

Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Overview: Part I

Regression Analysis: What for?

Regression analysis is a statistical technique to:

- (i) **Describe** the relationship between a response variable and a set of predictor variables;
- (ii) **Predict** the value of the response variable based on values of the predictor variables;

A Bit of History

- ▶ Francis Galton: Study of family resemblances, 1885
- ▶ 928 child-parent pairs: Height of the adult child and the “midparent height” (average height of the father and the mother)
- ▶ “**Regression to mean**”: Children’s heights tend to be more “moderate” than their parents

Child(in) Midparent(in)

1 61.57220 70.07404

2 61.24382 68.22505

3 61.90968 65.12639

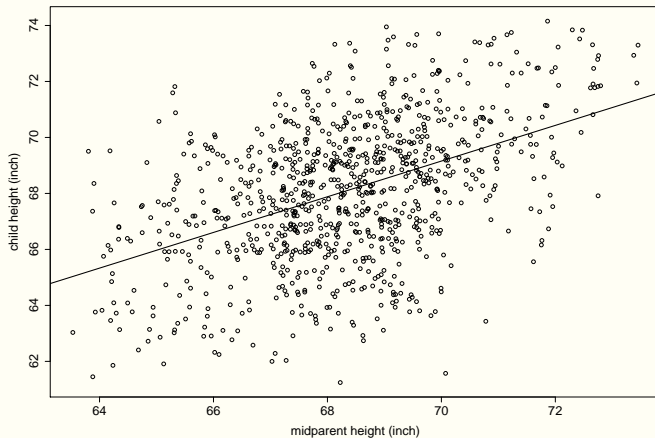
4 61.85769 64.23529

5 61.44986 63.88177

6 62.00005 67.02702

.

Figure: Scatter plot: Child's height versus midparent height



- ▶ Foot-ball shaped data cloud → linear relationship
- ▶ Fitted regression line:

$$Y = 24.54 + 0.637X$$

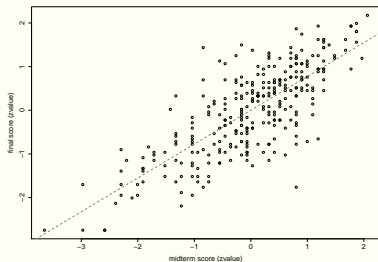
- ▶ If the midparent's height is 72in, then the child's height is predicted to be:

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

Overview: Part II

Exam Scores

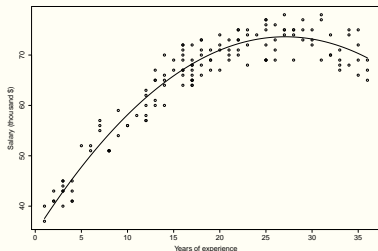
Figure: Standardized final score versus standardized midterm score



If a student's midterm score is 2 standard deviations above the class mean, then their predicted final score would be ...?

Salary

Figure: Salary versus years of experience



Would a straight line fit the data well?

Body Fat

Accurate measure of body fat is costly. It is desirable to use a set of easily obtainable measurements to predict the body fat. E.g., 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).

Are all these needed for predicting body fat? Are their effects linear?

Questions to Be Studied

- ▶ How to estimate the regression relationship?
- ▶ How reliable are the regression estimates?
- ▶ How reliable are the predictions?
- ▶ Does the model fit the data? Do model assumptions hold?
- ▶ How to choose X variables? How to choose between competing models? How to validate a model?

Review of Some Basics

Summation Operator: Σ

► $\sum_{i=1}^n Y_i = Y_1 + Y_2 + \cdots + Y_n$

► Variants:

$$\sum_{i=1}^n Y_i, \quad \sum_1^n Y_i, \quad \sum Y_i$$

► Useful facts:

$$\sum_{i=1}^n c = n \cdot c$$

$$\sum_{i=1}^n (Y_i + Z_i) = \sum_{i=1}^n Y_i + \sum_{i=1}^n Z_i$$

$$\sum_{i=1}^n c \cdot Y_i = c \cdot \sum_{i=1}^n Y_i$$

Expectation Operator: $E(\cdot)$

- ▶ Discrete random variable: $E(Y) = \sum_y y \cdot P(Y = y)$
- ▶ Continuous random variable: $E(Y) = \int y \cdot f(y) dy$
- ▶ Variants: expectation, mean, expected value
- ▶ Useful facts: (c is a constant and Y, Z are random variables)

$$E(c) = c$$

$$E(Y + Z) = E(Y) + E(Z)$$

$$E(c \cdot Y) = c \cdot E(Y)$$

Variance Operator: $Var(\cdot)$

- ▶ $Var(Y) = E(Y - E(Y))^2$
- ▶ Variants: $\sigma^2\{Y\}$
- ▶ Useful things: (c is a constant and Y, Z are random variables)

$$Var(Y) = E(Y^2) - (E(Y))^2$$

$$Var(c) = 0$$

$$Var(Y + Z) = Var(Y) + Var(Z), \text{ if } Y \text{ and } Z \text{ are uncorrelated}$$

$$Var(c + Y) = Var(Y)$$

$$Var(c \cdot Y) = c^2 \cdot Var(Y)$$

Covariance Operator: $\text{Cov}(\cdot, \cdot)$

- ▶ $\text{Cov}(Y, Z) = E\left(\left(Y - E(Y)\right) \cdot \left(Z - E(Z)\right)\right)$
- ▶ Variants: $\sigma\{Y, Z\}$
- ▶ Useful facts: (c_1, c_2 are constants and Y, Z are random variables)

$$\text{Cov}(Y, Z) = E(Y \cdot Z) - E(Y) \cdot E(Z)$$

$$\text{Cov}(Y, Y) = \text{Var}(Y)$$

$$\text{Cov}(c_1 + Y, c_2 + Z) = \text{Cov}(Y, Z)$$

$$\text{Cov}(c_1 \cdot Y, c_2 \cdot Z) = c_1 \cdot c_2 \cdot \text{Cov}(Y, Z)$$

$$\text{Var}(Y + Z) = \text{Var}(Y) + \text{Var}(Z) + 2 \cdot \text{Cov}(Y, Z)$$

Hypothesis Testing: Components

- ▶ **Null hypothesis** H_0 : “status quo”. E.g., H_0 : not guilty
 - ▶ Null hypothesis is assumed to be true unless there is strong evidence against it. E.g., A defendant is assumed not guilty unless proved otherwise by evidence that is beyond reasonable doubts.
- ▶ **Alternative hypothesis** H_a : “new theory”. E.g., H_a : guilty
 - ▶ Alternative hypothesis is being accepted only when one can reject the null hypothesis with strong evidence.
- ▶ A testing procedure is to ascertain if the evidence (in data) against the null hypothesis is strong enough to reject it.

Type I and Type II Errors

Decision \ Actual Fact	H_0 true	H_0 false
Accept H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

- ▶ **Significance level α :** (prespecified) maximum allowable **type I error rate**. E.g., $\alpha = 0.05$, $\alpha = 0.1$, $\alpha = 0.01$. **Type I error rate** is not to exceed the significance level α .
- ▶ **Type II error rate** is affected by the testing procedure, signal-to-noise ratio, sample size, significance level α , etc.

Point Estimator

A procedure to calculate a numerical quantity based on a (random) sample.

- ▶ E.g., Sample mean for the population mean; sample proportion for the population proportion.
- ▶ An estimator is a random variable.
- ▶ Error in the estimation: a numerical measure of the reliability of the estimation. E.g., standard error (SE)

Confidence Interval

A (random) interval that covers the parameter of interest with a (pre-specified) high probability.

- ▶ Confidence coefficient (level/coverage) : the pre-specified probability of coverage, denoted by $1 - \alpha$. E.g., $1 - \alpha = 90\%$, $1 - \alpha = 95\%$, $1 - \alpha = 99\%$.
- ▶ The corresponding interval is called a $(1 - \alpha)100\%$ -confidence interval.
- ▶ One common form:

$$\text{Estimator} \pm \text{multiplier}_{\alpha} \times SE(\text{Estimator})$$