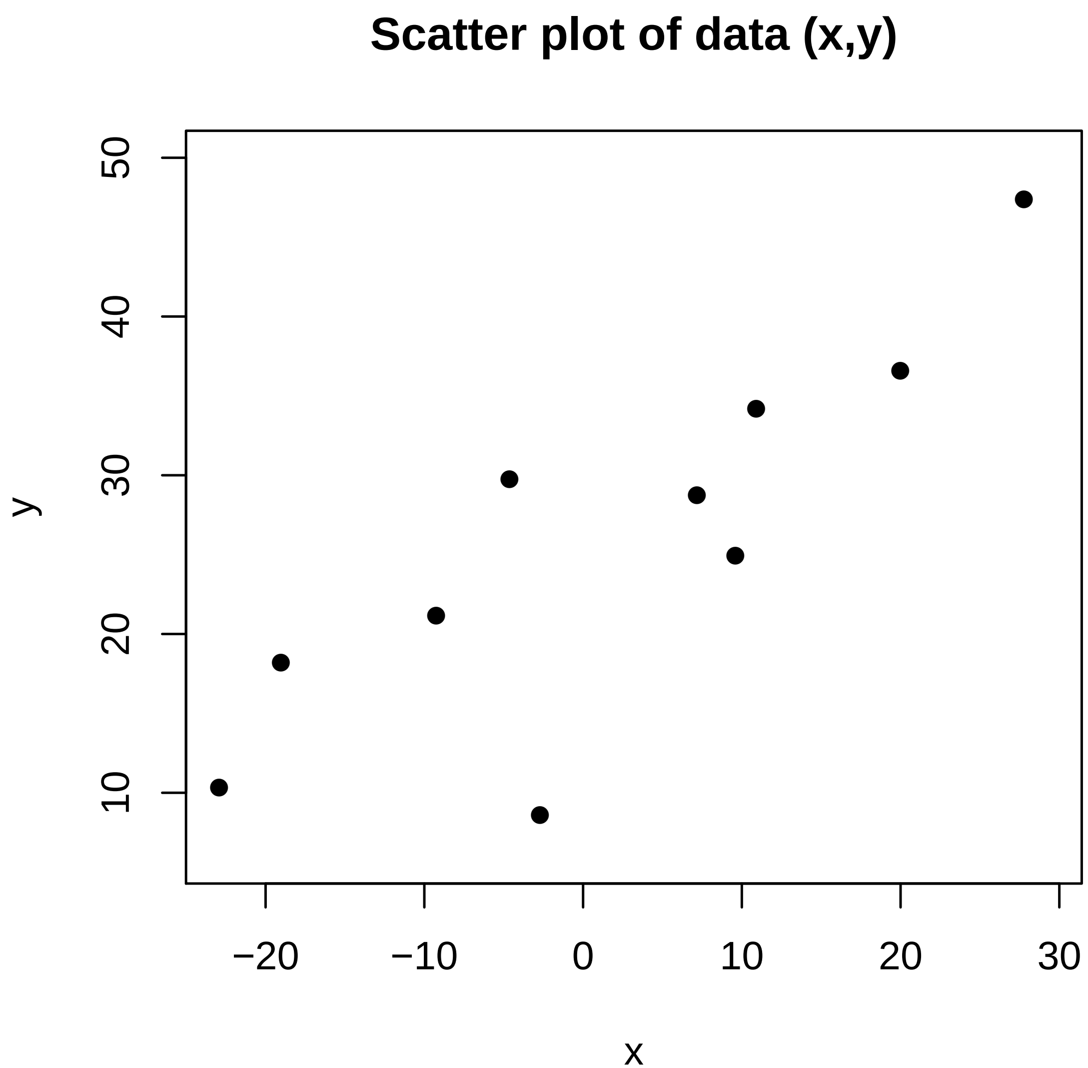
11. Regression and Least Squares

Prof. Tesler

Math 186
Winter 2019

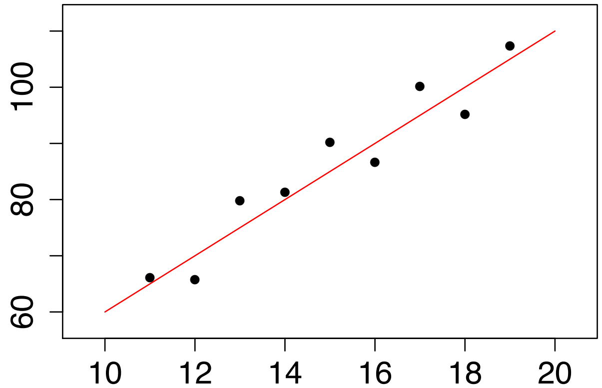
Regression

Given n points $(x_1, y_1), (x_2, y_2), \ldots,$ we want to determine a function $y = f(x)$ that is close to them.

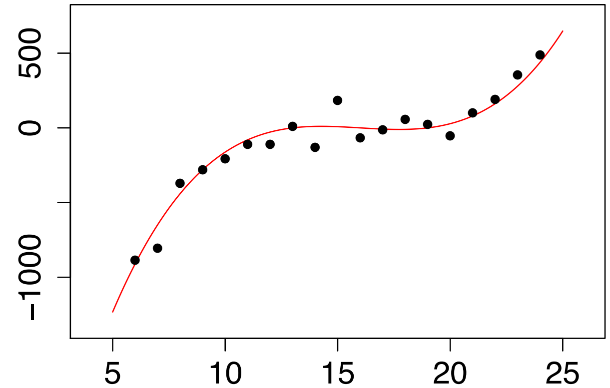


Based on knowledge of the underlying problem or on plotting the data, you have an idea of the general form of the function, such as:

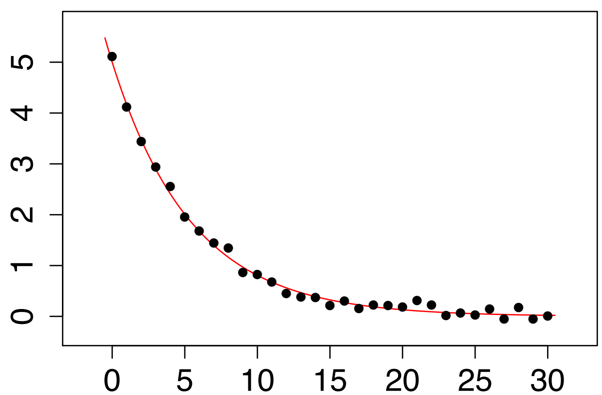
Line $y = \beta_0 + \beta_1 x$



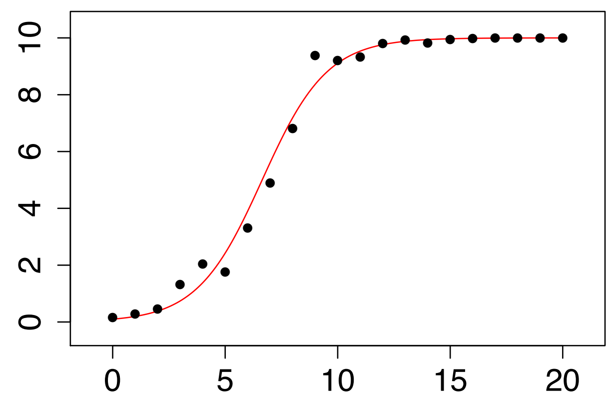
Polynomial $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$



Exponential Decay $y = Ae^{-Bx}$



Logistic Curve $y = A/(1 + B/C^x)$



Goal: Compute the parameters ($\beta_0, \beta_1, \ldots$ or $A, B, C, \ldots$) that give a "best fit" to the data.

Regression

- The methods we consider require the *parameters* to occur linearly. It is fine if (x, y) do not occur linearly.
  E.g., plugging $(x, y) = (2, 3)$ into $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
  gives $3 = \beta_0 + 2\beta_1 + 4\beta_2 + 8\beta_3$.

- For exponential decay, $y = Ae^{-Bx}$, parameter B does not occur linearly. Transform the equation to:

$$\ln y = \ln(A) - Bx = A' - Bx$$

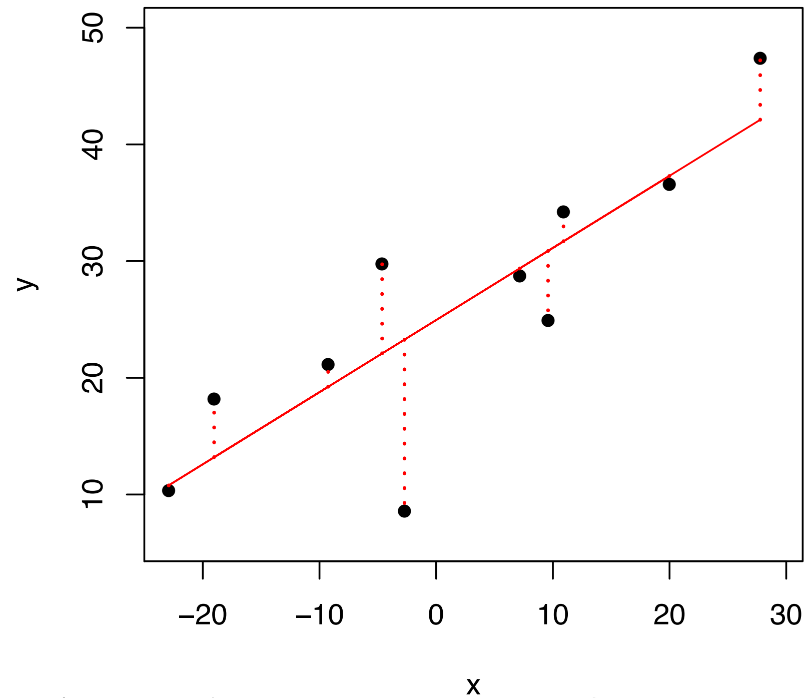
  When we plug in (x, y) values, the parameters A', B occur linearly.

- Transform the logistic curve $y = A/(1 + B/C^x)$ to:

$$\ln\left(\frac{A}{y} - 1\right) = \ln(B) - x\ln(C) = B' + C'x$$

  where A is determined from $A = \lim_{x\to\infty} y(x)$. Now B', C' occur linearly.

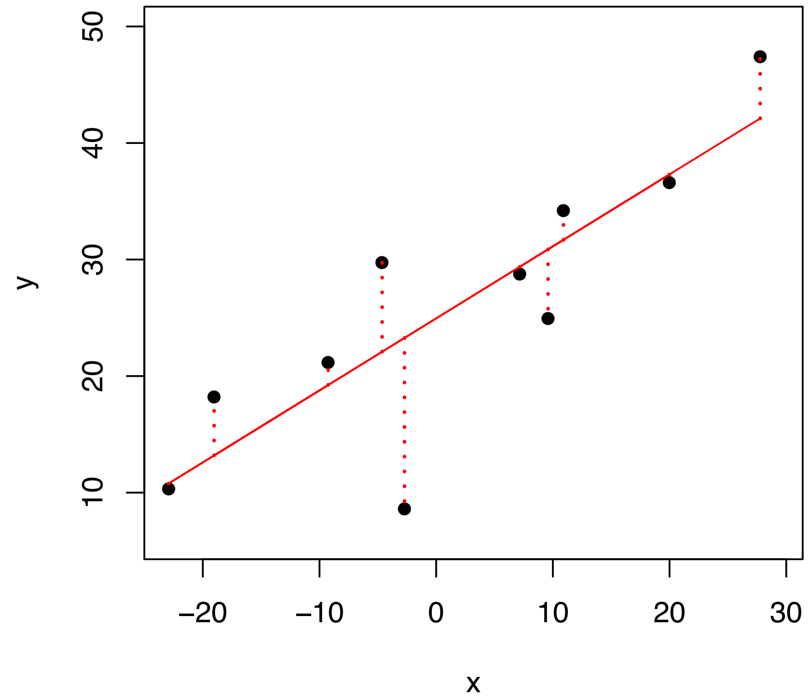
Least squares fit to a line



Given n points $(x_1, y_1), (x_2, y_2), \ldots,$ we will fit them to a line $\hat{y} = \beta_0 + \beta_1 x$:

- *Independent variable: x.* We assume the x's are known exactly or have negligible measurement errors.
- *Dependent variable: y.* We assume the y's depend on the x's but fluctuate due to a random process.
- We do not have $y = f(x)$, but instead, $y = f(x) + \text{error}$.

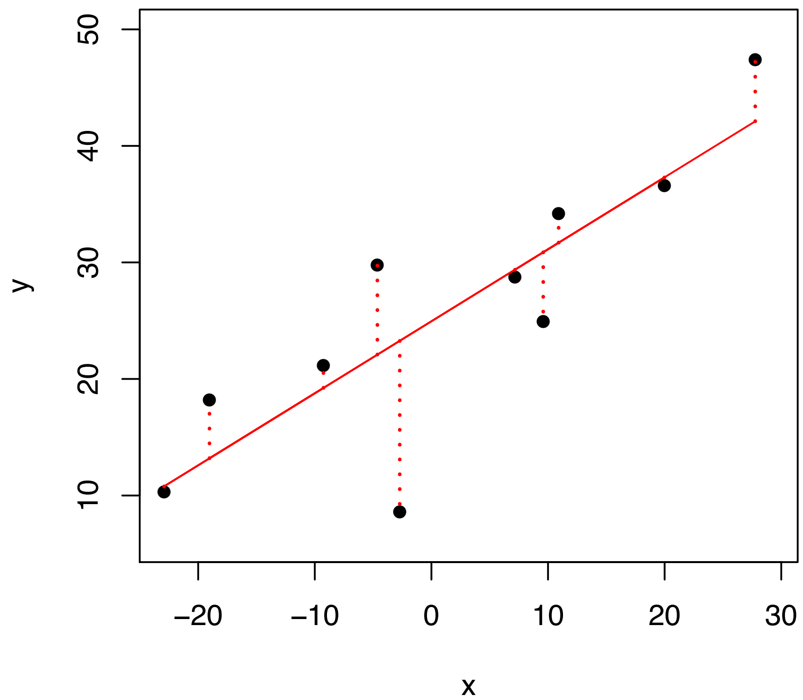
Least squares fit to a line



Given n points $(x_1, y_1), (x_2, y_2), \ldots,$ we will fit them to a line $\hat{y} = \beta_0 + \beta_1 x$:

Predicted y value (on the line):	$\hat{y}_i = \beta_0 + \beta_1 x_i$
Actual data (•):	$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
Residual (actual y minus prediction):	$\epsilon_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$

Least squares fit to a line



We will use the *least squares method*: pick parameters β_0, β_1 that minimize the sum of squares of the residuals.

$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Least squares fit to a line

$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

To find β_0, β_1 that minimize this, solve $\nabla L = \left(\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right) = (0,0)$:

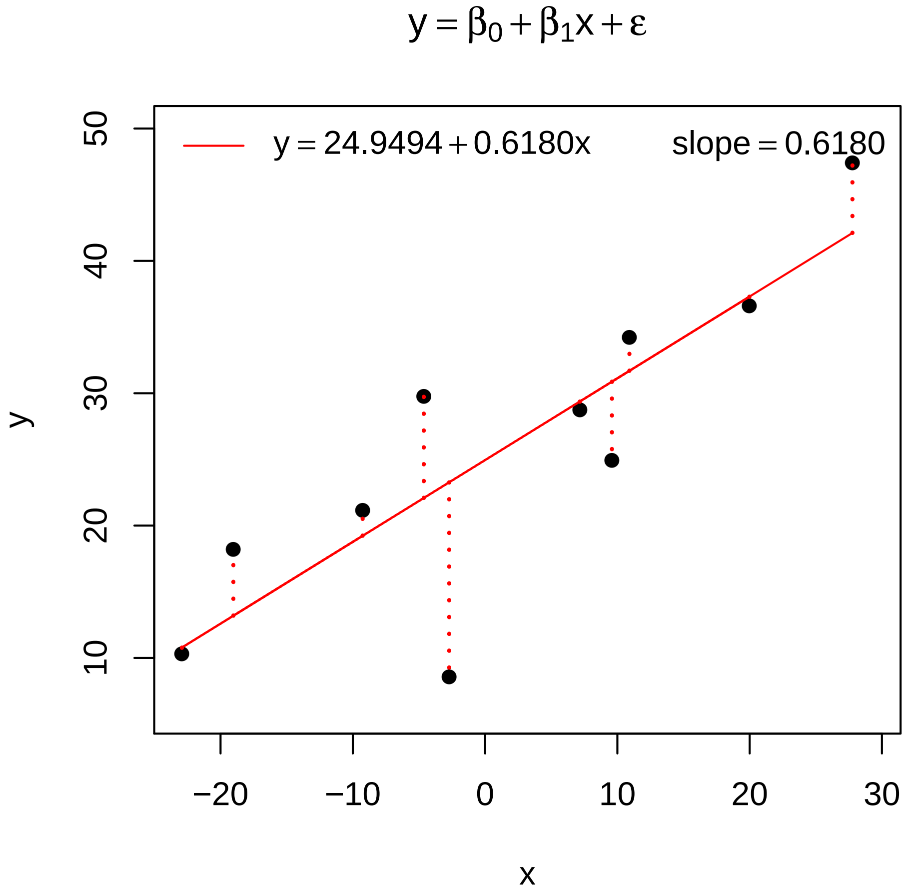
$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) = 0 \quad \Rightarrow \quad n\beta_0 + \left(\sum_{i=1}^{n} x_i\right) \beta_1 = \sum_{i=1}^{n} y_i$$

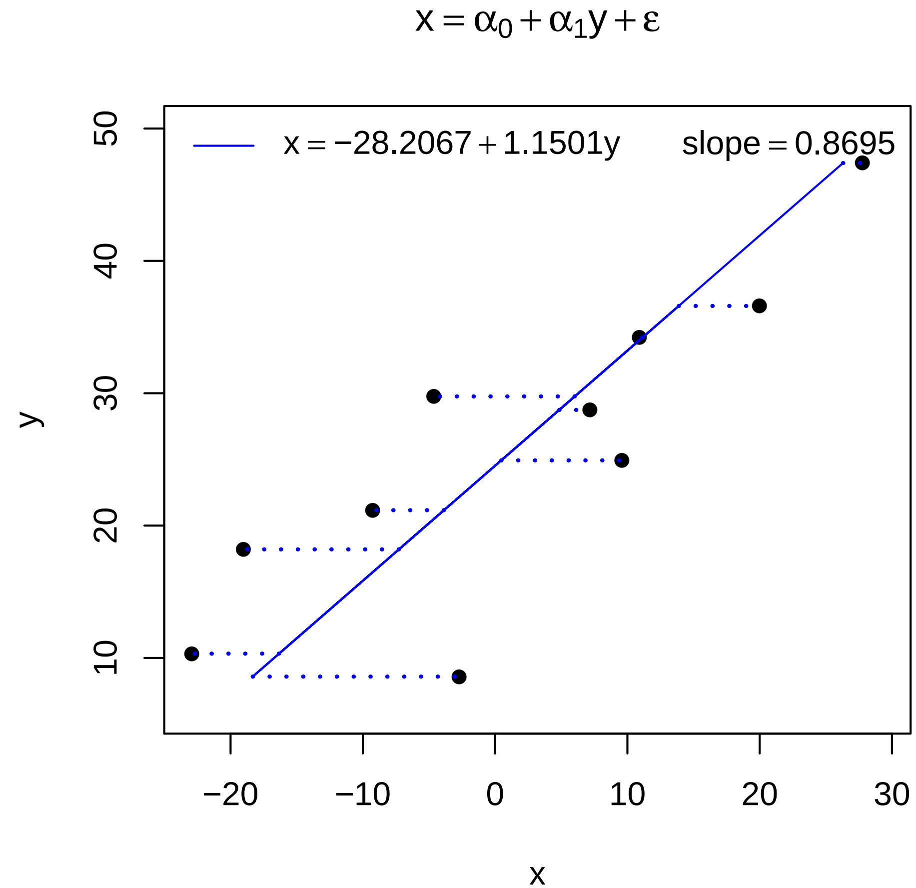
$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) x_i = 0 \quad \Rightarrow \quad \left(\sum_{i=1}^{n} x_i\right) \beta_0 + \left(\sum_{i=1}^{n} {x_i}^2\right) \beta_1 = \sum_{i=1}^{n} x_i y_i$$

which has solution (all sums are $i = 1$ to n)

$$\beta_1 = \frac{n\left(\sum_i x_i\, y_i\right) - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n\left(\sum_i {x_i}^2\right) - \left(\sum_i x_i\right)^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

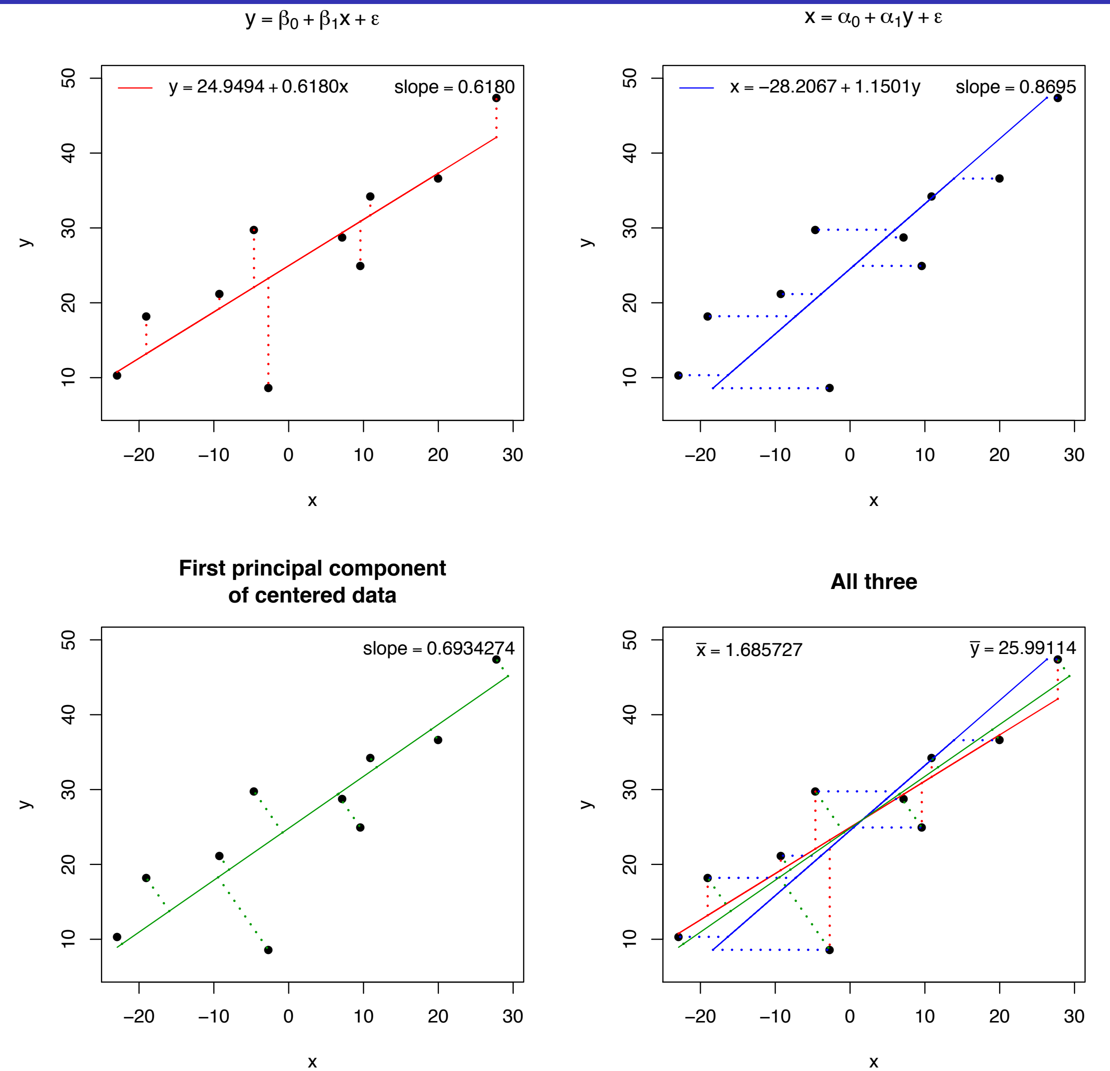
Not shown: use 2nd derivatives to confirm it's a minimum rather than a maximum or saddle point.





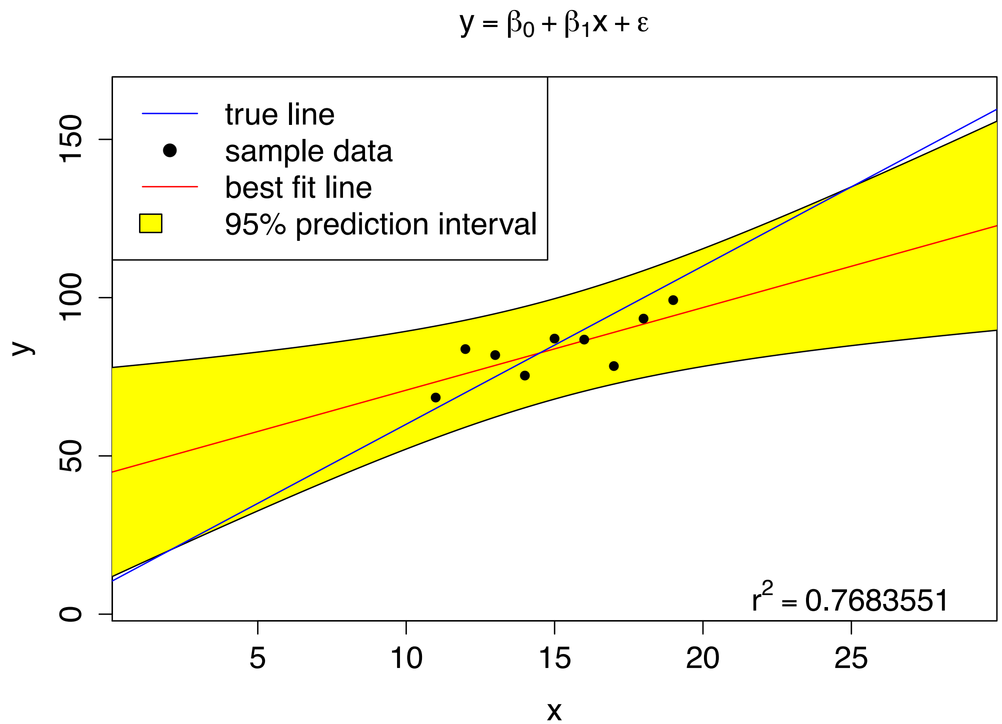
- The best fit for $y = \beta_0 + \beta_1 x +$ error or $x = \alpha_0 + \alpha_1 y +$ error give different lines!
- $y = \beta_0 + \beta_1 x +$ error assumes the x's are known exactly with no errors, while the y's have errors.
- $x = \alpha_0 + \alpha_1 y +$ error is the other way around.

Total Least Squares / Principal Components Analysis



- In many experiments, both x and y have measurement errors.
- Use *Total Least Squares* or *Principal Components Analysis*, in which the residuals are measured perpendicular to the line.
- Details require advanced linear algebra, beyond Math 18.

Confidence intervals



- The best fit line — is different than the true line —.
- We found point estimates of β_0 and β_1.
- Assuming errors are independent of x and normally distributed gives
  - Confidence intervals for β_0, β_1.
  - A *prediction interval* to extrapolate $y = f(x)$ at other x's.
    Warning: it may diverge from the true line when we go out too far.
  - **Not shown:** one can also do hypothesis tests on the values of β_0 and β_1, and on whether two samples give the same line.

Confidence intervals

- The method of least squares gave point estimates of β_0 and β_1:

$$\hat{\beta}_1 = \frac{n\sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n\left(\sum_i {x_i}^2\right) - \left(\sum_i x_i\right)^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- The sample variance of the residuals is

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 \quad \text{(with } df = n-2\text{).}$$

- $100(1-\alpha)\%$ confidence intervals:

$$\beta_0 : \quad \left(\hat{\beta}_0 - t_{\alpha/2,n-2}\frac{s\sqrt{\sum_i {x_i}^2}}{\sqrt{n\sum_i (x_i - \bar{x})}},\ \hat{\beta}_0 + t_{\alpha/2,n-2}\frac{s\sqrt{\sum_i {x_i}^2}}{\sqrt{n\sum_i (x_i - \bar{x})}}\right)$$

$$\beta_1 : \quad \left(\hat{\beta}_1 - t_{\alpha/2,n-2}\frac{s}{\sqrt{\sum_i (x_i - \bar{x})}},\ \hat{\beta}_1 + t_{\alpha/2,n-2}\frac{s}{\sqrt{\sum_i (x_i - \bar{x})}}\right)$$

$$y \text{ at new } x : \quad (\hat{y} - w, \hat{y} + w) \ \text{ with } \ \hat{y} = \beta_0 + \beta_1 x$$

$$\text{and } \ w = t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Covariance

- Let X and Y be random variables, possibly dependent.
- Let $\mu_X = E(X)$, $\mu_Y = E(Y)$
- $$\begin{aligned} \mathrm{Var}(X+Y) &= E((X+Y-\mu_X-\mu_Y)^2) = E\left(\left((X-\mu_X)+(Y-\mu_Y)\right)^2\right) \\ &= E\left((X-\mu_X)^2\right) + E\left((Y-\mu_Y)^2\right) + 2E\left((X-\mu_X)(Y-\mu_Y)\right) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) \end{aligned}$$
  where the *covariance* of X and Y is defined as
  $$\mathrm{Cov}(X,Y) = E\left((X-\mu_X)(Y-\mu_Y)\right) = E(XY) - E(X)E(Y)$$
- Independent variables have $E(XY) = E(X)E(Y)$, so $\mathrm{Cov}(X,Y) = 0$. But $\mathrm{Cov}(X,Y) = 0$ does not guarantee X and Y are independent.

Covariance and independence

- Independent variables have $E(XY) = E(X)E(Y)$, so $\text{Cov}(X, Y) = 0$. But $\text{Cov}(X, Y) = 0$ does not guarantee X and Y are independent.
- Consider the standard normal distribution, Z.
- Z and Z^2 are dependent.
- $\text{Cov}(Z, Z^2) = E(Z^3) - E(Z)E(Z^2)$.
- The standard normal distribution has mean 0: $E(Z) = 0$.
- $E(Z^3) = 0$ since Z^3 is an odd function and the pdf of Z is symmetric around $Z = 0$.
- So $\text{Cov}(Z, Z^2) = 0$.

Covariance properties

We have

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

where the *covariance* of X and Y is defined as

$$\mathrm{Cov}(X,Y) = E\left((X-\mu_X)(Y-\mu_Y)\right) = E(XY) - E(X)E(Y)$$

Additional properties of covariance

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$

Sign of covariance

$$\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$$

- ***When*** $\mathbf{Cov}(X,Y)$ ***is positive:***
  There is a tendency to have $X > \mu_X$ when $Y > \mu_Y$ and vice-versa, and $X < \mu_X$ when $Y < \mu_Y$ and vice-versa.
- ***When*** $\mathbf{Cov}(X,Y)$ ***is negative:***
  There is a tendency to have $X > \mu_X$ when $Y < \mu_Y$ and vice-versa, and $X < \mu_X$ when $Y > \mu_Y$ and vice-versa.
- ***When*** $\mathbf{Cov}(X,Y) = \mathbf{0}$***:***
  a) X and Y **might** be independent, but it's not guaranteed.
  b) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Sample variance and covariance

Variance of a random variable:

$$\sigma^2 = \mathrm{Var}(X) = E((X-\mu_X)^2) = E(X^2) - (E(X))^2$$

Sample variance from data $x_1, \ldots, x_n$ to estimate σ^2:

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

Covariance between random variables X, Y:

$$\sigma_{XY} = \mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(XY) - E(X)E(Y)$$

Sample covariance from data $(x_1,y_1), \ldots, (x_n,y_n)$ to estimate σ_{XY}:

$$s_{XY} = \mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\, y_i\right) - \frac{n}{n-1}\bar{x}\,\bar{y}$$

Correlation coefficient

Let X and Y be two random variables.
Their *correlation coefficient* is

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

- This is a normalized version of covariance, and is between ± 1.
- For a line $Y = aX + b$ with a, b constants ($a \neq 0$),

$$\rho(X,Y) = \frac{a\operatorname{Var}(X)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(aX)}} = \frac{a\sigma^2}{\sigma\cdot|a|\sigma} = \frac{a}{|a|} = \pm 1 \text{ (sign of } a\text{)}$$

- $\rho(X,Y) = \pm 1$ iff $Y = aX + b$ with a, b constants ($a \neq 0$).
- Closer to ± 1: more linear. Closer to 0: less linear.
- If X and Y are independent then $\rho(X,Y)=0$.
  The converse is not valid: dependent variables can have $\rho(X,Y)=0$.

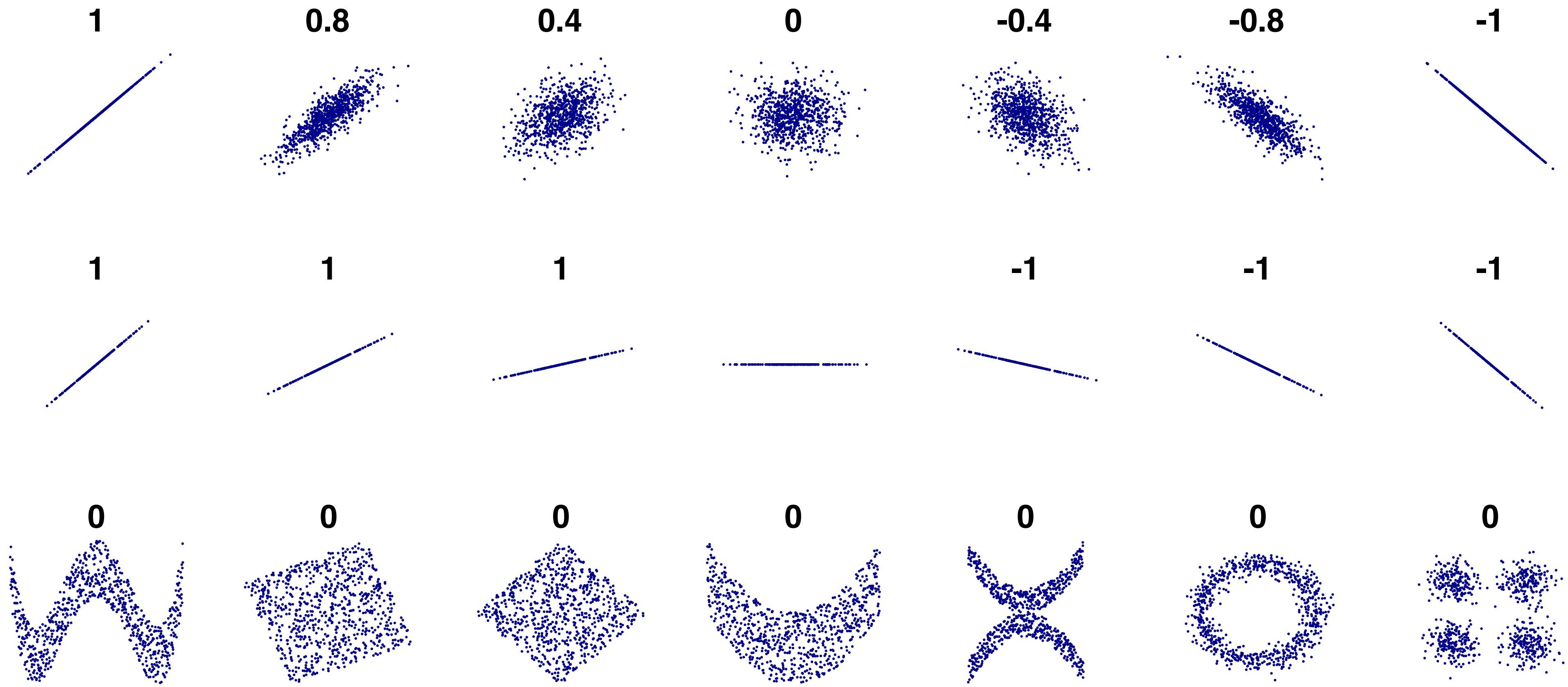
Correlation coefficient

- $\rho(X,Y)$ is estimated from data by the *sample correlation coefficient* (a.k.a. *Pearson product-moment correlation coefficient*):

$$r(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i-\bar{y})^2}}$$
$$= \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{n\sum_i {x_i}^2 - (\sum_i x_i)^2}\sqrt{n\sum_i {y_i}^2 - (\sum_i y_i)^2}}$$

- People often report r^2 (between 0 and 1) instead of r.

Sample correlation coefficient r



http://en.wikipedia.org/wiki/File:Correlation_examples2.svg
http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

- **Middle row:** Perfect linear relation $Y = aX + b$ gives

$r = 1$	for lines with positive slope $(a > 0)$
$r = -1$	for lines with negative slope $(a < 0)$
r undefined	for horizontal line $(Y = b)$

- **Other rows:** coming up!

Interpretation of r^2

- Let $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$
  be the predicted y-value for x_i based on the least squares line.
- Write the deviation of y_i from $\bar{y}$ as

$$\underbrace{y_i - \bar{y}}_{\text{Total deviation}} = \underbrace{(y_i - \hat{y}_i)}_{\text{Unexplained by line}} + \underbrace{(\hat{y}_i - \bar{y})}_{\text{Explained by line}}$$

- It can be shown that the sum of squared deviations for all y's is

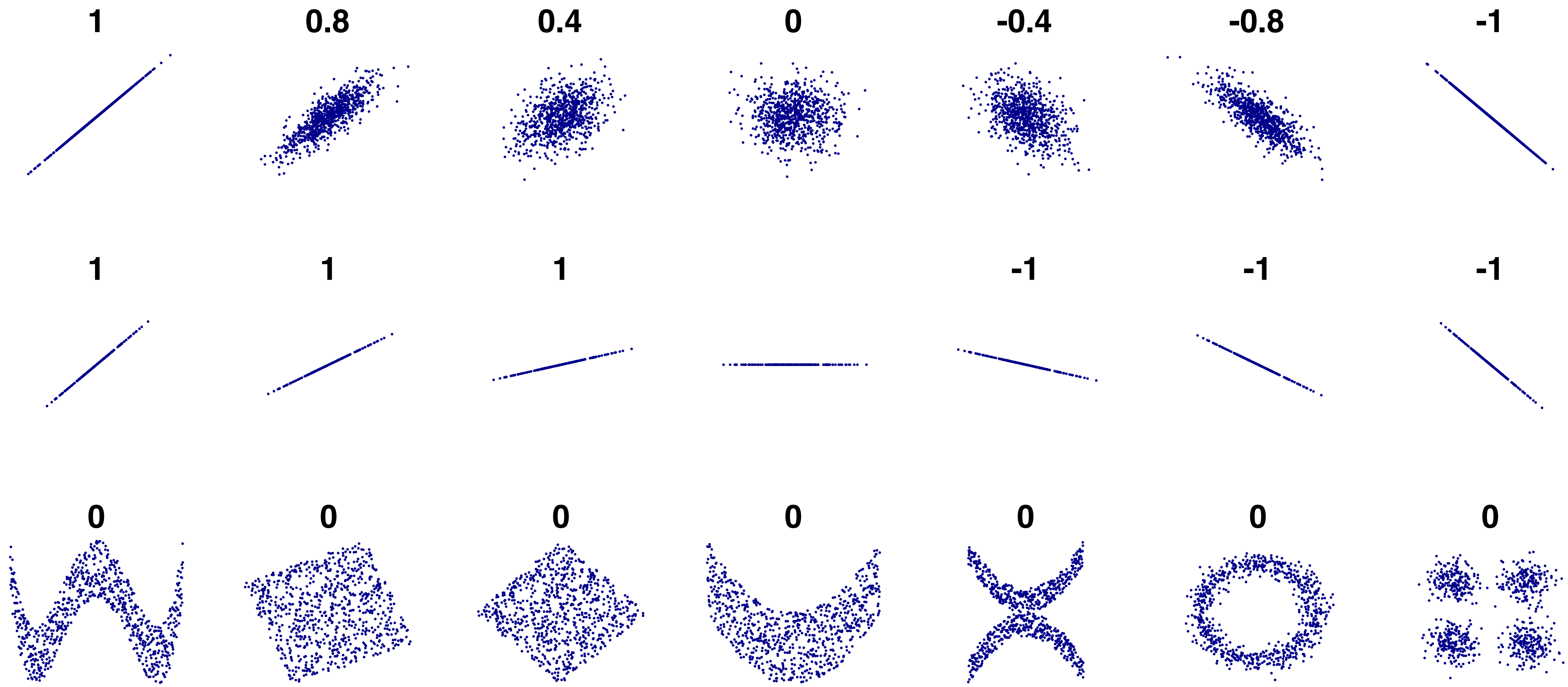
$$\underbrace{\textstyle\sum_i (y_i - \bar{y})^2}_{\text{Total variation}} = \underbrace{\textstyle\sum_i (y_i - \hat{y}_i)^2}_{\text{Unexplained variation}} + \underbrace{\textstyle\sum_i (\hat{y}_i - \bar{y})^2}_{\text{Explained variation}} + \underbrace{2 \textstyle\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\substack{= 0 \text{ by a miracle!} \\ \text{(Tedious algebra not shown)}}}$$

  and that

$$r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\text{Explained variation}}{\text{Total variation}}$$

- $r = 1$: 100% of the variation is explained by the line and 0% is due to other factors, and the slope is positive.
- $r = -.8$: 64% of the variation is explained by the line and 36% is due to other factors, and the slope is negative.

Sample correlation coefficient r



http://en.wikipedia.org/wiki/File:Correlation_examples2.svg
http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

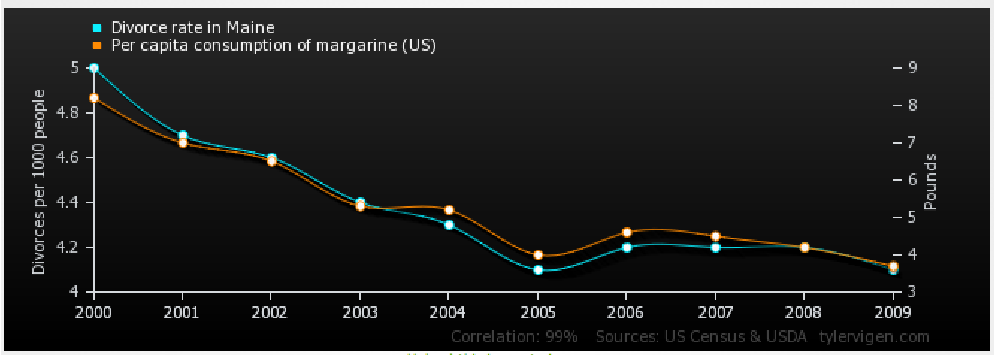
- **Top row:** Linear relations with varying r.
- **Bottom:** $r = 0$, yet X and Y are dependent in all of these (except possibly the last); it's just that the relationship is not a line.

Correlation does not imply causation

- High correlation between X and Y doesn't mean X causes Y or vice-versa. It could be a coincidence. Or they could both be caused by a third variable.
- Website tylervigen.com plots many data sets (various quantities by year) against each other to find spurious correlations:

spurious correlations

Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)



Upload this image to imgur

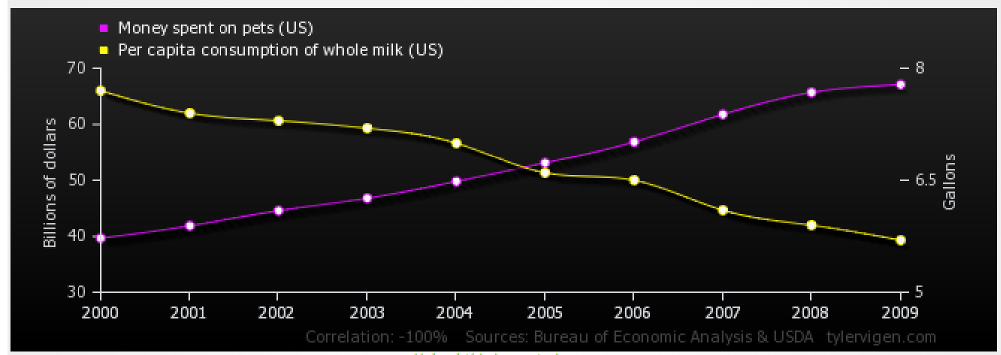
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

http://www.tylervigen.com/view_correlation?id=1703

spurious correlations

Money spent on pets (US)
inversely correlates with
Per capita consumption of whole milk (US)



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Money spent on pets (US) *Billions of dollars (Bureau of Economic Analysis)*	39.7	41.9	44.6	46.8	49.8	53.1	56.9	61.8	65.7	67.1
Per capita consumption of whole milk (US) *Gallons (USDA)*	7.7	7.4	7.3	7.2	7	6.6	6.5	6.1	5.9	5.7

Correlation: -0.995478

http://tylervigen.com/view_correlation?id=1759

More about interpretation of correlation

- Low r^2 does NOT guarantee independence; it just means that a line $y = \beta_0 + \beta_1 x$ is not a good fit to the data.

- r is an estimate of ρ. The estimate improves with higher n. With additional assumptions on the underlying joint distribution of X, Y, we can use r to test

  H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$ (or other values).

- Best fits and correlation generalize to other models, including

Polynomial regression	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$
Multiple linear regression	$y = \beta_0 + \beta_1 t + \beta_2 u + \cdots + \beta_p w$ $t, u, \ldots, w$: multiple independent variables y: dependent variable
Weighted versions	When the variance is different at each value of the independent variables