Statistical models

DAAG Chapter 3

Objectives

- Learn about statistical models as signal + noise
- Introduction to fitting models to data
- Introduction to models for the noise component
- Learn about simulation from models and sampling
- Introduction to checking model assumptions

Statistical models

- Statistical models are often compartmentalized into a signal component and a noise component
  - Signal component is deterministic
  - Noise component is random
- Signal component is often meaningful, but sometimes not (curve fitting)
- Noise component can be due to variation in process and/or variation in observation

Statistical models

- Model with additive error

$$y_i = f(x_i) + \varepsilon_i$$

- Simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Exponential model

$$y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i$$

Models for noise

- Bernouilli
- Binomial
- Poisson
- Normal
- …

Bernouilli

- Probability of “success” = p
- Probability of “failure” = 1-p
- X = 1 if “success”, X = 0 if “failure”
- Example: Did it rain on a given day?

Binomial

- Binomial is a sum of independent Bernouillis
- Expected value is np
- Variance is np(1-p)
- Example: Number of days it rained last month (what assumptions are made?)

Poisson

- The number of events that occur in a certain (time or space) interval.
- The events occur independently.
- λ is the expected value (mean) and variance.
- Example: number of cars that pass by Jeffery Hall in one minute.

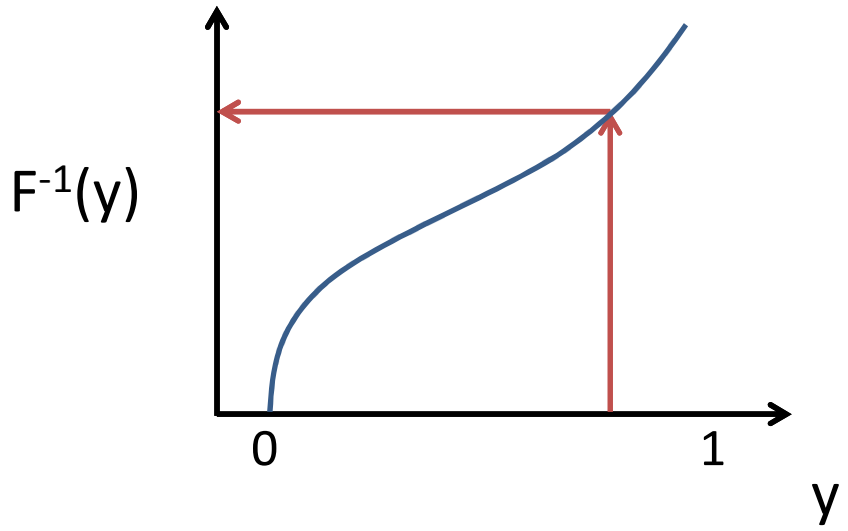
Normal

- Many processes follow a normal distribution, especially sums or averages (due to CLT).
- The distribution has nice properties, so it is used a lot in modeling.
- Parameters are mean μ and variance σ^2
- Example: Grades on an exam.
- Lognormal distribution is for when the log of a random variable follows a Normal. Often used for multiplicative processes.
  - Example: Return on an investment

Drawing from a distribution

- Generally, drawing from a statistical distribution in a computer involves two steps:

1. Draw a random uniform y~(0,1)
2. Map to a random variable quantity using the inverse cumulative distribution function $x \sim F^{-1}(y)$



Drawing from a distribution: resampling

- Suppose we have data that were drawn from an unknown distribution
  - Can't sample the distribution directly!
- Approach: Treat the sample as an approximation to the population (*empirical* distribution), and draw from the sample
- Known as resampling or bootstrapping

Simple bootstrapping

- Would like to know about the properties of a sample statistic $\Theta = g(\mathbf{X})$ (e.g. mean)
  - Don't know the distribution of the population, so use the data we have to approximate.

1. Draw n samples (with replacement!) from data, calculate sample statistic of interest Θ^*.
2. Repeat many (m) times to get m Θ^*
3. The resulting m Θ^* can be treated as samples for Θ.

Fitting a linear model

- Linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- Usually, these are fit using a least-squares criterion. That means we minimize $\sum_i \varepsilon_i^2$. The solution is $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$.
- We often assume that $\varepsilon_i \sim N(0, \sigma^2)$ for all *i.*
- In R, the lm() function will fit this model
- In SAS, PROC REG will fit this model
- Other criteria are possible, notably maximum likelihood

Fitting non-linear models

- In the non-linear case, $\boldsymbol{y} = g(X, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$
- We can again use a least-squares criterion, or possibly some other criterion. Call $h(\boldsymbol{y}, g(X, \boldsymbol{\beta}))$ the objective function.
- Our goal is to minimize this function by varying $\boldsymbol{\beta}$
- In R you can use optim() or optimize()
- In SAS you can use PROC IML

Checking statistical assumptions

- Fit a linear model…
  - Want to check the assumption that $\varepsilon_i \sim N(0, \sigma^2)$ for all *i.*
- Check on distributional assumption
- Check on homoscedasticity
- Check on independence

Quantile-Quantile plots