This regression time: ↑
next time: ANOVA

 homework 4 due Fri 22 Jul in class
 take-home final due Mon 25 Jul noon
 extra office hours:
 Mon 25 Jul 9A - 10A
 take elevator to 3rd floor Baskin: door open
 Sat 23 Jul noon - 1 pm
 Sun 24 Jul noon - 1 pm
 Sat 4D 23 Jul 2-3 pm
 Sun 4D 24 Jul 2-3 pm

y
Galton
(e. 1890)

SD line: best for capturing the trend of (x, y) pairs
regression line for predicting y from x
reg. line for prev. x from y
\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

make predictions for \( \hat{y} \)

\[ \text{sum} \left( y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2 \]

\[ \hat{\beta}_0, \hat{\beta}_1 \text{ to minimize} \]

\[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

Calculation: least squares line

least squares

least squares
harder to estimate \( \beta \). Here to judge whether a sample slope \( \beta \) is large in practical terms, use exactly the same line of reasoning as with the sample correlation \( r \).
how useful is the regression!

predictive task

0) ignore x
or don't measure it:

predict

\[
y: \hat{y} = \frac{y}{\text{no x}}
\]

with \( SE(\hat{y}) = \sigma_y \)
predictive task 2: use $x$ to predict $y$:

\[ \hat{y} = \beta_0 + \beta_1 x \]

with $\hat{SE}(\hat{y}_{pred}) = \hat{y} \sqrt{1-r^2}$

This is smaller than $SE(\hat{y}_{nox})$.
A residual plot

A scatter plot

Healthy residual plot: no trend or pattern
\[ y = \text{height} \]
\[ x = \text{trunk diameter} \]
\[ y = \beta_0 + \beta_1 x + \epsilon \]

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i \]

Can generalize least squares to get estimates:

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} \]
$$
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix}

\begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1k} \\
  x_{21} & x_{22} & \cdots & x_{2k} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}

\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_k
\end{bmatrix}

\text{if}
$$

Is the regression useful?

Compute

$$R^2 = \frac{\text{multiple R}^2}{\text{coefficient of determination}}$$

Want to