Stat 206: Linear Models

Lecture 1

Sept. 28, 2015

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 predictor variable (Part I).
- Multiple regression: more than one predictor variables (Part II).

History and Origin

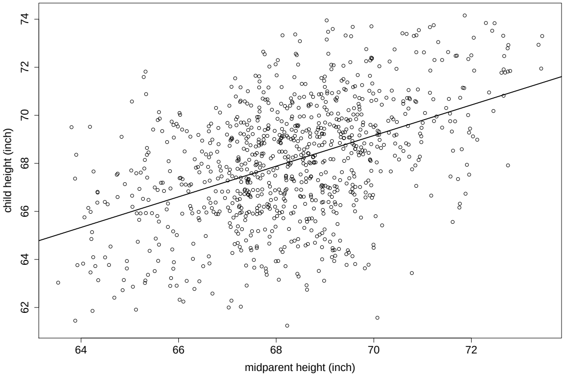
Regression type analysis was used by Galton (1822-1911) in his study of family resemblances. He noted that child's heights tend to be more moderate than their parents, an effect he called "**regression to mediocrity**".

- 1885 study of Francis Galton.
- Variables: 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.

[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

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

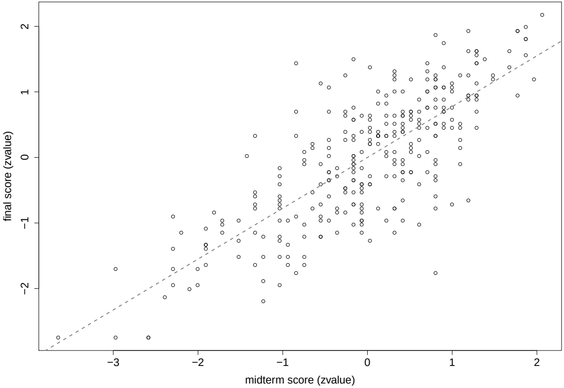
- Regression effect: For very tall parents, their children tended to be taller than their peers, but not to the extent as their parents compared with other parents.

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?*
- 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

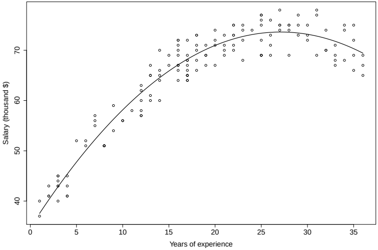


Salary

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 Answered

- How to estimate the regression relationship? Utilize Least-squares principle
- How good are the regression estimates? Quantified by Standard errors and confidence intervals
- How reliable are the predictions? Quantified Prediction intervals
- Does the model fit the data? Do model assumptions hold? Checked by Model diagnostics
- How to choose X variables? How to choose between competing models? How to validate a model? Utilize Model building and validation

Regression and Causality

- *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 X/fixed design.

Key ingredients of a regression model:

(i) There is 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

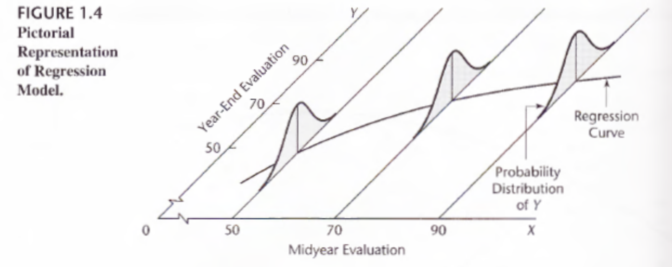
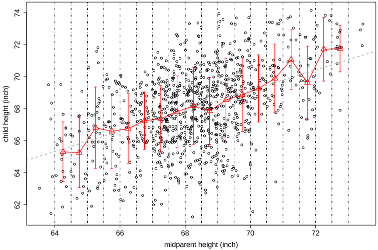


FIGURE 1.4 Pictorial Representation of Regression Model.

Heights

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



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

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 for 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:

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

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

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

Given X_is, the distributions of the responses Y_is have the following properties:

- The response Y_i is the sum of two terms:
  - a non-random term – the mean of Y_i:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

  - a (zero-mean) random term – the random error ϵ_i .
- Error terms ϵ_i have constant variance $\Longrightarrow$ Y_is have the same constant variance (regardless of the values of X_i):

$$\mathrm{Var}(Y_i) = \sigma^2, \quad i = 1, \cdots, n.$$

- Error terms are uncorrelated $\Longrightarrow$ Y_is are uncorrelated:

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

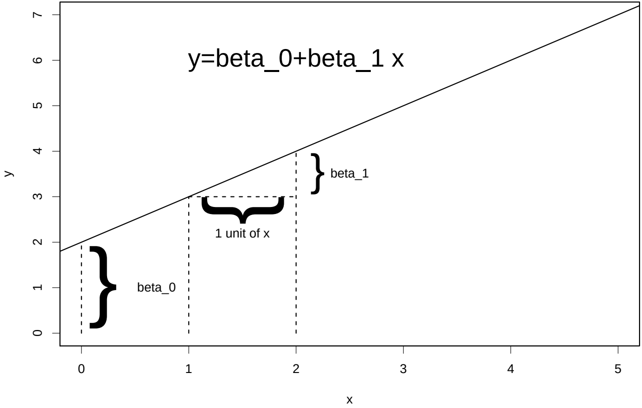
- In summary, the simple linear regression model says that the responses Y_i are random variables whose means are linear in X_i and whose variances are a constant. Moreover, two responses Y_i and Y_j $(i \neq j)$ are uncorrelated.
- Regression function:

$$y = \beta_0 + \beta_1 x$$

  is a straight line.
  - β_1 is the slope of the regression line: the change in $E(Y)$ per unit change of X.
  - β_0 is the intercept of the regression line: the value of $E(Y)$ when $X = 0$.

Are the distributions of the responses Y_i fully specified by the above model?

Figure: Regression line: $y = \beta_0 + \beta_1 x$



Least Squares Principle

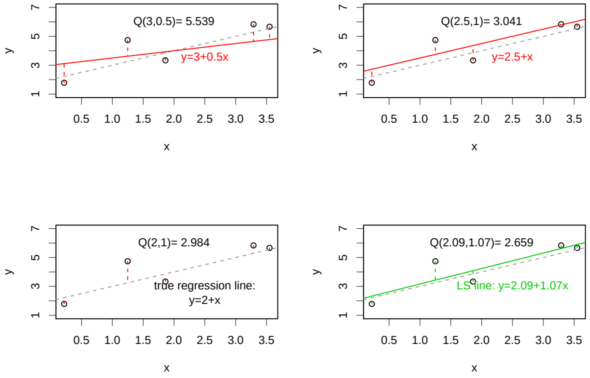
For a given line: $y = b_0 + b_1 x$, the *sum of squared vertical deviations* of the observations $\{(X_i, Y_i)\}_{i=1}^n$ from the corresponding points on the line is:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2 .$$

- $(X_i, b_0 + b_1 X_i)$ is the point on the line with the same x-coordinate as the ith observation point (X_i, Y_i).
- The *least squares (LS) principle* is to fit the observed data by minimizing the sum of squared vertical deviations.

LS line has the smallest sum of squared vertical deviations among all straight lines.

Figure: Illustration of LS principle



Which line has the smaller sum of squared vertical deviations, the LS line (a.k.a. the fitted regression line) or the true regression line?

Least Squares Estimators

LS estimators of β_0, β_1 are the pair of values b_0, b_1 that minimize the function $Q(\cdot, \cdot)$:

$$(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{argmin}_{\mathrm{b_0,b_1}} Q(b_0, b_1).$$

- LS estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \qquad (2)$$

- $\overline{X} = 1/n \sum_{i=1}^n X_i$, $\overline{Y} = 1/n \sum_{i=1}^n Y_i$ are the sample means.
- *Is there a situation such that the LS estimators are not defined?*