Linear Regression


"The Least-Squares Line didn't fit my data, so I decided to try the Hollywood Squares line."

Today's Class

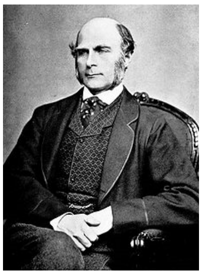
- Simple Linear Regression
- Least Square Estimates

Simple Linear Regression

- Analyzing correlated data
- "Fits" data to a line

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

  - x is the independent, predictor, or explanatory variable
  - y is the dependent or response variable



Francis Galton
(1822 – 1911)

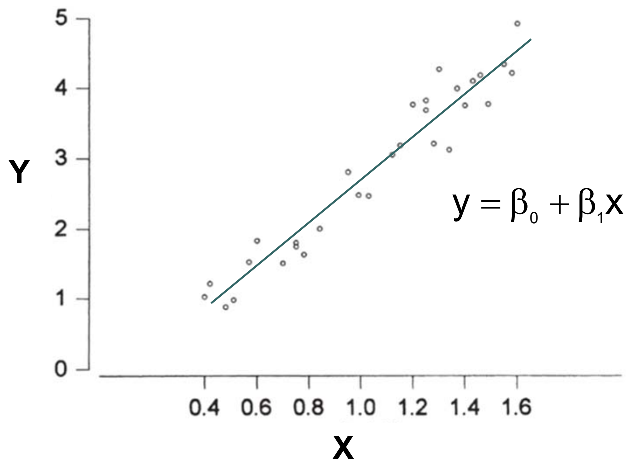
Example 12.1

- Vertical gaze direction as a source of eye strain and irritation.
  - y = ocular surface area (cm^2)
  - x = width of the palpebral fissure (i.e., the horizontal width of the eye opening, in cm)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	.40	.42	.48	.51	.57	.60	.70	.75	.75	.78	.84	.95	.99	1.03	1.12
y_i	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

Example 12.1, Cont'd



Linear Probabilistic Model

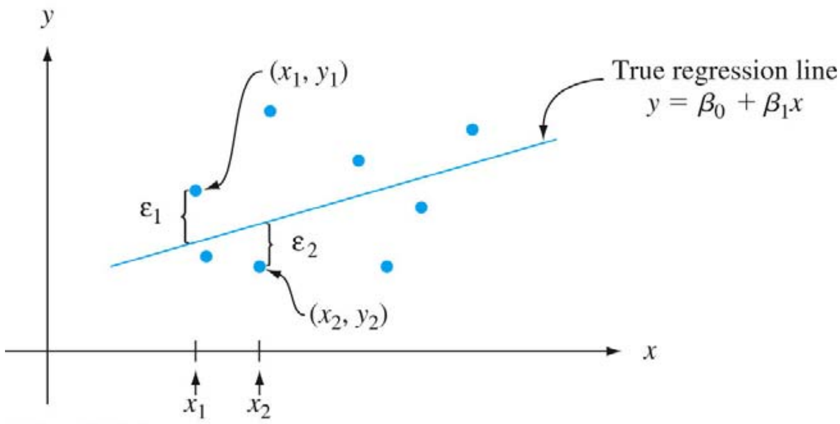
- Simple Linear Regression Model
  - Three parameters (β_0, β_1, and σ^2)
  - Model equation

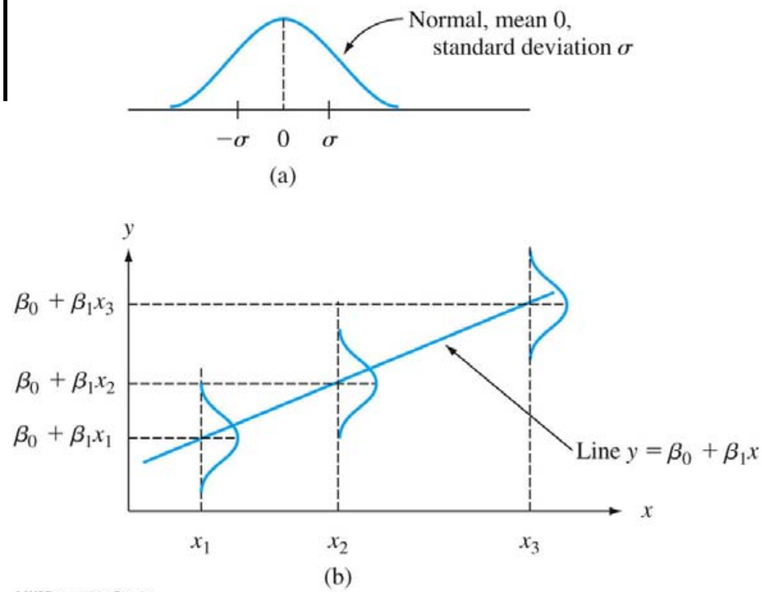
$$y = \beta_0 + \beta_1 x + \varepsilon$$

  - ε is a random variable assumed to be normally distributed with

$$E(\varepsilon) = 0 \quad \text{and} \quad V(\varepsilon) = \sigma^2$$

Simple Linear Regression





Example 12.3

- Suppose the relationship between applied stress x and time-to-failure y is described by the simple linear regression model with true regression line $y = 65 - 1.2x$ and $\sigma = 8$
- Then for any fixed value x of stress, time-to-failure has a normal distribution with mean value $65 - 1.2x$ and standard deviation 8
- In the population consisting of all (x, y) points, the magnitude of a typical deviation from the true regression line is about 8

Example 12.3 Cont'd

- For $x = 20$, Y has mean value

$$\mu_{Y \cdot 20} = 65 - 1.2(20) = 41$$

- $P(Y > 50 \text{ when } x = 20)$

$= P(z > \frac{50-41}{8})$

$= 1 - \Phi(1.13)$

$= 1 - 0.8708$

$= 0.1292$

Example 12.3, Cont'd

- For x = 25, Y has mean value

  $\mu_{Y \cdot 25}$ = 65 – 1.2(25)
  = 35

- $P(Y > 50$ when $x = 25)$

  $= P(z > \frac{50-35}{8})$

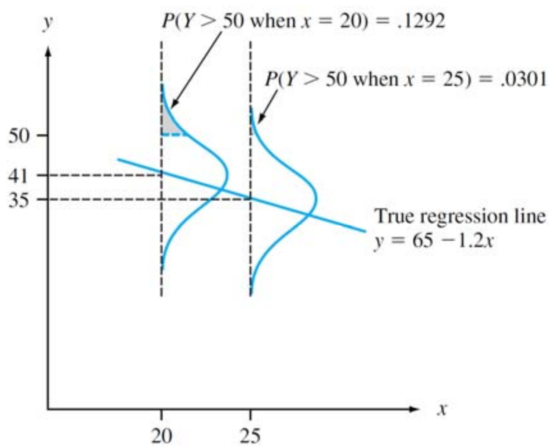
  $= 1 - \Phi(1.88)$

  = 1- 0.9699

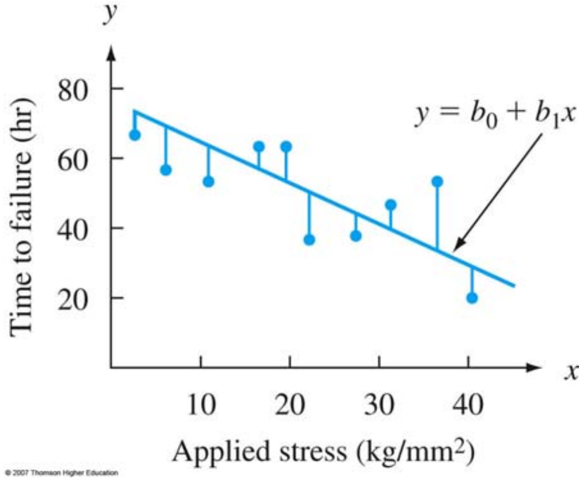
  = 0.0301

Example 12.3, Cont'd

- These probabilities are the shaded areas



Deviations of Observed Data



Least Squares Linear Regression

- Minimize deviation (residual)

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

$$\varepsilon = y - (\beta_0 + \beta_1 x)$$

- Sum of the squares of the deviation

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x)]^2$$

Least Squares Linear Regression

- To determine the minimizing values of β_0 and β_1, determine where the partial derivatives are equal to 0

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = \sum 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

Least Squares Linear Regression

- Simplifying the equations,

$$n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i$$

$$\left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 = \sum x_i y_i$$

Least Squares Estimates

- Slope Coefficient

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

where

$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n$$

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2/n$$

- Intercept

$$\widehat{\beta_0} = \frac{\sum y_i - \beta_1 \sum x_i}{n} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Example 12.4

- Find the linear regression of the following data.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

$\Sigma x_i = 1307.5$ $\Sigma y_i = 779.2$

$\Sigma x_i y_i = 71{,}347.30$

$\Sigma x_i^2 = 128{,}913.93$ $\Sigma y_i^2 = 43{,}745.22$

$\bar{x} = 93.392857$ $\bar{y} = 55.657143$

Solution

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum {x_i}^2 - (\sum x_i)^{2}/n}$$

$$= \frac{71347.30 - \dfrac{1307.5 \times 779.2}{14}}{128913.93 - \dfrac{1307.5^2}{14}}$$

$$= -.20938742$$

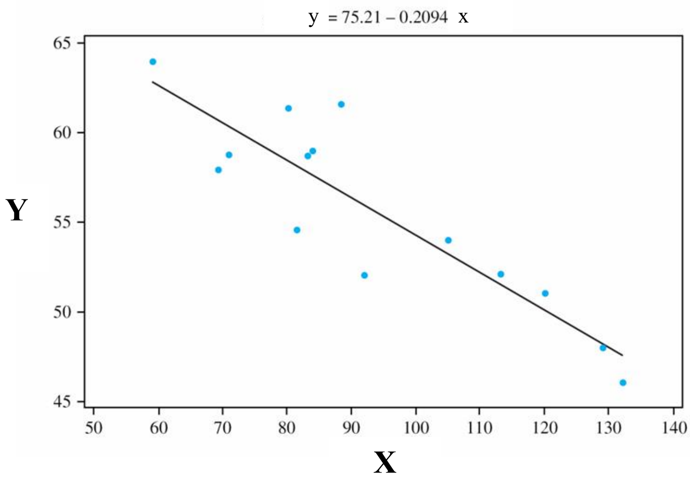
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

$$= 55.657143 - (-.20938742)(93.392857)$$

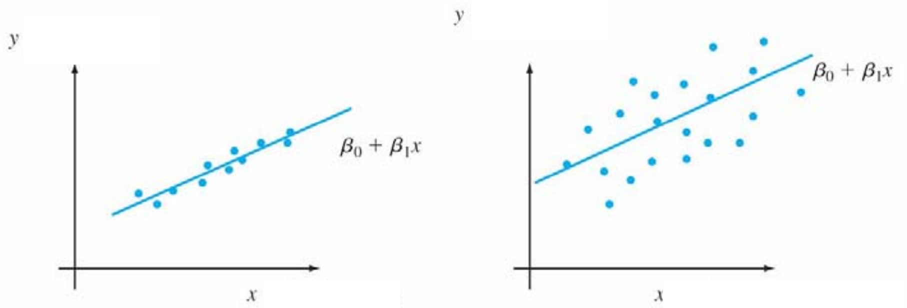
$$= 75.212432$$

Therefore, $y = 75.212 - .2094x$

Solution



Regression Comparison



Residuals

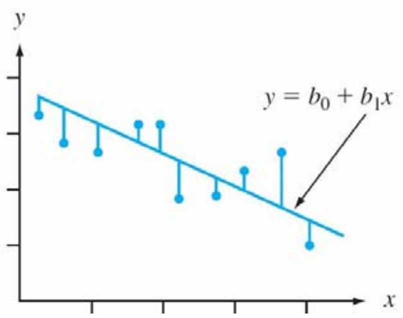
- The **fitted** (or **predicted**) **values** $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by successively substituting $x_1, \ldots, x_n$ into the equation of the estimated regression line:

$$\widehat{y_1} = \beta_0 + \beta_1 x_1, \ldots \widehat{y_n} = \beta_0 + \beta_1 x_n$$

- The **residuals** are the differences between the observed and fitted y values:

$$y_1 - \widehat{y_1}, \ldots y_n - \widehat{y_n}$$

Error Sum of Squares



$$SSE = \sum (y_i - \hat{y})^2$$

$$= \sum [y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)]^2$$

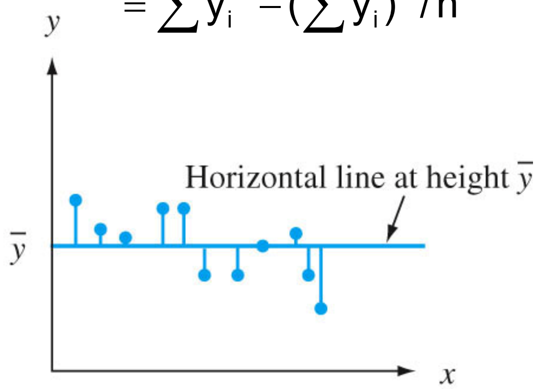
$$= \sum {y_i}^2 - \widehat{\beta_0} \sum y_i - \widehat{\beta_1} \sum x_i y_i$$

(Computational formula)

Total Sum of Squares

$$SST = S_{yy} = \sum (y_i - \overline{y})^2 = E[Y^2] - E[Y]^2$$

$$= \sum y_i^2 - (\sum y_i)^2 / n$$



The Coefficient of Determination

- r^2: the proportion of observed y variation that can be explained by the simple linear regression model

$$r^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\sum y_i^2 - \widehat{\beta_0}\sum y_i - \widehat{\beta_1}\sum x_i y_i}{\sum y_i^2 - \left(\sum y_i\right)^2/n}$$

$0 < r^2 < 1$

Example 12.4, revist

- Find the linear regression of the following data.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

$\Sigma x_i = 1307.5$ $\Sigma y_i = 779.2$

$\Sigma x_i y_i = 71{,}347.30$

$\Sigma x_i^2 = 128{,}913.93$ $\Sigma y_i^2 = 43{,}745.22$

$\bar{x} = 93.392857$ $\bar{y} = 55.657143$

- What is coefficient of determination?

Solution

$$r^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\sum y_i{}^2 - \widehat{\beta_0} \sum y_i - \widehat{\beta_1} \sum x_i y_i}{\sum y_i{}^2 - \left(\sum y_i\right)^2 / n}$$

$$= 1 - \frac{43745.22 - 75.21243 \times 779.2 - (-0.20939 \times 71347.3)}{43745.22 - \frac{(779.2)^2}{14}}$$

$$= 0.7902$$

Regression with Transformed Variables

- Useful intrinsically linear functions

Function	Transformation(s) to Linearize	Linear Form
a. Exponential: $y = \alpha e^{\beta x}$	$y' = \ln(y)$	$y' = \ln(\alpha) + \beta x$
b. Power: $y = \alpha x^{\beta}$	$y' = \log(y),\ x' = \log(x)$	$y' = \log(\alpha) + \beta x'$
c. $y = \alpha + \beta \cdot \log(x)$	$x' = \log(x)$	$y = \alpha + \beta x'$
d. Reciprocal: $y = \alpha + \beta \cdot \frac{1}{x}$	$x' = \frac{1}{x}$	$y = \alpha + \beta x'$