DATA AND BASICS OF MODELLING

Descriptive Analysis Summarize and describe the key characteristics of a dataset

Measures of central tendency The **mean**, **median** and **mode**

Measures of dispersion The **standard deviation**, **range** and interquartile range

Inferential Analysis Using a smaller sample size to draw conclusions about a larger population

Predictive Analysis Making predictions or forecasts about future events based on past or historical data

The Data Analysis Process

1. Develop
2. Identify
3. Collection
4. Processing
5. Data cleaning
6. Exploratory data analysis
7. Modelling
8. Communicating
9. Monitoring

Primary Source The data collected either from the source or through the original data collection process

Secondary Source Information that has already been collected, analyzed, and published by others

Cross-Sectional Data Recording values of the variable(s) of interest for each case in the sample at a single moment in time. It can be thought of as a snapshot of the data at a single moment in time

Longitudinal Data Recording values of the SAME subjects at intervals through time

Censored Data The value of a variable is only partially known

Truncated Data Measurements on some variables are not recorded so are completely unknown

Big Data Size, speed, variety, reliability of data

Reproducibility The ability to reproduce statistical analyses or models using the same data and methodology as the original study

Pros and Cons of Reproducibility

Pros	It is necessary for a complete technical work review
	Required by external regulators and auditors
	More easily extended to investigate the effect of changes to the analysis, or to incorporate new data
	Desirable to compare the results of an investigation with a similar one carried out in the past
	Lead to fewer errors that need correcting in the original work, greater efficiency
Cons	Reproducibility does not mean that the analysis is correct
	If activities involved in reproducibility only occurred at the end of an analysis, it may be too late to address any unforeseen problems

DISCRETE DISTRIBUTIONS

Discrete Uniform Distribution on sample space S, where $S = \{1, 2, \ldots, k\}$

$$P(X = x) = \frac{1}{k} \text{ for } x = 1, 2, 3, \ldots, k \qquad \mu = E[X] = \frac{k+1}{2} \qquad \sigma^2 = \textit{Var}[X] = \frac{k^2-1}{12}$$

Bernoulli Distribution on sample space S, where $S = \{s, f\}$

$$P(\{s\}) = p, \qquad P(\{f\}) = 1 - p \qquad \mu = p \qquad \sigma^2 = p - p^2 = p(1-p)$$

Binomial Distribution for n trials and success probability p

$$P(X = x) = \binom{n}{k} p^x (1-p)^{n-x} \quad \begin{matrix} x = 0, 1, 2, \ldots, n; \\ 0 < p < 1. \end{matrix} \qquad \mu = np \qquad \sigma^2 = np(1-p)$$

Geometric Distribution on the integers 0, 1, 2, ... and parameter p

$$P(X = x) = p(1-p)^{x-1} \quad \begin{matrix} x = 0, 1, 2, \ldots; \\ 0 < p < 1. \end{matrix} \qquad \mu = \frac{1}{p} \qquad \sigma^2 = \frac{(1-p)}{p^2}$$

Negative Binomial Distribution with X as the number of trials on which the k-th success occurs, or Y as the number of failures before the k-th success:

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \quad \begin{matrix} x = k, k+1, \ldots; \\ 0 < p < 1. \end{matrix} \qquad \mu = \frac{k}{p} \qquad \sigma^2 = \frac{k(1-p)}{p^2}$$

$$P(X = x) = \frac{x-1}{x-k}(1-\theta) P(X = x-1)$$

$$P(Y = y) = \binom{k+y-1}{y} p^k (1-p)^y, \; y = 0, 1, 2, 3, \ldots \qquad \mu = \frac{k(1-p)}{p}$$

Hypergeometric Distribution

$$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \quad \begin{matrix} x = 1, 2, 3, \ldots; \\ 0 < p < 1. \end{matrix} \qquad \mu = \frac{nk}{N} \qquad \sigma^2 = \frac{nk(N-k)(N-n)}{N^2(N-1)}$$

Poisson Distribution with parameter λ:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \begin{matrix} x = 0, 1, 2, \ldots; \\ \lambda > 0 \end{matrix} \qquad \mu = \lambda \qquad \sigma^2 = \lambda$$

$$P(X = x) = \frac{\lambda}{x} P(X = x-1)$$

CONTINUOUS DISTRIBUTIONS

Continuous Uniform Distribution on the interval $[\alpha, \beta]$

$$f_X(x) = \frac{1}{\beta - \alpha}, \; \alpha < x < \beta \qquad \mu = \frac{\alpha + \beta}{2} \qquad \sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

Gamma Distribution with parameters α and λ

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \text{ for } x > 0 \qquad \mu = \frac{\alpha}{\lambda} \qquad \sigma^2 = \frac{\alpha}{\lambda^2}$$

Gamma function: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ $\qquad \Gamma(1) = 1$ $\qquad \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for $\alpha > 1$

Exponential Distribution (Gamma with $\alpha = 1$)

$$f_X(x) = \lambda e^{-\lambda x}, \; x > 0 \qquad \mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$

Chi-square (χ^2) distribution with 'degrees of freedom' as its parameter

Gamma with $\alpha = v/2$ where v is a positive integer, and $\lambda = 1/2$ $\qquad \mu = v$ $\qquad \sigma^2 = 2v$

Beta Distribution $\{x : 0 < x < 1\}$ with parameters $\alpha > 0$ and $\beta > 0$

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } 0 < x < 1$$

$$\mu = \frac{\alpha}{\alpha+\beta} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Beta function: $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

Normal Distribution with parameters $N(\mu,\sigma)$ where mean=μ, and variance=σ^2

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty \quad F_X(x) = P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \text{ where } Z = \frac{X-\mu}{\sigma}$$

Standard Normal Distribution $N(0,1)$ where mean=0, and variance=1

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \text{ for } -\infty < x < \infty \qquad F_X(x) = \Phi(x)$$

Lognormal Distribution with parameters μ and σ. If $\ln(X) \sim N(\mu,\sigma^2)$ then we have

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2} \text{ for } 0 < x < \infty \qquad E[X] = e^{\mu+\frac{\sigma^2}{2}} \qquad Var[X] = E[Y]^2\left(e^{\sigma^2}-1\right)$$

***T*-distribution with 'degrees of freedom' parameter, v**

If $X \sim \chi^2_v$ and $Z \sim N(0,1)$, and X and Z are independent,
then $\dfrac{Z}{\sqrt{\frac{X}{v}}} \sim t$-distribution with degree of freedom v.

***F* distribution with 'degrees of freedom' parameters, n_1 and n_2**

If $X \sim \chi^2$ distribution with degree of freedom n_1 and $Y \sim \chi^2$ with degree of freedom n_2, and X and Y are independent, then $\dfrac{X/n_1}{Y/n_2} \sim F$ distribution with degrees of freedom n_1 and n_2.

POISSON PROCESS

Poisson Distribution $X \sim Poisson(\lambda) \rightarrow P(X=x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$, $x = 0,1,2,\ldots;$ $\lambda > 0$, $\mu = \lambda$, $\sigma^2 = \lambda$

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 (PP) $X(t) \sim Poisson(\lambda(t))$

$X(t) - X(s)$ 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.

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 $F_Y(y) = \Pr(Y \leq y)$.
3. Calculate $y_i = F_Y^{-1}(u_i)$

GENERATING FUNCTIONS

Moment Generating Functions (MGF)

MGF Moments

$$M_X(t) = E[e^{tX}] \qquad M_X(t) = 1 + tE[X] + \frac{t^2}{2!}E\left[X^2\right] + \frac{t^3}{3!}E\left[X^3\right] + \cdots$$

$$M'_X(0) = E[X] \qquad M''_X(0) = E\left[X^2\right]$$

Uniform

$$M_X(t) = E\left(e^{tX}\right) = \left(\frac{e^t}{k}\right)\left(\frac{1-e^{kt}}{1-e^t}\right)$$

Uniform (a, b)

$$M_X(t) = \int_a^b e^{tx}\frac{1}{b-a}dx = \frac{e^{bt}-e^{at}}{t(b-a)}$$

Binomial (n, p)
(Including Bernoulli, for which $n = 1$)

$$M_X(t) = \sum_{x=0}^{n}\binom{n}{x}\left(pe^t\right)^x q^{n-x} = \left(q+pe^t\right)^n$$

Negative Binomial (k, p)
(Including Geometric, for which $k = 1$)

$$M_X(t) = \sum_{x=k}^{\infty}\binom{x-1}{k-1}e^{tx}p^k q^{x-k} = \left[\frac{pe^t}{1-qe^t}\right]^k$$

Poisson (λ)

$$M_X(t) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{\lambda\left(e^t-1\right)}$$

Gamma (α, λ)

$$M_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty\left(\frac{1}{\lambda-t}\right)^\alpha y^{\alpha-1}e^{-y}dy = \left(\frac{\lambda}{\lambda-t}\right)^\alpha$$

Normal (μ, σ^2)

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

Cumulant Generating Functions(CGF) $C_X(t) = \ln M_X(t)$

CGF Moments

$$C'_X(t) = \frac{M'_X(t)}{M_X(t)}$$

$$C''_X(t) = \frac{M''_X(t)M_X(t) - \left(M'_X(t)\right)^2}{\left(M_X(t)\right)^2}$$

$$C'''_X(t) = \frac{M'''_X(t)\left(M_X(t)\right)^3 - 3\left(M_X(t)\right)^2 M'_X(t)M''_X(t) + 2M_X(t)\left(M'_X(t)\right)^3}{\left(M_X(t)\right)^4}$$

$$M_X(0) = 1$$

$$C'_X(0) = \frac{M'_X(0)}{M_X(0)} = E[X]$$

$$C''_X(0) = \frac{M''_X(0)M_X(0) - \left(M'_X(0)\right)^2}{M_X^2(0)} = \frac{E\left[X^2\right](1) - (E[X])^2}{1^2} = Var[X]$$

$$C'''_X(0) = \frac{M'''_X(0)\left(M_X(0)\right)^3 - 3\left(M_X(0)\right)^2 M'_X(0)M''_X(0) + 2M_X(0)\left(M'_X(0)\right)^3}{\left(M_X(0)\right)^4}$$

$$= skew(X)$$

Cumulants

The coefficient of $\frac{t^r}{r!}$ in the Maclaurin's series of $C_X(t) = \ln M_X(t)$ is called the rth cumulant and is denoted by κ_r

Linear function $Y = a + bX$ $M_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(a+bX)}\right] = e^{at}E\left[e^{btX}\right] = e^{at}M_X(bt)$

JOINT DISTRIBUTIONS

Joint Probability (Density) Functions

Discrete $p(x, y) = P(X = x, Y = y)$ $\sum_x \sum_y p(x, y) = 1$ $p(x, y) \geq 0$

Continuous $P(x_1 < X < x_2, y_1 < Y < y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f(x, y)dxdy$ $F(x, y) = P(X \leq x, Y \leq y)$

$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$ $\int_x \int_y f(x, y) = 1$ $f(x, y) \geq 0$

Marginal Distribution of X from joint distribution of X and Y:

$p_X(x) = \sum_y p(x, y)$ (discrete) $f_X(x) = \int_y f(x, y)\, dy$ (continuous)

Conditional Probability (Density) Functions

Discrete $p_{X|Y=y}(x|y) = P(X = x|Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$

Continuous $\int_{x=x_1}^{x_2} f_{X|Y=y}(x|y)dx = P(x_1 < X < x_2|Y = y)$

Independence of Random Variables

$f_Y(y) = f(y|x) = f(x, y)/f_X(x)$ i.e. $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

Discrete $P(X = x, Y = y) = P(X = x)P(Y = y)$

Continuous $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

Expectations

Discrete $E[g(X, Y)] = \sum_x \sum_y g(x, y)p_{X,Y}(x, y) = \sum_x \sum_y g(x, y)P(X = x, Y = y)$

Continuous $E[g(X, Y)] = \int_x \int_y g(x, y)f_{X,Y}(x, y)dxdy$

Expectations of Sums and Products

$E[ag(X) + bh(Y)] = aE[g(X)] + bE[h(Y)]$ where a and b are constants

$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for independent random variables X and Y

Covariance and Correlation Coefficient

$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

$\text{Cov}[aX + b, cY + d] = ac\,\text{Cov}[X, Y]$ $\text{Cov}[X, Y + Z] = \text{Cov}[X, Y] + \text{Cov}[X, Z]$

If X and Y are independent, $\text{Cov}[X, Y] = 0$

Variance of a Sum $V[X + Y] = V[X] + V[Y] + 2\,\text{Cov}[X, Y]$

$V[X + Y] = V[X] + V[Y]$ for independent random variables

Convolutions $p_Z(z) = \sum_x p(x, z - x)$ (discrete)

$p_Z(z) = \sum_x p_X(x)p_Y(z - x)$ for independent random variables X and Y

$f_Z(z) = \int_x f_X(x)f_Y(z - x)dx$ (continuous)

Moments of Linear Combinations of Random Variables

Mean $E(c_1X_1 + c_2X_2 + \ldots + c_nX_n) = c_1E(X_1) + c_2E(X_2) + \ldots + c_nE(X_n)$

$E\left(\sum_{i=1}^n c_iX_i\right) = \sum_{i=1}^n c_iE(X_i)$

Variance $V(Y) = \text{Cov}(Y,Y) = \sum_i c_i^2 \text{Cov}(X_i, X_i) + 2\sum_{i<j}\sum_j c_i c_j \text{Cov}(X_i, X_j)$

Independent r.v. $V(c_1X_1 + c_2X_2 + \ldots + c_nX_n) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_n^2V(X_n)$

$$V\left(\sum_{i=1}^n c_iX_i\right) = \sum_{i=1}^n c_i^2V(X_i)$$

Moment generating functions (MGFs):

X_1 and X_2 are independent random variables with MGFs $M_{X_1}(t)$ and $M_{X_2}(t)$ and $S = c_1X_1 + c_2X_2$

$$M_S(t) = E\left[e^{(c_1X_1+c_2X_2)t}\right] = E\left[e^{c_1X_1t}\right]E\left[e^{c_2X_2t}\right] = M_{X_1}(c_1t)\,M_{X_2}(c_2t)$$

$Y = X_1 + X_2 + \ldots + X_n$ where the X_i are independent and X_i has MGF $M_i(t)$

$$M_Y(t) = M_1(t)M_2(t)\ldots..M_n(t)$$

$Y = X_1 + X_2 + \ldots + X_n$ where the X_i's are identically distributed, each with MGF $M(t)$

$$M_Y(t) = [M(t)]^n$$

Bernoulli/Binomial $[q + pe^t]^n$ where $Y = X_1 + X_2 + \ldots + X_n$

with $X_i,\ i = 1, 2, \ldots, n$, be independent Bernoulli (p) variables and each has MGF $M(t) = q + pe^t$

Geometric/ Negative binomial $\left[\dfrac{pe^t}{1-qe^t}\right]^k$ where $Y = X_1 + X_2 + \ldots + X_k$

with $X_i,\ i = 1, 2, \ldots.., k$, be independent geometric (p) variables with MGF $M(t) = \dfrac{pe^t}{1-qe^t}$

Poisson $\exp\{(\lambda+\gamma)(e^t-1)\}$ with X and Z be independent Poisson (λ) and Poisson (γ) variables

X has MGF $M_X(t) = \exp\{\lambda(e^t-1)\}$, Z has MGF $M_Z(t) = \exp\{\gamma(e^t-1)\}$

Exponential/ Gamma $\left[\lambda(\lambda-t)^{-1}\right]^k$ where $Y = X_1 + X_2 + \ldots\ldots + X_k$ with $X_i,\ i = 1, 2, \ldots, k$, be independent exponential (λ) variables and each has MGF $M(t) = \lambda(\lambda-t)^{-1}$

Normal $\exp\left\{(\mu_X+\mu_Y)t + \frac{1}{2}\left(\sigma_x^2+\sigma_Y^2\right)t^2\right\}$ where $Z = X + Y$

with X has MGF $M_X(t) = \exp\left(\mu_Xt + \frac{1}{2}\sigma_X^2t^2\right)$, Y has MGF $M_Y(t) = \exp\left(\mu_Yt + \frac{1}{2}\sigma_Y^2t^2\right)$

Chi-square The sum of a chi-square (n) and an independent chi-square (m) is a chi-square $(n+m)$ variable

the sum of independent chi-square variables is a chi-square variable

CONDITIONAL EXPECTATION

Conditional expectation of Y given $X = x$ ($E[Y|X=x]$)

$$E[Y|X=x] = \begin{cases} \sum y \cdot f_{Y|X}(y|X=x) & \text{in the discrete case} \\ \int y \cdot f_{Y|X}(y|X=x)dy = \int y \cdot \frac{f(x,y)}{f_Y(y)}dy & \text{in the continuous case} \end{cases}$$

Conditional Variance of Y given $X = x$ ($Var[Y|X]$)

$$Var[Y|X] = E\left[Y^2|X\right] - (E[Y|X])^2$$

Double Expectation and Variance

$$E[E[Y|X]] = E[Y] \qquad Var[Y] = E[Var[Y|X]] + Var[E[Y|X]]$$

THE CENTRAL LIMIT THEOREM

Central Limit Theorem Suppose $X_1, X_2, \cdots, X_n$ are n independent random variables with with mean μ and variance σ^2

then the distribution of $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ approaches the standard normal distribution, $N(0,1)$, as $n \to \infty$

$$\bar{X} \sim N\left(\mu, \sigma^2/n\right) \qquad \sum X_i \sim N\left(n\mu, n\sigma^2\right)$$

Binomial Distribution	Binomial $(n,p) \sim N(np, np(1-p))$for large n	
Poisson Distribution	Poisson $(n\lambda) \sim N(n\lambda, n\lambda)$for large n	Poisson $(\lambda) \sim N(\lambda, \lambda)$
Gamma Distribution	Gamma $(n,\lambda) \sim N(n/\lambda, n/\lambda^2)$	$\chi^2_k \sim N(k, 2k)$
Continuity Correction	If n and m are integers, the probability $P[n \le X \le m]$ is approximated by using a normal random variable Y with the same mean and variance as X, and then finding the probability $P\left[n - \frac{1}{2} \le Y \le m + \frac{1}{2}\right]$	

RANDOM SAMPLING AND SAMPLING DISTRIBUTIONS

A **random sample**, $\underline{X} = (X_1, X_2, \ldots, X_n)$ is a collection of independent and identically distributed random variables,

with observed sample $\underline{X} = (X_1, X_2, \ldots, X_n)$ is $\underline{x} = (x_1, x_2, \ldots, x_n)$

Probability (Density) Function, $f(x;\theta)$, where θ denotes the parameter(s) of the distribution

Sample Mean $\bar{X} = \dfrac{\sum X_i}{n}$

Sample Variance $S^2 = \dfrac{\sum \left(X_i - \bar{X}\right)^2}{n-1}$

Sampling Distributions for the Normal

Sample Mean $\dfrac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ or $\bar{X} \sim \left(\mu, \sigma^2/n\right)$ Sample Variance $\dfrac{S^2}{\sigma^2}(n-1) \sim \chi^2_{n-1}$

Student's T-Distribution $t_k = \dfrac{N(0,1)}{\sqrt{\frac{\chi^2_k}{k}}} \rightarrow \dfrac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$

F-Distribution $F = \dfrac{U/v_1}{V/v_2}$, where U and V are independent χ^2 random variables with v_1 and v_2 degrees of freedom

$\dfrac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim F_{n_2-1, n_1-1}$ $F \sim F_{n_1-1, n_2-1} \Leftrightarrow \frac{1}{F} \sim F_{n_2-1, n_1-1}$

ESTIMATION AND ESTIMATORS

The Method of Moments

To estimate one parameter, use the first moment.
To estimate two parameters, use the first and second moments.

Complete Data $E[X] = \dfrac{\sum x_i}{n}$ $E\left[X^2\right] = \dfrac{\sum x_i^2}{n}$

Maximize the **likelihood function** $\boldsymbol{L}$.

The value(s) of parameter(s) that maximizes $\boldsymbol{L}$ is called the maximum likelihood estimate(s)

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. Set $l'(\theta) = 0$, obtain $\hat{\theta}$, which is the MLE

Complete Samples $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ for $x_1, x_2, \ldots, x_n$ from a population with density or probability function $f(x;\theta)$

Properties of Maximum Likelihood Estimation (MLE)

1. Invariance Property if $\hat{\theta}$ is the MLE of θ, then the MLE of a function $g(\theta)$ is $g(\hat{\theta})$
2. The consistency nature: Estimators approach true value with increase in sample size.
3. Asymptotic normality: As sample size increases, it converges to the normal distribution.
4. Efficiency: MlE achieves Cramer Rao as the sample size tends to infinity.

Incomplete Samples

Truncated Samples $L(\theta) = \left(\prod_{i=1}^{n} f(x_i, \theta)\right) \times (P(X > y))^m$ with n observations $(x_1, \ldots . x_n)$ and m observations $< y$

Censored Samples $L(\theta) = \left(\prod_{i=1}^{n} f(x_i, \theta)\right) / (P(X > z))^m$ with n observations $(x_1, \ldots . x_n)$ and no information about samples under z

Independent Samples For independent samples from two populations which share a common parameter the overall likelihood is the product of the two separate likelihoods

Bias The bias of an estimator is Bias $= \mathrm{E}[g(\underline{X})] - \theta$

An estimator is **unbiased** if Bias $= 0$

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

Cramer-Rao Lower Bound $CRLB = \dfrac{1}{-E\left[\frac{\partial^2}{\partial \theta^2} \log L(\theta, \underline{X})\right]}$

This is the lowest possible variance of an unbiased estimator $\hat{\theta}$

Alternative expressions $CRLB = \dfrac{1}{E\left\{\left[\frac{\partial}{\partial \theta} \log L(\theta, \underline{X})\right]^2\right\}} = \dfrac{1}{nE\left\{\left[\frac{\partial}{\partial \theta} \log f(X; \theta)\right]^2\right\}}$

Non-parametric (full) Bootstrap empirical distribution, $\hat{F}_n(y) = \dfrac{1}{n}\{\text{Number of } y_i \leq y\}$

1. Draw a sample of size n from $\hat{F}_n$
   This is the bootstrap sample $(y_1^*, y_2^*, \ldots, y_n^*)$ with y^* selected with replacement from $(y_1, y_2, \ldots, y_n)$
2. Obtain an estimate $\hat{\theta}^*$ from the bootstrap sample
3. Repeat steps 1 and 2 , say, B times

Empirical distribution of $\hat{\theta}^*$

An estimate of the sampling distribution of θ, and is referred to as the bootstrap empirical distribution of $\hat{\theta}$

$$y_1, y_2, \ldots, y_n \rightarrow \left.\begin{array}{l} \text{Sample 1: } (y_1^*, y_2^*, \ldots, y_n^*) \rightarrow \hat{\theta}_1^* \\ \text{Sample 2: } (y_1^*, y_2^*, \ldots, y_n^*) \rightarrow \hat{\theta}_2^* \\ \vdots \\ \text{Sample } B: \ y_1^*, y_2^*, \ldots, y_n^* \rightarrow \hat{\theta}_B^* \end{array}\right\} \rightarrow \text{Bootstrap empirical distribution of } \hat{\theta}.$$

$$\text{Mean}: \ \hat{E}(\hat{\theta}) = \frac{1}{B}\sum_{j=1}^{B} \hat{\theta}_j^* \qquad \text{Variance}: \ \widehat{Var}(\hat{\theta}) = \frac{1}{B-1}\left\{\sum_{j=1}^{B}\left(\hat{\theta}_j^*\right)^2 - \frac{1}{B}\left(\sum_{j=1}^{B}\hat{\theta}_j^*\right)^2\right\}$$

$(1-\alpha)\%$ confidence interval

$\left(k_{\alpha/2}, k_{1-\alpha/2}\right)$ where k_α denotes the αth empirical quantile of the bootstrap values $\hat{\theta}^*$

Parametric Bootstrap

first estimates parameters of the data-generating process and then simulates new values by drawing from this estimated distribution

HYPOTHESIS TESTING

$100(1-\alpha)\%$ confidence interval for θ: $\hat{\theta}_1(\underline{X}), \hat{\theta}_2(\underline{X})$ depending on the sample $\underline{X} = (X_1, \ldots, X_n)$ such that $P\left(\hat{\theta}_1(\underline{X}) < \theta < \hat{\theta}_2(\underline{X})\right) = 1 - \alpha$

The Pivotal Method

Pivotal quantity of the form $g(\underline{X}, \theta)$

1. It is a function of the sample values and the unknown parameter θ
2. Its distribution is completely known
3. It is monotonic in θ

$N(\mu,\ \sigma^2)$ with known σ^2	$\dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	Confidence interval: $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
$N(\mu,\ \sigma^2)$ with unknown σ^2	$\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$	Confidence interval: $\bar{X} \pm t_{1-\alpha/2,n-1}\dfrac{S}{\sqrt{n}}$
Estimation of normal variance σ^2	$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$	Confidence interval: $\left(\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$
$N(X_{n+1},\ \sigma^2)$ with unknown σ^2	$\dfrac{\bar{X} - X_{n+1}}{S\sqrt{1+1/n}} \sim t_{n-1}$	Prediction interval: $\bar{X} \pm t_{\alpha/2,n-1}S\sqrt{1+1/n}$
Binomial distribution with $\hat{p} = \dfrac{X}{n}$	$\dfrac{X - np}{\sqrt{np(1-p)}}$	Confidence interval: $\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
Poisson distribution	$\dfrac{\bar{X} - \lambda}{\sqrt{\lambda/n}}$	Confidence interval: $\bar{X} \pm z_{\alpha/2}\sqrt{\dfrac{\bar{X}}{n}}$

Two Normal means with known σ_1^2 and σ_2^2

Confidence interval: $\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Two Normal means with known or unknown σ_1^2 and σ_2^2

Confidence interval: $\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2,n_1+n_2-2} \cdot S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$

where $S_p^2 = \dfrac{(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2}{n_1+n_2-2}$

Two population variances

$\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$

Confidence interval: $\dfrac{S_1^2}{S_2^2} \cdot \dfrac{1}{F_{n_1-1,n_2-1}} < \dfrac{\sigma_1^2}{\sigma_2^2} < \dfrac{S_1^2}{S_2^2} \cdot F_{n_2-1,n_1-1}$

Two population proportions

$\dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

Two Poisson parameters

$\bar{X}_1 - \bar{X}_2 \to N\left(\lambda_1 - \lambda_2, \dfrac{\lambda_1}{n_1} + \dfrac{\lambda_2}{n_2}\right)$

Confidence interval: $\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\left(\frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}\right)}$

Paired data

$\dfrac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$

Confidence interval: $\bar{D} \pm t_{\alpha/2,n-1}\dfrac{S_D}{\sqrt{n}}$

Null hypothesis H_0

Alternative hypothesis H_1

The following are the same • 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

Type I error The error committed when a TRUE null hypothesis is rejected

Type II error The error committed when a FALSE null hypothesis is failed to be rejected.

Sensitivity The probability that an event that does occur is predicted

Specificity The probability that an event that does not occur is predicted to not occur

Neyman–Pearson Lemma $\frac{L(\theta_0)}{L(\theta_1)} \leq k$for all values $(x_1,\ldots,x_n) \in C$ $\frac{L(\theta_0)}{L(\theta_1)} > k$for all values $(x_1,\ldots,x_n) \notin C$.
rejection region C constitute a uniformly most powerful test

Likelihood Ratio Tests Mean $\mu \rightarrow H_0 : \mu = \mu_0$ Test statistic: $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$

variance $\sigma^2 \rightarrow H_0 : \sigma^2 = \sigma_0^2$ $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$

p-values The lowest level at which H_0 can be rejected

Population Mean $H_0 : \mu = \mu_0$ Test statistic: $\bar{X}$, and $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ under H_0 with σ known

Test statistic: $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$ under H_0 with σ unknown

Population Variance $H_0 : \sigma^2 = {\sigma_0}^2$ Test statistic: $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$ under H_0

Population Proportion $H_0 : p = p_0$ Test statistic: $X \sim$ binomial (n, p_0) under H_0

Mean of a Poisson Distribution $H_0 : \lambda = \lambda_0$ **Test statistic:** $\bar{X}$, and $\frac{\bar{X}-\lambda_0}{\sqrt{\lambda_0/n}} \sim N(0,1)$ under H_0

Difference Between Two Population Means $H_0 : \mu_1 - \mu_2 = \delta$ Test statistic: $z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ with σ_1^2, σ_2^2 known

Test statistic: $t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ with σ_1^2, σ_2^2 unknown

and $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

Ratio of Two Population Variances: $H_0 : \sigma_1^2 = \sigma_2^2$ v $H_1 : \sigma_1^2 \neq \sigma_2^2$ Test statistic: $S_1^2/S_2^2 \sim F_{n_1-1,n_2-1}$ under H_0

Difference Between Two Population Proportions $H_0 : p_1 = p_2$ Test statistic: $\frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}+\frac{\hat{p}(1-\hat{p})}{n_2}}} \sim N(0,1)$ under H_0

Difference Between Two Poisson Means $H_0 : \lambda_1 = \lambda_2$ Test statistic: $\frac{(\hat{\lambda}_1-\hat{\lambda}_2)}{\sqrt{\frac{\hat{\lambda}}{n_1}+\frac{\hat{\lambda}}{n_2}}} \sim N(0,1)$ under H_0

Paired Data $H_0 : \mu_D\,(=\mu_1-\mu_2) = \delta$ Test statistic: $\frac{\bar{D}-\delta}{S_D/\sqrt{n}} \sim t_{n-1}$ under H_0

Permutation Approach All possible permutations of the data subject to some criterion

Chi-square Tests Test statistic: $\sum \frac{(f_i-e_i)^2}{e_i}$

Contingency Table A two-way table of counts obtained when sample items are classified according to two category variables
The proportion of data in row i is $\sum_j f_{ij} / \sum_i \sum_j f_{ij}$
The number expected in cell (i,j) is $\left(\sum_j f_{ij} / \sum_i \sum_j f_{ij}\right) \times \left(\sum_i f_{ij}\right)$

Fisher's Exact Test

$$P\left(n_{X_1Y_1}\right) = \frac{\binom{n_{X_1}}{n_{X_1Y_1}}\binom{n_{X_2}}{n_{Y_1}-n_{X_1Y_1}}}{\binom{n}{n_{Y_1}}} \quad \text{for} \quad n_{X_1Y_1} \leq n_{X_1}, n_{Y_1}$$

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis

The process of analysing data to gain further insight into the nature of the data

Pearson Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$$

Sum of Squares

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n$$

$$S_{xx} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n$$

Sample variance of x and y: $s_x^2 = \frac{S_{xx}}{n-1}$ $s_y^2 = \frac{S_{yy}}{n-1}$ **Sample Covariance** $cv_{xy} = \frac{S_{xy}}{n-1}$

Spearman's rank Correlation coefficient

$$r_s = 1 - \frac{6\sum_i d_i^2}{n\left(n^2-1\right)}$$

where $d_i = r\left(X_i\right) - r\left(Y_i\right)$ and Pearson correlation coefficient $r\left(X_i\right)$ and $r\left(Y_i\right)$

Kendall rank Correlation coefficient

$\tau = \frac{n_c - n_d}{n(n-1)/2}$ where n_c is the number of concordant pairs, and n_d is the number of discordant pairs

Any pair of observations $(X_i, Y_i); (X_j, Y_j)$ where $i \neq j$, is **concordant** if the ranks for both elements agree, i.e. $X_i > X_j$ and $Y_i > Y_j$, or $X_i < X_j$ and $Y_i < Y_j$; otherwise **discordant**

Scatter Plot Matrix

Each entry of this matrix is a scatter plot for a pair of variables identified by corresponding row and column labels

Principal Component Analysis (PCA)

Eigenvalues of matrix A

The values λ such that $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ where $\mathbf{I}$ is the identity matrix
The corresponding eigenvector, $\mathbf{v}$, of an eigenvalue λ satisfies the equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0$

Covariance of the data $\mathbf{X}^T\mathbf{X}$

The principal components decomposition $\mathbf{P}$ of $\mathbf{X} \rightarrow \mathbf{P} = \mathbf{XW}$.

The explanatory power of each component $\rightarrow \mathbf{S} = \mathbf{P}^T\mathbf{P}$

LINEAR REGRESSION

Bivariate Model

$Y_i = \alpha + \beta x_i + e_i \quad i = 1, 2, \ldots, n$
$E[Y|x] = \alpha + \beta x$
α(intercept) and β(slope parameter) are regression coefficients, e_i is the random error term; e_i are independent with $\mathrm{E}[e_i] = 0$ and $\mathrm{Var}(e_i) = \sigma^2$

Fitted Regression

$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$

Statistic $\hat{B}$ $\quad E[\hat{B}] = \beta \qquad Var[\hat{B}] = \dfrac{\sigma^2}{S_{xx}}$

Error variance σ^2 $\quad \hat{\sigma}^2 = \dfrac{1}{n-2}\sum (y_i - \hat{y}_i)^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$

Partition Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{Total sum of quares}} = \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{Residual sum of squares}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{\text{Regression sum of squares}} + \underbrace{2\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\bar{y})}_{\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i)(\widehat{\beta}_0+\widehat{\beta}_1 x_i-\bar{y})=0}$$

Total Sum of Squares SS_{TOT}: Amount of variability inherent in the response prior to performing regression

Residual Sum of Squares SS_{RES}: Variation unexplained by the linear regression model

Regression Sum of Squares SS_{REG}: Variation explained by the linear regression model

Coefficient of Determination The proportion of the total variability of the responses 'explained' by a model

$$R^2 = \frac{SS_{REG}}{SS_{TOT}} = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{S_{xy}^2}{S_{xx}S_{yy}},$$

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

$R^2 = r^2$ (the square of the Pearson's correlation coefficient)

Adjusted R^2 $\quad$ Adjusted $R^2 = 1 - \dfrac{MSS_{RES}}{MSS_{TOT}} = 1 - \left(\dfrac{n-1}{n-k-1}\right)(1-R^2)$

ANOVA Table

Source of variation	degree of freedom	Sums of squares	Mean sums of squares
Regression	k	SS_{REG}	SS_{REG}/k
Residual	$n-k-1$	SS_{RES}	$SS_{REG}/(n-k-1)$
Total	$n-1$	SS_{TOT}	

Mean Response $\quad \mu_o = E[Y|x_o] = \alpha + \beta x_o \qquad Var(\widehat{u}_o) = \sigma^2\left(\dfrac{1}{n} + \dfrac{(x_o-\bar{x})^2}{S_{xx}}\right)$

$$\text{SE}(\widehat{\mu}_o) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_o-\bar{x})^2}{S_{xx}}\right)}$$

Individual Response

$$\text{SE}(\hat{y}_0) = \sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_o-\bar{x})^2}{S_{xx}}\right)}$$

Residual (Raw Residual) $\quad e_i = y_i - \widehat{y}_i =$ observed value $-$ fitted value

Sum-to-zero Constraints on Residuals

$$\sum_{i=1}^{n} e_i = 0, \qquad \sum_{i=1}^{n} x_i e_i = 0$$

Multivariate Model $\quad E[Y|x_1, x_2, \ldots, x_k] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

β_j is the regression coefficient attached to the jth predictor, for $j = 1, \ldots, k$

$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots \beta_k x_{ik} + e_i \quad i = 1 \ldots n$

Mean Response $\quad \mu_0 = E[Y|\mathbf{x_0}] = \alpha + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k}$

Individual Response $\quad \hat{y}_0 = \hat{\alpha} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}$

Forward Stepwise Selection There is a total of $1 + \frac{k(k+1)}{2}$ fitted models

1. Start with model having intercept only
2. Create $p+1$ predictor models by fitting a model with the current p predictors plus one of the $k-p$ unused predictors.
3. Select the best $p+1$ predictor model based on SS_{RES} or R^2
4. If $p+1 < k$, repeat steps $2-3$ with the $p+1$ parameter model.
5. Select the best model from the various models based on adjusted R^2 or AIC

Nested model the predictors in the p-predictor model are always a subset of the predictors in the $(p+1)$-predictor model

Backward Stepwise Selection There is a total of $1 + \frac{k(k+1)}{2}$ fitted models

1. Start with full model
2. create $p-1$ predictor models by fitting a model removing one of the parameters from the current p predictors
3. Select the best $p-1$ predictor model based on SS_{RES} or R^2
4. If $p-1 > 1$, repeat steps $2-3$ with the $p-1$ parameter model
5. Select the best model from the various models based on adjusted R^2 or AIC Backward selection cannot be implemented in the high-dimensional setting with $n \leq k$

Polynomial Regression: $Y = \alpha + \beta_1 x + \boldsymbol{\beta_2 x^2} + \cdots + \boldsymbol{\beta_m x^m} + \varepsilon$

$m = 2 \rightarrow$ quadratic regression, $m = 3 \rightarrow$ cubic polynomial

Regression with Interaction Term: $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \boldsymbol{\beta_3 x_1 x_2} + \varepsilon$

$x_1 x_2$ is called an **interaction term**

GENERALISED LINEAR MODELS

Exponential Family of Distributions $f_Y(y;\theta,\varphi) = \exp\left\{\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y,\varphi)\right\}$

- φ is the scale parameter.
- θ the 'natural' parameter

Mean and Variance $\mu = E[y] = b'(\theta)$ and $Var(y) = a(\varphi)b''(\theta)$

Distribution	θ	$b(\theta)$	φ
Normal, N (μ, σ^2)	μ	$\theta^2/2$	σ^2
Poisson (λ)	$\log\lambda$	e^θ	1
Binomial, Bin(n,μ)	$\log\left(\frac{\mu}{1-\mu}\right)$	$\log\left(1+e^\theta\right)$	n
Gamma	$-\frac{1}{\mu}$	$-\log(-\theta)$	α

Members of Exponential Family in Canonical Form

Distribution	Canonical Link Function	Mathematical Form
Normal	Identity	$g(\mu) = \mu$
Poisson	Log	$g(\mu) = \log\mu$
Binomial	Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Gamma	Inverse	$g(\mu) = \frac{1}{\mu}$

Obtaining the estimates: Maximising l with respect to the parameters in the linear predictor

Significance of the Parameters If $|\hat{\beta}| > 2$ standard error $(\hat{\beta})$, the parameter is significant and should be retained in the model

Deviance for Current Model D_M

Scaled Deviance A goodness-of-fit measure of how much the fitted GLM departs from the saturated model $2(l_{\text{SAT}} - l)$ scaled deviance $= \dfrac{D_M}{\varphi}$

Saturated Model The fitted values exactly equal the observed values, $\widehat{\mu}_i = y_i$ for all $i = 1, \ldots, n$ under the saturated model

Scaled Deviance Comparasion If $\dfrac{(S_1 - S_2)/q}{S_2/(n-(p+q))} >$ the 5% value for the $F_{q,n-p-q}$ distribution, model 2 is a significant improvement over Model 1 where Model 1 which has p parameters and scaled deviance S_1 and Model 2 has $p+q$ parameters and scaled deviance S_2

Akaike Information Criterion AIC $= -2 \times \log L_M + 2\times$ parameters, where $\log_{LM}$ is the log likelihood of the model under consideration the smaller the AIC, the better the model

Inverse of the Link Function $\mu = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$ for $\eta = \mathbf{x}'\boldsymbol{\beta}$.

Pearson residuals: $\dfrac{y - \hat{\mu}}{\sqrt{Var(\hat{\mu})}}$, $\dfrac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$ (Poisson) $\dfrac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i(1-\widehat{\mu}_i)}}$ (Bernoulli)

Deviance Residuals $\text{sign}(y - \hat{\mu})d_{i'}$ where $\sum_{i=1}^{n} d_i^2 = D^*$

Residual Plots Create residual plot (e) against fitted values $(\widehat{y})$ of the regression model

BAYESIAN STATISTICS

Bayes' Theorem: $P(B_r|A) = \dfrac{P(A|B_r)P(B_r)}{P(A)}$ where $P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$ for $r = 1, 2, \ldots, k$

Prior Density $f_\Theta(\theta)$ (continuous) $p_\Theta(\theta)$ (discrete)

Posterior Density $f(\theta \mid \underline{X}) = \dfrac{f(\theta, \underline{X})}{f(\underline{X})} = \dfrac{f(\underline{X} \mid \theta) f(\theta)}{f(\underline{X})}$ where $f(\underline{X}) = \int f(\underline{X} \mid \theta) f(\theta) d\theta$

Conjugate Prior The prior distribution leads to a posterior distribution belonging to the same family as the prior distribution

Quadratic Loss $L(g(\underline{x}), \theta) = [g(\underline{x}) - \theta]^2$

Absolute Error Loss $L(g(\underline{x}), \theta) = |g(\underline{x}) - \theta|$

'All-or-nothing' Loss $L(g(\underline{x}), \theta) = \begin{cases} 0 & \text{if } g(\underline{x}) = \theta \\ 1 & \text{if } g(\underline{x}) \neq \theta \end{cases}$

Bayesian Credible Interval $P(\theta \in A \mid x) = \int_A f(\theta \mid x) d\theta = 1 - \alpha$

CREDIBILITY THEORY

For any random variables X and Y: $E[X] = E[E(X \mid Y)]$

Two random variables X_1 and X_2 are conditionally independent given a third random variable Y:

$$E[X_1 X_2 \mid Y] = E[X_1 \mid Y] E[X_2 \mid Y]$$

Credibility Premium $Z\bar{X} + (1 - Z)\hat{\mu}$ where Z is credibility factor

Bayesian Credibility

Model	Posterior Distribution	Credibility Factor	Posterior Mean
Poisson(λ) /Gamma(α, β)	$\text{Gamma}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$	$\frac{n}{\beta + n}$	$Z\left[\sum_{i=1}^{n} x_i/n\right] + (1-Z)\alpha/\beta$
Normal/ Normal $N\left(\theta, \sigma_1^2\right)/N\left(\theta, \sigma_2^2\right)$	$N\left(\frac{\frac{n\bar{x}}{\sigma_1^2} + \frac{\mu}{\sigma_2^2}}{\frac{n}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \frac{1}{\frac{n}{\sigma_1^2} + \frac{1}{\sigma_2^2}}\right)$	$\frac{n/\sigma_1^2}{n/\sigma_1^2 + 1/\sigma_2^2} = \frac{n}{n + (\sigma_1^2/\sigma_2^2)}$	$Z\bar{x} + (1-Z)\mu$

Empirical Bayesian Credibility Theory

Model 1

$m(\theta) = E\left[X_j \mid \theta\right]$ $\quad s^2(\theta) = Var\left[X_j \mid \theta\right]$ $\quad \bar{X}_i = \dfrac{\sum_{j=1}^{n} X_{ij}}{n}$

$E[m(\theta)] = \bar{X}$ $\quad E\left[s^2(\theta)\right] = N^{-1}\sum_{i=1}^{N}(n-1)^{-1}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$

$Var[m(\theta)] = (N-1)^{-1}\sum_{i=1}^{N}\left(\bar{X}_i - \bar{X}\right)^2 - (Nn)^{-1}\sum_{i=1}^{N}(n-1)^{-1}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$

Credibility factor $\quad Z = \dfrac{n}{n + E\left[s^2(\theta)\right]/Var[m(\theta)]}$

Credibility premium $\quad Z\bar{X}_i + (1-Z)\,E[m(\theta)]$

Model 2

$m(\theta) = E\left[X_j \mid \theta\right]$ $\quad s^2(\theta) = P_j\,Var\left[X_j \mid \theta\right]$ $\quad \bar{X}_i = \dfrac{\sum_{j=1}^{n} P_j X_j}{\sum_{j=1}^{n} P_j}$

$E[m(\theta)] = \bar{X}$ $\quad E\left[s^2(\theta)\right] = N^{-1}\sum_{i=1}^{N}(n-1)^{-1}\sum_{j=1}^{n} P_{ij}\left(X_{ij} - \bar{X}_i\right)^2$

$Var[m(\theta)] = P^{*-1}\left((Nn-1)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}\left(X_{ij} - \bar{X}\right)^2 - N^{-1}\sum_{i=1}^{N}(n-1)^{-1}\sum_{j=1}^{n} P_{ij}\left(X_{ij} - \bar{X}_i\right)^2\right)$

Credibility Factor $\quad Z = \dfrac{\sum_{j=1}^{n} P_j}{\sum_{j=1}^{n} P_j + E\left[s^2(\theta)\right]/Var[m(\theta)]}$ $\quad Z_i = \dfrac{\sum_{j=1}^{n} P_{ij}}{\sum_{j=1}^{n} P_{ij} + E\left[s^2(\theta)\right]/\text{var}[m(\theta)]}$

Credibility Premium $\quad Z\bar{X}_i + (1-Z)\,E[m(\theta)]$