Lessons in linear regression

How do you teach linear regression to pupils aged 12–14? Direct proportionality is the key, advises **Miguel de Carvalho**. That and some sporting superstars…

> Part of my joy in learning is that it puts me in a position to teach.
>
> Seneca, *Letters from a Stoic*, Letter VI

Regression is a mainstay of statistics. Developed by Sir Francis Galton more than 130 years ago, it has since then been widely applied in a variety of sciences. Numerous regression models are being fitted as you read this sentence. But how would you introduce linear regression to pupils who are barely familiar with concepts such as the Cartesian coordinate system? In other words, is there a simple way to teach linear regression without relying on graphical representations?

I faced this question for the first time when I gave a masterclass to S2 pupils (second year of secondary school in Scotland, ages 12–14) with the goal of broadening their areas of mathematical knowledge. The trick was to resort to direct proportionality – a notion S2 pupils are familiar with – since regression through the origin can be simply regarded as a random version of it. Here's how I did it.

Ronaldo versus Messi

To be able to introduce linear regression – which was my main goal – I first had to cover the basics on means. I used data from footballing legends Ronaldo and Messi to introduce the notion of average. I recalled that the sample mean is defined as $\bar{x}_R = (x_1 + \ldots + x_n)/n$, where x_i is the number of goals scored by Ronaldo in his ith match, and with n denoting the total number of matches played. According to Wikipedia on 18 August 2023 (tinyurl.com/4skaz98y), Ronaldo scored a total of 721 goals for his club, while Messi scored 724 goals (tinyurl.com/3vvpjxn6). If we took only the number of goals to compare the performance of both players, then Messi would be a clear winner. Yet the same source also claims that Ronaldo has played 976 matches, while Messi only played 813, and thus a fairer comparison of performance would have to take this information on board: it follows that $\bar{x}_R = 0.739$, and $\bar{x}_M = 0.813$, and thus Messi scored more goals per match than Ronaldo. I gave pupils a few more comments about means, but as the focus of the talk was on regression, kept them short.

I used data from footballing legends Ronaldo and Messi to introduce the notion of average

Miguel de Carvalho is reader in statistics at the University of Edinburgh and holds a chair of statistics at University of Aveiro.

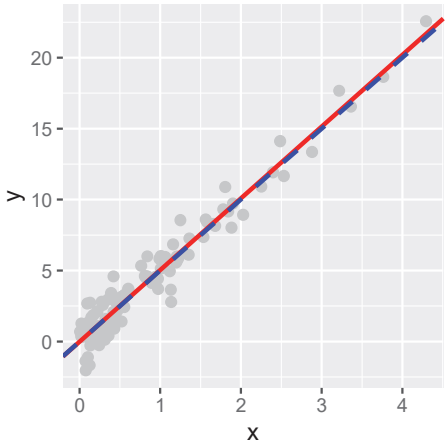

FIGURE 1: Scatterplot of simulated data from regression through the origin model with $\beta = 5$. The solid and dashed lines represent estimated ($\hat{y} = \hat{\beta}x$) and true ($y = 5x$) regression line, respectively.

Proportionality and regression through the origin

To introduce regression, I again resorted to data from the same source on Ronaldo and Messi; below, x is weight (in kilograms) and y is height (in centimetres). According to Wikipedia, for Ronaldo we have $(x_R, y_R) = (84, 187)$ and for Messi $(x_M, y_M) = (72, 170)$.

Before I was ready to elaborate on the statistical link between weight and height, I had to recall main concepts on proportionality. Particularly, I reminded pupils that two variables x and y are *directly proportional* if $y = \beta x$, where β is the *constant of proportionality*. Students are familiar with the fact that the perimeter of a circle (P) is proportional to the radius (r), with the constant of proportionality being 2π; indeed, $P = 2\pi r$. I spent some time reviewing these notions, and then challenged pupils with the following question:

> In the same way that the perimeter of the circle is proportional to the radius, can we claim that people's weight (x) tends to be "proportional" to their height (y)?

In other words:

> Does the relation $y = \beta x$ hold between the variables weight (x) and height (y), for a fixed β, for all subjects?

The Ronaldo versus Messi example can be used to underscore that weight (x) and height (y) are not proportional. Indeed, it follows that $y_R = \beta_R x_R$, $y_M = \beta_M x_M$, with $\beta_R \neq \beta_M$ ($\beta_R = 2.23$ and $\beta_M = 2.36$). These considerations indicate that height (in centimetres) could be about 2 times weight (in kilograms) – at least for Ronaldo and Messi. Motivated by this apparent *quasi-proportionality*, I introduced the idea of regression through the origin, and informally stated it as y being approximately the same as βx, that is, $y \approx \beta x$.

Our approach, which does not rely on visuals, may be beneficial for introducing linear regression to visually impaired students

I explained to pupils that regression models are a ubiquitous tool in modern statistics, and that they are widely used for predicting the value of a variable y as well as to understand how variables x and y relate statistically.

We then moved beyond Ronaldo and Messi and considered n players, so to consider the problem of learning about β from data. To make the set-up manageable, I focused on the case where x is positive. By noting that $y_1 \approx \beta x_1, \ldots, y_n \approx \beta x_n$, it follows that $y_1 + \ldots + y_n \approx \beta(x_1 + \ldots + x_n)$, that is, $\bar{y} = \beta \bar{x}$. This suggests the estimator:

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}}$$

(Clearly this naïve estimator is not as sound as the ordinary least squares (OLS) estimator; in a fixed design setting it is easy to show that it is unbiased, but it has larger variance than OLS. Yet it is straightforward to introduce to pupils!) Figure 1 depicts a cloud of points fitted with this estimator.

Take-home message and final comments

Pupils are familiar with the notion of proportionality; so it's convenient to take advantage of this while introducing ideas such as linear regression. Regression through the origin can be understood as a random version of the idea of direct proportionality, with the constant of proportionality acting as the slope parameter.

In statistical practice, the question of when it is more appropriate to focus on regression through the origin has been widely debated. With respect to this, George Casella once claimed that "The problem of deciding whether an intercept model or a no-intercept model is more appropriate for a given data is a problem with no simple solution".[1] Guidance on situations where a no-intercept regression is appropriate can be found in a 2003 paper by Eisenhauer.[2] And it is worth noting that the quantity $\bar{y}/\bar{x}$ appears in other contexts, being widely employed in sample surveys where it is known as the ratio of sample averages.[3]

As a by-product, our direct proportionality approach, which does not rely on visuals, may be beneficial for introducing linear regression to visually impaired students. See Stone *et al.*'s 2019 paper[4] on the challenges and approaches related to teaching statistics to visually impaired students.

Some final remarks on variations and suggestions for follow-up lectures are in order. First, for students familiar with visualisations it would be clearly beneficial to introduce scatterplots and residual plots as early as possible. Second, simple linear regression goes hand in hand with correlation and some of the ideas and principles in Zou *et al.*'s article on the subject[5] would be helpful for designing an accessible lecture on correlation and goodness of fit.

Finally, it is well known that many modern tools, such as neural networks and deep learning, can be understood as extensions of regression models.[6] Therefore, a follow-up lecture on artificial intelligence could build on the foundational knowledge established in this masterclass. Where Ronaldo and Messi might fit into that is up to you. ■

References

1. Casella, G. (1983) Leverage and regression through the origin. *The American Statistician*, **37**(2), 147–152.

2. Eisenhauer, J. G. (2003) Regression through the origin. *Teaching Statistics*, **25**(3), 76–80.

3. Barnett, V. (2002) *Sample survey principles and methods* (3rd edn). Chichester: Wiley.

4. Stone, B. W., Kay, D. and Reynolds, A. (2019) Teaching visually impaired college students in introductory statistics. *Journal of Statistics Education*, **27**(3), 225–237.

5. Zou, K. H., Tuncali, K. and Silverman, S. G. (2003) Correlation and simple linear regression. *Radiology* **227**(3), 617–622.

6. Efron, B. and Hastie, T. (2016) *Computer Age Statistical Inference*. New York: Cambridge University Press.