COMP229: Introduction to Data Science
Lecture 15: Simple Linear Regression (SLR)

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Lecture plan

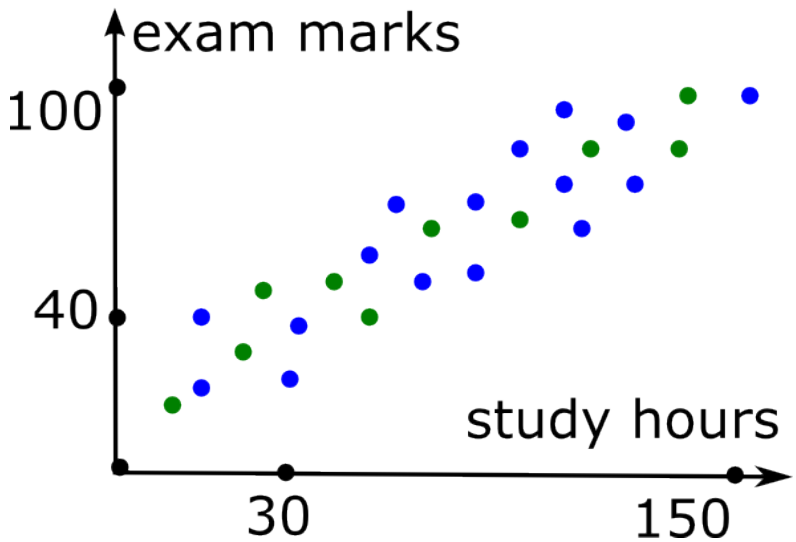
- Simple linear regression
- Applications
- Limitations

Reminder: correlation

- The *scatterplot* consists of points (x_i, y_i) whose coordinates are values of a data object.
- The *sample correlation* between variables x, y is $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \in [-1, 1]$.
- $r_{xy} > 0$ often means that y increases with x.
- $r_{xy} < 0$ often means that y decreases with x.
- $r_{xy} = 0$ means no *linear* relation between x, y.

Regression means "go backward"

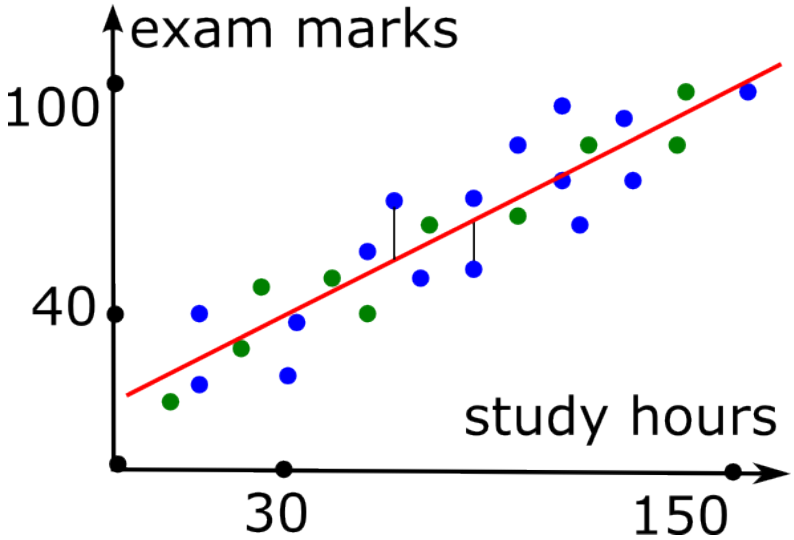
Sir Francis Galton (1822-1911) was the first scientist to apply regression to biological data. Taller-than-average parents tend to have children who also taller-than-average, but not as tall as their parents. It's called 'regression toward the mean'.



If a scatterplot looks linear, we can try to find the best straight line that fits or approximates this scatterplot.

The regression line for a scatterplot

Definition 15.1. For a scatterplot of n data points (x_i, y_i), the **least-squares regression line** has an equation $y = ax + b$ that minimises the sum of squares $f(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$. Here x_i, y_i are given samples, while a, b are unknown coefficients.

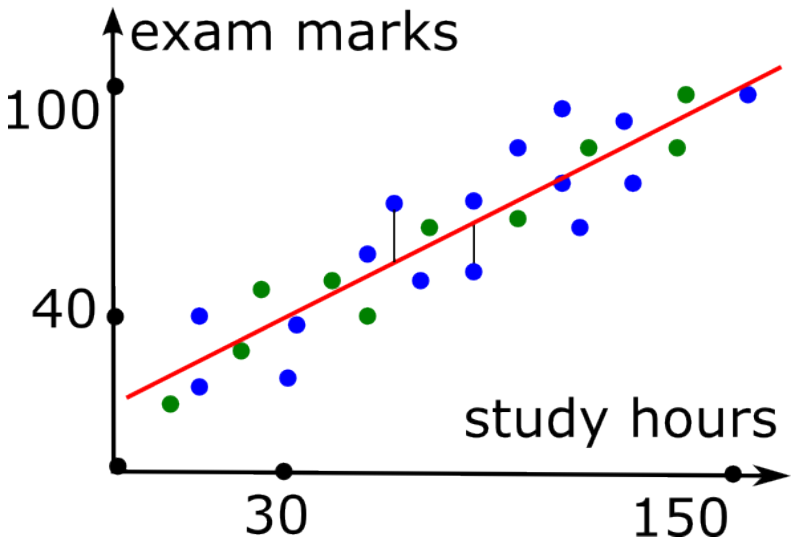


$|ax_i + b - y_i|$ is the **residual**, that is the *vertical* distance from the point (x_i, y_i) to the line $y = ax + b$, not along a perpendicular line.

The regression isn't symmetric

The squared distances are for easier differentiating to find simple formulae for the coefficients a, b.

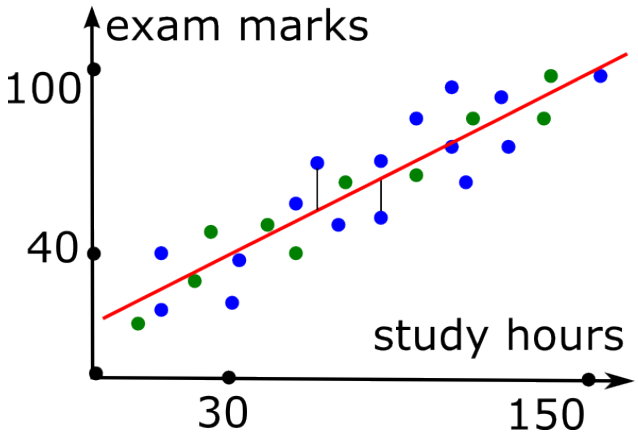
The distances along perpendiculars make sense and will lead to the more complicated PCA later.



If we swap x, y, the line can change, because the sum of squared *horizontal* distances to the line will be minimised.

Simple linear regression

Ordinary least squares (OLS) approximation: For n given data points (x_i, y_i), the **least-squares regression line** has an equation $y = ax + b$ that minimises the *sum of squared residuals* $f(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$ from points to the line.



The case of one independent variable x and one dependent variable y is called a **simple linear regression**.

y variable is called *dependent variable, response, outcome or label* (in Machine Learning). x variable is called *independent variable, explanatory variable, predictor, covariate or feature*.

Formulae for the regression line

Theorem 15.2. For n points $(x_i, y_i) \in \mathbb{R}^2$, the simple linear regression has the equation $y = ax + b$ with

$$a = r_{xy}\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \text{ and } b = \bar{y} - a\bar{x}.$$

Here $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y}$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},\ s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Visual explanation (6 minutes).

Another formula for the regression

Claim 15.3. The 1st coefficient is $a = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$, where each bar means the average over n samples.

Expand:

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})}{\sum_{i=1}^n (x_i^2 - 2\bar{x} x_i + \bar{x}^2)} =$$

$$\frac{\sum_{i=1}^n (x_i y_i) - \bar{x} \sum_{i=1}^n y_i - \bar{y} \sum_{i=1}^n x_i + n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2} =$$

$$\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}.$$

Example regression line

Problem 15.4.

x	1	2	3	4	5
y	1	3	5	7	9

Equation?

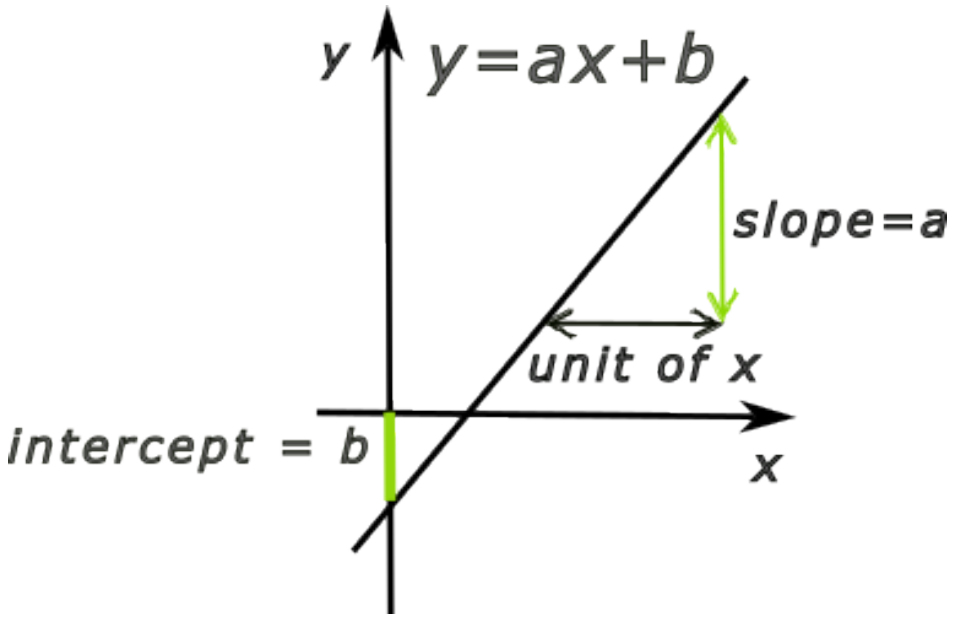
Solution 15.4. Sample means: $\bar{x} = 3$, $\bar{y} = 5$.

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{(-4)(-2)+(-2)(-1)+2\cdot 1+4\cdot 2}{(-2)^2+(-1)^2+1^2+2^2} = \frac{20}{10} = 2, \text{ then}$$

$$b = \bar{y} - a\bar{x} = 5 - 2 \cdot 3 = -1$$

The regression line $y = 2x - 1$ accidentally passes through all (x_i, y_i), so $f = \sum_{i=1}^{n}(ax_i + b - y_i)^2 = 0$.

Meaning of the line



Regression allows predict y from known values of x.

Linear regression for two points

Problem 15.5. Write down the regression line for the scatterplot of two points (x_1, y_1), (x_2, y_2).

Solution 15.5. The required line passes through both points, i.e. the sum of squared distances is 0.

The line is $y - y_1 = \dfrac{y_2 - y_1}{x_2 - x_1}(x - x_1)$ if $x_1 \neq x_2$.

There was no need to write the formulae for a, b.
Understanding is always better than memorising.

Real Life applications

Linear approximation was the first appriach to be used in ML algorithms (generalised as SVM).

Lines can be deceptive and can mask trends in data, or create such trends:

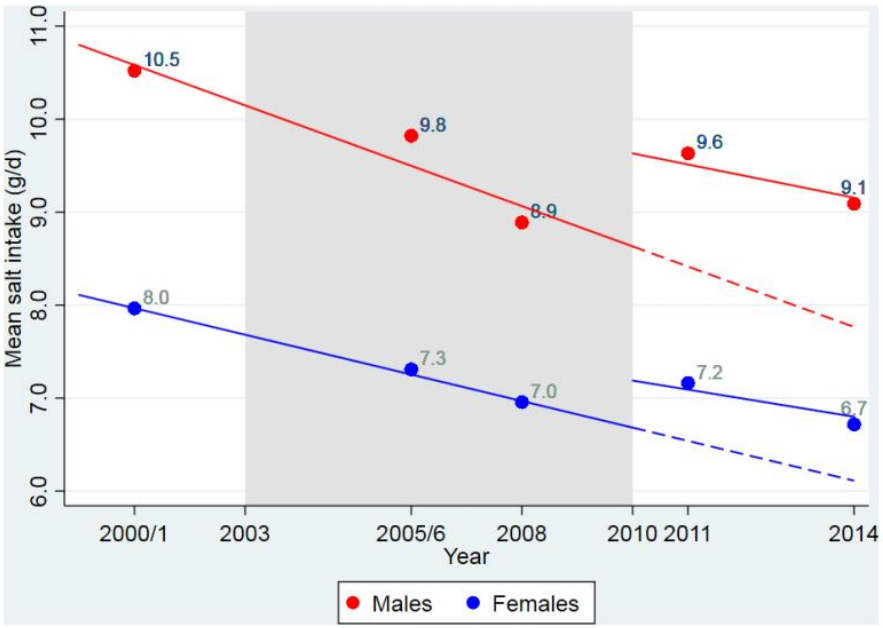


Figure 1 Pre- and post-Responsibility Deal trends of salt intake in England 2000/01 to 2014.

Laverty AA, etc. (2019) Quantifying the impact of the Public Health Responsibility Deal on salt intake, cardiovascular disease and gastric cancer burdens: Interrupted time series and microsimulation study.

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Example 15.6. These 4 well-known scatterplots have the same $\bar{x} = 9$, $\bar{y} = 7.50$

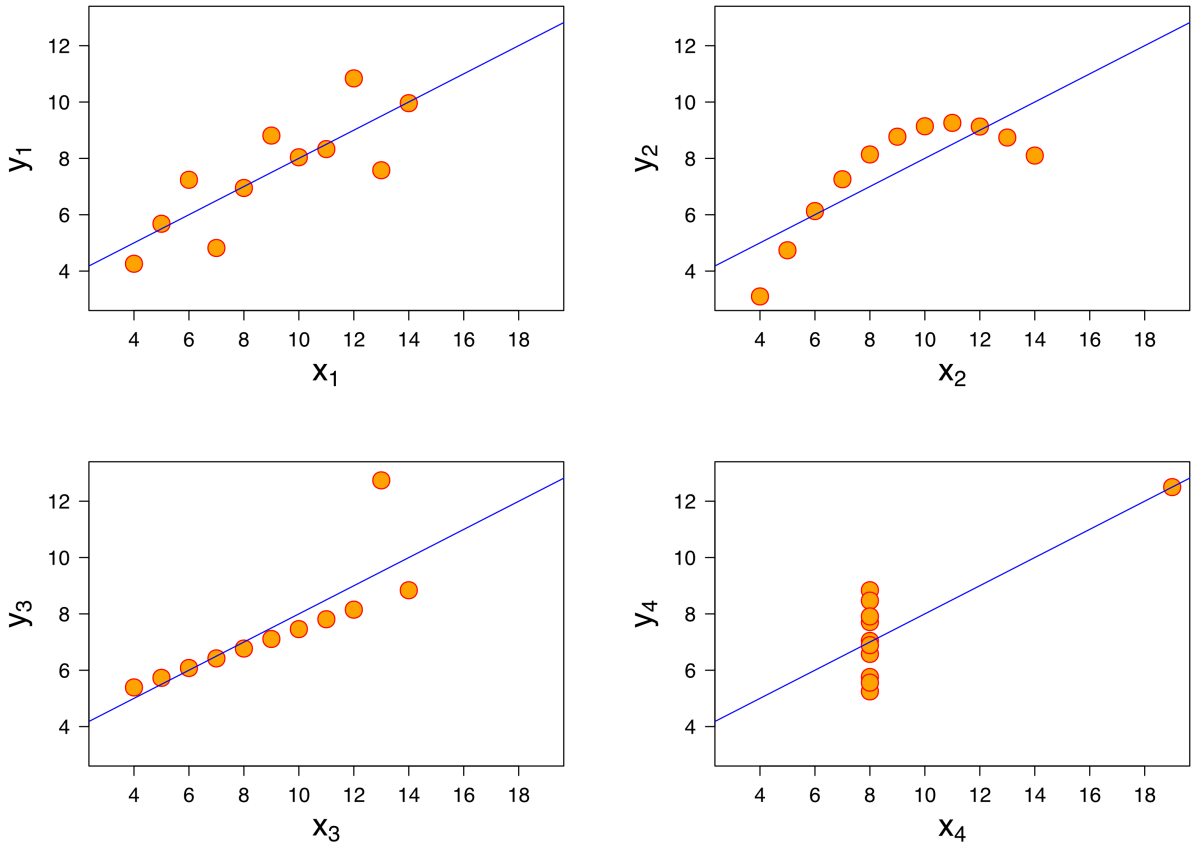
$s_x^2 = 11$, $s_y^2 = 4.12$

$r_{xy} = 0.816$, the same linear regression line

$y = 0.5x + 3$.

Are the samples similar?

Importance of visualizaition

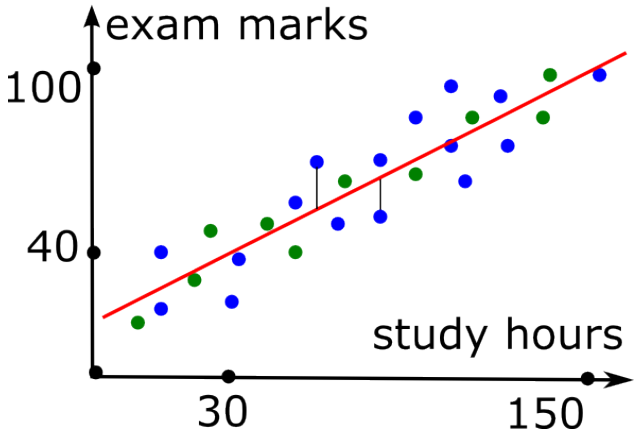


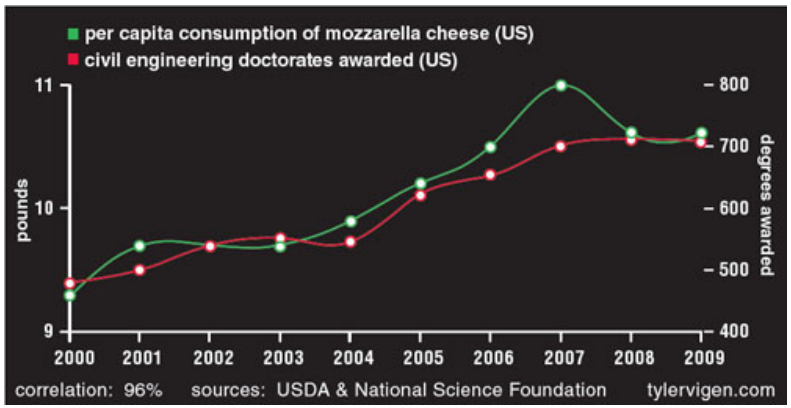
By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

Not everything is linear: Hans Rosling's 200 Countries, 200 Years, 4 Minutes

Limitations of linear regression

Predictions make sense only near given points: 0 hours don't guarantee mark 20.





A correlation is misleading if there is another variable affecting both variables.

The regression line is highly **sensitive to outliers**.

Deviations from the linear approximation should be independent and uniformly distributed (**homoscedasticity**).

Validation tests for a regression

To decide if the regression makes sense for our data:

- *fit* coefficient $r_{xy}^2 \in [0, 1]$: the larger the better;
- graphical analysis of *residuals*: should be randomly distributed, without outliers, Cook's distance is often used;
- *cross-validation*: randomly split the data into 5 groups, use 4 groups to find a regression line, use the remaining 20% group to test, compute the average error (equal to the squared vertical distance), repeat 4 times by swapping groups.

Time to revise and ask questions

- The *least-squares regression line* minimises the sum of squared vertical distances from points.
- The regression line $y = ax + b$ has $b = \bar{y} - a\bar{x}$, $a = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}, \bar{y}$ are the sample means.
- A regression line $y = ax + b$ may not be symmetric with respect to x, y, i.e. swapping x, y may give another regression $x = cy + d$.

Problem 15.7.

x	1	2	3	4	5
y	3	1	2	1	3

Find the regression line and compare to the regression line $x = cy + d$.

A regression may not be symmetric

Solution 15.7. $\bar{x} = 3$, $\bar{y} = 2$, $a = \dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2} =$

$$\frac{(-2)\cdot 1+(-1)\cdot(-1)+(3-3)(2-2)+(4-3)(1-2)+(5-3)(3-2)}{(-2)^2+(-1)^2+1^2+2^2} = 0$$

$b = \bar{y} - a\bar{x} = 2 - 0 \cdot 3 = 2$, the regression is $y = 2$.

Find the regression for the swapped points (y_i, x_i)

y	3	1	2	1	3
x	1	2	3	4	5

Then $a = \dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2} = 0$,

$b = \bar{x} - a\bar{y} = 3 - 0 \cdot 2 = 3$, the regression is $x = 3$.