Motivating Example: Description

- Let's begin with a simple example:
  - Construct a model of outputs y based on inputs X
  - Interested in estimating/describing $y|X$ and quantifying uncertainty in that estimate
- These data are based on actual experiments performed on cyber system
  - X: Bandwidth of servers on the network
  - y: Throughput of data across the network

Motivating Example: Sample Data

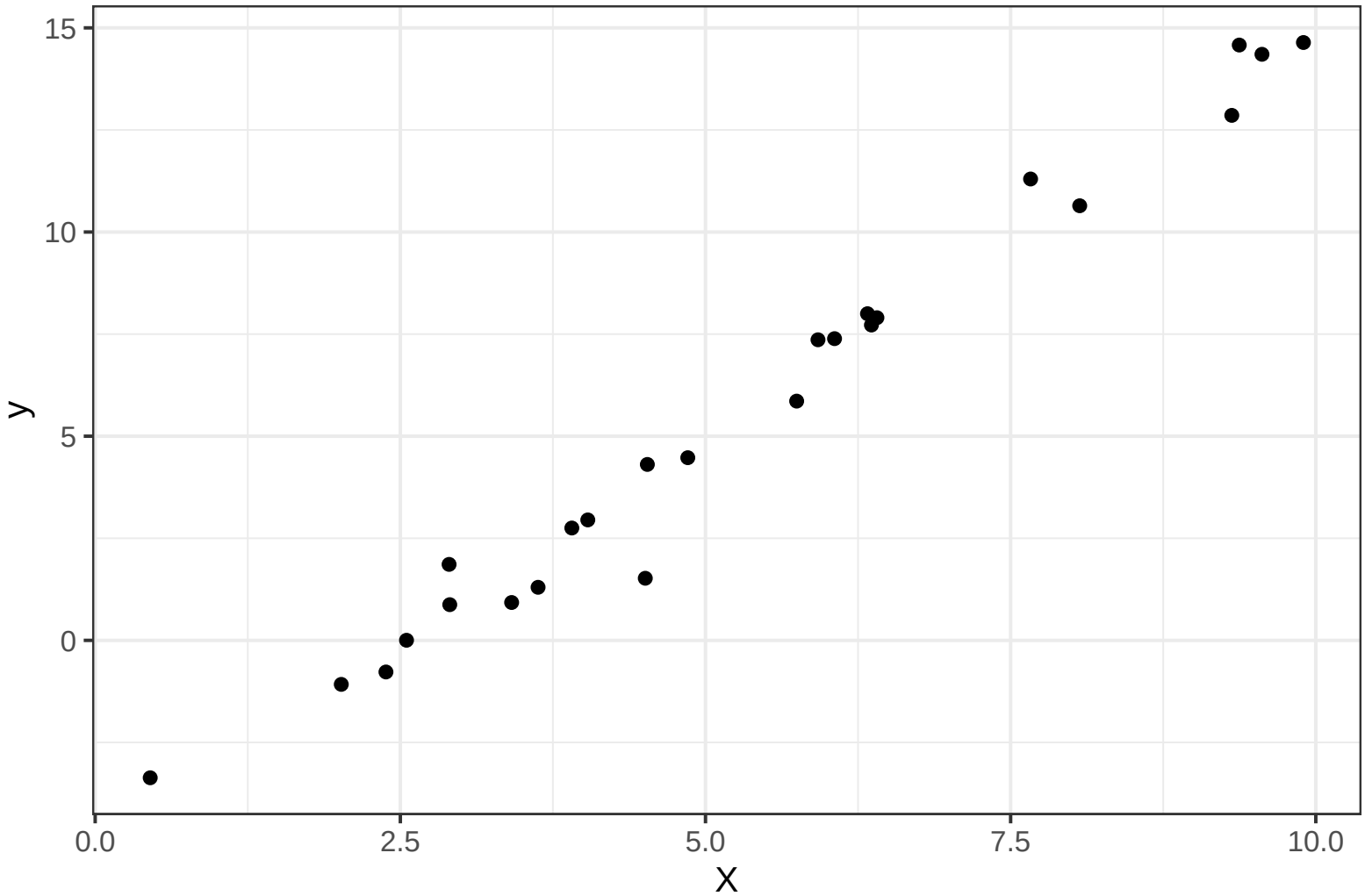


Figure: Sample Data

Motivating Example: Standard Approach

- Standard approach:
  1. Relationship between X and y looks linear
  2. Perform ordinary least squares (OLS) regression
  3. (Possibly) Examine diagnostics to check validity of assumptions

Motivating Example: Linear Regression

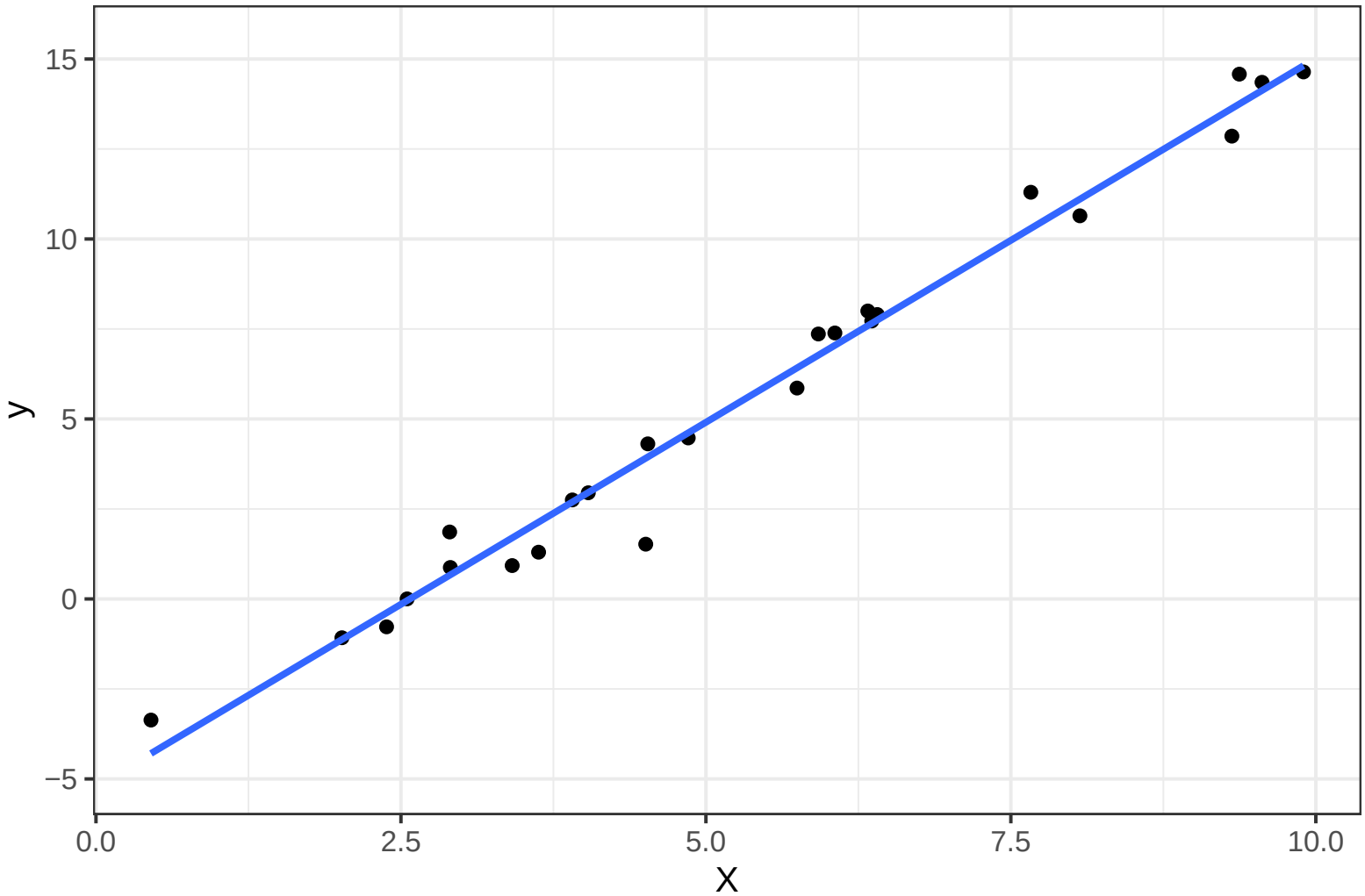


Figure: Fitted Sample Data

Motivating Example: Diagnostics

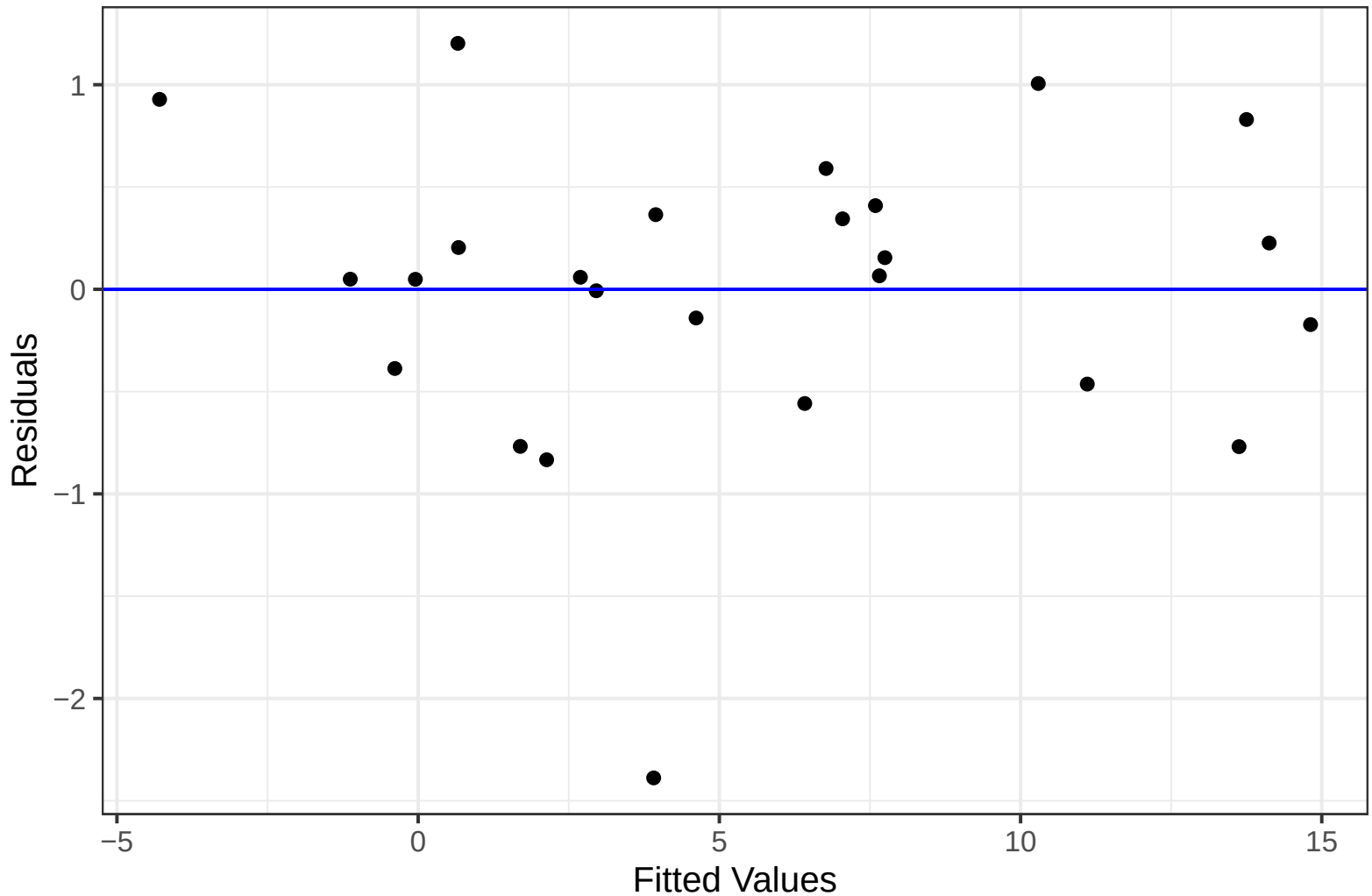


Figure: Residuals vs. Fitted Values

Motivating Example: Diagnostics

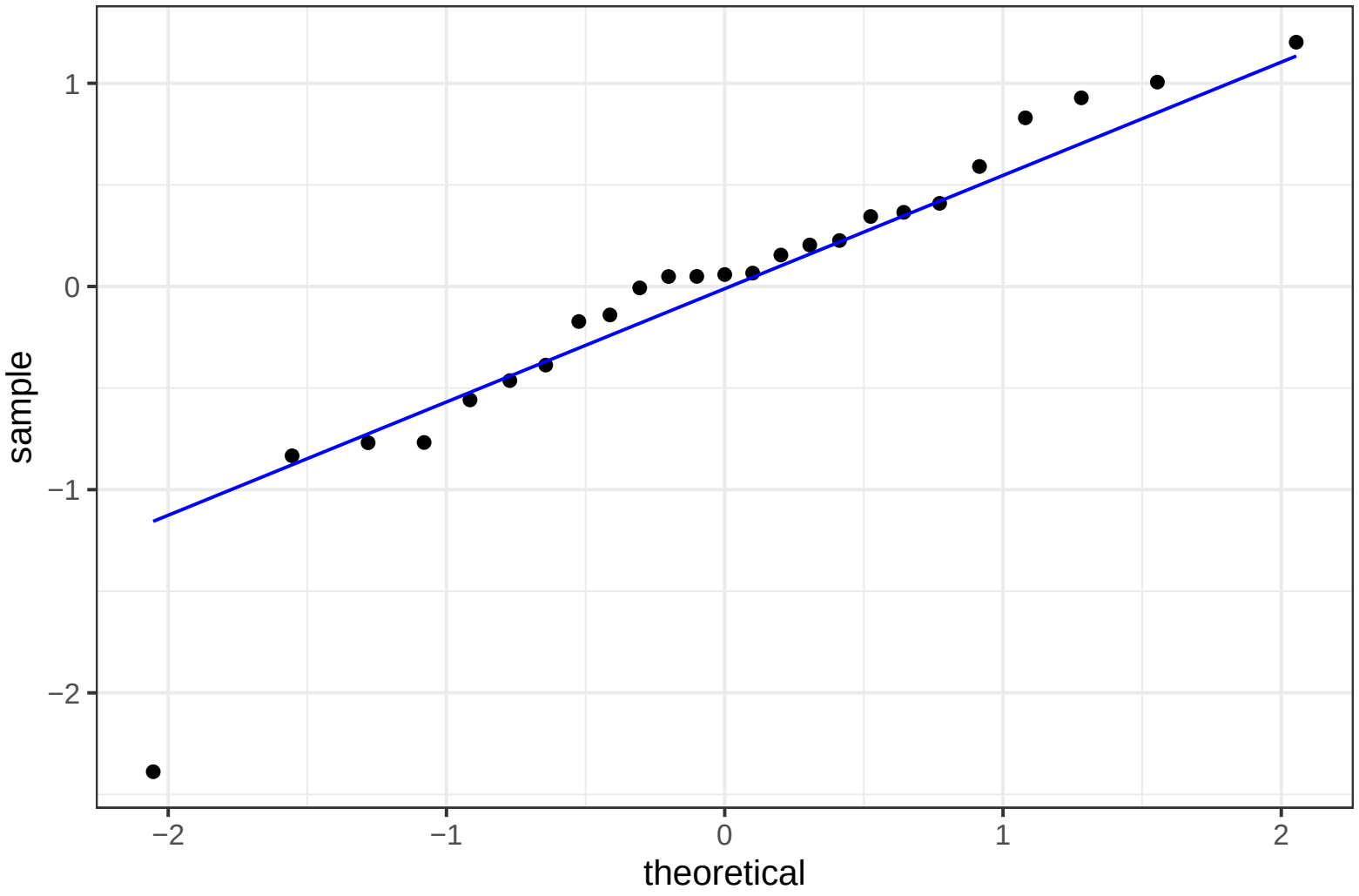


Figure: Normal Quantile Plot

Motivating Example: Standard Practice

- Many practitioners may not have run diagnostics
- If diagnostics were run, many practitioners will
  - use residuals vs. fitted values as evidence that errors are uncorrelated and homoskedastic (constant variance).
  - use quantile plot as evidence that errors are approximately normal, with exception of one outlier.
- An especially diligent practitioner may run test of normality of residuals
  - Shapiro-Wilk test p-value: 0.03917

Gauss-Markov Theorem

- However, most practitioners won't worry about normality
- Gauss-Markov theorem states OLS estimate is best linear unbiased estimate as long as
  - errors have mean zero
  - errors have constant finite variance
  - errors are uncorrelated
- While normality is needed for inference (e.g., significance, confidence intervals), it is not needed to produce unbiased parameter estimates or predictions

Finite Variance Assumption

- Gauss-Markov theorem requires that errors have constant **finite** variance
- Errors in the sample data are Cauchy distributed
  - Cauchy and normal distributions look *similar*
  - Cauchy distribution has fat tails, to the point where the mean and variance do not exist

Normal and Cauchy Distributions

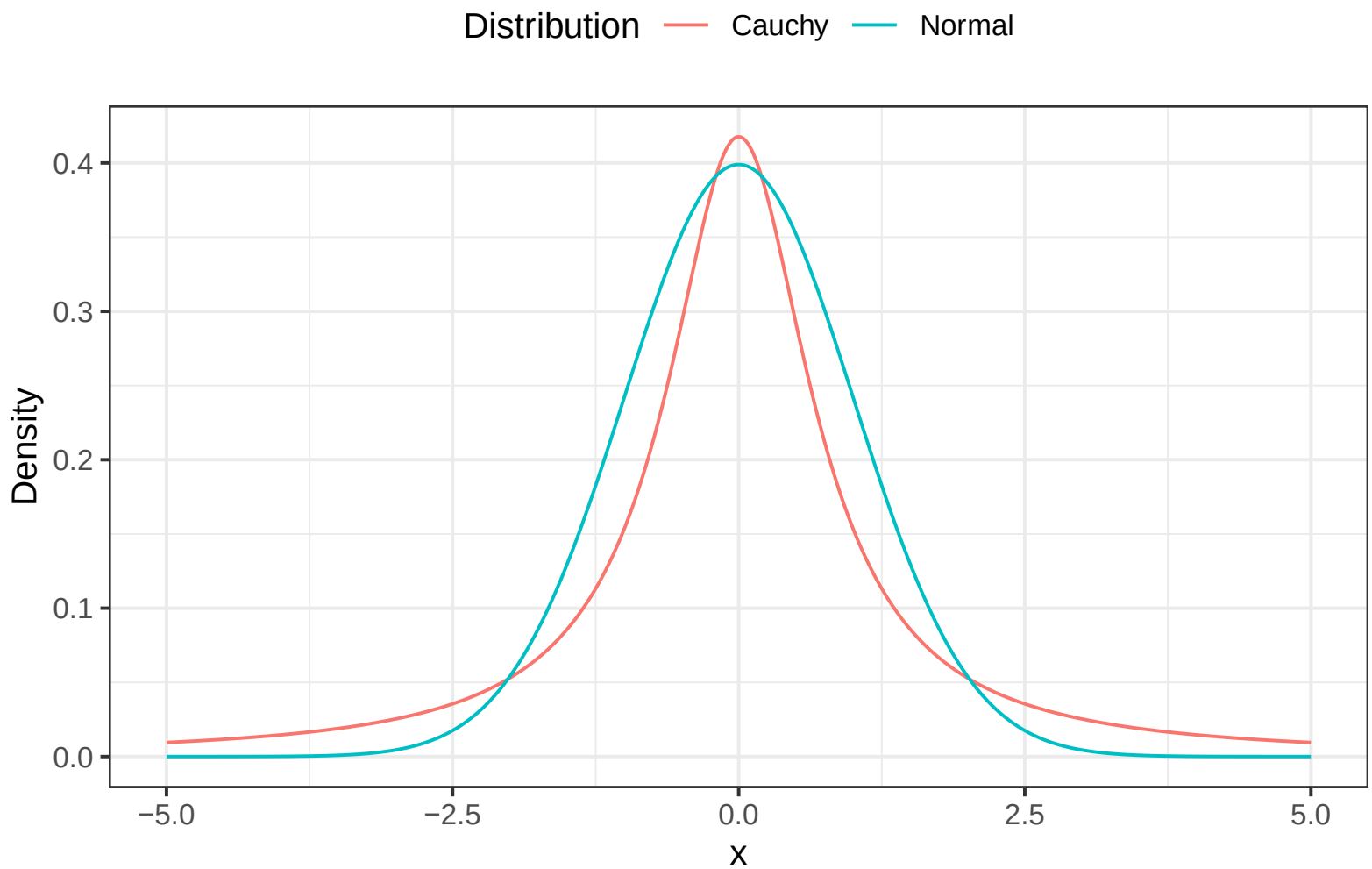
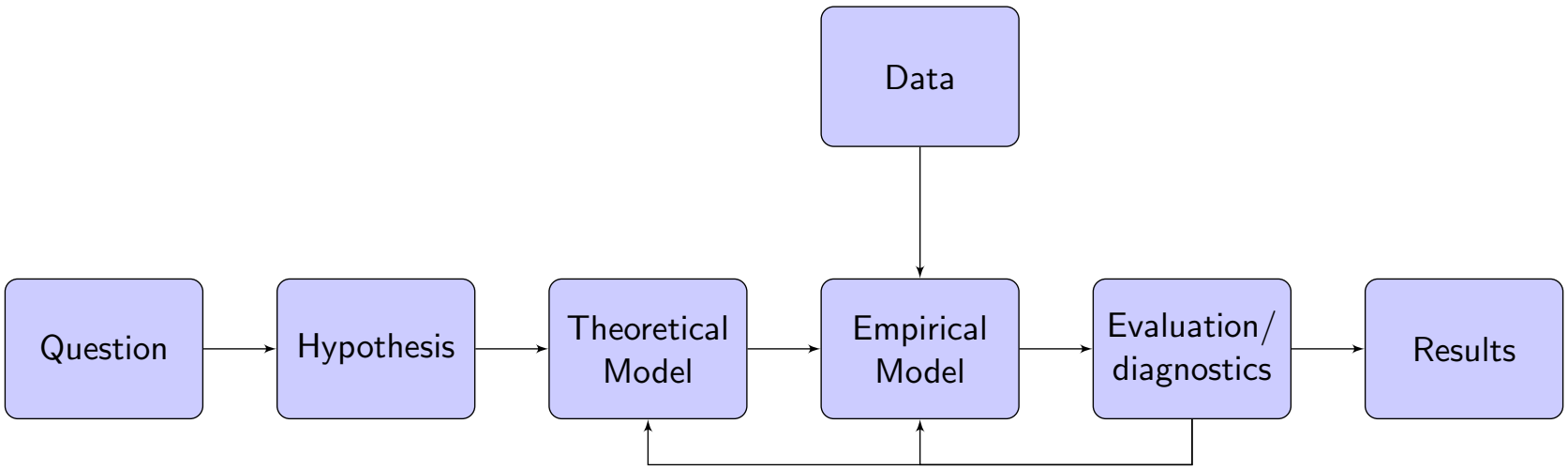


Figure: Normal and Cauchy Distributions

Infinite Variance Consequences

- When errors are Cauchy distributed
  - $E[y|X]$ does not exist, so cannot be estimated
  - Weak law of large numbers does not hold, so OLS parameter estimates do not converge to true values (or any other value) as sample size increases

Modeling Process



Replication Crisis

- In academic publishing, studies are reviewed for logical consistency and to ensure best practices are followed
- Peer review should enhance reproducibility of studies
- However, many published findings have proven difficult to replicate
  - Reviewers don't have time/resources to verify all details in theory/empirics
  - Often completely impossible to verify assumptions of empirical model (access to data/code, infeasible to re-run analysis)
  - Reviewers rarely see iterations on theory/empirics

Review of OUO/Classified Studies

- OUO and classified studies are rarely undergo any in-depth review
- When reviews are conducted, it can be especially difficult to obtain data and code used for analysis