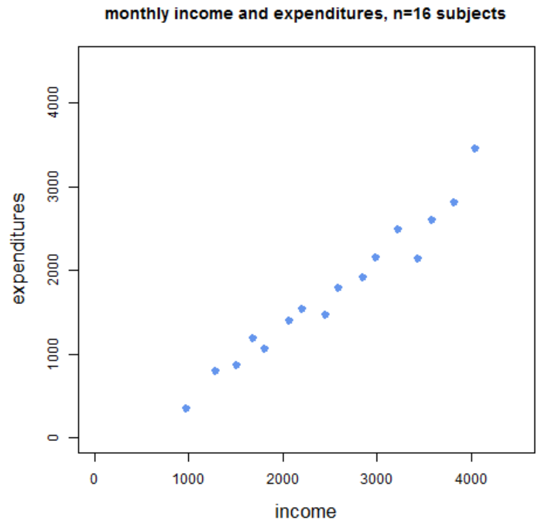
X. Linear Regression

- Linear Regression

Linear Regression (1/19)



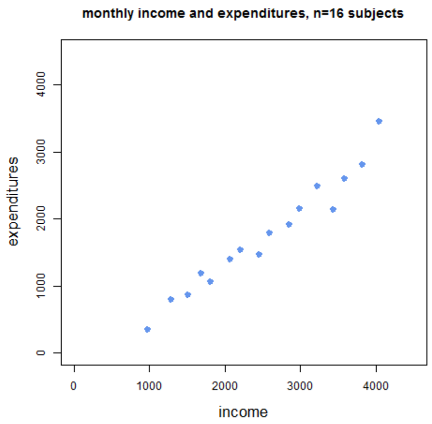
Linear Regression (2/19)

Mathematical Background

- We observed the data points $(x_1|y_1), (x_2|y_2), \cdots (x_n|y_n)$
- We want to find a linear function $\widehat{y} = \widehat{a} + \widehat{b} \cdot X$ that fits our data points *well.*
- **Idea (least squares method)**: Find $\widehat{a}$ and $\widehat{b}$ by minimizing the sum of the squared differences

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \to \text{min!}$$

(Minimize the squared differences between the observed and the fitted values The fitted values are the values on the straight line!)



Linear Regression (3/19)

Solving the problem

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \to \text{min!}$$

results in a formula for the estimators $\widehat{b}$ and $\widehat{a}$

Formula for the Estimator $\widehat{b}$:

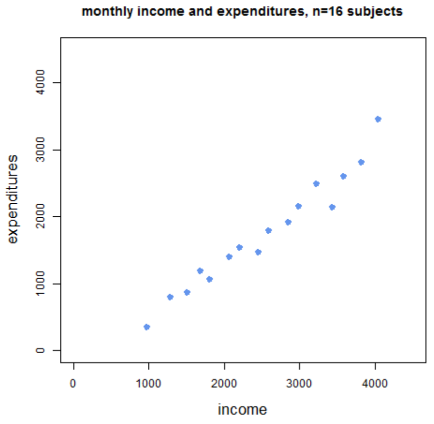
$$\widehat{b} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Formula for the Estimator $\widehat{a}$:

$$\widehat{a} = \overline{y} - \widehat{b} \cdot \overline{x}$$

Linear Regression (4/19)

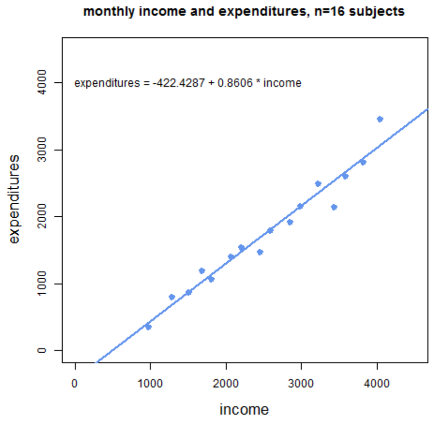
Example and Raw Data



	income	expenditures
1	967.00	348.95
2	1286.00	801.22
3	1506.00	872.88
4	1675.00	1187.74
5	1798.00	1060.32
6	2062.00	1402.01
7	2197.00	1542.92
8	2453.00	1470.94
9	2581.00	1799.03
10	2852.00	1915.53
11	2981.00	2158.98
12	3215.00	2491.62
13	3434.00	2143.57
14	3585.00	2602.28
15	3824.00	2806.59
16	4044.00	3456.30

Linear Regression (5/19)

Example, Raw Data and Regression Line



	income	expenditures
1	967.00	348.95
2	1286.00	801.22
3	1506.00	872.88
4	1675.00	1187.74
5	1798.00	1060.32
6	2062.00	1402.01
7	2197.00	1542.92
8	2453.00	1470.94
9	2581.00	1799.03
10	2852.00	1915.53
11	2981.00	2158.98
12	3215.00	2491.62
13	3434.00	2143.57
14	3585.00	2602.28
15	3824.00	2806.59
16	4044.00	3456.30

Linear Regression (6/19)

Assessing the Quality of a Regression Model (Overview)

1. correlation coefficient
2. coefficient of determination
3. standardized residuals
4. homoscedasticity and heteroscadasticity

Linear Regression (7/19)

Assessing the Quality of a Regression Model: Correlation Coefficient (1/3)

The Pearson correlation coefficient **r** is a measure of the linear correlation between two variables X and Y.

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{s_x \cdot s_y}$$

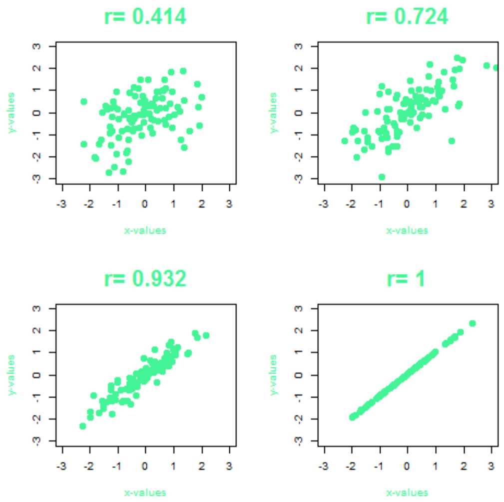
The expression in the nominator is called the **sample covariance** of x and y.

Properties of the Correlation Coefficient

- $-1 \leq r \leq 1$
- If r is positive then y tends to increase linearly as x increases (positive slope of the regression line). If r is negative, then y tends to decrease linearly as x increases (negative slope).
- A value of r close to -1 or +1 represents a **strong** linear relationship.
- A value of r close to 0 represents a **weak** linear relationship.
- Extreme case: If r is 1 (or -1), then all the data points are on the regression line with a positive (or a negative) slope. This would be a perfect linear relationship.

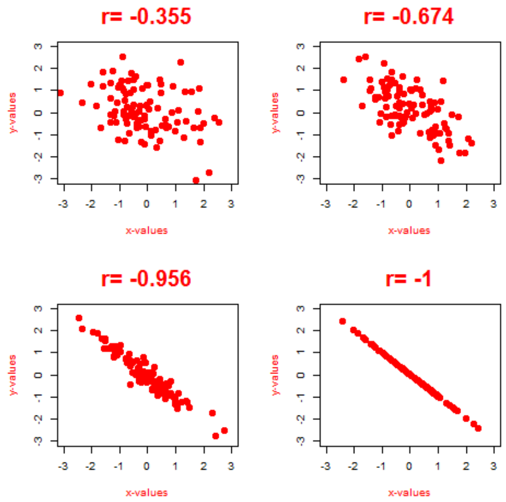
Linear Regression (8/19)

Assessing the Quality of a Regression Model: Correlation Coefficient (2/3): Examples of Positive Correlations.



Linear Regression (9/19)

Assessing the Quality of a Regression Model: Correlation Coefficient (3/3): Examples of Negative Correlations.



Linear Regression (10/19)

Assessing the Quality of a Regression Model: Coefficient of Determination

The coefficient of determination is the square of the correlation coefficient. r^2 is a statistic that will give some information about the goodness of fit of a model.

$$0 \leq r^2 \leq 1$$

r^2 is the fraction of the variance in Y that is explained by the regression model.

Example: Coefficient of Determination

If we measure a correlation between two random variables X and Y of r=-0.8 (a negative correlation), we get $r^2 = (-0.8)^2 = 0.64$. This means that our regression model accounts for 64% of the variance of the variable Y.

Linear Regression (11/19)

Assessing the Quality of a Regression Model: Standardized Residuals

The residuals ε_i are the differences between the observed and the fitted data points.
If we apply the z-Transformation

$$z\varepsilon_i = \frac{\varepsilon_i - \overline{\varepsilon}}{s_\varepsilon}$$

We get the **standardized residuals**. The standardized residuals have a mean of 0 and a standard deviation of 1. If the residuals follow a **normal distribution**, the standardized residuals will follow a **standard normal distribution**. It can be shown that in the linear regression model the sum of the residuals (and therefore their mean) equals 0, hence the formula can be simplified as follows:

$$z\varepsilon_i = \frac{\varepsilon_i}{s_\varepsilon}$$

A rule of thumb is that the regression model *fits well* if the standardized residuals are within the interval $[-2.5, 2.5]$. Data points outside this interval are considered to be outliers.

Linear Regression (12/19)

Assessing the Quality of a Regression Model: Homoscedasticity and Heteroscadasticity

- **Homoscadasticity:** The standard deviations of the error terms are constant and do not depend on the x-value. This is a **desirable feature** in regression analysis!
- **Heteroscedasticity:** One of the assumptions of the classical linear regression model is that there is no heteroscedasticity. Note that heteroscedasticity means that the variance of the error term correlates with the regressor. This clearly is an **unwanted feature** in regression analysis!

Linear Regression (13/19)

EXCEL and Linear Regression

=INTERCEPT(Yvalues; Xvalues)	*intercept*
=SLOPE(Yvalues; Xvalues)	*slope*
=CORREL(Matrix1; Matrix2)	*correlation coefficient*
=COVARIANCE(Matrix1; Matrix2)	*sample covariance of X and Y*

Alternatively, the regression procedure in Excel can be called directly from the `Data` Ribbon:

`Data` → `Data Analysis` → `Regression`

Linear Regression (14/19)

Excel: Calculating the Coefficients

	A	B	C
1	income (X)	expenditures (Y)	
2	€ 967,00	€ 348,95	
3	€ 1.286,00	€ 801,22	
4	€ 1.506,00	€ 872,88	
5	€ 1.675,00	€ 1.187,74	
6	€ 1.798,00	€ 1.060,32	
7	€ 2.062,00	€ 1.402,01	
8	€ 2.197,00	€ 1.542,92	
9	€ 2.453,00	€ 1.470,94	
10	€ 2.581,00	€ 1.799,03	
11	€ 2.852,00	€ 1.915,53	
12	€ 2.981,00	€ 2.158,98	
13	€ 3.215,00	€ 2.491,62	
14	€ 3.434,00	€ 2.143,57	
15	€ 3.585,00	€ 2.602,28	
16	€ 3.824,00	€ 2.806,59	
17	€ 4.044,00	€ 3.456,30	
18			
19			
20	INTERCEPT	-422,428728	=INTERCEPT(B2:B17; A2:A17)
21	SLOPE	0,86059663	=SLOPE(B2:B17; A2:A17)

Linear Regression (15/19)

Linear Regression Using Excel: Fitted Values and Residuals

	A	B	C	D
1	income (X)	expenditures (Y)	fitted values	residuals
2	€ 967,00	€ 348,95	€ 409,77	-€ 60,82
3	€ 1.286,00	€ 801,22	€ 684,30	€ 116,92
4	€ 1.506,00	€ 872,88	€ 873,63	-€ 0,75
5	€ 1.675,00	€ 1.187,74	€ 1.019,07	€ 168,67
6	€ 1.798,00	€ 1.060,32	€ 1.124,92	-€ 64,60
7	€ 2.062,00	€ 1.402,01	€ 1.352,12	€ 49,89
8	€ 2.197,00	€ 1.542,92	€ 1.468,30	€ 74,62
9	€ 2.453,00	€ 1.470,94	€ 1.688,61	-€ 217,67
10	€ 2.581,00	€ 1.799,03	€ 1.798,77	€ 0,26
11	€ 2.852,00	€ 1.915,53	€ 2.031,99	-€ 116,46
12	€ 2.981,00	€ 2.158,98	€ 2.143,01	€ 15,97
13	€ 3.215,00	€ 2.491,62	€ 2.344,39	€ 147,23
14	€ 3.434,00	€ 2.143,57	€ 2.532,86	-€ 389,29
15	€ 3.585,00	€ 2.602,28	€ 2.662,81	-€ 60,53
16	€ 3.824,00	€ 2.806,59	€ 2.868,49	-€ 61,90
17	€ 4.044,00	€ 3.456,30	€ 3.057,82	€ 398,48
18				
19				
20	INTERCEPT	-422,428728	=INTERCEPT(B2:B17; A2:A17)	
21	SLOPE	0,86059663	=SLOPE(B2:B17; A2:A17)	

Linear Regression (16/19)

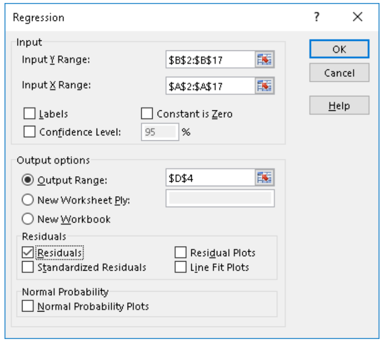
Linear Regression Using Excel: Fitted Values and Residuals (Formula View)

	A	B	C	D
1	income (X)	expenditures (Y)	fitted values	residuals
2	967	348,95	=$B$20+$B$21*A2	=B2-C2
3	1286	801,22	=$B$20+$B$21*A3	=B3-C3
4	1506	872,88	=$B$20+$B$21*A4	=B4-C4
5	1675	1187,74	=$B$20+$B$21*A5	=B5-C5
6	1798	1060,32	=$B$20+$B$21*A6	=B6-C6
7	2062	1402,01	=$B$20+$B$21*A7	=B7-C7
8	2197	1542,92	=$B$20+$B$21*A8	=B8-C8
9	2453	1470,94	=$B$20+$B$21*A9	=B9-C9
10	2581	1799,03	=$B$20+$B$21*A10	=B10-C10
11	2852	1915,53	=$B$20+$B$21*A11	=B11-C11
12	2981	2158,98	=$B$20+$B$21*A12	=B12-C12
13	3215	2491,62	=$B$20+$B$21*A13	=B13-C13
14	3434	2143,57	=$B$20+$B$21*A14	=B14-C14
15	3585	2602,28	=$B$20+$B$21*A15	=B15-C15
16	3824	2806,59	=$B$20+$B$21*A16	=B16-C16
17	4044	3456,3	=$B$20+$B$21*A17	=B17-C17
18				
19				
20	**INTERCEPT**	**-422,43**		
21	**SLOPE**	**0,86**		

Linear Regression (17/19)

Regression Using the Data Analysis Toolbox

	A	B
1	income (X)	expenditures (Y)
2	€ 967,00	€ 348,95
3	€ 1.286,00	€ 801,22
4	€ 1.506,00	€ 872,88
5	€ 1.675,00	€ 1.187,74
6	€ 1.798,00	€ 1.060,32
7	€ 2.062,00	€ 1.402,01
8	€ 2.197,00	€ 1.542,92
9	€ 2.453,00	€ 1.470,94
10	€ 2.581,00	€ 1.799,03
11	€ 2.852,00	€ 1.915,53
12	€ 2.981,00	€ 2.158,98
13	€ 3.215,00	€ 2.491,62
14	€ 3.434,00	€ 2.143,57
15	€ 3.585,00	€ 2.602,28
16	€ 3.824,00	€ 2.806,59
17	€ 4.044,00	€ 3.456,30



Linear Regression (18/19)

Linear Regression Using Excel: Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,978
R Square	0,956
Adjusted R Square	0,952
Standard Error	181,020
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1,00	9863317,17	9863317,17	301,00	7,32E-11
Residual	14,00	458753,38	32768,10		
Total	15,00	10322070,55			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-422,429	133,349	-3,168	0,007	-708,434	-136,423
X Variable 1	0,861	0,050	17,349	0,000	0,754	0,967

Linear Regression (19/19)

Possible Pitfalls in Regression Analysis

- The coefficient of determination is the squared correlation coefficient and is sometimes written as a percentage. If the correlation is e.g. r=0.9 ($r^2 = 81\%$) it is simply the amount of variance explained. It does **not** mean that 81% of the data points are on the regression line (!)
- A correlation of 0 means that there is no **linear** correlation beween the variables X and Y. It tells us nothing about non-linear relationships between the variables. Thus, the interpretation „there is no relationship“ is wrong.
- The residuals are the differences between the data points and the line in y-direction only. If you draw them in a plot, they are parallel to the y-axis (and **not** orthogonal to the regression line!)
- Before estimating a linear model always do a graph (a scatterplot) to see if a straight line is adequate for your data points.