UNIVERSITÄT
LEIPZIG

10-207-0003: Introduction to Stochastics

Regression

21.05.2025, Leipzig

Dr. Ing. Andreas Niekler

Computational
Humanities
UNIVERSITÄT LEIPZIG

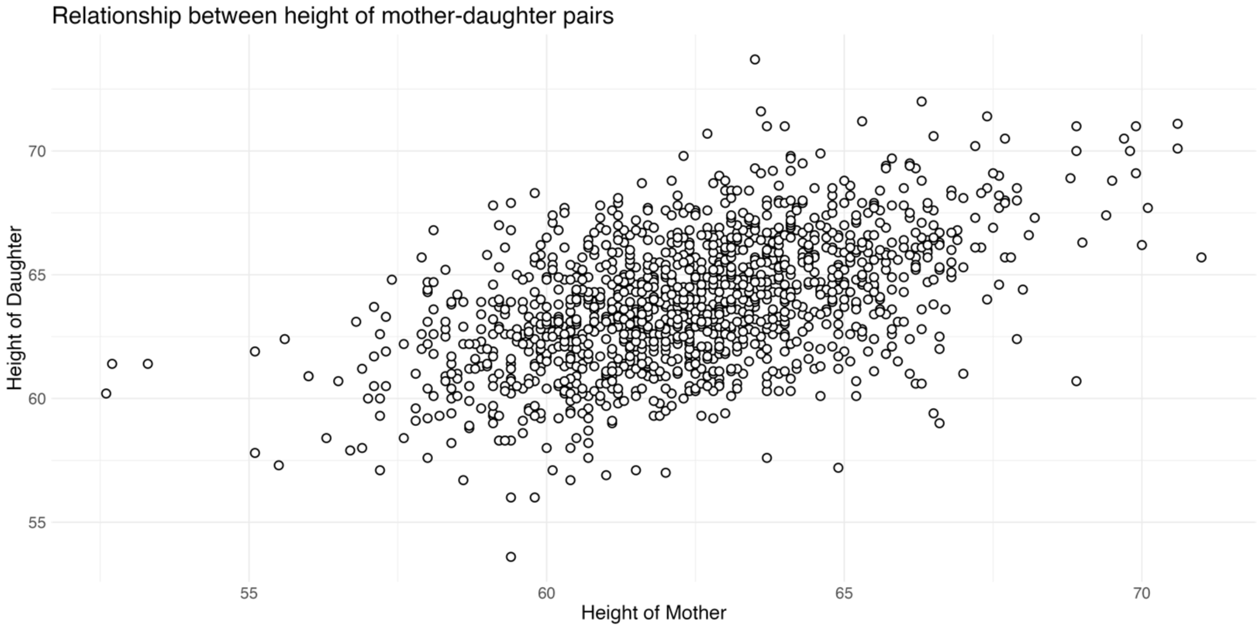
UNIVERSITÄT
LEIPZIG

SYLLABUS

1. Empirical research and scale levels
2. Univariate description and exploration of data
3. Graphical representation of characteristics / Explorative data analysis
4. Measures of data distribution
5. Multivariate Problems, Correlation
6. ***Regression***
7. Probability distributions
8. Central Limit Theorem
9. Confidences
10. Statistical testing
11. Linear Regression
12. Correlation and covariance
13. Logistic regression
14. Bayes theorem

Additional: Entropy, Mutual Information, Maximum Likelihood Estimator, Mathy Stuff

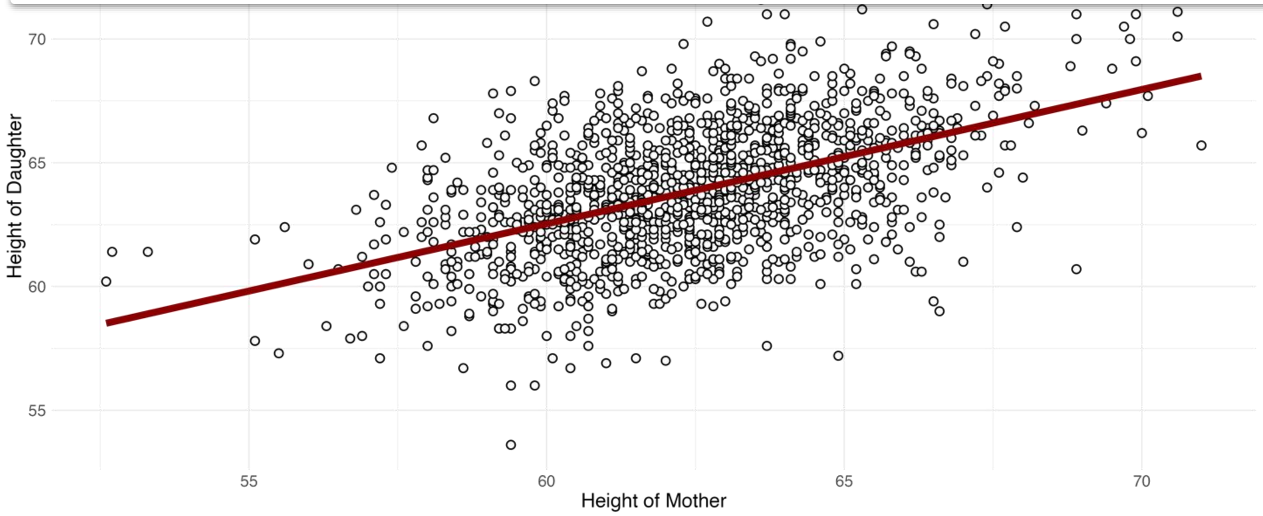
RECAP MULTIVARIATE DATE



Based on Pearson & Lee (1903)

RECAP MULTIVARIATE DATE

Positive linear relationship: It is **predominantly** true that the taller the mother, the taller her daughter.



Based on Pearson & Lee (1903)

RECAP CORRELATION COEFFICIENT

👉 The (linear) correlation or the so-called correlation coefficient r measures the strength and direction of the linear relationship between two variables X and Y and is defined as:

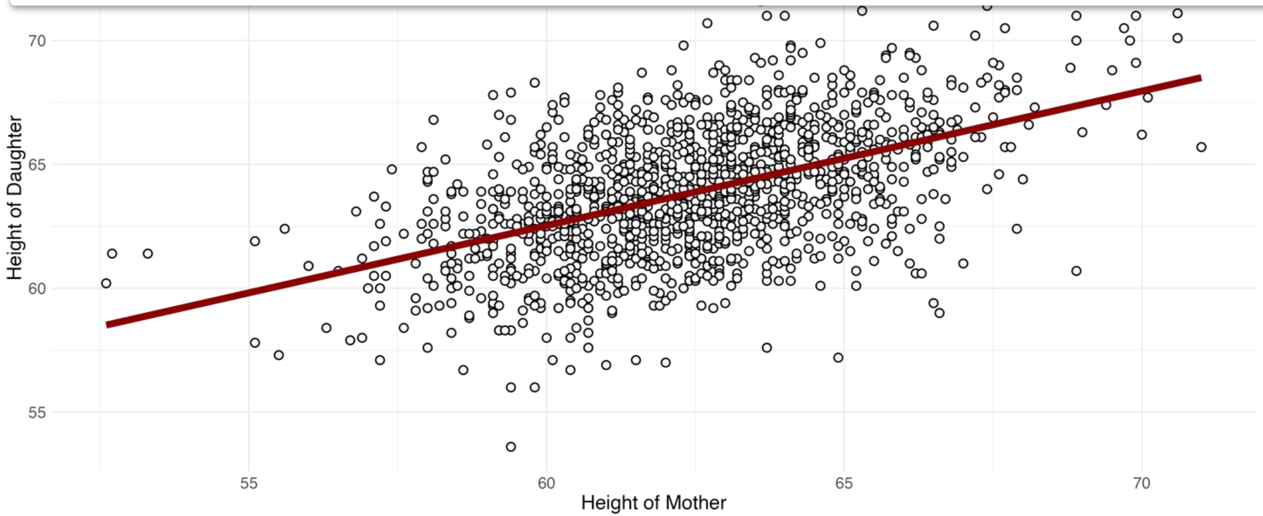
$$r = r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right),$$

where $\bar{x}$ and $\bar{y}$ are the means, and s_X and s_Y are the standard deviations of the two variables X and Y.

- One speaks of a positive (linear) correlation if $r > 0$.
- Analogously, one speaks of a negative (linear) correlation if $r < 0$.
- For $r \approx 0$, one says that X and Y are uncorrelated.

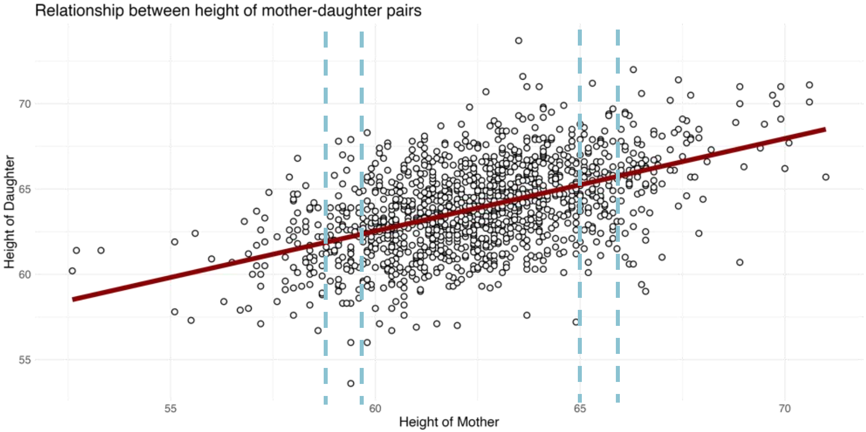
RECAP MULTIVARIATE DATE

r = 0.491 → There is actually a substantial positive linear relationship.



Based on Pearson & Lee (1903)

RECAP MULTIVARIATE DATE

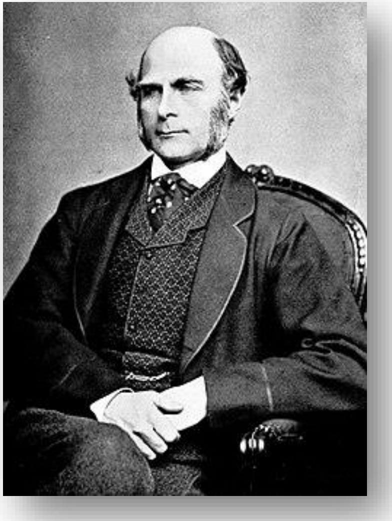


- **Phenomenon of Regression to the Mean (Lat. *regredi* for reverse, go back):**
  - The further the size of the mother being considered deviates from the mean, the larger is also the deviation of the average size of the corresponding daughters, although this deviation is less pronounced.
  - The term **Regression was therefore originally used in the sense of moving back / returning towards the mean**.
  - Today, the phenomenon of Regression to the Mean refers rather somewhat more generally to **situations of repeated measurements, where less extreme values occur after extreme values**.

Based on Pearson & Lee (1903)

SIR FRANCIS GALTON

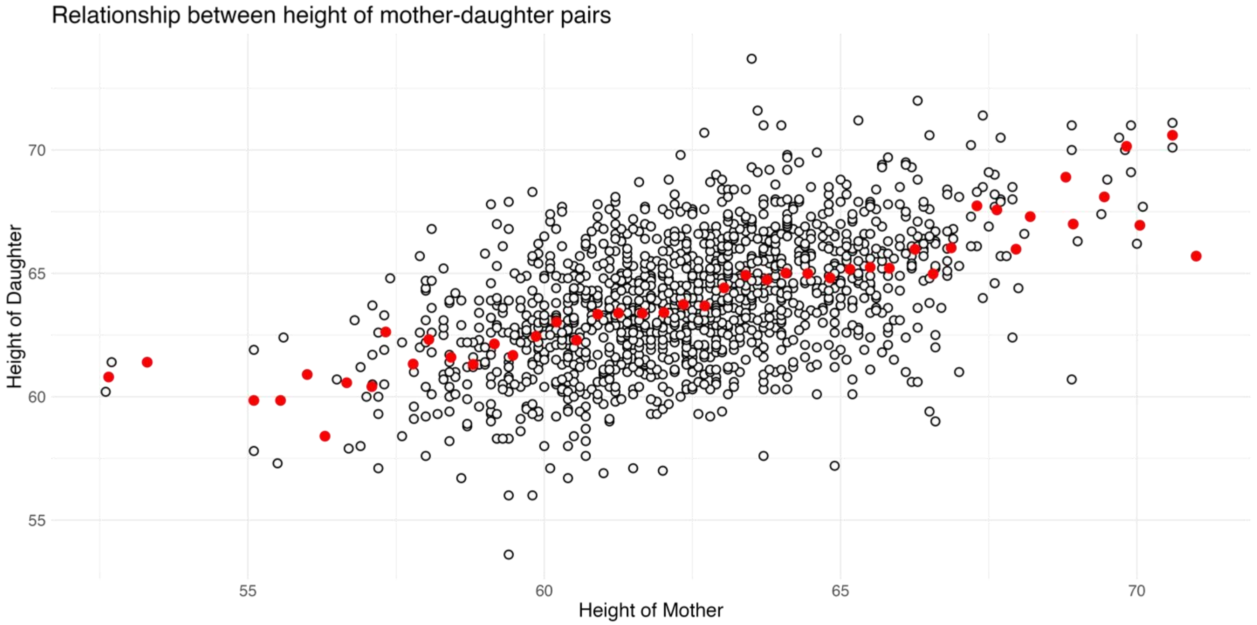
- Sir Francis Galton was an English polymath and statistician. A cousin of Charles Darwin, he was a pioneer in **applying statistical methods to study human traits**. He is best known for discovering and naming the phenomenon of "regression to the mean," which he **observed while studying the inheritance of characteristics like height**. His work laid foundational concepts in statistics, including correlation.



THE GALTON BOARD

- A Galton board is a device where beads fall through pegs, randomly bouncing left or right, and collect in bins to form a bell curve (normal distribution), illustrating how **random chance leads to a predictable pattern centered around the average.** This relates to regression to the mean because the extreme outcomes (beads in the far bins) are less likely and due to chance; **a subsequent drop of a bead that landed in an extreme spot would likely result in it landing closer to the average**, demonstrating how extreme results tend to revert towards the mean in processes influenced by randomness.

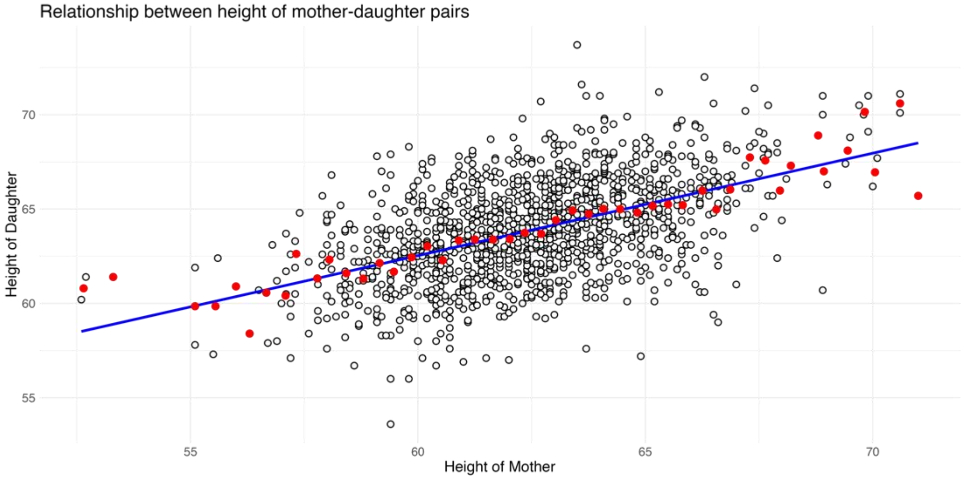
MOVING MEAN



– If one considers the conditional means, a linear relationship can be found for them.

Based on Pearson & Lee (1903)

MOVING MEAN



- Approximately the following regression line (blue line) can be determined: $\hat{y} = 76.0 + 0.542 \cdot x$
- **Interpretation:** If the mother's height increases by 1 cm, the average height of the daughter increases by 0.542 cm.
- **Phenomenon of Regression to the Mean:** If the mother's height moves away from the mean, the average height of the daughter increases or decreases by only about half.

Based on Pearson & Lee (1903)

REGRESSION ANALYSIS

- Although the phenomenon of regression to the mean was the namesake for the development of regression analysis, **today we understand the concept of regression or regression analysis a little differently**.

👉 **Regression Analysis**: is concerned with the study of the dependence of a variable Y, the so-called *dependent variable*, on other variables $X_1, \dots, X_p$, the so-called *explanatory variables*, with the intention of clarifying this dependency and **making predictions** for the dependent variable using the explanatory variables.

- Formulated somewhat more abstractly, the aim is to find a functional dependency

$$Y \approx f\left(X_1, \dots, X_p\right)$$

  such that Y can be well explained by $X_1, \dots, X_p$.

- We first consider only two metric variables X and Y, and restrict ourselves to clarifying linear relationships of the form

$$Y = b_0 + b_1 \cdot X.$$

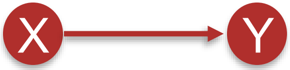
REGRESSION ANALYSIS

- In regression analysis, the symmetry of the relationship is generally abandoned, i.e., **a directed relationship of the form $X \rightarrow Y$ is usually considered, which is, for example, causally motivated.**

- Remark: The terms for X and Y in regression analysis are a little inconsistent, as they depend on the context:
- X|Y
  - independent variable | dependent variable
  - exogenous variable | endogenous variable
  - explanatory variable | variable to be explained
  - Stimulus | Response
  - Influencing variable | Target variable
  - Predictor | Target variable
  - Regressor | Regressand

CAUSALITY IN STATISTICS

- Causality is the direct link where one thing causes another, changing the effect when the cause changes. However, proving this definitively is often hard. Instead of absolute certainty, we **frequently reason about plausible causal links using theory, logic, and evidence, inferring likely cause-and-effect** based on the best available information.



REGRESSION ANALYSIS

- The Bravais-Pearson correlation coefficient measures the strength of the linear relationship between X and Y:
  - "**How well** can the data be described by a straight line?"
- Linear regression analysis now goes a step further:
  - "**What does the best fitting line look** like?"
- One goal is therefore the analysis and description of the **relationship using a suitable line**.
- Another goal is to make predictions about Y based on X:
  - "Which value $\hat{y}$ **is predicted** for a value x?”
- This is called **a model**.

RESIDUALS

- The predicted value for y is denoted by $\hat{y}$, as the actual and predicted values will generally not coincide.
  - In the chocolate/vanilla example, none of the data points lie on the line.
  - For some of the observed points, the predicted point is too large, and for others, it is too small.

👉 For each observed unit i, the difference e_i between the observed y_i and the corresponding predicted value $\hat{y}_i$, which represents the **prediction error**, is called the **residual**:

$$e_i := y_i - \hat{y}_i$$

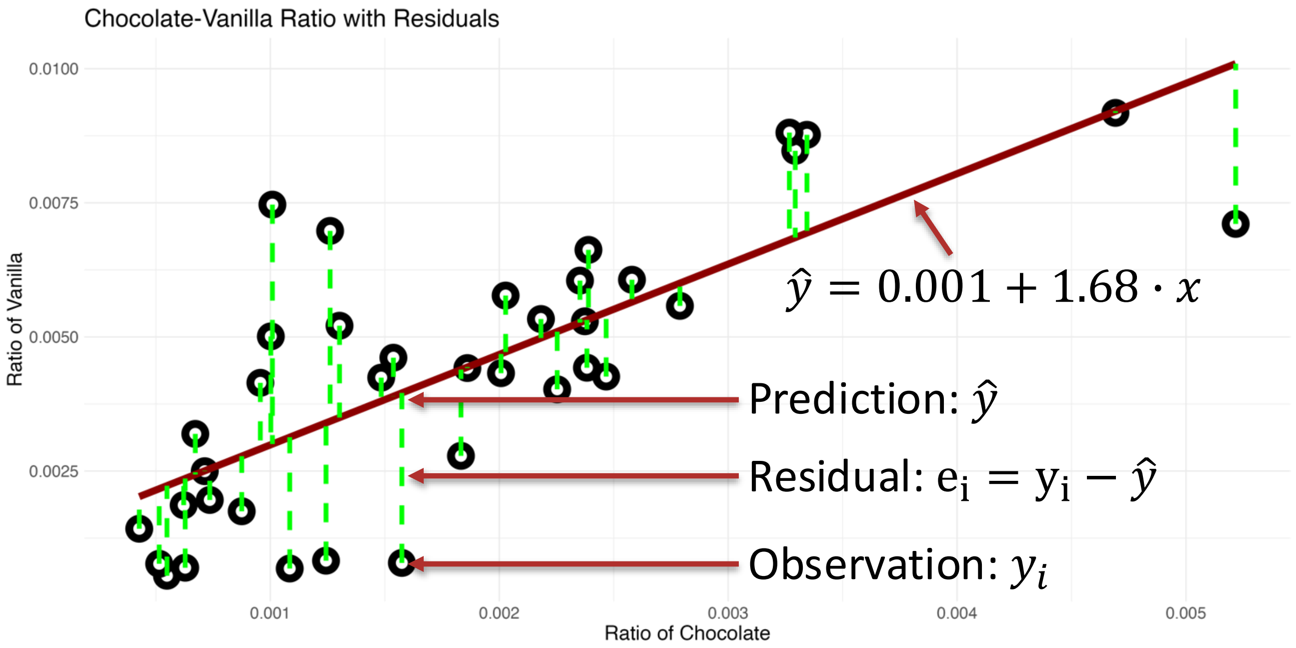
RESIDUALS

- The predicted value for y is denoted by $\hat{y}$, as the actual and predicted values will generally not coincide.
  - In the chocolate/vanilla example, none of the data points lie on the line.
  - For some of the observed points, the predicted point is too large, and for others, it is too small.

👉 For each observed unit i, the difference e_i between the observed y_i and the corresponding predicted value $\hat{y}_i$, which represents the **prediction error**, is called the **residual**:

$$e_i \coloneqq y_i - \hat{y}_i$$

RESIDUALS



$$\hat{y} = 0.001 + 1.68 \cdot x$$

Prediction: $\hat{y}$

Residual: $\mathrm{e_i} = \mathrm{y_i} - \hat{y}$

Observation: y_i

SIMPLE LINEAR REGRESSION

The Simple Linear Regression Model: Given two metric variables X and Y. The simple linear regression model of Y on X assumes a linear relationship of the form

$$Y = b_0 + b_1 \cdot X$$

where b_0 describes the intercept (also: constant, base level, bias) and b_1 describes the slope, such that for the data

$$y_i = b_0 + b_1 x_i + e_i, i = 1, \dots, n$$

holds, where the so-called residual e_i describes the prediction error for unit i.

- Remark: The term "linear" refers only to the two coefficients b_0 and b_1, i.e., these appear only to the first power.
- Our goal now is to determine the coefficients b_0 and b_1 in such a way that the **prediction errors are as small as possible**.

ORDINARY LEAST SQUARES

- One method to determine the parameters b_0 and b_1 is the so-called method of least squares or OLS method.
- OLS stands for **(o)**rdinary **(l)**east **(s)**quares.

👉**Regression using the OLS Method**: The OLS regression line of a simple linear regression model is the line

$$\hat{y} = b_0 + b_1 x$$

for which the sum of squared residuals

$$SSR := \sum_{i=1}^{n} e_i^2$$

is minimized (OLS criterion).

- SSR stands for (s)um of (s)quared (r)esiduals.

MOTIVATION

- **The Sum of Squared Residuals**

$$SSR := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

describes **how much the actual values y_i scatter around the predicted values**, in direct analogy to the variance s^2 or standard deviation s of a characteristic.

- The sum of squared residuals thus represents a measure of the **total prediction error**.
- The OLS criterion therefore states that the parameters b_0 and b_1 should be **chosen such that this total error or the variation is minimized**.

THEOREME

- The OLS regression line is given by

$$\hat{y} = b_0 + b_1 x,$$

where

$$b_1 = r \cdot \frac{s_X}{s_Y} \text{ and } \bar{y} = b_0 + b_1 \bar{x}.$$

- Here, r is the Pearson correlation coefficient, s_Y and s_X are the standard deviations of Y and X respectively, and $\bar{y}$ and $\bar{x}$ are the means of Y and X.
- Furthermore, it holds that:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) = 0$$

Interpretation: The residuals scatter around the OLS line on average (their sum and mean are zero).

$$\bar{y} = b_0 + b_1 \bar{x}$$

Interpretation: The OLS line runs through the centroid of the data $(\bar{x}, \bar{y})$, or rather, the centroid lies on the OLS line.

FICTIVE ECONOMIC EXAMPLE

- Coffee was sold at three flea markets:
  - X Number of cups of coffee sold
  - Y corresponding profit (price is negotiable)

Flea Market	1	2	3
x_i	10	15	5
y_i	9	21	0

- The OLS regression line is to be determined for the (linear) regression of Y on X, and the revenue expected for twelve cups sold is to be predicted.

FICTIVE ECONOMIC EXAMPLE

- Support Table

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(y_i - \bar{y})$	y_i
1	10	0	0	0	1	-1	9
2	15	5	25	55	121	11	21
3	5	-5	25	50	100	-10	0
Σ	30	0	50	105	222	0	30
	$\bar{x}$						$\bar{y}$

- Descriptive Statistics

$$\bar{x} = \frac{\sum_{i=1}^{3} x_i}{3} = 10, \; s_x = \sqrt{\frac{\sum_{i=1}^{3}(x_i - \bar{x})^2}{3}} \approx 4.08$$

$$\bar{y} = \frac{\sum_{i=1}^{3} y_i}{3} = 10, \qquad s_Y = \sqrt{\frac{\sum_{i=1}^{3}(y_i - \bar{y})^2}{3}} \approx 8.60$$

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right) = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{x})(\boldsymbol{y_i} - \bar{y})}{n \cdot S_x \cdot S_y} \approx \frac{105}{3 \cdot 4.08 \cdot 8.60} \approx 0.997$$

FICTIVE ECONOMIC EXAMPLE

- Regression Coefficient

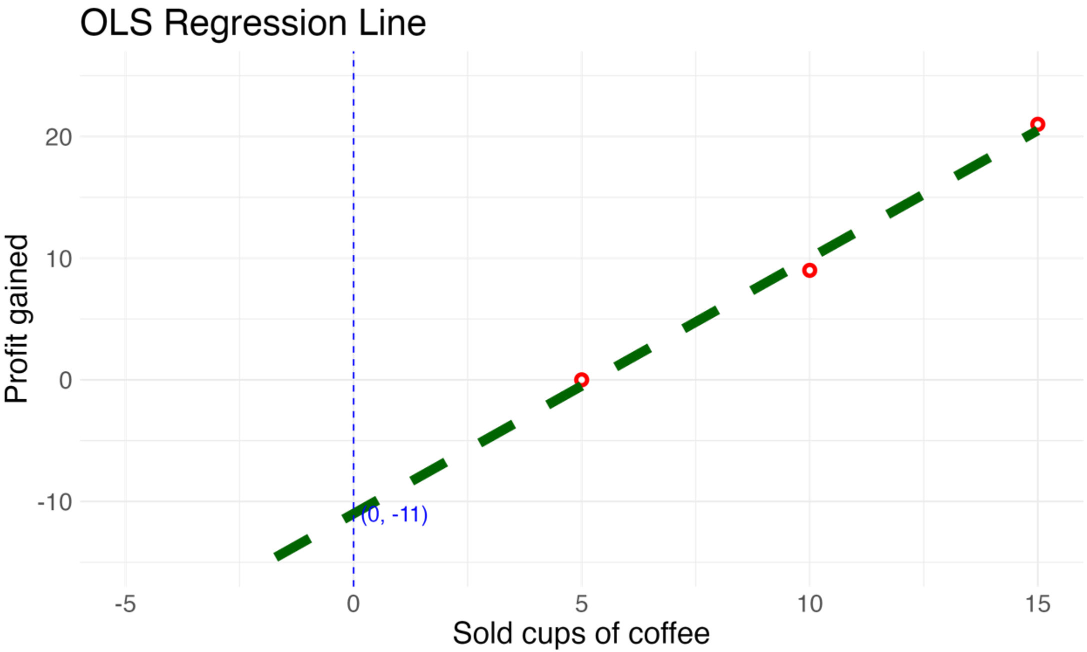
$$b_1 = r\frac{S_y}{S_x} \approx 0.997 \cdot \frac{8.60}{4.08} \approx 2.1$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 10 - 2.1 \cdot 10 = -11$$

- OLS-line

$$\hat{y} = b_0 + b_1 \cdot x = -11 + 2.1 \cdot x$$

FICTIVE ECONOMIC EXAMPLE



FICTIVE ECONOMIC EXAMPLE

- **The OLS Line:** $\hat{y} = b_0 + b_1 \cdot x = -11 + 2.1 \cdot x$ describes the best possible fit of a straight line to the data with respect to the OLS criterion.
- **Interpretation of the Regression Coefficients:**
  - $\boldsymbol{b_1 = 2.1}$**:** With an increase in the quantity X by one unit, the profit Y increases by 2.1 units. Thus, b_1 can be interpreted as a kind of marginal profit per additional cup of coffee sold.
  - $\boldsymbol{b_0 = -11}$**:** The predicted profit for 0 cups would thus be −11, meaning a loss of 11 Euros would be expected if no cups are sold. Therefore, b_0 can be interpreted here as the expected fixed costs (stall fees etc.).
- **Predicting Profit for Twelve Cups of Coffee:** For $x = 12$ cups of coffee, a profit of

$$\hat{y} = b_0 + b_1 \cdot x = -11 + 2.1 \cdot x = 14.2$$

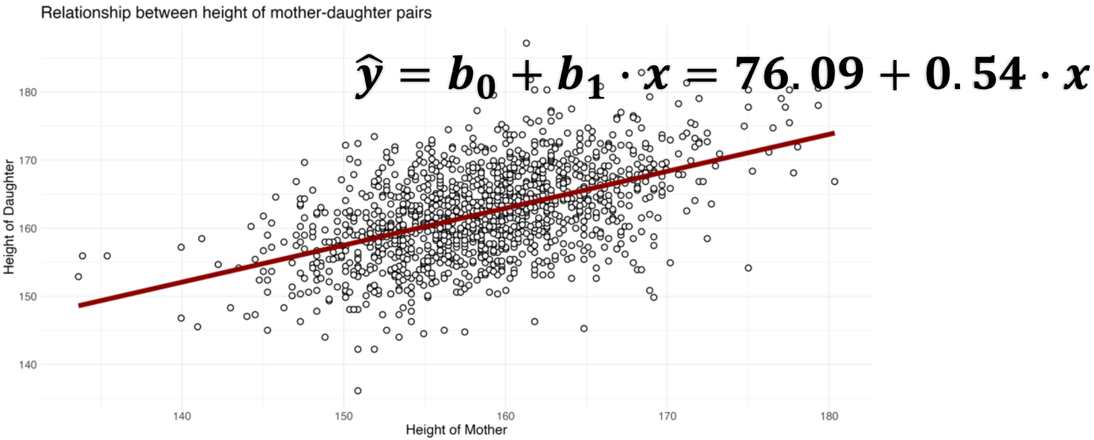
Euros is predicted.

REMARKS

- The computational effort is enormous for large data, especially for the auxiliary table. Leave such calculations to the computer; it does this faster and usually without errors.

- **Nevertheless, you should be able to perform a linear regression analysis using the OLS method either for a few data points or for given descriptive statistics.**

PEARSON-LEE MOTHER DAUGHTER DATA



$$\hat{y} = b_0 + b_1 \cdot x = 76.09 + 0.54 \cdot x$$

- Descriptive Statistics:

$$\bar{x} = 158.7 \; S_X = 6.25$$
$$\bar{y} = 162.2 \; S_Y = 6.67$$
$$r = 0.501 \quad n = 1375$$

Based on Pearson & Lee (1903)

PEARSON-LEE MOTHER DAUGHTER DATA

- **OLS Regression Coefficients for the Linear Regression of Y on X:**

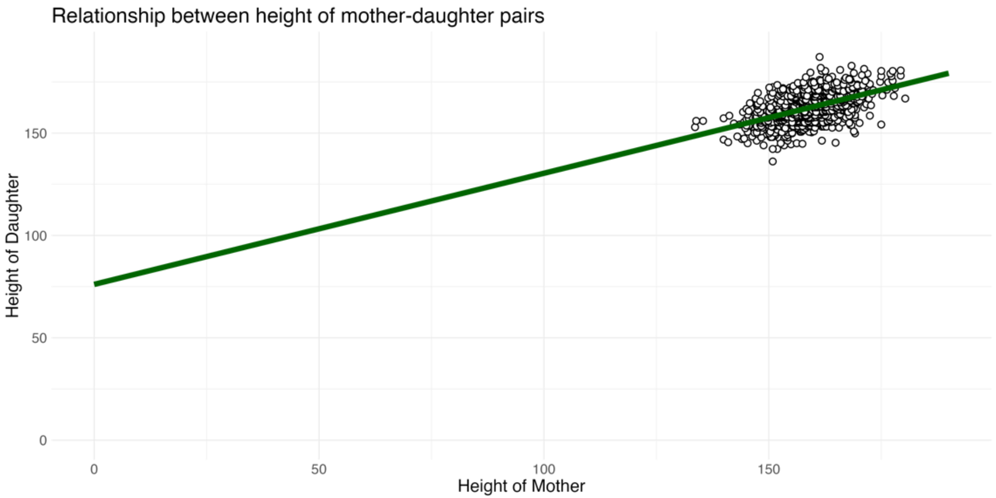
$$b_1 = r\frac{S_y}{S_X} \approx 0.501 \cdot \frac{6.67}{6.25} \approx 0.53$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 162.2 - 0.53 \cdot 158.7 = 78.09$$

- **Interpretation of $\boldsymbol{b_1 = 0.53}$:** If the height of fathers increases by 1 cm, the average height of sons increases by 0.53 cm.
- **Interpretation of $\boldsymbol{b_0}$:** $x = 0 \rightarrow \hat{y} = b_0 + b_1 \cdot 0 = 78.09$ The value $x = 0$ naturally makes no sense in this context, as it does not correspond to any observed value.
  - One could at most speculate that $b_0 \approx 78$ possibly describes the component of a son's height that is determined independently of the father's height.

Based on Pearson & Lee (1903)

PEARSON-LEE MOTHER DAUGHTER DATA



- However, if predictions are made for x-values that lie **far outside the observed data, this is called extrapolation**, and such predictions are generally not to be considered meaningful.
- If, on the other hand, predictions are made for x-values that **lie within or near the observed data, this is then also called interpolation.**
- Sensibly, one should limit predictions to interpolation or be able to put **forward theory-backed arguments for extrapolations.**

Based on Pearson & Lee (1903)

PEARSON-LEE MOTHER DAUGHTER DATA

- **OLS Regression Line for the Regression of Y on X:**

$$\hat{y} = 76.09 + 0.54 \cdot x$$

- **Question:** What is the OLS regression line for the regression of X on Y?
- **Naive Answer:**

$$\hat{y} = 76.09 + 0.54 \cdot x \Rightarrow x = \frac{\hat{y} - 76.09}{0.54} \approx -140{,}91 + 1.85 \cdot \hat{y}$$

- Fallacy: The equation on the right only describes how to determine an x such that an OLS regression of Y on X yields the prediction $\hat{y}$.
- What we want, however: The prediction $\hat{x}$ given y, if we regress X on Y.

Based on Pearson & Lee (1903)

PEARSON-LEE MOTHER DAUGHTER DATA

- **What is sought:** $\hat{x} = b'_0 + b'_1 y$
- **Regression Coefficients:**

$$b'_1 = r\frac{S_x}{S_y} \approx 0.501 \cdot \frac{6.25}{6.67} \approx 0.47$$

$$b'_0 = \bar{x} - b'_1 \cdot \bar{y} = 158.7 - 0.47 \cdot 162.2 = 82.47$$

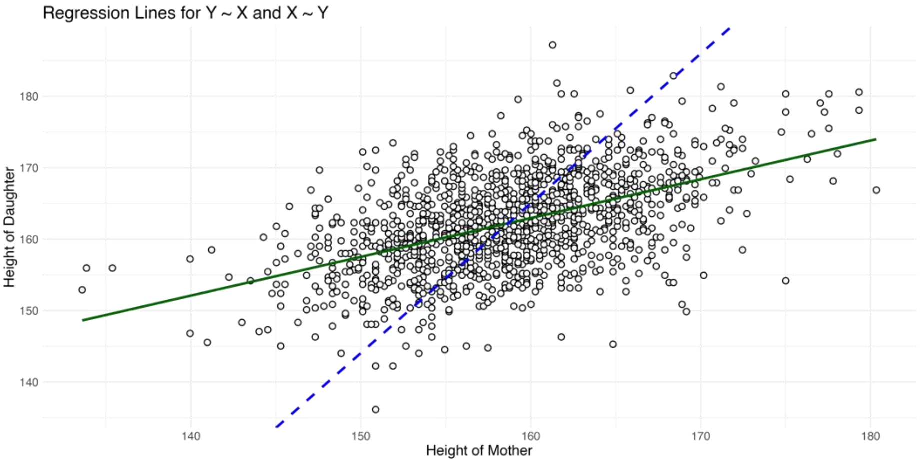
- **OLS Regression Line:**

$$\hat{x} = b'_0 + b'_1 \cdot y = 82.47 + 0.47 \cdot y$$

- The taller the son, the taller the father.

Based on Pearson & Lee (1903)

PEARSON-LEE MOTHER DAUGHTER DATA



Where does the difference come from?

- The OLS regression of Y on X minimizes the prediction errors for $X \rightarrow Y$

$$SSR := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

- whereas the OLS regression for $Y \rightarrow X$ minimizes the prediction errors

$$SSR' := \sum_{i=1}^{n} e'^2_i = \sum_{i=1}^{n} (x_i - (b_0 + b_1 y_i))^2$$

Based on Pearson & Lee (1903)

QUALITY ASSESSMENT

- How well can the dependent variable Y be explained or described by the independent variable X using the linear regression model?
- **Prediction Approach:** How much better can we predict Y if we additionally consider the knowledge about X (assuming the simple linear regression model)?
- **Model 1: Predicting Y without X**
  - If we did not know the value x_i for unit i a priori, then the mean of Y, $\bar{y}$, would in a way be our best estimate $\hat{y}_i$ for the specific value y_i that unit i takes for Y, i.e., $\hat{y}_i = \bar{y}$.
  - The **total error** made in this case (a posteriori) is: $SST := \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - This represents the **total variation** or total scatter of the y_i values, also known as the " (S)um of (S)quares (T)otal ."
- Remark: The mean, $\bar{y}$, indeed minimizes TSS.

QUALITY ASSESSMENT

- **Model 2: Predicting Y with X**
  - If we assume the simple linear regression model and determine the parameters using the OLS criterion, then our prediction is:

$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

  - The total error made in this case is:

$$SSR \coloneqq \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

    This is known as the **sum of squared residuals**(also referred to as residual variation or residual scatter).

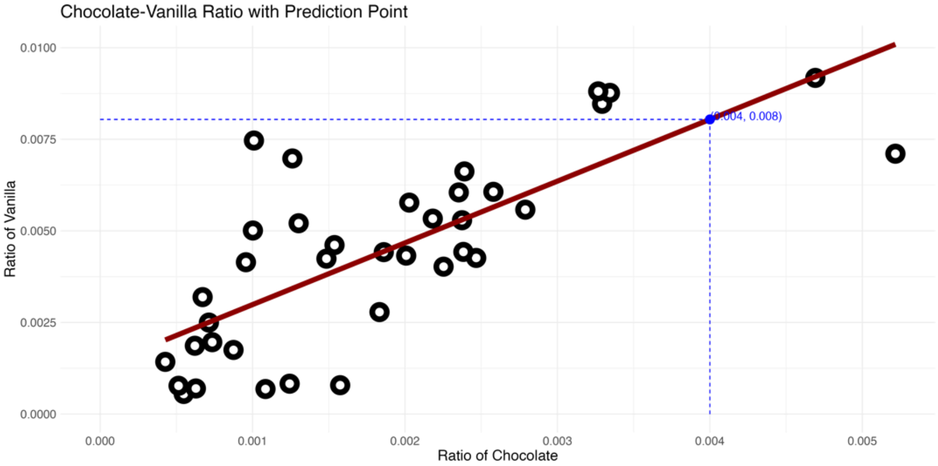
- The difference $SSE \coloneqq SST(without\ x) - SSR(with\ x)$ is called the **sum of squares explained**, representing the variation or variability in Y that is accounted for by the regression model.
- In the **best case**: $SSE = SST$ (meaning the model explains all the variation in Y) or $SSR = 0$ (meaning there are no prediction errors).
- In the **worst case**: $SSE = 0$ (meaning the model explains none of the variation in Y) or $SSR = SST$ (meaning the errors from the model are as large as the total variation in Y without the model).

QUALITY ASSESSMENT

- It holds that: $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$ and $SST = SSR + SSE$
- The prediction approach yields the **goodness-of-fit criterion**: $R^2 = \frac{SST-SSR}{SST} = \frac{SSE}{SST}$.
- R^2 is also called the **coefficient of determination**.
- It can be shown that $R^2 = \frac{SSE}{SST} = r^2$, meaning the **squared Bravais-Pearson correlation coefficient** can be regarded as a **goodness-of-fit measure** for how well the regression line adapts to or "fits" the data.

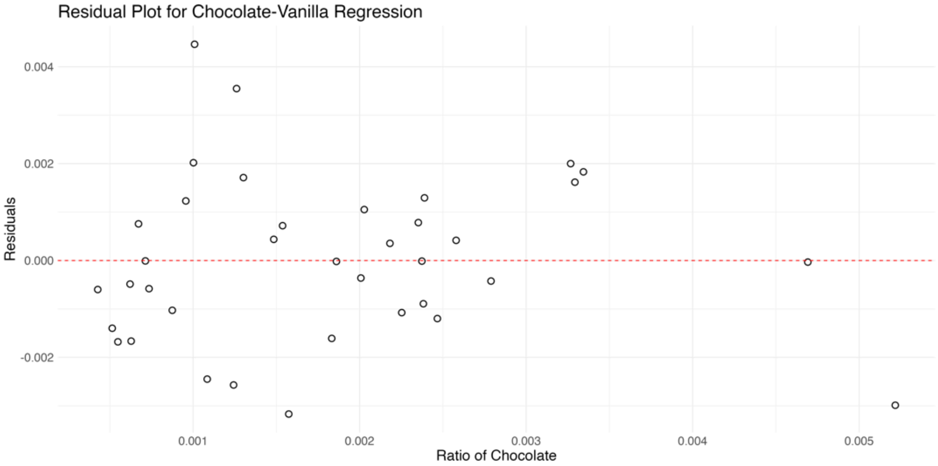
"VANILLA" EXAMPLE

- A coefficient of determination of $\boldsymbol{R^2 = 0.548}$ here means that 55% of the **variation** or **variability** in Vanilla usage rates can be explained by the (linear) regression on Chocolate usage rate.
- However, it can also be observed that there's a **systematic deviation** from the OLS line. These deviations can be better analyzed in a so-called **residual plot**.



“VANILLA” EXAMPLE

- A **residual plot** is a scatter plot showing the residuals (or prediction errors) e_i against their corresponding predictor values x_i.
- Although the residuals generally scatter around the zero line, for small x-values they tend to be larger, and for medium x-values they tend to be smaller.
- Interpretation? More dependence with more chocolate use?



UNIVERSITÄT
LEIPZIG

SEE YA‘LL NEXT WEEK!

Dr. Ing. Andreas Niekler

Computational Humanities

Paulinum, Augustusplatz 10, Raum P 616, 04109 Leipzig
T +49 341 97-32239
andreas.niekler@uni-leipzig.de
https://www.uni-leipzig.de/personenprofil/mitarbeiter/dr-andreas-niekler