Stat 213: Intro to Statistics 2

H. Kim

Fall 2007

bivariate data

- two variables are measured on a single experimental unit, (x_i, y_i)
  - is there a linear relationship between two variables ?
  - if a relationship exists between two variables, how strong is the relationship ?
  - if the relationship between variables is strong, can the existing relationship be used to predict what will happen in the future ?

bivariate data

- scatter plot: each pair of data values is plotted as a point in the two-dimensional graph
  - the best way to start observing the relationship between two quantitative variables: see patterns, trends, relationships, even the outliers sitting apart from the others

correlation coefficient, r

- correlation: measures the strength of the linear relationship between two quantitative variables (correlation coefficient, r)

$$r = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$$

  - where $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the standard deviation for the variable x and y
  - $\hat{\sigma}_{xy}$ is the covariance between x and y

correlation coefficient, r

- covariance between x and y:

$$\begin{aligned} \hat{\sigma}_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{n-1} \end{aligned}$$

- correlation coefficient:

$$r =$$

correlation coefficient, r

- looking for a linear relationship between two variables: positive or negative or neither ?
  - $-1 < r < 1$
  - $r > 0$: positive pattern
  - $r < 0$: negative pattern
  - $r = 1$ or $r = -1$: points lie exactly on a straight line
  - $r = 0$: no linear pattern
  - the closer the value of r to 1 or -1, the stronger the linear relationship between the two variables

example

data

x	2	3	4	5	6	7
y	3	5	5.5	6	8	9.5

two quantitative variables, x and y

- the two variables, X and Y, are related in a particular way
- it may be that the value of Y depends on the value X; Y (dependent or response variable) and X (independent or predictor variable)
- if the data exhibit a straight line pattern, it is possible to describe the relationship relating Y to X using a straight line given by the equation

$$Y = \beta_0 + \beta_1 X$$

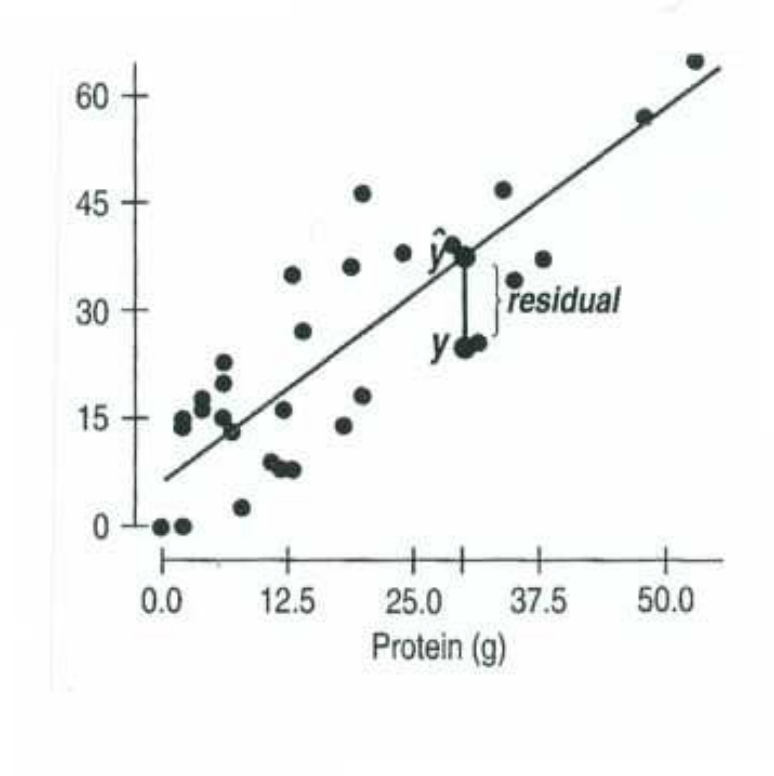
- how can we determine the straight line that "best" passes through the data point ?

least squares method

- statistical model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \cdots, n$
  - Y_i: response variable
  - X_i : predictor variable
  - β_0: y-intercept term: part of average value of y that does not change with changes in x
  - β_1: the slope term: change in average value of the response variable that corresponds to a unit change in the predictor variable
- for the observations of the data, $(x_1, y_1), \cdots, (x_n, y_n)$, the least squares method finds a values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that make the sum of squared deviations, $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ as small as possible

regression line

- the best fitting line relating y to x, called regression line is found by minimizing the sum of squared differences between the data points and the line itself



least squares line

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are calculated through the formula:

$$\begin{aligned} \hat{\beta}_1 &= r \times \frac{\hat{\sigma}_y}{\hat{\sigma}_x} = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} \\ \hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} \end{aligned}$$

- least squares line: an equation of a straight line through the data: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - $\hat{\beta}_0$: part of estimated average value of y that does not change with changes with x
  - $\hat{\beta}_1$: change in estimated average value of y per unit increase in x

example

data

x	2	3	4	5	6	7
y	3	5	5.5	6	8	9.5

- the best fitting line is:

coefficient determination

- choosing $\hat{\beta}_1$ minimizes the sum of the squared residuals, $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$, but how small does that sum get ?

  $1 - r^2$

- coefficient of determination: the percentage of variability in the response variable y that is explained by the predictor variable x: r^2 and $0 \leq r^2 \leq 1$
  - used to assess how well the linear model fits the data
- predicted value: the estimate made from a model: $\hat{y}$ given a new x value

example

data

x	2	3	4	5	6	7
y	3	5	5.5	6	8	9.5

- the best fitting line can be used to estimate or predict the value of the variable y when the value of x is known
  - if a new value $x = 3.5$ is observed, what would you predict for the value of y ? ( )
  - the percentage of variability in y that is explained by x ? ( )

example

- in order to develop an equation to predict sales of a product in terms of its selling price, the company's research department obtained the following data from a random sample of 10 of its outlets.

data

price	15	15.5	16	16.5	16.5	17.5	18	18.5	19	19.5
sales ($000)	19	11	11	14	10	10	8	9	6	5

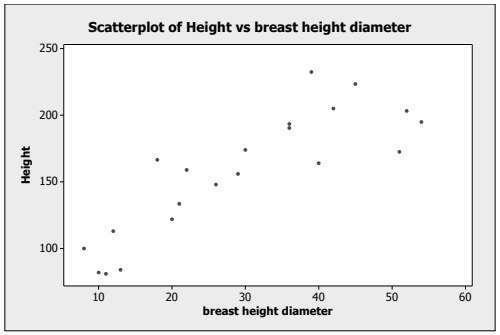
- the response variable Y: ( )
- the predict variable X: ( )

the scatter plot show a more or less straight line relationship between Y and X, find the fitted least squares line if $\sum x_i = 172$, $\sum y_i = 103, \sum y_i^2 = 1205, \sum x_i^2 = 2979.5, \sum x_i y_i = 1724.5$

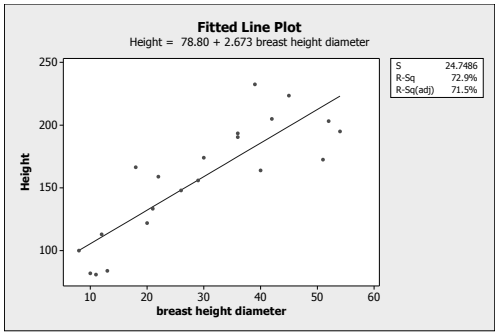
redwood data

- the **height** of a redwood tree is related to other characteristics of the the tree, the **diameter of the tree at the breast height** (dbh) of a person ($n = 21$)

scatter plot



fitted line



Minitab: Stat > Regression > Regression or Fitted line plot

output: redwood

```
The regression equation is
Height = 78.8 + 2.67 breast height diameter


Predictor                 Coef  SE Coef     T      P
Constant                 78.80    12.21  6.45  0.000
breast height diameter  2.6732   0.3741  7.15  0.000


S = 24.7486   R-Sq = 72.9%   R-Sq(adj) = 71.5%
```

questions

- find the regression coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$):

- state the least squares line (regression line):
- predict the height for a redwood tree that has a dbh of 30 inches:
- interpret the meaning of $\hat{\beta}_1$:

- find the coefficient of determination, and explaining its meaning in this problem: