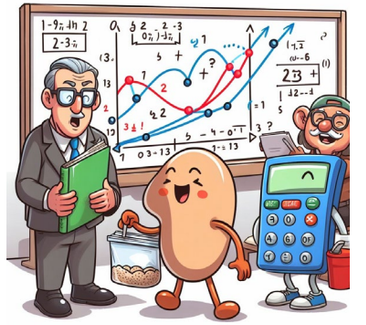
Introduction to probability, statistics and data handling

Agnieszka Obłąkowska-Mucha
based on lectures by Tomasz Szumlak

Faculty of Physics and Applied Computer Science
AGH University of Krakow

Curve fitting and regression



- In many practical problems, when collecting data, we may find that two (or more) R.V.s may **exhibit a relationship**

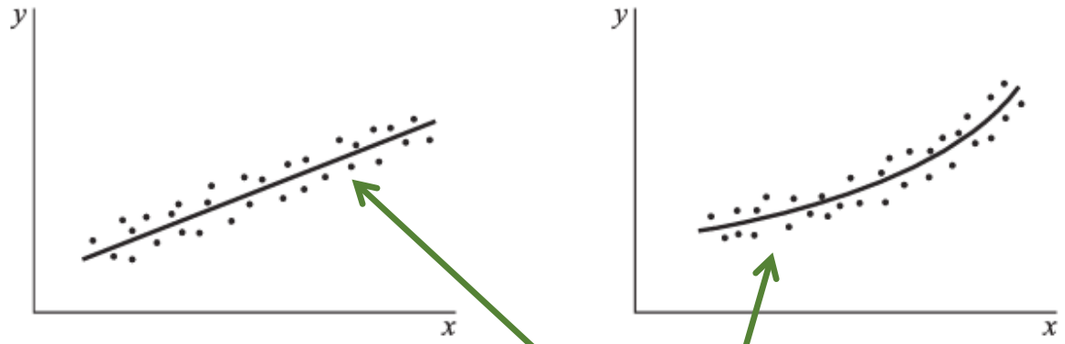
- It seems so natural to exploit this and express this fact using a mathematical function (model)
- The trick here would be to find the model that **FITS the best** our data (we also say that it connects the R.V.s)

> Regression is a statistical approach used to analyze the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). The objective is to determine the most suitable function that characterizes the connection between these variables.
> It seeks to find the best-fitting model, which can be utilized to make predictions or draw conclusions.

- Although this technique is well established and used, still **some experience is needed** when we want to choose the right model (this also may be driven by the physics of the phenomena, e.g., radioactive decay)
- NOTE, we may sometimes, when the relation is very complicated, or we are dealing with many dimensions, use the **machine learning approach** – in fact the linear regression can also be treated as machine learning.

Regression - first steps

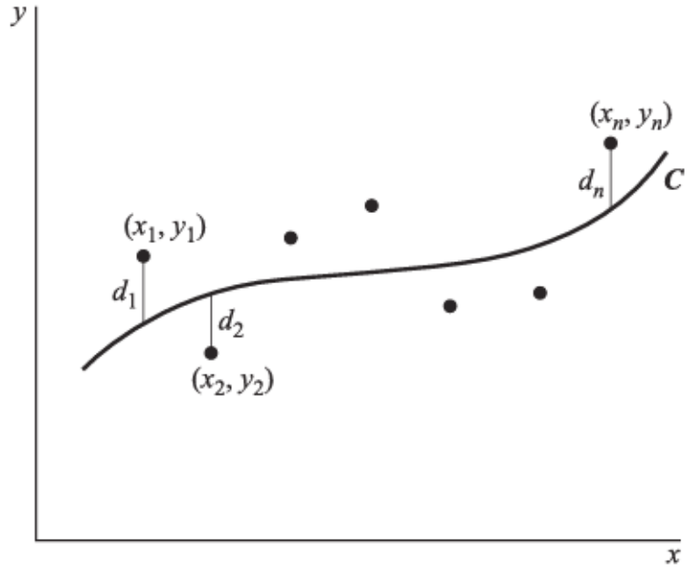
- Usually, we make first the **scatter plot** using collected data and take a look…



- The solid lines above are called approximating curves and the approximation of parameters
- What we need to work out is the **equation of this curve, ex.** $y = a + bx$**,** the variable ***x*** **is the independent variable, and *y* is the dependent variable.**
- That task is called **curve fitting.**
- Often to understand the relation, we may need to apply some transformation(s) to the variables

Least squares again!

- We have already encountered that technique when discussed general estimation theory. The idea is again the same we are going to minimise squares of residuals
- Again, the goal is to estimate a bunch of parameters but this time this is going to lead us to bit different result
- Lets define the data as pairs: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, next we make the scatter plot



$$(x_1, y_1) \rightarrow d_1$$

$$(x_2, y_2) \rightarrow d_2$$

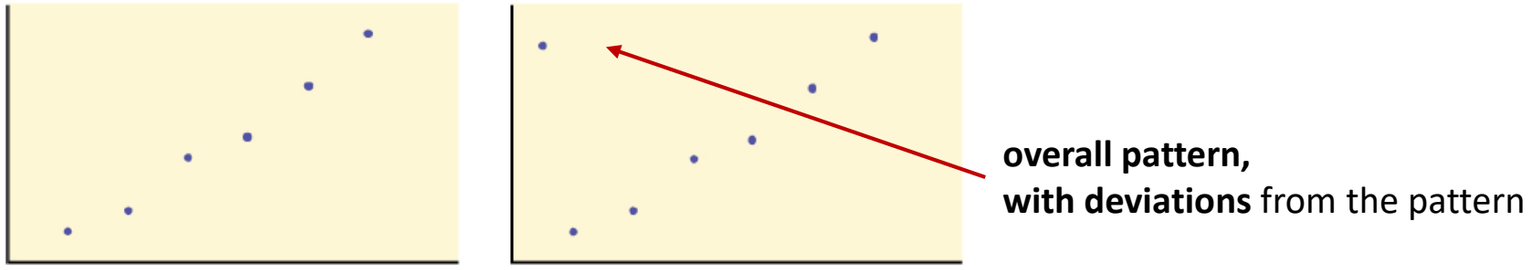
$$\vdots$$

$$(x_i, y_i) \rightarrow d_i$$

$$\Delta = min\left\{\sum_i d_i^2\right\}$$

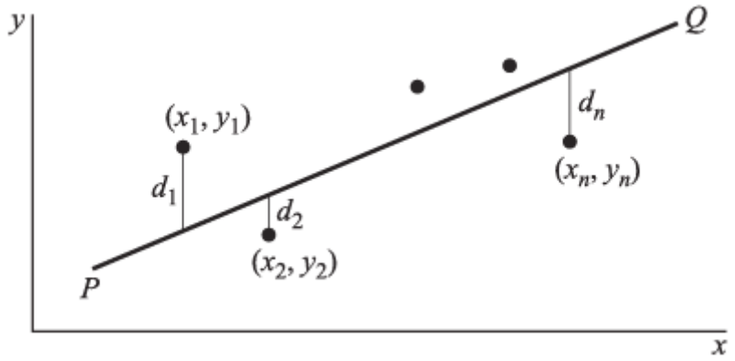
Least squares

- So, if we pick-up a given family of approximating curves the one with the property $\Delta = min\{\sum_i d_i^2\}$ will be **the best fitting or least-squares curve**
- Certainly, we can also discriminate between families (for instance the linear model or parabola)
- Silently, we assume that the uncertainties of the independent (x) variable is much smaller than on (y) variable
- Formally we can also switch the axes (treat the y variable as independent)
- Let's start discussing the linear model fit



LS line

- Here, we consider that our data set show linear dependency, which we denote as: $\boldsymbol{y = a_0 + a_1 x}$ (we will call a_0 the intercept and a_1 slope or gradient)
- To determine the parameters, we need to solve



regression line from the sample is our best estimate of this line in the population

$$d_i = a_0 + a_1 x_i - y_i$$

$$\Delta = \sum_i d_i^2 = \sum_i (a_0 + a_1 x_i - y_i)^2$$

$$\Delta = \Delta(a_0, a_1) \rightarrow \boldsymbol{\frac{\partial \Delta}{\partial a_0} = 0, \frac{\partial \Delta}{\partial a_1} = 0}$$

LS line – normal equations

- Searching for the extremum we get:

$$\frac{\partial \Delta}{\partial a_0} = \sum_i 2 \cdot (a_0 + a_1 x_i - y_i) = 0$$

$$\frac{\partial \Delta}{\partial a_1} = \sum_i 2 \cdot x \cdot (a_0 + a_1 x_i - y_i) = 0$$

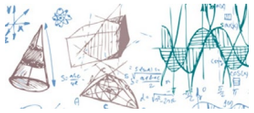
$$\sum_i y_i = a_0 n + a_1 \sum_i x_i \qquad \sum_i x_i y_i = a_0 \sum_i x_i + a_1 \sum_i x_i^2$$

- These two we call the normal equation for the LS line

$$\boldsymbol{a_0} = \frac{\sum y \cdot \sum x^2 - \sum x \cdot \sum xy}{n \cdot \sum x^2 - (\sum x)^2} \qquad \boldsymbol{a_1} = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2}$$

$$\sum x \equiv \sum_i x_i \, , etc.$$

LS line – normal equations



- The second equation can be written in more convenient way:

$$a_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

This „looks like" covariance

This „looks like" variance

- We can divide the first normal equation by n

$$\frac{1}{n}\sum_i y_i = \frac{1}{n}\left(a_0 n + a_1 \sum_i x_i\right) \rightarrow \bar{y} = a_0 + a_1\bar{x} \rightarrow \boldsymbol{a_0 = \bar{y} - a_1\bar{x}}$$

- And, we can write the LS line as:

$$\boldsymbol{y - \bar{y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}(x-\bar{x})}$$

- This is an interesting result, since it shows clearly that the LS line goes through the point $(\bar{x}, \bar{y})$- sample means - it is called **the centroid of the data**

LS line – simple(r) way

- The SL line equation can be simplified using the sample variance and covariance

$$s_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}, s_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}, s_{xy} = \frac{\sum_{i,j} (x_i - \bar{x})(y_j - \bar{y})}{n-1}$$

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \qquad x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

- And with the sample correlation coefficient one can express the slope: $r = \frac{s_{xy}}{s_x s_y}$

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right) \qquad \frac{x - \bar{x}}{s_x} = r\left(\frac{y - \bar{y}}{s_y}\right) \qquad z_y = r z_x$$

- This is of outmost interest – the lines that are obtained for (x, y) pairs will be in general different than for (y, x) pairs
- The equivalence is possible only when the correlation coefficient is $r = \pm 1$

Correlation Coefficient- interpretation

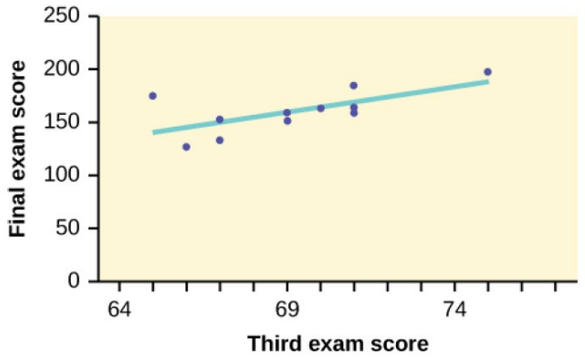
The variable r^2 is called the coefficient of determination

• r^2 , when expressed as a percent, represents the percent of variation in the dependent (predicted) variable *y* that can be explained by variation in the independent (explanatory) variable *x* using the regression (best-fit) line.

• $1 - r^2$, when expressed as a percentage, represents the percent of variation in *y* that is NOT explained by variation in *x* using the regression line

EX: The line of best fit is: $\hat{y} = -173.51 + 4.83x$

- ✓ The correlation coefficient is $r = 0.6631$
- ✓ The coefficient of determination is $r^2 = 0.66312 = 0.4397$



Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.

Therefore, approximately 56% of the variation (1 – 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

SL parabola

- Following the same idea we can get employ more complicated model, for instance if we use the parabola equation: $y = a_0 + a_1 x + a_2 x^2$
- Now the sum of the square of residuals will lead to three normal equations for each of the parameters a_i

$$\sum y = n a_0 + a_1 \sum x + a_2 \sum x^2$$

$$\sum xy = a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3$$

$$\sum x^2 y = a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4$$

- Usually, for more complicated models we use computer libraries to make the calculations for us or machine learning approach

Multiple regression

- It is just as easy to extend this idea to higher dimensions, for instance the dependence between 3 R.V.s

$$z = a + a_x x + a_y y, or\ x_3 = a_0 + a_1 x_1 + a_2 x_2$$

- Formally, this is a **plane equation**, thus, we call it **the regression plane**. Again we can use the least-squares principle to find our normal equations

$$\sum z = na + a_x \sum x + a_y \sum y$$

$$\sum xz = n \sum x + a_x \sum x^2 + a_y \sum xy$$

$$\sum yz = n \sum y + a_x \sum xy + a_y \sum y^2$$

- It is quite popular in the domain of machine learning

Estimate error

- As usual, we should take into account that the job is not yet done if we do not give error on the estimated parameters
- We can define the „standard” error of the estimate

$$s_{y|x} = \sqrt{\frac{\sum(y - y_{th})^2}{n-1}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-1}}$$

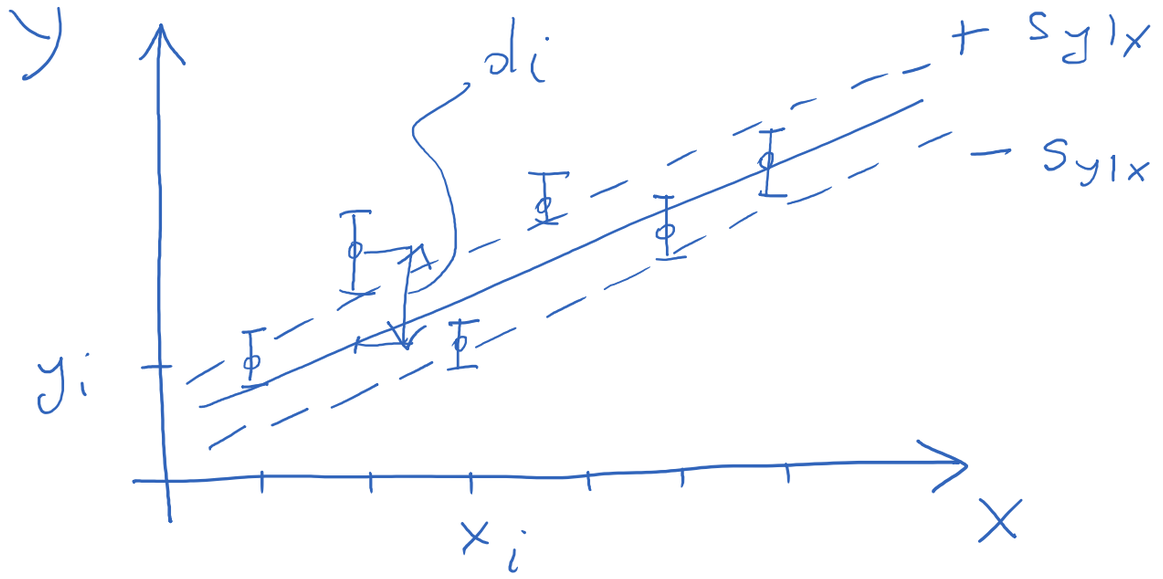
- Where y_{th} $(\hat{y})$ denotes the value calculated using the estimated line (sometimes it is called theory point)
- We see immediately that the LS curve will have the smallest standard error of estimate

$$s_{y|x}^2 = \frac{\sum y^2 - a_0 \sum y - a_1 \sum xy}{n-1}$$

- This estimator, has properties similar to those of standard deviation

Estimate error

- This analogy can be made a bit more intuitive: if we draw a pair of lines parallel to the LS line at respective vertical distances of $\pm s_{y|x}$ then we should expect that about 68% of the sampling point will be between them
- It is then easy to extend this for distances of $\pm 2s_{y|x}$ and $\pm 3s_{y|x}$



Linear correlation coefficient

- The square of the standard error of estimate can be written as:

$$s_{y|x}^2 = \frac{\sum(y-\bar{y})^2 - a_1 \sum(x-\bar{x})(y-\bar{y})}{n}$$

- And using the variance and correlation coefficient $s_{y|x}^2 = s_y^2(1-r^2)$
- Combining these definitions we have:

$$r^2 = 1 - \frac{\sum(y-y_{th})^2}{\sum(y-\bar{y})^2} = \frac{\sum(y-\bar{y})^2 - \sum(y-y_{th})^2}{\sum(y-\bar{y})^2}$$

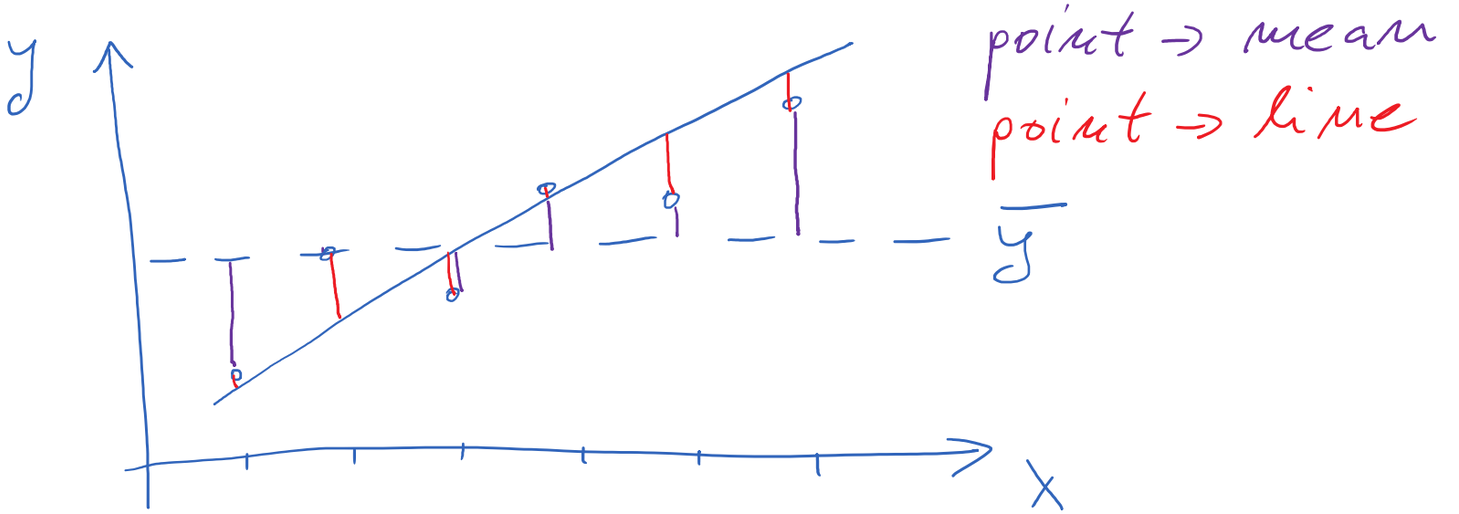
- Also: $\sum(y-\bar{y})^2 = \sum(y-y_{th})^2 + \sum(y_{th}-\bar{y})^2$, and combining with the above we get

$$r^2 = \frac{\sum(y_{th}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{explained\ variation}{total\ variation}$$

Explained variation

- Machine learning very often uses the r-squared as a **specific goodness of fit metric**. It is also much easier to interpret

$$r^2 = 1 - \frac{\sum(y - y_{th})^2}{\sum(y - \bar{y})^2} = \frac{\sum(y - \bar{y})^2 - \sum(y - y_{th})^2}{\sum(y - \bar{y})^2} = \frac{V[mean] - V[line]}{V[mean]}$$



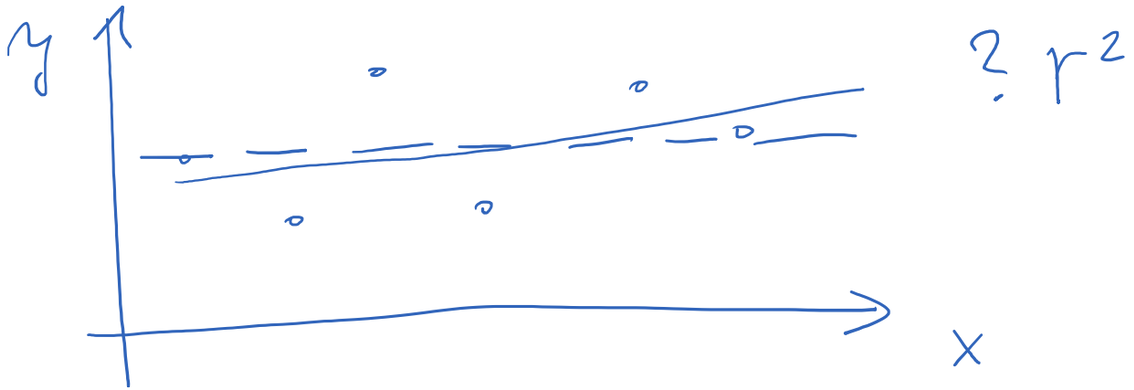
- The „red residuals” will never be larger than the „purple residuals”, so the r-squared will be always between 0 and 1 or we can measure it in percents

Explained variation

- Say, for the previous example we obtained: $V[mean] = 36$ and $V[line] = 8$, thus

$$r^2 = \frac{V[mean] - V[line]}{V[mean]} = \frac{36 - 8}{36} = 0.78$$

- We say, that most of the variation (78%) seen in our data can be explained by the line – or by the relationship between our variables. So, applying the model makes sense!



- Here making the fit does not make much sense…

Correlation Coefficient – test of significance

- The correlation coefficient, *r*, tells us about the strength and direction of the linear relationship between *x* and *y*.
- We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

ρ = population correlation coefficient (unknown)

r = sample correlation coefficient (known; calculated from sample data)

Null Hypothesis: $\boldsymbol{H_0: \rho = 0}$, The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between *x* and *y* in the population.

Alternate Hypothesis: $\boldsymbol{H_a: \rho \neq 0}$, The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between *x* and *y* in the population.

Correlation Coefficient – test of significance

Method 1: Using the *p*-value, we usually take significance level of 5%, $\alpha = 0.05$

- p-value is calculated with a t-distribution with n-2 degrees of freedom.
- test statistics is: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- If the *p*-value is less than the significance level ($\alpha = 0.05$):
  - ✓ Decision: Reject the null hypothesis.
  - ✓ Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is significantly different from zero”
- If the *p*-value is NOT less than the significance level ($\alpha = 0.05$)
  - ✓ Decision: DO NOT REJECT the null hypothesis.
  - ✓ Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between *x* and *y* because the correlation coefficient is NOT significantly different from zero."

EX: The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. The *p*-value is 0.026, what is the decision?

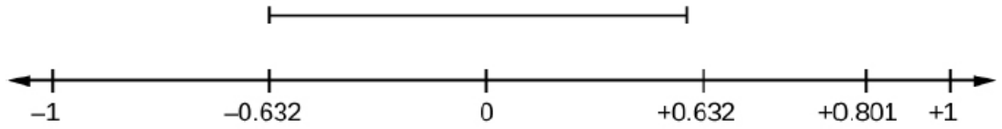
Correlation Coefficient – test of significance

Method 2: Using a table of critical values, $\alpha = 0.05$

The **95% Critical Values of the Sample Correlation Coefficient Table** can be used to give you a good idea of whether the computed value of ***r* is significant or not**.

If *r* is not between the positive and negative critical values, then the correlation coefficient is significant. If *r* is significant, then you may want to use the line for prediction.

EX: Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and + 0.632. If $r <$ negative critical value or $r >$ positive critical value, then ***r* is significant**.
Since $r = 0.801$ and $0.801 > 0.632$, *r* is significant and the line may be used for prediction.



Correlation Coefficient – test of significance

Method 2: Using a table of critical values, $\alpha = 0.05$

THIRD-EXAM vs FINAL-EXAM EXAMPLE: critical value method

Consider the **third exam/final exam example**. The line of best fit is: $\hat{y} = -173.51+4.83x$ with $r = 0.6631$ and there are $n = 11$ data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

$H_0: \rho = 0$
$H_a: \rho \neq 0$
$\alpha = 0.05$

- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$.
- The critical values are –0.602 and +0.602
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion:There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Sampling properties of LS

- The complete treatment of the estimated parameters' uncertainty is out of the scope of this lecture. I just give you somewhat simplified version – but still important and perfectly usable in practice!
- Assume that our model can be written as follow:

$$y_i = a_0 + a_1 x_i + d_i$$

- We assume that the residuals are independent of one another and are distributed with $\mu = 0$ and $V[d_i] = \sigma^2$
- Let's start discussing the constraint fit (with the point $(0,0)$), we know that the best estimate of the slope in that case is

$$\hat{a}_1 = \frac{\sum xy}{\sum x^2}$$

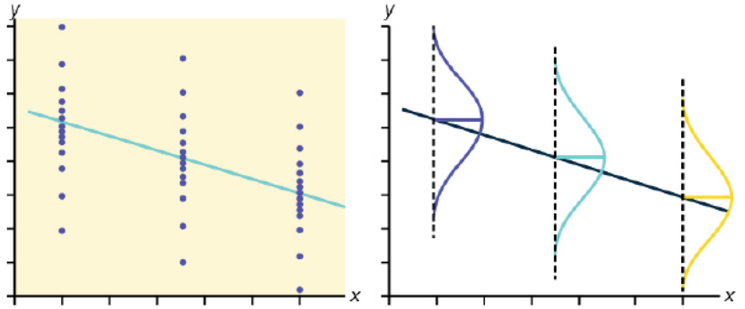
- It is quite easy to show, that this estimator is unbiased, $\boldsymbol{E[\hat{a}_1] = a_1}$ and its variance $\boldsymbol{V[\hat{a}_1] = \frac{\sigma^2}{\sum x^2}}$

Sampling properties of LS

- The variance of the slope estimator says a few interesting things – to improve the estimate, we should take more data, improve the precision of measurement of each data point (expressed as σ^2) and the most important measurements are done close to the origin point!

- If we assume that the residuals are normally distributed (not too crazy assumption), we find that the slope estimator is also normally distribu

$$\hat{a}_1 \sim \mathcal{N}\left(a_1, \frac{\sigma^2}{\sum_i x_i^2}\right)$$



The standard deviations of the population *y* values about the line are equal for each value of *x*. In other words, each of these normal distributions of *y* values has the same shape and spread about the line.

- In case, when the distribution is not Gaussian or the variance is not known one can use the regression estimation variance

$$s_{y|x} = \sqrt{\frac{\sum(y - y_{th})^2}{n-1}} \rightarrow s_{y|x}^2 = \frac{\sum(y - y_{th})^2}{n-1} \rightarrow \boldsymbol{s_{y|x}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-1}}$$

Sampling properties of LS

- We are familiar with this estimator already! It follows the χ^2 distribution:

$$\frac{(n-1)s^2_{y|x}}{\sigma^2} \sim \chi^2(n-1)$$

- And, in that case the distribution of the slope estimator is:

$$\frac{\hat{a}_1 - a_1}{s_{y|x} / \sqrt{\sum_i x_i^2}} \sim t(n-1)$$

- Now we can calculate the C.I. for the slope and even test hypothesis related to estimated slope!

C.I. for the slope

- Ex. 1 Let's assume that we performed the procedure of LS line estimation and obtained $\hat{a}_1 = 1.289$, the sum of residual squares are $\sum_i (y_i - \hat{y}_i)^2 = 107.30$ and the number of points in our data set was 20.
- The estimate variance:

$$s^2_{y|x} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-1} = \frac{107.30}{19} = 5.647$$

- So, the 90% C.I. for the slope can be constructed as follow:

$$C.I.^{t}_{90\%}(\hat{a}_1) = \left(\hat{a}_1 - t_{0.9} \frac{s_{y|x}}{\sqrt{\sum_i x_i^2}}, \hat{a}_1 + t_{0.9} \frac{s_{y|x}}{\sqrt{\sum_i x_i^2}} \right) =$$

$$= \left(1.289 - 1.729 \frac{\sqrt{5.647}}{\sqrt{6226.38}}, \hat{a}_1 + 1.729 \frac{\sqrt{5.647}}{\sqrt{6226.38}} \right) = (1.237, 1.341)$$

Unconstrained fit

- If there is no special points in our fitting procedure we need to proceed with the unconstrained fit. In this case it can be shown

$$E[\hat{a}_0] = \boldsymbol{a_0} \qquad V[\hat{a}_0] = \frac{\sigma^2}{n}$$

$$E[\hat{a}_1] = \boldsymbol{a_1} \qquad V[\hat{a}_1] = \frac{\sigma^2}{\sum_i (x_i - \bar{x}_i)^2}$$

Again, the larger the interval where we measure data the more precise the estimate

- The variance of the estimate

$$\boldsymbol{s^2_{y|x} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}}$$

- If we assume that the residuals are normally distributed

$$\hat{a}_1 \sim \mathcal{N}\left(a_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x}_i)^2}\right), \qquad \frac{(n-2)s^2_{y|x}}{\sigma^2} \sim \chi^2(n-2)$$

- Otherwise

$$\frac{\hat{a}_1 - a_1}{s_{y|x}/\sqrt{\sum_i (x_i - \bar{x}_i)^2}} \sim t(n-2)$$