Probability

REVIEW OF PROBABILITY

Cumulative distribution function:	$F(x) = \Pr(X \le x)$		
Survival function:	$S(x) = 1 - F(x) = \Pr(X > x)$		
Probability density function:	$f(x) = \frac{d}{dx}F(x) = -\frac{d}{dx}S(x)$		
Hazard rate function:	$\lambda(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx}\log S(x)$	$\rightarrow$	$S(x) = e^{-\int_{-\infty}^{x}\lambda(t)dt}$
Cumulative hazard function:	$\Lambda = \int_{-\infty}^{x}\lambda(t)dt$	$\rightarrow$	$S(x) = e^{-\Lambda(x)}$
Expectation of discrete X:	$E[X] = \sum x \Pr(X = x)$	$\rightarrow$	$E[g(X)] = \sum g(x) \Pr(X = x)$
Expectation of continuous X:	$E[X] = \int_{-\infty}^{\infty} xf(x)dx$	$\rightarrow$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$
Variance:	$Var(X) = E[X^2] - E[X]^2$		
Standard deviation:	$SD(X) = \sqrt{Var(X)}$		
Covariance:	$Cov(X, Y) = E[XY] - E[X]E[Y]$		
Correlation coefficient:	$Corr(X, Y) = \frac{Cov(X,Y)}{SD(X)\ SD(Y)}$		
Linear combination:	$E[W] = aE[X] + bE[Y]$		
$W = aX + bY$	$Var(W) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$		
Sum of iid X:	$E[S] = nE[X]$		
$S = X_1 + \cdots + X_n$	$Var(S) = nVar(X)$		
Conditional probability:	$\Pr(X = x\|Y = y) = \frac{\Pr(X=x,Y=y)}{\Pr(Y=y)} = \frac{\Pr(X=x)\Pr(Y=y\|X=x)}{\Pr(Y=y)}$		
Conditional density function:	$f(x\|y) = \frac{f(x,y)}{f(y)} = \frac{f(x)f(y\|x)}{f(y)}$		
Conditional expectations:	$E[X\|Y = y] = \sum x \Pr(X = x\|Y = y)$		Discrete X
	$E[X\|Y = y] = \int_{-\infty}^{\infty} xf(x\|y)dx$		Continuous X
Law of total probability:	$\Pr(X \le x) = E[\Pr(X \le x\|Y)]$		
Law of total expectation:	$E[X] = E[E[X\|Y]]$		
Law of total variance:	$Var(X) = E[Var(X\|Y)] + Var(E[X\|Y])$		

Normal approximation: $S \dot{\sim} N(\mu = E[S],\ \sigma^2 = Var(S)) \quad \rightarrow \quad \Pr(S \le k) = N(\frac{k-\mu}{\sigma})$

Continuity correction: $\Pr(X \le k) \rightarrow \Pr(X < k + 0.5) \quad \rightarrow \quad \Pr(X < k) \rightarrow \Pr(X < k - 0.5)$

$\Pr(X \ge k) \rightarrow \Pr(X > k - 0.5) \quad \rightarrow \quad \Pr(X > k) \rightarrow \Pr(X > k + 0.5)$

Order statistics pdf: $f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!}(F_X(x))^{k-1} f_X(x)(1 - F_X(x))^{n-k}$

For iid X_i's

Order statistics CDF: $F_{X_{(n)}}(x) = \Pr(X_1 \le x) \cdots \Pr(X_n \le x) = \Pr(X \le x)^n$

For iid X_i's $F_{X_{(1)}}(x) = 1 - \Pr(X_{(1)} > x) = 1 - \Pr(X_1 > x) \cdots \Pr(X_n > x) = 1 - \Pr(X > x)^n$

INSURANCE LOSSES AND PAYMENTS

Loss: X

Payment per loss:

$$Y^L = X - X \wedge d = \begin{cases} 0 & X < d \\ x - d, & X \ge d \end{cases}$$ with deductible d

$$Y^L = X \wedge u = \begin{cases} X, & X < u \\ u, & X \ge u \end{cases}$$ with maximum covered loss u

$$Y^L = X \wedge u - X \wedge d = \begin{cases} 0 & X < d \\ X - d, & d \le X < u \\ X - u, & X \ge u \end{cases}$$ with both u and d.

Payment per payment: $Y^P = Y^L | X > d \quad \rightarrow \quad E[Y^P] = \frac{E[Y^L]}{\Pr(X > d)}$

COMMONLY USED CONTINUOUS DISTRIBUTIONS

Uniform	$f(x) = \frac{1}{b-a}, a \le x \le b$	$F(x) = \frac{x-a}{b-a}$	$E[X] = \frac{a+b}{2}$	$Var(X) = \frac{(b-a)^2}{12}$
Exponential	$f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, x > 0$	$F(x) = 1 - e^{-\frac{x}{\theta}}$	$E[X] = \theta$	$Var(X) = \theta^2$
Weibull	$f(x) = \frac{\tau}{x}\left(\frac{x}{\theta}\right)^{\tau} e^{-\left(\frac{x}{\theta}\right)^{\tau}}, x > 0$	$F(x) = 1 - e^{-\left(\frac{x}{\theta}\right)^{\tau}}$		
Gamma	$f(x) = \frac{\left(\frac{1}{\theta}\right)^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\theta}}, x > 0$	$F(x) = \Pr(X^* \ge \alpha)$ X^* is Poisson with $\lambda = \frac{x}{\theta}$ If α is an integer.	$E[X] = \alpha\theta$	$Var(X) = \alpha\theta^2$
Beta	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, $0 < x < 1$		$E[X] = \frac{a}{a+b}$ if a and b are integers.	$E[X^2] = \frac{a(a+1)}{(a+b)(a+b+1)}$ if a and b are integers.

Pareto	$f(x)=\frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}}, x>0$	$F(x)=1-\left(\frac{\theta}{x+\theta}\right)^{\alpha}$	$E[X]=\frac{\theta}{\alpha-1}$ if a is integer.	$E[X^2]=\frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}$
Single P. Pareto	$f(x)=\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, x>\theta$	$F(x)=1-\left(\frac{\theta}{x}\right)^{\alpha}$	$E[X]=\frac{\alpha\theta}{\alpha-1}$	
Lognormal	$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$, $x>0$	$F(x)=N\left(\frac{\log x-\mu}{\sigma}\right)$	$E[X]=e^{\mu+\frac{\sigma^2}{2}}$	$E[X^2]=e^{2\mu+2\sigma^2}$

COMMONLY USED DISCRETE DISTRIBUTIONS

Poisson	$P(x)=\frac{e^{-\lambda}\lambda^x}{x!}, x=0,1,2,\ldots,\infty$	$E[X]=\lambda$	$Var(X)=\lambda$
Binomial	$P(x)=\binom{m}{x}q^x(1-q)^{m-x}, x=0,1,2,\ldots,\infty$	$E[X]=mq$	$Var(X)=mq(1-q)$
Geometric	$P(x)=\left(\frac{1}{1+\beta}\right)\left(\frac{\beta}{1+\beta}\right)^x, x=0,1,2,\ldots,\infty$	$E[X]=\beta$	$Var(X)=\beta(1+\beta)$
Negative Binomial	$P(x)=\binom{x+r-1}{x}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^x, x=0,1,2,\ldots,\infty$	$E[X]=r\beta$	$Var(X)=r\beta(1+\beta)$

Stochastic Processes

MARKOV CHAINS

Markov chain property: X_t only depends on X_{t-1}.

Transition probability: $P_{ij}=\Pr(X_1=j|X_0=i)=\Pr(X_{t+1}=j|X_t=i)$

Chapman-Kolmogorov Equations: $P_{ij}^t=\sum_{k=1}^{n}P_{ik}^u\,P_{kj}^{t-u}$

Gambler's ruin: Let $p=\Pr(\text{Win 1 chip at each round})$

Let $q=\Pr(\text{Lose 1 chip at each round})$

$$\Pr(\text{Win } N\text{chips}|\text{Currently have } j \text{ chips})=P_j=\begin{cases}\frac{j}{N}, & \frac{q}{p}=1\\ \frac{(\frac{q}{p})^j-1}{(\frac{q}{p})^N-1}, & \frac{q}{p}\neq 1\end{cases}$$

$\Pr(\text{Lose all chips}|\text{Currently have } j \text{ chips})=1-P_j$

Algorithmic efficiency: Let N_j = number of steps to the best solution, given that there are j-1 better solutions.

$E[N_j]=1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{j-1}$

$Var(N_j)=1(1-1)+\frac{1}{2}(1-\frac{1}{2})+\frac{1}{3}(1-\frac{1}{3})+\cdots+\frac{1}{j-1}(1-\frac{1}{j-1})$

CLASSIFICATION OF STATES

Classification of states:

- An **absorbing** state is one that cannot be exited.
- State j is **accessible** from state i if the probability of eventually going to state j from state i is greater than 0.
- Two states **communicate** if each state is accessible from the other.
- A **class** of states is a maximal set of states that communicate with each other. An absorbing state is a class by itself.
- A chain is **irreducible** if it has only one class.
- A state is **recurrent** if the probability of ever reentering the state is 1.
- A state is **transient** if it is not recurrent.

A **finite Markov chain** must have at least one recurrent class. If it is irreducible, then it is recurrent.

Reenter a transient state:

Let N be the number of times a life reenters a transient state

Let p be the probability of reentering a transient state.

$$\Pr(N > n) = p^{n+1}$$

$$E[N] = \sum_{n=0}^{\infty} \Pr(N > n) = \sum_{n=0}^{\infty} p^{n+1} = \frac{p}{1-p}$$

Random Walks:

Let p be the probability of moving up one state in each period.

Let $q = 1 - p$ be the probability of moving down one state in each period.

When $p = q = 0.5$, the chain is recurrent.

When $pq < 0.25$, the chain is transient.

LONG-RUN PROPORTION OF TIME

Consider a recurrent state i:

Let N_i be the number of transitions until the state recurs.

$E[N_i]$ is finite $\rightarrow$ The state is **positive recurrent.**

Otherwise $\rightarrow$ The state is **null recurrent.**

Consider an irreducible and recurrent Markov chain.

Long-run proportion of time a chain is in state i: $\pi_i = \sum_{j=1}^{k} \pi_j \, P_{ji}$

Note that: $\pi_1 + \pi_2 + \cdots + \pi_k = 1$

Consider an aperiodic positive recurrent irreducible Markov chain.

Limiting probability $\alpha_i = \lim_{n\to\infty} \Pr(X_n = i)$

a chain is in state i:

Note that: $\alpha_i = \pi_i$

Note the difference: The limiting probability of being in state i is π_i

The limiting probability of transitioning from state i to state j is $\pi_i \times P_{ij}$.

TIME REVERSIBILITY

Long-run proportions of time π_j are often called stationary probabilities.

If the Markov chain has been around for a long time and is ergodic, then the states have stationary probabilities π_j.

Recall Bayes' theorem: $\Pr(X_{t-1} = j|X_t = i) = \dfrac{\Pr(X_{t-1} = j)\Pr(X_t = i|X_{t-1} = j)}{\Pr(X_t = i)}$

Define the following: $Q_{ij} = \Pr(X_{t-1} = j|X_t = i)$ from i to j, go backward

$P_{ij} = \Pr(X_t = j|X_{t-1} = i)$ from i to j, go forward

We have: $Q_{ij} = \frac{\pi_j\ P_{ji}}{\pi_i} \rightarrow \pi_i\ Q_{ij} = \pi_j\ P_{ji}$

If Q = P, then P is said to be time-reversible.

If the matrix is time-reversible:

$$P_{12}\ P_{23}\ P_{31} = Q_{12}\ Q_{23}\ Q_{31} = \frac{\pi_2\ P_{21}}{\pi_1}\ \frac{\pi_3\ P_{32}}{\pi_2}\ \frac{\pi_1\ P_{13}}{\pi_3} = P_{13}\ P_{32}\ P_{21}$$

TIME IN TRANSIENT STATES

Inverting a matrix: $\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{pmatrix}$

Expected time in transient states:

$S = (I - P_T)^{-1}$ where P_T is the transient-state submatrix.

$S = \begin{pmatrix} s_{22} & s_{23} \\ s_{32} & s_{33} \end{pmatrix}$ where s_{ij} is the expected number of periods in state j given that one is currently in state i.

Note that s_{jj} includes the period of being in state j currently.

Probability of ever transitioning to state j for one in state i:

$$f_{ij} = P_{ij} + \sum_{k \neq j} P_{ik}\ f_{kj}$$

$$f_{ij} = \frac{s_{ij}}{s_{jj}} \qquad \text{since } s_{ij} = f_{ij}\ s_{jj}.$$

$$f_{ij} = \frac{s_{ij}-1}{s_{jj}} \qquad \text{since } s_{ij} = 1 + f_{ij}\ s_{jj}.$$

BRANCHING PROCESS

Let J be the number of offspring produced by an individual:

$$P_j = \Pr(J = j)$$

$$\mu = E[J]$$

$$\sigma^2 = Var(J)$$

Let X_n be the size of the n^{th} generation:

$$E[X_n | X_0 = x] = x\mu^n$$

$$Var(X_n | X_0 = x) = \begin{cases} x\sigma^2 \left(\frac{\mu^{n-1} - \mu^{2n-1}}{1-\mu}\right), & \mu \neq 1 \\ xn\sigma^2, & \mu = 1 \end{cases}$$

Probability of extinction, given that $X_0 = 1$:

$$\pi_{\{X_0=1\}} = \begin{cases} 1, & \mu \leq 1 \\ \sum_{j=0}^{\infty} P_j\ (\pi_{\{X_0=1\}})^j, & \mu > 1 \end{cases}$$

Probability of extinction, given that $X_0 = x$:

$$\pi_{\{X_0=x\}} = \left(\pi_{\{X_0=1\}}\right)^x$$

Poisson Process

EXPONENTIAL DISTRIBUTION

Exponential distribution: $X \sim Exp\left(\theta = \frac{1}{\lambda}\right) \rightarrow$

$$f(x) = \tfrac{1}{\theta} e^{-\frac{x}{\theta}},\ x > 0 \quad \text{or} \quad f(x) = \lambda e^{-\lambda x},\ x > 0$$

$$F(x) = 1 - e^{-\frac{x}{\theta}} \quad \text{or} \quad F(x) = 1 - e^{-\lambda x}$$

$$E[X] = \theta \quad \text{or} \quad E[X] = \tfrac{1}{\lambda}$$

$$\text{Var}[X] = \theta^2 \quad \text{or} \quad \text{Var}[X] = \left(\tfrac{1}{\lambda}\right)^2$$

Lack of memory: $X - k | X > k \sim Exp(\theta) \rightarrow \Pr(X > k + x | X > k) = \Pr(X > x)$

With limit u: $E\left[Y^L\right] = E[X \wedge u] = \theta\left(1 - e^{-\frac{u}{\theta}}\right)$

With deductible d: $E\left[Y^L\right] = E[X] - E[X \wedge d] = \theta - \theta\left(1 - e^{-\frac{d}{\theta}}\right) = \theta e^{-\frac{d}{\theta}} \rightarrow E\left[Y^P\right] = \frac{E\left[Y^L\right]}{\Pr(X>k)} = \theta$

Order statistics: $X_1, \ldots, X_n \overset{iid}{\sim} Exp(\theta) \rightarrow E[X_{(k)}] = \theta\left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + \frac{1}{n-k+1}\right)$

Minimum: $X_i \overset{ind}{\sim} Exp\left(\theta_i = \frac{1}{\lambda_i}\right) \rightarrow \min(X_1, \ldots, X_n) \sim Exp\left(\theta = \frac{1}{\frac{1}{\theta_1}+\frac{1}{\theta_2}+\cdots+\frac{1}{\theta_n}} = \frac{1}{\lambda_1+\lambda_2+\cdots+\lambda_n}\right)$

Greedy algorithm: There are n persons and k jobs to do. Cost of each person $X \sim Exp(\theta)$. $k \leq n$

Total expected cost $= \frac{\theta}{n} + \frac{\theta}{n-1} + \frac{\theta}{n-2} + \cdots + \frac{\theta}{n-k+1}$.

Sum: $X_1, ..., X_n \overset{iid}{\sim} Exp(\theta) \rightarrow X_1 + \cdots + X_n \sim Gamma(\alpha = n,\ \theta)$

Gamma distribution: $X \sim Gamma(\alpha,\ \theta) \rightarrow F_X(x; \alpha = 1) = 1 - e^{-\frac{x}{\theta}}$

$$F_X(x; \alpha = 2) = 1 - e^{-\frac{x}{\theta}} - \frac{e^{-\frac{x}{\theta}}\left(\frac{x}{\theta}\right)^1}{1!}$$

$$F_X(x; \alpha = 3) = 1 - e^{-\frac{x}{\theta}} - \frac{e^{-\frac{x}{\theta}}\left(\frac{x}{\theta}\right)^1}{1!} - \frac{e^{-\frac{x}{\theta}}\left(\frac{x}{\theta}\right)^2}{2!}$$

Chi-square: $X \sim Gamma\left(\alpha = \frac{n}{2},\ \theta = 2\right) \rightarrow X \sim \chi^2(n)$

POISSON PROCESS

Poisson distribution: $X \sim Poisson(\lambda) \rightarrow \Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots, \infty$

$$E[X] = Var(X) = \lambda$$

Sum: $X_i \overset{ind}{\sim} Poisson(\lambda_i) \rightarrow X_1 + \cdots + X_n \sim Poisson(\lambda_1 + \cdots + \lambda_n)$

Counting process: $X(t)$ is the number of events that occur at or before time t.

Poisson process: $X(t) \sim Poisson(\lambda(t))$

$X(t) - X(s)$ is independent of $X(v) - X(u)$. For $t > s > v > u > 0$.

$X(t+s) - X(t)$ is a Poisson random variable. For $s > 0$.

Homogeneous PP: $X(t) \sim Poisson(\lambda(t) = \lambda)$ λ is a constant.

TIME TO NEXT EVENT

Homogeneous PP with parameter λ: $X(t+s) - X(t) \sim Poi(\lambda s)$

Interarrival time/Time to next event: $T \sim Exp\left(\theta = \frac{1}{\lambda}\right) \rightarrow F_T(s) = 1 - e^{-\lambda s}$

Time to nth event: $T \sim Gamma\left(\alpha = n,\ \theta = \frac{1}{\lambda}\right)$

Nonhomogeneous PP with intensity function $\lambda(t)$:

$X(t+s) - X(t) \sim Poi(\lambda^*)$ where $\lambda^* = \int_t^{t+s} \lambda(r)\, dr$

Interarrival time/Time to next event: T is not exponential. $\rightarrow F_T(s) = 1 - e^{-\lambda^*}$

Note the difference: At time 0, $\Pr(\text{Next arrival between time 2 and time 4}) = e^{-\int_0^2 \lambda(r)dr} - e^{-\int_0^4 \lambda(r)dr}$

At time 2, $\Pr(\text{Next arrival before time 4}) = 1 - e^{-\int_2^4 \lambda(r)dr}$

Given that exactly k independent Poisson events occurred before time τ:

Time of each event $T_j \sim Unif(0,\ \tau)$.

The joint distribution of event times is $f_{T_1,\ldots,T_k}(t_1,\ldots,t_k) = \left(\frac{1}{\tau}\right)^k$.

The **expected time of the j^{th} event** is $E\left[X_{(j)}\right] = \frac{j\tau}{k+1}$.

SUMS, THINNING, AND MIXTURES

Sum of ind. PPs: $X_i(t)$'s are independent Poisson processes with $\lambda_i(t)$'s.

$X_1(t) + \cdots + X_n(t)$ is a Poisson process with $\lambda(t) = \lambda_1(t) + \cdots + \lambda_n(t)$.

Thinning/splitting: $X(t)$ is a Poisson process with intensity function $\lambda(t)$.

$X_A(t)$ is a Poisson process with intensity function $P(A) \times \lambda(t)$.

Sum of two ind. PPs: $X_A(t)$ is a Poisson process with parameter λ_A.

$X_B(t)$ is a Poisson process with parameter λ_B.

$X_A(t) + X_B(t)$ is a Poisson process with parameter $\lambda_A + \lambda_B$,
with $P(A) = \frac{\lambda_A}{\lambda_A+\lambda_B}$ and $P(B) = \frac{\lambda_B}{\lambda_A+\lambda_B}$.

Mixture of two PPs: $X_A(t)$ is a Poisson process with parameter λ_A.

$X_B(t)$ is a Poisson process with parameter λ_B.

However, $X(t)$ a mixture of $X_A(t)$ and $X_B(t)$ is NOT a Poisson process.

Negative binomial distribution: Let $X|\lambda \sim Poisson(\lambda)$, where $\lambda \sim Gamma(\alpha,\ \theta)$.

Then $X \sim NB(r = \alpha,\ \beta = \theta)$. ← which is NOT Poisson

COMPOUND POISSON PROCESS

Define the following:

N is a Poisson random variable. Primary r.v.

X_i's all follow the same distribution. Secondary r.v.

N and X_i's are all independent.

Compound Poisson r.v.: $S = X_1 + \cdots + X_N$

Mean: $E[S] = E[N]E[X] \rightarrow E[S] = \lambda E[X]$

Variance: $Var(S) = E[N]Var(X) + Var(N)E[X]^2 \rightarrow Var(S) = \lambda E[X^2]$

Reliability

STRUCTURE FUNCTIONS

State vector: $\boldsymbol{x} = (x_1, \ldots, x_n)$ where $x_i = 1$ if it functions, and $x_i = 0$ if it fails.
Note that $x_i^k = x_i$ for $k \geq 1$.
Assume that $x_1, \ldots, x_n$ are all mutually independent.

Structure function: $\phi(\boldsymbol{x}) = 1$ if the structure functions.

Minimal path sets: The system functions when all the components of a minimal path set $\boldsymbol{u}_j$ function.

Minimal cut sets: The system fails when all the components of a minimal cut set $\boldsymbol{v}_j$ fail.

Series system: Functions when n out of n components function.

$\phi(\boldsymbol{x}) = \prod x_i$

The entire system is a minimal path set.

Each component is a minimal cut set.

Parallel system: Functions when 1 out of n components functions.

$\phi(\boldsymbol{x}) = 1 - \prod(1 - x_i)$

Every component is a minimal path set.

The entire system is a minimal cut set.

k out of n system: Functions when k out of n components function.

There are $\binom{n}{k}$ minimal path sets.

There are $\binom{n}{k-k+1}$ minimal cut sets.

Two ways to express a system:

1. We have J minimal path sets $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_J \rightarrow \phi(\boldsymbol{u}_j) = \prod_i u_{ji} \rightarrow \phi(\boldsymbol{x}) = 1 - \prod_j (1 - \phi(\boldsymbol{u}_j))$
2. We have J minimal cut sets $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J \rightarrow \phi(\boldsymbol{v}_j) = 1 - \prod_i (1 - v_{ji}) \rightarrow \phi(\boldsymbol{x}) = \prod_j \phi(\boldsymbol{v}_j)$

PROBABILITIES

Bernoulli probability: $p_i = \Pr(x_i = 1) \rightarrow x_i \sim Bernoulli(p_i) \rightarrow E[x_i] = p_i$

Reliability function: $r(\boldsymbol{p}) = \Pr(\phi(\boldsymbol{x}) = 1) \rightarrow \phi(\boldsymbol{x}) \sim Bernoulli(r(\boldsymbol{p}))) \rightarrow E[\phi(\boldsymbol{x})] = r(\boldsymbol{p})$

Series system: $r(\boldsymbol{p}) = \prod p_i$

Parallel system: $r(\boldsymbol{p}) = 1 - \prod (1 - p_i)$

k out of n system: $r(\boldsymbol{p}) = \sum_{k \text{ to } n} \binom{n}{k} p^k (1-p)^{n-k}$ Assume that $p_1 = p_2 = \cdots = p_n = p$.

Example (Inclusion/Exclusion Bounds):

The minimal path sets are $\{1,\ 3\}, \{1,\ 4\}$ and $\{2\}$.

Structure function:
$$\begin{aligned}\varphi(x) &= 1 - (1 - x_1x_3)(1 - x_1x_4)(1 - x_2) \\ &= x_1x_3 + x_1x_4 + x_2 - x_1x_3x_4 - x_1x_2x_3 - x_1x_2x_4 + x_1x_2x_3x_4\end{aligned}$$

Reliability function: $r(p) = p_1p_3 + p_1p_4 + p_2 - p_1p_3p_4 - p_1p_2p_3 - p_1p_2p_4 + p_1p_2p_3p_4$

First upper bound: $r(p) \leq p_1p_3 + p_1p_4 + p_2$

First lower bound: $r(p) \geq p_1p_3 + p_1p_4 + p_2 - p_1p_3p_4 - p_1p_2p_3 - p_1p_2p_4$

Second upper bound: $r(p) \leq p_1p_3 + p_1p_4 + p_2 - p_1p_3p_4 - p_1p_2p_3 - p_1p_2p_4 + p_1p_2p_3p_4$

The minimal cut sets are $\{1,\ 2\}$ and $\{2,\ 3,\ 4\}$.

Structure function:
$$\begin{aligned}\varphi(x) &= (1 - (1 - x_1)(1 - x_2))(1 - (1 - x_2)(1 - x_3)(1 - x_4)) \\ &= 1 - (1 - x_1)(1 - x_2) - (1 - x_2)(1 - x_3)(1 - x_4) \\ &\quad + (1 - x_1)(1 - x_2)(1 - x_3)(1 - x_4)\end{aligned}$$

Reliability function:
$$\begin{aligned}r(p) &= 1 - (1 - p_1)(1 - p_2) - (1 - p_2)(1 - p_3)(1 - p_4) \\ &\quad + (1 - p_1)(1 - p_2)(1 - p_3)(1 - p_4)\end{aligned}$$

First lower bound: $r(p) \geq 1 - (1 - p_1)(1 - p_2) - (1 - p_2)(1 - p_3)(1 - p_4)$

First upper bound:
$$\begin{aligned}r(p) &\leq 1 - (1 - p_1)(1 - p_2) - (1 - p_2)(1 - p_3)(1 - p_4) \\ &\quad + (1 - p_1)(1 - p_2)(1 - p_3)(1 - p_4)\end{aligned}$$

Example (Intersections Bounds):

The minimal path sets are $\{1,\ 3\}, \{1,\ 4\}$ and $\{2\}$.

Structure function: $\phi(x) = 1 - (1 - x_1x_3)(1 - x_1x_4)(1 - x_2)$

Upper bound: $r(p) \leq 1 - (1 - p_1p_3)(1 - p_1p_4)(1 - p_2)$

The minimal cut sets are $\{1,\ 2\}$ and $\{2,\ 3,\ 4\}$.

Structure function: $\phi(x) = (1 - (1 - x_1)(1 - x_2))(1 - (1 - x_2)(1 - x_3)(1 - x_4))$

Lower bound: $r(p) \geq (1 - (1 - p_1)(1 - p_2))(1 - (1 - p_2)(1 - p_3)(1 - p_4))$

TIME TO FAILURE

Expectation formula: $E[T] = \int_0^\infty \Pr(T > t)dx$

Time to failure: T is a continuous random variable.

Survival probability: $p_i(t) = \Pr(T_i > t) \quad \rightarrow \quad E[T_i] = \int_0^\infty p_i(t)dx$

Reliability function at time t: $r(\boldsymbol{p}(t)) \quad \rightarrow \quad E[\text{ system life }] = \int_0^\infty r(\boldsymbol{p}(t))\,dx$

Note that T_i and x_i are NOT the same thing. We have $x_i = 1$ when $T_i > t$. Thus, $\Pr(x_i = 1) = \Pr(T_i > t) = p_i(t)$.

Series system: $r(\boldsymbol{p}(t)) = \prod p_i(t) = \prod S_i(t)$

Parallel system: $r(\boldsymbol{p}(t)) = 1 - \prod(1 - p_i(t)) = 1 - \prod F_i(t)$

For a k out of n system with $T_i \overset{iid}{\sim} Exp(\theta)$: $E[\text{ system life }] = E\left[T_{(n-k+1)}\right] = \theta\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{k}\right)$

Failure rate function: $\lambda(t) = \frac{f(t)}{S(t)} = \frac{-\frac{d}{dt}r(\boldsymbol{p}(t))}{r(\boldsymbol{p}(t))}$

Hazard function: $\Lambda(t) = \int_0^t \lambda(s)\,ds = \Lambda(t) = -\log r(\boldsymbol{p}(t)) \quad \rightarrow \quad r(\boldsymbol{p}(t)) = e^{-\Lambda(t)}$

Life Contingencies

SURVIVAL MODELS

Actuarial notation:

$${}_tp_x = \Pr(T_x > t)$$

$${}_tq_x = 1 - {}_tp_x = \Pr(T_x \leq t)$$

Formulas:

$${}_{t+u}p_x = {}_tp_x\ {}_up_{x+t}$$

$${}_{t|u}q_x = {}_tp_x\ {}_uq_{x+t} = {}_tp_x - {}_{t+u}p_x = {}_{t+u}q_x - {}_tq_x$$

Number of lives: $L_{x+t} \sim Binomial\left(l_x,\ {}_tp_x\right) \quad \rightarrow \quad E\left[L_{x+t}\right] = l_x\ {}_tp_x \qquad Var\left(L_{x+t}\right) = l_x\ {}_tp_x\ {}_tq_x$

Life table:

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

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

$${}_{t|u}q_x = \frac{l_{x+t} - l_{x+t+u}}{l_x}$$

Expectation:

$$e_x = \sum_{k=1}^{\infty} k_k p_x\ q_{x+k} = \sum_{k=1}^{\infty} {}_kp_x$$

INSURANCE

Endowment function:

$${}_nE_x = v^n\ {}_np_x$$ (Pure endowment of 1)

Insurance functions:

$$A_x = \sum_{k=0}^{\infty} v^{k+1}\ {}_kp_x\ q_{x+k}$$ (Whole life insurance of 1)

$$A_{\overset{1}{x:\overline{n}|}} = \sum_{k=0}^{n-1} v^{k+1}\ {}_kp_x\ q_{x+k}$$ (n-year term insurance of 1)

$$A_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^{k+1}\ {}_kp_x\ q_{x+k} + v^n\ {}_np_x$$ (n-year endowment insurance of 1)

Relations:

$$A_x = A_{\overset{1}{x:\overline{n}|}} + {}_nE_x\ A_{x+n}$$

$$A_{x:\overline{n}|} = A_{\overset{1}{x:\overline{n}|}} + {}_nE_x$$

$${}_{n|}A_x = {}_nE_x\ A_{x+n}$$ (n-year deferred whole life insurance of 1)

Recursive formulas:

$$A_x = vq_x + vp_x\ A_{x+1}$$

$$A_{\overset{1}{x:\overline{n}|}} = vq_x + vp_x\ A_{\overset{1}{x+1:\overline{n-1}|}}$$

ANNUITIES

Annuity functions:

$$\ddot{a}_x = \sum_{k=0}^{\infty} \left(\frac{1-v^{k+1}}{d}\right) {}_kp_x\ q_{x+k}$$ (Whole life annuity-due of 1 per year)

$$\ddot{a}_{x:\overline{n}|} = \sum_{k=0}^{n-1} \left(\frac{1-v^{k+1}}{d}\right) {}_kp_x\ q_{x+k} + \left(\frac{1-v^n}{\delta}\right) {}_np_x$$ (n-year term annuity-due of 1 per year)

Simpler formulas:

$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k\ {}_kp_x$$

$$\ddot{a}_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^k\ {}_kp_x$$

Relations:

$$\ddot{a}_x = \ddot{a}_{x:\overline{n}|} + {}_nE_x\ \ddot{a}_{x+n}$$

$${}_{n|}\ddot{a}_x = {}_nE_x\ \ddot{a}_{x+n}$$ (n-year deferred whole life annuity-due of 1 per year)

$$\ddot{a}_{\overline{x:\overline{n}|}} = \ddot{a}_{\overline{n}|} + {}_nE_x\ \ddot{a}_{x+n}$$ (n-year certain whole life annuity-due of 1 per year)

Recursive formulas: $\ddot{a}_x = 1 + vp_x\ \ddot{a}_{x+1}$

$\ddot{a}_{x:\overline{n}|} = 1 + vp_x\ \ddot{a}_{x+1:\overline{n-1}|}$

Insurance to annuity: $\ddot{a}_x = \frac{1-A_x}{d}$

$\ddot{a}_{x:\overline{n}|} = \frac{1-A_{x:\overline{n}|}}{d}$

PREMIUMS

Equivalence principle: EPV(Net Premiums)=EPV(Insurance Benefits)

Premium functions:

$P_x = \frac{A_x}{\ddot{a}_x}$ (Whole life insurance of 1, premiums payable for life)

$P_{\frac{1}{x:\overline{n}|}} = \frac{A^1_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}}$ (n-year term insurance of 1, premiums payable for n years)

$P_{x:\overline{n}|} = \frac{A_{x:\overline{n}|}}{\ddot{a}_{x:\overline{n}|}}$ (n-year endowment insurance of 1, premiums payable for n years)

Note that $v = \frac{1}{1+i}$, where i is the effective annual interest rate.

Simulation

INVERSE TRANSFORMATION METHOD

Simulation generator: $X_{n+1} = aX_n + c \pmod m$

To generate uniform numbers:

1. Specify an initial integer x_0 called the "seed".
2. Calculate $X_1 = ax_0 + c$.
3. Divide X_1 by m, obtain the first remainder x_1.
4. The first uniform number is $u_1 = \frac{x_1}{m}$.
5. Repeat steps 2-4 using x_1 to obtain the second remainder x_2 and the second uniform number $u_2 = \frac{x_2}{m}$. And so on...

Inverse transformation method:

1. Generate uniform numbers $u_1, \ldots, u_n$.
2. Specify a distribution function $F_Y(y) = \Pr(Y \le y)$.
3. Calculate $y_i = F_Y^{-1}(u_i)$.

REJECTION METHOD

Specify the following:

$f(x)$ is the density function of variable to simulate, $F(x)$ is the corresponding CDF.

$g(x)$ is the base density function, $G(x)$ is the corresponding CDF.

General method:

1. Determine $c = \max \frac{f(x)}{g(x)}$.
2. Generate two uniform numbers u_1 and u_2.
3. Calculate $x_1 = G^{-1}(u_1)$.
4. Accept x_1 if $u_2 \leq \frac{f(x_1)/g(x_1)}{c}$.

Gamma distribution:

$f(x)$ is the density function of gamma distribution with mean $\alpha\theta$.

$g(x)$ is the density function of exponential distribution with mean $\theta^* = \alpha\theta$.

Use the general method.

In step 1, $c = \max \frac{f(x)}{g(x)}$ at $x = \theta^* = \alpha\theta$.

Normal distribution:

$f(x)$ is the density function of standard normal distribution.

$g(x)$ is the density function of exponential distribution with mean 1.

Use the following method:

1. Generate three uniform numbers u_1, u_2 and u_3.
2. Calculate $x_1 = -\log u_1$ and $x_2 = -\log u_2$.
3. Accept x_1 if $x_2 \geq \frac{(x_1-1)^2}{2}$.
4. If $u_3 < 0.5 \rightarrow$ use x_1

   If $u_3 \geq 0.5 \rightarrow$ use $-x_1$ Note that x_1 or $-x_1$ is standard normal number.

Estimation

KERNEL DENSITY ESTIMATION

Suppose we are given n observations and the base distribution function $\beta(x)$.

For each observation x_i, we set $k_{x_i}(x) = \frac{1}{b}\beta\left(\frac{x-x_i}{b}\right)$.

The kernel distribution is an equally weighted mixture of the n distributions: $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_{x_i}(x)$

$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x_i}(x)$

Rectangular Kernel	Triangle Kernel	Gaussian Kernel
$\beta(x)=\begin{cases}\frac{1}{2}, & -1\le x\le 1\\ 0, & \text{otherwise}\end{cases}$	$\beta(x)=\begin{cases}x+1, & -1\le x\le 0\\ 1-x, & 0\le x\le 1\\ 0, & \text{otherwise}\end{cases}$	$\beta(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
$k_{x_i}(x)=\begin{cases}\frac{1}{2b}, & x_i-b\le x\le x_i+b\\ 0, & \text{otherwise}\end{cases}$	$k_{x_i}(x)=\begin{cases}\frac{x-(x_i-b)}{b^2}, & x_i-b\le x\le x_i\\ \frac{(x_i+b)-x}{b^2}, & x_i<x\le x_i+b\\ 0 & \text{otherwise}\end{cases}$	$X_i\sim N(\mu=x_i,\ \sigma^2=b^2)$ $k_{x_i}(x)=\frac{1}{\sqrt{2\pi b^2}}e^{-\frac{(x-x_i)^2}{2b^2}}$
$K_{x_i}(x)=\begin{cases}0, & x\le x_i-b\\ \frac{x-(x_i-b)}{2b}, & x_i-b\le x\le x_i+b\\ 1, & x>x_i+b\end{cases}$	$K_{x_i}(x)=\begin{cases}0, & x\le x_i-b\\ \frac{(x-(x_i-b))^2}{2b^2}, & x_i-b\le x\le x_i\\ 1-\frac{((x_i+b)-x)^2}{2b^2}, & x_i<x\le x_i+b\\ 1, & \text{otherwise}\end{cases}$	$K_{x_i}(x)=N\left(\frac{x-x_i}{b}\right)$
$E[X\|X_i=x_i]=x_i$ $Var(X\|X_i=x_i)=\frac{(2b)^2}{12}=\frac{b^2}{3}$	$E[X\|X_i=x_i]=x_i$ $Var(X\|X_i=x_i)=\cdots=\frac{b^2}{6}$	$E[X\|X_i=x_i]=x_i$ $Var(X\|X_i=x_i)=b^2$
$E[X]=\frac{\sum x_i}{n}$ $Var(X)=\frac{b^2}{3}+\left(\frac{\sum x_i^2}{n}-\left(\frac{\sum x_i}{n}\right)^2\right)$	$E[X]=\frac{\sum x_i}{n}$ $Var(X)=\frac{b^2}{6}+\left(\frac{\sum x_i^2}{n}-\left(\frac{\sum x_i}{n}\right)^2\right)$	$E[X]=\frac{\sum x_i}{n}$ $Var(X)=b^2+\left(\frac{\sum x_i^2}{n}-\left(\frac{\sum x_i}{n}\right)^2\right)$

METHOD OF MOMENTS

Set the theoretical moments equal to the sample moments.

To estimate 1 parameter, use the first moment. To estimate 2 parameters, use the first and second moments.

Complete data: $E[X]=\frac{\sum x_i}{n}$ where x_i's are complete losses data.

$E[X^2]=\frac{\sum x_i^2}{n}$

Left-truncated data: $E[X|X>d]=\frac{\sum x_i}{n}$ where x_i's are incomplete losses data.

$E[X-d|X>d]=\frac{\sum y_i}{n}$ where y_i's are claims data.

Right-censored data: $E[X\wedge u]=\frac{\sum x_i}{n}=\frac{\sum y_i}{n}$ where x_i's are incomplete losses data, and y_i's are claims data.

PERCENTILE MATCHING

Set the theoretical percentiles equal to the smoothed empirical percentile.

To estimate 1 parameter, use one percentile. To estimate 2 parameters, use two percentiles.

Smoothed empirical percentile: $\hat{\pi}_p = x_{(k)}$ where $k = p(n+1)$.

If k is not an integer, use interpolation.

Complete data:	$\Pr(X \le \hat{\pi}_p) = p$	where $\hat{\pi}_p$ is calculated using complete losses data.
Left-truncated data:	$\Pr(X \le \hat{\pi}_p \mid X > d) = p$	where $\hat{\pi}_p$ is calculated using incomplete losses data.
	$\Pr(X - d \le \hat{\pi}_p \mid X > d) = p$	where $\hat{\pi}_p$ is calculated using claims data.
Right-censored data:	$\Pr(X \le \hat{\pi}_p) = p$	where $\hat{\pi}_p$ is calculated using incomplete losses data, which are the claims data.

MAXIMUM LIKELIHOOD ESTIMATION

Maximize the **likelihood function** L. The value(s) of parameter(s) that maximizes the likelihood function is called the **maximum likelihood estimate(s)**.

Complete data:	$L = f(x_1) \cdots f(x_n)$	$f = \Pr$ for discrete distribution.
Grouped data:	$L = (F(c_1) - F(c_0)) \cdots (F(c_n) - F(c_{n-1}))$	where $c_0 < c_1 < \ldots < c_n$ are interval boundaries.
Left-truncated data:	$L = \frac{f(x_1)}{S(d)} \cdots \frac{f(x_n)}{S(d)}$	There are n observations with exact values. They are all greater than d.
Right-censored data:	$L = f(x_1) \cdots f(x_n) S(u)^m$	There are n observations with exact values and m observations greater than u.
Left-truncated and Right-censored data:	$L = \frac{f(x_1)}{S(d)} \cdots \frac{f(x_n)}{S(d)} \left(\frac{S(u)}{S(d)}\right)^m$	There are n observations with exact values and m observations greater than u. These $n+m$ observations are all greater than d.

For distributions that belong to the exponential family:

1. Determine $L(\theta)$.
2. Apply natural logarithm, obtain $l(\theta) = \log L(\theta)$.
3. Take the first derivative with respect to the parameter, obtain $l'(\theta)$.
4. Find the value of θ such that $l'(\theta) = 0$.

With complete data:

Distribution	Maximum likelihood estimate(s)	Method of moment estimate(s)
Exponential	$\hat{\theta} = \bar{x}$	Same.
Normal	$\hat{\mu} = \bar{x}$ $\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \hat{\mu})^2$	Same.
Lognormal	$\hat{\mu} = \frac{1}{n}\sum \log x_i$ $\hat{\sigma}^2 = \frac{1}{n}\sum (\log x_i - \hat{\mu})^2$	
Uniform	$\hat{b} = \max(x_1, \ldots, x_n)$	
Binomial	$\hat{q} = \frac{\bar{x}}{m}$	Same.
Poisson	$\hat{\lambda} = \bar{x}$	Same.

ESTIMATOR QUALITY

Estimator vs estimate: Prior to knowing the observations, one must first obtain an estimator using MOM, PM, MLE, etc.

An estimator is a random variable, and an estimate is just an outcome of the estimator.

Bias: The bias of an estimator is $\text{Bias} = E\left[\hat{\theta}\right] - \theta$.

An estimator is **unbiased** if $\text{Bias} = 0$.

An estimator is **asymptotically unbiased** if $\lim_{n\to\infty} \text{Bias} = 0$.

Consistency: An estimator is (weakly) consistent if $\lim_{n\to\infty} \Pr\left(\left|\hat{\theta} - \theta\right| < \epsilon\right) = 1$.

If $\lim_{n\to\infty} \text{Bias}\left(\hat{\theta}\right) = 0$ and $\lim_{n\to\infty} Var\left(\hat{\theta}\right) = 0$, then the estimator is consistent. However the converse is not true.

Efficiency: An estimator is more efficient than another estimator of its variance is lower.

The **relative efficiency** of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is $\text{RE} = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$.

The **mean square error** of an estimator is $\text{MSE} = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var\left(\hat{\theta}\right) + \text{Bias}^2$.

A uniformly minimum variance unbiased estimator (MVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.

RAO-CRAMER LOWER BOUND

Fisher information matrix: $I = -E\left[\frac{d^2 l(\boldsymbol{x};\theta)}{d\theta^2}\right] = E\left[\left(\frac{d\ l(\boldsymbol{x};\theta)}{d\theta}\right)^2\right]$ where $l(\boldsymbol{x};\theta) = \sum \log f(x_i;\theta)$ is the log-likelihood function.

Rao-Cramer lower bound: I^{-1} This is the lowest possible variance of an unbiased estimator $\hat{\theta}$.

The efficiency of an unbiased estimator is: $E = \frac{I^{-1}}{Var(\hat{\theta})}$

An unbiased estimator is efficient if: $E = 1$ In this case, this estimator is the MVUE.

SUFFICIENT STATISTICS

A **sufficient statistic** is a function $T(\boldsymbol{x})$ whose value contains all the information needed to compute any estimate of the parameter.

Rao-Blackwell Theorem: For any unbiased estimator $\hat{\theta}$ and sufficient statistic $T(\boldsymbol{x})$, the estimator $E\left[\hat{\theta}\middle|T(\boldsymbol{x})\right]$ is the MVUE.

This means that $E\left[\hat{\theta}\middle|T(\boldsymbol{x})\right]$ is at least as efficient as $\hat{\theta}$.

Maximum likelihood function: $L(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)\,h(\boldsymbol{x})$ where $h(\boldsymbol{x})$ is a function that does not depend on θ.

The MLE (if unique) is a function of a sufficient statistic $T(\boldsymbol{x})$:

1. Determine the MLE.
2. Determine the mean of MLE.
   - If the MLE is unbiased, the MLE is the MVUE.
   - If the MLE is not unbiased, obtain an unbiased estimator by adjusting the MLE. This new estimator is the MVUE.

Exponential family: $f(x;\theta) = e^{p(\theta)q(x)+r(\theta)+s(x)}$ for $a < x < b$

It is a regular case if:

- a and b do not depend on θ.
- $p(\theta)$ is nontrivial and continuous.
- $q(x)$ is nontrivial.

For such a family, $\hat{\theta} = \sum q(x_i)$ is a sufficient statistic.

BOOTSTRAP METHOD

A method for estimating the standard error of an estimator.

Suppose we have n observations:

1. Select n observations, with replacement, at random. Calculate the statistic $\hat{\alpha}$.
2. Repeat a total of m times to obtain $\hat{\alpha}_i$ for $i = 1, 2, \ldots, m$.
3. Estimate the standard error of $\hat{\alpha}$.

Estimate of standard error: $SE(\hat{\alpha}) = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\alpha}_i - \bar{\alpha})^2}$ where $\bar{\alpha} = \frac{1}{m}\sum_{i=1}^{m}\hat{\alpha}_i$

Hypothesis Testing

HYPOTHESES

Null hypothesis: H_0

Alternative hypothesis: H_1

The following are the same thing:

- Probability of Type I error
- Size of critical region
- Significance level
- α

Terminology:

	Accept H_0	Reject H_0
H_0 true	$1-\alpha$	α Pr(Type I error)
H_1 true	β Pr(Type II error)	$1-\beta$ Power of test

Probabilities:

- $\alpha = \Pr(\text{Reject } H_0 | H_0 \text{ true})$
- P-value $= \Pr(\text{Observations} | H_0 \text{ true})$
  P-value $= 2\Pr(\text{Observations} | H_0 \text{ true})$ for two-tailed test.
- Power $= \Pr(\text{Reject } H_0 | H_1 \text{ true})$

Reject H_0 if:

- Observations fall within the critical region.
- Test statistic is beyond the critical point.
- p-value $\leq \alpha$

Note that α increases $\rightarrow$ Power increases $\rightarrow$ β decreases. One can decrease both α and β by increasing the sample size.

HYPOTHESIS TESTS FOR MEANS, PROPORTIONS AND VARIANCES

Normal distribution: $X \sim N(\mu, \sigma^2) \rightarrow \frac{X-\mu}{\sigma} \sim N(0,1) \rightarrow \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

χ^2 distribution: $\sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$

$\sum_{i=1}^{n} \left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1) \rightarrow \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$

Student's t distribution: $\frac{N(0,1)}{\sqrt{\frac{\chi^2(n)}{n}}} \sim t(n) \rightarrow \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$

F distribution: $\frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2} \sim F(n_1,\ n_2) \rightarrow \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n_X-1,\ n_Y-1)$

Distribution	Null hypothesis	Conditions	Statistic	Confidence Interval
$X \sim N(\mu,\ \sigma^2)$	$\mu = \mu_0$	σ^2 known	$z = \frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}$	$\bar{x} \pm z\sqrt{\frac{\sigma^2}{n}}$
$X \sim N(\mu,\ \sigma^2)$	$\mu = \mu_0$	σ^2 unknown	$t = \frac{\bar{x}-\mu_0}{\sqrt{\frac{S^2}{n}}}$ DOF $= n-1$	$\bar{x} \pm t\sqrt{\frac{s^2}{n}}$
$X \sim N(\mu_X,\ \sigma_X^2)$ $Y \sim N(\mu_Y,\ \sigma_Y^2)$	$\mu_X - \mu_Y = 0$	σ_X^2 known σ_Y^2 known	$z = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}}$	$(\bar{x}-\bar{y}) \pm z\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$
$X \sim N(\mu_X,\ \sigma_X^2)$ $Y \sim N(\mu_Y,\ \sigma_Y^2)$	$\mu_X - \mu_Y = 0$	σ_X^2 unknown σ_Y^2 unknown $\sigma_X^2 = \sigma_Y^2$	$t = \frac{\bar{x}-\bar{y}}{\sqrt{s^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$ $s^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$ DOF $= n_x+n_y-2$	$(\bar{x}-\bar{y}) \pm t\sqrt{S^2\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}$
$X \sim Bin(n,\ p)$	$p = p_0$		$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
$X \sim Bin(n_X,\ p_X)$ $Y \sim Bin(n_Y,\ p_Y)$	$p_X - p_Y = 0$		$z = \frac{\hat{p}_x-\hat{p}_y}{\sqrt{\widetilde{p}(1-\widetilde{p})\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$ $\widetilde{p} = \frac{x+y}{n_x+n_y}$	$(\hat{p}_x-\hat{p}_y) \pm z\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$
$X \sim N(\mu,\ \sigma^2)$	$\sigma^2 = \sigma_0^2$	μ unknown	$\chi^2 = \frac{s^2}{\sigma_0^2}(n-1)$ DOF $= n-1$	$\left(\frac{s^2}{\chi^2_{n-1,\ 1-\frac{\alpha}{2}}}(n-1),\ \frac{s^2}{\chi^2_{n-1,\ \frac{\alpha}{2}}}(n-1)\right)$
$X \sim N(\mu_X,\ \sigma_X^2)$ $Y \sim N(\mu_Y,\ \sigma_Y^2)$	$\frac{\sigma_X^2}{\sigma_Y^2} = 1$	μ_X unknown μ_Y unknown	$F = \frac{s_x^2}{s_y^2}$ DOF $= (n_x-1,\ n_y-1)$	$\left(\frac{1}{F_{n_x-1,n_y-1,1-\frac{\alpha}{2}}}\frac{s_x^2}{s_y^2},\ \frac{1}{F_{n_x-1,n_y-1,\frac{\alpha}{2}}}\frac{s_x^2}{s_y^2}\right)$ $\frac{1}{F_{n_x-1,n_y-1,1-\frac{\alpha}{2}}} = F_{n_y-1,n_x-1,\frac{\alpha}{2}}$

KOLMOGOROV-SMIRNOV TEST

Null hypothesis: The parametric model fits its data well

Example:

Observations are 1.7, 1.6, 1.6 and 1.9.

Parametric CDF is $F_\theta(x) = \frac{x^2}{4}$.

x	$F_e(x^-) - F_e(x)$	$F_\theta(x)$	MAX Difference
1.6	0.00-0.50	0.64	0.64
1.7	0.50-0.75	0.72	0.22
1.9	0.75-1.00	0.90	0.15

So, $D = 0.64$.

1-tailed test. Reject H_0 if $D > c$.

CHI-SQUARE TEST

Null hypothesis: The parametric model fits its data well / The mean is the same across categories

Suppose there are k categories:

Where:

Test statistic: $Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$

O_i is the observed value for category i.

1-tailed test. Reject H_0 if $Q > c$.

E_i is the observed value for category i.

Suppose there are $k_1 \times k_2$ categories:

Test statistic: $Q = \sum_{j=1}^{k_2} \sum_{i=1}^{k_1} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((k_1-1)(k_2-1))$

1-tailed test. Reject H_0 if $Q > c$.

Note: Subtract additional 1 degree of freedom for each parameter fitted from the data.

LIKELIHOOD RATIO TEST

$H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

$R = \frac{L(X|\theta_0)}{L(X|\theta_{MLE})}$

where θ_{MLE} is the maximum likelihood estimate of θ.

Test statistic: $-2\log R \sim \chi^2(1)$

The DOF depends on the number of constraints in H_0 and H_1.

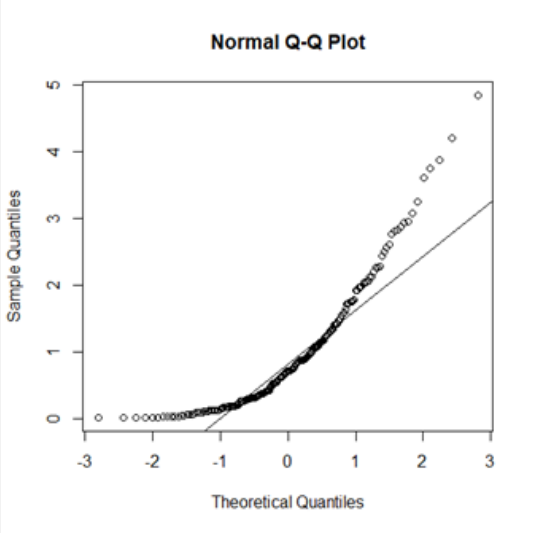
1-tailed test. Reject H_0 if $-2\log R > c$.

Q-Q PLOT

QQ plot: x-coordinates for quantiles of fitted distribution.

y-coordinates for quantiles of observations.

Example:



SCORE TEST

Suppose X follows a parametric distribution with parameter θ.

Loglikelihood function: $l(\theta)$

Score: $U(\theta) = l'(\theta) = 0 \quad \rightarrow \quad$ The MLE is $\hat{\theta}$.

Information: $I = -E[U'(\theta)] = -E[l''(\theta)] \quad \rightarrow \quad Var(U(\theta)) = I \qquad Var\left(\hat{\theta}\right) = I^{-1}$

$H_0: \ \theta = \theta_0 \quad$ vs $\quad H_1: \theta \neq \theta_0$

Test statistic: $\dfrac{U(\theta_0)}{\sqrt{Var(U(\theta_0))}} \dot{\sim} N(0,\ 1)$

CI for θ : $\hat{\theta} \pm z\sqrt{Var\left(\hat{\theta}\right)}$ $\quad Var\left(\hat{\theta}\right)$ can be estimated using $\hat{\theta}$.

Normal Linear Model

NORMAL LINEAR MODEL

Normal linear model: $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$ $\quad Y$ is the response variable

$x's$ are the explanatory variables

$\varepsilon \sim N\left(0,\ \sigma^2\right)$

Properties: $E[Y] = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

$Var(Y) = Var(\varepsilon) = \sigma^2$

Box-Cox transformation: $Y^* = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log Y, & \lambda = 0 \end{cases}$ Thus, $Y^* = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$.

Types of variables: Continuous variables

Count variables

Categorical variables Base category coefficient is 0.

ESTIMATING PARAMETERS

Define the following matrices:

We have n observations of p explanatory variables: $\boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$

which correspond to n observations of the response variables: $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$

We want to estimate p coefficients: $\boldsymbol{beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$

The estimates are: $\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$

Linear regression for "full" model:

Model: $\mathbf{E}[\mathbf{Y}] = \mathbf{x}\boldsymbol{\beta}$

To minimize: $\sum (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{b})^2$

Estimates: $\boldsymbol{b} = (\boldsymbol{x}^T\boldsymbol{x})^{-1}(\boldsymbol{x}^T\boldsymbol{y})$

Fitted values: $\hat{\boldsymbol{y}} = \boldsymbol{x}\boldsymbol{b}$

Properties: $E[\boldsymbol{b}] = \boldsymbol{\beta}$ $VCOV(\boldsymbol{b}) = \boldsymbol{I}^{-1}$ $\boldsymbol{I}^{-1} = \sigma^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$ is the inverse of information matrix.

$E[\hat{\boldsymbol{y}}] = \boldsymbol{x}\boldsymbol{\beta}$ $Var(\hat{\boldsymbol{y}}) = \sigma^2\boldsymbol{H}$ $\boldsymbol{H} = \boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T$ is the hat matrix.

Linear regression for two-variable model:

Model: $E[Y] = \beta_1 + \beta_2 x$

To minimize: $\sum_{i=1}^{n} (y_i - (b_1 + b_2 x_i))^2$

Estimates: $b_1 = \bar{y} - b_2\bar{x}$ $b_2 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$

Fitted values: $\hat{y} = b_1 + b_2 x$

Properties: $E[b_1] = \beta_1$ $Var(b_1) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$

$E[b_2] = \beta_2$ $Var(b_2) = \sigma^2\left(\frac{1}{\sum(x_i-\bar{x})^2}\right)$ $Cov(b_1, b_2) = \sigma^2\left(\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right)$

$E[\hat{y}_i] = \beta_1 + \beta_2 x_i$ $Var(\hat{y}_i) = \sigma^2\left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum_{k=1}^{n}(x_k-\bar{x})^2}\right)$

MEASURES OF FIT

ANOVA table:

Source	Sum of Squares	DF	Mean square
Error	SSE= $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p$	MSE= $\frac{\text{SSE}}{n-p}$
Regression	SSR= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$p-1$	MSR= $\frac{\text{SSR}}{p-1}$
Total	Total SST= $\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	MST= $\frac{\text{SST}}{n-1}$

Error sum of squares:

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

SSE is also called the residual sum of squares, RSS=$\sum\hat{\varepsilon}_i^2$. where $\hat{\varepsilon}_i = y_i - \hat{y}_i$ is the residual.

SSE is also called the scaled deviance, $\sigma^2 D$. where D is the deviance.

Standard error of the regression:

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

s is also called the residual standard error. $\boldsymbol{s^2}$ is an unbiased estimator of $\boldsymbol{\sigma^2 = Var(Y) = Var(\varepsilon)}$.

Coefficient of determination:

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

R^2 is the proportion of variation in y that was explained by variation in x.

For full model: $R^2 = (Corr(\boldsymbol{y}, \hat{\boldsymbol{y}}))^2$

For two-variable model: $R^2 = (Corr(\boldsymbol{y}, \boldsymbol{x}))^2$

In general, the higher the R^2, the better the model fits your data.

However, adding more variables to the model always increases R^2.

Also, R^2 cannot be evaluated statistically.

t test:

$H_0 : \beta_j = \beta$ vs $H_1 : \beta_j \neq \beta$

Test statistic: $t = \frac{b_j - \beta}{\sqrt{\widehat{Var}(b_j)}} \sim t(n-p)$ where $\widehat{VCOV}(\boldsymbol{b}) = s^2(\boldsymbol{x}^T\boldsymbol{x})^{-1}$

CI for β_j: $b_j \pm t_{n-p}\sqrt{\widehat{Var}(b_j)}$ If σ^2 is known, then use $\frac{b_j - \beta}{\sqrt{Var(b_j)}} \sim N(0,\ 1)$.

F test (minimal model):

H_0: Minimal model, with $\beta_2 = \cdots = \beta_p = 0$.

Test statistic: $F = \frac{(\text{SST} - SSE)/(p-1)}{SSE/(n-p)} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \sim F(p-1,\ n-p)$

1-tailed test. Reject H_0 if $F > F_{p-1,n-p,1-\alpha}$.

For the two-variable model, the F statistic is the square of the t statistic for $\beta_2 = 0$.

That is: $F = \left(\frac{b_2 - 0}{\sqrt{\widehat{Var}(b_2)}}\right)^2 = \frac{(b_2)^2}{\widehat{Var}(b_2)} \sim F(1,\ n-2)$

F test (reduced model):

H_0 : Reduced model, with q parameters equal to 0.

Test statistic: $F = \frac{(\text{SSE}^R - \text{SSE}^F)/q}{\text{SSE}^F/(n-p)} = \frac{(R^{F2} - R^{R2})/q}{(1-R^{F2})/(n-p)} \sim F(q,\ n-p)$

1-tailed test. Reject H_0 if $F > F_{q,n-p,1-\alpha}$.

DOF of SSE^R is $n-(p-q)$.

DOF of SSE^F is $n-p$.

Thus, DOF of $\text{SSE}^R - \text{SSE}^F is\,(n-(p-q))-(n-p)=q$.

Collinearity of explanatory variables:

$VIF = \frac{1}{1-R^2_{(j)}}$

where $R^2_{(j)}$ is the coefficient of determination from regression of $x_j = \sum_{k\neq j} \beta_k x_k$.

The larger the VIF, the more collinear the variable j is.

ANOVA & ANCOVA

Here, the **ANOVA** is for normal linear model with categorical variables.

In **ANCOVA**, continuous variables are added to the model.

Note that if the model has an intercept, the base coefficient of a categorial variable is 0.

One-factor ANOVA:

In one-factor ANOVA, one is given observations of J treatments.

Number of observations are $n_1, n_2, \ldots, n_J$. Total number of observations is $n = n_1 + \cdots + n_J$.

The full model is: $Y = \mu_j + \varepsilon$ $j = 1, 2, \ldots, J$ $\mu_1 \neq 0$ is the base parameter.

Source	Sum of Squares	DF	Mean square
Error Within treatment	$\text{SSE}^F = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_j)^2$	$n-J$	$\text{MSE} = \frac{\text{SSE}}{n-J}$
Treatment Between treatment	$\text{SStr} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{y}_j-\bar{y})^2$	$J-1$	$\text{MStr} = \frac{\text{SStr}}{J-1}$
Total	$\text{SST} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2$	$n-1$	$\text{MST} = \frac{\text{SST}}{n-1}$

$H_0 : Y = \mu + \varepsilon$ Reduced model with $J-1$ dummy variables removed.

Test statistic: $F = \frac{(\text{SSE}^R - \text{SSE}^F)/(J-1)}{\text{SSE}^F/(n-J)} = \frac{\text{SStr}/(J-1)}{\text{SSE}^F/(n-J)} \sim F(J-1,\ n-J)$

Two-factor ANOVA without replication:

In two-factor ANOVA, there are J treatments applied to K blocks.

So there is a total of $n = JK$ observations.

The full model is: $Y = \mu + \alpha_j + \beta_k + \varepsilon \quad j = 1, 2, \ldots, J \quad k = 1, 2, \ldots, K$.

Source	Sum of Squares	DF	Mean square
Error	$\text{SSE}^F = \text{SST} - \text{SStr} - \text{SSb}$	$(J-1)(K-1)$	$MSE = \frac{\text{SSE}}{(J-1)(K-1)}$
Treatments α	$\text{SStr} = \sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_j - \bar{y})^2$	$J-1$	$\text{MStr} = \frac{\text{SStr}}{J-1}$
Blocks β	$\text{SSb} = \sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_k - \bar{y})^2$	$K-1$	$\text{MSb} = \frac{\text{SSb}}{K-1}$
Total	$\text{SST} = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk} - \bar{y})^2$	$JK-1$	$\text{MST} = \frac{\text{SST}}{JK-1}$

$H_0 : Y = \mu + \beta_k + \varepsilon$ Reduced model with $J-1$ dummy variables removed.

Test statistic: $F = \frac{(\text{SSE}^R - \text{SSE}^F)/(J-1)}{\text{SSE}^F/((J-1)(K-1))} = \frac{\text{SStr}/(J-1)}{\text{SSE}^F/((J-1)(K-1))} \sim F(J-1,\ (J-1)(K-1))$

$H_0 : Y = \mu + \alpha_j + \varepsilon$ Reduced model with $K-1$ dummy variables removed.

Test statistic: $F = \frac{(\text{SSE}^R - \text{SSE}^F)/(K-1)}{\text{SSE}^F/((J-1)(K-1))} = \frac{\text{SSb}/(K-1)}{\text{SSE}^F/((J-1)(K-1))} \sim F(K-1,\ (J-1)(K-1))$

Two-factor ANOVA with replication:

In two-factor ANOVA, there are J treatments applied to K blocks, and there are L replications.

So there is a total of $n = JKL$ observations.

The full model is: $Y = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon \quad j = 1, 2, \ldots, J \quad k = 1, 2, \ldots, K$.

Source	Sum of Squares	DF	Mean square
Error	SSE^F	$JK(L-1)$	MSE
Treatments α	SStr	$J-1$	MStr
Blocks β	SSb	$K-1$	MSb
Interaction γ	SSint	$(J-1)(K-1)$	MSint
Total	SST	$JKL-1$	MST

$H_0 : Y = \mu + \alpha_j + \beta_k + \varepsilon$ Compared to full model.

Test statistic: $F = \frac{(\text{SSE}^{\alpha+\beta} - \text{SSE}^F)/((J-1)(K-1))}{\text{SSE}^F/(JK(L-1))} = \frac{\text{SSint}/((J-1)(K-1))}{\text{SSE}^F/(JK(L-1))} \sim F((J-1)(K-1), JK(L-1))$

$H_0 : Y = \mu + \beta_k + \varepsilon$ Compared to $\alpha + \beta$ model.

Test statistic: $F = \frac{\left(\text{SSE}^{\beta} - \text{SSE}^{\alpha+\beta}\right)/(J-1)}{\text{SSE}^{F}/(JK(L-1))} = \frac{\text{SStr}/(J-1)}{\text{SSE}^{F}/(JK(L-1))} \sim F\left(J-1, JK\left(L-1\right)\right)$

$H_0 : Y = \mu + \alpha_j + \varepsilon$ Compared to $\alpha + \beta$ model.

Test statistic: $F = \frac{\left(\text{SSE}^{\alpha} - \text{SSE}^{\alpha+\beta}\right)/(K-1)}{\text{SSE}^{F}/(JK(L-1))} = \frac{\text{SSb}/(K-1)}{\text{SSE}^{F}/(JK(L-1))} \sim F\left(K-1, JK\left(L-1\right)\right)$

$H_0 : Y = \mu + \varepsilon$ Compared to α model.

Test statistic: $F = \frac{\left(\text{SSE}^{M} - \text{SSE}^{\alpha}\right)/(J-1)}{\text{SSE}^{F}/(JK(L-1))} = \frac{\text{SStr}/(J-1)}{\text{SSE}^{F}/(JK(L-1))} \sim F\left(J-1, JK\left(L-1\right)\right)$ ← This is the same as the second one.

Model	**SSE = Scaled Deviance**	**Deviance**	**SSE DF**
$Y = \mu + \alpha_j + \beta_k + \gamma_l + \varepsilon$ $\text{SSE}^F + \text{SStr} + \text{SSb} + \text{SSint} = \text{SST}$	SSE^F	D^F	$JK\left(L-1\right)$
$Y = \mu + \alpha_j + \beta_k + \varepsilon$ $\text{SSE}^{\alpha+\beta} + \text{SStr} + \text{SSb} = \text{SST}$	$\text{SSE}^{\alpha+\beta}$	$D^{\alpha+\beta}$	$JKL - J - K + 1$
$Y = \mu + \alpha_j + \varepsilon$ $\text{SSE}^{\alpha} + \text{SStr} = \text{SST}$	SSE^{α}	D^{α}	$JKL - J$
$Y = \mu + \beta_k + \varepsilon$ $\text{SSE}^{\beta} + \text{SSb} = \text{SST}$	SSE^{β}	D^{β}	$JKL - K$
$Y = \mu + \varepsilon$ $\text{SSE}^{M} = \text{SST}$	SSE^{M}	D^{M}	$JKL - 1$

One-factor ANCOVA with a continuous variable added to the model:

The full model is: $Y = \beta_0 + \beta_1 x + \beta_{2j} + \varepsilon,$ where $\varepsilon \sim N\left(0,\ \sigma^2\right).$ $j = 1, 2, \ldots, J.$

Source	**Sum of Squares**	**DF**	**Mean square**
Error	SSE	$n - 1 - J$	MSE
Regression	SSR	$1 + (J-1) = J$	MSR
Total	SST	$n - 1$	MST

$H_0 : Y = \beta_0 + \beta_1 x + \varepsilon$ Reduced model with $J - 1$ dummy variables removed.

Test statistic: $F = \frac{\left(\text{SSE}^{R} - \text{SSE}^{F}\right)/(J-1)}{\text{SSE}^{F}/(n-1-J)} \sim F\left(J-1,\ n-1-J\right)$

VALIDATION

Residual: $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} \rightarrow \hat{\varepsilon}_i = y_i - \hat{y}_i$

Variance: $VCOV(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H}) \rightarrow Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}) \rightarrow \widehat{Var}(\hat{\varepsilon}_i) = s^2(1 - h_{ii})$

$Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(0 - h_{ii}) \rightarrow Cov(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \sigma^2(0 - h_{ii})$

Standardized residual: $r_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_{ii}}}$

We need to validate model assumptions by checking the pattern of residuals.

1. Response is linear: Plot $\hat{y}$ or $\hat{\varepsilon}$ against each x_j.
2. Response is normal: Plot r against $Z \sim N(0,1)$.
3. Response has constant variance (homoscedasticity): Plot r or $\hat{\varepsilon}$ against $\hat{y}$.
4. Observations are independent: Plot $\hat{\varepsilon}$ in the order.

Influential points: Some observations may have a strong influence on the predicted value of y.

Outliers: Outliers have unusually high/low residuals.

Absolute value of r_i: $|r_i| = \left|\frac{\varepsilon_i}{s\sqrt{1-h_{ii}}}\right|$ $|r_i| > 2$ *or* $3 \rightarrow$ outlier

Leverage: h_{ii}

Leverage for 2-variable model: $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$ $h_{ii} > 2\left(\frac{p}{n}\right)$ *or* $3\left(\frac{p}{n}\right) \rightarrow$ Influential point

DFITS: $\text{DFITS}_i = r_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}$ $|r_i| > 2$ *or* $3 \rightarrow$ outlier

Cook's distance: $D_i = r_i^2\left(\frac{h_{ii}}{1-h_{ii}}\right)\left(\frac{1}{p}\right)$ $D_i > 1 \rightarrow$ outlier

PREDICTIONS

Two-variable model: $y_* = \beta_1 + \beta_2 x_* + \varepsilon \rightarrow E[y_*] = \beta_1 + \beta_2 x_*$ $Var(y_*) = \sigma^2$

Predicted value: $\hat{y}_* = b_1 + b_2 x_* \rightarrow E[\hat{y}_*] = \beta_1 + \beta_2 x_*$ $Var(\hat{y}_*) = \sigma^2\left(\frac{1}{n} + \frac{(x_*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

Note that we use $\hat{y}_*$ to predict the mean $E[y_*]$, but the actual value y_* is normally distributed around the mean.

Confidence interval for $E[y_*]$: $\hat{y}_* \pm t_{n-1}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$

Prediction interval for y_*: $\hat{y}_* \pm t_{n-1}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_*-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$ Thus, PI is wider than CI.

SUBSET SELECTION

Leave One Out Cross-Validation (LOOCV) $= CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{MSE}_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\right)^2$

MSE_i is calculated using the test data.

One observation as test data, the rest as training data. There is a total of n fits.

LOOCV minimizes bias but maximizes variance.

k-fold cross-validation $= CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_\text{i}$ MSE_i is calculated using the test data.

The data is randomly partitioned into k subsets, each subset as test data, the rest as training data.

There is a total of k fits.

k-fold CV has higher bias but lower variance. k-fold CV has computational advantage over LOOCV.

Mallow's $C_p = \frac{1}{n}\left(\text{SSE}^S + 2(p-1)\sigma^2\right)$ SSE^S is the SSE of the subset model.

$\sigma^2 = Var(\varepsilon)$ can be estimated using $s^2 =$ MSE of the full model.

p is the number of parameters including the intercept.

Alternative AIC $= \frac{1}{n\sigma^2}\left(\text{SSE}^S + 2(p-1)\sigma^2\right)$

Alternative BIC $= \frac{1}{n\sigma^2}\left(\text{SSE}^S + (\log n)(p-1)\sigma^2\right)$

Adjusted $R^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}$ Adjusted $R^2 = 1 - (1-R^2)\left(\frac{n-1}{n-p}\right)$

Models with the same number of predictors:

- The one with the lower SSE is better.
- The one with the higher R^2 is better.

Models with different number of predictors:

- The one with lower cross-validation is better.
- The one with lower Mallow's C_p is better.
- The one with lower AIC/BIC is better.
- The one with higher adjusted R^2 is better.

Suppose there are k possible predictors:

- **Best subset selection:** For each number of predictors, find the best candidate. Then, select the best among these candidates. There is a total of 2^k fitted models.
- **Forward stepwise selection:** Begin with minimal model. Each time, add the best predictor to the model. Then, select the best among these candidates. There is a total of $1 + \frac{k(k+1)}{2}$ fitted models.
- **Backward stepwise selection:** Begin with full model. Each time, remove the worst predictor from model. Then, select the best among these candidates. There is a total of $1 + \frac{k(k+1)}{2}$ fitted models.

SHRINKAGE AND DIMENSION REDUCTION METHODS

Shrinkage methods:

Ridge regression and the lasso apply a specific penalty function to the sum of square differences.

Minimizing the modified function results in shrinking coefficients of less important variables.

Ridge regression:

$$\text{Minimize } \sum_{i=1}^{n}\left(y_i - \beta_1 - \sum_{j=2}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=2}^{p}\beta_j^2$$

Or

$$\text{Minimize } \sum_{i=1}^{n}\left(y_i - \beta_1 - \sum_{j=2}^{p}\beta_j x_{ij}\right)^2 \qquad \text{subject to } \sum_{j=2}^{p}\beta_j^2 \le s.$$

Ridge regression shrinks the coefficients but does not set them equal to 0. Thus all variables are left in the regression, but the less important ones have small coefficients.

The lasso:

$$\text{Minimize } \sum_{i=1}^{n}\left(y_i - \beta_1 - \sum_{j=2}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=2}^{p}|\beta_j|$$

Or

$$\text{Minimize } \sum_{i=1}^{n}\left(y_i - \beta_1 - \sum_{j=2}^{p}\beta_j x_{ij}\right)^2 \qquad \text{subject to } \sum_{j=2}^{p}|\beta_j| \le s.$$

The lasso shrinks the coefficients and forces them to equal 0. It drops the less important variables from the model. In other words, the lasso performs feature selection.

Both ridge regression and the lasso decrease the MSE of the estimate on the test data.

λ is a tuning parameter selected using cross-validation. λ increases $\rightarrow$ β_j decrease $\rightarrow$ Bias increases, variance decreases.

Dimension reduction methods:

PCR and PLS create new variables that are linear combinations (but not necessarily weighted average) of the original variables. These new variables capture the most important information from the original variables.

Principal components regression: Unsupervised method, higher bias, lower variance.

Partial least squares: Supervised method, lower bias, higher variance.

EXTENSION TO THE LINEAR MODEL

Basis function: $b_j(x) = x^j$ or $b_j(x) = I_{\{c_j \le x < c_{j+1}\}}$ $\rightarrow$ $Y = \beta_0 + \sum_j \beta_j b_j(x) + \varepsilon$

Splines:

Cubic splines match function values, first derivatives, and second derivatives at knots.

Natural splines have second derivative equal to 0 at endpoints.

Regression splines have smaller number of knots and are fitted with least squares.

Smoothing splines $g(x)$ have knots for each point of training data and are fitted by minimizing:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

where λ is again a tuning parameter selected using cross-validation.

When $\lambda = 0 \rightarrow$ smoothing spline is natural spline, with n effective DOF.

When $\lambda = \infty \rightarrow$ it is least squares regression line, with 2 effective DOF.

Local regression: The value of the predictor variable $\hat{y}$ is calculated using a different linear regression at each point x:

1. Select $0 \le s \le 1$, then span. This determines the number of points to use for each regression.
2. Assign weights to these pints. The point furthest away gest a weight of 0.
3. Perform a weighted regression of y on x. The regression may be constant, linear, or quadratic.

Generalized Additive Models: We can generalize to models with any number of predictors:

$$Y = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_{p-1}(x_{p-1}) + \varepsilon$$

The model is additive in that there is no interaction between the predictors.

Generalized Linear Models

GENERALIZED LINEAR MODELS

Generalized linear model: $g(\mu) = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ We model $\mu = E[Y]$ rather than Y. g is the link function.

Exponential family: Y may have a pdf in the form of $f(y;\theta) = e^{p(\theta)q(y)+r(\theta)+s(y)}$.
The GLM requires $q(y) = y$, which is called the canonical form.

Canonical link function: The link function that makes the GLM estimates unbiased.

Tweedie distributions: A family of distributions for which $Var(Y) = aE[Y]^b$,
where a does not contain the GLM parameter.

Members of exponential family in canonical form:

Distribution	Canonical Link Function		Tweedie Distribution
Exponential/Gamma	$g(\mu) = -\frac{1}{\mu}$	Negative inverse link	$b = 2$
Normal	$g(\mu) = \mu$	Identity link	$b = 0$
Inverse Gaussian	$g(\mu) = \frac{1}{\mu^2}$	Inverse squared link	$b = 3$
Poisson	$g(\mu) = \log \mu$	Log link	$b = 1$
Binomial	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	Logit link	
Negative Binomial			

POISSON RESPONSE

Let $Y_i \sim Poisson(\lambda)$ $\rightarrow$ Then $S = Y_1 + \cdots + Y_n \sim Poisson(n\lambda)$

With log link: $\log E[S] = \log n + \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p \rightarrow \sum_{i=1}^{p} \beta_i x_i$ produces an estimate of $\log \lambda$.

$\rightarrow$ $\log n$ is called an offset.

CATEGORICAL RESPONSE

Let $\eta = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Binomial Response: There are two categories, Yes or No.

Probabilities are q and $1 - q$. Treat $E[Y] = q$.

Logit link	$\log \frac{q}{1-q} = \eta$	$q = \frac{e^{\eta}}{1+e^{\eta}}$
Probit link	$N^{-1}(q) = \eta$	$q = N(\eta)$
Complementary log-log link	$\log(-\log(1-q)) = \eta$	$q = 1 - e^{-e^{\eta}}$

Note that for Logit link, the **odds of YES** is Odds=$\frac{q}{1-q} = e^{\eta}$.

Or we can write $q = \frac{\text{Odds}}{1+\text{Odds}}$.

Nominal Response: There are multiple categories with no particular order.

Probability is q_j for category j.

Category 1 is the base category.

Logit link	$\log \frac{q_j}{q_1} = \eta_j \quad j \neq 1$	$q_1 = \frac{1}{1+e^{\eta_2}+e^{\eta_3}+\cdots}$ $q_j = \frac{e^{\eta_j}}{1+e^{\eta_2}+e^{\eta_3}+\cdots} \quad j \neq 1$

For a binary explanatory variable x_2, assume all other variables are held constant.

The **odds ratio** of category j to the based category is:

$$\text{OR}_j = \frac{\text{Odds}_{j,x_2=1}}{\text{Odds}_{1,x_2=1}} = \frac{\left(\frac{q_{j,x_2=1}}{q_{j,x_2=0}}\right)}{\left(\frac{q_{1,x_2=1}}{q_{1,x_2=0}}\right)} = \frac{\left(\frac{q_{j,x_2=1}}{q_{1,x_2=1}}\right)}{\left(\frac{q_{j,x_2=0}}{q_{1,x_2=0}}\right)} = \frac{e^{\beta_{1j}+\beta_{2j}+\sum_{i=3}^{p}\beta_{ij}x_i}}{e^{\beta_{1j}+0+\sum_{i=3}^{p}\beta_{ij}x_i}} = e^{\beta_{2j}}$$

It does not vary with the values of other explanatory variables.

Ordinary Response: There are multiple categories that follow a logical order.

Probability is q_j for category j.

There is no base category.

Cumulative logit link	$\log \frac{q_1+\cdots+q_j}{q_{j+1}+\cdots+q_J} = \beta_{1j} + \sum_{i=2}^{p}\beta_{ij}x_i$ $\log \frac{q_1+\cdots+q_j}{q_{j+1}+\cdots+q_J} = \beta_{1j} + \sum_{i=2}^{p}\beta_i x_i$ (Proportional odds model)
Adjacent categories logit link	$\log \frac{q_j}{q_{j+1}} = \beta_{1j} + \sum_{i=2}^{p}\beta_{ij}x_i$ $\log \frac{q_j}{q_{j+1}} = \beta_{1j} + \sum_{i=2}^{p}\beta_i x_i$
Continuation ratio logit link	$\log \frac{q_j}{q_{j+1}+\cdots+q_J} = \beta_{1j} + \sum_{i=2}^{p}\beta_{ij}x_i$ $\log \frac{q_j}{q_{j+1}+\cdots+q_J} = \beta_{1j} + \sum_{i=2}^{p}\beta_i x_i$

For proportional odds model, suppose there are two sets of values of x_i.

The (cumulative) **odds ratio** for category j is:

$$\text{OR}_j = \frac{C.\text{Odds}_{j,x_i=k_i}}{C.\text{Odds}_{j,x_i=g_i}} = \frac{\left(\frac{q_1+\cdots+q_j}{q_{j+1}+\cdots+q_J}\right)_{\{x_i=k_i\}}}{\left(\frac{q_1+\cdots+q_j}{q_{j+1}+\cdots+q_J}\right)_{\{x_i=g_i\}}} = \frac{e^{\beta_{1j}+\sum_{i=2}^{p}\beta_i k_i}}{e^{\beta_{1j}+\sum_{i=2}^{p}\beta_i g_i}} = e^{\sum_{i=2}^{p}\beta_i(k_i-g_i)}$$

This is independent of category j.

ESTIMATING PARAMETERS

Method of scoring:

With link function: $g(\mu) = \sum_{j=1}^{p}\beta_j x_j$ $\mu = E[Y]$ $v = Var(Y)$ Refer to **Tweedie distributions**.

The initial estimation: $\boldsymbol{b}^{(0)}$

So we have: $g(\boldsymbol{\mu}) = \boldsymbol{x}\boldsymbol{b}^{(0)} \rightarrow \boldsymbol{\mu} = g^{-1}\left(\boldsymbol{x}\boldsymbol{b}^{(0)}\right)$

We calculate: $\boldsymbol{z} = g(\boldsymbol{\mu}) + \boldsymbol{G}(\boldsymbol{y}-\boldsymbol{\mu})$ $\boldsymbol{G}$ is diagonal with $G_{ii} = \left(\left.\frac{\partial g}{\partial \mu}\right|_{\mu_i}\right)$

$\boldsymbol{w}$ $\boldsymbol{w}$ is diagonal with $w_{ii} = \left(\left(\left.\frac{\partial g}{\partial \mu}\right|_{\mu_i}\right)^2 \times v_i\right)^{-1}$

The next estimate: $\boldsymbol{b}^{(1)} = \left(\boldsymbol{x}^T\boldsymbol{w}\boldsymbol{x}\right)^{-1}\left(\boldsymbol{x}^T\boldsymbol{w}\boldsymbol{z}\right)$

We can write: $\boldsymbol{b}^{(1)} = \boldsymbol{b}^{(0)} + \left(\boldsymbol{I}^{(0)}\right)^{-1}\boldsymbol{U}^{(0)}$ $\boldsymbol{I} = \boldsymbol{x}^T\boldsymbol{w}\boldsymbol{x}$ is the information matrix.

$\boldsymbol{U}$ with $U_j = \sum_{i=1}^{n}\frac{y_i-\mu_i}{\left(\left.\frac{\partial g}{\partial \mu}\right|_{\mu_i}\right)v_i}x_{ij}$ is the score vector.

MEASURES OF FIT

Suppose Y follows a parametric distribution with parameter θ.

The parameter θ can be estimated using MLE, GLM, etc.

Using MLE, the loglikelihood function is: $l = \sum_{i=1}^{n}\log f\left(y_i;\hat{\theta}\right)$ There is only one estimate, $\hat{\theta}$.

Using GLM, the loglikelihood function is: $l = \sum_{i=1}^{n}\log f\left(y_i;\hat{\theta}_i\right)$ There are n estimates, each $\hat{\theta}_i$ depends on the resulting $\hat{y}_i$.

Response	Loglikelihood function	Saturated model with n parameters	Your model with p parameters	Minimal model with an intercept only
Poisson	$l = \sum_{i=1}^{n}\left(-\hat{\lambda}_i + y_i\log\hat{\lambda}_i - \log y_i!\right)$	Set $\hat{\lambda}_i = y_i$	Set $\hat{\lambda}_i = \hat{y}_i$.	Set $\hat{\lambda}_i = \bar{y}$.
Binomial	$l = \sum_{i=1}^{n}\left(\log\binom{m_i}{y_i} + y_i\log\hat{q}_i + (m_i-y_i)\log(1-\hat{q}_i)\right)$	Set $\hat{q}_i = \frac{y_i}{m_i}$	Set $\hat{q}_i = \frac{\hat{y}_i}{m_i}$	Set $\hat{q}_i = \frac{\sum y_i}{\sum m_i}$

Deviance test:

H_0: Model under consideration with l, with p parameters.

H_1: Saturated model with l^s, with n parameters.

Test statistic: $D = -2(l - l^s) \sim \chi^2(n-p)$

1-tailed test Reject H_0 if $D > c$

Likelihood ratio test (minimal model):

H_0: Minimal model with l^m, with only an intercept.

H_1: Model under consideration with l, with p parameters.

Test statistic: $C = -2(l^m - l) \sim \chi^2(p-1)$ $\rightarrow$ Pseudo $R^2 = \frac{l^m - l}{l^m} = 1 - \frac{l}{l^m}$

1-tailed test Reject H_0 if $C > c$

Likelihood ratio test (constrained model):

H_0: Constrained model with l^c, with q parameters removed from model under consideration.

H_1: Model under consideration with l, with p parameters.

Test statistic: $C = -2(l^c - l) \sim \chi^2(q)$

1-tailed test Reject H_0 if $C > c$

Pearson chi-square test:

H_0: No significant difference between the observed and the expected values.

H_1: Significant difference between the observed and the expected values.

Binomial: $Q = \sum \frac{(O_i - E_i)^2}{V_i} = \sum \frac{(y_i - \hat{y}_i)^2}{m_i \hat{q}_i (1 - \hat{q}_i)} \sim \chi^2(n-p)$ where $\hat{q}_i = \frac{\hat{y}_i}{m_i}$

Poisson: $Q = \sum \frac{(O_i - E_i)^2}{V_i} = \sum \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \sim \chi^2(n-p)$

1-tailed test Reject H_0 if $Q > c$

Wald test:

$H_0 : \beta_j = \beta$ vs $H_1 : \beta_j \neq \beta$

Test statistic: $\frac{(b_j - \beta)^2}{Var(b_j)} \sim \chi^2(1)$ where $VCOV(\boldsymbol{b}) = \boldsymbol{I}^{-1} = (\boldsymbol{x}^T \boldsymbol{w} \boldsymbol{x})^{-1}$

1-tailed test Reject H_0 if statistic $> c$

CI for β_j : $b_j \pm z\sqrt{Var\,(b_j)}$ since $\dfrac{b_j-\beta}{\sqrt{Var\,(b_j)}} \dot{\sim} N\,(0,\ 1)$

Type I/Type III tests:

Type I tests are sequential, adding one variable (or a group of variables) at a time to the model in a prescribed order.

Type III tests check one variable (or a group of variables) assuming all other variables are in the model.

Penalized loglikelihood tests:

AIC$=-2l+2p$

BIC$=-2l+p\log n$

Lower AIC or BIC $\rightarrow$ Better model

VALIDATION

Pearson/chi-square residuals:

Binomial response: $X_i=\dfrac{y_i-\hat{y}_i}{\sqrt{V_i}}$ $V_i=m_i\hat{q}_i\,(1-\hat{q}_i)$ $\hat{q}_i=\dfrac{\hat{y}_i}{m_i}$

Poisson response: $X_i=\dfrac{y_i-\hat{y}_i}{\sqrt{V_i}}$ $V_i=\hat{\lambda}_i$ $\hat{\lambda}_i=\hat{y}_i$

Standardized Pearson residuals: $r_i=\dfrac{X_i}{\sqrt{1-h_{ii}}}$

Deviance residuals:

Binomial response: $\pm\sqrt{d_i}$ where $D=-2\,(l-l^s)=\sum_{i=1}^{n} d_i$ is the deviance test statistic.

Poisson response: $\pm\sqrt{d_i}$ where $D=-2\,(l-l^s)=\sum_{i=1}^{n} d_i$ is the deviance test statistic.

Standardized deviance residuals: $r_i=\dfrac{\pm\sqrt{d_i}}{\sqrt{1-h_{ii}}}$ Use $+\sqrt{d_i}$ if $y_i-\hat{y}_i>0$, and $-\sqrt{d_i}$ if $y_i-\hat{y}_{ii}<0$.

Validation:

- Standardized residuals can be plotted to check normality, serial correlation, and linearity.
- High deviance may indicate overdispersion.