Relationships between variables

•••

11 March 2019
PLSC 309

Review

- So far we have learned about single variables
- We have learned how to express variables as an infinite series of events
  - Probability distributions
- We have learned how to compare samples of a variable
  - Parametric: Z-score; t-test
  - Non-parametric: bootstrap, permutation test

Preview

- Statistical modelling is about finding relationships between multiple variables
- The easiest kind of relationship to find is a linear relationship
- Linear relationships are additive and proportional

Preview: linear regression

Today: 30,000 ft view -- conceptual introduction

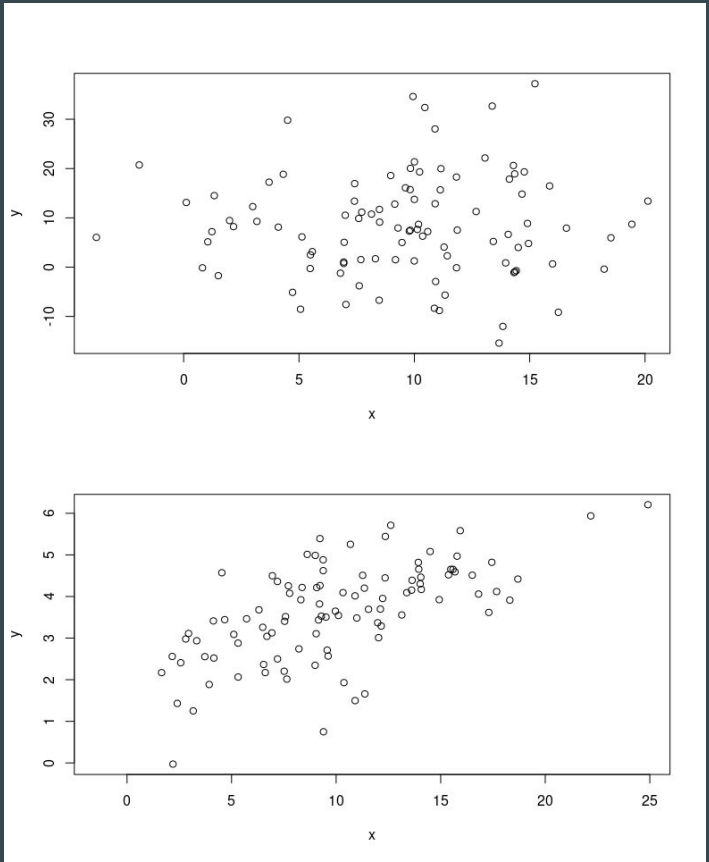
Wednesday: nuts and bolts; inference for regression

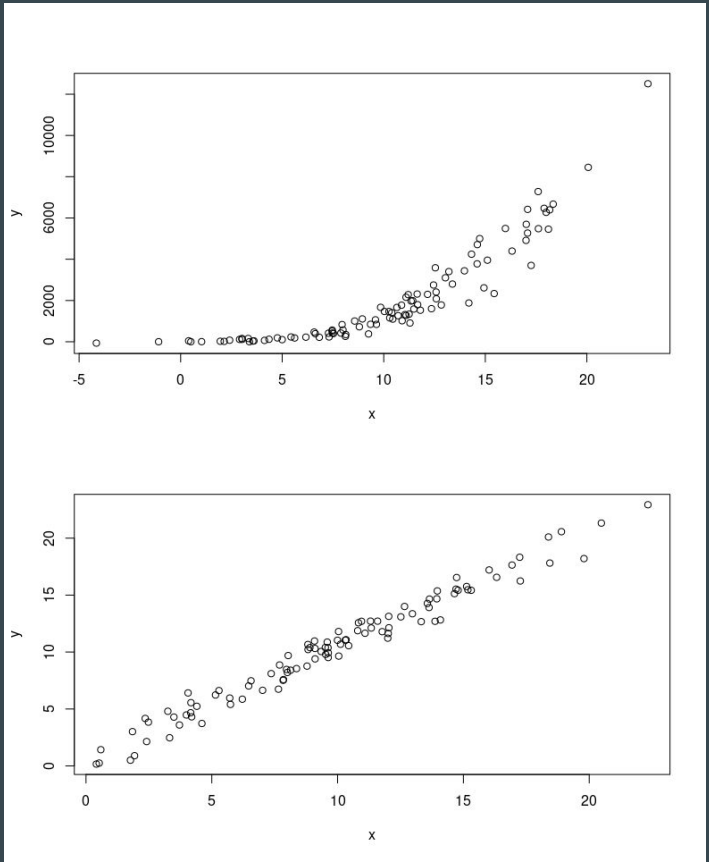
Friday: get our hands dirty

Explanations and Outcomes

- In order to create a statistical model, we must first decide what we're studying
  - This is the *outcome* or *response*
  - Symbolized by Y (vector)
- Additionally, we can use one or more variables to account for variation in Y
  - These are *explanatory variables* or *features*
  - Symbolized by X or **X** (vector or **matrix**)

Relationship between X and Y





Correlation

- The correlation represents the strength of linear relationships between two variables
- Ranges from -1 to 1
  - 1 = perfect positive linear relationship
  - -1 = perfect negative relationship
  - 0 = no linear relationship whatsoever

How to calculate correlation

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

- x_i, y_i = individual observation
- x-bar, y-bar = means of x and y
- s_x, s_y = standard deviation for x and y
- n = number of observations

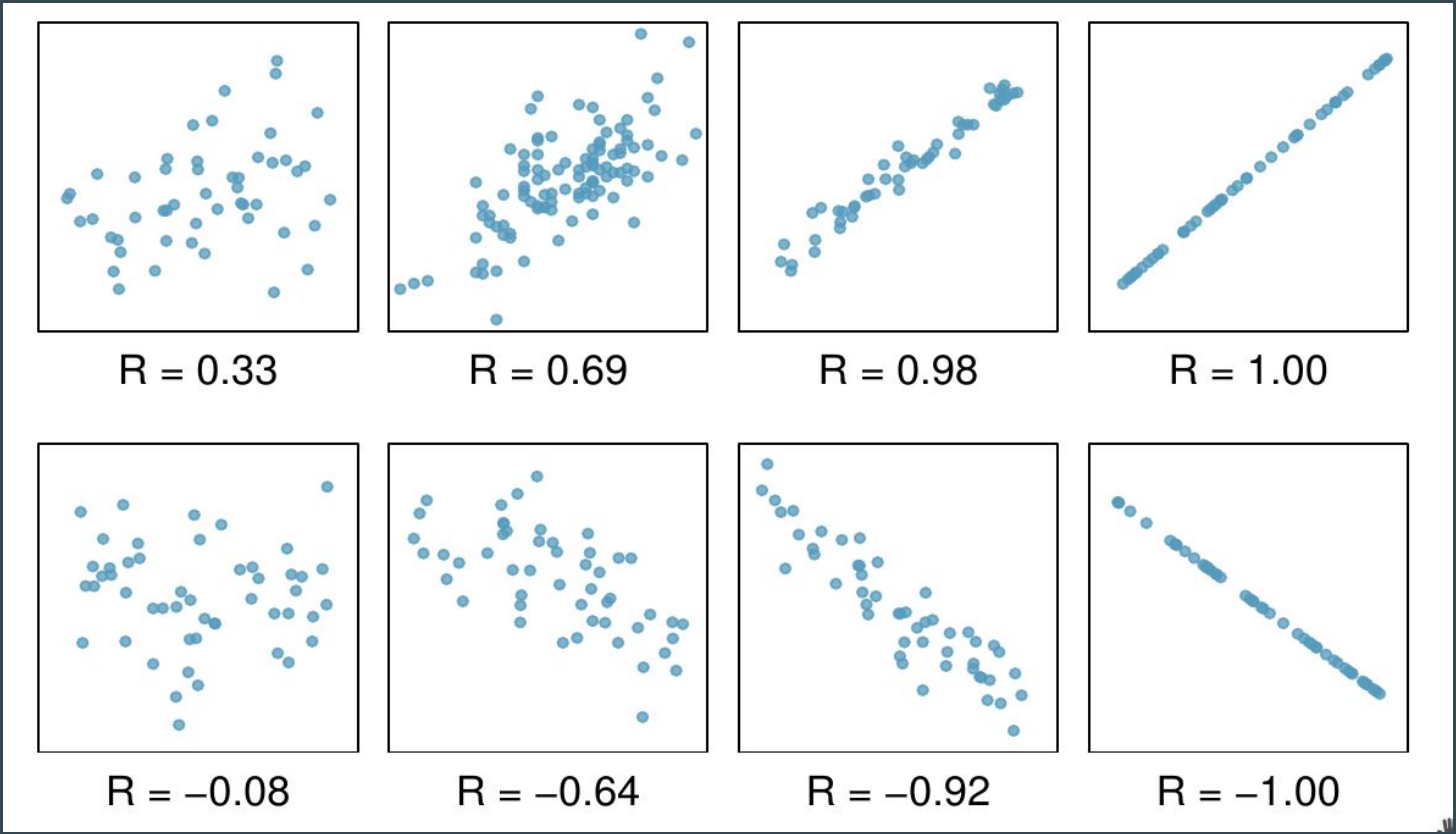
How to calculate correlation

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

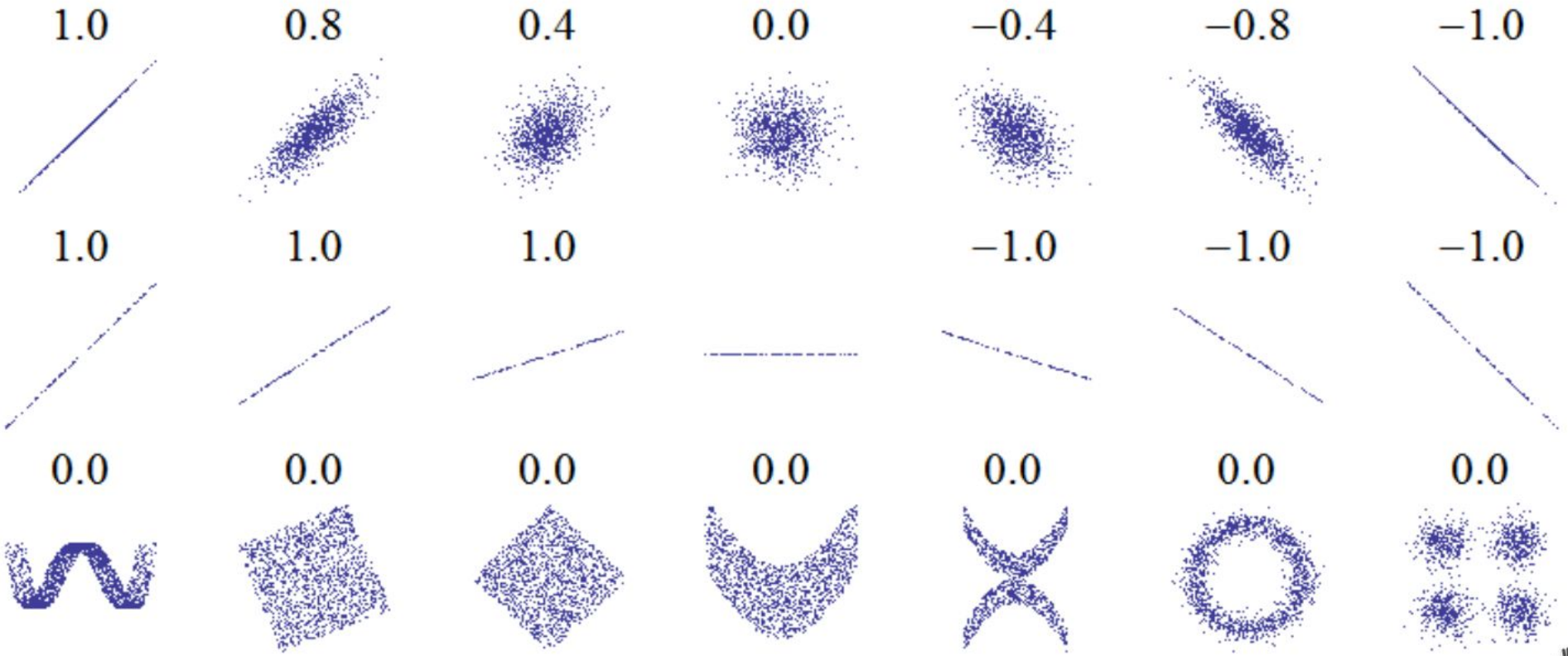
An average of deviations from the mean, scaled by their standard deviation

- Greater standard deviations = smaller correlation
- Greater deviations from average *for the same observation* = larger correlations

Correlations visualized

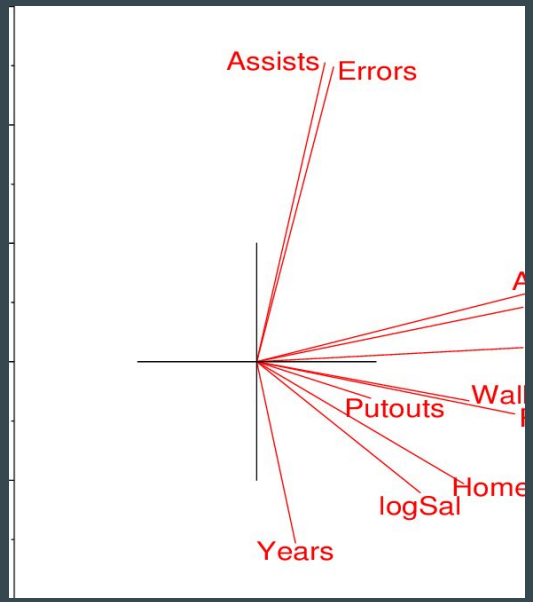


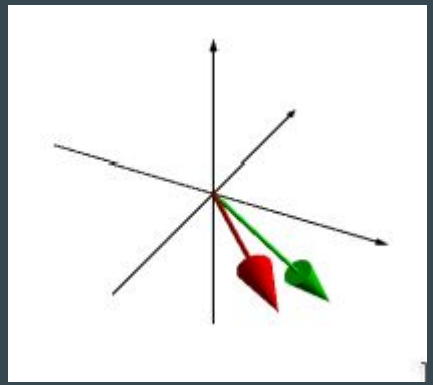
Correlations visualized



A different view of correlations

- Remember, a variable is a vector, or a list of numbers, with as many dimensions as observations





A different view of correlation

The correlation is actual the cosine of the angle of two variables (represented as vectors)

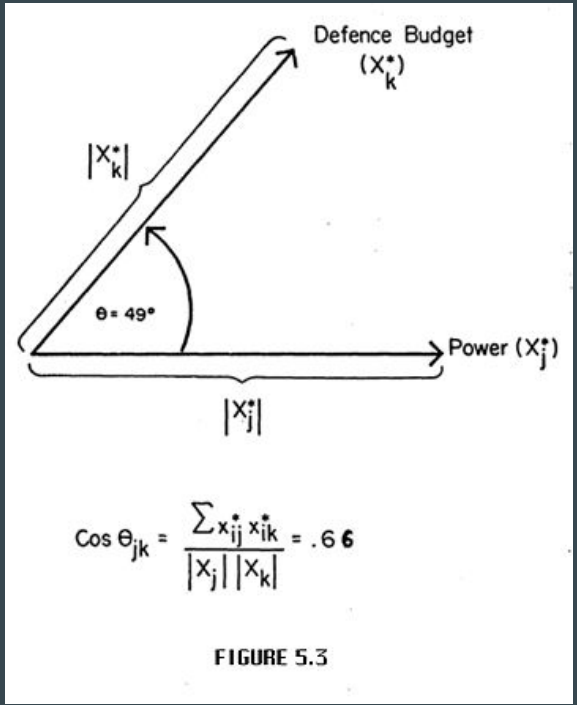


FIGURE 5.3

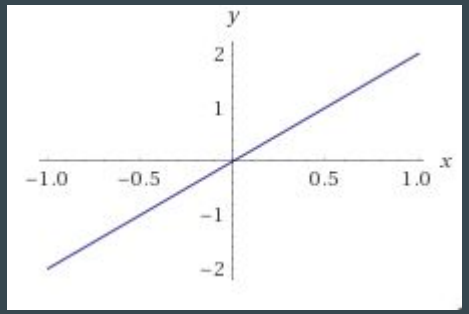
Review: why statistical modeling?

- We are interested in statistical modelling because we are interested in outcomes
- We want to use some set of data to *model* another variable
- In other words, we want to take our data x, input it into some function f(x), and get a result, our outcome
- This allows us to the answer the question, if we had different values of x, what would we get for y?
- In other words, statistical modeling is about *prediction*

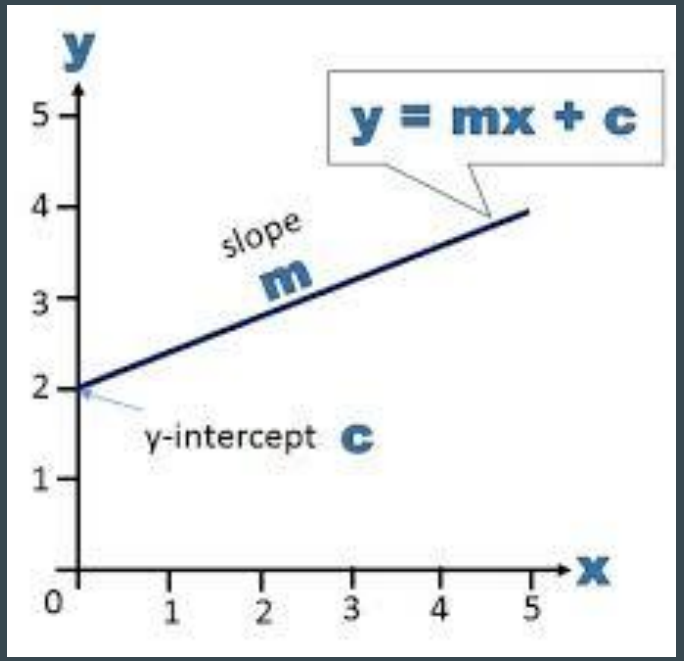
Statistical models are functions

- A function has three parts:
  - Input: **X**
  - Transformation: f(**X**)
  - Output: Y
- Example: f(x) = 2x

Input	Transformation	Output
2	2(2)	4
3	2(3)	6
4	2(4)	8

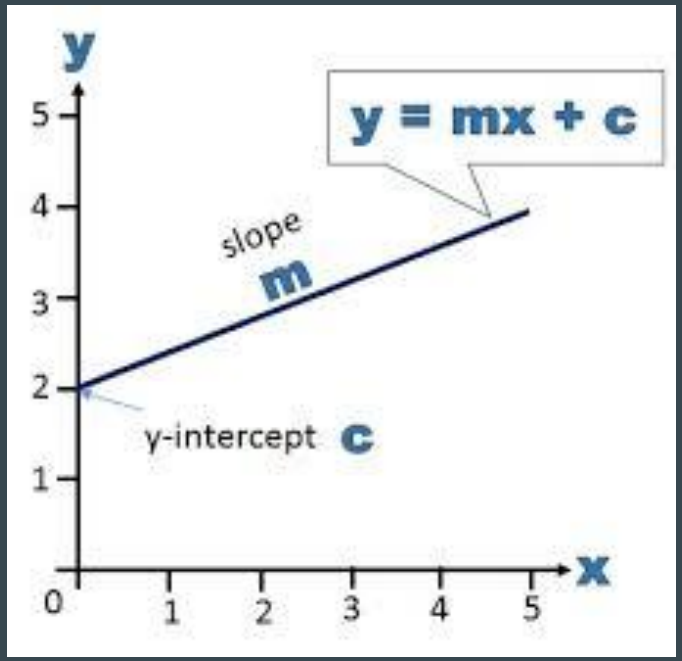


We'll start with linear functions



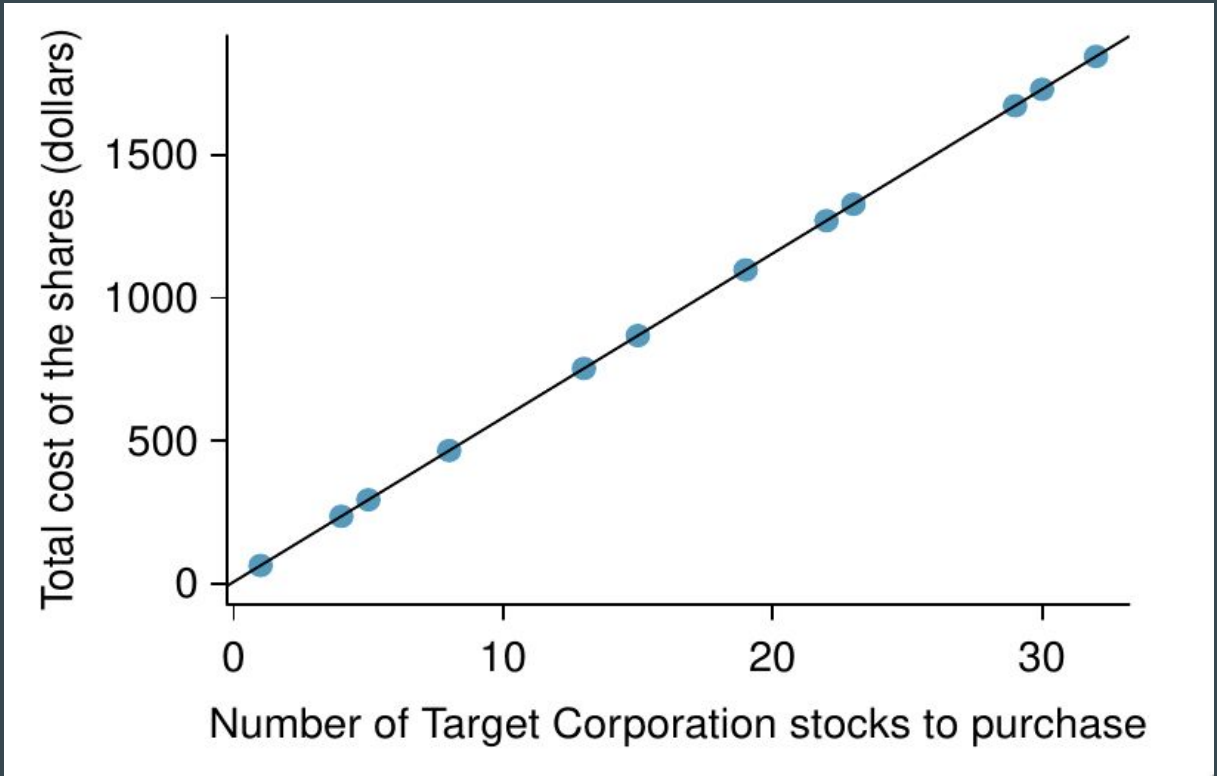
- y and x are variables
- m = slope
- c = y-intercept

Equation of a line in statistics terminology



- y and x are variables
- β = slope
- α = y-intercept
- β and α are parameters for our function

Perfect linear equations

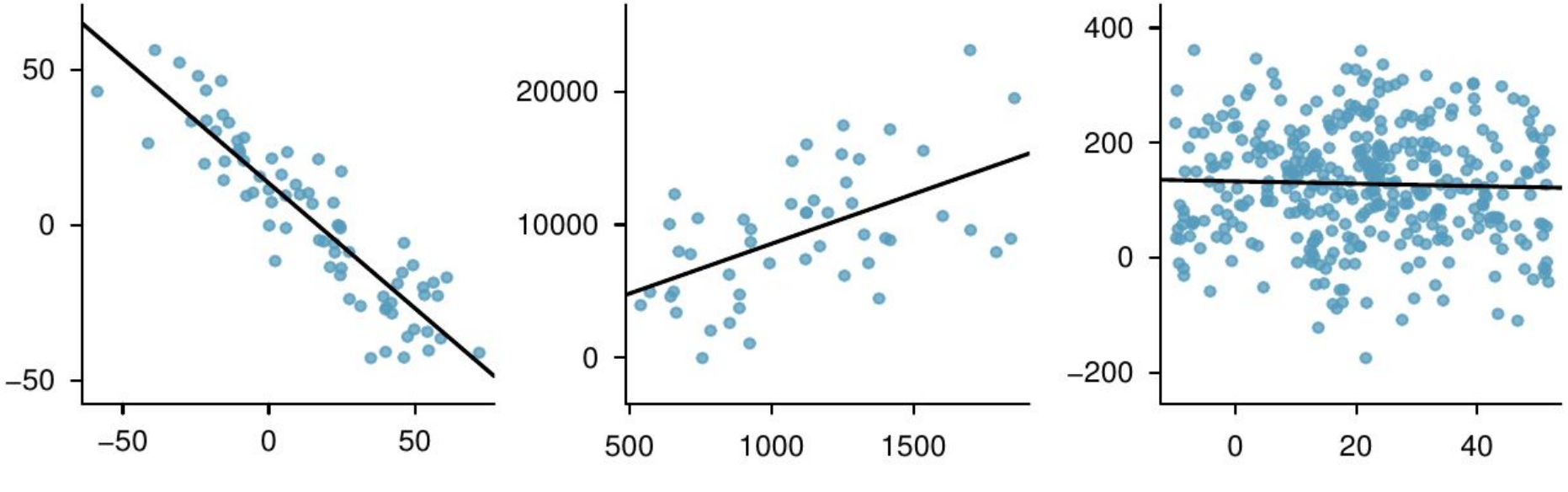


- Equation can be solved with two data points
- Two unknowns
  - Y-intercept
  - Slope
- $Y = \beta X + \alpha$
  - $500 = \beta 9 + \alpha$
  - $1500 = \beta 27 + \alpha$

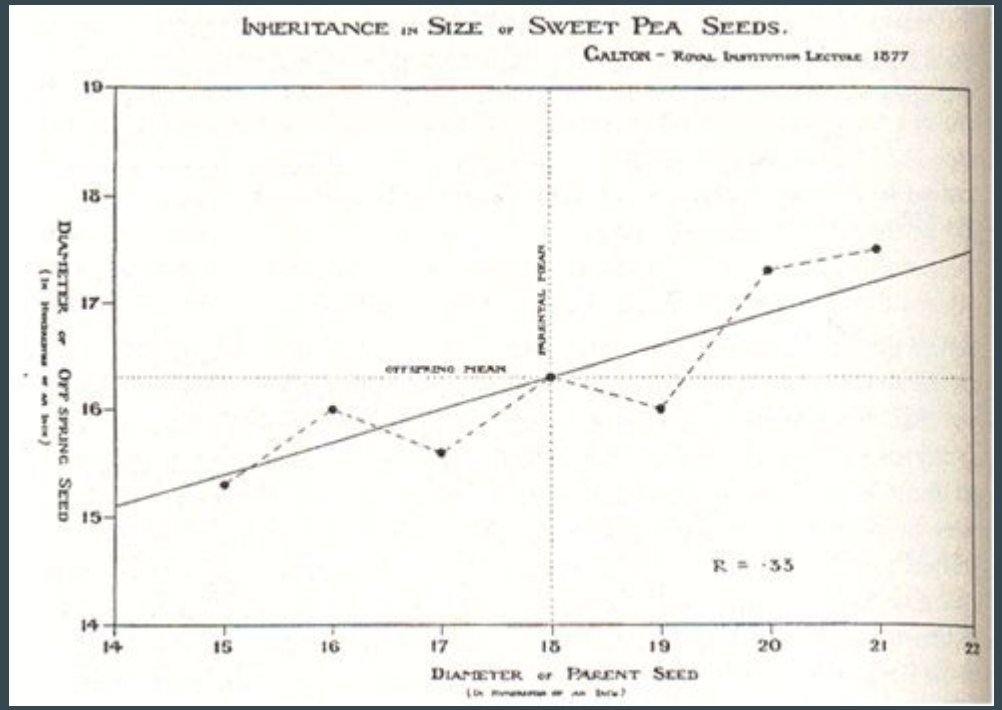
The parameters of the linear model: β and α

- α is the y-intercept
  - The value our outcome takes when all explanatory variables are 0
  - If you are modelling the impact of year of college on income, α can be interpreted as the predicted income for an individual with no college
- β is the slope
  - We have as many β's as we have explanatory variables
  - Describes the magnitude and direction of a relationship between X and Y
  - If you are modelling the impact of education on income, β for education can be interpreted as the direction (positive or negative) and magnitude (absolute size) of the relationship

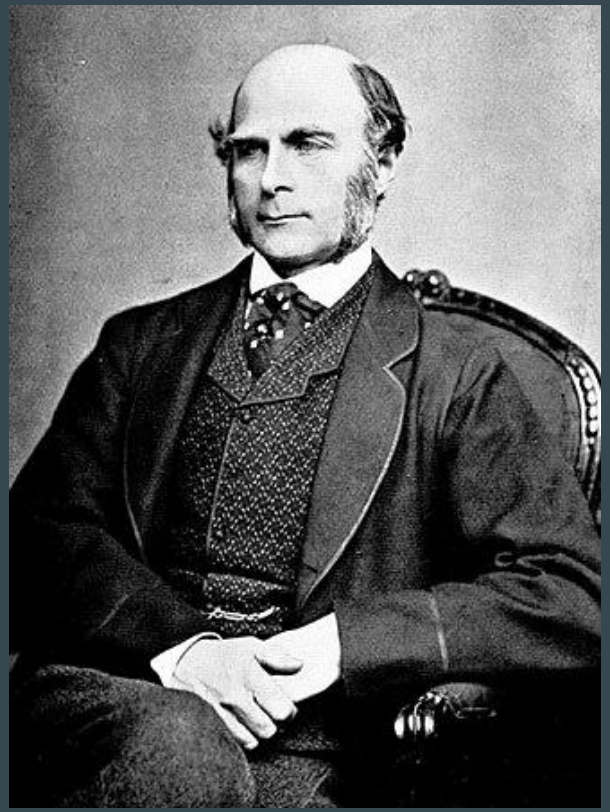
From math to statistics



Fundamental problem: how do we measure a relationship when there is no perfect linear solution?



Sir Francis Galton



- Inventor of regression, correlation, and standard deviation
- First to use the phrase “eugenics” and “nature vs. nurture”
- Invented modern fingerprinting classification
- Invented the first weather map
- Introduced sleeping bags to Europe

Fitting an *imperfect* linear relationship is known as regression

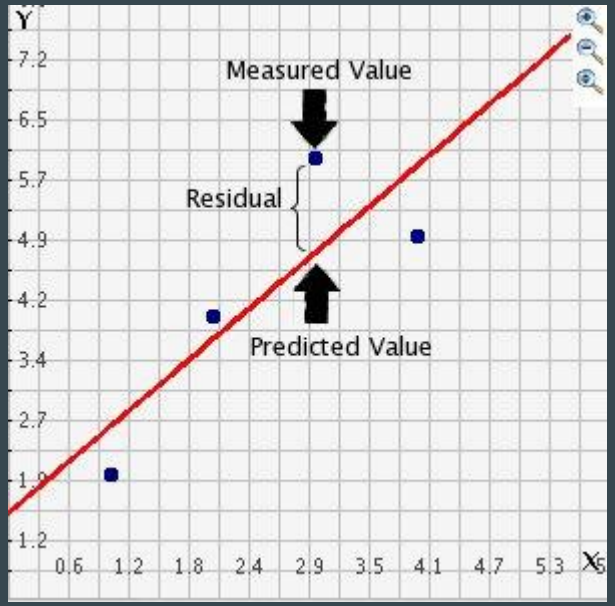
- Regression implies there is some true linear relationship between X and Y, f(X)
- However, due to random chance, observed data will generally fall above or below that line
- Similar logic to CLT and probability distributions

Solution: model the noise directly

- $Y = \beta X + \alpha$ now becomes...
- $Y = \beta X + \alpha + \varepsilon$
- ε = error term
- In other words, we're saying the data we get is some combination of a true, linear relationship, plus some random noise

An error is called a residual

- The residual is the observed value (our actual data) minus our predicted value
- $e = Y - \hat{Y}$



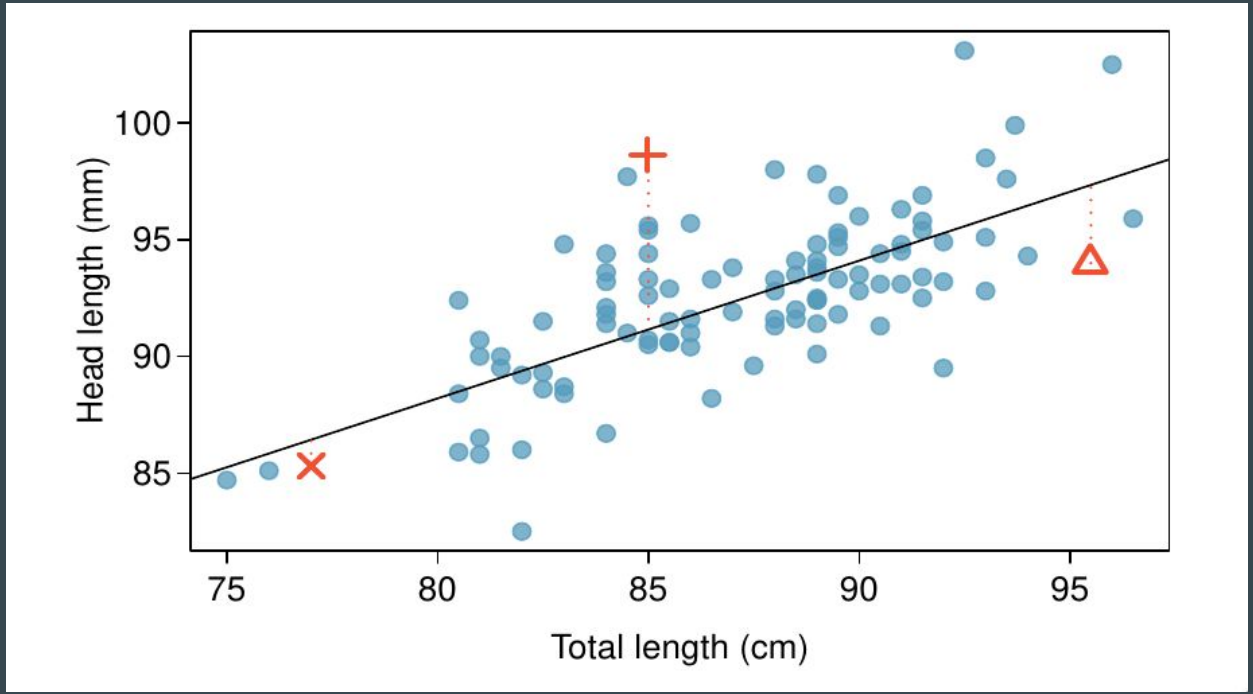
Calculating residuals

- We estimate a linear model with
  - $\alpha = 0$
  - $\beta = 3$

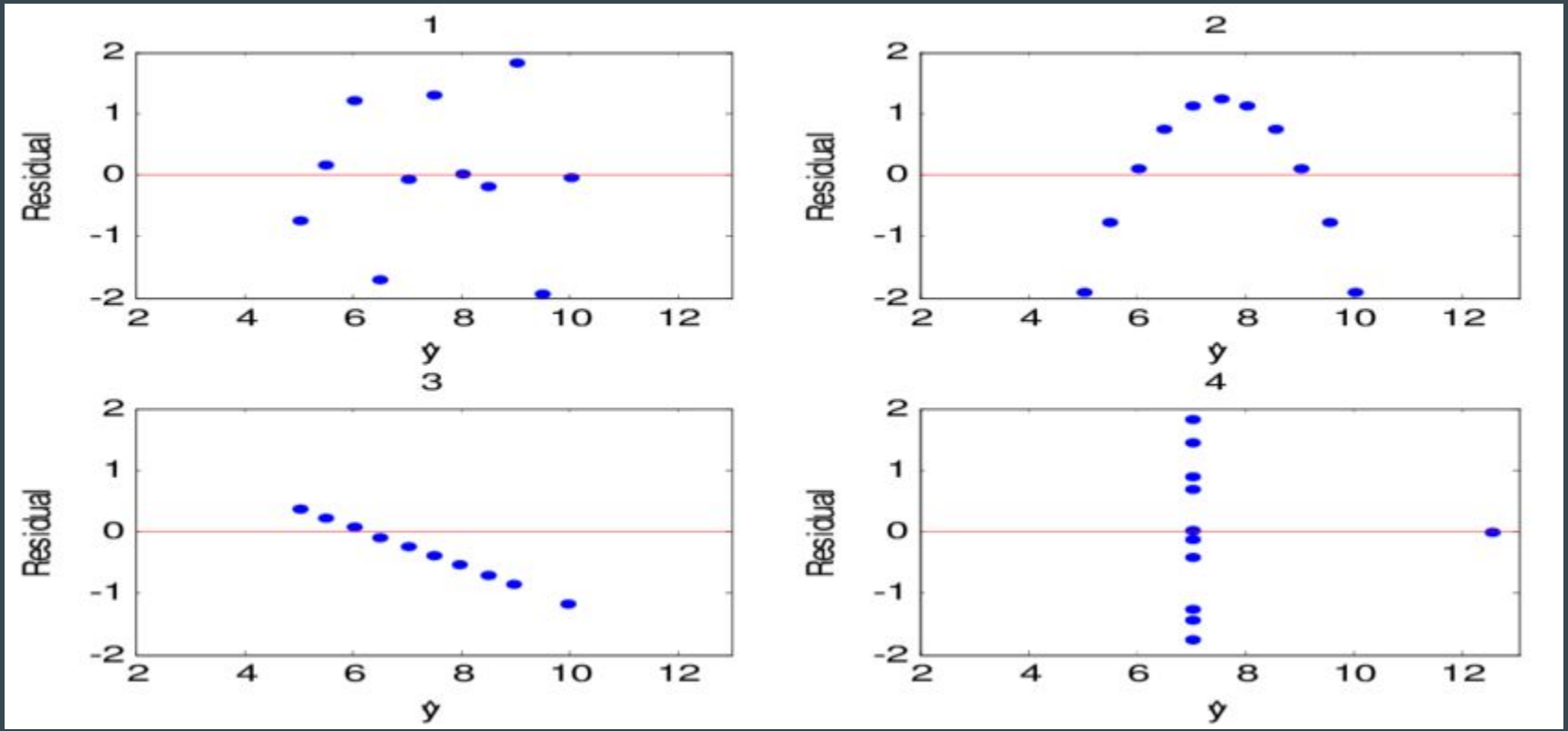
Y	X	e
10	3	-1
7.5	3	1.5
5	2	-1
12	4	0

Residuals

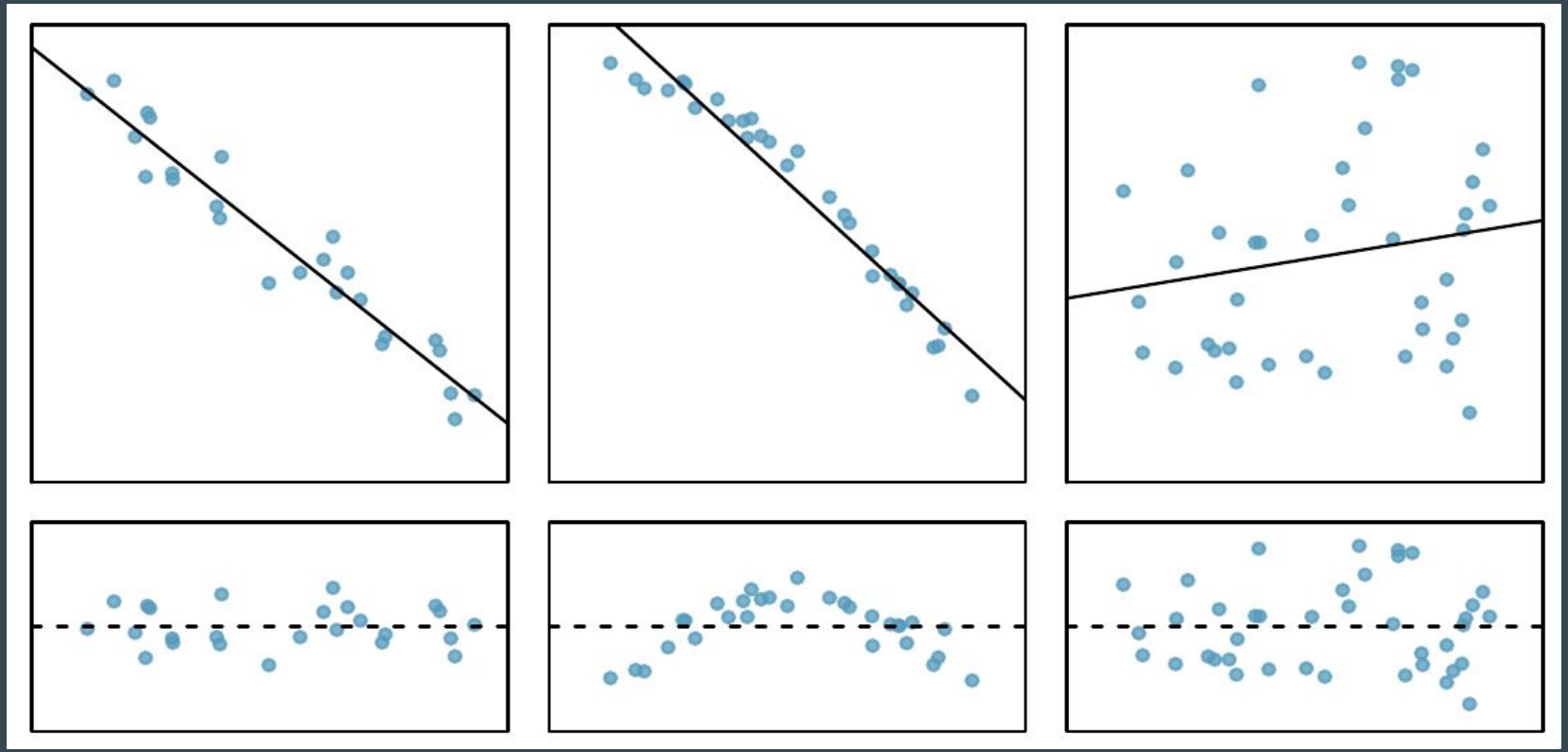
- Residuals can be large or small, positive or negative



Residuals

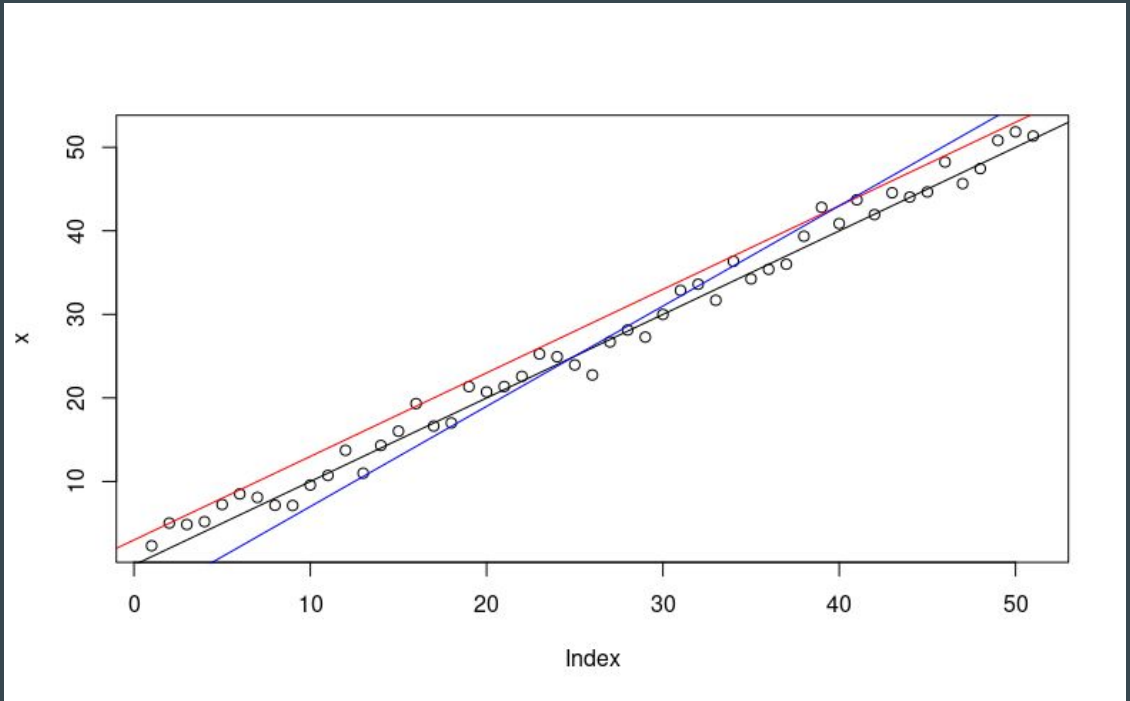


Patterns in residuals



Our goal is to find the best linear fit

- We want to find a line that minimizes the residuals



Best linear fit

- This can be done by eye
  - Find the straight line that passes through the most points
- But formally, it can be done by minimizing the residuals
- In other words, picking the line that finds the smallest average residual

Calculating model error

- The best square fit wants to find the smallest residuals in either direction
- We can do this by first squaring each of our residuals
  - $e_1^2, e_2^2, e_3^2, \ldots$
- And then summing the residuals
  - $e_1^2 + e_2^2 + e_3^2 + \ldots$
- This is known as the *sum of least squares*
- A regression model calculated with the sum of least squares is known as *Ordinary Least Squares (OLS) Regression*

Ordinary Least Squares

- We estimate a linear model with
  - $\alpha = 0$
  - $\beta = 3$

Y	X	e
10	3	-1
7.5	3	1.5
5	2	-1
12	4	0

$\sum e = (-1)^2 + (1.5)^2 + (-1)^2 + (0)^2 = 4.25$

Ordinary Least Squares

- We could try a different line
  - $\alpha = 0$
  - $\beta = 3.5$

Y	X	e
10	3	0.5
7.5	2	0.5
5	1	-1.5
12	3	-1.5

$\sum e = (0.5)^2 + (0.5)^2 + (-1.5)^2 + (-1.5)^2 = 5$

- β=3 is a better fit than β=2.5

Assumptions for OLS

There are four assumptions that we have to make for linear regression:

1. Linearity / additivity
2. Our residuals must be:
    a. Independent
    b. Homoscedastic (constant variance)
    c. Normally distriburted

Additivity assumption

Lines are “additive” in the sense that each variable is only raised to the power of one and added (never multiplied or divided)

- Additive: $Y = b_1X_1 + b_2X_2$
- Non additive: $Y = b_1(X_1 * X_2)$
- Non additive: $Y = b_1(X_1^2)$

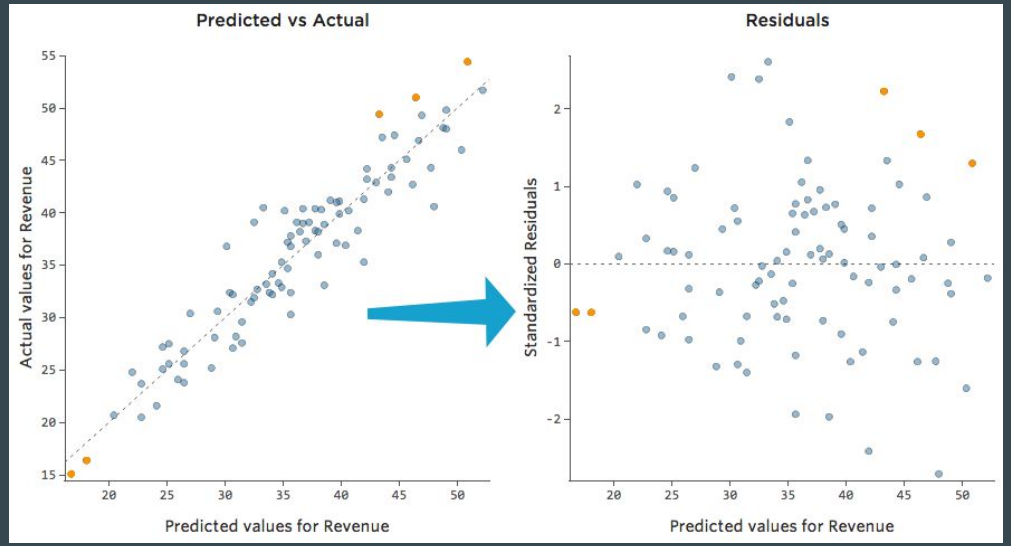
Additivity assumption

Lines are “additive” in the sense that each variable is only raised to the power of one and added (never multiplied or divided)

- An additive function means the slope one variable does not depend on the slope of another
- In other words, we could *add or subtract* variables without changing our estimates of the model parameters

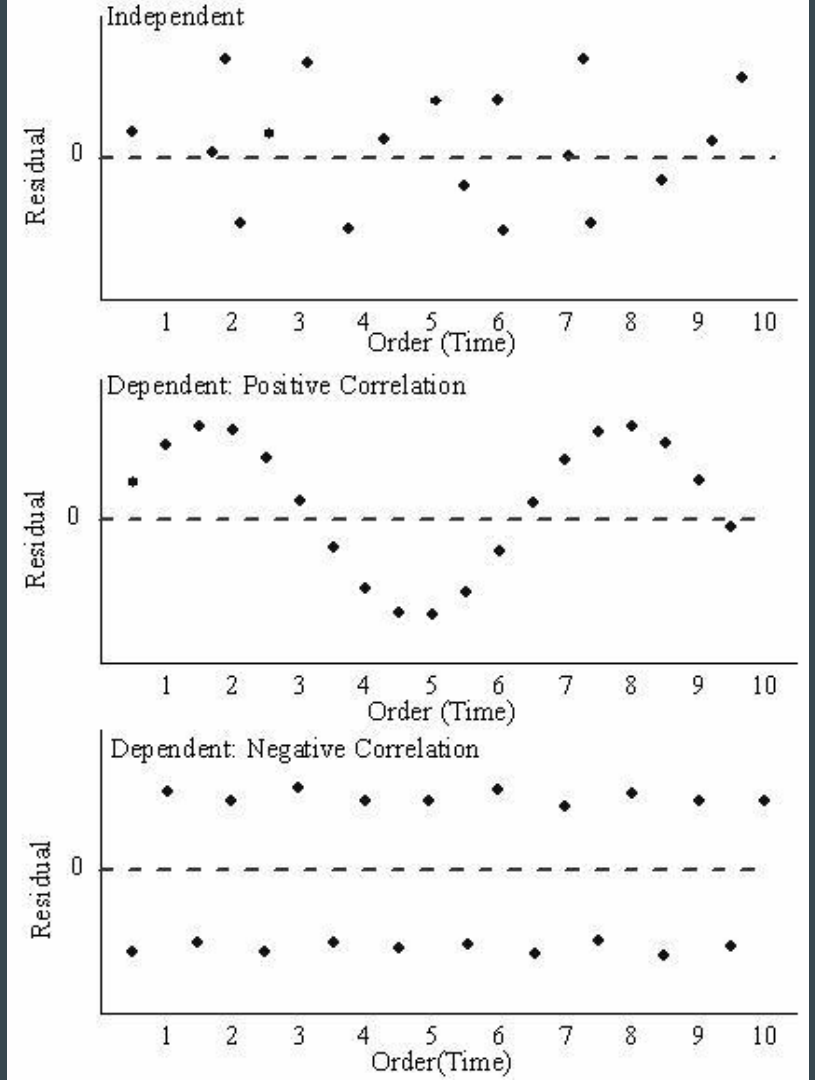
“Random” errors

- For OLS results to be valid, we do not expect to find any patterns in our residuals
  - Independent
  - Homoscedastic
  - Normally distributed



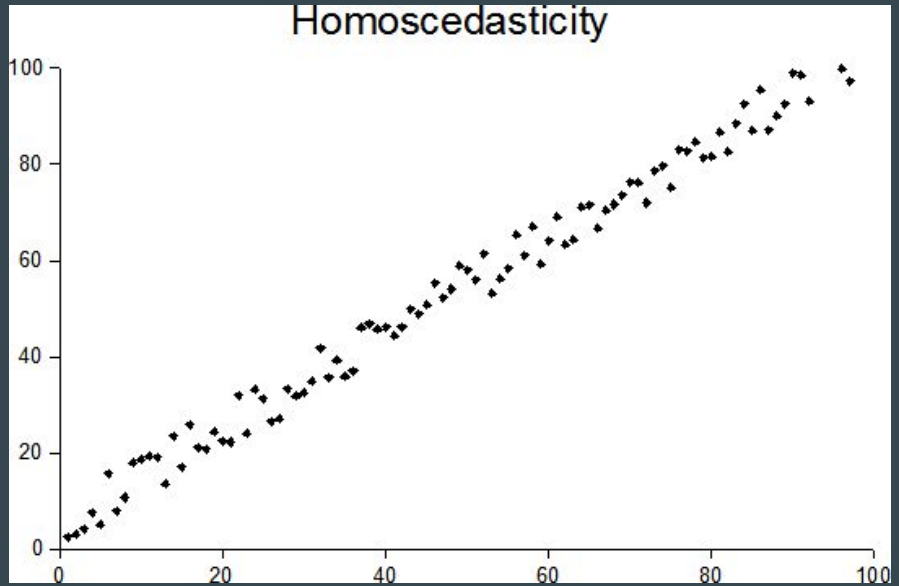
Dependent errors

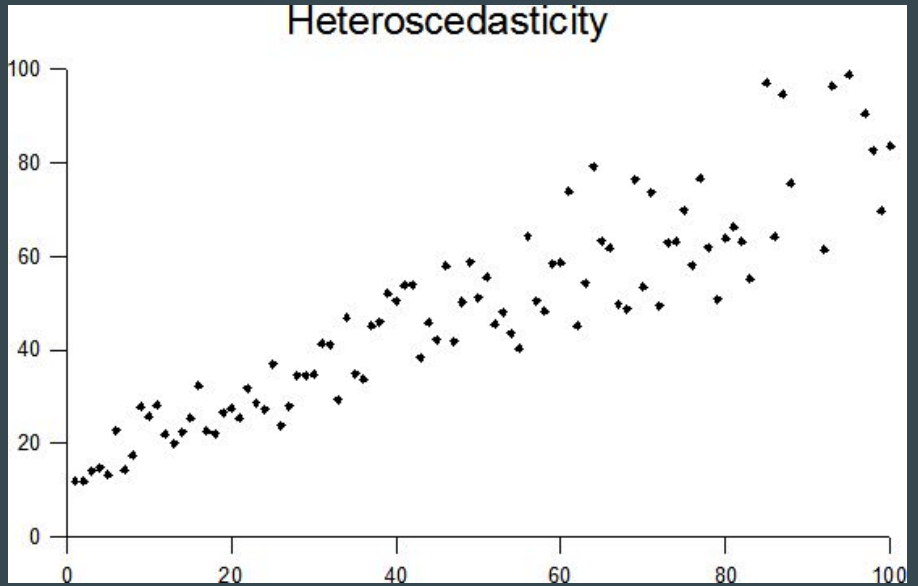
- This occurs when our residuals correlate with another variable



Errors have constant variance

- Errors with constant variance are *homoscedastic*; those without are *heteroscedastic*

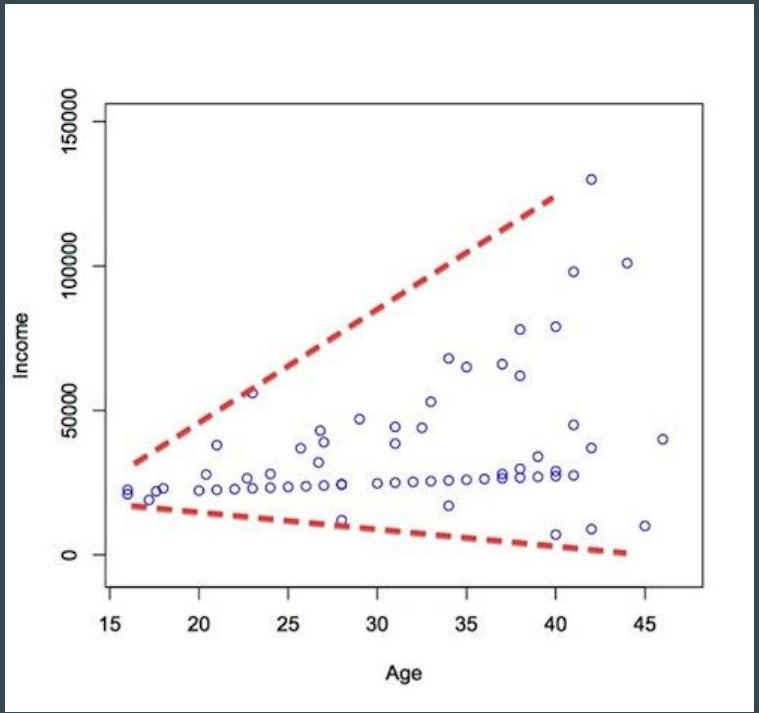




Heteroscedastic relationships

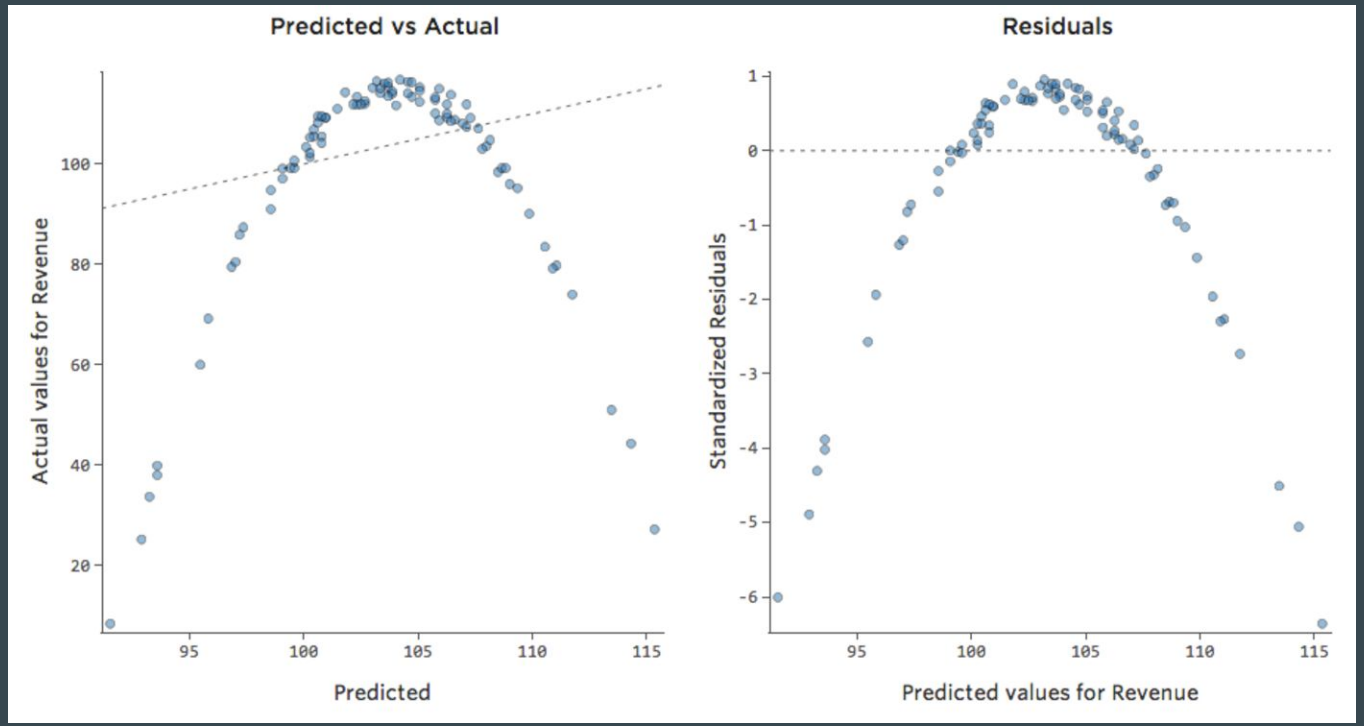
Consider using a regression to explain income by age

- At younger ages, even people who will become rich at not really that wealthy
- At older ages, the impacts of individual features are more clear



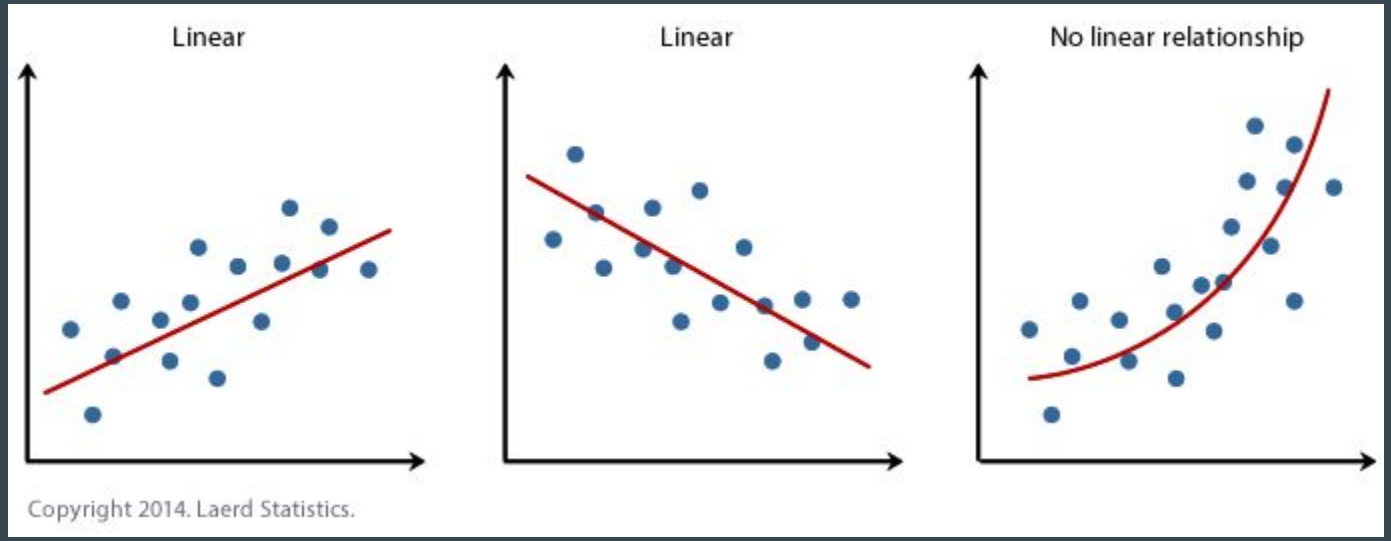
Residuals are normally distributed

- Non-normal residuals follow a “pattern”



Residuals are normally distributed

- Errors are non-normally distributed when the relationships modelled are non-linear



Non-linearity vs. Heteroscedasticity

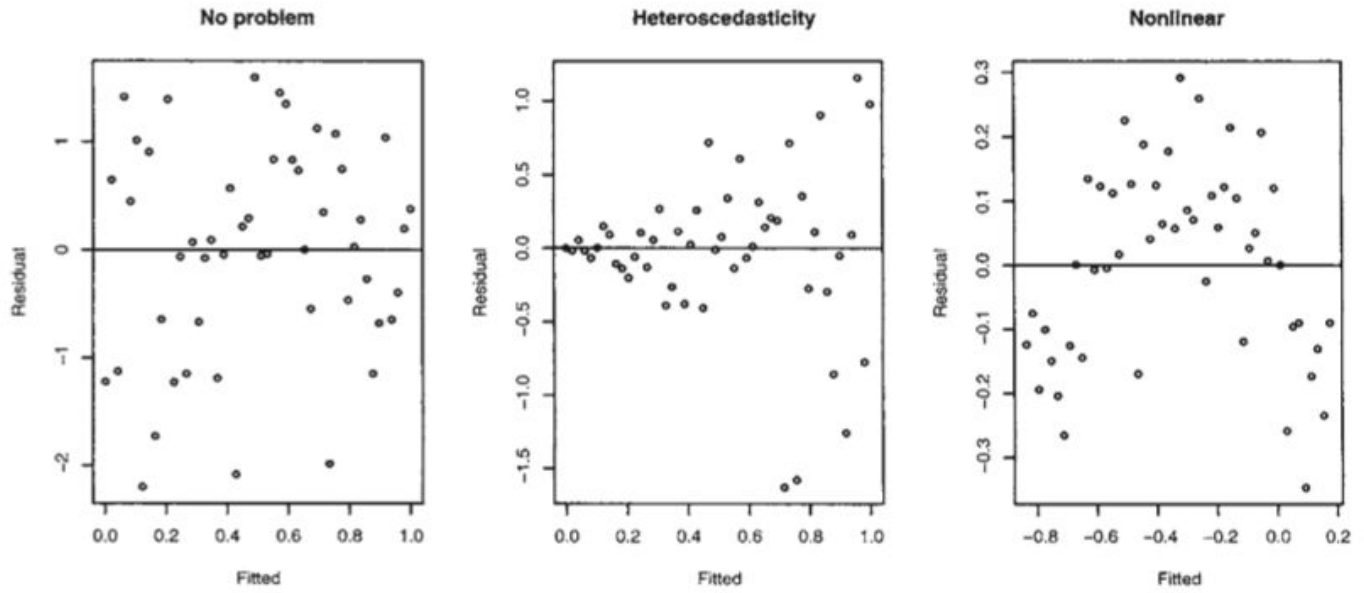


Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

Where do these assumptions come from?

- We are not just estimating *any* relationship, f(x) between **X** and Y
- We are estimating a very specific relationship
  - Linear / additive
- Why linear relationships?
  - The math is easy
  - Central Limit Theorem implies a linear relationship
- Because we are relying on the C.L.T., *i.i.d.* is a fundamental assumption
  - Independent
  - Identically distributed

Review: definitions and notation

- What we're interested in: Y, outcome, response
- What we're using to explain what we're interested in: **X**, explanatory variables
- Parameters of a linear model
  - Slope: β_i, b_i
  - Y-intercept: α, β_0, b_0
  - Errors or residuals: ε, e

Review: Linear relationships between variables

- We learned our first statistical model
  - Take some data X (explanatory variable) and Y (outcome/response)
  - Estimate a function f(X) with certain parameters
  - Minimize the loss to predict values of $\hat{Y}$
- Ordinary Least Squares
  - Take some data X (explanatory variable) and Y (outcome/response)
  - Estimate a *linear* (straight line) f(X) with parameters β and α
  - Find the parameters that minimize the *sum of squared errors*

Review: Ordinary Least Squares

- Find the line (i.e slope and intercept) that minimizes the squared differences ($\hat{Y}$-Y)
- $\hat{Y}$-Y are known as errors or residuals