Class 3: One Health, BioStatistics II: Statistical Regression Modeling

1 Regression Modeling Intro

1.1 Objectives & Agenda

1.2 Regression Models

1.2.1 Overview

- Workhorse of data science.
- Interpretable model fits, contrasting with ML algorithms.
- Simplicity, parsimony and intrepretability.
- First tool of choice for any practical problem.

Regression models are the workhorse of data science. They are the most well described, practical and theoretically understood models in statistics. A data scientist well versed in regression models will be able to solve an incredible array of problems.

Perhaps the key insight for regression models is that they produce highly interpretable model fits. This is unlike machine learning algorithms, which often sacrifice interpretability for improved prediction performance or automation. These are, of course, valuable attributes in their own rights. However, the benefit of simplicity, parsimony and intrepretability offered by regression models (and their close generalizations) should make them a first tool of choice for any practical problem.

1.2.2 Historical

- The earliest form of regression was the method of least squares, published by Legendre in 1805, and by Gauss in 1809.
  - Problem of determining, from astronomical observations, the orbits of bodies about the Sun.

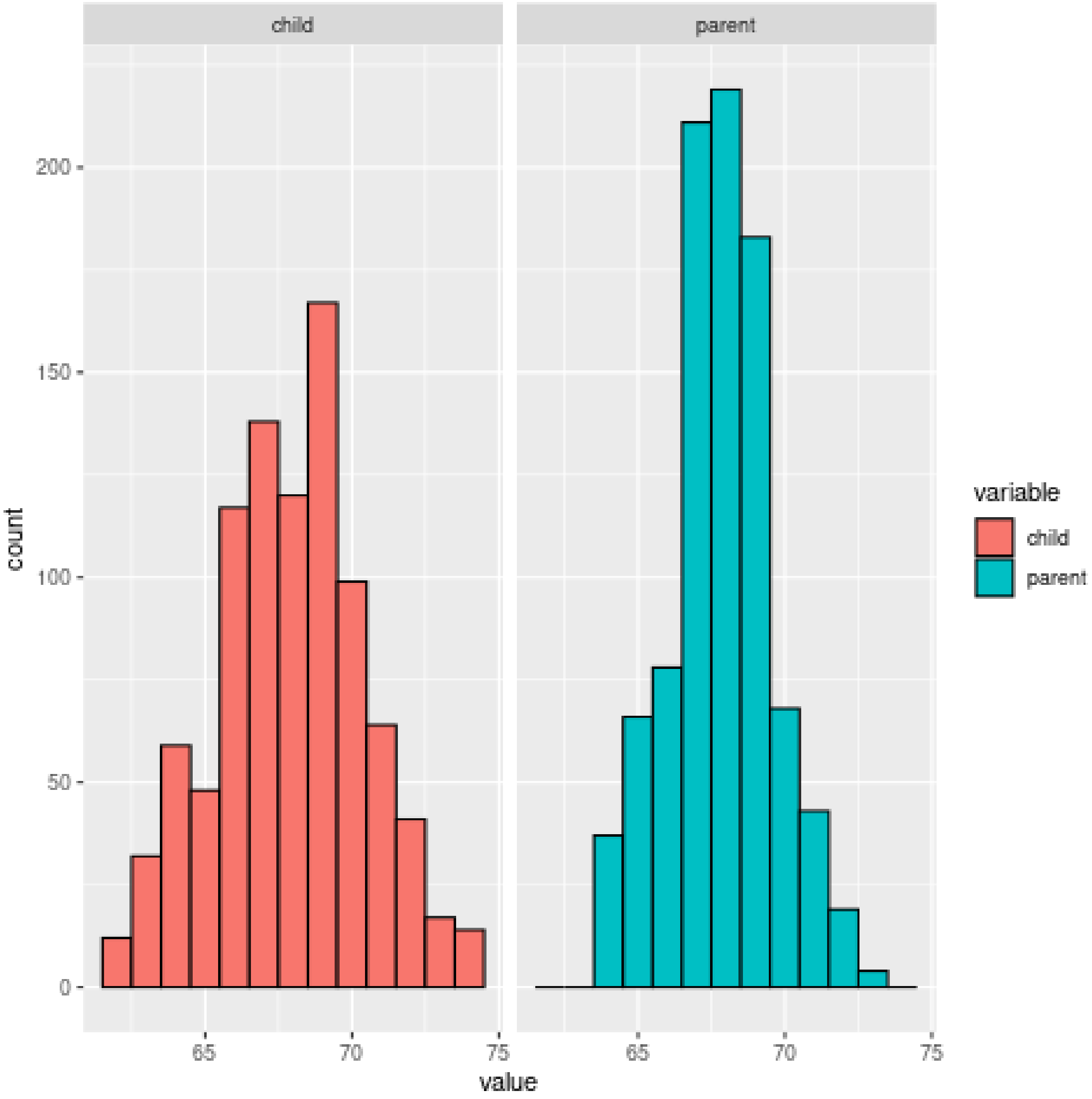
1.2.3 Regression to the Mean

- Dicovered by Francis Galton (1886)
  - "Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.(Galton, 1888, p 135)

1.2.4 Statistical Effect

What is that?.

- In statistics, regression toward the mean (or regression to the mean) arises if a sample point of a random variable is extreme (nearly an outlier), a future point will be closer to the mean or average on further measurements.



1.2.5 For a pair of random variables X, Y

- The largest the first one, by chance; the probability of smaller for the second is high.
- $P(Y > x|X = x)$ get bigger as x heads to very small values.

1.2.6 Conversely

- $P(Y < x|X = x)$ get bigger as x heads to very large values

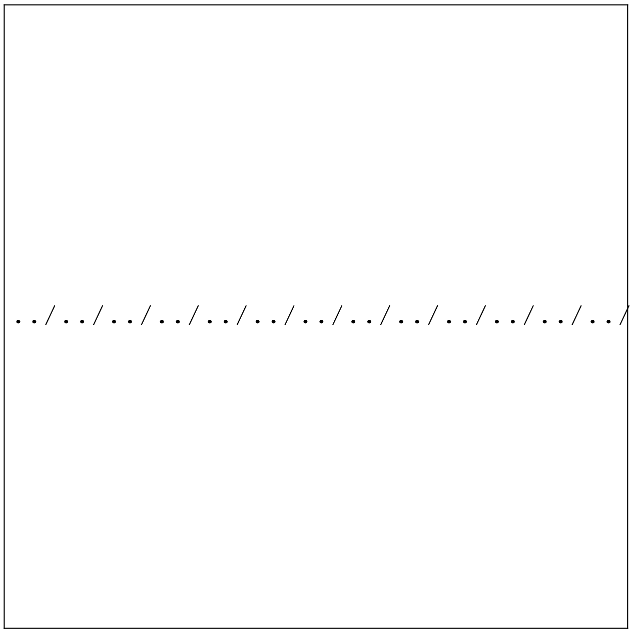

../../../../../../../../../../../../../org/Projects/Lecturing/Coursepac

Figure 1: Linear Regression

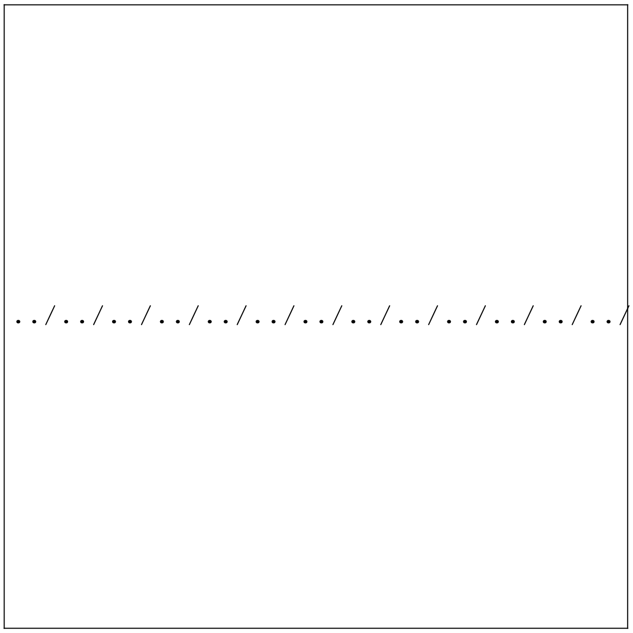

../../../../../../../../../../../../../org/Projects/Lecturing/Coursepac

Figure 2: Linear Regression

`../../../../../../../../../../../../../../org/Projects/Lecturing/Coursepac`

Figure 3: Linear Regression

2 Linear Regression Modeling

2.1 Objectives & Agenda