Signal and Noise

Zack Treisman

Spring 2026

Philosophy

Basic scenario: x is a **predictor** and y the **response**.

We want to build and understand a model

$$y = f(x) + \epsilon$$

where

- $f(x)$ is **signal**
- ϵ is **noise**.

xkcd

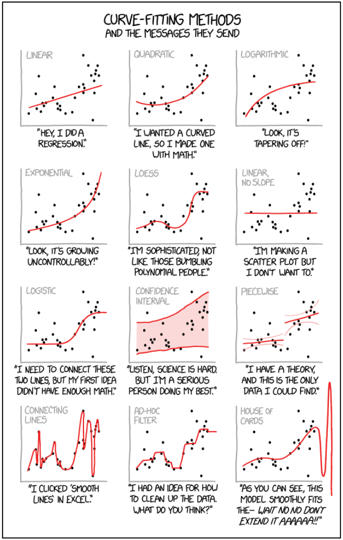


Figure 1: https://xkcd.com/2048

Parametric vs Non-parametric models

- **parametric** model: defined using arithmetic and analytic functions.
  - Coefficients, exponents *et cetera* defining the function are called the **parameters**.
  - Often more interpretable and meaningful.
- **non-parametric** model: decision trees, random forests, neural networks etc.
  - Often have better predictive power and fit data more effectively.
  - Might be a black box. Not as explanatory.

Reducible and irreducible error

- $\mu(x)$ = true[1] expected value of y given x.

$$y = \mu(x) + \varepsilon$$

- ε is the **irreducible error** or **intrinsic variance**.
- $E = f(x) - \mu(x)$ is the **reducible error**.

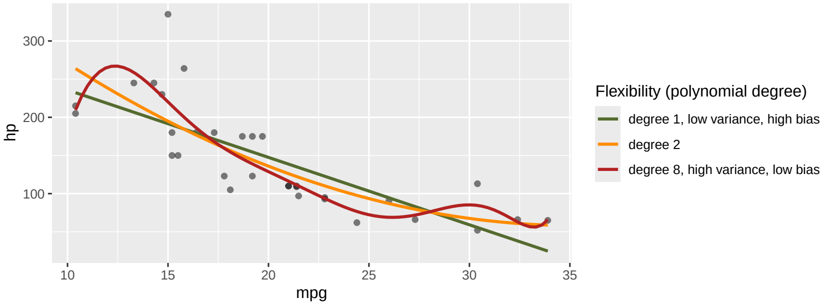
[1] requires perfect knowledge

Bias and Variance

$E = \text{Bias} + \text{Variance}$

- **Bias**: model can't change when it needs to. $\rightarrow$ **underfit**
- **Variance**: model adapts to match particular data. $\rightarrow$ **overfit**

bias-variance tradeoff $\leftarrow$ consideration for many model types



Breaking up the noise

Model based error decomposition:

- $\epsilon = \varepsilon + E$
- $E = \text{Bias} + \text{Variance}$

Observation based error decomposition:

- **Measurement error** - Unavoidable, but hopefully minimal.
  - Structured measurement error $\rightarrow$ problems to solve
  - eg. distance sampling
- **Process noise** - Natural demographic and environmental variability.
  - Minimized with large samples and stable environments.
  - The main input to the stochastic part of a model.

Conditional distributions

Alternative notation to $y = f(x) + \epsilon$:

$$Y \sim \mathbb{P}(f(X))$$

- $f(X)$: expected value of Y as a function of X.
- $\mathbb{P}$: probability distribution of the error

For example a linear model is:

$$Y \sim \text{Norm}(\text{mean} = \beta_0 + \beta_1 X, \text{sd} = \sigma)$$

The parameters of this model are:

- β_0 - intercept
- β_1 - slope
- σ^2 - residual variance (describes ϵ)