coaching actuaries

MAS-I

Updated 4/9/24

Probability Models

Basics

CDFs, Survival Functions, and Hazard Functions

$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$

$S(x) = \Pr(X > x) = \int_{x}^{\infty} f(t)\,\mathrm{d}t$

$h(x) = \dfrac{f(x)}{S(x)}$

$H(x) = \int_{-\infty}^{x} h(t)\,\mathrm{d}t = -\ln S(x)$

$S(x) = e^{-H(x)}$

Percentiles

$100q^{\text{th}}$ percentile is π_q where $F(\pi_q) = q$.

Mode

Mode is x that maximizes $f(x)$.

Moments

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\cdot f(x)\,\mathrm{d}x$$
$$= \int_{0}^{\infty} g'(x)\cdot S(x)\,\mathrm{d}x$$

$\mathrm{Var}[g(X)] = \mathrm{E}[g(X)^2] - \mathrm{E}[g(X)]^2$

$\mu'_k = \mathrm{E}[X^k]$

$\mu = \mu'_1 = \mathrm{E}[X]$

$\mu_k = \mathrm{E}[(X-\mu)^k]$

$\sigma^2 = \mu_2 = \mathrm{Var}[X]$

$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\cdot\mathrm{E}[Y]$

$\mathrm{Cov}[X,X] = \mathrm{Var}[X]$

Coefficient of variation, $CV = \dfrac{\sigma}{\mu}$

Skewness $= \dfrac{\mu_3}{\sigma^3}$

Kurtosis $= \dfrac{\mu_4}{\sigma^4}$

Moment Generating Function (MGF)

$M_X(z) = \mathrm{E}[e^{zX}]$

$M_X^{(n)}(0) = \mathrm{E}[X^n]$

where $M_X^{(n)}$ is the n^{th} derivative

Probability Generating Function (PGF)

$P_X(z) = \mathrm{E}[z^X]$

$P_X^{(n)}(1) = \mathrm{E}[X(X-1)\dots(X-n+1)]$

where $P_X^{(n)}$ is the n^{th} derivative

Conditional Distribution

$\Pr(A \mid B) = \dfrac{\Pr(A\cap B)}{\Pr(B)} = \dfrac{\Pr(B\mid A)\Pr(A)}{\Pr(B)}$

$f_{X\mid j<X<k}(x) = \dfrac{f_X(x)}{\Pr(j<X<k)}$

where $j < x < k$

Law of Total Probability

$\Pr(X = x) = \mathrm{E}_Y[\Pr(X = x \mid Y)]$

Law of Total Expectation

$\mathrm{E}_X[X] = \mathrm{E}_Y\big[\mathrm{E}_X[X\mid Y]\big]$

Law of Total Variance

$$\mathrm{Var}_X[X] = \mathrm{E}_Y\big[\mathrm{Var}_X[X\mid Y]\big] + \mathrm{Var}_Y\big[\mathrm{E}_X[X\mid Y]\big]$$

Independence

$\Pr(A\cap B) = \Pr(A)\cdot\Pr(B)$

For independent X and Y:

- $f_{X,Y}(x,y) = f_X(x)\cdot f_Y(y)$
- $\mathrm{E}[g(X)\cdot h(Y)] = \mathrm{E}[g(X)]\cdot\mathrm{E}[h(Y)]$

Claim Severity Distributions

Common Distributions

S-P Pareto$(\alpha,\theta) \sim$ Pareto$(\alpha,\theta)+\theta$

Beta$(a=1, b=1, \theta) \sim$ Uniform$(0,\theta)$

Weibull$(\theta, \tau=1) \sim$ Exponential(θ)

Gamma$(\alpha=1,\theta) \sim$ Exponential(θ)

Gamma CDF Shortcut

$F_X(x) = 1 - \Pr(N < \alpha)$

- α is a positive integer
- $X \sim$ Gamma(α,θ)
- $N \sim$ Poisson(x/θ)

Properties of Exponential Distribution

$X_i \sim$ Exponential(θ_i)

$\mathrm{E}[X] = \theta$

$h(x) = 1/\theta = \lambda$

$\Pr(X > t+s \mid X > t) = \Pr(X > s)$

$\Pr(X_1 < X_2) = \dfrac{\lambda_1}{\lambda_1+\lambda_2}$

$\min(X_1, X_2, \dots, X_n) \sim \text{Exponential}\left(\dfrac{1}{\sum_{i=1}^{n}\lambda_i}\right)$

$\sum_{i=1}^{n} X_i \sim$Gamma(n,θ) where $\theta_i = \theta$

Greedy Algorithms

Algorithm A

For $i = 1, 2, \dots, n$:

1. Choose the assignment with the lowest cost, i.e., $\min_j C_{i,j}$, among all $n-i+1$ possible assignments.
2. Assign that job to that employee.
3. Remove that employee and that job from their respective sets.

Algorithm B

For $k = n^2, (n-1)^2, \dots, 1^2$:

1. Choose the assignment with the lowest cost, i.e., $\min_{i,j} C_{i,j}$, among all k possible assignments.
2. Assign that job to that employee.
3. Remove that employee and that job from their respective sets.

$\mathrm{E}[\text{Total Cost}] = \theta\sum_{i=1}^{n}\dfrac{1}{i}$

where $C_{i,j} \sim$ Exponential(θ)

Transformations

- Scaling
  θ is a scale parameter for all continuous distributions on the exam table, except lognormal, inverse Gaussian, and log-t.
- CDF Method
- PDF Method
- MGF Method

Mixtures

Discrete Mixture

$$f_Y(y) = \sum_{i=1}^{n} w_i \cdot f_{X_i}(y)\text{, where } \sum_{i=1}^{n} w_i = 1$$

$$F_Y(y) = \sum_{i=1}^{n} w_i \cdot F_{X_i}(y)$$

$$S_Y(y) = \sum_{i=1}^{n} w_i \cdot S_{X_i}(y)$$

$$\mathrm{E}[Y^k] = \sum_{i=1}^{n} w_i \cdot \mathrm{E}[X_i^k]$$

Continuous Mixture

- *Poisson-Gamma Mixture*
  $X|\Lambda \sim \text{Poisson}(\Lambda)$
  $\Lambda \sim \text{Gamma}(\alpha, \theta)$
  $X \sim \text{Negative Binomial}(r = \alpha, \beta = \theta)$
- *Exponential-Gamma Mixture*
  $X|\Lambda \sim \text{Exponential}(\Lambda)$
  $\Lambda \sim \text{Inverse Gamma}(\alpha, \theta)$
  $X \sim \text{Pareto}(\alpha, \theta)$

Splices

$$f_Y(y) = \begin{cases} c_1 \cdot f_{X_1}(y), & a_0 < y < a_1 \\ c_2 \cdot f_{X_2}(y), & a_1 < y < a_2 \\ \vdots & \vdots \\ c_n \cdot f_{X_n}(y), & a_{n-1} < y < a_n \end{cases}$$

where $\sum_{i=1}^{n} c_i$ does not need to equal 1.

Bernoulli Shortcut

$\mathrm{Var}[X] = (a-b)^2 q(1-q)$

where $X = \begin{cases} a, & \Pr(X=a) = q \\ b, & \Pr(X=b) = 1-q \end{cases}$

Insurance Applications

Y^L: payment per loss

Policy Limits, u

$$Y^L = X \wedge u = \min(X, u) = \begin{cases} X, & X < u \\ u, & X \ge u \end{cases}$$

$$\begin{aligned} \mathrm{E}[(Y^L)^k] &= \mathrm{E}[(X \wedge u)^k] \\ &= \int_0^u x^k f(x)\,\mathrm{d}x + u^k \cdot S(u) \\ &= \int_0^u k x^{k-1} S(x)\,\mathrm{d}x \end{aligned}$$

Deductibles, d

Ordinary deductible:

$$Y^L = (X-d)_+ = \begin{cases} 0, & X < d \\ X - d, & X \ge d \end{cases}$$

$$\mathrm{E}[Y^L] = \mathrm{E}[(X-d)_+] = \mathrm{E}[X] - \mathrm{E}[X \wedge d]$$

$$\begin{aligned} \mathrm{E}[(Y^L)^k] &= \mathrm{E}\left[(X-d)_+^k\right] \\ &= \int_d^\infty (x-d)^k f(x)\,\mathrm{d}x \\ &= \int_d^\infty k(x-d)^{k-1} S(x)\,\mathrm{d}x \end{aligned}$$

Loss elimination ratio:

$$\mathrm{LER} = \frac{\mathrm{E}[X \wedge d]}{\mathrm{E}[X]}$$

Franchise deductible:

$$Y^L = \begin{cases} 0, & X < d \\ X, & X \ge d \end{cases}$$

$$\mathrm{E}[Y^L] = \mathrm{E}[(X-d)_+] + d \cdot S(d)$$

Payment per Payment

Y^P: payment per payment

$$\begin{aligned} \mathrm{E}[Y^P] &= e(d) = \mathrm{E}[\,X - d \mid X > d\,] \\ &= \frac{\mathrm{E}[Y^L]}{S(d)} = \frac{\mathrm{E}[(X-d)_+]}{S(d)} \end{aligned}$$

Special Cases for $e(d)$

Loss	*Excess Loss*
Exponential(θ)	Exponential(θ)
Uniform(a, b)	Uniform$(0, b-d)$
Pareto(α, θ)	Pareto$(\alpha, \theta + d)$
S-P Pareto(α, θ)	Pareto(α, d)
Beta$(1, b, \theta)$	Beta$(1, b, \theta - d)$

Impact of Deductibles on Claim Frequency

For $v = \Pr(X > d)$, the # of payments N':

	N	N'
Poisson	λ	$v\lambda$
Binomial	m, q	m, vq
Neg. Binomial	r, β	$r, v\beta$

The Ultimate Formula for Insurance

$$\mathrm{E}[Y^L] = \alpha(1+r)\left(\mathrm{E}\left[X \wedge \frac{m}{1+r}\right] - \mathrm{E}\left[X \wedge \frac{d}{1+r}\right]\right)$$

where

d: deductible (set to 0 if not applicable)

u: policy limit (set to ∞ if not applicable)

α: coinsurance (set to 1 if not applicable)

r: inflation rate (set to 0 if not applicable)

m: maximum covered loss $= \dfrac{u}{\alpha} + d$

Tail Properties of Distributions

q quantile

$\pi_q = F_X^{-1}(q)$

Conditional Tail Expectation (CTE)

$1 - q$: tolerance probability

$$\begin{aligned} \mathrm{CTE}_q(X) &= \mathrm{E}\left[X \mid X > \pi_q\right] \\ &= \pi_q + \frac{\mathrm{E}[X] - \mathrm{E}[X \wedge \pi_q]}{1-q} \end{aligned}$$

	$\mathrm{CTE}_q(X)$
Normal	$\mu + \sigma\left[\dfrac{\phi(z_q)}{1-q}\right]$
Lognormal	$\mathrm{E}[X] \cdot \left[\dfrac{\Phi(\sigma - z_q)}{1-q}\right]$

Tail Weight

- The fewer positive raw moments that exist, the greater the tail weight.
- If the ratio of the survival functions or the density functions approaches infinity as x increases, the numerator has a heavier tail.
- If the hazard rate function decreases with x, the distribution has a heavy tail.
- The larger a given CTE or quantile is, the greater the tail weight.

Poisson Processes

Counting process where non-overlapping Poisson increments are independent

$N(t+h) - N(t) \sim \text{Poisson}(\lambda)$

where $\lambda = \int_t^{t+h} \lambda(u)\,\mathrm{d}u$

- Homogeneous if $\lambda(t)$ is constant
- Non-homogeneous if $\lambda(t)$ varies with t

Time between Events

T_k : Time until the k^{th} event occurs

$V_k = T_k - T_{k-1}$

Homogeneous Poisson process:

- $V_k \sim \text{Exponential}(\theta = 1/\lambda)$
- $T_k \sim \text{Gamma}(\alpha = k, \theta = 1/\lambda)$

Conditional Distribution of Arrival Times

- Given that $N(t) = n$, past events $T_1, T_2, \dots, T_n$ are order statistics of i.i.d. $\text{Uniform}(0, t)$.
- Given that $T_n = t$, past events $T_1, T_2, \dots, T_{n-1}$ are order statistics of i.i.d. $\text{Uniform}(0, t)$.

Other Properties

- Subprocesses are Poisson processes with proportional rates.
- Sum of Poisson processes:
  $$\sum_{i=1}^{n} N_i \sim \text{Poisson}\left(\sum_{i=1}^{n} \lambda_i\right)$$
- Probability of observing n events from N_1 before m events from N_2 is:
  $$\sum_{i=n}^{n+m-1} \binom{n+m-1}{i} q^i (1-q)^{n+m-1-i}$$
  $$\sum_{j=0}^{m-1} \binom{n-1+j}{n-1} q^n (1-q)^j$$
  where $q = \dfrac{\lambda_1}{\lambda_1 + \lambda_2}$

Compound Poisson Processes

$$S(t) = \sum_{i=1}^{N(t)} X_i$$

$\mathrm{E}[S(t)] = \lambda t \cdot \mathrm{E}[X]$

$\mathrm{Var}[S(t)] = \lambda t \cdot \mathrm{E}[X^2]$

- Use normal approximation to calculate probabilities of events in $S(t)$.
- Continuity correction is needed if $S(t)$ is discrete.

Reliability Theory*

- A parallel system functions as long as one of the components functions.
- A series system functions only when all components function.
- A k-out-of-n system functions only when at least k out of n components function.
- A minimal path set, A_j, is a minimal set of components whose functioning guarantees the functioning of the system.
- A minimal cut set, C_j, is a minimal set of components whose failure guarantees the failure of the system.

Combining Systems

	Placement of Systems	*Action*
# of Minimal Path Sets	Parallel	Sum
	Series	Product
# of Minimal Cut Sets	Parallel	Product
	Series	Sum

Reliability of Systems

$r(\mathbf{p}) = \Pr[\phi(\mathbf{X}) = 1] = \mathrm{E}[\phi(\mathbf{X})]$

Bounds on Reliability Function

Method of Inclusion and Exclusion:

First two bounds using minimal path sets:

$$r(\mathbf{p}) \le \sum_{j=1}^{s} \left(\prod_{i \in A_j} p_i\right)$$

$$r(\mathbf{p}) \ge \sum_{j=1}^{s} \left(\prod_{i \in A_j} p_i\right) - \sum_{j=1}^{s} \sum_{k>j} \left(\prod_{i \in A_j \cup A_k} p_i\right)$$

First two bounds using minimal cut sets:

$$1 - r(\mathbf{p}) \le \sum_{j=1}^{m} \left(\prod_{i \in C_j} (1 - p_i)\right)$$

$$1 - r(\mathbf{p}) \ge \sum_{j=1}^{m} \left(\prod_{i \in C_j} (1 - p_i)\right) - \sum_{j=1}^{m} \sum_{k>j} \left(\prod_{i \in C_j \cup C_k} (1 - p_i)\right)$$

Method of Intersection:

$$r(\mathbf{p}) \le 1 - \prod_{j=1}^{s} \left[1 - \prod_{i \in A_j} p_i\right]$$

$$r(\mathbf{p}) \ge \prod_{j=1}^{m} \left[1 - \prod_{i \in C_j} (1 - p_i)\right]$$

Random Graphs

- n^{n-2} minimal path sets
- $2^{n-1} - 1$ minimal cut sets
- $P_{i,j}$ is the probability nodes i and j are connected.
- P_n is the probability that a random graph is connected, where all $P_{i,j} = p$.

$$P_n = \begin{cases} 1, & n = 1 \\ p, & n = 2 \\ 1 - \displaystyle\sum_{k=1}^{n-1} \binom{n-1}{k-1} q^{k(n-k)} P_k, & n > 2 \end{cases}$$

$1 - P_n \le (n+1) q^{n-1}$

$1 - P_n \ge n q^{n-1} - \binom{n}{2} q^{2n-3}$

$P_n \approx 1 - n q^{n-1}$

Lifetime of Systems

$\Pr(T > t) = r[\mathbf{S}(t)]$

$$\mathrm{E}[T] = \int_0^{\infty} r[\mathbf{S}(t)]\,\mathrm{d}t$$

For k-out-of-n systems whose components are $r_i \sim \text{Exponential}(\theta)$:

$$\mathrm{E}[T] = \mathrm{E}\left[X_{(n-k+1)}\right] = \theta \sum_{i=k}^{n} \frac{1}{i}$$

Increasing Failure Rate (IFR) Distribution

$h(x)$ is an increasing function of x.

Decreasing Failure Rate (DFR) Distribution

$h(x)$ is a decreasing function of x.

Increasing Failure on the Average (IFRA)

$H(x)/x$ is an increasing function of x.

- An IFR distribution is also IFRA.
- A monotone system's lifetime distribution is IFRA if the lifetimes of all components are IFRA.

*Key information on Reliability Theory is on page 5.

Discrete Markov Chains

Multiple-Step Transition Probabilities

- Chapman-Kolmogorov Probabilities

$$P_{i,j}^{n+m} = \sum_{k=1}^{\infty} P_{i,k}^{n} P_{k,j}^{m}$$

- Unconditional probability of being in state j at time n:

$$\Pr(X_n = j) = \sum_{i=1}^{\infty} \alpha_i P_{i,j}^{n}$$

where α_i is the probability of being in state i at time 0.

- The probability of entering state j at time m, starting at state i without entering any state in set $\mathcal{A}$:

State i	*State j*	*Desired Probability*
$i \notin \mathcal{A}$	$j \notin \mathcal{A}$	$Q_{i,j}^{m}$
$i \notin \mathcal{A}$	$j \in \mathcal{A}$	$\sum_{r \notin \mathcal{A}} Q_{i,r}^{m-1} P_{r,j}$
$i \in \mathcal{A}$	$j \notin \mathcal{A}$	$\sum_{r \notin \mathcal{A}} P_{i,r} Q_{r,j}^{m-1}$
$i \in \mathcal{A}$	$j \in \mathcal{A}$	$\sum_{r \notin \mathcal{A}} \sum_{k \notin \mathcal{A}} P_{i,r} Q_{r,k}^{m-2} P_{k,j}$

where:

$Q_{i,j} = P_{i,j}$, if $i \notin \mathcal{A}, j \notin \mathcal{A}$

$Q_{i,A} = \sum_{j \in \mathcal{A}} P_{i,j}$ if $i \notin \mathcal{A}$

$Q_{A,i} = 0$ if $i \notin \mathcal{A}$

$Q_{A,A} = 1$

Classification of States

- *Absorbing*: State that cannot be left once it is entered
- *Accessible*: State that can be entered from another state
- *Communicating*: Two states are accessible to each other
- *Class*: A set of communicating states
- *Irreducible*: A chain with only one class
- *Recurrent*: Probability of re-entering state is 1, $f_i = 1$
- *Transient*: Probability of re-entering state is less than 1, $f_i < 1$
  - Given that a process starts in a transient state i, the number of times the process re-enters state i, $n \geq 0$, has a geometric distribution with $\beta = \frac{f_i}{1-f_i}$
- *Positive recurrent*: Finite expected # of transitions for a chain to return to state j given it started in that state
- *Null recurrent*: Infinite expected # of transitions for a chain to return to state j given it started in that state
- *Aperiodic*: A chain that has limiting probabilities
- *Periodic*: A chain that does not have limiting probabilities
- *Ergodic*: A chain that is irreducible, positive recurrent, and aperiodic

Long-Run Proportions (Stationary Probabilities)

$$\pi_j = \sum_{i=1}^{n} \pi_i P_{i,j} \quad , \quad \sum_{j=1}^{n} \pi_j = 1$$

- The reciprocal of π_j is the expected time spent to return to state j.
- For aperiodic chains, long-run proportions equal limiting probabilities.

Time Spent in Transient States

$\mathbf{S} = (\mathbf{I} - \mathbf{P}_T)^{-1}$

$$f_{i,j} = \frac{s_{i,j} - \delta_{i,j}}{s_{j,j}}$$

$$\delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases}$$

- $s_{i,j}$ is the expected time spent in state j given it starts in state i.
- $f_{i,j}$ is the probability of ever transitioning to state j from state i.

Time Reversibility

$$R_{i,j} = \frac{\pi_j P_{j,i}}{\pi_i}$$

A Markov chain is time reversible if $R_{i,j} = P_{i,j}$ for every i and j.

Random Walk

All random walk models are transient except for one-dimensional and two-dimensional symmetric random walks.

Gambler's Ruin Problem

Probability of reaching j starting with i is:

$$P_i = \begin{cases} \dfrac{1-(q/p)^i}{1-(q/p)^j}, & p \neq \dfrac{1}{2} \\ \dfrac{i}{j}, & p = \dfrac{1}{2} \end{cases}$$

Branching Processes

$$\mu = \sum_{j=0}^{\infty} j \cdot P_j$$

$$\sigma^2 = \sum_{j=0}^{\infty} (j-\mu)^2 \cdot P_j$$

For $X_0 = 1$:

$\mathrm{E}[X_n] = \mu^n$

$$\mathrm{Var}[X_n] = \begin{cases} \sigma^2 \mu^{n-1} \left(\dfrac{1-\mu^n}{1-\mu} \right), & \mu \neq 1 \\ n\sigma^2, & \mu = 1 \end{cases}$$

$$\pi_0 = \begin{cases} 1, & \mu \leq 1 \\ \sum_{j=0}^{\infty} \pi_0^j P_j, & \mu > 1 \end{cases}$$

Life Contingencies

Number of Deaths

$d_x = l_x - l_{x+1}$

Probability of Survival

$${}_tp_x = \frac{l_{x+t}}{l_x}$$

Probability of Death

$${}_tq_x = \frac{l_x - l_{x+t}}{l_x}$$

Curtate Life Expectancy

$$e_x = \sum_{k=1}^{\infty} {}_kp_x$$
$$= p_x(1 + e_{x+1})$$

Complete Expectation of Life

$$0.5 + \sum_{k=1}^{\infty} {}_kp_x$$

Whole Life Insurance

$$A_x = \sum_{k=0}^{\infty} v^{k+1} \cdot {}_kp_x \cdot q_{x+k}$$
$$= vq_x + vp_xA_{x+1}$$

- The APV of whole life insurance is the sum of the APV of term life and deferred whole life.
- The APV of endowment insurance is the sum of the APV of term life and pure endowment.

Whole Life Annuity

$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k \cdot {}_kp_x$$
$$= 1 + vp_x \cdot \ddot{a}_{x+1}$$
$$= \frac{1 - A_x}{d}$$

Mortality Discount Factor

$${}_tE_x = v^t\ {}_tp_x$$

Joint Lives

$$\ddot{a}_x + \ddot{a}_y = \ddot{a}_{xy} + \ddot{a}_{\overline{xy}}$$

Equivalence Principle

$$\text{APV}_{\text{Premium}} = \text{APV}_{\text{Benefit}}$$

Simulation

$U \sim \text{Uniform}(0, 1)$

Uniform Number Generation

$$X_{n+1} = (aX_n + c) \bmod m, \qquad n \geq 0$$
$$U = \frac{X_{n+1}}{m}$$

Inversion Method

$X = F_X^{-1}(U)$

Acceptance-Rejection Method

1. Find constant c that satisfies:
$$\frac{f(x)}{g(x)} \leq c, \qquad \text{for all } x$$
2. Simulate U and a random number Y with density function g.
3. Accept the value Y if
$$U \leq \frac{f(Y)}{cg(Y)}$$
Otherwise, reject and return to step 2.

Key Information for Reliability Theory

	$\phi(\mathbf{x})$	# of Minimal Path Sets	# of Minimal Cut Sets	$r(\mathbf{p})$
Parallel	$\max(x_i) = 1 - \prod_{i=1}^{n}(1 - x_i)$	n	1	$1 - \prod_{i=1}^{n}(1 - p_i)$
Series	$\min(x_i) = \prod_{i=1}^{n} x_i$	1	n	$\prod_{i=1}^{n} p_i$
k-out-of-n	–	$\binom{n}{k}$	$\binom{n}{n-k+1}$	$\sum_{i=k}^{n}\binom{n}{i} p^i(1-p)^{n-i}$ where $p_i = p$ for all i
Minimal Path Sets	$\max_j \prod_{i \in A_j} x_i$	–	–	–
Minimal Cut Sets	$\prod_{j=1}^{m} \max_{i \in C_j} x_i$	–	–	–

Statistics

Parameter and Density Estimation

Method of Moments

To fit an r-parameter distribution, set:

$$\mathrm{E}[X^k] = \frac{\sum_{i=1}^{n} x_i^k}{n}, \qquad k = 1, 2, \ldots, r$$

Percentile Matching

- Estimate parameters by setting the theoretical percentiles equal to the sample percentiles

Smoothed Empirical Percentile – Unique Values

$\hat{\pi}_q = [q(n+1)]^{\text{th}}$ smallest observed value

- If $q(n+1)$ is a non-integer, calculate $\hat{\pi}_q$ by interpolating between the order statistics before and after.

Maximum Likelihood Estimation

$$L(\theta) = \prod_{i=1}^{n} f(x_i)$$

- Estimate θ as the value that maximizes $L(\theta)$ or $l(\theta) = \ln L(\theta)$
- Invariance property

Incomplete Data

Case	*Likelihood*
Right-censored at m	$\Pr(X \geq m)$
Left-truncated at d	$\dfrac{f(x)}{\Pr(X > d)}$
Grouped data on interval $(a, b]$	$\Pr(a < X \leq b)$

Special Cases – Complete Data

Distribution	*Shortcut*
Gamma, fixed α	$\hat{\theta} = \dfrac{\bar{x}}{\alpha}$
Normal	$\hat{\mu} = \bar{x}$ $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} x_i^2}{n} - \hat{\mu}^2$
Lognormal	$\hat{\mu} = \dfrac{\sum_{i=1}^{n} \ln x_i}{n}$ $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (\ln x_i)^2}{n} - \hat{\mu}^2$
Poisson	$\hat{\lambda} = \bar{x}$
Binomial, fixed m	$\hat{q} = \dfrac{\bar{x}}{m}$
Negative Binomial, fixed r	$\hat{\beta} = \dfrac{\bar{x}}{r}$
Uniform $[0, \theta]$	$\hat{\theta} = \max(x_1, \ldots, x_n)$

Special Cases – Incomplete Data

Pareto, fixed θ
$\hat{\alpha} = \dfrac{n}{\sum_{i=1}^{n+c} [\ln(x_i + \theta) - \ln(d_i + \theta)]}$
S-P Pareto, fixed θ
$\hat{\alpha} = \dfrac{n}{\sum_{i=1}^{n+c} \{\ln x_i - \ln[\max(\theta, d_i)]\}}$
Exponential
$\hat{\theta} = \dfrac{\sum_{i=1}^{n+c} (x_i - d_i)}{n}$
Weibull, fixed τ
$\hat{\theta} = \left(\dfrac{\sum_{i=1}^{n+c} x_i^\tau - \sum_{i=1}^{n+c} d_i^\tau}{n}\right)^{1/\tau}$

where:

- n: # of uncensored data points
- c: # of censored data points
- x_i: i^{th} observed value, or the censoring point for censored data points
- d_i: truncation point for the i^{th} observation

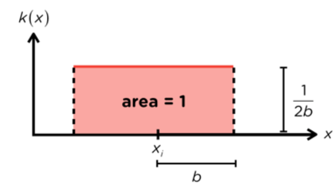
Kernel Density Estimation

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} k_i(x)$$

- b: Bandwidth
- x_i: i^{th} observed value
- $k_i(x)$: Kernel density function for x_i, evaluated at x
- $\tilde{f}(x)$: PDF of the kernel-smoothed distribution

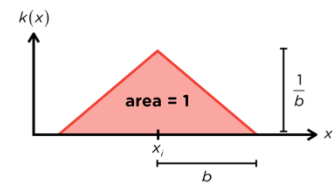
Rectangular Kernels

$$k_i(x) = \begin{cases} \dfrac{1}{2b}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{otherwise} \end{cases}$$



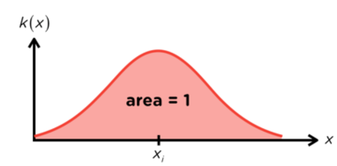
Triangular Kernels

$$k_i(x) = \begin{cases} \dfrac{b - |x - x_i|}{b^2}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{otherwise} \end{cases}$$



Gaussian Kernels

$$k_i(x) = \frac{1}{b\sqrt{2\pi}} \exp\left[-\frac{(x - x_i)^2}{2b^2}\right], \quad -\infty < x < \infty$$



Estimator Quality

Statistics and Estimators

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

For a random sample:

- $\mathrm{E}[\bar{X}] = \mathrm{E}[X]$
- $\mathrm{Var}[\bar{X}] = \frac{\mathrm{Var}[X]}{n}$

Bias

$$\mathrm{Bias}[\hat{\theta}] = \mathrm{E}[\hat{\theta}] - \theta$$

- If $\lim_{n\to\infty} \mathrm{Bias}[\hat{\theta}] = 0$, then $\hat{\theta}$ is asymptotically unbiased.

Variance

$$\mathrm{Var}[\hat{\theta}] = \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}[\hat{\theta}]\right)^2\right]$$

Mean Squared Error

$$\mathrm{MSE}[\hat{\theta}] = \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathrm{Var}[\hat{\theta}] + \left(\mathrm{Bias}[\hat{\theta}]\right)^2$$

Consistency

$$\lim_{n\to\infty} \Pr\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) = 0 \text{ for all } \varepsilon > 0$$

- If $\lim_{n\to\infty} \mathrm{Bias}[\hat{\theta}] = 0$ and $\lim_{n\to\infty} \mathrm{Var}[\hat{\theta}] = 0$, then $\hat{\theta}$ is consistent.

Efficiency

$$\mathrm{Eff}[\hat{\theta}] = \frac{[I(\theta)]^{-1}}{\mathrm{Var}[\hat{\theta}]}$$

- If $\mathrm{Eff}[\hat{\theta}] = 1$, then $\hat{\theta}$ is efficient.

Fisher Information

$$I(\theta) = -\mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} l(\theta)\right] = -n \cdot \mathrm{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln f(X)\right]$$

- $[I(\theta)]^{-1}$ is the Rao-Cramér lower bound.
- $I(\theta) \cdot g'(\theta)^{-2}$ is the Fisher information for $g(\theta)$.

Minimum Variance Unbiased Estimator

- The MVUE is an unbiased estimator with the smallest variance among all *unbiased* estimators.
- If Y is a complete sufficient statistic for θ and $\varphi(Y)$ is an unbiased estimator of θ, then the MVUE of θ is $\varphi(Y)$.

Sufficiency

- Y is a sufficient statistic for θ if and only if $f(x_1, \dots, x_n | y) = h(x_1, \dots, x_n)$ where $h(x_1, \dots, x_n)$ does not depend on θ.
- By factorization theorem, Y is sufficient if and only if $f(x_1, \dots, x_n) = h_1(y, \theta) \cdot h_2(x_1, \dots, x_n)$ for non-negative functions h_1 and h_2 where $h_2(x_1, \dots, x_n)$ does not depend on θ.
- $g(Y)$ is a sufficient statistic for θ if $g(\cdot)$ is a one-to-one function of sufficient Y.
- By Rao-Blackwell theorem, the variance of the unbiased estimator $\mathrm{E}_Z[Z|Y]$ is at most the variance of any unbiased estimator Z for sufficient Y. The MVUE $\varphi(Y)$ is $\mathrm{E}_Z[Z|Y]$.

Exponential Class of Distributions

$$f(x) = \exp[a(x) \cdot b(\theta) + c(\theta) + d(x)]$$

- $\sum_{i=1}^{n} a(X_i)$ is a complete sufficient statistic for θ.

Maximum Likelihood Estimators

Under specific circumstances, the MLE of θ:

- Consistent estimator
- Asymptotically follows a normal distribution with mean θ and variance $[I(\theta)]^{-1}$; its exact variance may equal the asymptotic variance
- Function of sufficient statistic Y

Key Results for Distributions in the Exponential Class

Distribution	Parameter of Interest	$\sum_{i=1}^{n} a(X_i)$	MVUE
Binomial	q	$\sum_{i=1}^{n} X_i$	$\frac{1}{m}\bar{X}$
Normal	μ	$\sum_{i=1}^{n} X_i$	$\bar{X}$
Normal	σ^2	$\sum_{i=1}^{n}(X_i - \mu)^2$	$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$
Poisson	λ	$\sum_{i=1}^{n} X_i$	$\bar{X}$
Gamma	θ	$\sum_{i=1}^{n} X_i$	$\frac{1}{\alpha}\bar{X}$
Inverse Gaussian	μ	$\sum_{i=1}^{n} X_i$	$\bar{X}$
Negative Binomial	β	$\sum_{i=1}^{n} X_i$	$\frac{1}{r}\bar{X}$

Hypothesis Testing

Terminology

- *Test statistic*: A value calculated from data that assumes H_0 is true
- *Critical region*: The range of test statistic values where H_0 is rejected
- *Critical value*: A value that borders the critical region
- *Two-tailed test*: A test that includes both tails in its critical region
- *Right-tailed test*: A test that only includes the right tail in its critical region
- *Left-tailed test*: A test that only includes the left tail in its critical region
- *Significance level, α*: The probability of rejecting H_0, assuming it is true
- *Power*: The probability of rejecting H_0, assuming it is false
- *p-value*: The probability of observing the test statistic or a more extreme value, assuming H_0 is true

	H_0 is true	H_0 is false
Reject H_0	Type I Error	Correct Decision
Fail to reject H_0	Correct Decision	Type II Error

- For all hypothesis tests, reject H_0 if p-value $\le \alpha$.

Tests for Means

- When variance is known, we apply the Central Limit Theorem.
- When variance is unknown, the random sample must be drawn from a normal distribution.

Critical Regions – Known Variance

Test Type	*Critical Region*
Left-tailed	$t.s. \le -z_{1-\alpha}$
Two-tailed	$\lvert t.s.\rvert \ge z_{1-\alpha/2}$
Right-tailed	$t.s. \ge z_{1-\alpha}$

Critical Regions – Unknown Variance

Test Type	*Critical Region*
Left-tailed	$t.s. \le -t_{2\alpha,\mathrm{df}}$
Two-tailed	$\lvert t.s.\rvert \ge t_{\alpha,\mathrm{df}}$
Right-tailed	$t.s. \ge t_{2\alpha,\mathrm{df}}$

One Sample

- $\mathrm{df} = n - 1$

Two Samples

$$s_\mathrm{p}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

- $\sigma_1^2 = \sigma_2^2$
- $\mathrm{df} = n_1 + n_2 - 2$

Two Samples – Paired

- Samples are not independent; observations form pairs.
- Identical to one sample of observed differences
- $n_* = n_1 = n_2$
- $\mathrm{df} = n_* - 1$

Tests for Proportions

$$\hat{q} = \frac{\text{\# of successes from } n \text{ trials}}{n}$$

- Critical regions are the same as those for testing means with known variance.

Tests for Variances – One Sample

Test Type	*Critical Region*
Left-tailed	$t.s. \le \chi^2_{\alpha,n-1}$
Two-tailed	$\left[t.s. \le \chi^2_{\alpha/2,n-1}\right] \cup \left[t.s. \ge \chi^2_{1-\alpha/2,n-1}\right]$
Right-tailed	$t.s. \ge \chi^2_{1-\alpha,n-1}$

Tests for Variances – Two Samples

Test Type	*Critical Region*
Left-tailed	$t.s. \le F_{1-\alpha,n_1-1,n_2-1}$
Two-tailed	$\left[t.s. \le \left(F_{\alpha/2,n_2-1,n_1-1}\right)^{-1}\right] \cup \left[t.s. \ge F_{\alpha/2,n_1-1,n_2-1}\right]$
Right-tailed	$t.s. \ge F_{\alpha,n_1-1,n_2-1}$

- A left-tailed test can be performed by writing H_0 in terms of σ_2^2/σ_1^2 instead and doing a right-tailed test.
- $F_{q,v_2,v_1} = \left(F_{1-q,v_1,v_2}\right)^{-1}$

Summary for Hypothesis Testing

Parameter	# of Samples	H_0	Variance	$t.s.$
Means	One	$\mu = h$	Known	$\dfrac{\bar{x}-h}{\sigma/\sqrt{n}}$
			Unknown	$\dfrac{\bar{x}-h}{s/\sqrt{n}}$
	Two	$\mu_1 - \mu_2 = h$	Known	$\dfrac{\bar{x}_1-\bar{x}_2-h}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$
			Unknown	$\dfrac{\bar{x}_1-\bar{x}_2-h}{s_{\text{p}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$
	Two, Paired	$\mu_1 - \mu_2 = h$	Known	$\dfrac{\bar{d}-h}{\sigma_D/\sqrt{n_*}}$
			Unknown	$\dfrac{\bar{d}-h}{s_D/\sqrt{n_*}}$
Proportions	One	$q = h$	–	$\dfrac{\hat{q}-h}{\sqrt{\dfrac{h(1-h)}{n}}}$
	Two	$q_1 - q_2 = h$	–	$\dfrac{\hat{q}_1-\hat{q}_2-h}{\sqrt{\dfrac{\hat{q}_1(1-\hat{q}_1)}{n_1}+\dfrac{\hat{q}_2(1-\hat{q}_2)}{n_2}}}$
Variances	One	$\sigma^2 = h$	–	$\dfrac{(n-1)s^2}{h}$
	Two	$\dfrac{\sigma_1^2}{\sigma_2^2} = h$	–	$\dfrac{s_1^2}{s_2^2}\cdot\dfrac{1}{h}$

Intervals for Means

Parameter	Scenario	Type	$100k\%$ Confidence Interval
μ	Known Variance	Two-sided	$\bar{x} \pm z_{(1+k)/2} \cdot \frac{\sigma}{\sqrt{n}}$
		Left-sided	$\left(-\infty, \bar{x} + z_k \cdot \frac{\sigma}{\sqrt{n}}\right)$
		Right-sided	$\left(\bar{x} - z_k \cdot \frac{\sigma}{\sqrt{n}}, \infty\right)$
	Unknown Variance	Two-sided	$\bar{x} \pm t_{1-k,n-1} \cdot \frac{s}{\sqrt{n}}$
		Left-sided	$\left(-\infty, \bar{x} + t_{2(1-k),n-1} \cdot \frac{s}{\sqrt{n}}\right)$
		Right-sided	$\left(\bar{x} - t_{2(1-k),n-1} \cdot \frac{s}{\sqrt{n}}, \infty\right)$
$\mu_1 - \mu_2$	Known Variances	Two-sided	$\bar{x}_1 - \bar{x}_2 \pm z_{(1+k)/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
		Left-sided	$\left(-\infty, \bar{x}_1 - \bar{x}_2 + z_k\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
		Right-sided	$\left(\bar{x}_1 - \bar{x}_2 - z_k\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \infty\right)$
	Unknown Variances	Two-sided	$\bar{x}_1 - \bar{x}_2 \pm t_{1-k,n_1+n_2-2} \cdot s_{\text{p}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
		Left-sided	$\left(-\infty, \bar{x}_1 - \bar{x}_2 + t_{2(1-k),n_1+n_2-2} \cdot s_{\text{p}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$
		Right-sided	$\left(\bar{x}_1 - \bar{x}_2 - t_{2(1-k),n_1+n_2-2} \cdot s_{\text{p}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty\right)$
	Paired	All	Identical to the one-sample case

Intervals for Proportions

Parameter	Type	$100k\%$ Confidence Interval
q	Two-sided	$\hat{q} \pm z_{(1+k)/2}\sqrt{\dfrac{\hat{q}(1-\hat{q})}{n}}$
	Left-sided	$\left(-\infty, \hat{q} + z_k\sqrt{\dfrac{\hat{q}(1-\hat{q})}{n}}\right)$
	Right-sided	$\left(\hat{q} - z_k\sqrt{\dfrac{\hat{q}(1-\hat{q})}{n}}, \infty\right)$
$q_1 - q_2$	Two-sided	$\hat{q}_1 - \hat{q}_2 \pm z_{(1+k)/2}\sqrt{\dfrac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \dfrac{\hat{q}_2(1-\hat{q}_2)}{n_2}}$
	Left-sided	$\left(-\infty, \hat{q}_1 - \hat{q}_2 + z_k\sqrt{\dfrac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \dfrac{\hat{q}_2(1-\hat{q}_2)}{n_2}}\right)$
	Right-sided	$\left(\hat{q}_1 - \hat{q}_2 - z_k\sqrt{\dfrac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \dfrac{\hat{q}_2(1-\hat{q}_2)}{n_2}}, \infty\right)$

Intervals for Variances

Parameter	Type	$100k\%$ Confidence Interval
σ^2	Two-sided	$\left(\dfrac{(n-1)s^2}{\chi^2_{(1+k)/2,n-1}}, \dfrac{(n-1)s^2}{\chi^2_{(1-k)/2,n-1}}\right)$
	Left-sided	$\left(0, \dfrac{(n-1)s^2}{\chi^2_{1-k,n-1}}\right)$
	Right-sided	$\left(\dfrac{(n-1)s^2}{\chi^2_{k,n-1}}, \infty\right)$
$\dfrac{\sigma_1^2}{\sigma_2^2}$	Two-sided	$\left(\dfrac{s_1^2}{s_2^2} \cdot \left(F_{(1-k)/2,n_1-1,n_2-1}\right)^{-1}, \dfrac{s_1^2}{s_2^2} \cdot F_{(1-k)/2,n_2-1,n_1-1}\right)$
	Left-sided	$\left(0, \dfrac{s_1^2}{s_2^2} \cdot F_{1-k,n_2-1,n_1-1}\right)$
	Right-sided	$\left(\dfrac{s_1^2}{s_2^2} \cdot \left(F_{1-k,n_1-1,n_2-1}\right)^{-1}, \infty\right)$

Most Powerful Tests

Terminology

- *Simple*: Fully specifies the distribution(s)
- *Composite*: Does not fully specify the distribution(s)

Most Powerful Test

When H_0 and H_1 are both simple, the most powerful test of size α has the largest power among all tests with the same α.

Neyman-Pearson Theorem

The best critical region is embedded in

$$\frac{L(h_0)}{L(h_1)} \le k$$

where H_0 and H_1 are both simple.

Uniformly Most Powerful (UMP) Tests

- For a simple H_0 and composite H_1, a test is UMP when the best critical region is the same for testing H_0 against each simple hypothesis in H_1.
- For composite hypotheses $H_0 : \theta \le h$ and $H_1 : \theta > h$, a test is UMP if there is a monotone likelihood ratio in a statistic y.

Goodness of Fit Tests

Kolmogorov-Smirnov Test

$t.s. = D =$ maximum absolute difference between $F^*(x)$ and $\hat{F}(x)$

- Reject H_0 if $t.s. \ge$ critical value
- $F^*(x)$: CDF of the proposed distribution
- $\hat{F}(x)$: Empirical distribution function

$$\hat{F}(x) = \frac{\#\text{ of observations} \le x}{n}$$

Left-Truncated at d

$$F^*(x) = \frac{F(x) - F(d)}{1 - F(d)}$$

Right-Censored at m

$\hat{F}(m)$ is undefined.

Chi-Square Goodness-of-Fit Test

$$t.s. = \sum_{j=1}^{k} \frac{\left(n_j - nq_j\right)^2}{nq_j}$$

- Reject H_0 if $t.s. \ge \chi^2_{1-\alpha,k-1-r}$
- k: # of mutually exclusive intervals
- q_j: probability of being in interval j
- n_j: # of observed values in interval j
- r: # of free parameters

Chi-Square Test of Independence

$$t.s. = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(n_{ij}n - n_{i\bullet}n_{\bullet j}\right)^2}{n_{i\bullet}n_{\bullet j}}$$

- Reject H_0 if $t.s. \ge \chi^2_{1-\alpha,(a-1)(b-1)}$
- a: # of categories for first variable
- b: # of categories for second variable
- n_{ij}: # of observations in first variable's category i and second variable's category j
- $n_{i\bullet}$: subtotal # of observations in category i, across all categories of the second variable
- $n_{\bullet j}$: subtotal # of observations in category j, across all categories of the first variable

Likelihood Ratio Test

$$t.s. = -2\ln\left(\frac{L_0}{L_1}\right) = 2(l_1 - l_0)$$

- Reject H_0 if $t.s. \ge \chi^2_{1-\alpha,r_1-r_0}$
- r_0: # of free parameters in distribution under H_0
- r_1: # of free parameters in distribution under H_1
- L_0: Maximized likelihood under H_0
- L_1: Maximized likelihood under H_1
- $l_0 = \ln L_0$
- $l_1 = \ln L_1$

Confidence Intervals

- For means and proportions, the two-sided general form is estimate $\pm$ (percentile)(standard error)
- H_0 will fail to be rejected at α if h is within the $100(1-\alpha)\%$ confidence interval.

Order Statistics

$X_{(k)} = k^{\text{th}}$ order statistic
$X_{(1)} = \min(X_1, \ldots, X_n)$
$X_{(n)} = \max(X_1, \ldots, X_n)$

First Principles

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,(n-k)!} \cdot [F_X(x)]^{k-1} \cdot f_X(x) \cdot [S_X(x)]^{n-k}$$

Special Cases

Uniform (a, b)
$\mathrm{E}\left[X_{(k)}\right] = a + \dfrac{k(b-a)}{n+1}$
Uniform $(0, \theta)$
$X_{(k)} \sim$ Beta $(k, n-k+1, \theta)$
Exponential (θ)
$\mathrm{E}\left[X_{(k)}\right] = \theta \displaystyle\sum_{i=n-k+1}^{n} \frac{1}{i}$

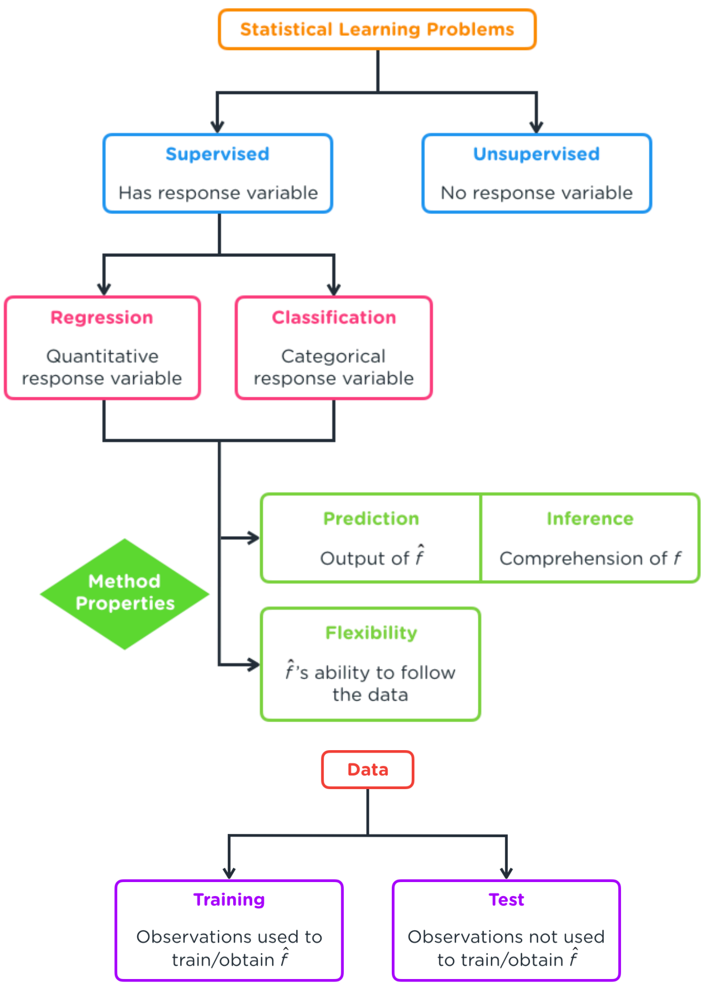
Extended Linear Models

Introduction to Statistical Learning

Types of Variables

- *Response*: A variable of primary interest
- *Explanatory*: A variable used to study the response variable
- *Count*: A quantitative variable valid on non-negative integers
- *Continuous*: A quantitative variable valid on real numbers
- *Nominal*: A qualitative variable having categories without a meaningful or logical order
- *Ordinal*: A qualitative variable having categories with a meaningful or logical order

Contrasting Statistical Learning Elements



Model Accuracy

$Y = f(x_1, \dots, x_p) + \varepsilon, \quad \mathrm{E}[\varepsilon] = 0$

Test MSE $= \mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$ can be estimated using $\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

For fixed inputs $x_1, \dots, x_p$, the test MSE is

$$\underbrace{\mathrm{Var}\left[\hat{f}(x_1, \dots, x_p)\right] + \left(\mathrm{Bias}\left[\hat{f}(x_1, \dots, x_p)\right]\right)^2}_{\text{reducible error}} + \underbrace{\mathrm{Var}[\varepsilon]}_{\text{irreducible error}}$$

- If training data y_i's are used, training MSE is computed instead.
- As flexibility increases, the training MSE decreases, but the test MSE follows a u-shaped pattern.
- Low flexibility leads to a method with low variance and high bias; high flexibility leads to a method with high variance and low bias.

Numerical Summaries

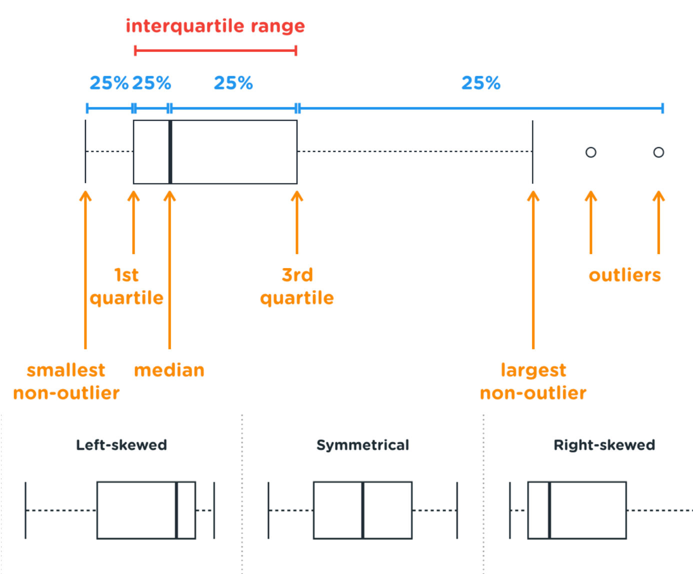
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$cov_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$r_{x,y} = \frac{cov_{x,y}}{s_x \cdot s_y}, \qquad -1 \le r_{x,y} \le 1$$

Graphical Summaries

- A scatterplot plots values of two variables to investigate their relationship.
- A box plot captures a variable's distribution using its median, 1st and 3rd quartiles, and distribution tails.
- A QQ plot plots sample percentiles against theoretical percentiles to determine whether the sample and theoretical distributions have similar shapes.



Simple Linear Regression (SLR)

Special case of MLR where $p = 1$

Estimation

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Standard Errors

$$se(\hat{\beta}_0) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$se(\hat{y}) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}$$

$$se(\hat{y}_{n+1}) = \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}$$

Other Numerical Results

$R^2 = r_{x,y}^2$

Multiple Linear Regression (MLR)

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Assumptions

1. $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$
2. $x_{i,j}$'s are non-random
3. $\mathrm{E}[\varepsilon_i] = 0$
4. $\mathrm{Var}[\varepsilon_i] = \sigma^2$
5. ε_i's are independent
6. ε_i's are normally distributed
7. The predictor x_j is not a linear combination of the other p predictors, for $j = 0, 1, \ldots, p$

Estimation – Ordinary Least Squares (OLS)

$$\begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1}$$

residual standard error $= \sqrt{\text{MSE}}$

Other Numerical Results

$$e = y - \hat{y}$$

$$\text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2 = \text{SSR} + \text{SSE}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$R^2_{\text{adj.}} = 1 - \frac{\text{MSE}}{s_y^2} = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

Other Key Ideas

- R^2 is a poor measure for model comparison because it will increase simply by adding more predictors to a model.
- Polynomials do not change consistently by unit increases of its variable, i.e., no constant slope.
- Only $w - 1$ dummy variables are needed to represent w classes of a categorical predictor; one of them acts as the baseline class.
- In effect, dummy variables define a distinct intercept for each class. Without the interaction between a dummy variable and a predictor, the dummy variable cannot additionally affect that predictor's regression coefficient.

Standard Errors

$$\widehat{\mathrm{Var}}[\hat{\boldsymbol{\beta}}] = \text{MSE}(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \widehat{\mathrm{Var}}[\hat{\beta}_0] & \cdots & \widehat{\mathrm{Cov}}[\hat{\beta}_0, \hat{\beta}_p] \\ \vdots & \ddots & \vdots \\ \widehat{\mathrm{Cov}}[\hat{\beta}_0, \hat{\beta}_p] & \cdots & \widehat{\mathrm{Var}}[\hat{\beta}_p] \end{bmatrix}$$

$$se(\hat{\beta}_j) = \sqrt{\widehat{\mathrm{Var}}[\hat{\beta}_j]}$$

Confidence Intervals

$$\hat{\beta}_j \pm t_{1-k,n-p-1} \cdot se(\hat{\beta}_j)$$

$$\hat{y} \pm t_{1-k,n-p-1} \cdot se(\hat{y})$$

Prediction Intervals

$$\hat{y}_{n+1} \pm t_{1-k,n-p-1} \cdot se(\hat{y}_{n+1})$$

t Tests

$$t.s. = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

Test Type	*Critical Region*
Left-tailed	$t.s. \le -t_{2\alpha,n-p-1}$
Two-tailed	$\lvert t.s.\rvert \ge t_{\alpha,n-p-1}$
Right-tailed	$t.s. \ge t_{2\alpha,n-p-1}$

F Tests

$$t.s. = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR} \div p}{\text{SSE} \div (n - p - 1)}$$

- Reject H_0 if $t.s. \ge F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = p$
- $\text{ddf} = n - p - 1$
- If $p = 1$, $t.s.$ is the squared test statistic of the t test with the same H_0.

Source	*SS*	*df*	*MS*
Regression	SSR	p	MSR
Error	SSE	$n - p - 1$	MSE
Total	SST	$n - 1$	s_y^2

Partial *F* Tests

$$t.s. = \frac{\overbrace{(\text{SSE}_r - \text{SSE}_f)}^{\text{reduction in variability}} \div \overbrace{(p_f - p_r)}^{\text{additional df spent}}}{\text{SSE}_f \div (n - p_f - 1)} = \frac{(R_f^2 - R_r^2) \div (p_f - p_r)}{(1 - R_f^2) \div (n - p_f - 1)}$$

- Reject H_0 if $t.s. \ge F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = p_f - p_r$
- $\text{ddf} = n - p_f - 1$

Source	*SS*	*df*
Reduced Regression	SSR_r	p_r
Difference	$\text{SSE}_r - \text{SSE}_f$ or $\text{SSR}_f - \text{SSR}_r$	$p_f - p_r$
Full Error	SSE_f	$n - p_f - 1$
Total	SST	$n - 1$

Bootstrapping

The bootstrapped $se(\hat{\beta}_j)$ is the unbiased sample standard deviation of the β_j bootstrap estimates.

Analysis of Variance (ANOVA)

One-Way ANOVA

$Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}$

- $i = 1, \ldots, n_j$
- Factor has w levels, $j = 1, \ldots, w$

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j}$$

$$\text{SSR} = \sum_{j=1}^{w}\sum_{i=1}^{n_j}(\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{w} n_j(\bar{y}_j - \bar{y})^2$$

$$\text{SSE} = \sum_{j=1}^{w}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y}_j)^2$$

$$\text{SST} = \sum_{j=1}^{w}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y})^2$$

Source	*SS*	*df*
Factor	SSR	$w-1$
Error	SSE	$n-w$
Total	SST	$n-1$

Testing the Significance of Factor

$$t.s. = \frac{\text{SSR} \div (w-1)}{\text{SSE} \div (n-w)}$$

- Reject H_0 if $t.s. \geq F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = w - 1$
- $\text{ddf} = n - w$

Two-Way ANOVA – Additive Model

$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{i,j,k}$

- Factor A has w levels, $i = 1, \ldots, n_*$
- Factor B has v levels, $j = 1, \ldots, w$
- $k = 1, \ldots, v$

$$\begin{aligned}\text{SSR}_\text{B} &= \text{SSE}_\text{A} - \text{SSE}_\text{add} \\ &= \text{SSR}_\text{add} - \text{SSR}_\text{A}\end{aligned}$$

Source	*SS*	*df*
Factor A	SSR_A	$w-1$
Factor B	SSR_B	$v-1$
Error	SSE_add	$n-w-v+1$
Total	SST	$n-1$

Testing the Significance of Factor A

$$t.s. = \frac{\text{SSR}_\text{A} \div (w-1)}{\text{SSE}_\text{add} \div (n-w-v+1)}$$

- Reject H_0 if $t.s. \geq F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = w - 1$
- $\text{ddf} = n - w - v + 1$

Testing the Significance of Factor B

$$t.s. = \frac{\text{SSR}_\text{B} \div (v-1)}{\text{SSE}_\text{add} \div (n-w-v+1)}$$

- Reject H_0 if $t.s. \geq F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = v - 1$
- $\text{ddf} = n - w - v + 1$

Two-Way ANOVA – Additive Model without Replication

$Y_{j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{j,k}$

- $n_* = 1$
- $j = 1, \ldots, w$
- $k = 1, \ldots, v$

$$\bar{y}_{j\bullet} = \frac{1}{v}\sum_{k=1}^{v} y_{j,k}\,, \qquad \bar{y}_{\bullet k} = \frac{1}{w}\sum_{j=1}^{w} y_{j,k}$$

$$\text{SSR}_\text{A} = \sum_{k=1}^{v}\sum_{j=1}^{w}(\bar{y}_{j\bullet} - \bar{y})^2 = \sum_{j=1}^{w} v(\bar{y}_{j\bullet} - \bar{y})^2$$

$$\text{SSR}_\text{B} = \sum_{k=1}^{v}\sum_{j=1}^{w}(\bar{y}_{\bullet k} - \bar{y})^2 = \sum_{k=1}^{v} w(\bar{y}_{\bullet k} - \bar{y})^2$$

$$\text{SSE}_\text{add} = \sum_{k=1}^{v}\sum_{j=1}^{w}(y_{j,k} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y})^2$$

$$\text{SST} = \sum_{k=1}^{v}\sum_{j=1}^{w}(y_{j,k} - \bar{y})^2$$

Two-Way ANOVA – Model with Interactions

$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k} + \varepsilon_{i,j,k}$

- $i = 1, \ldots, n_*$
- $j = 1, \ldots, w$
- $k = 1, \ldots, v$

$$\begin{aligned}\text{SS}_\text{diff} &= \text{SSE}_\text{add} - \text{SSE}_\text{int} \\ &= \text{SSR}_\text{int} - \text{SSR}_\text{add}\end{aligned}$$

Source	*SS*	*df*
Factor A	SSR_A	$w-1$
Factor B	SSR_B	$v-1$
Interaction	SS_diff	$(w-1)(v-1)$
Error	SSE_int	$n-wv$
Total	SST	$n-1$

Testing the Significance of Interactions

$$t.s. = \frac{\text{SS}_\text{diff} \div [(w-1)(v-1)]}{\text{SSE}_\text{int} \div (n-wv)}$$

- Reject H_0 if $t.s. \geq F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = (w-1)(v-1)$
- $\text{ddf} = n - wv$

Testing the Significance of Factor A

$$t.s. = \frac{\text{SSR}_\text{A} \div (w-1)}{\text{SSE}_\text{int} \div (n-wv)}$$

- Reject H_0 if $t.s. \geq F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = w - 1$
- $\text{ddf} = n - wv$

Testing the Significance of Factor B

$$t.s. = \frac{\text{SSR}_\text{B} \div (v-1)}{\text{SSE}_\text{int} \div (n-wv)}$$

- Reject H_0 if $t.s. \geq F_{\alpha,\text{ndf},\text{ddf}}$
- $\text{ndf} = v - 1$
- $\text{ddf} = n - wv$

Other Key Ideas

- In testing whether a source is significant, the test statistic is the mean square of that source divided by the MSE of the model that has the most predictors.
- ANCOVA models have both quantitative and qualitative predictors.
- The uncorrected total sum of squares is $\sum_{i=1}^{n} y_i^2$. The sources of an ANOVA/ANCOVA table may sum to the uncorrected table rather than the corrected total.

Linear Model Assumptions

Leverage

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{u=1}^{n}(x_u - \bar{x})^2} \text{ for SLR}$$

- h_i is the i^{th} diagonal entry of $\mathbf{H}$.
- $\sum_{i=1}^{n} h_i = p + 1$

Standardized Residuals

$$e_{\text{sta},i} = \frac{e_i}{\sqrt{\text{MSE}(1 - h_i)}}$$

DFITS

$$\text{DFITS}_i = e_{\text{sta},i}\sqrt{\frac{h_i}{1 - h_i}}$$

Cook's Distance

$$d_i = \frac{\text{DFITS}_i^2}{p+1} = \frac{e_{\text{sta},i}^2 h_i}{(p+1)(1-h_i)} = \frac{e_i^2 h_i}{\text{MSE}(p+1)(1-h_i)^2}$$

Plots of Residuals

- e versus $\hat{y}$
  Residuals are well-behaved if
  - Points appear to be randomly scattered
  - Residuals seem to average to 0
  - Spread of residuals does not change
- e versus i
  Detects dependence of error terms
- QQ plot of e

Variance Inflation Factor

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

$\text{VIF}_j > 5$ indicates multicollinearity.

Curse of Dimensionality

Having many predictors in a model increases the risk of including noise predictors that are not associated with the response.

Model Selection

- g: Total # of predictors in consideration
- p: # of predictors for a specific model
- MSE_g: MSE of the model that uses all g predictors
- M_p: The "best" model with p predictors

Best Subset Selection

1. For $p = 0, 1, \ldots, g$, fit all $\binom{g}{p}$ models with p predictors. The model with the largest R^2 is M_p.
2. Choose the best model among $\text{M}_0, \ldots, \text{M}_g$ using a selection criterion of choice.

Forward Stepwise Selection

1. Fit all g simple linear regression models. The model with the largest R^2 is M_1.
2. For $p = 2, \ldots, g$, fit the models that add one of the remaining predictors to M_{p-1}. The model with the largest R^2 is M_p.
3. Choose the best model among $\text{M}_0, \ldots, \text{M}_g$ using a selection criterion of choice.

Backward Stepwise Selection

1. Fit the model with all g predictors, M_g.
2. For $p = g - 1, \ldots, 1$, fit the models that drop one of the predictors from M_{p+1}. The model with the largest R^2 is M_p.
3. Choose the best model among $\text{M}_0, \ldots, \text{M}_g$ using a selection criterion of choice.

Selection Criteria

- Adjusted R^2
- Mallows' C_p
  $$C_p = \frac{1}{n}\left(\text{SSE} + 2p \cdot \text{MSE}_g\right)$$
- Akaike information criterion
  $$\text{AIC} = \frac{1}{n}\left(\text{SSE} + 2p \cdot \text{MSE}_g\right)$$
- Bayesian information criterion
  $$\text{BIC} = \frac{1}{n}\left(\text{SSE} + \ln n \cdot p \cdot \text{MSE}_g\right)$$
- Cross-validation error

Validation Set

- Randomly splits all available observations into two groups: the training set and the validation set.
- Only the observations in the training set are used to attain the fitted model, and those in validation set are used to estimate the test MSE.

k-fold Cross-Validation

1. Randomly divide all available observations into k folds.
2. For $v = 1, \ldots, k$, obtain the v^{th} fit by training with all observations except those in the v^{th} fold.
3. For $v = 1, \ldots, k$, use $\hat{y}$ from the v^{th} fit to calculate a test MSE estimate with observations in the v^{th} fold.
4. To calculate CV error, average the k test MSE estimates in the previous step.

Leave-One-Out Cross-Validation (LOOCV)

- Calculate LOOCV error as a special case of k-fold cross-validation where $k = n$.

$$\text{LOOCV Error} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \text{ for MLR}$$

Key Ideas on Cross-Validation

- The validation set approach has unstable results and will tend to overestimate the test MSE. The two other approaches mitigate these issues.
- With respect to bias, LOOCV < k-fold CV < Validation Set.
- With respect to variance, LOOCV > k-fold CV > Validation Set.

Other Linear Regression Approaches

Standardizing Variables

- A centered variable is the result of subtracting the sample mean from a variable.
- A scaled variable is the result of dividing a variable by its standard deviation.
- A standardized variable is the result of first centering a variable, then scaling it.

Shrinkage Methods

	Ridge	Lasso
Minimize	SSE $+\lambda\sum_{j=1}^{p}\hat{\beta}_j^2$	SSE $+\lambda\sum_{j=1}^{p}\lvert\hat{\beta}_j\rvert$
	SSE subject to $\sum_{j=1}^{p}\hat{\beta}_j^2\le a$	SSE subject to $\sum_{j=1}^{p}\lvert\hat{\beta}_j\rvert\le a$
ℓ norm	$\lVert\hat{\boldsymbol{\beta}}\rVert_2=\sqrt{\sum_{j=1}^{p}\hat{\beta}_j^2}$	$\lVert\hat{\boldsymbol{\beta}}\rVert_1=\sum_{j=1}^{p}\lvert\hat{\beta}_j\rvert$

- λ: Tuning parameter
- a: Budget parameter
- $x_1,\dots,x_p$ are scaled predictors.
- λ is inversely related to flexibility.
- With a finite λ, none of the ridge estimates will equal 0, but the lasso estimates could equal 0.

Principal Components

$$z_m=\sum_{j=1}^{p}\phi_{j,m}x_j$$

$$\sum_{j=1}^{p}\phi_{j,m}^2=1$$

$$\sum_{j=1}^{p}\phi_{j,m}\cdot\phi_{j,u}=0,\qquad m\neq u$$

- Unsupervised technique that performs dimension reduction on p variables
- The variability explained by each subsequent principal component is always less than the variability explained by its previous principal component.
- Principal components form the lower dimension surface that is closest to the observations in p-dimensional space.
- Standardized variables affect the loadings by becoming resistant to varying scales among the original variables.

Principal Components Regression

- Uses the first k principal components that are orthogonal as predictors in an MLR.
- k is a measure of flexibility.
- When $k=p$, PCR is equivalent to performing MLR with the p original variables as predictors.

Partial Least Squares

- Supervised technique that performs dimension reduction on p variables
- Uses the first k PLS directions that are orthogonal as predictors in an MLR.
- k is a measure of flexibility.
- When $k=p$, PLS is equivalent to performing MLR with the p original variables as predictors.
- The first PLS direction is a linear combination of the p standardized predictors, with coefficients that are based on the response y.
- Every subsequent PLS direction is calculated iteratively as a linear combination of "updated predictors" which are the residuals of fits with the "previous predictors" explained by the previous direction.

Generalized Linear Models

Exponential Family*

$f(y) = \exp[a(y) \cdot b(\theta) + c(\theta) + d(y)]$

$\mathrm{E}[a(Y)] = -\dfrac{c'(\theta)}{b'(\theta)}$

$\mathrm{Var}[a(Y)] = \dfrac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$

Canonical Form

- $a(y) = y$
- $b(\theta)$ is the natural parameter
- $\mu = \mathrm{E}[Y]$ is a function of θ
- $\mathrm{Var}[Y]$ is a function of μ

*Key results on Exponential Family is on page 21.

Model Framework

$g(\mu) = \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Function Name	$g(\mu)$
Identity	μ
Logit	$\ln\left(\dfrac{\mu}{1-\mu}\right)$
Logarithmic	$\ln \mu$
Inverse	$\dfrac{1}{\mu}$
Power	μ^d

Distribution	*Canonical Link*
Normal	Identity
Binomial	Logit
Poisson	Logarithmic
Gamma	Inverse
Inverse Gaussian	Inverse squared

Parameter Estimation

$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i)]$

$\hat{\mu} = g^{-1}\left(\mathbf{x}^T\hat{\boldsymbol{\beta}}\right)$

$u_j = \sum_{i=1}^{n} \dfrac{(y_i - \mu_i)\, x_{i,j}}{\mathrm{Var}[Y_i] \cdot g'(\mu_i)}$

$\mathbf{I} = \sum_{i=1}^{n} \dfrac{\mathbf{x}_i \mathbf{x}_i^T}{\mathrm{Var}[Y_i] \cdot g'(\mu_i)^2}$

Parameter Estimation – Method of Scoring

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m)} &= \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\mathbf{I}^{(m-1)}\right]^{-1}\mathbf{u}^{(m-1)} \\ &= \left(\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}\end{aligned}$$

$w_i = \dfrac{1}{\mathrm{Var}[Y_i] \cdot g'(\mu_i)^2}$

$z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$

Numerical Results

$D = 2\left[l_{\text{sat}} - l(\hat{\boldsymbol{\beta}})\right]$

$R^2_{\text{pse.}} = 1 - \dfrac{l(\hat{\boldsymbol{\beta}})}{l_{\text{null}}}$

$\mathrm{AIC} = -2 \cdot l(\hat{\boldsymbol{\beta}}) + 2k$

$\mathrm{BIC} = -2 \cdot l(\hat{\boldsymbol{\beta}}) + k \ln n$

where k is the # of estimated parameters

Residuals

Raw Residual

$e_i = y_i - \hat{\mu}_i$

Pearson Residual

$e_i^P = \dfrac{e_i}{\sqrt{\widehat{\mathrm{Var}}[Y_i]}}$

$e_{\text{sta},i}^P = \dfrac{e_i^P}{\sqrt{1 - h_i}}$

- Pearson chi-square statistic is $\sum_{i=1}^{n}\left(e_i^P\right)^2$.

Deviance Residual

$e_i^D = \pm\sqrt{D_i}$

whose sign follows the i^{th} raw residual

$e_{\text{sta},i}^D = \dfrac{e_i^D}{\sqrt{1 - h_i}}$

- Deviance is $\sum_{i=1}^{n}\left(e_i^D\right)^2$.

Inference

- Score statistics $\mathbf{U}$ asymptotically follow a multivariate normal distribution with mean $\mathbf{0}$ and asymptotic variance-covariance matrix $\mathbf{I}$. Thus, $\mathbf{U}^T\mathbf{I}^{-1}\mathbf{U}$ follows an approximate chi-square distribution with $p + 1$ degrees of freedom.
- Maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ asymptotically follow a multivariate normal distribution with mean $\boldsymbol{\beta}$ and asymptotic variance-covariance matrix $\mathbf{I}^{-1}$.
- Overdispersion can be addressed by quasi-likelihood method, which changes the variance to: $\mathrm{Var}[Y_i] = \phi \cdot \text{original variance}$

Likelihood Ratio Test

$$\begin{aligned}t.s. &= 2\left[l\left(\hat{\boldsymbol{\beta}}_f\right) - l\left(\hat{\boldsymbol{\beta}}_r\right)\right] \\ &= D_r - D_f\end{aligned}$$

- Reject H_0 if $t.s. \geq \chi^2_{1-\alpha, p_f - p_r}$

Wald Test

$t.s. = \left[\dfrac{\hat{\beta}_j - h}{se(\hat{\beta}_j)}\right]^2$

- Reject H_0 if $t.s. \geq \chi^2_{1-\alpha,1}$
- $\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\mathbf{I}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ follows an approximate chi-square distribution with $p + 1$ degrees of freedom.

Tweedie Distributions

$\mathrm{Var}[Y] = a \cdot \mathrm{E}[Y]^d$

Distribution	*d*
Normal	0
Poisson	1
Compound Poisson-Gamma	(1, 2)
Gamma	2
Inverse Gaussian	3

Connection with MLR

- A GLM with a normally distributed response, identity link, and homoscedasticity is the same as MLR.
- MLE estimates = OLS estimates
- $\sigma^2 D = \mathrm{SSE}$

Binomial and Categorical Response Regression

Binomial Response Variable

- The odds of an event are the ratio of the probability that the event will occur to the probability that the event will not occur, i.e.,
$$\text{odds} = \frac{q}{1-q}$$
- The odds ratio is the ratio of the odds of an event with the presence of a characteristic to the odds of the same event without the presence of that characteristic.

Function Name	$g(q)$
Logit	$\ln\left(\dfrac{q}{1-q}\right)$
Probit	$\Phi^{-1}(q)$
Complementary log-log	$\ln[-\ln(1-q)]$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \ln\left(\frac{q_i}{1-q_i}\right) + m_i \ln(1-q_i) + \ln\binom{m_i}{y_i}\right]$$

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right]$$

$$e_i^P = \frac{y_i - m_i\hat{q}_i}{\sqrt{m_i\hat{q}_i(1-\hat{q}_i)}}$$

$$\text{Pearson chi-square stat.} = \sum_{i=1}^{n}\frac{(y_i - m_i\hat{q}_i)^2}{m_i\hat{q}_i(1-\hat{q}_i)}$$

Logistic Regression

$$q_i = \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}$$

$$u_j = \sum_{i=1}^{n}(y_i - \mu_i)x_{i,j}$$

$$\mathbf{I} = \sum_{i=1}^{n} m_i q_i (1-q_i)\mathbf{x}_i\mathbf{x}_i^T$$

Nominal Response

Let $\pi_{i,c}$ be the probability that the i^{th} observation is classified as category c. k is the reference category.

$$\ln\left(\frac{\pi_{i,t}}{\pi_{i,k}}\right) = \mathbf{x}_i^T\boldsymbol{\beta}_t$$

$$\pi_{i,c} = \begin{cases} \dfrac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_c\right)}{1+\sum_{\text{all } t}\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_t\right)}, & c \neq k \\ \dfrac{1}{1+\sum_{\text{all } t}\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_t\right)}, & c = k \end{cases}$$

Ordinal Response – Proportional Odds Cumulative

$$\ln\left(\frac{\Pi_{i,c}}{1-\Pi_{i,c}}\right) = \beta_{0,c} + \mathbf{x}_i^T\boldsymbol{\beta}$$

$$\Pi_c = \pi_1 + \cdots + \pi_c$$

$$\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

A ratio of cumulative odds is not a function of the predictor values, e.g.,

$$\frac{\hat{\Pi}_1 \div \left(1-\hat{\Pi}_1\right)}{\hat{\Pi}_2 \div \left(1-\hat{\Pi}_2\right)} = \exp\left(\hat{\beta}_{0,1} - \hat{\beta}_{0,2}\right)$$

Poisson Response Regression

$$\mu_i = a_i \cdot \exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)$$

where a_i is the exposure amount

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \ln \mu_i - \mu_i - \ln(y_i!)]$$

$$u_j = \sum_{i=1}^{n}(y_i - \mu_i)\, x_{i,j}$$

$$\mathbf{I} = \sum_{i=1}^{n}\mu_i\mathbf{x}_i\mathbf{x}_i^T$$

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

$$= 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right)$$

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

$$\text{Pearson chi-square stat.} = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Log-Linear Models

- Assess whether there is an association or dependence between two factors.
- The response is the count in each cell of the contingency table created by the two factors.
- Key results of the multinomial model and the product multinomial model are shared with the Poisson model.
- In testing the interaction effects with a likelihood ratio test, the reduced model does not have the interaction terms as predictors, while the full model has the interaction terms.

Generalized Additive Models

The # of degrees of freedom used is the # of regression coefficients, i.e., $p+1$.

Basis Functions

$Y = \beta_0 + \beta_1 b_1(x) + \cdots + \beta_p b_p(x) + \varepsilon$

Step Functions

$$b_j(x) = \begin{cases} I(\xi_j \le x < \xi_{j+1}), & j = 1, \dots, k-1 \\ I(x \ge \xi_k), & j = k \end{cases}$$

Piecewise Polynomial Regression

The basis functions are:

- $x, x^2, \dots, x^d$
- k step functions
- dk interaction terms

Regression Splines

- A degree-d spline is a continuous piecewise degree-d polynomial with continuity in derivatives up to degree $d-1$ at each knot.
- The basis functions of a cubic spline can be $x, x^2, x^3, (x-\xi_1)_+^3, \dots, (x-\xi_k)_+^3$.
- A natural spline is a regression spline that is linear instead of a polynomial in the boundary regions.

Smoothing Splines

$$\text{Minimize } \sum_{i=1}^{n} [y_i - g(x_i)]^2 + \lambda \int_{-\infty}^{\infty} g''(t)^2 \, dt$$

- Smoothing parameter λ is inversely related to flexibility.
- $g(x)$ has the same form as the fitted natural cubic spline with knots at the n values of x.
- Effective degrees of freedom measures flexibility as the sum of the diagonal entries of $\mathbf{S}_\lambda$, where $\hat{\mathbf{y}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$.

Local Regression

- Calculates the fitted value for a specific input by mimicking weighted least squares, i.e., minimize $\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$.
- Weights are determined by the span and the weighting function, such that observations nearer to the input are given larger weights.
- Span is inversely related to flexibility.
- Does not perform well in high dimension.

Generalized Additive Models

- Each explanatory variable contributes to the mean response independently of the other explanatory variables; no interactions are considered.
- The effect of each explanatory variable on the response can be investigated individually, assuming the other variables are held constant.
- Backfitting can be used for fitting if ordinary least squares cannot.

Key Results for Distributions in the Exponential Family

Distribution	θ	Natural Parameter, $b(\theta)$	$c(\theta)$
Binomial, fixed m	q	$\ln\left(\frac{q}{1-q}\right)$	$m\ln(1-q)$
Normal, fixed σ^2	μ	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2}$
Poisson	λ	$\ln\lambda$	$-\lambda$
Gamma, fixed α	θ	$-\frac{1}{\theta}$	$-\alpha\ln\theta$
Inverse Gaussian, fixed θ	μ	$-\frac{\theta}{2\mu^2}$	$\frac{\theta}{\mu}$
Negative Binomial, fixed r	β	$\ln\left(\frac{\beta}{1+\beta}\right)$	$-r\ln(1+\beta)$

Number of Predictors for GAMs with a d^{th} degree polynomial and k knots

Model	# of Predictors, p
Polynomial	d
Piecewise constant	k
Piecewise polynomial	$d+k+dk$
Continuous piecewise polynomial	$d+dk$
Cubic spline	$3+k$
Natural cubic spline	$k-1$

Notation

$X \sim \text{Name(parameters)}$ represents X follows a "Name" distribution with "parameters" following the parametrization on the exam table.

Probability Models

Symbol	*Description*
$\mathbf{A}^T$	Transpose of matrix **A**
$\mathbf{A}^{-1}$	Inverse of matrix **A**

Statistics

Symbol	*Description*
H_0	Null hypothesis
H_1	Alternative hypothesis
α	Significance level
$t.s.$	Test statistic
h	Hypothesized value
df	Degrees of freedom
ndf	Numerator degrees of freedom
ddf	Denominator degrees of freedom
$t_{2(1-q),\text{df}}$	$100q^{\text{th}}$ percentile of a t-distribution
$F_{1-q,\text{ndf},\text{ddf}}$	$100q^{\text{th}}$ percentile of an F-distribution
$\chi^2_{q,\text{df}}$	$100q^{\text{th}}$ percentile of a chi-square distribution
se	Estimated standard error

Extended Linear Models

Symbol	*Description*
n	# of observation
p	# of predictors
SST	Total sum of squares
SSR	Regression sum of squares
SSE/RSS	Error sum of squares
SS	Sum of squares
MS	Mean square
$\text{E}[Y]$, μ	Mean response
$g(\mu)$	Link function
$l(\hat{\boldsymbol{\beta}})$	Maximized log-likelihood
l_{null}	Maximized log-likelihood for null model
l_{sat}	Maximized log-likelihood for saturated model
$\mathbf{I}$	Information matrix
D	Deviance statistic