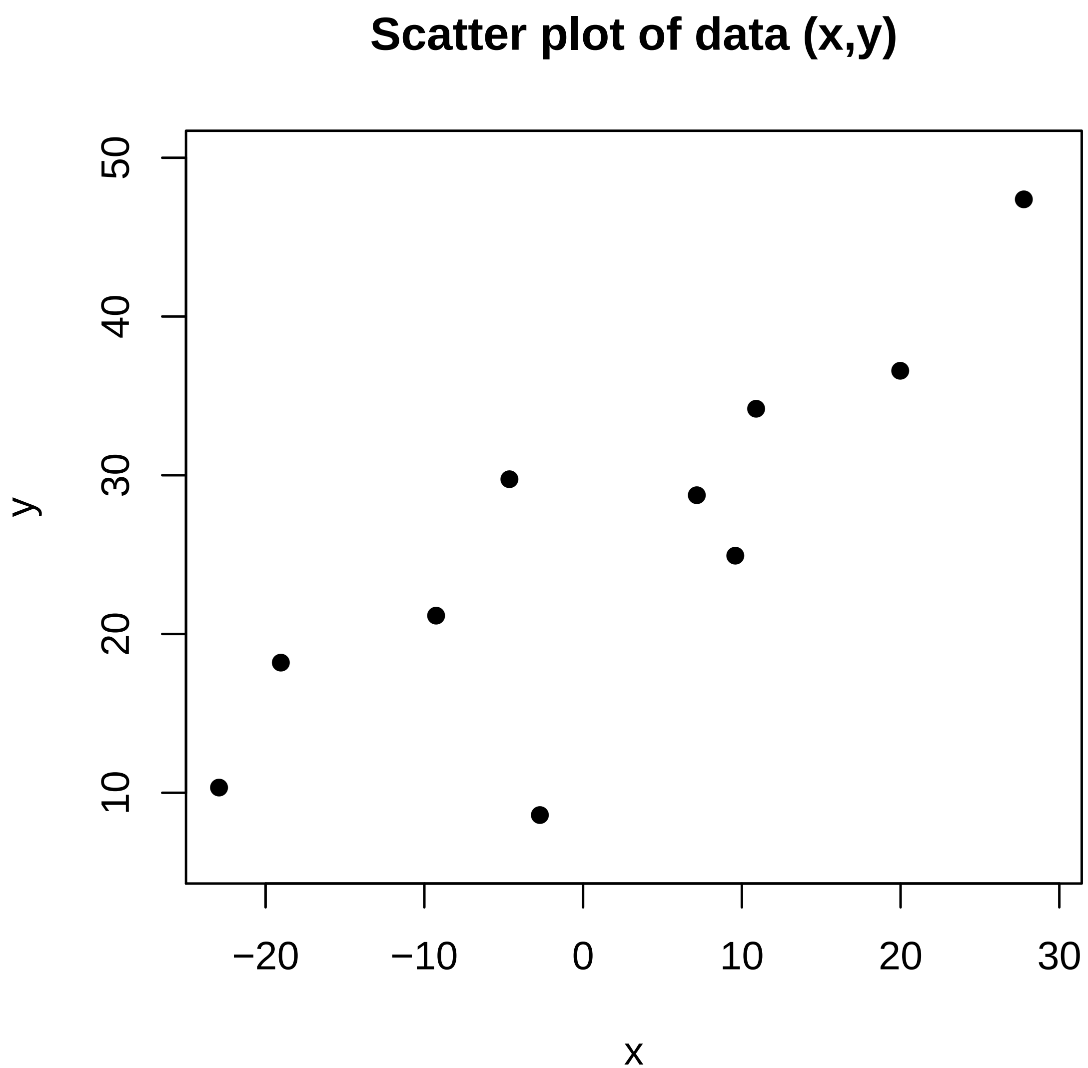
8.4.3 Linear Regression

Prof. Tesler

Math 283
Fall 2019

Regression

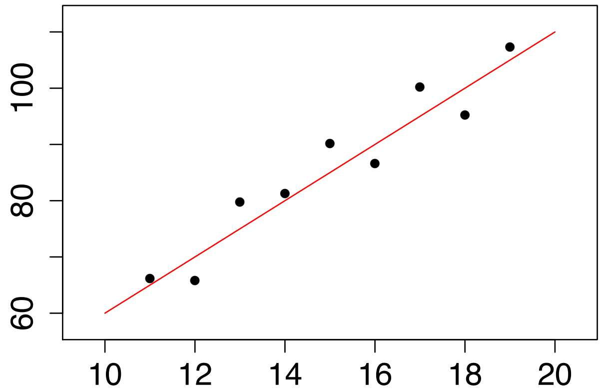
Given n points $(x_1, y_1), (x_2, y_2), \ldots$, we want to determine a function $y = f(x)$ that is close to them.



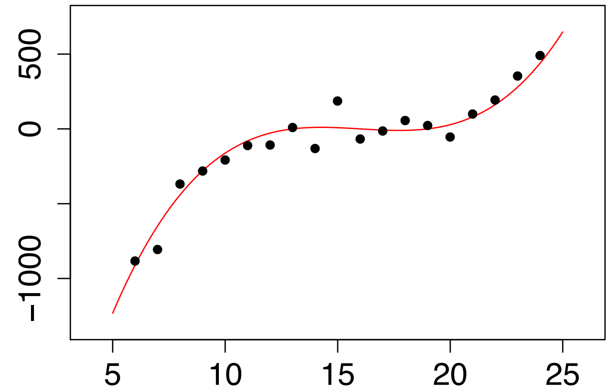
Regression

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:

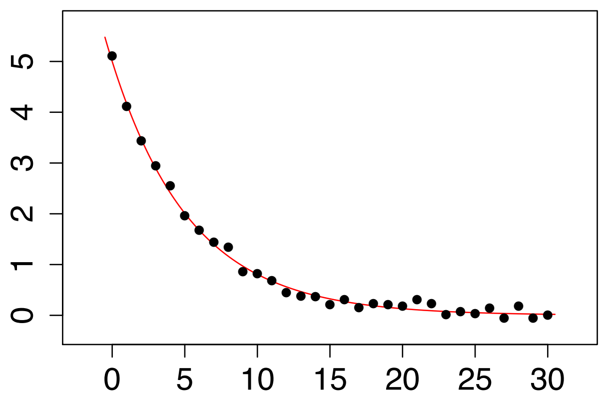
Line $y = \beta_0 + \beta_1 x$



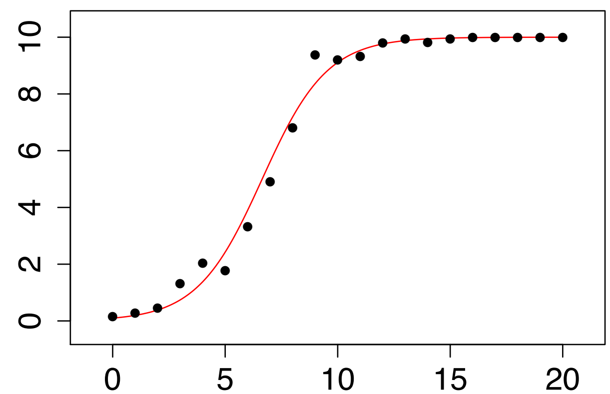
Polynomial $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$



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 in some sense (least squares or MLEs).

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:

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

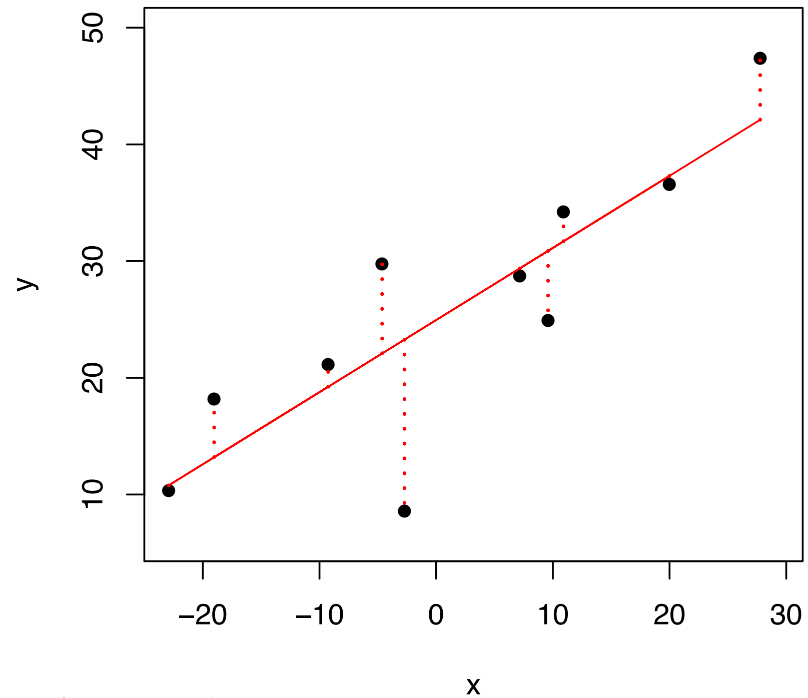
  When we plug in (x,y) values, the parameters A',B occur linearly.

- 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.

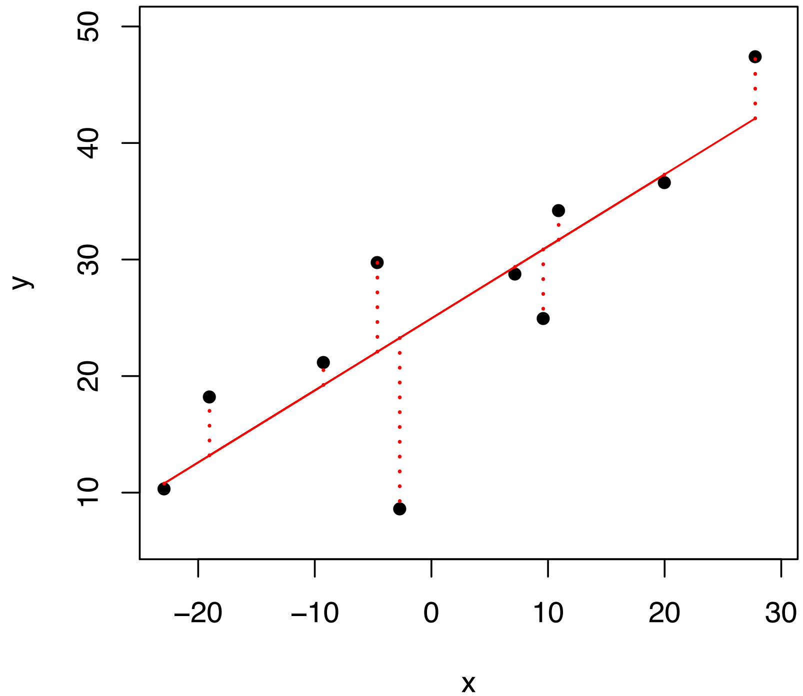
Least squares fit to a line



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, $y = f(x) + \text{error}$.

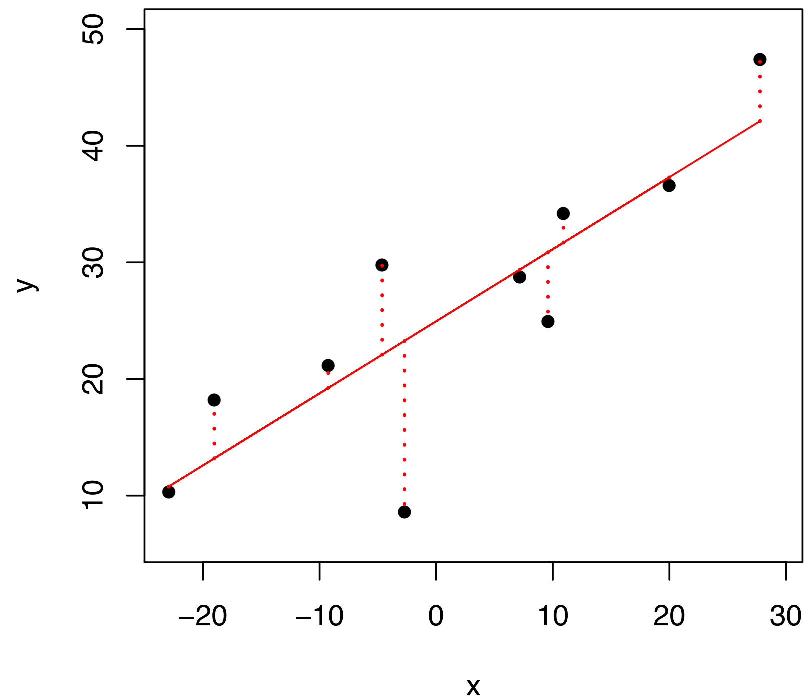
Least squares fit to a line



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 $\nabla L = \left(\frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}\right) = (0,0)$:

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

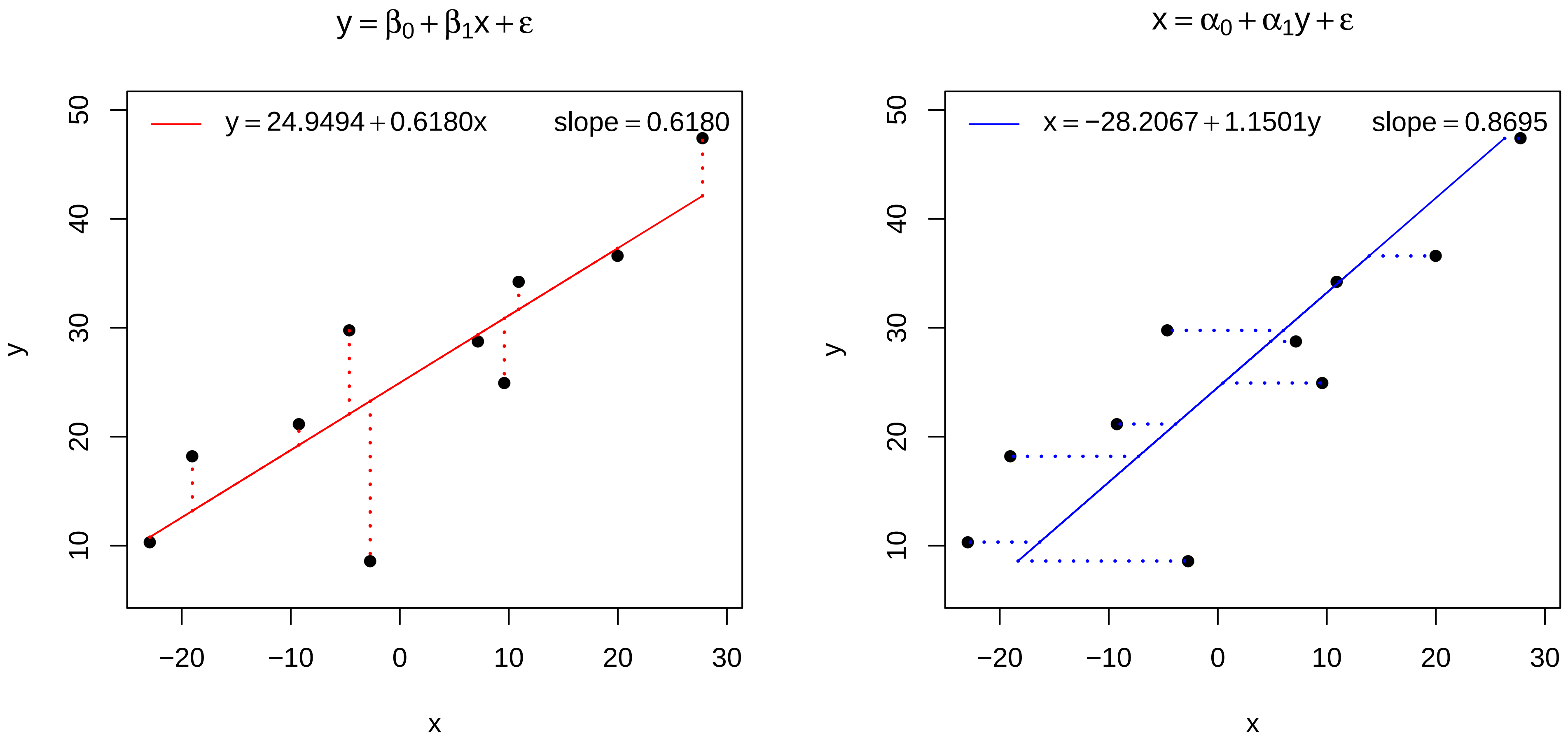
$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) x_i = 0 \qquad \Rightarrow \qquad \left(\sum_{i=1}^{n} x_i\right) \beta_0 + \left(\sum_{i=1}^{n} {x_i}^2\right) \beta_1 = \sum_{i=1}^{n} x_i y_i$$

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

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

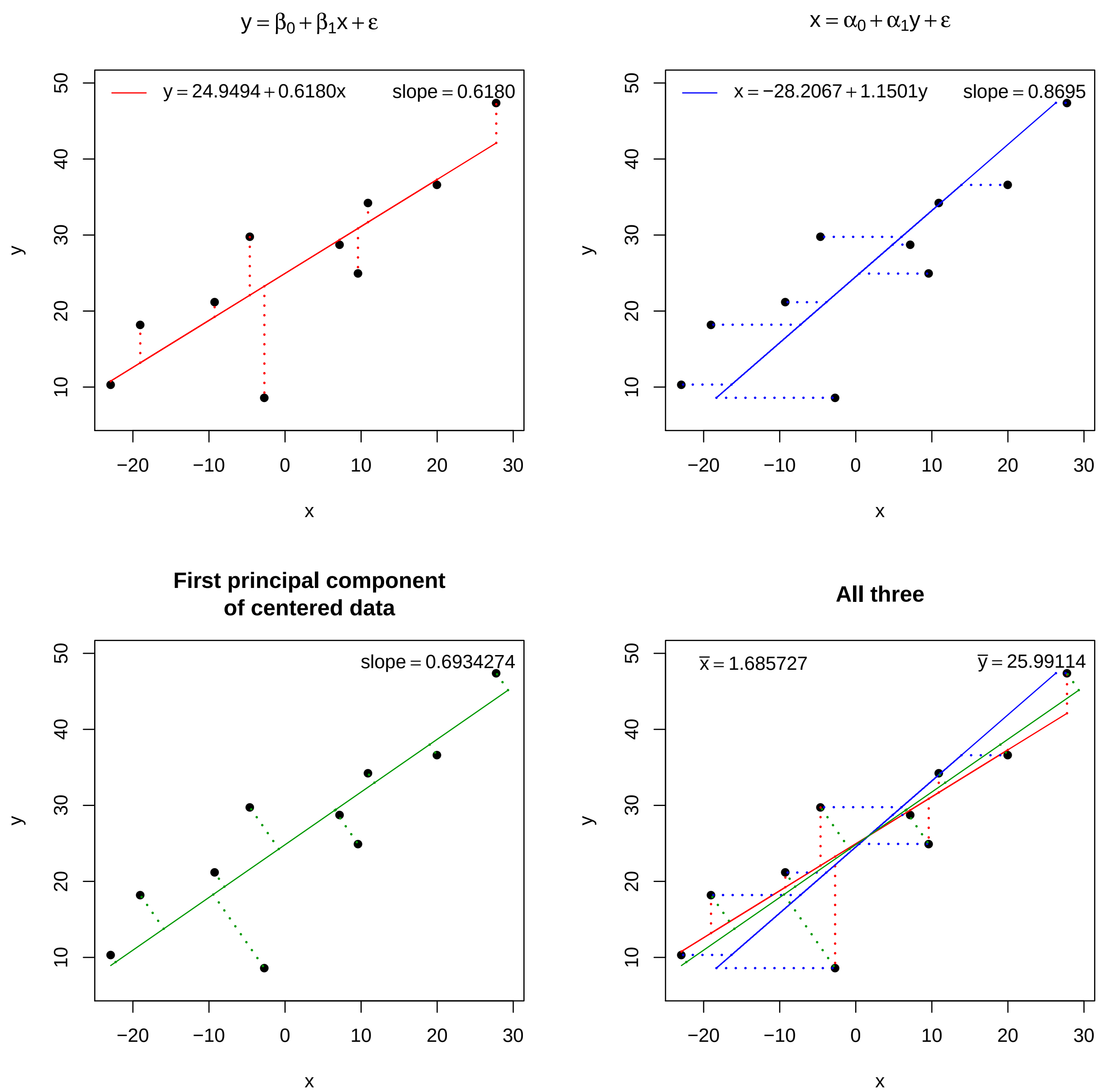
Not shown: use 2nd derivatives to confirm it's a minimum rather than a maximum or saddle point.

Best fitting line



- 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



Least squares vs. PCA

Errors in data:

- **Least squares:** $y = \beta_0 + \beta_1 x +$ error
  assumes x's have no errors while y's have errors.
- **PCA:** assumes all coordinates have errors.

For (x_i, y_i) data, we minimize the sum of ...

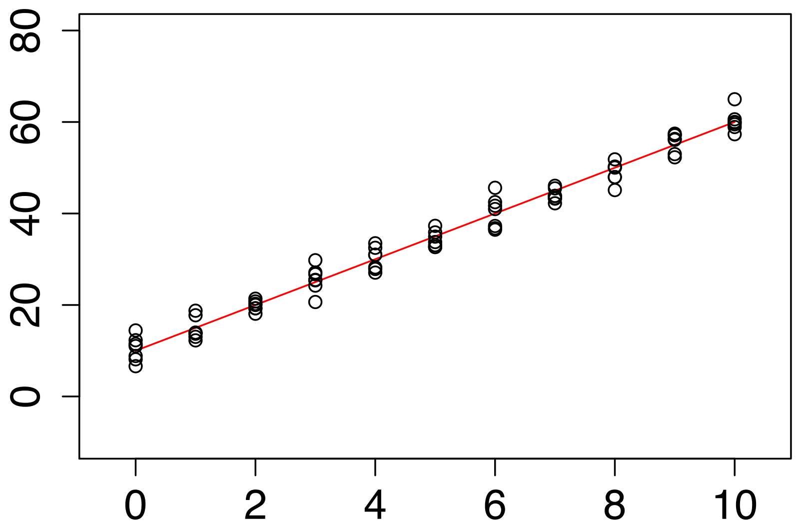
- **Least squares:** squared vertical distances from points to the line.
- **PCA:** squared orthogonal distances from points to the line.
- Due to centering data, the lines all go through $(\bar{x}, \bar{y})$.
- For multivariate data, lines are replaced by planes, etc.

Different units/scaling on inputs (x) and outputs (y):

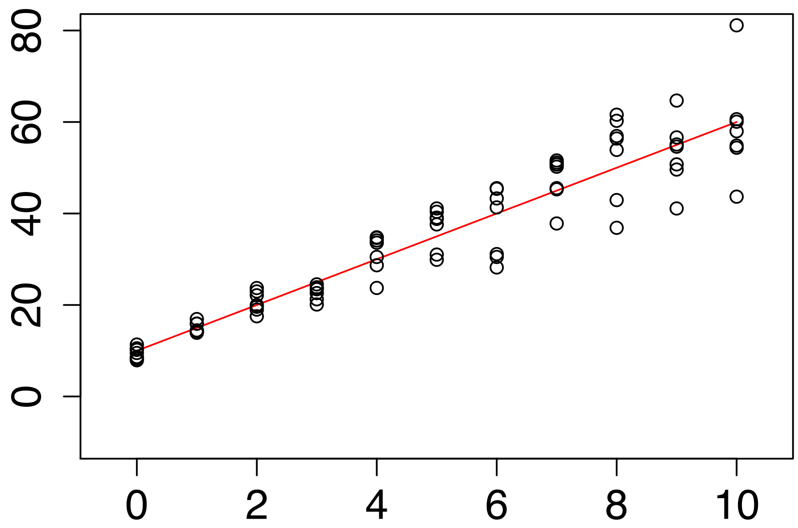
- Least squares gives equivalent solutions if you change units or scaling, while PCA is sensitive to changes in these.
- **Example:** (a) x in seconds, y in cm vs. (b) x in seconds, y in mm give equivalent results for least squares, inequivalent for PCA.
- For PCA, a workaround is to convert coordinates to Z-scores.

Distribution of values at each x

(a) Homoscedastic



(b) Heteroscedastic



- On repeated trials, at each x we get a distribution of values of y rather than a single value.
- In (a), the error term is a normal distribution with the same variance for every x. This is the case we will study. Assume the errors are independent of x and have a normal distribution with mean 0, SD σ.
- In (b), the variance changes for different values of x.
  Use a generalization called *Weighted Least Squares*.

Maximum Likelihood Estimate for best fitting line

- The method of least squares uses a geometrical perspective.
- Now we'll assume the data has certain statistical properties.
- Simple linear model:

$$Y = \beta_0 + \beta_1 x + \mathcal{E}$$

  Assume the x's are known (so lowercase) and $\mathcal{E}$ is Gaussian with mean 0 and standard deviation σ, making $\mathcal{E}, Y$ random variables.
- At each x, there is a distribution of possible y's, giving a *conditional distribution*: $f_{Y|X=x}(y)$.
- Assume conditional distributions for different x's are independent.
- The means of these conditional distributions form a line

$$y = E(Y|X = x) = \beta_0 + \beta_1 x.$$

- Denote the MLE values by $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ to distinguish them from the true (hidden) values.

Maximum Likelihood Estimate for best fitting line

Given data $(x_1, y_1), \ldots, (x_n, y_n)$, we have

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

where

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

has a normal distribution with mean 0 and standard deviation σ.
The likelihood of the data is the product of the pdf of the normal distribution at ϵ_i over all i:

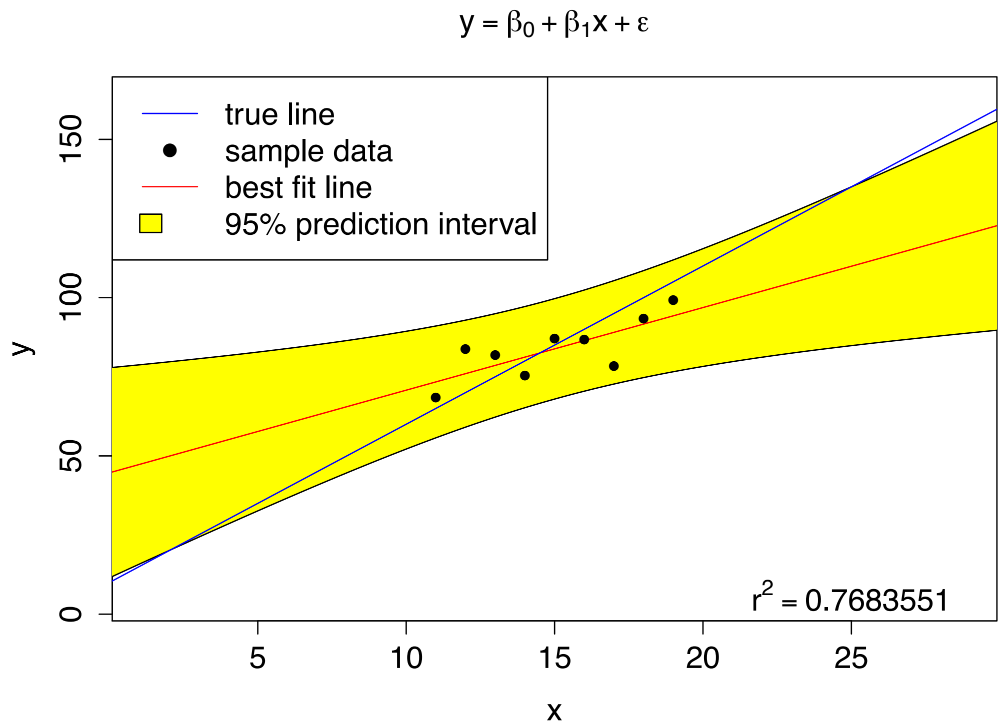
$$L = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2} \right)$$

Finding β_0, β_1 that maximize L (or $\log L$) is equivalent to minimizing

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

so we get the same answer as using least squares!

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

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

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

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

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

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

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

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 $\rho(X,Y){=}0$.

Sample correlation coefficient r

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

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

- People often report r^2 (between 0 and 1) instead of r.
- The slopes of the least squares lines are

$$y = \beta_1 x + \beta_0 + \epsilon \qquad\qquad x = \alpha_1 y + \alpha_0 + \epsilon'$$

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

(slope in normal orientation is $1/\hat{\alpha}_1$)

so $r = \pm\sqrt{\hat{\alpha}_1\hat{\beta}_1} = \pm\sqrt{\hat{\beta}_1/(1/\hat{\alpha}_1)}$ (with same $\pm$ sign as slopes) is the square root of the ratio of the slopes of the lines.

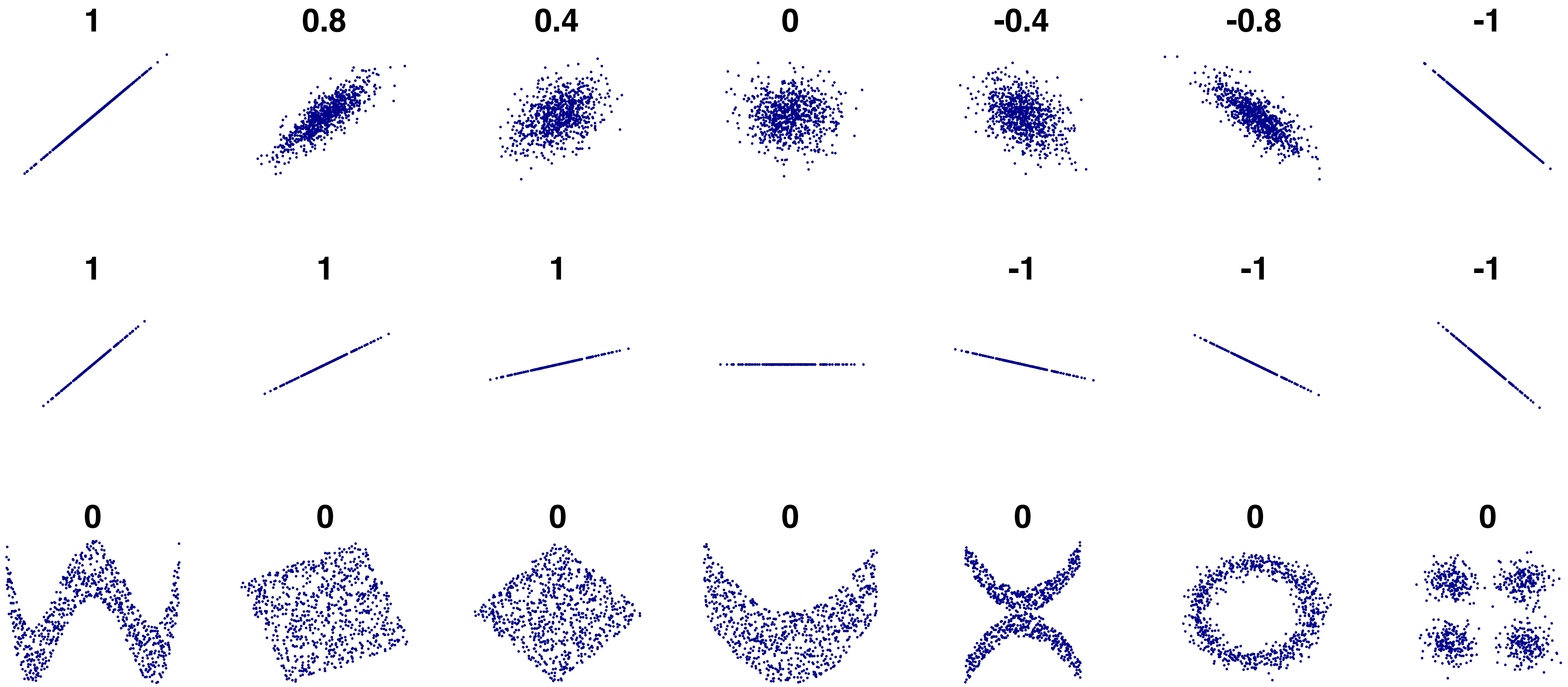
- An aside: $\hat{\beta}_1 = \operatorname{cov}(x,y)/\operatorname{var}(x)$.

Sample correlation coefficient r

- r^2 is a biased estimator of ρ^2.
- If the data comes from a bivariate normal distribution, then for large n, the estimate is good (asymptotically unbiased and efficient).
- See this Wikipedia article for more information on exceptions.

https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#Sample_size

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

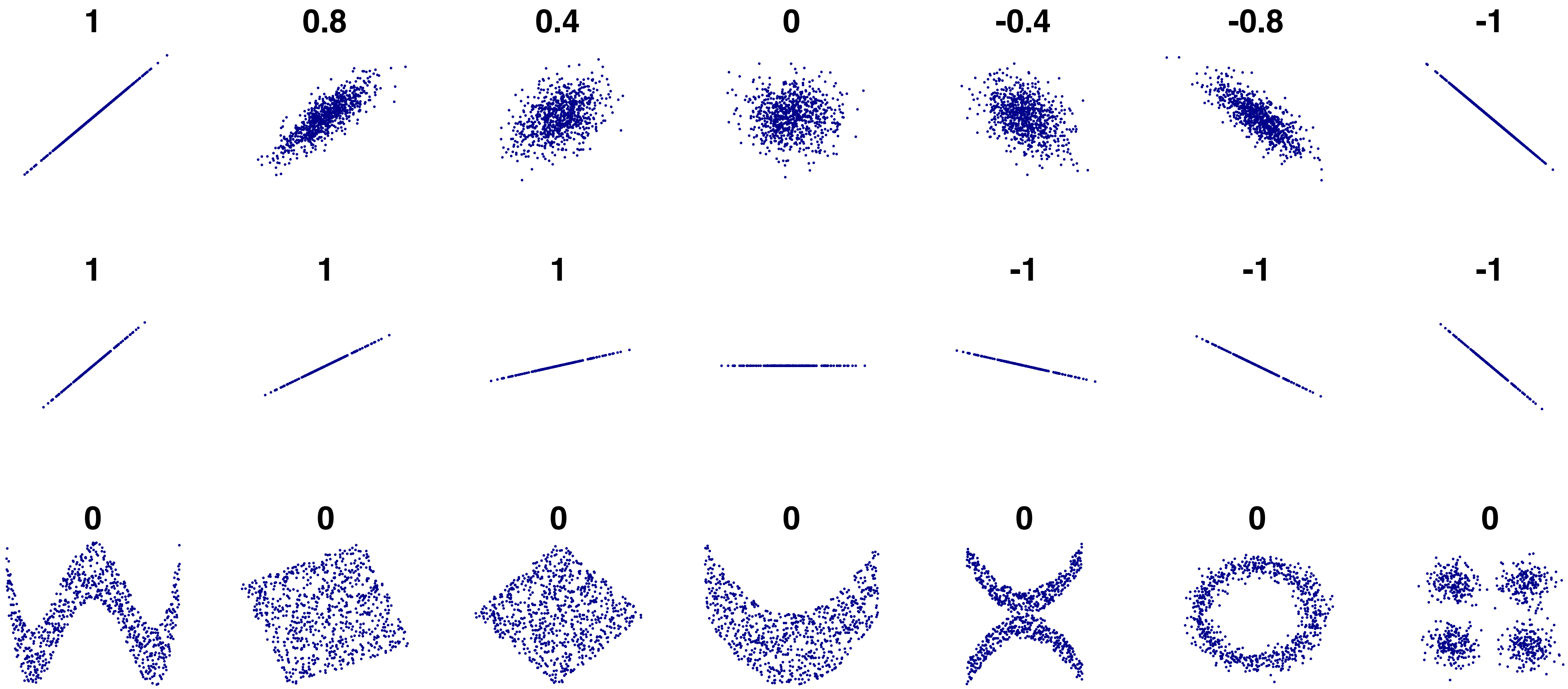
$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{Total variation}} = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{Unexplained variation}} + \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{Explained variation}} + \underbrace{2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\substack{= 0 \text{ by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

  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

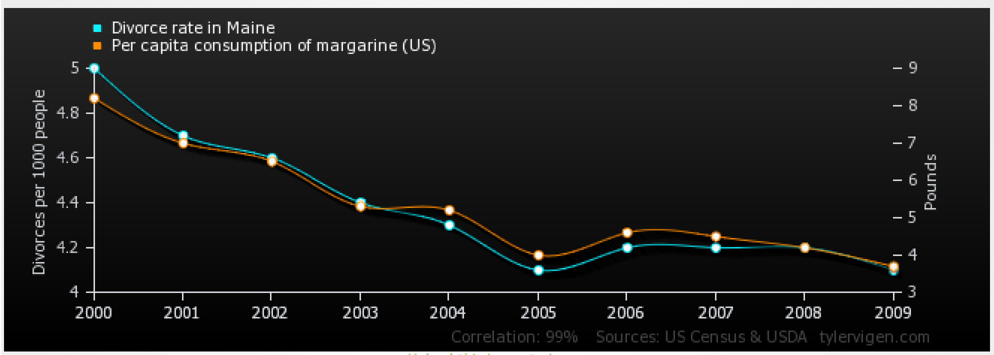
- **Top row:** Linear relations with varying r.
- **Bottom:** $r = 0$, yet X and Y are dependent in all of these (except possibly the last); it's just that the relationship is not a line.

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

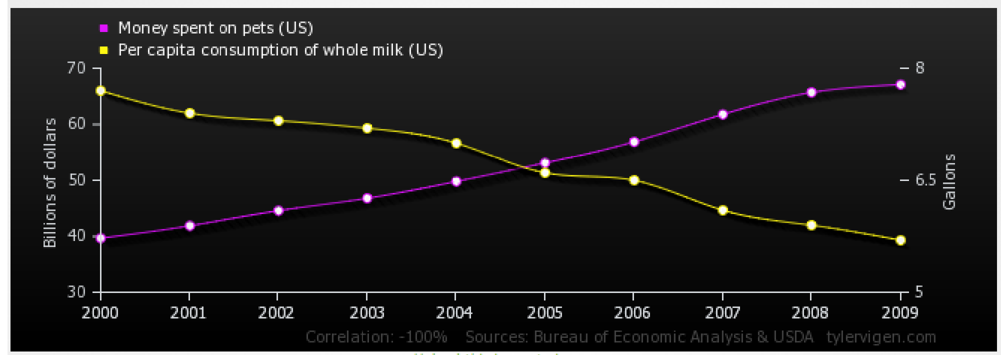
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

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
  $$H_0\colon \rho = 0 \quad \text{vs.} \quad H_1\colon \rho \neq 0 \quad \text{(or other values).}$$

- 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

Polynomial regression

- Model y as a polynomial in x of degree p.

$$y = \beta_0 + \beta_1 x \quad + \beta_2 x^2 \quad + \cdots + \beta_p x^p$$

- The ith observation (x_i, y_i) gives
$$y_i = \beta_0 + \beta_1 x_i \; + \beta_2 {x_i}^2 + \cdots + \beta_p {x_i}^p + \epsilon_i$$

- Matrix notation: $\vec{y} = X\vec{\beta} + \vec{\epsilon}$

$$\begin{array}{ccccccc}
\vec{y} & = & X \text{ (design matrix)} & \cdot & \vec{\beta} & + & \vec{\epsilon} \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & x_1 & {x_1}^2 & \cdots & {x_1}^p \\ 1 & x_2 & {x_2}^2 & \cdots & {x_2}^p \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & {x_n}^2 & \cdots & {x_n}^p \end{bmatrix} & \cdot & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} & + & \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
n \times 1 & & n \times (p+1) & & (p+1) \times 1 & & n \times 1
\end{array}$$

- MLE point estimate of $\vec{\beta}$ is $\widehat{\vec{\beta}} = (X'X)^{-1}X'\vec{y}$.
  Need $X'X$ to be non-singular and $n \geqslant p+1$ (usually a lot bigger).

Multiple linear regression

- Model one dependent variable as constant + linear combination of p independent variables. Goal is a best fit for
$$y = \beta_0 + \beta_1 x_{(1)} + \beta_2 x_{(2)} + \cdots + \beta_p x_{(p)}$$
- The ith observation $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ gives
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$
- Matrix notation: $\vec{y} = X\vec{\beta} + \vec{\epsilon}$

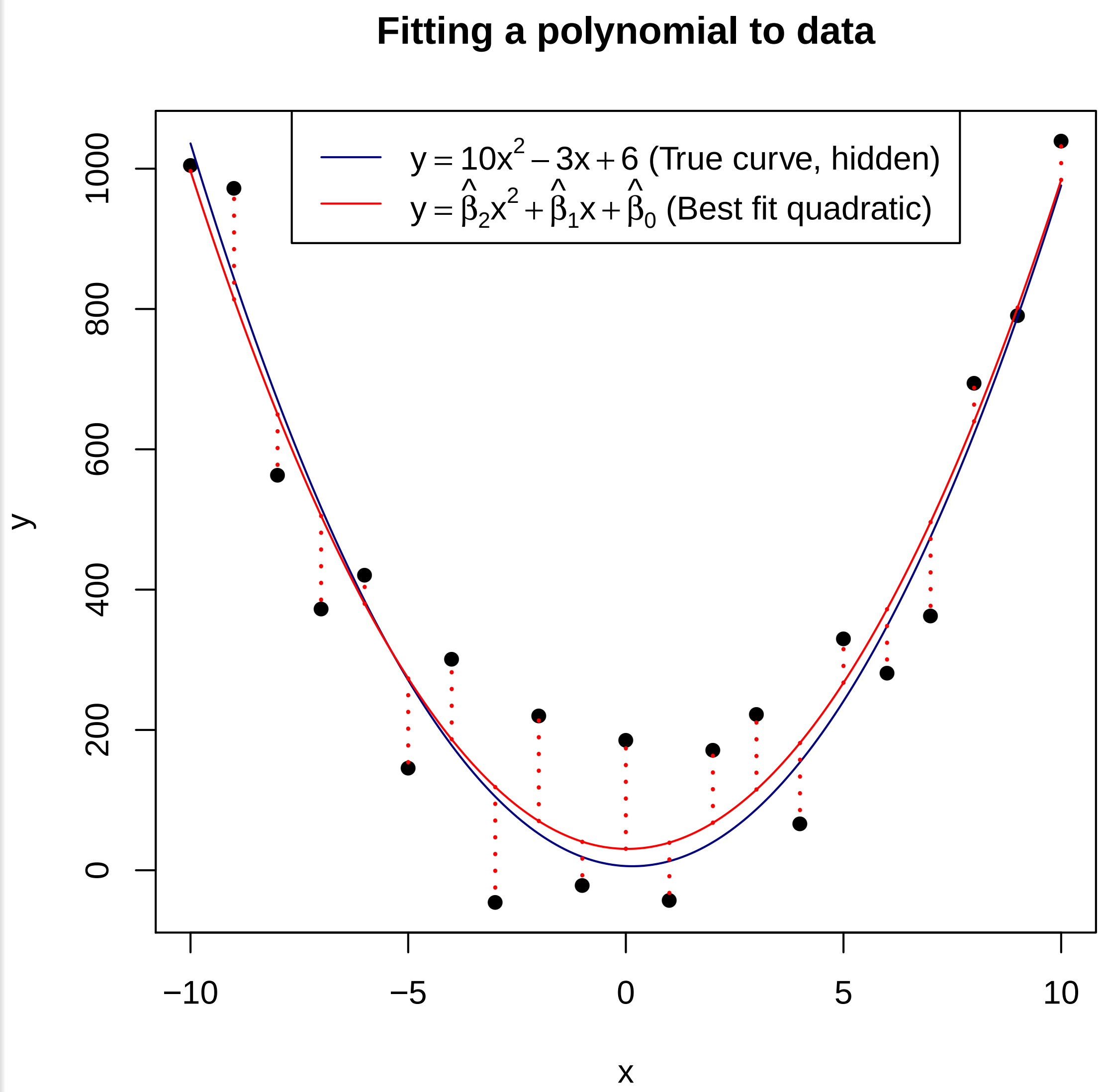
$$\begin{array}{ccccccc}
\vec{y} & = & X \text{ (design matrix)} & \cdot & \vec{\beta} & + & \vec{\epsilon} \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} & \cdot & \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} & + & \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\
n \times 1 & & n \times (p+1) & & (p+1) \times 1 & & n \times 1
\end{array}$$

- MLE point estimate of $\vec{\beta}$ is $\widehat{\vec{\beta}} = (X'X)^{-1}X'\vec{y}$.
Need $X'X$ to be non-singular and $n \geqslant p+1$ (usually a lot bigger).

Example in Matlab

```
>> # Generate data with known x
>> # but random errors in y
>> x = (-10:10)'; # column vector
>> err = normrnd(0, 100, size(x));
>> y = 10*(x.^2) - 3*x + 6 + err;

>> # Point estimate (no conf. int.):
>> polyfit(x,y,2)
    9.5968   -0.6319   30.5096

>> # Interval estimate (with conf. int.)
>> # Create the design matrix
>> Xdesign = [ones(size(x)), x, x.^2]
Xdesign =
     1   -10   100
     1    -9    81
...
     1    10   100

>> [b, bint] = regress(y, Xdesign)
b =
   30.5096
   -0.6319
    9.5968

bint =
  -48.6394  109.6587
   -9.3294    8.0655
    7.9854   11.2082
```

Fit is $y = 9.5968x^2 - 0.6319x + 30.5096$



Example in R

Fit is $y = 9.5968040x^2 - 0.6319475x + 30.5096087$

Example in R

```
> # Generate data with known x
> # but random errors in y
> x = -10:10;
> n = length(x);
> err = rnorm(n, 0, 100);
> y = 10*x^2 - 3*x + 6 + err;

> # Fit to y = b0 + b1*x + b2*x^2
> # intercept b0 is implied:
> bestfit = lm(y ~ I(x) + I(x^2));

> coefficients(bestfit)
(Intercept)        I(x)      I(x^2)
 30.5096087  -0.6319475   9.5968040

> confint(bestfit)
                 2.5 %     97.5 %
(Intercept) -48.639445 109.658662
I(x)         -9.329402   8.065507
I(x^2)        7.985427  11.208181
```