STAT 425

Introduction

Spring 2020

Where to start?

Statistical Analysis:

Problem$\rightarrow$ Collection of Data $\rightarrow$ Data Analysis $\rightarrow$ Conclusions

Two important steps:

- Problem formulation
- Data collection

Problem formulation:

- Understand the physical background
- Understand the objective
- Learn what the *client* wants
- Set the problem in statistical terms

How the data were collected:

- Observational vs. Experimental. Convenience sampling vs. design sampling survey.
- Is there a missing response?
- Are there missing values?

Initial Data Analysis

- Summary Statistics: This is a very important step!!!
- Single variables: Boxplots, histograms, density plots, etc.
- Bi-variate and multivariate: scatter plots, interactive graphics, etc.
- Look for outliers, typing errors, skewed distributions (are the prior distributions as expected?)

Example

School expenditure and test scores from USA in 1994-95
Data Description:

The **sat data frame** has 50 rows and 7 columns. Data were collected to study the relationship between expenditures on public education and test results.

This data frame contains the following columns:

- **expend:** Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)
- **ratio**: Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
- **salary**: Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)
- **takers:** Percentage of all eligible students taking the SAT, 1994-95
- **verbal**: Average verbal SAT score, 1994-95
- **math**: Average math SAT score, 1994-95
- total: Average total score on the SAT, 1994-95

Source:

"Getting What You Pay For: The Debate Over Equity in Public School Expenditures" D. Guber, Journal of Statistics Education, 1999

Linear modeling

It is used for explaining or modeling the relationship between variable Y and one or more variables: $X_1, X_2, \ldots, X_p$.

Y: dependent variable, response, outcome, output variables
$X_1, X_2, \ldots, X_p$: independent, predictor, input, explanatory

variables.

WARNING: Avoid using the terms *independent* and *dependent* variables for variables X and Y since these terms are used in another broader context. Another term used is Regression

Analysis.

- Response variable Y must be continuous
- Explanatory variables $X_1, X_2, \ldots, X_p$ can be continuous, discrete or categorical.

Model types

- Simple Regression: $p = 1$
- Multiple Regression: $p > 1$
- Multivariate multiple regression: More than one response variable (not covered in this class)
- Mixture of quantitative and qualitative explanatory: Analysis of Covariance (ANCOVA).
- Qualitative explanatory variables: Analysis of Variance (ANOVA)

Regression Analysis Objectives

- Prediction of futures values of the response for specified values of the predictors
- Assessment of the relationship between the explanatory variables and the response. Is there a causal relationship???
- Summarize the relationship between variables.

Francis Galton example

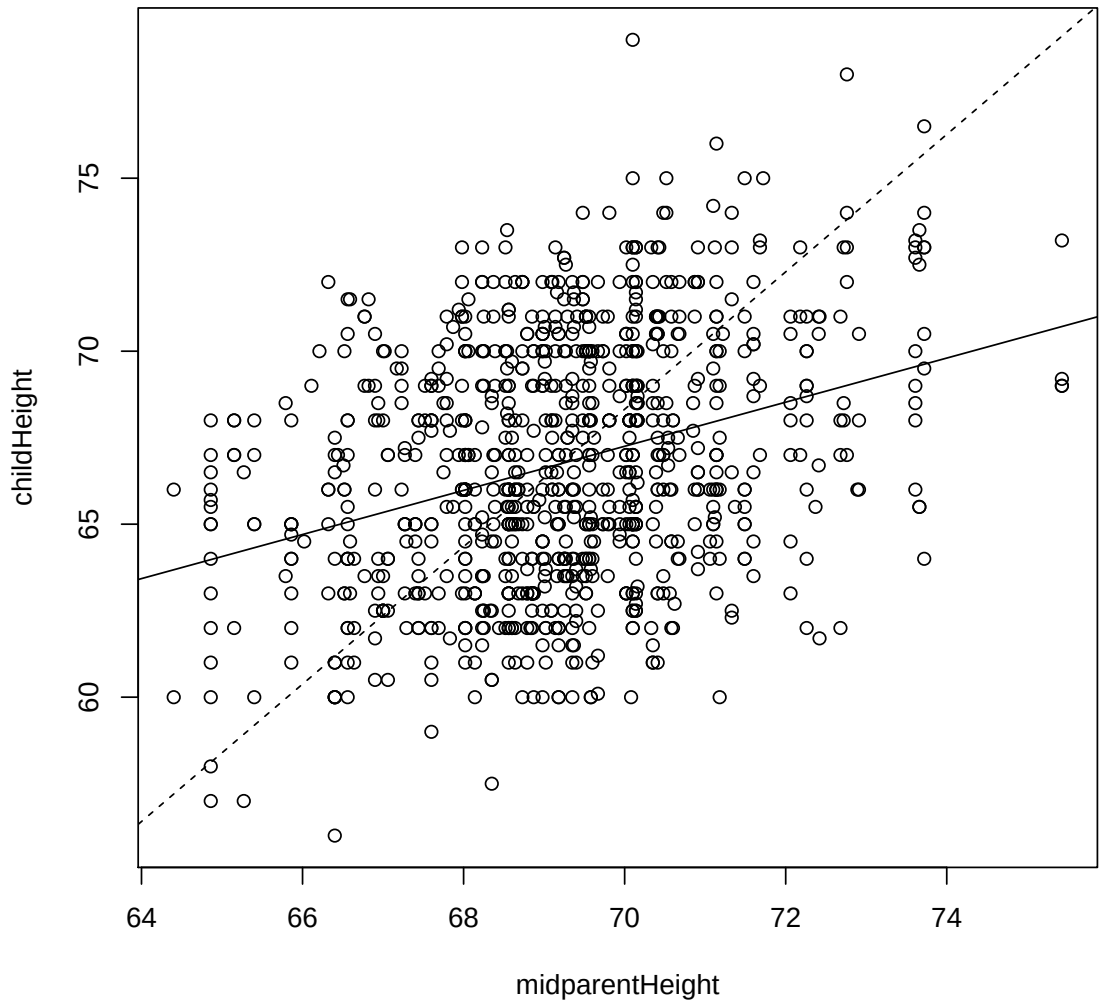
Francis Galton was a nephew of Charles Darwin. He coined the term *regression to mediocrity* in 1875, where the term regression comes from. For a response y and a single predictor x we can write the equation:

$$\frac{y - \bar{y}}{SD_y} = r\frac{x - \bar{x}}{SD_x}$$

r is the correlation between x and y. The response in standard units is the correlation r times the predictor in standard units. This equation produces the same results, by rearranging the equation in the form:

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

The height of the child is plotted against a combined parents height: (father's height + 1.08 mother's height)/2.



You would expect that a child from tall parents (height above the average), to be also with height above the average, but this is not the case unless the correlation r is close to 1 (dotted line). That is why Galton talks about *Regression to mediocrity.* or *Regression to the mean*.

More details on: `https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2011.00509.x`