Correlation and Simple Linear Regression

Berlin Chen

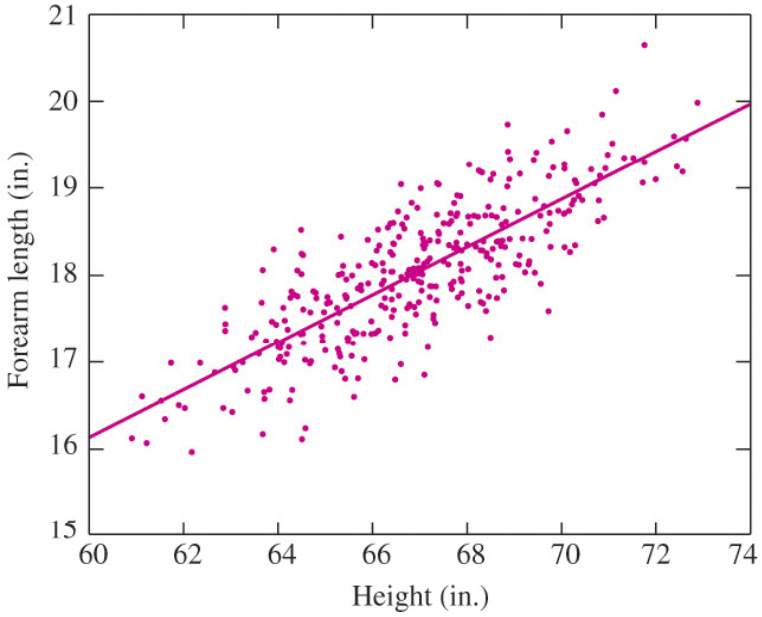
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:

1. W. Navidi. *Statistics for Engineering and Scientists*. Chapter 7 (7.1-7.3) & Teaching Material

Introduction (1/2)

- Often, scientists and engineers collect data in order to determine the nature of the relationship between two quantities
  - An example: heights and forearm lengths of men



  - The points tend to slop upward and to the right, indicating that taller men tend to have longer forearms
  - A “positive association” between height and forearm length

Introduction (2/2)

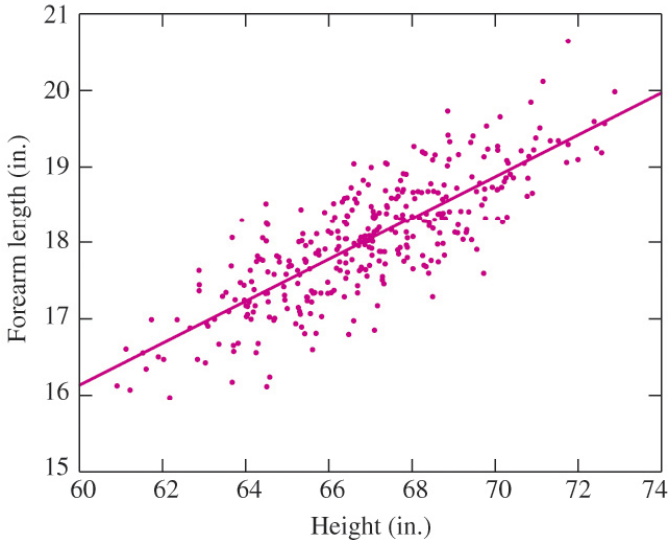
- Many times, this ordered pairs of measurements fall approximately along a straight line when plotted
  - In those situations, the data can be used to compute an equation for the line that “best fits” the data
  - This line can be used for various things:
    - Summarize the data
    - Predict for future (unseen) values

Correlation

- Something we may be interested in is how closely related two physical characteristics are
  - For example, height and weight of a two-year-old child
- The quantity called the correlation coefficient is a measure of this
- We look at the direction of the relationship, positive or negative, strength of relationship, and then we find a line that best fits the data
- In computing correlation, we can only use quantitative data (instead of qualitative data)

Example

- This is a plot of height vs. forearm length for men



  - We say that there is a positive association between height and forearm length
  - This is because the plot indicates that taller men tend to have longer forearms
  - The slope is roughly constant throughout the plot, indicating that the points are clustered around a straight line
  - The line superimposed on the plot is a special line known as the least-squares line

Correlation Coefficient

- The degree to which the points in a scatterplot tend to cluster around a line reflects the strength of the linear relationship between *x* and *y*
- The correlation coefficient is a numerical measure of the strength of the linear relationship between two variables
- The correlation coefficient is usually denoted by the letter *r*
- Also called "sample correlation" (cf. population correlation $\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbf{E}[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])]}{\sqrt{\mathbf{E}[X^2]-(\mathbf{E}[X])^2}\sqrt{\mathbf{E}[Y^2]-(\mathbf{E}[Y])^2}}$)

Computing Correlation Coefficient *r*

- Let $(x_1, y_1),\ldots,(x_n, y_n)$ represent *n* points on a scatterplot
- Compute the means and the standard deviations of the *x*'s and *y*'s
- Then convert each *x* and *y* to standard units. That is, compute the *z*-scores: $(x_i - \bar{x})/s_x$ and $(y_i - \bar{y})/s_y$.
- The correlation coefficient is the average of the products of the *z*-scores, except that we divide by $n - 1$ instead of *n*

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Sometimes, this computation is more useful

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

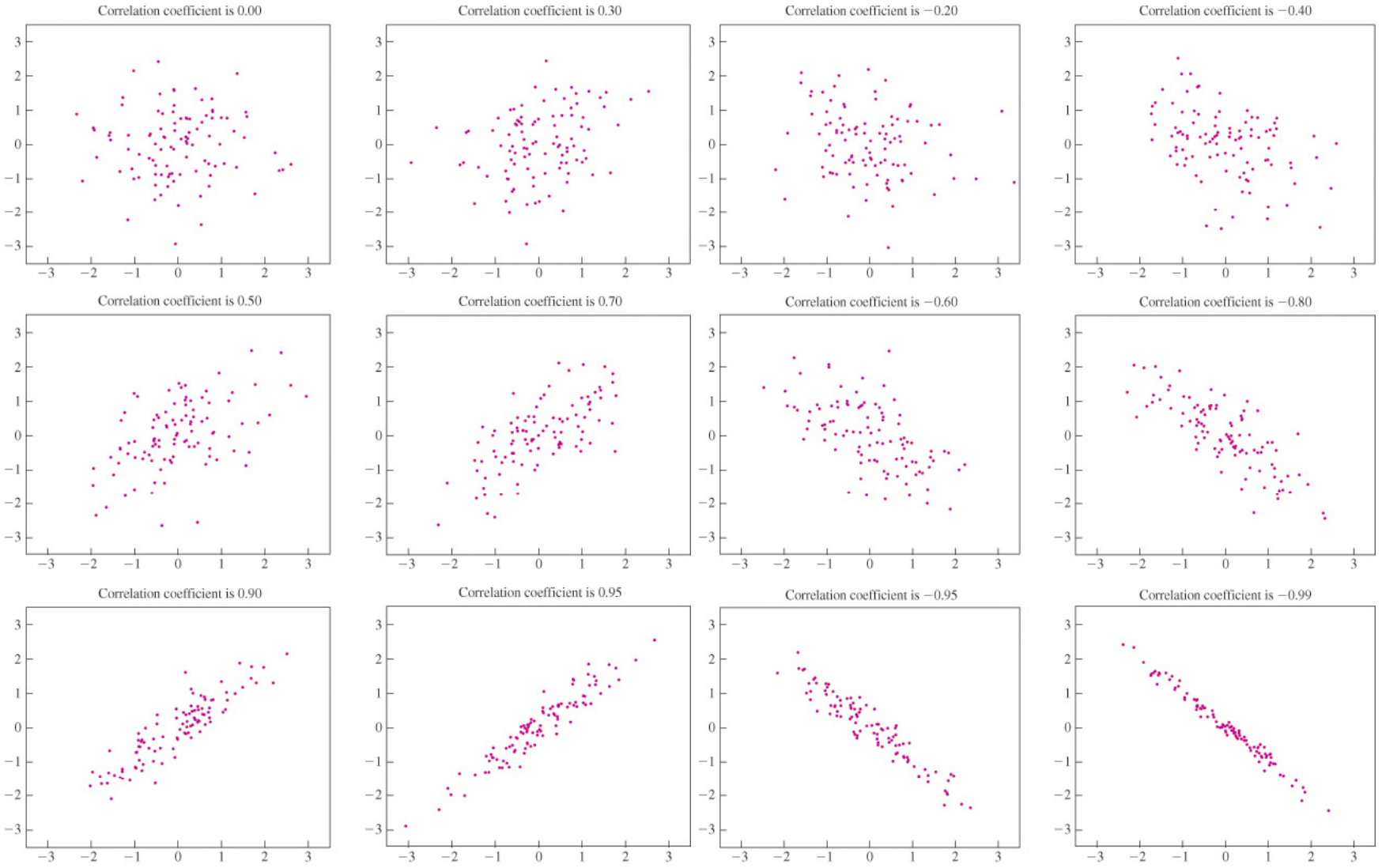
$$s_x = \sqrt{\textstyle\sum_{i=1}^{n}(x_i - \bar{x})/n-1}$$

$$s_y = \sqrt{\textstyle\sum_{i=1}^{n}(y_i - \bar{y})/n-1}$$

Comments on Correlation Coefficient

- In principle, the correlation coefficient can be calculated for any set of points
- In many cases, the points constitute a random sample from a population of points
- In this case, the correlation coefficient is called the sample correlation, and it is an estimate of the population correlation
- It is a fact that *r* is always between -1 and 1
- Positive values of *r* indicate that the least-squares line has a positive slope. The greater values of one variable are associated with greater values of the other
- Negative values of *r* indicate that the least-squares line has a negative slope. The greater values of one variable are associated with lesser values of the other

Examples of Various Levels of Correlation



More Comments

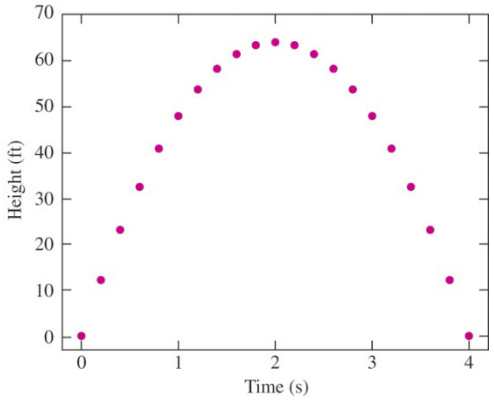
- Values of *r* close to -1 or 1 indicate a strong linear relationship
- Values of *r* close to 0 indicate a weak linear relationship
- When *r* is equal to -1 or 1, then all the points on the scatterplot lie exactly on a straight line
- If the points lie exactly on a horizontal or vertical line, then *r* is undefined
- If $r \neq 0$, then *x* and *y* are said to be correlated. If $r = 0$, then *x* and *y* are uncorrelated

Properties of Correlation Coefficient *r* (1/2)

- An important feature of *r* is that it is unitless. It is a pure number that can be compared between different samples
- *r* remains unchanged under each of the following operations:
  - Multiplying each value of a variable by a positive constant
  - Adding a constant to each value of a variable
  - Interchanging the values of *x* and *y*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- If $r = 0$, this does not imply that there is not a relationship between *x* and *y*. It just indicates that there is no *linear* relationship

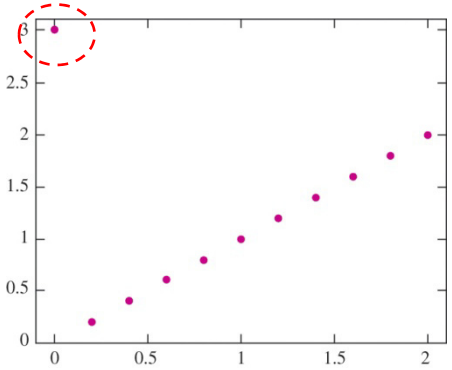


$$y = 64x - 16x^2$$

quadratic relationship

Properties of Correlation Coefficient *r* (2/2)

- Outliers can greatly distort *r*, especially, in small data sets, and present a serious problem for data analysts



correlation coefficient *r*=0.26

- Correlation is not causation
  - For example, vocabulary size is strongly correlated with shoe size, but this is because both increase with age. Learning more words does not cause feet to grow and vice versus. Age is confounding the results

Inference on the Population Correlation

- If the random variables *X* and *Y* have a certain joint distribution called a bivariate normal distribution, then the sample correlation *r* can be used to construct confidence intervals and perform hypothesis tests on the population correlation, ρ. The following results make this possible
  - Let $(x_1, y_1),\ldots,(x_n, y_n)$ be a random sample from the joint distribution of *X* and *Y* and *r* is the sample correlation of the *n* points. Then the quantity

$$W = \frac{1}{2}\ln\frac{1+r}{1-r} \qquad \text{(a function of } r\text{)}$$

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}, \mu_Z = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

$$\Sigma_Z = \begin{bmatrix} \sigma_{X,X} & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_{X,X} \end{bmatrix}$$

is approximately normal with mean $\mu_W = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$

and variance $\sigma_W^2 = \frac{1}{n-3}$.

Example 7.3

- Question: Find a 95% confidence for the correlation between the reaction time of visual stimulus (x) and that of audio stimulus (y), given the following sample

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

The sample correlation between x and y is $r = 0.8159$

$$\Rightarrow \quad W = \frac{1}{2}\ln\frac{1+r}{1-r} = \frac{1}{2}\ln\frac{1+0.8159}{1-0.8159} = 1.1444$$

$$\sigma_W = \sqrt{1/(10-3)} = 0.3780$$

A 95% (two - sided) confidence interval for μ_W is given by

$$1.1444 - 1.96(0.3780) \le \mu_W \le 1.1444 + 1.96(0.3780)$$

$$0.4036 \le \mu_W \le 1.8852$$

Note that the population correlation ρ can be expressed as

$$\rho = \frac{e^{2\mu_W} - 1}{e^{2\mu_W} + 1}$$

The corresponding 95% confidence interval for ρ

$$\Rightarrow \quad \frac{e^{2\cdot 0.4036} - 1}{e^{2\cdot 0.4036} + 1} \le \frac{e^{2\mu_W} - 1}{e^{2\mu_W} + 1} \le \frac{e^{2\cdot 1.8852} - 1}{e^{2\cdot 1.8852} + 1}$$

$$\Rightarrow \quad 0.383 \le \rho \le 0.955$$

Linear Model

- When two variables have a linear relationship, the scatterplot tends to be clustered around a line known as the least-squares line
  - The line that we are trying to fit is

ideal value $l_i = \beta_0 + \beta_1 x_i$

measured value $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (measurement error ε_i)

dependent variable

independent variable

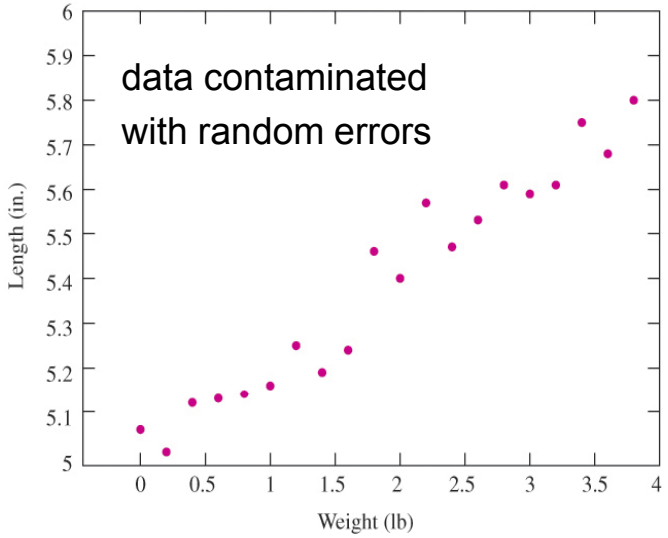
    - β_0 and β_1 are called the regression coefficients
    - We only know the values of *x* and *y*, we must estimate the other quantities
  - This is what we call simple linear regression
    - With only one independent variable
  - We use the data to estimate these quantities

The Least-Squares Line

- β_0 and β_1 cannot be determined because of measurement error, but they can be estimated by calculating the least-squares line

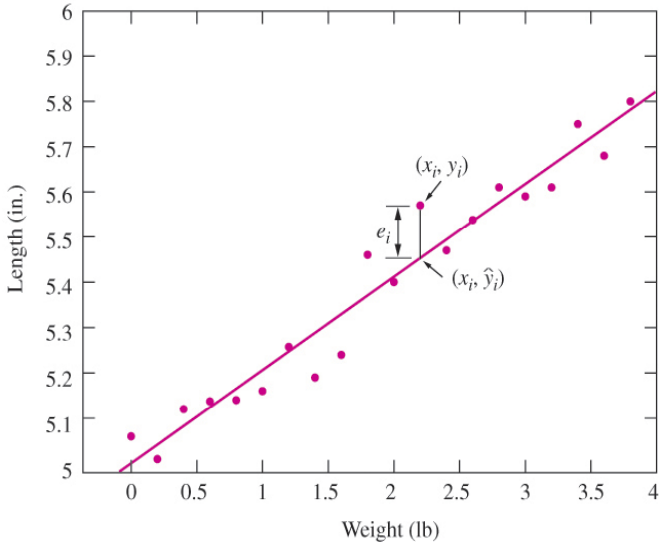
$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least-squares coefficients
  - The least-squares line is the line that fits the data "best" (?)



fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

residual $e_i = y - \hat{y}_i$



Residuals

- For each data point (x_i, y_i), the vertical distance to the point $(x_i, \hat{y}_i)$ on the least squares line is $e_i = y_i - \hat{y}_i$. The quantity $\hat{y}_i$ is called the fitted value and the quantity e_i is called the residual associated with the point (x_i, y_i)
  - Points above the least-squares line have positive residuals
  - Points below the line have negative residuals
- The closer the residuals are to 0, the closer the fitted values are to the observations and the better the line fits the data
- The least-squares line is the one that minimizes the sum of squared residuals

$$S = \sum_{i=1}^{n} e_i^2$$

Finding the Equation of the Line

- To find the least-squares line, we must determine estimates for the slope β_0 and β_1 intercept that minimize the sum of the squared residuals

$$E^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- These quantities are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)\left(y_i - \overline{y} \right)}{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Some Shortcut Formulas

- The expressions on the right are equivalent to those on the left, and are often easier to compute

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

Cautions (1/2)

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are not the same as the true values β_0 and β_1
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables
  - β_0 and β_1 are constants whose values are unknown
- The residuals e_i are not the same as the errors ε_i
  - The residuals e_i can be computed, which are the differences between y_i and $\hat{y}_i$

$$e_i = y_i - \hat{y}_i$$

  - The errors ε_i can not be computed, since the true values β_0 and β_1 are unknown (or l_i is unknown)

$$\varepsilon_i = y_i - l_i$$

Cautions (2/2)

- Do not extrapolate the fitted line (such as the least-squares line) outside the range of the data. The linear relationship may not hold there

- We learned that we should not use the correlation coefficient when the relationship between *x* and *y* is not linear. The same holds for the least-squares line. When the scatterplot follows a curved pattern, it does not make sense to summarize it with a straight line

- If the relationship is curved, then we would want to fit a regression line that contain squared terms (i.e., polynomial regression)

Another Representation of the Line

- Another way to compute an estimate of β_1 is

$$\hat{\beta}_1 = r\frac{s_y}{s_x} \qquad s_y \Big/ s_x = \sqrt{\textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2} \Big/ \sqrt{\textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - The units (but not the value) of $\hat{\beta}_1$ must be same as *y*/*x*
  - The slope is proportional to the correlation coefficient

- The least-sqaures line can be rewritten as

$$\hat{y} - \bar{y} = r\frac{s_x}{s_y}(x - \bar{x})$$

  - So the line passes through the center of the mass of the scatterplot with slope $r(s_x/s_y)$

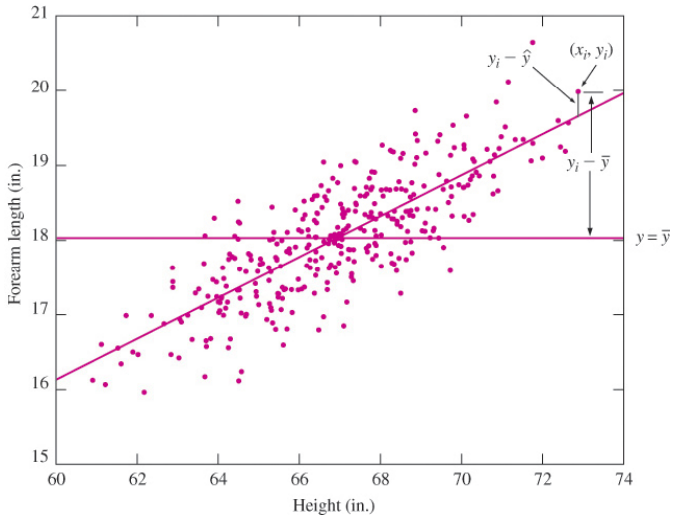
Measures of Goodness of Fit

- A goodness of fit statistic is a quantity that measures how well a model explains a given set of data
- The quantity r^2 is the square of the correlation coefficient and we call it the coefficient of determination

error sum of squares

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

total sum of squares



  - The proportion of variance in y explained by regression is the interpretation of r^2
  - $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ measures the reduction of spread of the points obtained by using the leas-squares line rather than $y = \bar{y}$
    - A goodness-of-fit statistic that has units

Sums of Squares (1/2)

- $\sum_{i=1}^{n}(y_i - \hat{y})^2$ is the error sum of squares (SSE) and measures the overall spread of the points around the least-squares line
- $\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares (SST) and measures the overall spread of the points around the line $y = \bar{y}$
- The difference $\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is called the regression sum of squares (SSR)
- Clearly, the following relationship holds:

Total sum of squares (SST) = regression sum of squares (SSR)
+ error sum of squares (SSE)

Sums of Squares (2/2)

- The analysis of variance identity has the form

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

- r^2 is also called the proportion of the variance in y explained by regression

$$r^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Uncertainties in the Least-Squares Coefficients

- Assumptions for Errors in Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- In the simplest situation, the following assumptions are satisfied:
  1. The errors $\varepsilon_1,\ldots,\varepsilon_n$ are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}
  2. The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have mean 0
  3. The errors $\varepsilon_1,\ldots,\varepsilon_n$ all have the same variance, which we denote by σ^2 (variance of the error)
  4. The errors $\varepsilon_1,\ldots,\varepsilon_n$ are normally distributed

Distribution of y_i

- In the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, under assumptions 1 through 4, the observations $y_1,\ldots,y_n$ are independent random variables that follow the normal distribution. The mean and variance of y_i are given by

$$\mu_{y_i} = l_i = \beta_0 + \beta_1 x_i$$

$$\sigma^2_{y_i} = \sigma^2.$$

- The slope β_1 represents the change in the mean of y associated with an increase in one unit in the value of x

Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (1/2)

- Under assumptions 1 – 4:
  - The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed random variables

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}y_i - \left(\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\bar{y}$$

0

$$= \sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]y_i$$

- After further manipulation, we have

$$\mu_{\hat{\beta}_0} = \beta_0$$

$$\mu_{\hat{\beta}_1} = \beta_1$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates

Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (2/2)

- The standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated with

$$\sigma_{\hat{\beta}_0} = \sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$\sigma \overset{?}{\approx} s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}} = \sqrt{\frac{(1-r^2)\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-2}}$$

The residuals tends to be a little smaller than the errors

  - s is an estimate of the error standard deviation σ

Notes

1. Since there is a measure of variation of *x* in the denominator in both of the uncertainties we just defined, the more spread out *x*'s are the smaller the uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

2. Use caution, if the range of *x* values extends beyond the range where the linear model holds, the results will not be valid

3. The quantities $\left(\hat{\beta}_0 - \beta_0\right)/s_{\hat{\beta}_0}$ and $\left(\hat{\beta}_1 - \beta_1\right)/s_{\hat{\beta}_1}$ have Student's *t* distribution with $n - 2$ degrees of freedom

Confidence Intervals for β_0 and β_1

- Level 100(1-α)% confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} s_{\hat{\beta}_0}$$

  - and

two-sided confidence intervals

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} s_{\hat{\beta}_1}$$

Summary

- We discussed
  - Correlation
  - Least-squares line / regression
  - Uncertainties in the least-squares coefficients
  - Confidence intervals (and hypothesis tests) for least-squares coefficients