

Class #7 & #8 - Wed 18 Sept & Mon 23 Sept
Paired Data: Linear Regression

Readings: Ross, Sections 12.1-12.3; see also <https://www.khanacademy.org/math/probability/regression>

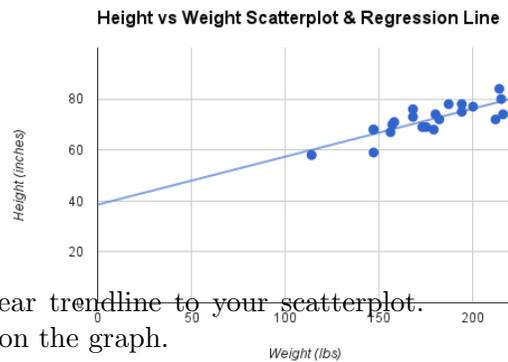
Intro: Last time we discussed two ways of looking at the relationship between two variables given a paired data set $\{(x_i, y_i) : 1 \leq i \leq n\}$:

- a scatterplot to visualize the data set
- the correlation coefficient r which quantifies to what extent the variables are positively or negatively correlated:

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

For example, we generated the scatterplot for the height-weight data from the last handout, which seemed to indicate a strong positive correlation between height and weight. We also calculated the correlation coefficient to be $r \approx 0.84$, which verifies that there is a strong positive correlation in the data.

We can also manually add a line which seems to fit the data—but we want a systematic way of doing so. In fact, the spreadsheet has a built-in option to add a “linear trendline” to any scatterplot.



Spreadsheet Exercise: Find the option to add a linear trendline to your scatterplot. There is also an option to show the equation of the line on the graph.

How is this “trendline” defined? This process is called *linear regression*. We “regress” the y -data (usually called the dependent or response variable) against the x -data (the independent or input variable) to find a linear relationship for y in terms of x , or what is called a linear model for y in terms of x :

$$y = \alpha + \beta x$$

This is the slope-intercept equation of a line, where α is the y -intercept of the line and β is the slope; α and β are called the linear regression parameters.

Now the question is: how do we find α and β , i.e., what are the best choices of y -intercept and slope so that line that best fits the data?

We will choose α and β so that the sum of squares of the “errors” (or “residuals”) is minimized, where the error ϵ_i is the difference between the data point $\{x_i, y_i\}$ and the point on the line. Hence linear regression is also called “linear least squares” or “ordinary least squares” (OLS).

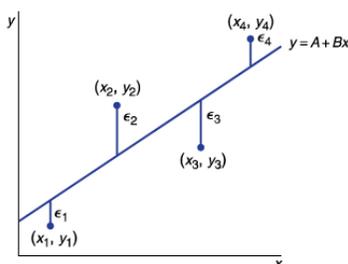


FIGURE 12.2
The errors.

“Least-squares” linear regression: The error for an observed data point (x_i, y_i) with a linear model $y = \alpha + \beta x$ is

$$\epsilon_i = y_i - (\alpha + \beta x_i)$$

The least-squares linear regression model is given by choosing α and β such that the sum of squared errors (often abbreviated “SSE”)

$$\text{SSE} = \sum_{i=1}^n \epsilon_i^2$$

is as small as possible. It can be shown (using calculus) that such α and β are given by:

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

Spreadsheet exercise: Calculate the linear regression parameters for the height-weight data from the previous handout by implementing the formulas above. First calculate β . Note that the numerator in the formula for β ,

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

is the same as the numerator in the correlation coefficient r , so you should have already calculated that in your spreadsheet. Once you’ve calculated β , it is straightforward to calculate α .

Spreadsheet functions =slope, =intercept & =linest: Check your calculations above by directly calculating the linear regression parameters using the built-in functions

`=slope(data_y, data_x) & =intercept(data_y, data_x)`

which output the two linear regression parameters. Alternatively, you can use the “linear estimator” function which will output both regression parameters:

`=linest(data_y, data_x)`

(Like `=frequency`, `=linest` is an “array function” in Excel, and so you will need to first select the output cells and hit “Control+Shift+Enter” to enter the formula. Another way to do linear regression in Excel is to use “Tools → Data Analysis → Regression”.)