Statistically Speaking: *Linear Regression*

DAVID L. TABB, PH.D.

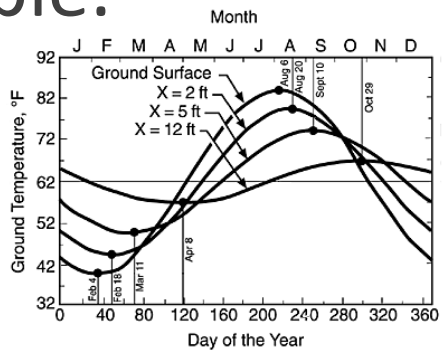
SEPTEMBER 14, 2017

Overview

- Independent and dependent variables
- History: astronomers, biologists, statisticians
- Slopes, intercepts, and residuals
- R code for inferring linear models

Farewell, commutative property!

- Correlation assumes that the two metrics have a symmetric relationship.
- Linear models assume that one variable (typically *y*) depends at least in part on the value of another (typically *x*).
- We might say that the dependent variable *is a function of* the independent variable.



http://www.builditsolar.com/Projects/Cooling/EarthTemperatures.htm

https://www.sciencebuddies.org/science-fair-projects/science-fair/variables

Legendre in his own time

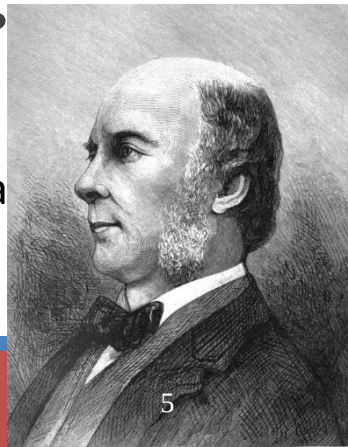
- Born in 1752, Adrien-Marie Legendre served as a professor in the École Militaire at Paris, France. His studies in the paths of projectiles led him to study the heavens.
- His "method of least squares" seeks the linear relationship that minimizes the squared error between modeled and observed points.

Wikimedia Commons

http://www.stat.ucla.edu/history/legendre.pdf

Galton and "regression to the mean"

- Born in 1822, Galton published on an astonishing array of topics. He coined the term "eugenics." He inferred from population studies that offspring traits reflect those of their parents most and more distant ancestors to a lesser degree.
- *"It appeared from these experiments that the offspring did not tend to resemble their parents in size, but always to be more mediocre than they – to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were small."*



Wikimedia Commons

https://select-statistics.co.uk/blog/regression-to-the-mean-as-relevant-today-as-it-was-in-the-1900s/

UNIVERSITEIT STELLENBOSCH UNIVERSITY

George Udny Yule on interpreting correlation

- "when an association leads to the inference of a *direct causal relation* where none exists it is *misleading*. The inferred causal relation is an *illusion*. It is a *fallacy* to interpret an association as if it were necessarily due to such a relation."
- Yule argued that mathematics employed in eugenics literature occasionally slanted the answers in favor of eugenic theories.
- Yule brought linear regression to time series.

Yule nicknamed free-spirited researchers the "loafers of the world."

Terence C. Mills. *Statistical Biography of George Udny Yule*…

J. Aldrich. *Statistical Science* (1995) 10: 364-376.

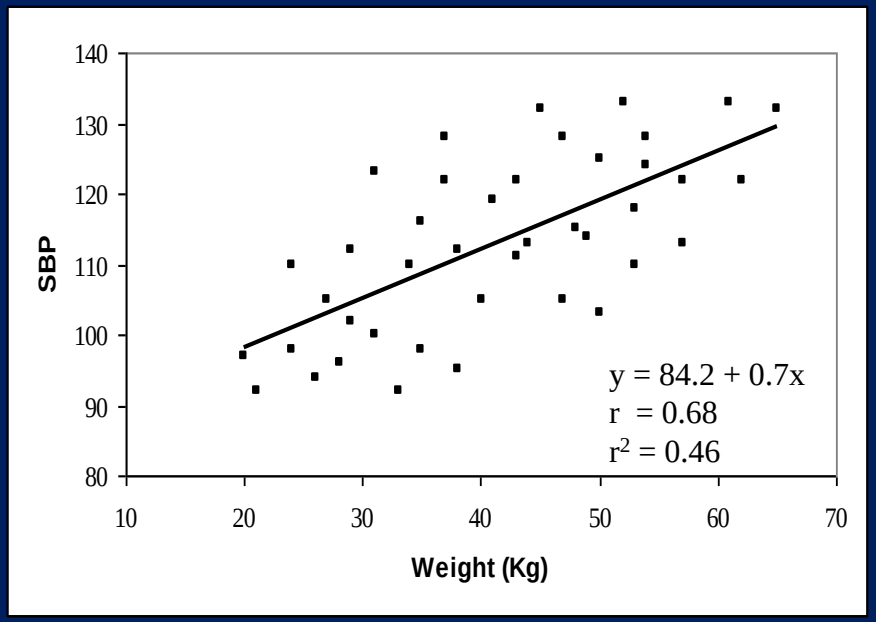
Linear models

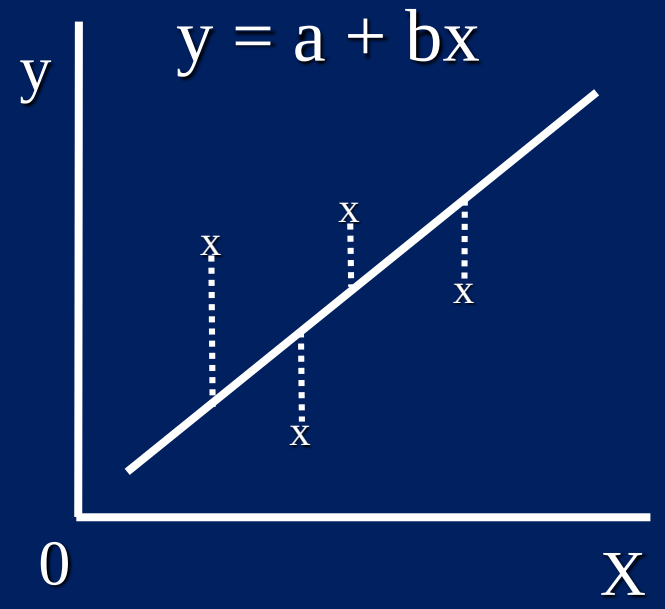
- What is our expected value of the dependent variable (for example, grade) given the values of several variables that bear on it? (e.g. time studying, IQ, hours required for part-time job)
- A variable may be more weighted because it is important or because it has a smaller range.
- Each independent variable is present at only the first power; this is no quadratic!

The simple case

- From geometry: $y = mx + b$
- Statistical form: $y_i = \alpha + \beta x_i + \varepsilon_i$
- y_i: the ith value from the vector of y values.
- α: the intercept (crossing point on the y-axis)
- β: the slope or coefficient (also known as the steepness of line or “weight” of factor)
- ε_i: the error seen versus the model for item i (also called a “residual”).

Calculation of Regression and Correlation Coefficients





Having inferred a linear model, we see that data points are some distance from the line that we can plot through them. We call the vertical distance between the line and each data point a "residual."

Figures from Tonya Esterhuizen

The solution

- Pick the α and β that minimize the sum of squared error ($\sum \varepsilon_i^2$)
- Remember definition of variance:

$$s^2 = \frac{\sum (x_i - \bar{X})^2}{n - 1}$$

```
x <- runif(100)
y <- 6 + 4*x + rnorm(100)
#Set the true alpha to 6
#We set the true beta to 4
#We model Gaussian error
model <- lm(y~x)

#Example runs:
alpha 5.797 beta 4.236
alpha 6.218 beta 3.658
alpha 6.103 beta 3.874
alpha 5.754 beta 4.408
```

Takeaways

- Linear models infer the linear relationship between a dependent variable and the independent variables that may influence it.
- Minimizing the sum of squared error is a standard approach to find alpha and beta values (or intercept and slope values).