0
22	12.0	abt 7.0	13.0, 11.0	7.0
23	12.0	5.0	14.2, 10.5, 9.5	6.0, 5.5, 5.0, 5.0
24	12.0	5.5		5.5

Regression: At the very beginning ...



Regression: Pioneer Francis Galton

The British Association for the Advancement of Science

Presidential address (1885): "Regression toward mediocrity in hereditary stature"

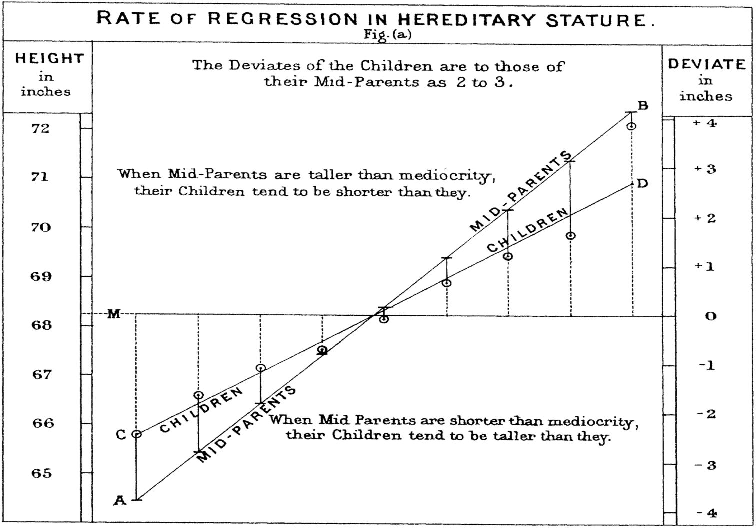
TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

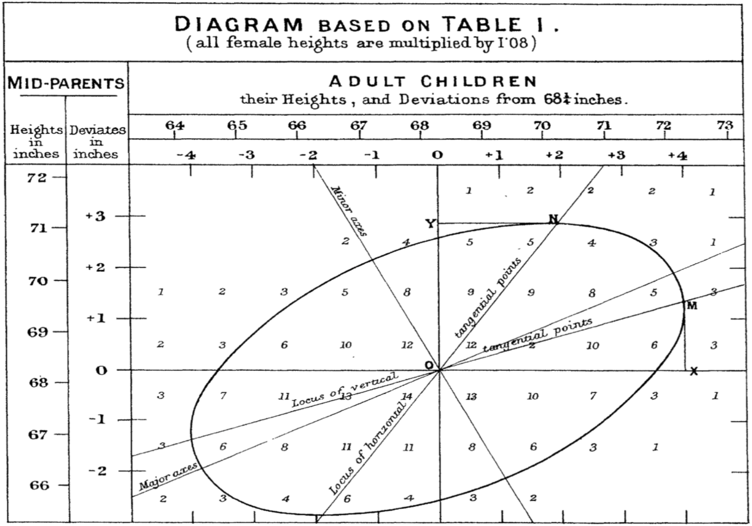
Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Regression: Regressing towards mediocrity



Regression: Dependent vs. independent



Regression: General concept

- An accidental word invented by Francis Galton (1822 – 1911) because the heights of sons, while following the tendency of their parents (tall parents have tall sons, small parents small sons), tend to return – “regress” – towards the mediocrity/median/average (population stability).
- Nowadays, ”regression” is understood as a technique for fitting functional relationships (not necessarily linear) to data (regardless of whether the slope is less or greater than 1).
- Some sources understand regression as a study of the mean (expectation) conditionally on predictors. Our understanding is broader – beyond conditional expectations, and beyond least squares.

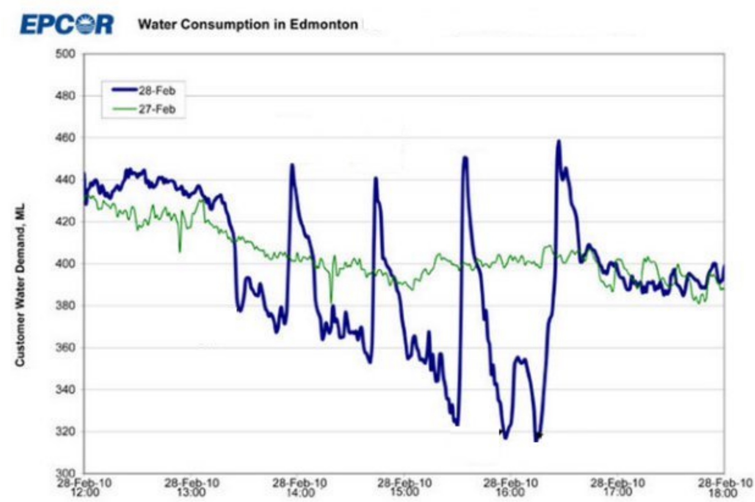
- **The primary goal of regression is to understand, as far as possible with the available data, how the conditional distribution of the response varies across subpopulations determined by possible values of the predictor(s)** (repeating observations under different conditions).

  (R. D. Cook and S. Weisberg, Applied Regression Including Computing and Graphics, p. 27)

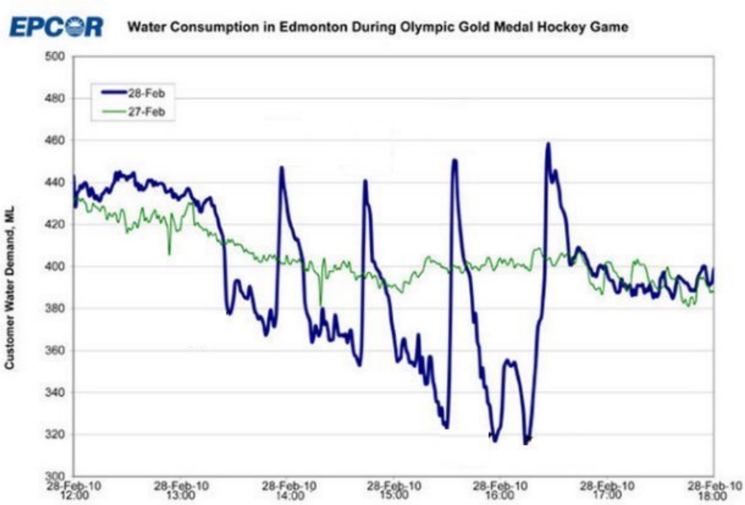
Regression: Three main tasks of regression

- **Cognition – understanding the given data**
  - What data actually is? What is the nature of data?
  - How data is collected and represented?
  - How data is connected/shared/stored/integrated?
- **Calibration – quantification of the relationship**
  - What is our believe about the underlying data structure?
  - What methodology should be applied to access the information in data?
  - Which (regression) model is suitable for the data generation?
- **Prediction/forecasting future observations**
  - Can the model be utilized for prediction/forecast?
  - What is the model potential in prediction/forecast?
  - What is the reliability of the prediction/forecast?

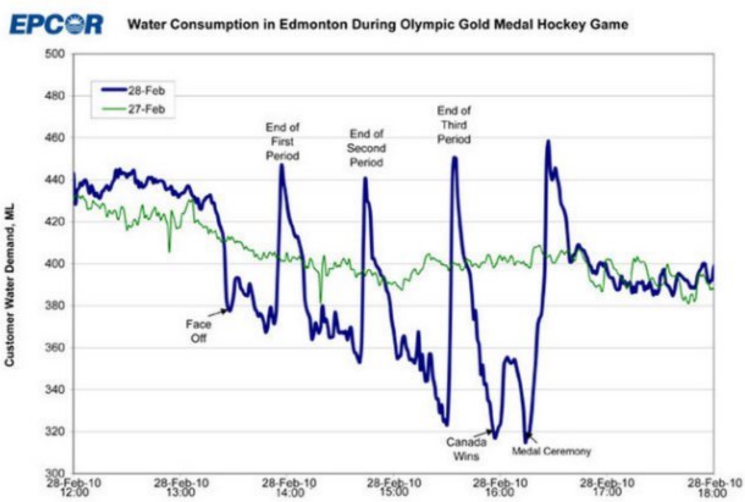
1. Cognition: Understanding the data



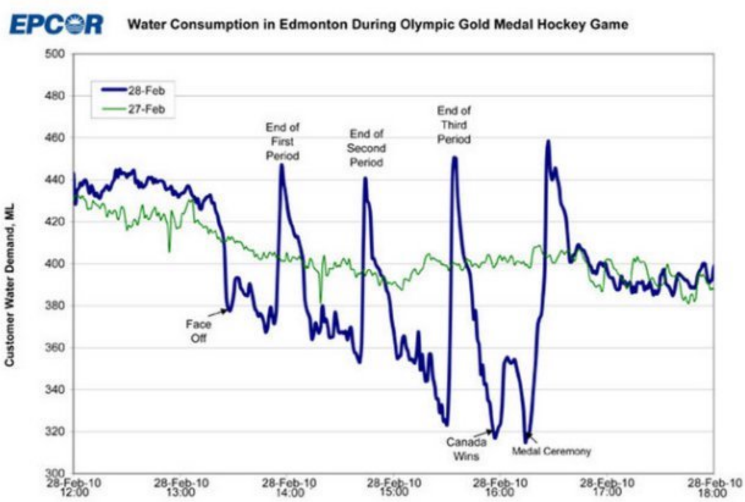
1. Cognition: Understanding the data



1. Cognition: Understanding the data



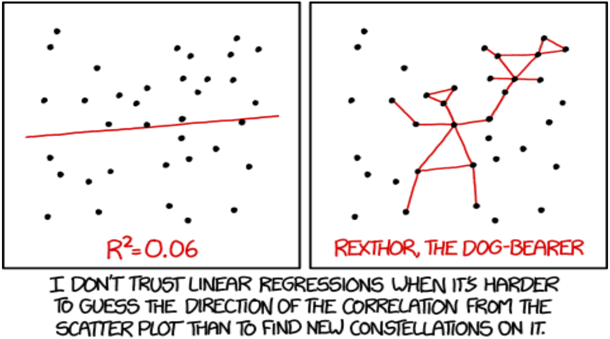
1. Cognition: Understanding the data



Canada : USA (3:2) ot | Rogers Arena, Vancouver, BC

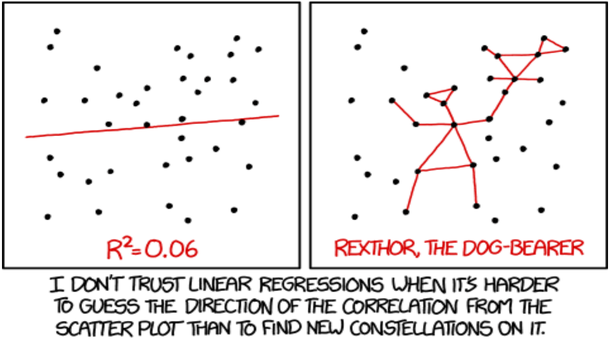
2. Calibration: Model specification (linearity)

- **Linear regression model: Where the linearity comes from?**



2. Calibration: Model specification (linearity)

- **Linear regression model: Where the linearity comes from?**

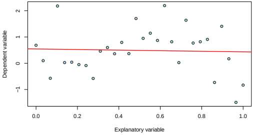


- Regression is about fitting functional relationships within the data, not geometric objects (not "fitting a line" through data).
- There is a lot of geometry in regression, but of a high-dimensional nature. (projections within $\mathbb{R}^n$ dimensional linear space into a finite dimensional subspace)

2. Calibration: Model specification (parametric structure)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

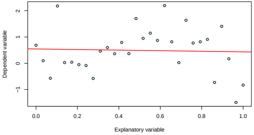
$$E[Y|x] = \beta_0 + \beta_1 x$$

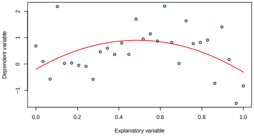


2. Calibration: Model specification (parametric structure)

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x \qquad E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

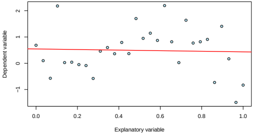


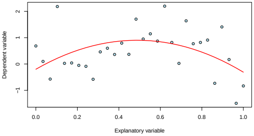


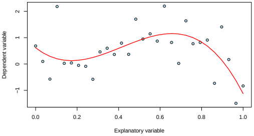
2. Calibration: Model specification (parametric structure)

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \qquad Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x \qquad E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 \qquad E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



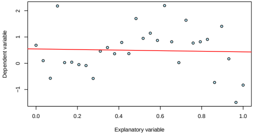




2. Calibration: Model specification (parametric structure)

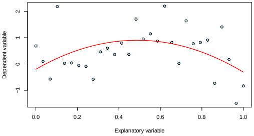
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x$$



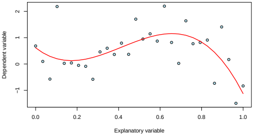
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$



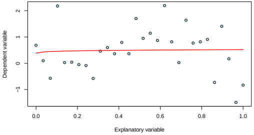
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

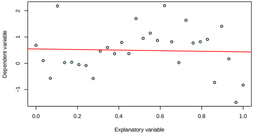
$$E[Y|x] = \beta_0 + \beta_1 \log(x)$$



2. Calibration: Model specification (parametric structure)

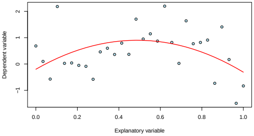
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x$$



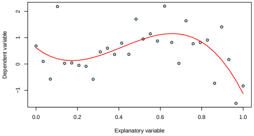
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$



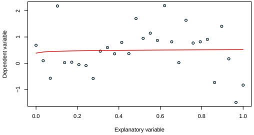
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



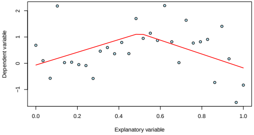
$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 \log(x)$$



$$Y = \beta_0 + \beta_1 X + \beta_2 (X - 0.5)_+ + \varepsilon$$

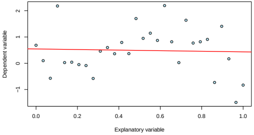
$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 (x - 0.5)_+$$



2. Calibration: Model specification (parametric structure)

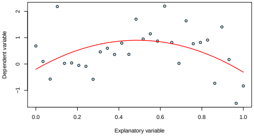
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x$$



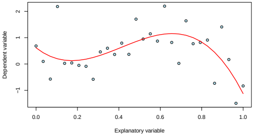
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$



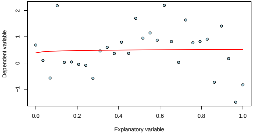
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



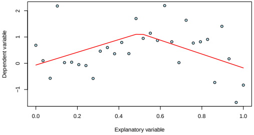
$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 \log(x)$$



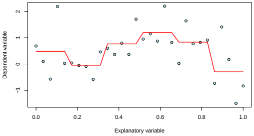
$$Y = \beta_0 + \beta_1 X + \beta_2 (X - 0.5)_+ + \varepsilon$$

$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 (x - 0.5)_+$$



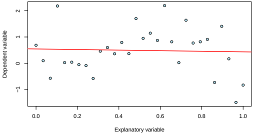
$$Y = \beta_0 + \sum_{i=1}^{5} \beta_i \mathbb{I}_{(\xi_1, \xi_{i+1})}(X) + \varepsilon$$

$$E[Y|x] = \beta_0 + \sum_{i=1}^{5} \beta_i \mathbb{I}_{(\xi_1, \xi_{i+1})}(x)$$

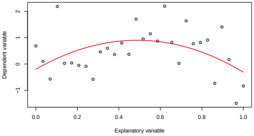


2. Calibration: Model specification (parametric structure)

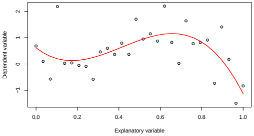
$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$E[Y|x] = \beta_0 + \beta_1 x$$



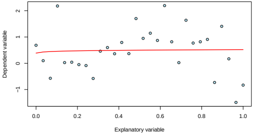
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$
$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$



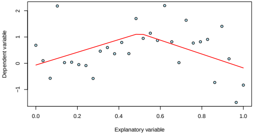
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$
$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$



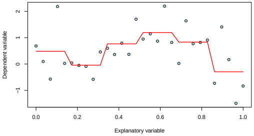
$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$
$$E[Y|x] = \beta_0 + \beta_1 \log(x)$$



$$Y = \beta_0 + \beta_1 X + \beta_2 (X - 0.5)_+ + \varepsilon$$
$$E[Y|x] = \beta_0 + \beta_1 x + \beta_2 (x - 0.5)_+$$

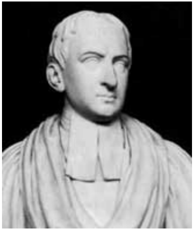


$$Y = \beta_0 + \sum_{i=1}^{5} \beta_i \mathbb{I}_{(\xi_1, \xi_{i+1})}(X) + \varepsilon$$
$$E[Y|x] = \beta_0 + \sum_{i=1}^{5} \beta_i \mathbb{I}_{(\xi_1, \xi_{i+1})}(x)$$



- ❑ Infinitely many options how to define the underlying (parametric) structure of the linear regression model using the given data points only;

2. Calibration: Pioneers before least squares



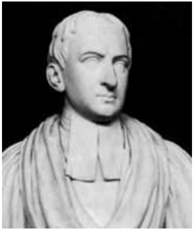
o Roger Cotes (1682 – 1716)

o Tobias Mayer (1723 – 1762)

o Roger Joseph Boscovich (1711 – 1787)

o Pierre-Simon Laplace (1749–1827)

2. Calibration: Pioneers before least squares



o Roger Cotes (1682 – 1716)

o Tobias Mayer (1723 – 1762)

o Roger Joseph Boscovich (1711 – 1787)

o Pierre-Simon Laplace (1749–1827)

- 1722 – combination of different observations taken under the same conditions instead of trying one's best to observe a single observation accurately (method of averages);
- 1750 – studying the librations of the moon in 1750 by Tobias Mayer and exploring the motion of Jupiter and Saturn by Laplace;
- 1757 – combination of different observations taken under different conditions to study the shape of the earth by Boscovich (least absolute deviations);
- 1799 – combination of the method with a symmetric two-sided exponential distribution by Laplace for studying the same problem as Boscovich (discovering median instead of average);

2. Calibration: Model estimation approaches

- **Method of averages** – multiple observations of the same event observed with random error rather than just one precise measurement;
- **Least absolute deviation** – ancient method developed by Roger Joseph Boscovich in 1757 (about 50 years before the least squares);
- **Least squares** – developed in 19th century (Legendre in 1805 and Gauss in 1809) for describing the behavior of celestial bodies used for astronomy, ships' navigation, and geodesy – connection with the normal distribution;
- **Maximum likelihood** – first ideas by Bernoulli in 1713 for analyzing Bernoulli trials, however, its widespread use arose between 1912 and 1922 due to Ronald Fisher;
- **Robust estimation** – estimation approach less sensitive to outlying observations, developed by Huber in 1964;
- **Other methods** – for instance, based on different risk assessment, atomic pursuit estimation and sparsity, non-convex problems;

Calibration by the method of least squares

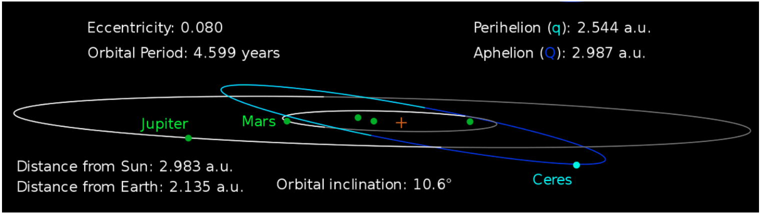
Adrien-Marie Legendre (1752 – 1833)

Johann Carl Friedrich Gauss (1777 – 1855)

- **Legendre** used the technique for fitting linear equations to data while demonstrating the new method by analyzing the same data as Laplace for the shape of the earth. The method is described as an algebraic procedure.
- **Gauss** claimed to know the method since 1795. He connected the method of least squares with the principles of the theory of probability and defined the estimation method that minimizes the error – normal distribution.

Proving the least squares: Ceres rediscovery

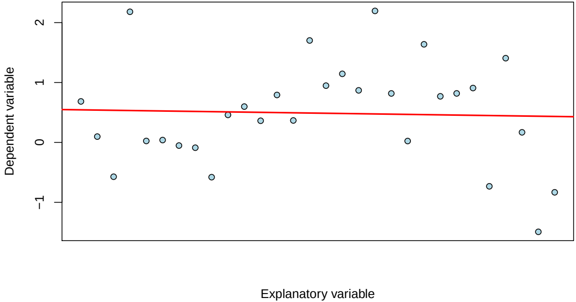
- Italian astronomer Giuseppe Piazzi discovered Ceres on 1st January 1801 and followed it for 40 days before it was lost in the glare of the sun – until the last observation (out of 24) taken on 11 February 1801.
- Given the data, astronomers desired to determine the location of Ceres after it emerged from behind the sun without solving Kepler's complicated nonlinear equations of planetary motion.
- Using the information published in *Monatliche Correspondenz* in September 1801, J.C.F.Gauss (24 years old at that time) was the only one to successfully predicted the Ceres position.
- Hungarian astronomer Heinrich W. M. Olbers found Ceres at the predicted location on 31st December 1801.



Calibration by the method of least squares

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

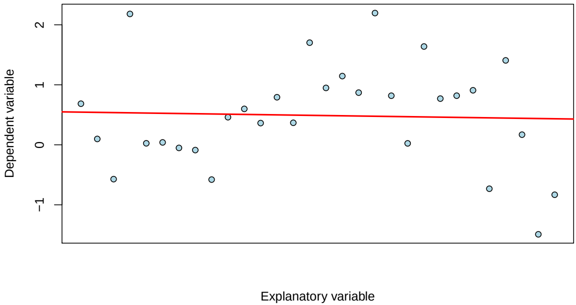
$$E[Y|x] = \beta_0 + \beta_1 x$$



Calibration by the method of least squares

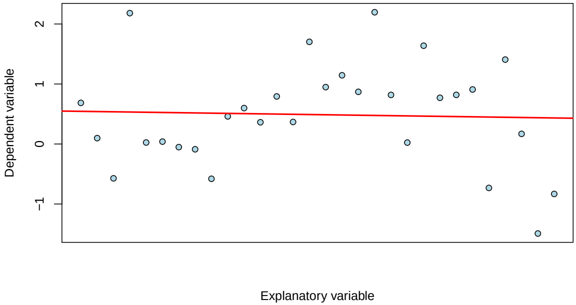
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\hookrightarrow \text{for all } i = 1, \dots, n$$



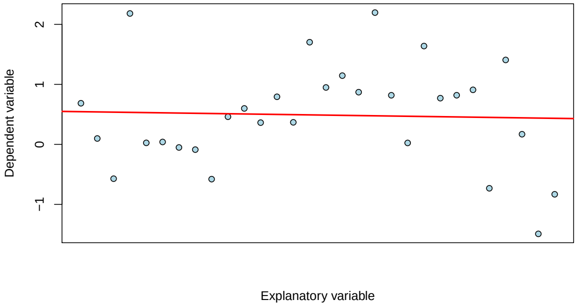
Calibration by the method of least squares

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$



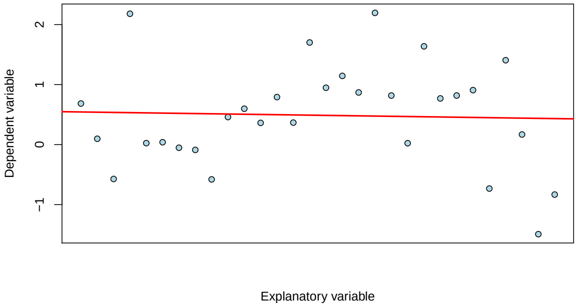
Calibration by the method of least squares

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



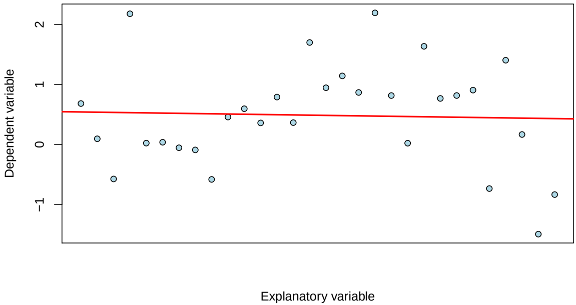
Calibration by the method of least squares

$$\boldsymbol{Y} = model + \varepsilon$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \beta \in \mathbb{R}^2$$



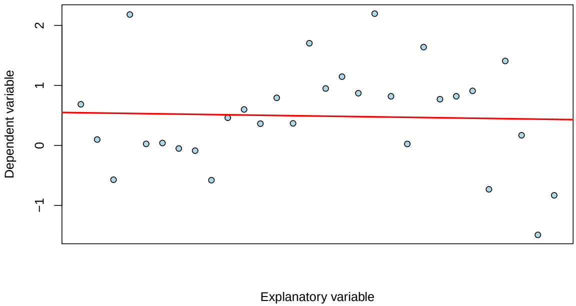
Calibration by the method of least squares

$$\boldsymbol{Y} = model + error$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \beta \in \mathbb{R}^2$$



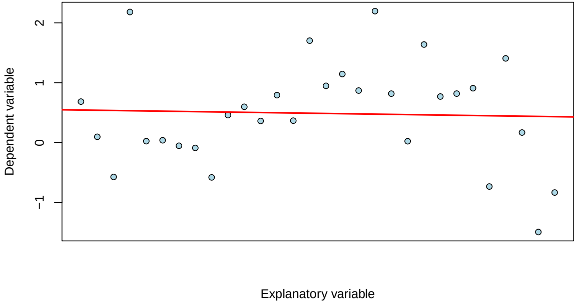
Calibration by the method of least squares

$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



Calibration by the method of least squares

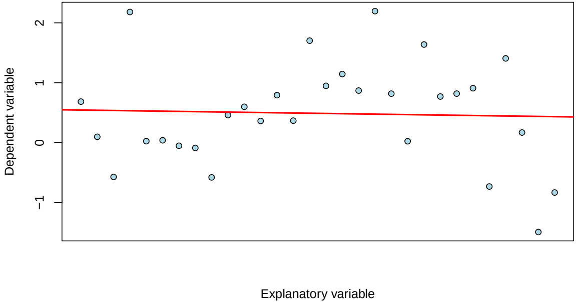
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \beta \in \mathbb{R}^2$$



The errors $(\mathbb{I} - \mathbb{P})\boldsymbol{Y}$ should be minimal in some sense!

Calibration by the method of least squares

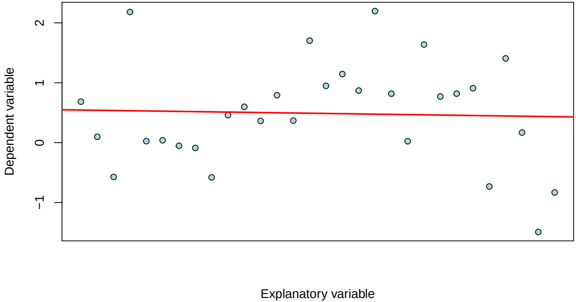
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \beta \in \mathbb{R}^2$$



The errors $\boldsymbol{Y} - \mathbb{P}\boldsymbol{Y}$ should be minimal in some sense!

Calibration by the method of least squares

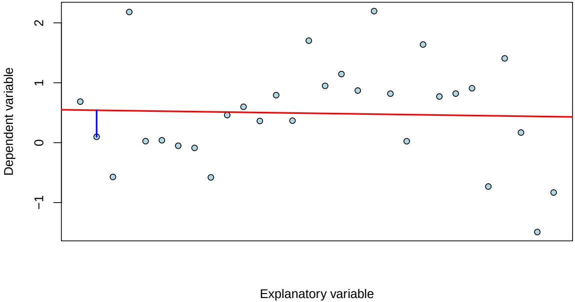
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $Y_i - (\beta_0 + \beta_1 X_i)$ should be minimal in some sense!
(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

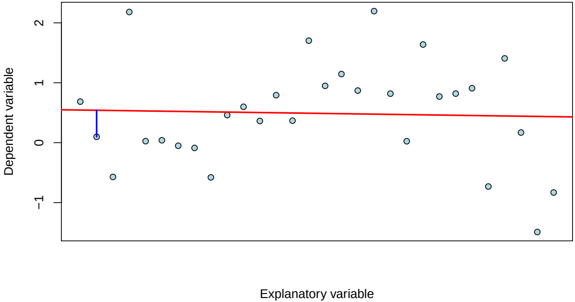
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $Y_i - (\beta_0 + \beta_1 X_i)$ should be minimal in some sense!
(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

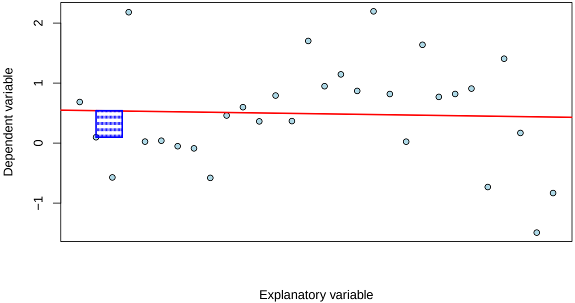
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $Y_i - (\beta_0 + \beta_1 X_i)$ should be minimal in some sense!
(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

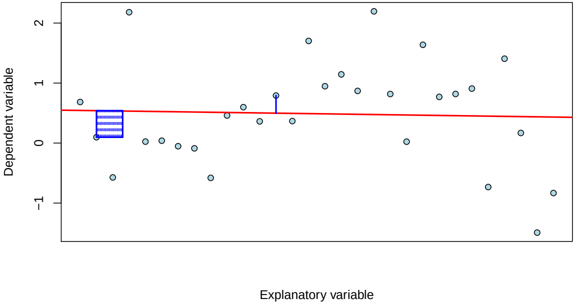
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $\left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$ should be minimal in some sense!
(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

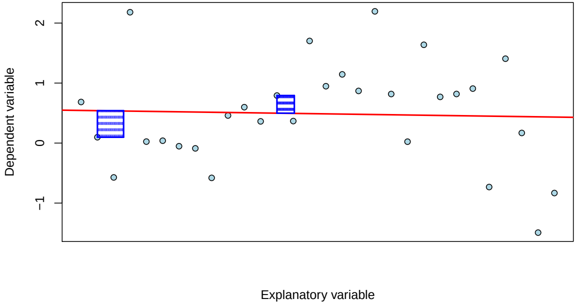
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $\Big[Y_i - (\beta_0 + \beta_1 X_i)\Big]^2$ should be minimal in some sense!
(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

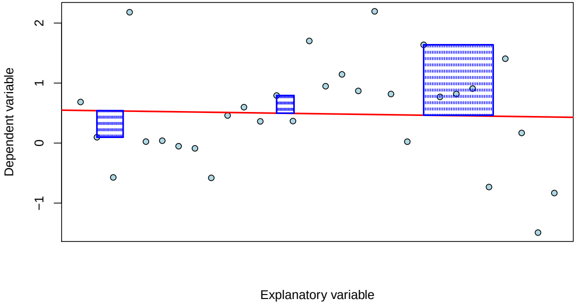
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $\left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$ should be minimal in some sense!
(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$

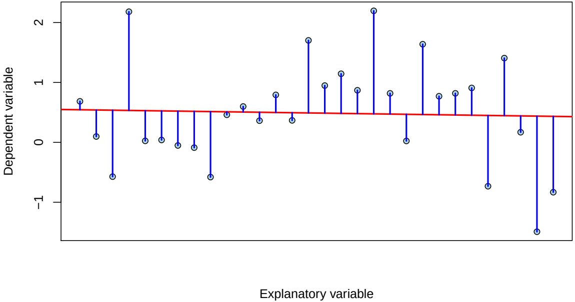


The errors $\left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$ should be minimal in some sense!

(for all indexes $i = 1, \ldots, n$)

Calibration by the method of least squares

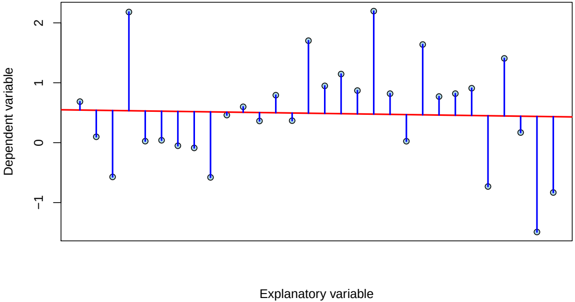
$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^2$$



The errors $\sum_{i=1}^{n} \Big[Y_i - (\beta_0 + \beta_1 X_i) \Big]^2$ should be minimal in some sense!
(for all indexes $i = 1, \dots, n$)

Calibration by the method of least squares

$$\boldsymbol{Y} = \mathbb{P}\boldsymbol{Y} + (\mathbb{I} - \mathbb{P})\boldsymbol{Y}$$
$$\boldsymbol{Y} \in \mathbb{R}^n, \beta \in \mathbb{R}^2$$



The errors $\|\boldsymbol{Y} - \mathbb{X}\beta\|_2^2$ should be minimal in some sense!

Calibration by the method of least squares

- The model parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top \in \mathbb{R}^{p+1}$ are obtained/estimated by solving the minimization problem

$$\widehat{\beta}_n = \underset{\beta \,\in\, \mathbb{R}^{p+1}}{Argmin} \quad \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$$

Calibration by the method of least squares

- The model parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$ are obtained/estimated by solving the minimization problem

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{Argmin} \quad \| \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} \|_2^2$$

- It is easy to verify that this is a convex minimization problem – the effective solution exists and it can be obtained in an explicit form;

Calibration by the method of least squares

- The model parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$ are obtained/estimated by solving the minimization problem

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{Argmin} \quad \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$$

- It is easy to verify that this is a convex minimization problem – the effective solution exists and it can be obtained in an explicit form;

- Taking partial derivatives with respect to $\beta_0, \dots, \beta_p$ and setting the derivatives to be equal to zero, the system of linear equations is obtains:

$$\mathbb{X}^\top \mathbb{X}\boldsymbol{\beta} = \mathbb{X}^\top \boldsymbol{Y}$$

Calibration by the method of least squares

- The model parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$ are obtained/estimated by solving the minimization problem

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{Argmin} \quad \|\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$$

- It is easy to verify that this is a convex minimization problem – the effective solution exists and it can be obtained in an explicit form;

- Taking partial derivatives with respect to $\beta_0, \dots, \beta_p$ and setting the derivatives to be equal to zero, the system of linear equations is obtains:

$$\mathbb{X}^\top \mathbb{X} \boldsymbol{\beta} = \mathbb{X}^\top \boldsymbol{Y}$$

- If the matrix $\mathbb{X}^\top \mathbb{X}$ is invertible, then the solution is explicitly expressed as

$$\widehat{\boldsymbol{\beta}}_n = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{Y}$$

Calibration by the method of least squares

- The model parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^\top \in \mathbb{R}^{p+1}$ are obtained/estimated by solving the minimization problem

$$\widehat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{Argmin} \quad \| \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} \|_2^2$$

- It is easy to verify that this is a convex minimization problem – the effective solution exists and it can be obtained in an explicit form;

- Taking partial derivatives with respect to $\beta_0, \dots, \beta_p$ and setting the derivatives to be equal to zero, the system of linear equations is obtains:

$$\mathbb{X}^\top \mathbb{X} \boldsymbol{\beta} = \mathbb{X}^\top \boldsymbol{Y}$$

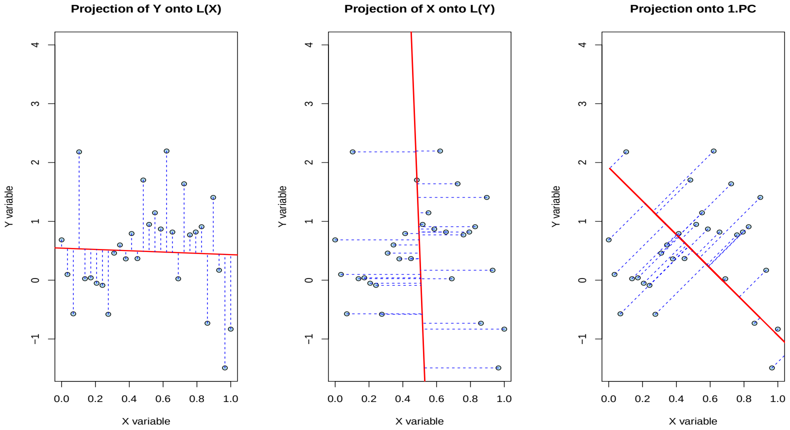
- If the matrix $\mathbb{X}^\top \mathbb{X}$ is invertible, then the solution is explicitly expressed as

$$\widehat{\boldsymbol{\beta}}_n = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{Y}$$

- The estimated model $\mathbb{X}\widehat{\boldsymbol{\beta}}_n$ is actually a projection into a linear subspace generated by the columns of the matrix $\mathbb{X}$ (i.e. $\mathbb{P} = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top$).

Some alternative calibration techniques

- However, we can still do better... (SVD, EIV);



Probabilistic model and the role of statistics

For practical utilization of the model (linear regression) we need much more than just some algebraic calculations, partial derivatives, and numerical algorithms to find the solution... The goal is to do **inference**!

- **Probabilistic model (usually imposed on the error terms)**
  - this allows to derive some useful properties for $\widehat{\beta}_n$ (the model);
  - the most common probabilistic model: the normal regression model;
  - BLUE, consistency, normality or asymptotic normality, etc.;
- **Statistical data which corresponds with the underlying theory**
  - not the data should be enhanced but the model must suit the data;
  - various statistical tools to verify underlying theoretical assumptions;
  - this is, however, not performed by the black-box software automatically!

3. Prediction/Forecasting: Model utilization

"The regression model describes the relationship between one or more 'input' variables and an 'output' variable. It gives us an equation to predict values for the 'output' variable, by plugging in the corresponding values for the 'input' variables."

- **Prediction**
  Formal statement which can be validated or falsified with just one single observation (the prediction was true or false);
  - A calibrated regression model is needed to make a prediction;
  - Algebraic procedures and numerical algorithms needed to calibrate model;

- **Forecasting**
  Multiple observations are needed to determine confidence level – it is characterized by calculating probabilities;
  - The regression model and the nature of the data is needed for forecasting;
  - **Probability theory and statistical inference tools!**

Regression: Some useful jargon

- If we believe to know the underlying model – we believe in some specific form of an analytic functional relationship which we know up to some few values of parameters – then the regression is called **parametric**;
- Otherwise, the regression is called **nonparametric**;

- If the unknown parameters enter the model in a linear way, we speak about **linear regression**.
- Otherwise, we speak about **nonlinear regression**;

- A linear regression is called **simple** if we fit a linear dependence of a response on just one single predictor;
- Otherwise, the linear regression is called **multiple**;

- If we believe that the nature of the data follows the normal distribution, we speak about **normal linear regression**;
- Otherwise, the regression is **general**;

Common problems
when fitting regression models


CURIOSITY BUILDS
MYSTERY BOX
WONDER AWAITS

What can go wrong in regression?

- **Model specification**
  (incorrect specification of the unknown underlying structure)
- **Inconsistent calibration**
  (wrong method used for the model estimation)
- **False prediction/forecasting**
  (violated assumptions needed for the proper inference)
- **Model selection**
  (incorrect covariates used for explaining the dependent variable)
- **Multicolinearity**
  (the estimated parameters, the calibrated model respectively, is not stable)
- **Dependence**
  (analyzing dependent data instead of independent)
- **... ... ...**

Model selection: Variable screening

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20*. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."*

(JAMA Surgery, 2016)

Model selection: Variable screening

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20*. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."*

(JAMA Surgery, 2016)

- **Sifnificant covariate in a univariate regression may turn non-significant in a multivariate regression;**
- **Non-significant covariate in a univariate regression may turn significant in a multivariate regression;**

Missing important covariate

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20*. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."*

(JAMA Surgery, 2016)

- Three independent (standard normal) covariates: X_1, X_2, X_3;
- Standard normal error terms (indelendent of X covariates) $\varepsilon \sim N(0, \sigma^2)$;
- Additional covariate X_4 defined as: $X_4 = \beta_1 X_1 + \beta_2 X_2 = 2X_1 + X_2$;
- True underlying model of the form: $Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \varepsilon$;
- Univariate regression slope for $Y \sim X_4$: $\frac{Cov(Y, X_4)}{Var X_4} = \alpha_4 + \frac{\alpha_2 \beta_2}{\beta_1^2 + \beta_2^2}$

Missing important covariate

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."

(JAMA Surgery, 2016)

- Three independent (standard normal) covariates: X_1, X_2, X_3;
- Standard normal error terms (indelendent of X covariates) $\varepsilon \sim N(0, \sigma^2)$;
- Additional covariate X_4 defined as: $X_4 = \beta_1 X_1 + \beta_2 X_2 = 2X_1 + X_2$;
- True underlying model of the form: $Y = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \varepsilon$;
- Univariate regression slope for $Y \sim X_4$: $\frac{Cov(Y, X_4)}{VarX_4} = \alpha_4 + \frac{\alpha_2 \beta_2}{\beta_1^2 + \beta_2^2}$

Underlying model:

$$Y = 1 + 3X_2 + 4X_3 - 0.6X_4 + \varepsilon$$

Missing important covariate

- Simulation results based on 10.000 Monte Carlo repetitions;

n	**Univariate Regression X_4** Estimate (Std.Err.)	**Multiple Regression X_4** Estimate (Std.Error)	**Regression on X_2 and X_3** Estimates (Std. Errs)
30	-0.0005 (0.4225)	-0.6010 (0.0988)	2.4074 (0.3030) | 4.0028 (0.3047)
50	-0.0016 (0.3194)	-0.6003 (0.0748)	2.3990 (0.2281) | 4.0014 (0.2302)
100	-0.0009 (0.2226)	-0.6003 (0.0513)	2.4020 (0.1611) | 3.9992 (0.1581)
200	0.0002 (0.1485)	-0.6002 (0.0357)	2.3999 (0.1111) | 4.0019 (0.1126)
500	0.0005 (0.0965)	-0.6005 (0.0226)	2.4002 (0.0703) | 4.0005 (0.0705)
1000	-0.0002 (0.0691)	-0.6000 (0.0160)	2.4002 (0.0498) | 3.9993 (0.0492)

- Underlying model:

$$Y = 1 + 3X_2 + 4X_3 - 0.6X_4 + \varepsilon$$

Missing important covariate

- Simulation results based on 10.000 Monte Carlo repetitions;

n	Univariate Regression X_4 Estimate (Std.Err.)	Multiple Regression X_4 Estimate (Std.Error)	Regression on X_2 and X_3 Estimates (Std. Errs)
30	-0.0005 (0.4225)	-0.6010 (0.0988)	2.4074 (0.3030) | 4.0028 (0.3047)
50	-0.0016 (0.3194)	-0.6003 (0.0748)	2.3990 (0.2281) | 4.0014 (0.2302)
100	-0.0009 (0.2226)	-0.6003 (0.0513)	2.4020 (0.1611) | 3.9992 (0.1581)
200	0.0002 (0.1485)	-0.6002 (0.0357)	2.3999 (0.1111) | 4.0019 (0.1126)
500	0.0005 (0.0965)	-0.6005 (0.0226)	2.4002 (0.0703) | 4.0005 (0.0705)
1000	-0.0002 (0.0691)	-0.6000 (0.0160)	2.4002 (0.0498) | 3.9993 (0.0492)
	$\longrightarrow_P 0$		

- Underlying model:

$$Y = 1 + 3X_2 + 4X_3 - 0.6X_4 + \varepsilon$$

Missing important covariate

- Simulation results based on 10.000 Monte Carlo repetitions;

n	**Univariate Regression X_4** Estimate (Std.Err.)	**Multiple Regression X_4** Estimate (Std.Error)	**Regression on X_2 and X_3** Estimates (Std. Errs)
30	-0.0005 (0.4225)	-0.6010 (0.0988)	2.4074 (0.3030) | 4.0028 (0.3047)
50	-0.0016 (0.3194)	-0.6003 (0.0748)	2.3990 (0.2281) | 4.0014 (0.2302)
100	-0.0009 (0.2226)	-0.6003 (0.0513)	2.4020 (0.1611) | 3.9992 (0.1581)
200	0.0002 (0.1485)	-0.6002 (0.0357)	2.3999 (0.1111) | 4.0019 (0.1126)
500	0.0005 (0.0965)	-0.6005 (0.0226)	2.4002 (0.0703) | 4.0005 (0.0705)
1000	-0.0002 (0.0691)	-0.6000 (0.0160)	2.4002 (0.0498) | 3.9993 (0.0492)
	$\longrightarrow_P 0$	$\longrightarrow_P -3/5$	

- Underlying model:

$$Y = 1 + 3X_2 + 4X_3 - 0.6X_4 + \varepsilon$$

Missing important covariate

- Simulation results based on 10.000 Monte Carlo repetitions;

n	**Univariate Regression X_4** Estimate (Std.Err.)	**Multiple Regression X_4** Estimate (Std.Error)	**Regression on X_2 and X_3** Estimates (Std. Errs)
30	-0.0005 (0.4225)	-0.6010 (0.0988)	2.4074 (0.3030) | 4.0028 (0.3047)
50	-0.0016 (0.3194)	-0.6003 (0.0748)	2.3990 (0.2281) | 4.0014 (0.2302)
100	-0.0009 (0.2226)	-0.6003 (0.0513)	2.4020 (0.1611) | 3.9992 (0.1581)
200	0.0002 (0.1485)	-0.6002 (0.0357)	2.3999 (0.1111) | 4.0019 (0.1126)
500	0.0005 (0.0965)	-0.6005 (0.0226)	2.4002 (0.0703) | 4.0005 (0.0705)
1000	-0.0002 (0.0691)	-0.6000 (0.0160)	2.4002 (0.0498) | 3.9993 (0.0492)
	$\longrightarrow_P 0$	$\longrightarrow_P -3/5$	$\longrightarrow_P 4$

- Underlying model:

$$Y = 1 + 3X_2 + 4X_3 - 0.6X_4 + \varepsilon$$

Missing important covariate

- Simulation results based on 10.000 Monte Carlo repetitions;

n	**Univariate Regression X_4** Estimate (Std.Err.)	**Multiple Regression X_4** Estimate (Std.Error)	**Regression on X_2 and X_3** Estimates (Std. Errs)
30	-0.0005 (0.4225)	-0.6010 (0.0988)	2.4074 (0.3030) | 4.0028 (0.3047)
50	-0.0016 (0.3194)	-0.6003 (0.0748)	2.3990 (0.2281) | 4.0014 (0.2302)
100	-0.0009 (0.2226)	-0.6003 (0.0513)	2.4020 (0.1611) | 3.9992 (0.1581)
200	0.0002 (0.1485)	-0.6002 (0.0357)	2.3999 (0.1111) | 4.0019 (0.1126)
500	0.0005 (0.0965)	-0.6005 (0.0226)	2.4002 (0.0703) | 4.0005 (0.0705)
1000	-0.0002 (0.0691)	-0.6000 (0.0160)	2.4002 (0.0498) | 3.9993 (0.0492)
	$\longrightarrow_P 0$	$\longrightarrow_P -3/5$	$\not\rightarrow_P 3 \mid \longrightarrow_P 4$

- Underlying model:

$$Y = 1 + 3X_2 + 4X_3 - 0.6X_4 + \varepsilon$$

Irrelevant covariate passing the screening

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20*. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."*

(JAMA Surgery, 2016)

Irrelevant covariate passing the screening

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20*. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."*

(JAMA Surgery, 2016)

- Three independent (standard normal) covariates: X_1, X_2, X_3;
- Standard normal error terms (indelendent of X covariates) $\varepsilon \sim N(0, \sigma^2)$;
- Additional covariate X_4 defined as: $X_4 = \beta_1 X_1 + \beta_2 X_3 = X_1 + X_3$;
- Consider the true model of the form: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon$;
- Extended model of the form: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_4 + \varepsilon$
- Alternatively: $Y = \alpha_0 + (\alpha_1 + \alpha_3 \beta_1) X_1 + \alpha_2 X_2 + \alpha_3 \beta_2 X_3 + \varepsilon$

Irrelevant covariate passing the screening

"The administrative database was evaluated by means of univariate and multivariate regression. First, we identified variables that were associated with the dependent variable with p-value < 0.20. These potential confounders were then entered in multivariate regression in a stepwise backward fitting approach."

(JAMA Surgery, 2016)

- Three independent (standard normal) covariates: X_1, X_2, X_3;
- Standard normal error terms (indelendent of X covariates) $\varepsilon \sim N(0, \sigma^2)$;
- Additional covariate X_4 defined as: $X_4 = \beta_1 X_1 + \beta_2 X_3 = X_1 + X_3$;
- Consider the true model of the form: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon$;
- Extended model of the form: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_4 + \varepsilon$
- Alternatively: $Y = \alpha_0 + (\alpha_1 + \alpha_3 \beta_1) X_1 + \alpha_2 X_2 + \alpha_3 \beta_2 X_3 + \varepsilon$

Underlying model:

$$Y = X_1 + 2X_2 + \varepsilon$$

Irrelevant covariate passing the screening

- Simulation results based on 10.000 Monte Carlo repetitions;

n	**Univariate Regression X_4** Estimate (Std.Err.)	**Multiple Regression X_4** Estimate (Std.Error)
30	1.0024 (0.4723)	0.0038 (0.2014)
50	0.9975 (0.3564)	-0.0008 (0.1496)
100	0.9995 (0.2469)	-0.0015 (0.1032)
200	0.9982 (0.1733)	0.0005 (0.0723)
500	0.9999 (0.1101)	0.0005 (0.0452)
1000	0.9995 (0.0776)	0.0004 (0.0318)

- Underlying model:

$$Y = X_1 + 2X_2 + \varepsilon$$

Correlation transitivity

"Since factor A is highly correlated with outcome Y, and factor A and factor B are highly correlated, then B should be also correlated with Y."

(JAMA Surgery, 2016)

Correlation transitivity

"Since factor A is highly correlated with outcome Y, and factor A and factor B are highly correlated, then B should be also correlated with Y."

(JAMA Surgery, 2016)

- Random variables X and Z are independent standard normal;
- Let the variable Y be defined as $Y = X + Z$;
- The correlation between Y and X is: **0.707**;
- The correlation between Y and Z is again **0.707**;
- However, the correlation between X and Z is zero;

Correlation transitivity

"Since factor A is highly correlated with outcome Y, and factor A and factor B are highly correlated, then B should be also correlated with Y."

(JAMA Surgery, 2016)

- Random variables X and Z are independent standard normal;
- Let the variable Y be defined as $Y = X + Z$;
- The correlation between Y and X is: **0.707**;
- The correlation between Y and Z is again **0.707**;
- However, the correlation between X and Z is zero;

- Example before: the correlation between X_4 and X_1 is 0.707;
- Example before: the correlation between X_1 and Y is 0.408;
- However, X_4 has no role in the multiple regression model;

Stability of the estimates and *p*-values

- Available covariates: `height`, `weight`, `age`, `gender`, `bmi`, `wh-ratio`;
- body fat vs. subject's height:

```
lm(formula = fat ~ height, data = Policie)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -47.6791    23.9707  -1.989   0.0524 .
height        0.3405     0.1343   2.535   0.0146 *
```

Stability of the estimates and *p*-values

- Available covariates: `height`, `weight`, `age`, `gender`, `bmi`, `wh-ratio`;
- body fat vs. subject's height:

```
lm(formula = fat ~ height, data = Policie)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -47.6791    23.9707  -1.989   0.0524 .
height        0.3405     0.1343   2.535   0.0146 *
```

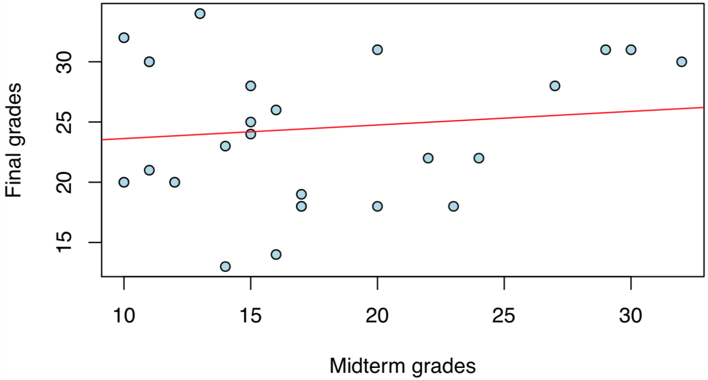
- body fat vs. subject's height and weight:

```
lm(formula = fat ~ height + weight, data = Policie)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.55309   15.24621   1.086   0.2831
height      -0.24362    0.09728  -2.504   0.0158 *
weight       0.50418    0.05095   9.896 4.49e-13 ***
```

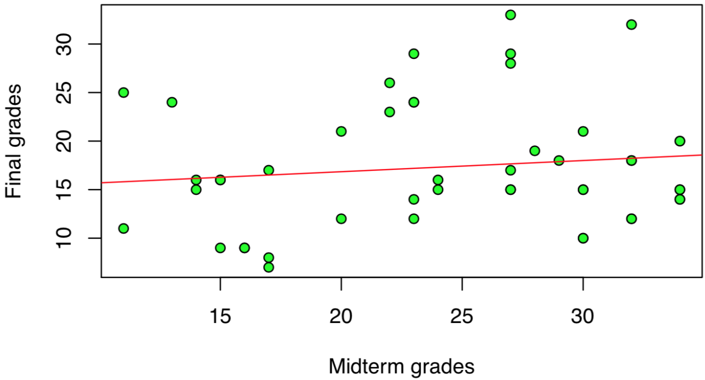
Paradox: Ecological fallacy

- Stat235 classes at University of Alberta in Fall 2012/2013;
- Students' performance for midterm exams and final exams;



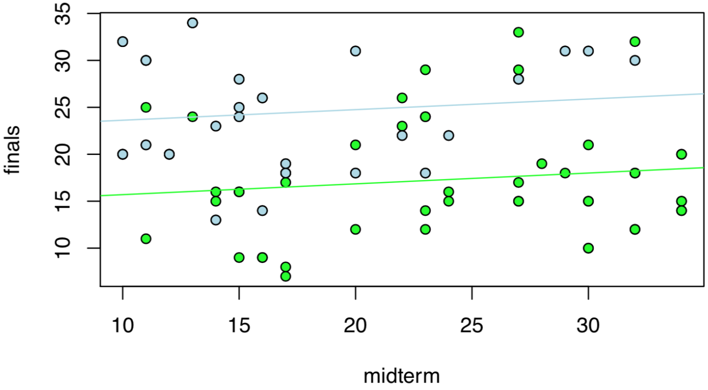
Paradox: Ecological fallacy

- Stat235 classes at University of Alberta in Fall 2012/2013;
- Students' performance for midterm exams and final exams;



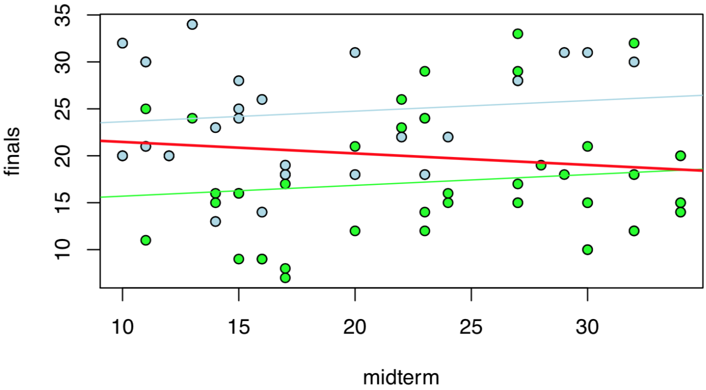
Paradox: Ecological fallacy

- Stat235 classes at University of Alberta in Fall 2012/2013;
- Students' performance for midterm exams and final exams;

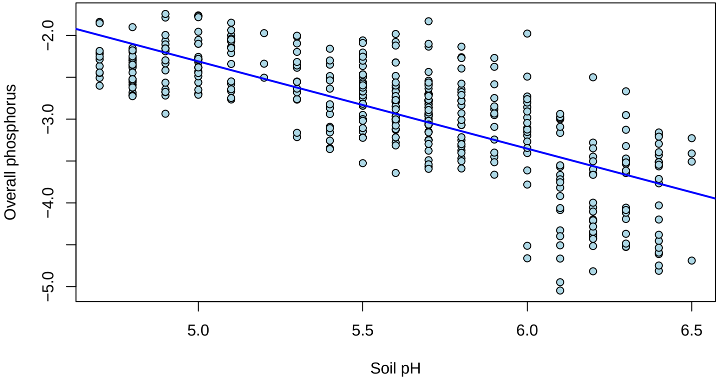


Paradox: Ecological fallacy

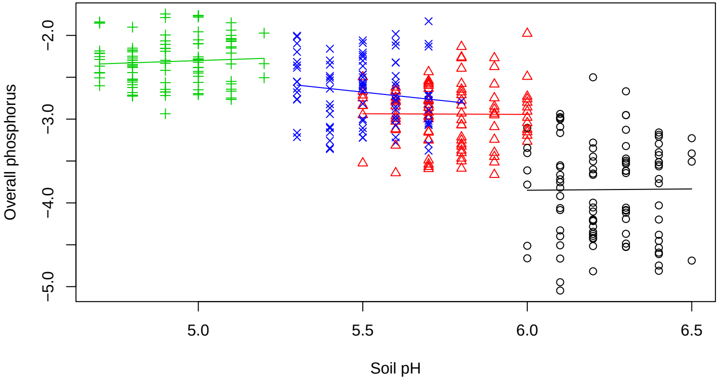
- Stat235 classes at University of Alberta in Fall 2012/2013;
- Students' performance for midterm exams and final exams;



Paradox: Ecological fallacy

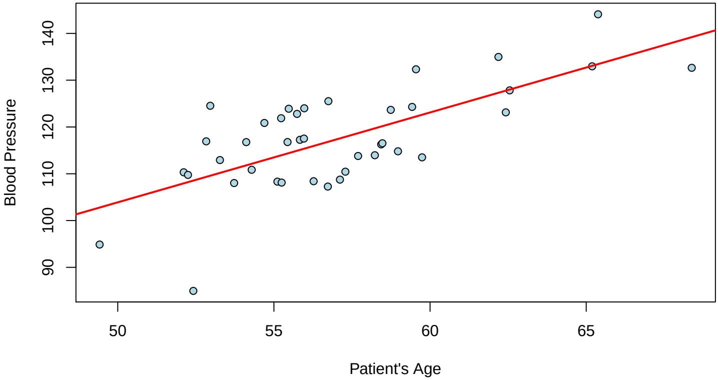


Paradox: Ecological fallacy



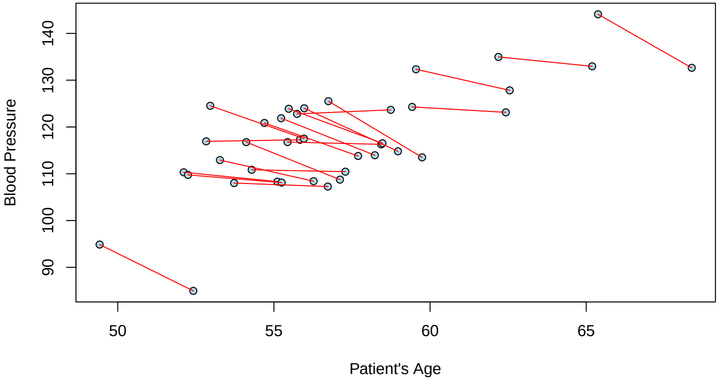
Dependent and independent observations

- Random sample:
  Independent and identically distributed random observations/variables1;



Dependent and independent observations

- Random sample:
  Independent and identically distributed random observations/variables1;




CURIOSITY BUILDS
MYSTERY
BOX
WONDER AWAITS

Linear regression: Formally and correctly

- Linear regression – a probabilistic model which requires a specific set of assumptions to be satisfied to obtain reliable results at the end;

Linear regression: Formally and correctly

- Linear regression – a probabilistic model which requires a specific set of assumptions to be satisfied to obtain reliable results at the end;
  - **Available data form a random sample;**
    (independent and identically distributed random observations)
  - **Correct model specification;**
    (the parametric form of the estimated structure must be correctly defined)
  - **Normally distributed error terms;**
    (especially if there is some interest in a consequent statistical inference)
  - **Equal variance $\equiv$ homoscedasticity;**
    (all error terms should have same variance)
  - **Well defined set of explanatory variables;**
    for instance, no linear dependence among covariates or multicolinearity

Linear regression: Formally and correctly

- Linear regression – a probabilistic model which requires a specific set of assumptions to be satisfied to obtain reliable results at the end;
    - **Available data form a random sample;**
      (independent and identically distributed random observations)
    - **Correct model specification;**
      (the parametric form of the estimated structure must be correctly defined)
    - **Normally distributed error terms;**
      (especially if there is some interest in a consequent statistical inference)
    - **Equal variance $\equiv$ homoscedasticity;**
      (all error terms should have same variance)
    - **Well defined set of explanatory variables;**
      for instance, no linear dependence among covariates or multicolinearity
- Straightforward extensions of the linear model (easy ones or quite complex) for handling violated assumptions;

Linear regression: Formally and correctly

- Linear regression – a probabilistic model which requires a specific set of assumptions to be satisfied to obtain reliable results at the end;
  - **Available data form a random sample;**
    (independent and identically distributed random observations)
  - **Correct model specification;**
    (the parametric form of the estimated structure must be correctly defined)
  - **Normally distributed error terms;**
    (especially if there is some interest in a consequent statistical inference)
  - **Equal variance $\equiv$ homoscedasticity;**
    (all error terms should have same variance)
  - **Well defined set of explanatory variables;**
    for instance, no linear dependence among covariates or multicolinearity
- Straightforward extensions of the linear model (easy ones or quite complex) for handling violated assumptions;
- However, applying a standard linear regression model in such cases causes incorrect results and false conclusions;