Data Science Career Track

The Art of Statistics, Chapter 5: Modeling relationships using regression

Take-Away Notes

This chapter introduces us to **statistical models**, and in particular, regression models. Models are maps - simplifications of the territory - and not the territory itself. The British statistician George Box wrote that 'all models are wrong, but some are useful.' Statistical models assume that:
*Observation = deterministic model (**the signal**) + residual error (**the noise**)*

That is, what we see is the sum of a mathematical idealization, and some random contribution that can't yet be explained.

Regression models predict a quantitative variable (known as the **dependent** or **response** variable, typically on the vertical *y*-axis) with a set of **explanatory** variables (typically on the *x*-axis). The gradient of the line is known as the **regression coefficient**. If there is more than one variable in the set of explanatory variables, we're doing **multiple linear regression**; we often do this when we want to adjust for the potential impact of other variables.

A linear regression model typically draws the **least-squares** fitted line over the plot of the explanatory variables against the response variable; that is, a line of 'best fit' that minimizes the error between each observed data-point and the line.

The **regression coefficient** differs from the **Pearson correlation coefficient,** which runs between -1 and 1, and expresses how close to a straight line the data-points fall. The regression

coefficient, the Pearson correlation coefficient, and the standard deviations of the variables are related systematically in the mathematical equation for the least-squares fitted line. If the standard deviations of the independent and dependent variables are equal, then the gradient is just the Pearson correlation coefficient.

The meaning of the gradient (or regression coefficient) depends on what we assume to be the relationship (i.e., correlational or causal) between the variables in question.

Regression to the mean occurs when an extreme observation is succeeded by a less extreme one, through the process of natural variation. The phenomenon occurs because part of the cause of the initial extreme is chance, and this is unlikely to repeat to the same degree. We need to be careful about attributing causal efficacy to interventions that had no effect on an outcome that was just brought about by regression to the mean.