Stat 206: Linear Models

Lecture 1

Sept. 25, 2019

Overview of Regression Analysis

Regression analysis is a statistical methodology to

(i) **describe** the relationship between a response variable Y and a set of predictor variables X and to

(ii) **predict** the values of the response variable based on those of the predictor variables.

- Simple regression: only one X variable.
- Multiple regression: more than one X variables.

History and Origin

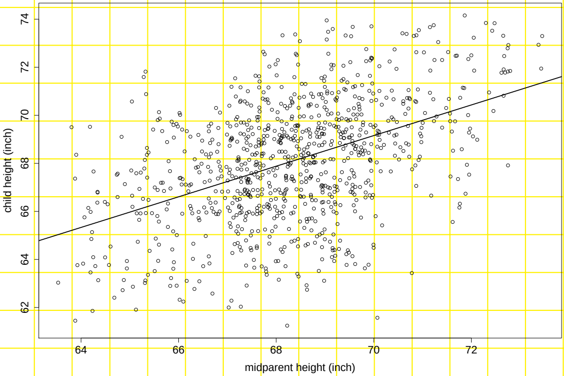
- 1885 study of Francis Galton of family resemblances.
- Height of the adult child, the midparent height – average of the height of the father and the adjusted height of the mother[1].
- Cases: 928 child-parent pairs.
- "**regression to mediocrity**": child's heights tend to be more moderate than their parents

[1] Heights of women were adjusted by multiplying 1.08 such that men's and women's heights would have the same mean.

	Child(inch)	Midparent(inch)
1	61.57220	70.07404
2	61.24382	68.22505
3	61.90968	65.12639
4	61.85769	64.23529
5	61.44986	63.88177
6	62.00005	67.02702

......

Figure: Scatter plot of child's height against parent's height



- Foot-ball shaped scatter plot $\Longrightarrow$ relationship between child's height (Y) and parent's height (X) appears to be linear.
- Fitted regression line:

$$Y = 24.54 + 0.637X$$

- Prediction: If the parent's height is 72in, then the child's height is predicted to be

$$24.54 + 0.637 \times 72in = 70.4in.$$

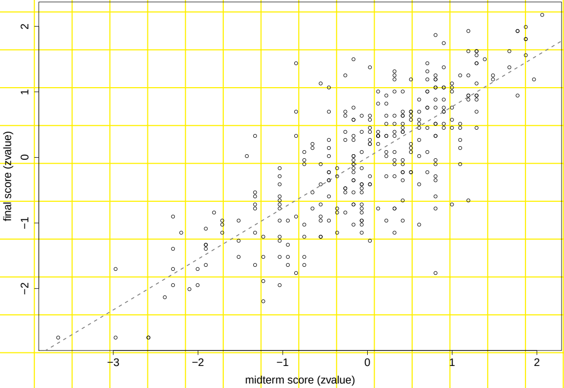
- Regression effect: children of very tall parents tended to be taller than their peers, but in a lesser degree than their parents compared to other parents.

Exam Scores

What is the relation between midterm score and final score?

- Variables: Standardized midterm exam score (X) and standardized final exam score (Y).
- Cases: 301 students from an elementary statistics class.
- Scatter plot: The relationship appears to be linear.
- Fitted regression line: $Y = 0.775X$. *Why is there no intercept? Why is the slope less than one?*
- Regression effect: If a student's midterm score is 2 standard deviations above the class mean, then his predicted final score would be 1.55 standard deviations above the class mean.

Figure: Scatter plot of final score against midterm score

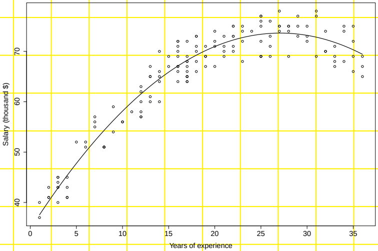


Salary

Salary survey of professional organizations relates salary to years of experience.[2]

- Variables: Years of experience (X) and salary (Y).
- Cases: 143 organizations.

Figure: Scatter plot of salary against years of experience



[2] Source of data: Tryfos (1998): Methods for business analysis and forecasting

Case	Salary	Experience
1	71	26
2	69	19
3	73	22
4	69	17
5	65	13
6	75	25
...,	....	

- The relationship appears to be: **curvilinear** (not linear).
- Fitted polynomial regression line:

$$Y = 34.721 + 2.872X - 0.053X^2.$$

Body Fat

Accurate measure of body fat is costly. It is desirable to use a set of easily obtainable measurements to estimate the body fat. [3]

- Variables: Percent body fat (Y) and 13 predictors (X): Age (years), Weight (lbs), Height (inches), Neck circumference (cm), Chest (cm), Abdomen 2 (cm), Hip (cm), Thigh (cm), Knee(cm), Ankle (cm), Biceps (cm), Forearm (cm), Wrist (cm).
- Cases: 252 men.
- A multiple regression model can be fitted to this data and then used for prediction of body fat of a future case :

$$Y = \hat{\beta}_0 + \sum \hat{\beta}_k X_k.$$

- *Are all 13 predictors needed for predicting Y? Are the effects of all predictors linear?*

[3] Source of data: `lib.stat.cmu.edu/datasets/`

Questions to Be Studied

- How to estimate the regression relationship? Least-squares principle
- How reliable are the regression estimates? Hypothesis testing and confidence intervals
- How reliable are the predictions? Prediction intervals
- Does the model fit the data? Do model assumptions hold? Model diagnostics
- How to choose X variables? How to choose between competing models? How to validate a model? Model building and validation.

Regression and Causation

- *Does ‘good midterm score” cause “good final score”?*
- A data on size of vocabulary (X) and writing speed (Y) for a group of children aged 5-10 showed a positive relationship.
- *Does this imply that an increase in vocabulary causes a faster writing speed? Can you think about other factors that may lead to such an association?*

- Regression analysis by itself does not imply casual-and-effect relation.
- A strong regression relation neither implies “X causes Y” nor implies “Y causes X”. It only means that there is a strong **association** between X and Y.
- Additional information, often through controlled experiments, is needed to draw cause-and-effect conclusions.

Basic Ingredients of Regression Model

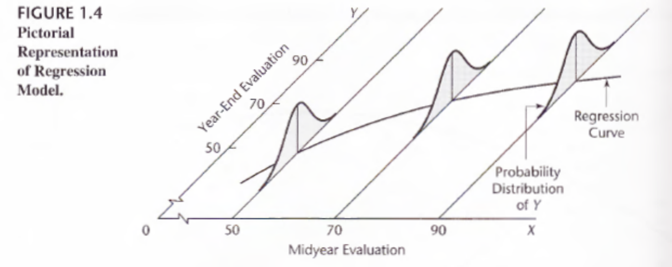
In this course, most analysis are conditioned on the values of the X variables such that they are treated as non-random $\Longrightarrow$ fixed design.

Key ingredients of a regression model:

(i) A probability distribution of the response variable Y for each given set of values of the X variables.

(ii) The means of these probability distributions vary in a systematic fashion with X.

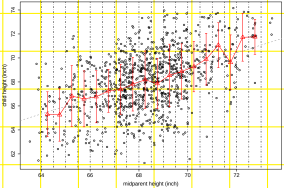
Figure: Illustration of regression model



From *"Applied Linear Statistical Models by Kutner, Nachtsheim, Neter and Li"*

Heights

Figure: Scatter plot of child's height against parent's height



- The average of the points falling in each vertical strip (bin) lies approximately on a straight line.
- The degree of dispersion of the points falling in each vertical strip is roughly the same.

The technique used here is called ”binning”. *Can you think another application of binning?*

Notations and definitions.

- Mean of a random variable Y, denoted by $E(Y)$.
- Variance of a random variable Y, denoted by $Var(Y)$ or $\sigma^2\{Y\}$.
- Covariance between two random variables Y, Z, denoted by $Cov(Y, Z)$ or $\sigma\{Y, Z\}$.

Check out appendix A.3 to review definitions of random variables, mean (a.k.a. expected value), variance and covariance.

Simple Linear Regression Model

n **cases** (trials/subjects): Y_i – the value of the response variable in the *ith* case; X_i – the value of the predictor variable in the *ith* case.

- **Model equation**:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1}$$

- **Model assumptions**:
  - ϵ_is are uncorrelated, zero-mean, equal-variance random variables:

$$E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad 1 \leq i \neq j \leq n.$$

- **Unknown parameters**:
  - β_0 – regression intercept; β_1 – regression slope
  - σ^2: error variance