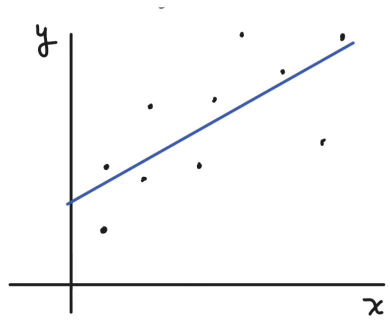
Regression

We often want to establish a relationship between an independent variable x, and a response (dependent) variable y by fitting a line through data points. The line of best fit is chosen to minimize the mean squared error between the line and the data.



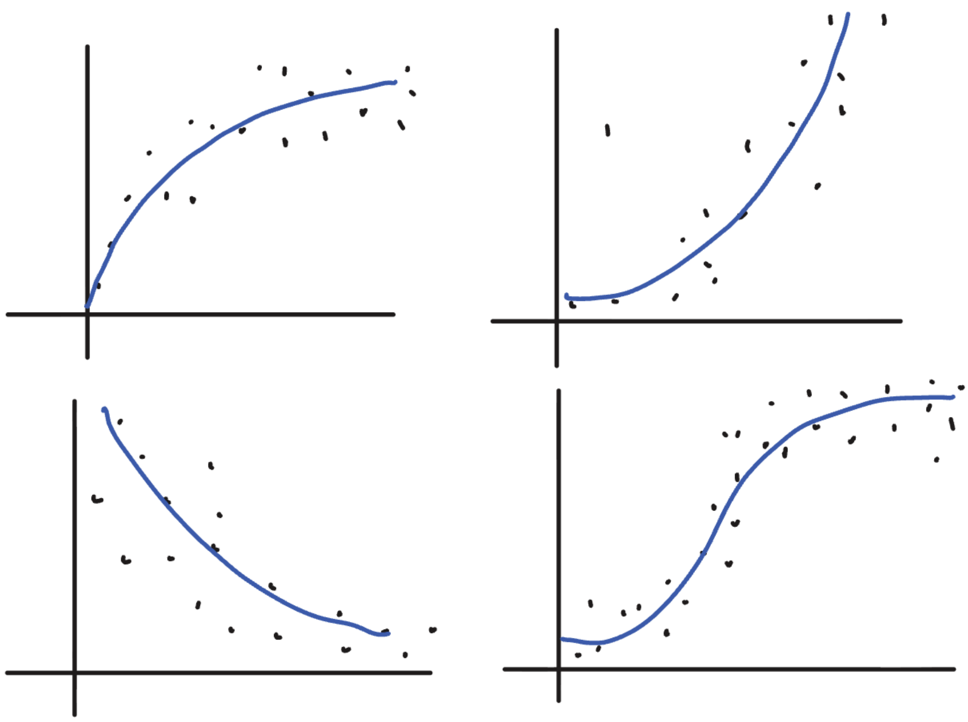
So here, $y = mx + b$ is a model. How do we compute this?

- Details: STAT 101, STAT 231, ...
- In this class, we will use software

Assessing the fit (2 ways):

- Qualitative: Plot data with a regression line
- Quantitative: R^2 is the coefficient of determination. This is the proportion of variance in y that is explained by the variation in x. $R^2 \in [0, 1]$, this is the "goodness of fit".

It is useful to simplify relationships within a model, or as a standalone tool to establish a relationship. These relationships can be linear, like above, or nonlinear, like below.

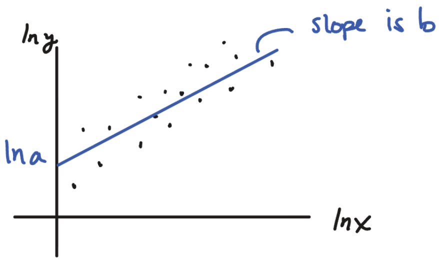


Either data, or prior information can suggest linear vs. nonlinear formulation. Let's learn how to fit linear and nonlinear regression to a dataset.

Consider the power function $y = ax^b$ (useful for determining scaling laws)

$$\begin{aligned} y &= ax^b \\ \ln(y) &= \ln\left(ax^b\right) \\ \ln(y) &= \ln(a) + \ln\left(x^b\right) \\ \ln(y) &= \ln(a) + b \cdot \ln(x) \end{aligned}$$

This is linear in variables $\ln(x)$, $\ln(y)$.



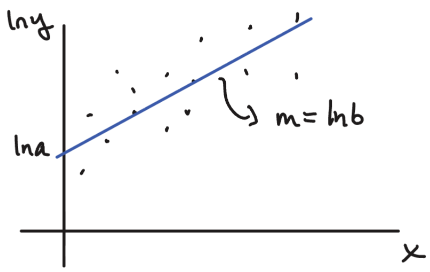
This is called a loglog plot. The slope of the line of best fit is the scaling power b in $y = ax^b$, and the y-intercept is $\ln(a)$

- `[coeffs, S] = polyfit(log(x), log(y), 1)`
- `coeffs(1) = slope = b`
- `coeffs(2) = intercept = ln(a)`

Consider the exponential function $y = ab^x$

$$\begin{aligned} y &= ab^x \\ \ln(y) &= \ln(ab^x) \\ \ln(y) &= \ln(a) + \ln(b^x) \\ \ln(y) &= \ln(a) + x \cdot \ln(b) \end{aligned}$$

This is linear in the variables x, $\ln(y)$.



The slope of the line of best fit is $\ln(b)$ and the y-intercept is $\ln(a)$.

- `[coeffs, S] = polyfit(x, log(y), 1)`
- `coeffs(1) = slope = ln(b)`
- `coeffs(2) = intercept = ln(a)`