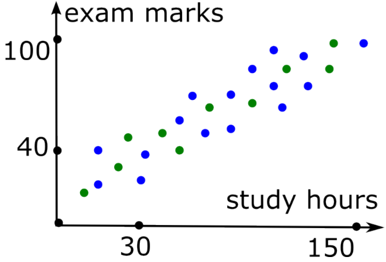
COMP229: Introduction to Data Science
Lecture 15: a simple linear regression

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

Regression means "go backward"

Sir Francis Galton (1822-1911) was the first scientist to apply regression to biological data.

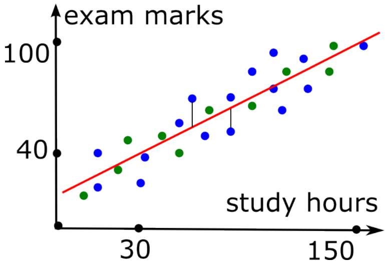
He noticed that taller-than-average parents tend to have children who also taller-than-average, but not as tall as their parents. He called this observation "regression toward the mean".



If a scatterplot looks linear, we can try to find the best line that fits (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 values, while a, b are unknown coefficients.



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

Formulae for the regression line

Vertical distances help to get an easy solution.

Claim 15.2. For n points (x_i, y_i) the regression line is $y = ax + b$ with $a = r_{xy}\dfrac{s_y}{s_x}$ and $b = \bar{y} - a\bar{x}$.

Here $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ are sample means.

$s_x = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $s_y = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ are sample standard deviations, and the sample correlation is $r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$.

Example regression line

Find the regression line for

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

Sample means: $\bar{x} = 3$, $\bar{y} = 5$. Sample standard deviations: $s_x = \sqrt{\frac{1}{4}(2 \cdot 2^2 + 2 \cdot 1^2)} = \sqrt{2.5}$, $s_y = \sqrt{\frac{1}{4}(2 \cdot 4^2 + 2 \cdot 2^2)} = \sqrt{10}$. The correlation is $r_{xy} = \frac{(-2)(-4) + (-1)(-2) + 1 \cdot 2 + 2 \cdot 4}{4 s_x s_y} = \frac{20}{4\sqrt{25}} = 1$ (no calculator needed). Use Claim 15.2.

Error of the regression line

By Claim 15.2 the regression line $y = ax + b$ has the gradient (slope) $a = r_{xy}\frac{s_y}{s_x} = \frac{\sqrt{10}}{\sqrt{2.5}} = \sqrt{4} = 2$ and $b = \bar{y} - a\bar{x} = 5 - 2 \cdot 3 = -1$.

The resulting regression line $y = 2x - 1$ accidentally passes through all the given points (x_i, y_i). Hence the minimum of the error function $f(a, b) = \sum_{i=1}^{n}(ax_i + b - y_i)^2$ is 0 (all terms vanish).

A proof of the formula for b

Proof of Claim 15.2. We find a, b that minimise the quadratic function $f(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2 \geq 0$.

At any local minimum the derivatives with respect to the variables a, b must be zero: $\frac{\partial f}{\partial a} = 0 = \frac{\partial f}{\partial b}$.

With respect to b, the derivative of $(ax_i + b - y_i)^2$ is $2(ax_i + b - y_i)$. Take the sum over $i = 1, \ldots, n$.

$$0 = \frac{\partial f}{\partial b} = 2\sum_{i=1}^{n}(ax_i + b - y_i),\ a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i.$$

After dividing by n, we get $a\bar{x} + b = \bar{y}$, $b = \bar{y} - a\bar{x}$.

Shift the plane: $(\bar{x}, \bar{y}) \mapsto (0, 0)$

$b = \bar{y} - a\bar{x}$ means that $(\bar{x}, \bar{y})$ is in $y = ax + b$.

If we shift the plane by moving $(\bar{x}, \bar{y})$ to $(0, 0)$, the line $y = ax + (\bar{y} - a\bar{x})$ becomes $y = ax$, so we may assume that $\bar{x} = 0 = \bar{y}$ and $b = 0$.

Since $r_{xy} = \dfrac{\sum_{i=1}^{n} x_i y_i}{(n-1)s_x s_y}$ and $(n-1)s_x^2 = \sum_{i=1}^{n} x_i^2$,

the formula $a = r_{xy}\dfrac{s_y}{s_x}$ becomes $a = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$.

Notice that $y = ax + b$ isn't symmetric with respect to x, y, i.e. swapping x, y will give another regression line $x = cy + d$.

A proof of the formula for a

It remains to use $\dfrac{\partial f}{\partial a} = 0$ for $f(a,0) = \sum_{i=1}^{n}(ax_i - y_i)^2$.
With respect to a, the derivative of $(ax_i - y_i)^2$ is $2(ax_i - y_i)x_i$. The extra factor x_i is due to the chain rule (as the derivative of the internal function). Take the sum over $i = 1, \ldots, n$.

$$0 = \frac{\partial f}{\partial a} = 2\sum_{i=1}^{n}(ax_i - y_i)x_i, \quad a\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Finally, we get $a = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ as required. $\square$

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2–3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Find the regression line for

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

Another regression line

The sample means are $\bar{x} = 3$ and $\bar{y} = 2$.

The sample standard deviations are

$$s_x = \sqrt{\frac{2 \cdot 2^2 + 2 \cdot 1^2}{4}} = \sqrt{2.5},\ s_y = \sqrt{\frac{4 \cdot 1^2}{4}} = 1.$$

The correlation is $r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} =$

$$\frac{(-2) \cdot 1 + (-1)^2 + 1 \cdot (-1) + 2 \cdot 1}{4\sqrt{2.5}} = 0.$$

The regression line $y = ax + b$ has the coefficients $a = r_{xy}\dfrac{s_x}{s_y}$ and $b = \bar{y} - a\bar{x}$.

Answer to the quiz and summary

Answer. The regression line for the x, y data

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

has the equation $y(x) = 2$.

- The *least-squares regression line* minimises the sum of squared vertical distances.
- The regression line $y = ax + b$ has the coefficients $a = r_{xy}\dfrac{s_x}{s_y}$ and $b = \bar{y} - a\bar{x}$ and passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}, \bar{y}$ are sample means; s_x, s_y are the sample standard deviations; r_{xy} is the correlation.