Notes: Statistics and Probabilistic Models

Johnew Zhang

October 25, 2015

Acknowledgement

Thank you for Craig Sloss's contribution to most of portion of Probability Models and Statistics Section.

Contents

1 Probability Models (Stochastic Process and Survival Methods)

1.1 Poisson Processes

Learning Objectives

Understand and apply the properties of Poisson processes:

- For increments in the homogeneous case
- For interval times in the homogeneous case
- For increments in the non-homogeneous case
- Resulting from special types of events in the Poisson process
- Resulting from sums of independent Poisson processes

1.1.1 Background

Definition 1.1. *A* ***Poisson random variable*** *X with mean λ is a random variable with the following probability mass function:*

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $x \in \mathbb{N}$ and $Var(X) = E[X] = \lambda$.

Definition 1.2. *An* ***Exponential random variable*** *X with mean $1/\lambda$ is a random variable with the following probability density function:*

$$f(x) = \begin{cases} \lambda e^{-x\lambda} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The cumulative function of X can be easily derived as:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Definition 1.3. *A random variable X is said to be without memory, or* ***memoryless****, if*

$$P(X > s + t | X > t) = P(X > s), \forall s, t \geq 0$$

Here, both Poisson and Exponential random variables have the memoryless property. In other words, $\tilde{F}(t+s) = \tilde{F}(t)\tilde{F}(s)$.

Definition 1.4. *A **Gamma random variable** X with parameters (n, λ) is a random variable with the following probability density function*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} & t \geq 0 \\ 0 & O.W. \end{cases}$$

Proposition 1.1. The sum of n independent exponential random variables, each having parameter λ is a gamma random variable with parameters (n, λ).

1.1.2 Definition of Poisson Process

Definition 1.5. *A stochastic process $\{N(t), t \geq 0\}$ is said to be a **counting process** if $N(t)$ represents the total number of "events" that occur by time t. Intuitively, counting process has the following properties:*

1. *$N(t) \in \mathbb{N}$*
2. *If $s < t$, then $N(s) \leq N(t)$*
3. *For $s < t$, $N(t) - N(s) = N\{(s, t]\}$.*

*A counting process is said to possess **independent increments** if the numbers of events that occur in disjoint time intervals are independent.*

*A counting process is said to possess **stationary increments** if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval.*

The final ingredient for a definition of a Poisson process is the rate at which events are, on average, occurring. The **rate (or intensity) function** λ gives the rate as $\lambda(t)$ at time t.

Definition 1.6. *A **Poisson process** N with rate function λ has the following properties:*

1. *N is a counting process where $N(0) = 0$, and for $t > 0$, $N(t)$ is non-decreasing and takes on natural numbers.*
2. *N has independent increments.*

3. *$\forall t \geq 0$ and $h > 0$, the increment $N(t+h) - N(t)$ is a Poisson Random variable with mean $\lambda = \int_t^{t+h} \lambda(z)dz$.*

4. *the function m defined by $m(t) = \int_0^t \lambda(z)dz$ is called the mean-value function, since $E[N(t)] = m(t)$. m is often called the operational time.*

5. *If the rate function is in fact a constant, then N is called a homogenous Poisson process[1]; otherwise, non-homogeneous.*

1.1.3 Homogenous Poisson Process

First, let's suppose function $f(.)$ is said to be $o(h)$ if $\lim_{h\to 0} f(h)/h = 0$. Note $N(h) = N(t+h) - N(t) \sim Poisson(\lambda h)$. Hence

$$P\{N(h) = n\} = e^{-\lambda h}(\lambda h)^n/n!, n \in \mathbb{N}.$$

Then from definition 1.6, we can derive the following:

- $P[N(h) = 1] = \lambda h + o(h)$ (Taylor expansion)
- $P(N(h) \geq 2] = o(h)$ (Easy; just expand)

1.2 Inter-arrival and Waiting Distribution Associated with the Poisson Process

Learning Objectives

For any Poisson process and inter-arrival and waiting distributions associated with the Poisson process, calculate:

- Expected Values
- Variance
- Probability

[1] In the context, only Poisson process is mentioned. One shall regard it as the homogenous Poisson process

1.2.1 Interarrival and Waiting Time Distribution

Consider a Poisson process, and let us denote the time of the first event by T_1. Further, for $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and the nth event. The sequence $\{T_n, n = 1, 2, \cdots\}$ is called the sequence of inter-arrival times.

Proposition 1.2. $T_n, n = 1, 2, \cdots$ are independent identically distributed exponential random variables having mean $1/\lambda$.

The assumption of stationary and independent increments is basically equivalent to asserting that, at any point in time, the process probabilistically restarts itself. That is, the process from any point on is independent of all that has previously occurred and also has the same distribution as the original process. In other wards, the process has no memory. Hence exponential is expected.

In general, the probability that T_1 is at least x can be expressed as the following:

$$F_{T_1}(x) = 1 - e^{-1 \int_0^x \lambda(t)dt}$$

Anther quantity of interest is S_n, the arrival time of the nth event, also called the waiting time until the nth event. It is easily seen that

$$S_n = \sum_{i=1}^{n} T_i, n \geq 1$$

Hence $S_n \sim Gamma(n, \lambda)$ by Prop. 1.2 and Prop. 1.1

1.2.2 Further Properties of Poisson Processes

Let $N_1(t)$ and $N_2(t)$ denote respectively the number of type I and type II events occurring in $[0, t]$, Note $N(t) = N_1(t) + N_2(t)$.

Proposition 1.3. $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are both Poisson processes having respective rates λp and $\lambda(1-p)$. Furthermore, the two processes are independent.

Proposition 1.4. Suppose we are given two Poisson processes with rates λ_1 and λ_2. Then the probability that an event from the first process occurs before an event from the second process is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2}$$

To calculate the probability that k_1 events from process 1 occur before k_2 events from process 2, calculate the probability that at least k_1 of the next k_1+k_2-1 events are from process 1. The probability that an event is from process 1 is $\lambda_1/(\lambda_1+\lambda_2)$, so the probability at least k_1 out of k_1+k_2-1 vents are from process 1 is a sum of binomial probabilities of $k_1, k_1+1, \cdots, k_1+k_2-1$. The parameters of the binomial distribution of k_1+k_2-1 and $\lambda_1/(\lambda_1+\lambda_2)$.

1.2.3 Conditional Distribution of the Arrival Times

Suppose we are told that exactly one event of a Poisson process has taken place by time t, and we are asked to determine the distribution of the time at which the event occurred. Now, since a Poisson process possesses stationary and independent increments it seems reasonable that each interval in $[0,t]$ of equal length should have the same probability of containing the event. In other words, the time of the events should be uniformly distributed over $[0,t]$. We can generalize this result but before doing so, we will introduce the concept of order statistics.

Let $Y_1, Y_2, \cdots, Y_n$ be n random variables. We say that $Y_{(1)}, Y_{(2)}, \cdots, Y_{(n)}$ are the order statistics corresponding $Y_1, Y_2, \cdots, Y_n$ if $Y_{(k)}$ is the kth smallest value among $Y_1, \cdots, Y_n, k = 1, 2, \cdots, n$. If the $Y_i, i = 1, \cdots, n$ are independent identically distributed continuous random variables with probability density f, then the joint density of the order statistics is given by

$$f(y_1, y_2, \cdots, y_n) = n! \prod_{i=1}^{n} f(y_i), y_1 < y_2 < \cdots < y_n$$

The preceding follows since

1. $(Y_{(1)}, Y_{(2)}, \cdots, Y_{(n)})$ will equal $(y_1, y_2, \cdots, y_n)$ if $(Y_1, Y_2, \cdots, Y_n)$ is equal to any of the $n!$ permutations of $(y_1, y_2, \cdots, y_n)$;

2. The probability density th

   at $(Y_1, Y_2, \cdots, Y_n)$ is equal to $y_{i_1}, \cdots, y_{i_n}$ is $\prod_{j=1}^{n} f(y_{i_j}) = \prod_{j=1}^{n} f(y_i)$ when $i_1, \cdots, i_n$ is a permutation of $1, 2, \cdots, n$. If Y_i are uniformly distributed over $(0,t)$, then $f(y_1, y_2, \cdots, y_n) = \frac{n!}{t^n}$.

Theorem 1.1. *Given that $N(t) = n$, the n arrival times $S_1, \cdots, S_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval $(0,t)$.*

Proposition 1.5. If $N_i(t), i = 1, \cdots, k$, represents the number of type i events occurring by time t then $N_i(t), i = 1, \cdots, k$, are independent Poisson random variables having means

$$E[N_i(t)] = \lambda \int_0^t P_i(s) ds$$

where $P_i(s)$ is the probability that ith event occurs during $(0, s)$.

Hence, the expected value of the jth variable is $jt/(k+1)$ where t is the waiting time and k is the number of Poisson events.

Proposition 1.6. Given that $S_n = t$, the set $S_1, \cdots, S_{n-1}$ has the distribution of a set of $n-1$ independent uniform $(0, t)$ random variables.

1.2.4 Greedy algorithms

Greedy algorithms can be analyzed by considering the density function of the minimum of a collection of exponential random variables (provided the costs are independent and exponentially distributed). Note that

$$\begin{aligned} P(\min(X_i) > x) &= \prod_{1 \le i \le n} P(X_i > x) \\ &= \exp\left(-x \sum_{1 \le i \le n} \mu_i\right). \end{aligned}$$

In other words, the distribution of the minimum is also exponential, and the rate is obtained as the sum of the rates of the individual variables.

This can be applied to determine the expected cost of the result of a Greedy Algorithms by:

1. Considering the number of options available at each step.
2. If all job-worker pairs are considered at each stage, add the expected value of the preceding choices (accounting for the memoryless property of the exponential distribution).

1.2.5 Generalizations of Poisson Process

As we discussed above, the non-homogeneous Poisson process does not have constant intensity functions $\lambda(t)$. Now let's look at some generalized properties below

Proposition 1.7. Let $\{N(t), t \geq 0\}$ and $\{M(t), t \geq 0\}$, be independent non homogeneous Poisson processes, with respective intensity function $\lambda(t)$ and $\mu(t)$, and let $N^*(t) = N(t) + M(t)$, Then, the followings are true

- $\{N^*(t), t \geq 0\}$ is a non homogeneous Poisson process with intensity function $\lambda(t) + \mu(t)$.
- Given that an event of the $\{N^*(t)\}$ process occurs at time t then, independent of what occurred prior to t, the event at t was from the $\{N(t)\}$ process with probability $\frac{\lambda(t)}{\lambda(t)+\mu(t)}$.

1.3 Compound Poisson Processes

Learning Objectives: For a compound Poisson process, calculate the moments associated with the value of the process at a given time.

A *compound Poisson process* S is of the form

$$S(t) = \sum_{1 \leq j \leq N(t)} X_j,$$

where $N(t)$ is a Poisson process, and the set $\{X_j\}$ are independent and identically distributed random variables.

The expected value of a compound Poisson process can be calculated as follows:

$$\begin{aligned}
E[S(t)] &= \sum_{n \geq 1} P(N(t) = n) E[\sum_{1 \leq j \leq n} X_j] \\
&= \sum_{n \geq 1} e^{-\lambda t} \frac{(\lambda t)^n}{n!} n E[X_1] \\
&= \lambda t E[X_1] e^{-\lambda t} \sum_{n \geq 1} \frac{(\lambda t)^{n-1}}{(n-1)!} \\
&= \lambda t E[X_1].
\end{aligned}$$

The variance can be determined using the conditional variance formula as follows:

$$\begin{aligned}
\mathrm{Var}(S(t)) &= E(\mathrm{Var}(S(t)|N(t))) + \mathrm{Var}(E[S(t)|N(t)]) \\
&= \sum_{n \geq 1} P(N(t) = n) n \mathrm{Var}(X_1) + \mathrm{Var}(N(t) E[X_1]) \\
&= (E[X_1^2] - E[X_1]^2) \sum_{n \geq 1} e^{-\lambda t} \frac{(\lambda t)^n}{(n-1)!} + \lambda t E[X_1]^2 \\
&= \lambda t E[X_1^2].
\end{aligned}$$

1.4 Hazard Functions

Learning Objectives: Apply Poisson Process concepts to calculate the hazard function and related survival model concepts:

- Relationship between hazard rate, probability density function, and cumulative distribution function
- Effect of memoryless nature of Poisson distribution on survival time estimation

Suppose that X is a random variable representing the "failure time" of some object. Let $f(x)$ be its density function, and $F(x) = \int_0^x f(t)dt$ be its cumulative distribution function. The *hazard rate function* is the density function $\lambda(t)$ of the failure time, conditional on the object having survived up to time t. In symbols,

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{\overline{F}(t)}.$$

Note that if the failure time is exponentially distributed with $\theta = 1/\lambda$, then

$$\lambda(t) = \frac{\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})} = \lambda,$$

consistent with the memoryless property of the exponential distribution.

The hazard rate function is also called the *failure rate function*. The system has *increasing (decreasing) failure rate* if $\lambda(t)$ is an increasing (decreasing) function of t.

By integrating the hazard rate function, we obtain:

$$\int_0^t \lambda(s)ds = \int_0^t \frac{F'(s)}{1 - F(s)}ds = -\log(\overline{F}(t)).$$

This means that

$$\overline{F}(t) = e^{-\Lambda(t)},$$

where

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

The function Λ is called the *hazard function* for F.

1.4.1 Failure time random variables

For a failure time random variable in which the failure rate does not change over time, use the exponential distribution. If there are multiple parts that can fail, with distinct rates $\lambda_1, \ldots, \lambda_n$, then use a hypoexponential distribution (which is a convolution of the denstiy functions for each component). This has density function

$$f(t) = \sum_{1 \leq i \leq n} C_i \lambda_i e^{-\lambda_i t},$$

where

$$C_i = \prod_{j \neq i} \frac{\lambda_i}{\lambda_j - \lambda_i}.$$

A key property of the hypoexponential distribution is that

$$\lim_{t \to \infty} r(t) = \min(\lambda_1, \ldots, \lambda_n).$$

In other words, for large t, the failure rate approaches the failure rate of the longest-lived component.

1.4.2 Cumulative distribution functions

Given a hazard rate function

$$r(t) = \frac{\frac{d}{dt} F(t)}{1 - F(t)},$$

we can determine $F(t)$ by integrating both sides:

$$\int_0^t r(s) ds = -\log(1 - F(t)) + \log(1 - F(0)).$$

Simplifying,

$$F(t) = 1 - \exp\left(-\int_0^t r(s) ds\right).$$

1.4.3 Probability density functions

Given a hazard rate function, the corresponding density function can be obtained by differentiating the cumulative distribution function:

$$f(x) = r(t) \exp\left(-\int_0^t r(s) ds\right).$$

1.4.4 Hazard functions and relationship to Exponential distribution

A hazard rate function is constant if and only if it corresponds to an exponential distribution. In this case, the mean of the exponential distribution is $\theta = \frac{1}{r}$.

1.5 Reliability Theory

Learning Objectives: Given the joint distribution of more than one source of failure in a system (or life) and using Poisson Process assumptions:

- Calculate probabilities and moments associated with functions of these random variables.
- Understand difference between a series system (joint life) and parallel system (last survivor) when calculating expected time to failure or probability of failure by a certain time.
- Understand the effect of multiple sources of failure (mulitple decrement) on expected time to failure (expected lifetime).

For a system with n components, we use the binary variable x_i to specify whether component i is functioning or not. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the *state vector* of the system. Associated with each system is a *structure function*, $\phi(\mathbf{x})$, which is equal to 1 if the system is functioning, and 0 otherwise. Note that a structure function must be monotonic, i.e. if $\mathbf{y} \leq \mathbf{x}$, then $\phi(\mathbf{y}) \leq \phi(\mathbf{x})$.

A *series structure* functions if and only if all of its components are functioning; in other words,

$$\phi(\mathbf{x}) = \prod_{1 \leq i \leq n} x_i.$$

A *parallel structure* functions if and only if at least one of its components is functioning. In this case,

$$\phi(\mathbf{x}) = 1 - \prod_{1 \leq i \leq n} (1 - x_i).$$

Terminology that is used to analyze system reliability is as follows:

- If $\phi(\mathbf{x}) = 1$, then $\mathbf{x}$ is a *path vector.*
- If $\mathbf{x}$ is a path vector and $\phi(\mathbf{y}) = 0$ for all $\mathbf{y} < \mathbf{x}$, then $\mathbf{x}$ is a *minimal path vector.*
- If $\mathbf{x}$ is a minimal path vector, then $A = \{i : x_i = 1\}$ is a *minimal path set.*

- If $\phi(\mathbf{x}) = 0$, then $\mathbf{x}$ is a *cut vector.*
- If $\mathbf{x}$ is a cut vector, and $\phi(\mathbf{y}) = 1$ for all $\mathbf{y} > \mathbf{x}$, then $\mathbf{x}$ is a *minimal cut vector.*
- If $\mathbf{x}$ is a minimal cut vector, then $C = \{i : x_i = 0\}$ is a *minimal cut set.*

Minimal path sets and minimal cut sets can be used to express an arbitrary system as a parallel arrangement of series systems, or a series arrangement of parallel systems, respectively:

$$\phi(\mathbf{x}) = 1 - \prod_j \left(1 - \prod_{i \in A_j} x_i\right) = \prod_j \left(1 - \prod_{i \in C_j} (1 - x_i)\right).$$

If we assume that the components are independent, and that component i is functioning with probability p_i, let $r(\mathbf{p})$ denote the probability that the system is functioning; this is called a *reliability function.* Key facts about the reliability function include:

- $r(\mathbf{p}) = E[\phi(\mathbf{X})]$. (This is due to the fact that the expected value of a binary random variable is equal to the probability that the variable is one.)
- $r(\mathbf{p})$ is an increasing function of $\mathbf{p}$.

For a random variable X with density function f and cumulative distribution function F, its *tail distribution function* is

$$\overline{F}(t) = P(X > t) = 1 - F(t) = \int_t^\infty f(s)ds.$$

When studying the failure time of a system, we assume that once one of its components has failed, it remains in that state. (Note that the event "lifetime is greater than t" is equivalent to "system is functioning at time t.") Let $\overline{F}_i$ be the tail distribution for component i, and let $\overline{F}$ be the tail distribution for the whole system. We can determine $\overline{F}$ by using the reliability function:

$$\overline{F}(t) = r(\overline{F}_1(t), \ldots, \overline{F}_n(t)).$$

An important result that facilitates calculation of expected failure time is

$$E[X] = \int_0^\infty \overline{F}(t)dt.$$

The reason for this is:

$$\begin{aligned} E[X] &= \int_0^\infty y f(y) dy \\ &= \int_0^\infty \int_0^y f(y) dx dy \\ &= \int_0^\infty \int_x^\infty f(y) dx dy \\ &= \int_0^\infty \overline{F}(y) dy. \end{aligned}$$

1.5.1 Probabilities and moments

For the case of independent components, given the probability p_i that component i is functioning, the probability that the system is functioning can be determined as follows:

1. Determine the structure function for the system.
2. Determine the expected value of $\phi(X_1, \ldots, X_n)$. When doing this, first express the structure function as a linear combination of monomials, using $X_i^2 = X_i$ to simplify.

The minimum path sets can be used to derive an upper bound on the probability that the system functions. To do this, determine the probability that at least one component fails in each of the minimum path sets. Their product is a lower bound on the failure probability, and hence an upper bound on the probability that the system functions.

1.5.2 Time until failure of the system (life)

To determine the distribution of a system:

1. Determine the structure function $\phi(\mathbf{x})$ of the system.
2. Determine the reliability function as $r(\mathbf{p}) = E[\phi(\mathbf{X})]$.
3. Evaluate the reliability function at the tail distribution functions for each individual component.
4. If the density function is needed, differentiate the result and multiply by -1.

1.5.3 Time until failure of the system (life) for parallel or series systems with multiple components

For a series system,

$$\overline{F}(t) = \prod_{1\leq i\leq n} \overline{F}_i(t).$$

For a parallel system,

$$\overline{F}(t) = 1 - \prod_{1\leq i\leq n} (1 - \overline{F}_i(t)) = 1 - \prod_{1\leq i\leq n} F_i(t).$$

1.5.4 Paths that lead to parallel or series system failure for systems with multiple components

For a system in which k out of n components must function, a simple way to construct the structure function is to begin with the monomials corresponding to the k-subsets, then subtracting the duplicate $(k+1)$-subsets, adding back in the $(k+2)$-subsets, as appropriate. (We want each subset to be represented exactly once.)

1.5.5 Random graphs and defining path to failure

For a random graph with n vertices, we can interpret each potential edge as a component. Suppose we are interested in determining the probability that the graph is connected. The minimal cut sets consist of edges with one end in each side of a partition of the form (X, X^c), so there are $2^{n-1} - 1$ minimal cut sets. The minimal path sets are spanning trees, so there are n^{n-2} minimal path sets.

1.5.6 Method of inclusion and exclusion as applied to failure time estimates

The principal of inclusion-exclusion can be used determine bounds on the reliability function by considering the minimal path sets A_i. If E_i is the event in which all components of A_i function, then

$$r(\mathbf{p}) = P\left(\bigcup E_i\right) \leq \sum_i P(E_i).$$

Subtracting the two-way intersection terms reverses the inequality, adding the three-way intersection terms reverses the inequality again, etc.

The second way for obtaining upper and lower bounds for the reliability function is to express the probability of functioning or failure as an intersection of events.

The system will fail if every minimal path fail. Let D_i be the event that at least one component of a minimal path set A_1 fails. Then

$$1 - r(\mathbf{p}) = Pr(D_1)Pr(D_2|D_1)Pr(D_3|D_1 \cap D_2)\cdots$$

However, $Pr(D_2)$ is close than $Pr(D_2|D_1)$. If we are giving D_1, then some component of A_1 failed, and this component may be a component of A_2 as well. So the failure of D_1 increases the probability of failure of D_2. The same can be said about the conditions of the later factors. If s is the number of minimal path sets, then

$$r(\mathbf{p}) \leq 1 - \prod_{i=1}^{s}(1 - \prod_{l \in A_i} p_l)$$

The system will function if every minimal cut set has a functioning component. Let U_i be the event that minimal cut set C_i has a functioning component. The probability that a set is functioning given that a component of another set is functioning is greater than the unconditional probability that a set is functioning. Therefore, if there are r minimal cut sets,

$$r(\mathbf{p}) \geq \prod_{i=1}^{r}(1 - \prod_{l \in C_i}(1 - p_l))$$

1.5.7 Expected system lifetime as a function of component lifetime and properties of expected lifetime estimates

To determine the expected lifetime of a system,

1. Determine the reliability function r of the system.

2. Determine $\overline{F}(t)$ by evaluating r at the tail distributions for the component parts.

3. Integrate $\overline{F}(t)$ from 0 to ∞ to determine the expected failure time.

1.6 Markov Chains

Learning Objectives: for discrete and continuous Markov Chains under both homogeneous and non-homogeneous states:

- Definition of a Markov Chain

- Chapman-Kolmogorov Equations for n-step transition calculations
- Accessible states
- Ergodic Markov Chains and limiting probabilities
- Markov Chain Monte Carlo Methods

A Markov Chain is a stochastic process $\{X_n\}_{n\geq 0}$ in which each X_n takes on a countable number of *states*, $\{0, 1, 2, 3, \ldots\}$, and the probability that $X_{n+1} = j$ depends only on the state of X_n, and $P(X_{n+1} = j|X_n = i) = P_{ij}$. These will be analyzed using the matrix

$$\mathbf{P} = [P_{ij}].$$

(Note that the rows of this matrix correspond to the state before the transition, and the columns correspond to the state after the transition.) Note that the interpretation of P_{ij} as a probability means that all entries of this matrix are non-negative, and the sum of each row is 1. By interpreting $\mathbf{P}$ as an adjacency matrix, we can think of a Markov Chain as a directed graph with edge weights.

The n-step transition probabilities, denoted by P_{ij}^n, are the probabilities that a process, starting in state i, is in state j after exactly n steps. This probability can be calculated as the i, j-entry of $\mathbf{P}^n$. (It's just a random walk on the corresponding graph.)

Terminology used to identify special states is as follows:

- An *absorbing state* is a state that is never left after it is entered. This corresponds to a row of the matrix having a 1 in the diagonal entry, and zeroes elsewhere. Graphically, the vertex for this state has no outgoing edges. (One application of an absorbing state is to calculate the probability that a set of states A is reached *at some point* – add a single absorbing state to take the place of the states in A.)
- A state j is *accessible* from state i if $P_{ij}^n > 0$ for some n. Graphically, this means there is a directed path from i to j.
- Two states i and j are said to *communicate* if i is accessible from j, and j is accessible from i. Note that communication in a Markov chain is an equivalence relation, so communication partitions the states into equivalence classes. Graphically, the classes are the strongly connected components of the graph.

- A Markov Chain is *irreducible* if there is only one class. In this case, the corresponding graph is strongly connected.

- A state is *recurrent* if the probability of eventually re-entering state i, having started in state i, is 1. The state is *transient* if the probability is strictly less than 1. A recurrent state will be entered infinitely often, but the expected number of times a transient state will be entered is finite. Both recurrence and transience are class properties.

For an irreducible, aperiodic Markov chain, $\lim_{n\to\infty} P_{ij}^n$ exists. This quantity is independent of i, so we will let $\pi_j = \lim_{n\to\infty} P_{ij}^n$; these are referred to as the *stationary probabilities.* π_j can be determined by finding the stochastic principal left eigenvector of $\mathbf{P}$:

$$\pi = \pi\mathbf{P},$$

where $\pi_1 + \cdots + \pi_k = 1$.

Beginning with a stationary Markov chain with stationary probabilities π_j, we can reverse the process of to get a sequence of states $X_n, X_{n-1}, \ldots, X_0$. This is also a Markov chain with transition probabilities

$$Q_{ij} = P(X_n = j|X_{n+1} = i) = \frac{\pi_j P_{ji}}{\pi_i}.$$

The Markov chain is *time reversible* if $Q_{ij} = P_{ij}$, or, in other words, if

$$\pi_j P_{ji} = \pi_i P_{ij}.$$

The intuition behind this equation is that, in the long run, the probability of seeing a move from i to j is the same as the probability of seeing a move from j to i.

A *continuous-time Markov Chain* is a stochastic process with the property that $X(t+s)$ depends only on $X(s)$, and not on $X(u)$ for $u < s$. The time that a variable spends in any one state must therefore be exponentially distributed. Such a process is defined by two sets of parameters:

1. The mean time spent in state i before making a transition, $1/v_i$.

2. The probability that the process enters state j after leaving state i, denoted by P_{ij}.

The *instantaneous transition rate*, $q_{ij} = v_i P_{ij}$, is the rate at which the process makes a transition into state j, given that it starts in state i. These quantities must satisfy the *Kolmogorov Backward Equations*

$$P'_{ij}(t) = \sum_{k\neq i} q_{ik} P_{kj}(t) - v_i P_{ij}(t),$$

as well as the *Kolmogorov Forward Equations*

$$P'_{ij}(t) = \sum_{k \neq j} q_{kj} P_{ik}(t) - v_j P_{ij}(t).$$

Note the Chapman-Kolmogorov Equation is the following

$$P^t_{ij} = \sum_{k=1} P^{\mu}_{ik} P^{t-\mu}_{kj}$$

The limiting probability is as follow:

$$\sum_{k \neq i} P_{jk} q_{kj} = P_i v_j$$

In other words, the probability of transitioning out is the same as the probability of transitioning in.

1.6.1 Random Walk

A random walk is a Markov Chain in which the states are integers, and there is a fixed probability p such that

$$P_{i,i+1} = p = 1 - P_{i,i-1}.$$

In other words, the state moves up one unit with probability p, and down with probability $1-p$. If $p = \frac{1}{2}$, then the walk is *symmetric.*

1.6.2 Classification of states and classes of states (absorbing, accessible, transition, irreducible, and recurrent)

To determine which states communicate with each other, construct the directed graph associated with the Markov chain. The classes will correspond to the strongly connected components of this graph.

1.6.3 Transition step probabilities

To determine the n-step transition probabilities, compute $\mathbf{P}^n$ and take the i, j entry. This process may be made more efficient by applying the square-and-multiply algorithm if n is large. Alternately, rather than multiplying matrices, left-multiply by a state vector corresponding to the initial state and repeat. (In other words, multiply a vector and a matrix, rather than two matrices.)

An important example involves analysis of system in which there are N states, and from state j, the chain moves to one of the lower $j-1$ states with equal probability. In other words,

$$P_{ij} = \frac{1}{i-1}$$

when $j < i$ and $i > 1$, with $P_{11} = 1$. We will analyze the number of transitions T_i to get from state i to state 1. To analyze this variable, let

$$I_j = \begin{cases} 1 & \text{if the process ever enters } j, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $T_N = \sum_{1 \le j \le N-1} I_j$. Using the fact that $P(I_j = 1) = 1/j$, $E[T_n]$ and $\mathrm{Var}(T_N)$ can be determined.

By conditioning on the first transition, we obtain

$$E[T_i] = 1 + \frac{1}{i-1} \sum_{1 \le j \le i-1} E[T_j].$$

Let N_i be the number of transition until the sate recur. Then $E[N_i] = m_i$. When $m_i = \infty$, it is called null recurrent; if it is finite, then we call it positive recurrent.

1.6.4 Stationary probabilities

To determine the stationary probabilities, solve

$$\pi = \pi \mathbf{P},$$

along with $\pi_1 + \cdots + \pi_k = 1$.

Determining stationary probabilities is the first step in finding a reverse-time Markov chain, since

$$\pi_i P_{ij} = \pi_j Q_{ji}.$$

Note that the left side of this equation has a useful interpretation: it is the long-term probability of seeing a transition from state i to state j.

1.6.5 Recurrent vs. transient states

When identifying recurrent vs. transient states, recall that these are both class properties, so the determination only needs to be made for one vertex in each strongly connected component of the graph.

The matrix

$$\mathbf{S} = (\mathbf{I} - \mathbf{P}_T)^{-1} = \sum_{n\geq 0} \mathbf{P}_T^n$$

can be used to determine the expected time s_{ij} in a transient state j, starting in state i, where $\mathbf{P}_T$ is the submatrix of $\mathbf{P}$ corresponding to the transient states.

The probability f_{ij} that a transition into state j occurs, given that the initial state is i, can be determined by using the conditional formula for s_{ij}:

$$\begin{aligned} s_{ij} &= (\delta_{ij} + s_{jj}) f_{ij} + \delta_{ij}(1 - f_{ij}) \\ &= \delta_{ij} + f_{ij} s_{jj}. \end{aligned}$$

where

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

1.6.6 Gambler's ruin problem

In the Gambler's ruin problem, the gambler wins one dollar with probability p, and loses it with probability $q = 1 - p$. If the gambler quits playing upon reaching state N, this can be modelled as a Markov chain with two absorbing states, 0 and N. To analyze this problem, let P_i denote the probability of reaching N, if the gambler currently has i dollars. Then

$$P_i = pP_{i+1} + qP_{i-1},$$

so

$$P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1}).$$

Applying this recursively, and using the fact that $P_0 = 0$,

$$P_{i+1} - P_i = \left(\frac{q}{p}\right)^i P_1.$$

We can simply this system of equations as a telescoping sum:

$$P_i = \sum_{0\leq j\leq i-1} \left(\frac{q}{p}\right)^j P_1.$$

The analysis now differs depending on whether $q = p = \frac{1}{2}$ or not. In either case, determine P_1 by setting $i = N$ and using the fact that $P_N = 1$.

In short,

$$P_i = \frac{1 - (q/p)^i}{1 - (q/p)^N}$$

1.6.7 Branching Processes

In a *branching process*, we study a population in which each individual produces j offspring with probability P_j. This is a Markov chain in which state i is the number of individuals in generation i. Problems are generally concerned with determining the probability that the population will eventually die out. Let

$$\mu = \sum_{j \geq 0} j P_j.$$

Let π_0 denote the probability that the population will eventually die out, assuming $X_0 = 1$. If $\mu < 1$, then

$$P(X_n \geq 1) = \sum_{j \geq 1} P(X_n = j) \leq E[X_n] = \mu^n,$$

so $P(X_n \geq 1) \leftarrow 0$ as $n \to \infty$.

When $\mu > 1$, the probability that the population dies out can be determined by conditioning on the number of individuals in generation 1. This produces a polynomial for π_0 that can be solved:

$$\pi_0 = \sum_{j \geq 0} P_j \pi_0^j.$$

The idea behind this formula is that if j individuals are born in the first generation, we think of this as j subpopulations that must all die out in order for the population to go to zero.

Let X_n denote the size of the population in the nth generation. Let μ be the mean number of offspring produced by each individual, and let σ^2 denote its variance. The expectation and variance of X_n can be calculated by conditioning on X_{n-1}:

$$\begin{aligned} E[X_n] &= E[E[X_n | X_{n-1}]] \\ &= E[\mu X_{n-1}] \\ &= \mu E[X_{n-1}] \\ &= \mu^n, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(X_n) &= \mathrm{Var}(E[X_n | X_{n-1}]) + E[\mathrm{Var}(X_n | X_{n-1})] \\ &= \mathrm{Var}(\mu X_{n-1}) + E[\sigma^2 X_{n-1}] \\ &= \mu^2 \mathrm{Var}(X_{n-1}) + \sigma^2 \mu^{n-1}. \end{aligned}$$

Applying this repeatedly, we obtain:

$$\begin{aligned}
\mathrm{Var}(X_n) &= \sigma^2\mu^{n-1} + \mu^2\mathrm{Var}(X_{n-1}) \\
&= \sigma^2\mu^{n-1} + \sigma^2\mu^n + \mu^4\mathrm{Var}(X_{n-2}) \\
&= \sigma^2\mu^{n-1} + \sigma^2\mu^n + \sigma^2\mu^{n+1} + \mu^6\mathrm{Var}(X_{n-3}) + \cdots \\
&= \begin{cases} \sigma^2(\frac{\mu^{n-1}-\mu^{2n-1}}{1-\mu}) & \mu \neq 1 \\ n\sigma^2 & \mu = 1 \end{cases}
\end{aligned}$$

1.6.8 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is used to produce a time-reversible Markov chain with specified stationary probabilities, starting from a given Markov chain with transition probabilities Q_{ij}. The process is:

1. Compute
$$\alpha_{ij} = \min\left(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right).$$
2. For $i \neq j$, determine $P_{ij} = Q_{ij}\alpha_{ij}$.
3. Determine P_{ii} such that the rows sum to 1.

1.6.9 Gibbs sampler

The Gibbs sampler is a modification of the Metropolis-Hastings algorithm. It uses a Markov chain in which state changes are defined as follows: (1) select a component of the vector, uniformly at random, then (2) select a new value for that component, uniformly at random, conditional on all the other components. This implies that $\alpha_{ij} = 1$ for all i.

1.6.10 Birth-Death process

In a *birth and death process*, the states are the number of people currently in the system, n. New arrivals enter at an exponential rate λ_n, and leave at an exponential rate μ_n. In this case, the Markov chain parameters are:

- $v_0 = \lambda_0$
- $v_i = \lambda_i + \mu_i$
- $P_{01} = 1$

- $P_{i,i+1} = \frac{\lambda_i}{\lambda_i+\mu_i}$
- $P_{i,i-1} = \frac{\mu_i}{\lambda_i+\mu_i}$

Let T_i denote the amount of time, starting from state i, that it takes for the process to enter state $i+1$. We can determine T_i recursively through the following argument. Clearly, $E[T_0] = \frac{1}{\lambda_0}$. Conditioning on whether the first transition from i is to $i+1$ or $i-1$, we obtain

$$E[T_i] = \frac{\lambda_i}{(\lambda_i+\mu_i)^2} + \frac{\mu_i}{(\lambda_i+\mu_i)^2} + \frac{E[T_{i-1}]+E[T_i]}{\lambda_i+\mu_i}$$

This can be simplified algebraically to

$$E[T_i] = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i}E[T_{i-1}].$$

Note that the expected time to enter state j for the first time can be determined by summing T_i for $i < j$.

We use $P_{ij}(t)$ to denote the probability that the Markov chain $X(t)$ is in state j at time t, provided that it was in state i at time 0. For a pure birth process, $P_{i,i+1}(t)$ can be determined as follows. Note that $X(t) < i+1$ if and only if $T_i > t$, so

$$\begin{aligned} P(X(t) < i+1 | X(0) = i) &= P(T_i > t) \\ &= e^{-\lambda_i t}, \end{aligned}$$

and

$$\begin{aligned} P(X(t) < i+2 | X(0) = i) &= P(T_i + T_{i+1} > t) \\ &= \frac{\lambda_{i+1}}{\lambda_{i+1}-\lambda_i}e^{-\lambda_i t} + \frac{\lambda_i}{\lambda_i - \lambda_{i+1}}e^{-\lambda_{i+1} t}. \end{aligned}$$

Therefore,

$$P(X(t) = i : X(0) = i) = \frac{\lambda_i}{\lambda_{i+1}-\lambda_i}(e^{-\lambda_i t} - e^{-\lambda_{i+1} t}).$$

1.6.11 Homogeneous transition probabilities

1.6.12 Memoryless property of Markov Chains

1.6.13 Limiting probabilities

Limiting probabilities for continuous-time Markov Chains can be derived by taking the limit of the Kolmogorov Forward Equation, and observing that $\lim_{t\to\infty} P_{ij}(t) = 0$, to obtain

$$v_i P_j = \sum_{k \neq j} q_{kj} P_k$$

for all states j. Along with the fact that $\sum_j P_j = 1$, these equations can be used to determine P_i. Note that the left side can be interpreted as the long-run rate of leaving state j, and the right side can be interpreted as the long-run rate of entering state j. For birth and death processes, these equations simplify to

$$\begin{aligned} \lambda_0 P_0 &= \mu_1 P_1 \\ \mu_j P_j + \lambda_j P_j &= \mu_{j+1} P_{j+1} + \lambda_{j-1} P_{j-1}. \end{aligned}$$

Substituting the first equation into the second gives

$$P_{j+1} = \frac{\lambda_j}{\mu_{j+1}} P_j,$$

which can be used to determine P_j, along with the restriction $\sum_{j \geq 0} P_j = 1$.

1.7 Life Contingencies

Learning Objectives: Solve Life Contingency problems using a life table in a spreadsheet as the combined result of discount, probability of payment, and amount of payment vectors. Understand the linkage between the life table and the corresponding probability models.

- Calculate annuities for discrete time
- Calculate life insurance single net premiums (or property/casualty pure premiums) for discrete time
- Solve for net level premiums (not including fractional lives)

Life contingency problems involve tables containing the following information:

- x, representing the year / age. The notation (x) refers to an individual whose exact age is x.
- ℓ_x, the number of individuals alive at the beginning of year x
- $d_x = \ell_x - \ell_{x+1}$, the number of individuals who die in year x

- $q_x = d_x/\ell_x$, known as the *mortality.* This is the probability that a person of age x dies before reaching age $x+1$.
- $p_x = \ell_{x+1}/\ell_x$ is the probability that a person of age x survives to age $x+1$.
- The notation ${}_np_x$ is the probability that a person of age x survives to age $x+n$. Note that
$$ {}_np_x = p_x p_{x+1} \cdots p_{x+n} = \frac{\ell_{x+n}}{\ell_x}. $$

In life contingency problems, we are typically concerned with calculating the actuarial present value of a future payment that is contingent on the death of an individual. In *term life insurance*, the length of the contract is fixed, and a payment is made at the end of the year in which the individual dies, provided it is within the term of the contract. Life insurance contracts with unlimited term are called *whole life* policies. An *endowment insurance* policy is similar to term life insurance, except that it pays \$1 at expiration of the term. A *pure endowment policy* is a policy that pays \$1 if the insured survives to the end of the term; it is the difference between an endowment policy and a term policy. In the sample tables, the value of a pure endowment of duration n for an individual of age x is denoted by ${}_nE_x$. These types of policies are related: a term insurance policy is equivalent to the difference between a whole life policy and a deferred whole life policy for someone of an older age:

$$ A_x - {}_nE_x A_{x+n}. $$

Let A_x denote the actuarial present value of a whole life insurance policy which pays \$1 at the end of the year in which an individual of age x dies. Let $v = 1/(1+i)$ be the present value of \$1 payable in one year. Then

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

by conditioning on whether the individual lives or dies, and by noting that if the individual lives, then the value of the policy in one year is A_{x+1}. Applying this formula recursively, we obtain

$$ A_x = \sum_{n\geq 0} v^{n+1} \, {}_np_x q_{x+n} = \sum_{n\geq 0} v^{n+1} \frac{d_{x+n}}{\ell_x}. $$

A similar argument can be used to determine the *curtate life expectancy*, which is the expected number of whole years of life remaining for an individual of age x. Denote this by e_x. Then

$$ e_x = p_x(1 + e_{x+1}). $$

The premiums are assumed to be paid at the beginning of the year, so they form an annuity-due and their actuarial present value is denoted by by $\ddot{a}_x$. If the payments are made at the end of the year, they form an annuity-immediate, and their actuarial present value is denoted by a_x. Life annuities differ from ordinary annuities in that their duration is not known – payments cease when an individual dies. Therefore,

$$\ddot{a}_x = 1 + vp_x\ddot{a}_{x+1}.$$

Applying this formula recursively,

$$\ddot{a}_x = \sum_{n\geq 0} v^n \;{}_np_x.$$

1.7.1 Discounted cash flow

The term *actuarial present value* refers to the expected present value of a stream of cash flows that depends on the value of a random variable. We assume that yield curves are level for simplicity.

1.7.2 Relationship between annuity values and insurance premiums

Insurance premiums can be determined through an arbitrage argument: an investor is ambivalent between receiving $1 today, and receiving interest on that dollar annually until the insured dies, receiving $1 at the end of that year. Therefore,

$$1 = d\ddot{a}_x + A_x,$$

where $d = 1 - v$ is the rate of discount. This can be solved to determine $\ddot{a}_x$. From this, the premium P can be determined by requiring that

$$A_x = P\ddot{a}_x.$$

Premiums for more complicated contracts can be determined by constructing a replicating portfolio, for example, an n-year annuity-certain and a life annuity, deliverable in n years, replicates a workers' compensation policy that pays for a minimum of n years, until the insured returns to work.

1.7.3 Life table linkage to probability models

The life tables can be used to determine the probability that an individual of age x dies in the given year (q_x) and the probability that they survive (p_x). Let X_i be an indicator variable equal to 1 if and only if individual i survives. Since this is an indicator random variable, then

$$E[X_i] = p_x,$$

and

$$\mathrm{Var}(X_i) = E[X_i^2] - E[X_i]^2 = p_x - p_x^2 = p_x q_x.$$

In the sample tables, ${}^2A_x$ is the expectation of the square of the present value, and it can be used to determine the variance of the value of the life insurance policy.

1.7.4 Equivalence property

2 Statistics

2.1 Parameter Estimation

Learning Objectives: Perform point estimation of statistical parameters using Maximum likelihood estimation (”MLE”). Apply criteria to the estimates such as:

- Consistency
- Unbiasedness
- Sufficiency
- Efficiency
- Minimum variance
- Mean square error

2.1.1 Background

Suppose we have a n samples, $X_1, \ldots, X_n$, from a collection of independent and identically distributed random variables whose probability density function $f(x;\theta)$ depends on an unknown parameter $\theta \in \Omega$. The *likelihood function* is

$$L(\theta; \mathbf{x}) = \prod_{1 \le i \le n} f(x_i; \theta).$$

In many applications, it is more useful to work with the *log-likelihood function,*

$$\ell(\theta) = \log(L(\theta)) = \sum_{1 \le i \le n} \log(f(x_i; \theta)).$$

The *maximum likelihood estimator* $\hat{\theta}$ the function of $x_1, \ldots, x_n$ that maximizes $L(\theta)$, or equivalently, $\ell(\theta)$. We will use the notation θ_0 to denote the true value of θ.

The quantity $\frac{\partial \log f(x;\theta)}{\partial \theta}$ is called the *score function* – the estimating equations for the MLE are sums of score functions. Define the *Fisher information* to be the expectation of the square of the score function, namely,

$$I(\theta) = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right]$$

It can be shown that the expectation of the score function is zero, which means that Fisher information is the variance of the score function.

The quality of statistics that are used to estimate unknown parameters is evaluated using a variety of criteria:

- **Consistency**: A statistic T_n is a consistent estimator of θ if T_n converges in probability to θ; that is, $\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) = 1$ for any $\epsilon > 0$.

- **Unbiased**: If T is a statistic based on the values of n independent and identically distributed random variables, then we say that T is an unbiased estimator of θ if $E[T] = \theta$. If an estimator is biased, we define $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ to be the *bias*.

- **Mean Square Error**: For an estimator $\hat{\theta}$, define

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + B(\hat{\theta})^2.$$

- **Efficiency**: an unbiased estimator Y is an *efficient estimator* if and only if $\text{Var}(Y) = \frac{1}{nI(\theta)}$; in other words, if the Rao-Cramer bound is achieved. The ratio of the Rao-Cramer bound to the actual variance of an unbiased estimator is called the *efficiency* of the estimator.

- **Sufficiency**: a statistic $Y = u(X_1, \ldots, X_n)$ is *sufficient* if the probability of observing $X_1, \ldots, X_n$, given that

$$(X_1, \ldots, X_n) \in \{(x_1, \ldots, x_n) : u(x_1, \ldots, x_n) = y\}$$

does not depend on θ. Intuitively, Y exhausts all the information about θ, and given $Y = y$, no other statistic can draw any inference about θ. If f_Y is the density function of Y, then Y is a sufficient statistic for θ if and only if the ratio

$$\frac{f(x_1;\theta)\cdots f(x_n;\theta)}{f_Y(u(x_1,\ldots,x_n);\theta)}$$

does not depend on θ.

A statistic Y is called a *minimum variance unbiased estimator (MVUE)* of θ if it is unbiased and the variance of Y is less than or equal to the variance of every other unbiased estimator of θ.

Theorems about the MLE are stated in terms of several regularity conditions:

(R0) If $\theta \neq \theta'$ then $f(x_i;\theta) \neq f(x_i;\theta')$.

(R1) The density functions have common support for all θ.

(R2) The point θ_0 is an interior point of Ω.

(R3) $f(x;\theta)$ is twice-differentiable as a function of θ.

(R4) $\int f(x;\theta)dx$ can be differentiated twice, under the integral sign, as a function of θ.

Facts about the MLE include:

1. If (R0) and (R1) hold, then the likelihood function is asymptotically maximized at θ_0.
2. If $\eta = g(\theta)$ is a parameter, then $g(\hat{\theta})$ is the minimum likelihood estimator of η.
3. If regularity conditions (R0) through (R4) hold, and Y is an estimator for θ based on n sample points, with $E[Y] = k(\theta)$, then

$$\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

This is called the *Rao-Cramer bound.* In particular, if Y is an unbiased estimator of θ, then $\mathrm{Var}(Y) \geq \frac{1}{nI(\theta)}$.

2.1.2 Equations for MLE of mean, variance from a sample

2.1.3 Estimation of mean and variance based on a sample

The *sample mean* is defined by

$$\overline{X} := \frac{1}{n} \sum_{1 \leq i \leq n} X_i.$$

Note that $E[\overline{X}] = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$. The *sample variance* is

$$S^2 = \frac{1}{n-1} \sum_{1 \leq i \leq n} (X_i - \overline{X})^2 = \frac{1}{n-1} \left(\sum_{1 \leq i \leq n} X_i^2 - n\overline{X}^2 \right),$$

and $E[S^2] = \sigma^2$.

2.1.4 General equations for MLE of parameters

To find a MLE, solve the equation

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0.$$

One application of MLE estimation is to determine the parameter θ for exponentially distributed claims, when observed losses are censored. Suppose that claims are capped at C. Then the probability distribution of a claim X is a hybrid discrete/continuous distribution. With probability $e^{-C/\theta}$, $X = C$. Otherwise, the distribution is continuous with density function $\frac{1}{\theta}e^{-x/\theta}$ for $0 \leq x \leq C$. Using this function allows us to determine the MLE for θ — it should be equal to the total claim amount, divided by the number of claims that were not subject to a cap.

2.1.5 Recognition of consistency property of estimators and alternative measures of consistency

Some methods of identifying consistent estimators include:

- The sample mean and sample variance converge in probability to μ and σ, respectively. Consequently, any continuous function of these estimators will also be consistent.
- If $\hat{\theta}$ is unbiased and $\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$, then $\hat{\theta}$ is consistent.
- If regularity conditions (R0), (R1), and (R2) are satisfied, then the maximum likelihood estimator is consistent.

2.1.6 Application of criteria for measurement when estimating parameters through minimization of variance, mean square error

2.1.7 Definition of statistical bias and recognition of estimators that are unbiased or biased

If $E[T] = \lambda\theta$ for value $\lambda \neq 1$, then the biased estimator T can be converted into an unbiased estimator for θ, $T' = \frac{1}{\lambda}T$.

2.1.8 Application of Rao-Cramer Lower Bound and Efficiency

To calculate efficiency of an unbiased estimator, first calculate

$$I(\theta) = -E\left[\frac{\partial^2 f(x;\theta)}{\partial \theta^2}\right].$$

Efficiency is equal to

$$\frac{1}{nI(\theta)\mathrm{Var}(Y)}.$$

If this equal to 1, then Y is an efficient estimator.

2.1.9 Relationship between Sufficiency and Minimum Variance

The *Rao-Blackwell Theorem* tells us that if Y_1 is a sufficient statistic for θ, and Y_2 is an unbiased estimator of θ that is not a function of Y_1 alone, then the function $\phi(y_1) = E(Y_2|y_1)$ is an unbiased estimator of θ whose variance is less than or equal to that of Y_2. In other words, when searching for a MVUE, we can limit our search to functions of Y_1 alone.

For a complete family of density functions, if Y is a sufficient statistic for θ, then a function of Y that is an unbiased estimator of θ is the unique MVUE of θ.

2.1.10 Develop and estimate a sufficient statistic for a distribution

A sufficient statistic can often be determined from inspection: look at $f(x_1;\theta)\cdots f(x_n;\theta)$, and select Y so that it can be written as $k_1(Y,\theta)k_2(x_1,\ldots,x_n)$.

2.1.11 Factorization Criterion for sufficiency

When determining the density function for a statistic that is a linear combination of the observations, it may be useful to multiply the moment generating functions.

The need to find the density function of the estimator can be avoided by using the *factorization theorem*: Y is a sufficient statistic for θ if and only if there exist two non-negative functions, k_1 and k_2, such that

$$f(x_1;\theta)\cdots f(x_n;\theta) = k_1(Y;\theta)k_2(x_1,\ldots,x_n),$$

where $k_2(x_1,\ldots,x_n)$ does not depend on θ. (Be careful to check that the domain of k_2 does not depend on θ.)

2.1.12 Application of Rao-Cramer Lower Bound and Fisher Information

Fisher information can be used to determine the asymptotic variance of the maximum likelihood estimator, as $\frac{1}{nI(\theta)}$.

2.1.13 Application of MVUE for the exponential class of distributions

The *regular exponential class* of probability density functions has a density / mass function of the form

$$f(x;\theta) = \begin{cases} \exp(p(\theta)K(x) + H(x) + q(\theta)) & \text{if } x \in \mathcal{S} \\ 0 & \text{otherwise,} \end{cases}$$

where

- $\gamma < \theta < \delta$
- $\mathcal{S}$ does not depend on θ
- $p(\theta)$ is a nontrivial continuous function
- If X is continuous, then $K'(x) \not\equiv 0$ and $H(x)$ is continuous. If X is discrete, then $K(x)$ is a nontrivial function of $x \in \mathcal{S}$.

Key facts about this class include:

- $Y = \sum_{1 \le i \le n} K(x_i)$ is a complete sufficient statistic for θ.
- The density function for Y is of the form

  $$R(y)\exp(p(\theta)y + nq(\theta)),$$

  where $R(y)$ is some function that does not depend on θ.
- $E[Y] = -n\frac{p'(\theta)}{q'(\theta)}$.
- $\mathrm{Var}(Y) = \frac{n}{p'(\theta)^3}(p''(\theta)q'(\theta) - q''(\theta)p'(\theta))$

2.2 Hypothesis Testing

Learning Objectives:

- Test statistical hypotheses including Type I and Type II errors using:
  - Neyman-Pearson lemma
  - Likelihood ratio tests
  - First principles
- Apply Neyman-Pearson lemma to construct likelihood ratio equation
- Use critical values from a sampling distribution to test means and variances.

2.2.1 Background

Let X be a random variable with density function $f(x;\theta)$, where $\theta \in \Omega$. Given a partition (ω_0, ω_1) of Ω, we consider two hypotheses: the *null hypothesis*, denoted by H_0, in which $\theta \in \omega_0$, and the *alternative hypothesis*, denoted by H_1, in which $\theta \in \omega_1$. We test the hypotheses by taking a random sample $X_1, \ldots, X_n$ by specifying a subset C of the sample space, called the *critical region*, with the following decision rule:

- Reject H_0 (accept H_1) if $(X_1, \ldots, X_n) \in C$
- Retain H_0 (reject H_1) if $(X_1, \ldots, X_n) \notin C$.

A typical strategy is to select critical regions that bound the probability of Type I error, and among these regions, select the one that minimizes the probability of a Type II error. The *size* of a critical region (also called the *significance* of the test) is defined to be

$$\alpha := \max_{\theta \in \omega_0} P_\theta((X_1, \ldots, X_n) \in C).$$

Note that $P_\theta((X_1, \ldots, X_n) \in C)$ is the probability of not making a Type II error. This is called the *power* of the test at θ. The *power function* of a critical region is

$$\gamma_C(\theta) = P_\theta((X_1, \ldots, X_n) \in C))$$

for $\theta \in \omega_1$.

Theorem 2.1. *Let $X_i, i = 1, \cdots, n$ be a random sample from a normal distribution with mean μ and variance σ^2. Let $Q = \sum_{i=1}^{n}(X_i - \mu)^2$. Then Q/σ^2 has a $\chi^2(n)$ distribution.*

Theorem 2.2. *Let $X_i, i = 1, \cdots, n, n \geq 2$ be a random sample from a normal distribution with variance σ^2. Let $\bar{X}$ be the sample mean. Let $W = \sum_{i=1}^{n}(X_i - \bar{X})^2$. Then W/σ^2 has a $\chi^2(n-1)$ distribution. Moreover, W and $\bar{X}$ are independent.*

In a *likelihood ratio test*, to test the hypothesis $\theta = \theta_0$, we use the test statistic

$$\Lambda = \frac{L(\theta_0)}{L(\hat{\theta})},$$

where L is the likelihood function and $\hat{\theta}$ is the maximum likelihood estimator. We use decision rules of the form "Reject H_0 in favour of H_1 if $\Lambda < c$." When the underlying distribution is normal, $-2\log\Lambda$ is $\chi^2(1)$, and in general, for any distribution, $-2\log\Lambda$ is asymptotically $\chi^2(1)$. The Neyman-Pearson Theorem states that a critical region

C defined by $\Lambda \leq k$ will be a best critical region of size $\alpha = P_{H_0}(x \in C)$ when H_0 and H_1 are both simple. (Compound H_1 can be handled by considering each simple sub-hypothesis; if they all provide a consistent result, then we have a uniformly most powerful critical region.)

More generally, the likelihood ratio test can be used to compare distributional assumptions: the numerator is the likelihood function under a null hypothesis that the density function is f, and the denominator is the likelihood function under the alternative hypothesis that the distribution is a generalization of f. **Sample problem**: Exam 4-21

Similar tests include:

- In a *Wald-type test*, the test statistic is
$$\chi^2_W = nI(\hat{\theta})(\hat{\theta} - \theta_0)^2,$$
which is asymptotically $\chi^2(1)$.
- In *Rao's score test*, the test statistic is
$$\chi^2_R = \frac{\ell'(\theta_0)^2}{nI(\theta - 0)},$$
which is also asymptotically $\chi^2(1)$.

When testing a simple hypothesis $H_0 : \theta = \theta'$ against $H_1 : \theta = \theta''$, a *best critical region* of size α is a critical region such that $P_{\theta'}(X \in C) = \alpha$, and for every subset A of the sample space such that $P_{\theta'}(X \in A)$, we have $P_{\theta''}(X \in C) \geq P_{\theta''}(X \in A)$. The Neyman-Pearson Theorem may be used to construct best critical regions as follows: $x \in C$ if and only if
$$\frac{L(\theta'; x)}{L(\theta''; x)} \leq k,$$
where k is selected such that $P_{\theta'}(X \in C) = \alpha$. A critical region is *uniformly most powerful* for testing a simple hypothesis H_0 against a composite hypothesis H_1 if it is a best critical region for every simple hypothesis in H_1.

2.2.2 Presentation of fundamental inequalities based on general assumptions and normal assumptions

Given a random variable X with mean μ and finite variance σ^2, suppose we are testing the hypothesis
$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0.$$

Without a distributional assumption, we can use the Central Limit Theorem to conclude that $\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim N(0,1)$. We will develop a test in which we accept H_0 if and only if $\overline{X} < L$, for some value of L to be determined. To develop a test with significance α, we want to find a value L such that

$$P(\overline{X} \geq L) = \alpha.$$

This is equivalent to

$$P(\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \geq \frac{L-\mu}{S/\sqrt{n}}) = \alpha,$$

so we want

$$\frac{L-\mu}{S/\sqrt{n}} = \Phi^{-1}(1-\alpha) = z_\alpha$$

In other words, we accept H_0 if and only if

$$\frac{\overline{X}-\mu_0}{S/\sqrt{n}} < z_\alpha.$$

The power function for this test is given by

$$\begin{aligned} \gamma_(\mu) &= P(\overline{X} > \mu_0 + \sigma z_\alpha/\sqrt{n}) \\ &= P(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > (\mu_0-\mu)\sqrt{n}/\sigma + z_\alpha) \end{aligned}$$

2.2.3 Definition of Type I and Type II errors

Type I error: occurs when we decide $\theta \in \omega_1$ when in fact $\theta \in \omega_0$. In other words, we reject the null hypotheses when it is correct. To determine the probability of a Type I error, assume H_0 is true, then calculate the probability that the test rejects H_0.

Type II error: occurs when we decide $\theta \in \omega_0$ when in fact $\theta \in \omega_1$. In other words, we accept the null hypothesis when it is incorrect.

2.2.4 Significance levels

An *observed significance level*, also called a p-value, is the probability of observing a statistic at least as extreme as the actual value that was observed.

2.2.5 One-sided versus two-sided tests

In a one-sided test, H_1 is of the form $\mu > \mu_0$ or $\mu < \mu_0$. In a two-sided test, H_1 is of the form $\mu \neq \mu_0$. In the two-sided case, we use a rule of the form "Reject H_0 in favour of H_1 if either $\overline{X} \leq h$ or $\overline{X} \geq k$." Then the probability of a Type 1 error is

$$P(\overline{X} \leq h) + P(\overline{X} \geq k).$$

There is flexibility in terms of how much of α to assign to each term in this expression, but typically $\alpha/2$ is used for each, especially if the distribution of X is symmetric.

2.2.6 Definition and measurement of likelihood ratio tests

1. Calculate the maximum likelihood as a function of the observations under H_0. If H_0 is simple, there is only one likelihood for each test statistic.

2. Calculate the maximum likelihood as a function of the observations under H_1. If H_1 is composite, this means finding the simple sub-hypothesis which maximizes the likelihood.

3. The critical region is the set of observations for which the ratio of the first expression over the second expression is below a constant k.

2.2.7 Recognizing when to apply likelihood ratio tests versus chi-square or other goodness of fit tests

In a goodness-of-fit test, we want to test whether a given discrete distribution is appropriate. Specifically, if f is the probability mass function that we want to test, H_0 is $f(i) = p_{i0}$ for $1 \leq i \leq k$. When H_0 is true, the random variable

$$Q_{k-1} = \sum_{1 \leq i \leq k} \frac{(X_i - np_{i0})^2}{np_{i0}}$$

is $\chi^2(k-1)$.

2.2.8 Test for difference in variance under Normal distribution between two samples through application of F-test

Suppose we have two sets of samples, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, from normal distributions and we want to test $H_0 : \sigma_X = \sigma_Y$ against $H_1 : \sigma_X \neq \sigma_Y$. The likelihood ratio is a function of the statistic

$$F = \frac{(m-1)\sum_{1\leq i\leq n}(X_i - \overline{X})^2}{(n-1)\sum_{1\leq j\leq m}(Y_j - \overline{Y})^2}.$$

This has an F-distirbution with $(n-1)$ and $(m-1)$ degrees of freedom, so we can select constants c_1 and c_2 such that

$$P(F \leq c_1) = P(F \geq c_2) = \frac{\alpha}{2}.$$

2.2.9 Test for significance of means under Normal distribution assumption in both large and small sample cases

Given random variables X and Y that are normally distributed with the same variance, when testing the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$:

- Calculate $\overline{X}$, $\overline{Y}$, and the *pooled estimator* of the variance:

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

- Use the t-distribution with $n-2$ degrees of freedom to construct a confidence interval around $\overline{X} - \overline{Y}$. (Remember that instead of using $\sqrt{1/n}$, $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is used.)

- Accept H_0 if and only if 0 lies in the confidence interval.

2.2.10 Test for significance of difference in proportions between two samples under Binomial distribution assumption in both large and small sample case

Testing a binomial distribution proceeds similarly to testing under the assumption of normality. In this case, the estimate of the variance can either be $\hat{p}(1-\hat{p})$ or $p_0(1-p_0)$ (if H_0 is of the form $p = p_0$).

For small sample sizes, use the fact that

$$Y = \frac{X - np}{\sqrt{np(1-p)}}$$

is such that $Y^2 \sim \chi^2(1)$.

2.2.11 Application of contingency tables to test independence between effects

Suppose we are observing two discrete variables, the first of which has a levels and the second of which has b levels. Let p_{ij} denote the probability that an observation is at level i in the first variable, and level j in the second; let X_{ij} denote the corresponding observed frequencies.

$$Q_{ab-1} = \sum_{1 \le j \le b} \sum_{1 \le i \le a} \frac{(X_{ij} - np_{ij})^2}{np_{ij}}$$

is approximately $\chi^2(ab-1)$ when n is large. To test independence, we are testing the hypothesis $p_{ij} = p_i p_j$. If we replace p_i and p_j with the observed frequencies, then this variable becomes $\chi^2((a-1)(b-1))$.

2.2.12 Application of Neyman-Pearson lemma to Uniformly Most Powerful hypothesis tests

Lemma 2.1. *The Neyman-Pearson lemma states that for tests of one simple hypothesis against another, a best critical region exists for each significance level and that selecting all points below a fixed likelihood ratio determines the region.*

2.2.13 Equivalence between critical regions and confidence intervals

A two-sided test can be rephrased as: "Accept H_0 if and only if μ_0 is in a $(1-\alpha)$ confidence interval around $\overline{X}$.'

2.3 Non-parametric Statistics

Learning Objectives: Calculate the order statistics of a sample for a given distribution and use non-parametric statistics to describe a data set.

2.3.1 Background

Given a sample $X_1, \ldots, X_n$ of a continuous random variable X with support $[a, b]$, the *order statistics* are $Y_1 < Y_2 < \ldots < Y_n$, where Y_i is the ith smallest element of $\{X_1, \ldots, X_n\}$. If $f(x)$ is the density function of X, then the density function of the order statistics is

$$g(y_1, \ldots, y_n) = \begin{cases} n! \prod_{1 \le i \le n} f(y_i) & \text{if } a < y_1 < y_2 < \ldots y_n < b, \\ 0 & \text{otherwise.} \end{cases}$$

The density functions of individual order statistics can be obtained by integrating to obtain a marginal density function. This process can be simplified by using the fact that

$$\int_a^x F(w)^{\alpha-1} f(w)dw = \frac{1}{\alpha}F(x)^\alpha,$$

and

$$\int_y^b (1-F(w))^{\beta-1} f(w)dw = \frac{1}{\beta}(1-F(w))^\beta,$$

where F is the cumulative distribution function of X. In particular, the density function of the kth order statistic is:

$$g_k(y_k) = \begin{cases} \frac{n!}{(k-1)!(n-k)!}F(y_k)^{k-1}(1-F(y_k))^{n-k}f(y_k) & \text{if } a < y_k < b, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < p < 1$, the pth *quantile* of X is $\xi_p = F^{-1}(p)$. The *median* is the 0.5 quantile. It can be shown that

$$E(F(Y_k)) = \frac{k}{n+1} \approx p,$$

where $k = \lfloor p(n+1) \rfloor$. Therefore, Y_k is a reasonable estimator of ξ_p, called the *pth sample quantile*, or the $100p$th *percentile of the sample.*

For a random variable Z with known cumulative distribution function $F(z)$, let $X = bZ + a$, where a and b are parameters to be estimated — then the distribution function of X is of the form $F((x-a)/b)$. In this case, if $\xi_{X,p}$ and $\xi_{Z,p}$ are the quantiles of X and Z, respectively, then $\xi_{X,p} = b\xi_{Z,p} + a$. If $Y_1 < \ldots < Y_n$ are the order statistics of a random sample of X, and $p_k = k/(n+1)$, then a plot of Y_k versus ξ_{Z,p_k} is called a *q-q plot.* If the plot is linear, then in indicates that the the distribution function of X is of the form $F((x-a)/b)$.

Given a random sample $X_1, \ldots, X_n$ where $X_i = \theta + \epsilon_i$, where ϵ_i are independent and identically distributed with median 0, define the *sign statistic* $S(\theta_0)$ to be the number of observations greater than θ_0. This can be used to test a null hypothesis of the form $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$. The test is:

$$\text{Reject } H_0 \text{ if } S(\theta_0) \geq c,$$

where c is selected to achieve a desired significance level. Note that when the null hypothesis is true, $S(\theta_0)$ has a binomial distribution with $p = \frac{1}{2}$. For large n, the normal approximation may be used. Confidence intervals for the median θ can be determined by selecting c_1 such that $P(S(\theta) \leq c_1) = \frac{\alpha}{2}$. (This may be determined either using the binomial distribution or normal approximation). Then $[Y_{c_1+1}, Y_{n-c_1})$ is a $(1-\alpha)$-confidence interval.

If we assume that the density function for ϵ_i is even, then we can use a *signed-rank Wilcoxon* test. Without loss of generality, we work with a null hypothesis of the form $H_0 : \theta = 0$, against $H_1 : \theta > 0$. (If the null hypothesis is $\theta = \theta_0$, subtract θ_0 from all observations.) Define

$$T = \sum_{1 \leq i \leq n} \operatorname{sgn}(X_i) R|X_i|,$$

where $R|X_i|$ is the rank of X_i among $|X_1|, \ldots, |X_n|$. The decision rule is of the form

$$\text{Reject } H_0 \text{ if } T \geq c.$$

By re-ordering the summation, we can determine the moment generating function for T:

$$\begin{aligned} E[\exp(sT)] &= \prod_{1 \leq j \leq n} E[\exp(sj\operatorname{sgn}(X_{i_j})] \\ &= \prod_{1 \leq j \leq n} \left(\frac{1}{2} e^{-sj} + \frac{1}{2} e^{sj} \right) \\ &= \frac{1}{2^n} \prod_{1 \leq j \leq n} (e^{-sj} + e^{sj}). \end{aligned}$$

Note that T has mean zero, and variance $\sum_{1 \leq i \leq n} i^2 = \frac{n(n+1)(2n+1)}{6}$ (since $\operatorname{Var}(\operatorname{sgn}(X)) = 1$).

The *Mann-Whitney-Wilcoxon Procedure* involves two random samples, $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, with cumulative distribution functions $F(x)$ and $G(x)$, respectively. We assume that $G(x) = F(x - \Delta)$, and consider a null hypothesis $H_0 : \Delta = 0$ against $H_1 : \Delta > 0$. Rank the combined sample of X_i and Y_i, and calculate

$$W = \sum_{1 \leq j \leq n_2} R(Y_j).$$

The decision rule is of the form

$$\text{Reject } H_0 \text{ if } W \geq c.$$

Under the null hypothesis, the ranks are uniformly distributed on $\{1, 2, \ldots, n\}$, so

$$E(W) = \frac{n_2(n+1)}{2}$$

and

$$\mathrm{Var}(W) = \frac{n_1 n_2(n+1)}{12}.$$

The test statistic can also be written as

$$W = U + \frac{n_2(n_2+1)}{2},$$

where

$$U = |\{(i,j) : Y_j > X_i\}|.$$

Given two pairs of samples, (X_1, Y_1) and (X_2, Y_2), we say that the pairs are *concordant* if $\mathrm{sgn}(X_1 - X_2)(Y_1 - Y_2) = 1$, and *discordant* if the sign is -1. *Kendall's* τ is

$$\tau = P(\text{concordant}) - P(\text{discordant}).$$

If $\tau = 0$, then X and Y are independent. We can develop a test for the hypothesis $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$ by using the test statistic:

$$K = \binom{n}{2}^{-1}(c - d),$$

where c is the number of concordant pairs in the sample, and d is the number of discordant pairs. *Spearman's* ρ is another measure of correlation that is obtained by replacing observations with their ranks in the usual correlation coefficient formula:

$$\rho = \frac{\sum_{1 \le i \le n}(R(X_i) - \frac{n+1}{2})(R(Y_i) - \frac{n+1}{2})}{n(n^2-1)/12}.$$

If X and Y are independent, then $E[\rho] = 0$.

Summary of Non-parametric Statistical Tests

In the following, the null hypothesis is that X and Y have the same distribution; in particular, they have the same median and $P(X > Y) = \frac{1}{2}$.

Name of test	Context	Test statistic	Determining p-value
Sign test	Matched pairs of samples (X_i, Y_i). Taking Y_i equal to the median is a special case.	Number of i such that $X_i > Y_i$. Remove ties from sample.	Binomial distribution or normal approximation
Wilcoxon Signed-Rank "T" Test	Matched pairs of samples (X_i, Y_i)	Remove ties and rank $\lvert X_i - Y_i \rvert$. Two-sided test statistic T is the smaller of the sum of positive ranks and the sum of negative ranks. Reject H_0 when T is small.	Tables or normal approximation, $\mu = \frac{n(n+1)}{4}$, $\sigma^2 = \mu\frac{2n+1}{6}$
Wilcoxon Rank-Sum "W" Test	Independent (non-paired) samples of size n_1 and n_2	Combine samples and rank. Sum the ranks of the second sample.	Normal approximation, $\mu = \frac{n_2(n_1+n_2+1)}{2}$, $\sigma^2 = \mu\frac{n_1}{6}$.
Mann-Whitney "U" Test	Independent (non-paired) samples of size n_1 and n_2	Combine and order the samples. For each point from the second sample, count the number of first-sample points that precede it, and sum the results to get U.	Tables provided. Reject if U is very large (population 1 too small) or very small (population 1 too large). Normal approximation uses $\mu = \frac{n_1 n_2}{2}$, $\sigma^2 = \mu\frac{n_1+n_2+1}{6}$.

2.3.2 Application to a given distributional form

If the density and cumulative distribution function are known, then the density function

$$\frac{n!}{(k-1)!(n-k)!}[F(y)]^{k-1}[1-F(y)]^{n-k}f(y)$$

can be used to answer questions about the kth order statistic in a sample of n.

Confidence intervals for order statistics can be determined by using the binomial distribution, and the fact that each observation is less than the pth quantile with probability p. In this case, $P(Y_i < \xi_p < Y_j)$ is equal to the probability that at least i observations, and at most $j-1$, fall below the pth quantile. A similar approach can be used to calculate $P(Y_i < t)$, using $p = F(t)$.

2.3.3 Calculate Spearman's Rho and Kendall's Tau and understand how those correlation measures differ from the Pearson correlation coefficient

An easy way to calculate Kendall's Tau is to convert all observations into their rank, sort the x_i values in ascending order, and use a table indicating how many concordant and discordant pairs are accounted for by each row. In order to count concordant / discordant pairs, it suffices to look at the values in the y column below the current row.

Spearman's ρ

2.3.4 Apply rank order statistics using Sign-Rank Wilcoxon for matched pair tests

To apply the Signed-Rank Wilcoxon test to the hypothesis $H_0 : \theta = \theta_0$:

1. Subtract θ_0 from each observation.
2. Compute $T = \sum_{1\leq i\leq n} \operatorname{sgn}(X_i)R|X_i|$.
3. Determine the variance of T, $\frac{n(n+1)(2n+1)}{6}$.
4. The variable $T/\sqrt{\operatorname{Var}(T)}$ is approximately normally distributed. Use this to determine the probability that the T value is greater than / less than the observed.

2.3.5 Apply rank order statistics using Sign Test for matched pairs and comparison of distributions with or without Normal approximation

To determine a confidence interval for the median, the key idea is that the probability that the median is between Y_i and Y_{i+1} is equal to the probability that i samples are below the median, and that the probability of a sample being below the median is 0.5.

1. Find a confidence interval around the mean of $S(\theta)$. This can be done either using the binomial distribution or normal approximation. Expand to an integral interval, $[a, b]$.

2. Convert this to a confidence interval for the median, $[Y_a, Y_{b+1}]$.

Given a sample of n pairs (x_i, y_i), we can test the hypothesis that x_i and y_i have the same median by testing that $d_i := x_i - y_i$ has a median of zero.

2.3.6 Apply rank order statistics using Mann-Whitney-Wilcoxon procedure

To test a hypothesis that two samples come from the same distribution using a U-test:

1. By convention, assume that the "second sample" is the one with more observations.

2. Calculate U: write all observations in order, then for each observation in the "second sample," add the number of first-sample observations less than it.

3. Use the tables to look up $P(U \leq U_0)$. If $P(U \geq U_0)$ is needed, use the fact that U is symmetric around its mean $n_1 n_2 / 2$.

Some questions may require the use of expressions for the expectation and variance of W and U:

$$\begin{aligned} E[W] &= \frac{n_2(n+1)}{2} \\ \operatorname{Var}(W) &= \frac{n_1 n_2 (n+1)}{12} \\ E[U] &= \frac{n_1 n_2}{2} \\ \operatorname{Var}(U) &= \frac{n_1 n_2 (n+1)}{12}. \end{aligned}$$

Instead of using the U-statistic tables, the normal approximation to W can be used.

2.3.7 Application of QQ plots to evaluate goodness of fit

Given a set of n observations, to determine the x coordinate corresponding to the kth order statistic, do the following:

1. If needed, estimate any unknown parameters in the fitted distribution.

2. Calculate $F^{-1}(k/(n+1))$, where $F(x)$ is the cumulative distribution function of the fitted distribution.

To analyze a normal q-q plot, use the fact that $\xi_Z = \frac{\xi_X - \mu}{\sigma}$, where X is the fitted (x-axis) and Z is the variable being sampled (y-axis). Note that:

1. If the plot is a line, then the normal distribution is a good fit.

2. If the y-intercept is negative, then $\mu > 0$, so the observed mean is smaller than the fitted mean.

3. If the slope is greater than 1, then $\sigma < 1$, so the observed mean has greater variance than the fitted mean.

2.4 Bayesian Parameter Estimation

Learning Objectives: Bayesian Statistics parameter estimation for conjugate prior and posterior distributions:

- Beta-Binomial

- Normal-Normal

- Gamma-Poisson

2.4.1 Background

Recall that Bayes' Theorem states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}.$$

In Bayesian statistics, we treat the parameter θ as a random variable Θ with density function $h(\theta)$, called the *prior density function* of θ. We will regard the density function of a random variable X as a conditional density function, $f(x|\theta)$, given $\Theta = \theta$. Given a sample $X_1, \ldots, X_n$, we use the following functions:

- the joint conditional density function of X, is $L(\mathbf{x}|\theta) = \prod_{1 \leq i \leq n} f(x_i|\theta)$,
- the joint density function of $\mathbf{X}$ and Θ, $g(\mathbf{x}, \theta) = L(\mathbf{x}|\theta)h(\theta)$,
- and the marginal density function of $\mathbf{X}$, $g_1(x) = \int_{-\infty}^{\infty} g(\mathbf{x}, \theta) d\theta$.

Then the *posterior density function* is

$$k(\theta|\mathbf{x}) = \frac{L(\mathbf{x}|\theta)h(\theta)}{g_1(\mathbf{x})}.$$

A shortcut for deriving the posterior density function can be derived based on the observation that $g_1(\mathbf{x})$ does not depend on θ, so we can write

$$k(\theta|\mathbf{x}) = c(\mathbf{x})s(\mathbf{x}|\theta) \propto s(\mathbf{x}|\theta),$$

where $c(\mathbf{x})$ is whatever is needed to make k into a density function. The function $s(\mathbf{x}|\theta)$ can be obtained from $L(\mathbf{x}|\theta)h(\theta)$ by dropping any factors depending only on $\mathbf{x}$. This approach can be strengthened if we have a sufficient statistic Y for θ, because then

$$L(\mathbf{x}|\theta) = g(y|\theta)h(\mathbf{x}),$$

so

$$k(\theta|y) \propto g(y|\theta)h(\theta).$$

The posterior distributions are obtained by modifying the original parameters as follows:

- **Gamma-Poisson**: $\alpha^* = \alpha + \sum_{1 \leq i \leq n} x_i$ and $\beta^* = \beta/(n\beta + 1)$.
- **Normal-Normal**: for $y = \frac{1}{n}\sum_{1 \leq i \leq n} x_i$, prior distribution $N(\theta_0, \sigma_0^2)$, and known variance σ^2, the posterior distribution has parameters:

  $$\theta^* = \left(\frac{\sigma_0^2}{\sigma_0^2 + (\sigma^2/n)}\right) y + \left(\frac{\sigma^2/n}{\sigma_0^2 + (\sigma^2/n)}\right) \theta_0,$$

  and

  $$\sigma^* = \frac{\sigma^2 \sigma_0^2/n}{\sigma_0^2 + (\sigma^2/n)}.$$

- **Beta-Binomial**: $\alpha^* = \alpha + \sum_{1\leq i\leq n} x_i$ and $\beta^* = \beta + n - \sum_{1\leq i\leq n} x_i$.

These are all examples of *conjugate families of distributions*, since the posterior distribution is part of the same family as the prior.

Point estimation of parameters is done relative to a *decision function* $\delta(\mathbf{x})$ and a *loss function* $\mathcal{L}(\theta, \delta(\mathbf{x}))$. A *Bayes estimate* is a decision function δ that minimizes

$$E[\mathcal{L}(\Theta, \delta(\mathbf{x})|\mathbf{X} = \mathbf{x}] = \int_{-\infty}^{\infty} \mathcal{L}(\theta, \delta(\mathbf{x}))k(\theta|\mathbf{x})d\theta.$$

For the loss function $\mathcal{L}(\theta, \delta(\mathbf{x})) = (\theta - \delta(\mathbf{x}))^2$, note that $\delta(\mathbf{x}) = E[\Theta|\mathbf{x}]$, since $E[(W - b)^2]$ is minimized when $b = E[W]$. This is called a square-error loss function.

In addition, there are two other type of loss functions.

1. zero-one loss function $\mathcal{L}(\theta, \delta(\mathbf{x})) = \begin{cases} 0 & \text{if } \delta(\mathbf{x}) = \theta \\ 1 & ow \end{cases}$. $\delta(\mathbf{x})$ is the mode.
2. Abosulte value loss function $\mathcal{L}(\theta, \delta(\mathbf{x})) = |\theta - \delta(\mathbf{x})|$. $\delta(\mathbf{x})$ is median.

2.4.2 Calculate Bayesian Point Estimates for the three conjugate prior distributions listed on the Learning Objective

1. Determine the parameters of the posterior distribution. To simplify this, use $k(\theta|\mathbf{x}) = L(\mathbf{x}|\theta)h(\theta)$, dropping any factors that do not involve θ.
2. If a minimum square error estimate is desired, use the mean from the posterior distribution as the estimate.

As a special case, if the prior distribution is uniform on $[0, 1]$, then this can be regarded as a beta distribution with parameters $\alpha = 1, \beta = 1$.

2.4.3 Calculate Bayesian Interval estimates for the Normal-Normal distribution and special cases of the other conjugate prior distributions listed on the Learning Objective

To determine an Bayesian *credible interval estimate*, determine $u()$ and $v(\mathbf{x})$ such that

$$\int_u^v k(\theta|\mathbf{x})d\theta$$

is large, *e.g.* 0.95.

For the Normal-Normal case, this is just a matter of finding the usual confidence interval for the mean of the posterior distribution.

The Gamma-Poisson case can be handled by using the fact that $\Gamma(n, 2) = \chi^2(2n)$.

2.4.4 Posterior mean as weighted average of posterior distribution and observations

Normal-normal:

$$\left(\frac{\sigma_0^2}{\sigma_0^2+(\sigma^2/n)}\right)y+\left(\frac{\sigma^2/n}{\sigma_0^2+(\sigma^2/n)}\right)\theta_0$$

3 Extended Linear Models

3.1 Introduction of the Generalized Linear Models

Learning Objectives

Understand the assumptions behind different forms of the Generalized Linear Model under the exponential family assuming independent and identically distributed observations and be able to select the appropriate model from list below:

- Ordinary Least Squares
- Generalized Linear Model
- ANOVA

Background

Generalized linear modeling is a methodology for modeling relationships between variables. It generalizes the classical normal linear model, by relaxing some of its restrictive assumptions, and provides methods for the analysis of non-normal data. With the GLM, the variability in one variable is explained by the changes in one or more other variables. The variable being explained is called the dependent or response variable, while the variables that are doing the explaining are the explanatory variables. In some contexts these are called risk factors or drivers of risk. The model explains the connection between the response and the explanatory variables.

3.1.1 Classical Linear Model

Classical linear modeling details explicitly with the approximation involved in by assuming

$$E(y|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

In other words, it is expected value that is being explained by a linear combination of the explanatory variables. Any particular observation on y will deviate from this average. The formulation emphasizes y is regarded as random whereas the x's are considered given or fixed.

One can rewrite Equation (1) as follow:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon E(\epsilon) = 0 \tag{2}$$

Here ϵ is called the error term. In econometrics the term "disturbance" is used. Equation (2) emphasizes y is viewed as determined by the x's, with an error ϵ serving to mask or "disturb" the relationship.

Further assumptions are imposed

- Homoskedastic: The variance of ϵ is finite and does not vary with x variables, at least over the considered range: $Var(\epsilon) = \sigma^2$.
- Normal: The distribution of ϵ is normal.
- Uncorrelated: Each observation on y is uncorrelated with all the other observations.

Addition assumptions are summarized in matrix notation:

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)'$, I is the $n \times n$ identity matrix, $\sigma^2 I$ is the covariance matrix of ϵ, and the distribution of ϵ is multivariate normal. An equivalent way of stating it is in terms of the conditional distribution of y, given X:

$$y|X \sim N(X\beta, \sigma^2 I)$$

Least squares properties under the classical linear model

Under the assumption of the classical linear model,

$$\hat{\beta} \sim N\{\beta, \sigma^2 (X'X)^{-1}\}$$

This statement summarizes the following properties:

- Unbiased. $E(\hat{\beta}) = \beta$.
- Maximum likelihood. This states $\hat{\beta}$ is the MLE of β.
- Minimum variance, invariance, consistency.

Here are some formulas for β:

- Single Variable linear model:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= r_{xy} \frac{s_y}{s_x} \end{aligned}$$

- Multiple variable linear model

$$\hat{\beta} = (X'X)^{-1}X'y$$

Weighted Least Squares

In regression certain observation may be less precise than others, and in a fit it is reasonable to assign them less weight. Weighting cases according to their precision leads to weighted least squares. The precision w_i of case i is signaled through $Var(\epsilon_i) = \sigma^2/w_i$. Thus $w_i \to 0$ indicates no precision and $w_i \to \infty$ indicates complete precision. Multiplying case i by $\sqrt{w_i}$ yields

$$y_i^* = \beta_0\sqrt{w_i} + \beta_1 x_{i1}^* + \cdots \beta_p x_{ip}^* + \epsilon_i^*$$

where $y_i^* \equiv \sqrt{w_i}y_i, x_{ij}^* \equiv \sqrt{w_i}x_{ij}, \epsilon_i^* \equiv \sqrt{w_i}\epsilon_i$ Thus we can get

$$\hat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (X'WX)^{-1}X'Wy$$

where W is the diagonal matrix with diagonal entries w_i. The estimator $\hat{\beta}^*$ is called the weighted least squares estimator. Hence it follows

$$\hat{\beta}^* \sim N\{\beta, \sigma^2(X'WX)^{-1}\}$$

If $Var(y) = \phi E[y]^p$, the following transformation $g(y)$ stablizes the variance:

$$g(y) = \begin{cases} y^{1-p/2} & p \neq 2 \\ \ln y & p = 2 \end{cases}$$

This is known as Box-Cox transformation. To obtain the above, one can simply use the first order taylor expansion to approximate $g(y)$ and then find the function $g(y)$ to eliminate the effect of $E[y]^p$.

3.1.2 General Linear models

Generalized linear models are important in the analysis of insurance data. With insurance data, the assumption of the normal model are frequently not applicable. For example, claim sizes, claim frequencies and the occurrence of a claim on a single policy are all outcomes which are not normal. Also, the relationship between outcomes and drivers of risk is often multiplicative rather than additive.

Given a response y, the generalized linear model (GLM) is

$$f(y) = c(y, \phi) exp\{\frac{y\theta - a(\theta)}{\phi}\}, g(\mu) = x'\beta$$

The equation for $f(y)$ specifies that the distribution of the response is in the exponential family. The second equation specifies that a transformation of the mean, $g(\mu)$, is linearly related to explanatory variables contained in x.

- The choice of $a(\theta)$ determines the response distribution. Note, $E[y] = a'(\theta)$ and $Var(y) = \phi a''(\theta)$. Here ϕ is called dispersion parameter since varying it allows varying the relationship of the variance to the mean.
- The choice of $g(\mu)$, called the link, determines how the mean is related to the explanatory variables x. Note g is a monotonic differentiable function, such as log or square root.
- Given x, μ is determined through $g(\mu)$. Given μ, θ is determined through $a'(\theta) = \mu$. Finally given θ, y is determined as a draw from the exponential density specified in $a(\theta)$.
- Observations on y are assumed to be independent.

Steps in generalized linear modeling

1. Choose a response distribution $f(y)$ and hence choose $a(\theta)$. The response distribution is tailored to the given situation.
2. Choose a link $g(\mu)$. This choice is sometimes simplified by choosing the so called "canonical" link corresponding to each response distribution.
3. Choose explanatory variable x in terms of which $g(\mu)$ is to be modeled.
4. Collect observations $y_1, \cdots, y_n$ on the response y and corresponding values $x_1, \cdots, x_n$ on the explanatory variables x. Successive observations are assumed to be independent.
5. Fit the model by estimating β and, if unknown, ϕ.
6. Given the estimate of β, generate predictions of y for different settings of x and examine how well the model fits by examining the departure of the fitted values from actual values, as well as other model diagnostics.

Link Functions

If $g(\mu) = \theta$ then g is called the canonical link corresponding to $a(\theta)$. In this case, $\theta = x'\beta$.

The list are commonly used canonical links.

Link Function	$g(\mu)$	Canonical link for
identity	μ	normal
log	$\ln \mu$	Poisson
power	μ^p	gamma ($p = -1$); inverse Gaussian ($p = -2$)
Square root	$\sqrt{\mu}$	
logit	$\ln \frac{\mu}{1-\mu}$	binomial

Modeling counts such as the number of claims or deaths in a risk group requires correction for the number n exposed to risk. If μ is the mean of the count y, then the occurrence rate μ/n of interest and

$$g(\mu/n) = x'\beta$$

When g is the log function, this becomes

$$\ln(\mu/n) = x'\beta \implies \ln \mu = \ln n + x'\beta$$

The variable n is called the exposure and $\ln n$ is called an “offset”. An offset is effectively another x variable in the regression with a β coefficient equal to one. Offsets are used to correct for group size or differing time periods of observations.

Exponential Family

Parameters of commonly used exponential family distributions: (In all cases, $E[y] = \mu$)

Distribution	θ	μ	$a(\theta)$	ϕ	$V(\mu) = a''(\theta)$
Binomial	$\ln \frac{\mu}{n-\mu}$	$\frac{ne^\theta}{1+e^\theta}$	$n \ln(1 + e^\theta)$	1	$\mu(1 - \mu/n)$
Binomial in terms of π	$\ln \frac{\pi}{1-\pi}$	$n\pi$	$n \ln(1 + e^\theta)$	1	$n\pi(1 - \pi)$
Poisson	$\ln \mu$	e^θ	e^θ	1	μ
Negative binomial	$\ln \frac{\kappa\mu}{1+\kappa\mu}$	$\frac{e^\theta}{(1-e^\theta)}$	$-\frac{1}{k} \ln(1 - e^\theta)$	1	$\mu(1 + \kappa\mu)$
Normal	mu	θ	$\frac{1}{2}\theta^2$	σ^2	1
Gamma	$-\frac{1}{\mu}$	$-\frac{1}{\theta}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	μ^2
Inverse Gaussian	$-\frac{1}{2\mu^2}$	$\frac{1}{\sqrt{-2\theta}}$	$-\sqrt{-2\theta}$	σ^2	μ^3

Tweedie Distribution

Suppose

$$y = z_1 + \cdots + z_c$$

with $y = 0$ if $c = 0$.

If c is Poisson and the z_j are independent gamma random variables then y has the Tweedie distribution. The distribution has a non-zero probability at $y = 0$ equal to the Poisson probability of no claims. The rest of the distribution is similar to the gamma. The Tweedie distribution is a member of the exponential family and $Var(y)\phi\mu^p$ where $1 < p < 2$. it is similar to the zero-adjusted inverse Gaussian distribution.

3.2 Model Evaluation

3.2.1 Standard Error of regression

Let's look at a simple linear model. Suppose y is the response variable, $\hat{y}$ is the estimate and $\bar{y}$ is the mean. Then We can express the difference between y_i and the sample mean into

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Squarng both sides and adding up over all i, we get

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

where the total sum of squares consists of two parts, the first term on the right hand side is called error sum of squares and the second term is regression sum of squares.

The sum of squares information can be summarized in an analysis of variance (ANOVA) table. The table looks like this:

Source	Sum of Squares	df	Mean square
Regression	SSR	p	SSR/p
Error	SSE	$n-(p+1)$	$SSE/(n-(p+1))$
Total	SST	$n-1$	$SST/(n-1)$

The F statistic, which tests the significance of the entire regression, is the quotient of the mean square of the regression over the mean square error:

$$F_{p,n-(p+1)} = \frac{SSR/p}{SSE/(n-(p+1))}$$

It tests whether the model as a whole is significant. In other words, the null hypothesis for the F test is $H_0 : \beta_1 = \beta_2 = \cdots = 0$. In a 1-variable model, the $F_{1,n-2}$ statistic is the square of the t_{n-2} statistic for β_1.

3.2.2 R^2:the coefficient of determination

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

where is the positive square root of R^2 is called the multiple correlation coefficient and it is the square of the correlation coefficient of y and $\hat{y}$.

3.2.3 t statistics

$\hat{\beta}$ is an unbiased estimator of β and has the minimum variance of all unbiased estimators. Its covariance matrix is

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Since σ^2 is unknown, we will use the standard errors of regression s^2. Hence the variance of $\hat{\beta}$ is estimated as $s^2_{\hat{\beta}_i} = s^2\psi_i$ where ψ_i is the it diagonal element of $(X'X)^{-1}$. The square root of the it component of the estimated variance of $\hat{\beta}$ is called the standard error of β_i. To test the null hypothesis that $\hat{\beta}_i = 0$, we use the t statistic β_i/s_β, which has $n-(p+1)$ degrees of freedom. More generally, to test the null hypothesis that $_i = \beta^*$, we use

$$t_{n-(p+1)} = \frac{\hat{\beta}_i - \beta^*}{s_{\beta_i}}$$

A 100p% confidence interval for β_i may be constructed as $\hat{\beta}_i \pm ts_\beta$, where t is the $100(1+p)/2$ percentile of a t-distribution with the appropriate number of degrees of freedom.

3.2.4 Confidence Internal

Given values of the explanatory variables x, the estimated value of the mean of y is $\hat{\mu}$ where $g(\hat{\mu}) = x'\hat{\beta}$. A confidence interval around the estimate is used to indicate precision. The computation of the confidence interval requires the sampling distribution of $\hat{\mu}$. The variance of the linear predictor $x'\beta$ is

$$Var(x'\hat{\beta}) = \phi x'(X'WX)^{-1}x$$

Thus an approximate confidence interval for the mean is (μ_l, μ_u) where

$$g(\mu_l) = x'\hat{\beta} - z\sqrt{\phi x'(X'WX)^{-1}x}$$

and μ_u similarly defined with a plus sign. Hence z is the appropriate point on the $N(0, 1)$ distribution. The dispersion ϕ is replaced by an estimate.

The estimate $\hat{\mu}$ is unbiased when using an identity link. For other links it is biased.

3.2.5 Goodness of Fit

Deviance

One way of assessing the fit of a given model is to compare it to the model with the best possible fit. The best fit will obtain when there are as many parameters as observations: this is called a saturated model. The saturated model will ensure there is complete flexibility in fitting θ_i. Since

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - \mu_i}{\phi} = \frac{y_i - a'(\theta_i)}{\phi},$$

the MLE of θ_i under the saturated model is $\check{\theta}_i$, where $a'(\check{\theta}_i) = y_i$. Thus each fitted value is equal to the observation and the saturated model fits perfectly.

The value of the saturated log-likelihood is

$$\check{l} \equiv \sum_{i=1}^{n} \{\ln c(y_i, \phi) + \frac{y_i\check{\theta} - a(\check{\theta})}{\phi}\}$$

which is the maximum possible log-likelihood for y given the response distribution specified by $a(\theta)$. This value is compared to $\hat{l}$, the value of the maximum of the log-likelihood based on y and the given explanatory variables. The "deviance" denoted as Δ, is defined as a measure of distance between the saturated and fitted models:

$$\Delta = 2(\check{l} - \hat{l})$$

- The smaller, the better fit
- The size of Δ is assessed relative to the χ^2_{n-p} distribution.
- A direct calculation shows that for the exponential family

$$\Delta = 2\sum_{i=1}^{n} \{\frac{y_i(\check{\theta}_i - \hat{\theta}_i) - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi}\}$$

where $\check{\theta}_i$ and $\hat{\theta}_i$ are such that $a'(\check{\theta}_i) = y_i$ and $g\{a'(\hat{\theta}_i)\} = x_i'\hat{\beta}$, respectively.

- When ϕ is unknown and estimated, then the χ^2_{n-p} distribution for the deviance is compromised. In case of the Poisson, $\phi = 1$ and the χ^2 approximation is useful. When it is normal and σ^2 is known, then χ^2 is exact; however, if σ^2 is estimated, then it is not reliable. Some paper shows deviance is useful for testing the significance of explanatory variables in nested models.

Distribution	deviance
Normal	$\frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^n \{y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$
Binomial	$2\sum_{i=1}^n n_i\{y_i \ln(y_i/\hat{\mu}_i) + (n_i - y_i)\ln(\frac{n_i - y_i}{1-\hat{\mu}_i})\}$
Gamma	$2\nu\sum_{i=1}^n \{-\ln(y_i/\hat{\mu}_i) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\}$
Inverse Gaussian	$1/\sigma^2 \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i}$
Negative binomial	$2\sum_{i=1}^n \{y_i \ln(y_i/\hat{\mu}_i - (y_i + 1/\kappa)\ln(\frac{y_i + 1/\kappa}{\hat{\mu}_i + 1/\kappa})\}$

AIC Akaike Information Criterion: The penalty is 2 for each parameter. $AIC = -2l + 2q$, where p is number of parameters.

BIC Bayesian Information Criterion: The penalty varies with the number of observations and is $\ln n$ for each parameter. The formula is $BIC = -2l + q\ln n$.

3.2.6 Model Validation

In the classical linear model setup, let's define hat matrix as

$$H = X(X'X)^{-1}X'$$

. Then $\hat{y} = Hy$. Since $y - \hat{y} = \hat{\epsilon}$, it follows that $\hat{\epsilon} = (I - H)y)$. Let $M = I - H$, then $\hat{\epsilon} = My$. Note H and M are self-adjoint. Hence it follows that

$$Var(\hat{\epsilon}) = M^2\sigma^2$$

This is the covariance matrix for $\hat{\epsilon}$. The variance $\hat{\epsilon}_i$ is estimated by $(1 - h_{ii})s^2$, where h_{ii} is the it diagonal element of the H matrix. The standardized or studentized residuals are $\hat{\epsilon}_i/(s\sqrt{1 - h_{ii}})$.

The classical linear model makes the following assumptions:

- linearity: To construct an added variable plot, also known as a partial regression or partial leverage plot, first regress y and x_k on the other variables. The added variable plot plots the residuals of the y regression on the residuals of the x_j regression. If these are more or less linear with nonzero slope, then x_j

should be in the model. If the slope is 0, then x_j should not be in the model. If the plot isn't linear, perhaps x_j should be transformed.

To construct a partial residual plot, add $\hat{\beta}_j x_j$ to $\hat{\epsilon}$. The partial residuals $\hat{\epsilon}_i^*$ are defined by

$$\hat{\epsilon}_i^* = \hat{\epsilon}_i + \hat{\beta}_j x_{ij}$$

where the regression has all variables including x_j. Plot the partial residuals against $x_i j$. The plot will have slope $\hat{\beta}_j$, but if the plot is not close to a line, a nonlinear relationship is indicated.

- Normal: To check normality, construct a pp-plot or a qq-plot of the studentized residuals against a standard normal distribution. Closeness to the 45 line indicated normality.

- constant variance: To check that the variance is constant, construct a scatter plot of the residuals or studentized residuals against each explanatory variable x_j, or against y. A random scatter indicates homoskedasticity. A non-linear pater may indicate that the relationship between the explanatory variable and the output variable is not linear

- independence: plot the ϵ_i in the order for which correlation is expected, such as order of occurrence. Alternatively, examine M. Correlations between residuals are indicated by off-diagonal numbers, which should be small.

Influence points

If some points having unusually high residuals, they are outliers. If the model is valid, with residual is normally distribution with mean 0 and variance $s^2(1 - h_{ii}) =$, Here h_{ii} are called leverage. They must be between $1/n$ and 1, and sum up to $p + 1$. Thus on average they are $(p+1)/n$. If leverage is more than 2 or 3 times the average oval, the data should be reviewed.

Cook's distance combines outliers and high-leverage points into a single measurement. Let $\hat{y}_{j(i)}$ be the fitted value of y_j if the it observation is removed from data set. Let D_i be Cook's distance for the it variable. Then

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)s^2} = (\frac{\hat{\epsilon}_i}{se(\hat{\epsilon}_i)})^2 \frac{h_{ii}}{(p+1)(1-h_{ii})}$$

where $se(\hat{\epsilon}_i)$ is the standard error of $\hat{\epsilon}_i$. The first factor of the last expression measures how much of an outlier the it point is. The second factor measures leverage. Rules of thumb for determined whether D_i is too high are $D_i > 1, D_i > 4/(n-(p+1))$ and $D_i > 4/n$.

Predictive accuracy of model

$$SSPE = \sum_{i=n_1+1}^{n} (y_i - \hat{y}_i)^2$$

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} (\frac{\hat{\epsilon}_i}{1-h_{ii}})^2$$

3.2.7 Testing the significance of explanatory variables

As before, hypotheses are written as $C\beta = r$ where C is known matrix - sometimes called the hypothesis matrix - and r is a set of given values.

There are three main approaches to testing hypotheses of the form $C\beta = r$. Each of these approaches considers the likelihood or log-likelihood. Write $\hat{\beta}$ as the unrestricted MLE of β, and let $\tilde{\beta}$ denote the MLE of β when l is maximized subject to the restrictions $C\beta = r$. Further, write $\hat{l}$ as the value of l at $\hat{\beta}$, and $\tilde{l}$ the value of l at $\tilde{\beta}$. Obviously, $\hat{l} \geq \tilde{l}$. Also let q be the number of rows of C.

Likelihood ratio test Let $\hat{l}$ be the log likelihood of the unconstrained model and $\tilde{l}$ the loglikelihood of the constrained model. Then the likelihood ratio statistic

$$2(\hat{l} - \tilde{l})$$

has a chi-square distribution with q degrees of freedom. Often the dispersion parameter ϕ has to be estimated, but as long as it is estimated consistently in the two models the statistic should be valid. The likelihood ratio statistic may also be computed as the difference of the deviances of the two models as long as the same estimate for ϕ is used in both log-likelihoods.

Wald Test This measures how far $C\hat{\beta}$ is from r, with a large difference $C\hat{\beta} - r$ providing evidence against the restrictions. The estimate $\hat{\beta}$ is required, but not $\tilde{\beta}$. If $C\beta = r$ then since $\hat{\beta} \sim N\{\beta, \phi(X'WX)^{-1}$ it follows that

$$C\hat{\beta} - r \sim N\{0, \phi C(X'WX)^{-1}C'\}$$

This leads to the Wald statistic for testing $C\beta = r$:

$$(C\hat{\beta} - r)'\{\phi C(X'WX)^{-1}C'\}^{-1}(C\hat{\beta} - r) \sim \chi^2_q$$

For testing a single coefficient $\beta_j = 0$, the Wald statistic reduces to

$$\frac{(\hat{\beta}_j - r)^2}{\phi\psi_j} \sim \chi_1^2$$

where ψ_j is the $j+1$ st diagonal entry in the matrix $\phi(X'WX)^{-1}$.

One can use normal approximation as

$$\frac{\hat{\beta}_j - r}{\sqrt{\phi\psi_j}}$$

Score Test This test only requires $\tilde{\beta}$ but not $\hat{\beta}$. It checks whether the slope of the log likelihood at $\tilde{\beta}$ is high; if it is, it is an indication that additional variables should be added to the model. “Score” is a term that refers to the first derivative of the log likelihood. The score statistic is

$$\dot{l}'(\tilde{\beta})(Var(\dot{l}(\beta))^{-1}\dot{l}(\tilde{\beta}) \sim \chi_q^2$$

with $\dot{l}(\tilde{\beta}) = \phi^{-1}X'WG(y - \tilde{\mu}), Var(\dot{l}(\beta)) = \phi^{-1}(X'WX)$.

When the identity link is used, the statistic reduces to

$$\phi^{-1}(\hat{\beta} - \tilde{\beta})'(X'WX)(\hat{\beta} - \tilde{\beta})$$

Additional Information Type I tests are sequential tests. We start off with no variables and add them in one by one, or several at a time if necessary, as with categorical variables. Each variable is tested assuming the previously added variables are in the model. This testing procedure deepens on the order the variables are added to the model.

Type II tests are tests on variables or sets of variables that assume that all other variables are present in the model. This testing procedure does not depend on the order the variables are added to the model. Walk tests are type II by their very nature; they test the significance of a variable or group of variables assuming that all the other variables are in the model. Likelihood ratio test may be Type I or Type III.

Likelihood ratio tests are superior statistically to Wald tests, so when the two lead to opposite conclusions, the likelihood ratio test result should be preferred.

3.2.8 Other Diagnostic Tools

Deviance Residuals

$$\delta_i^2 = \sqrt{\frac{2(y_i(\check{\theta}_i - \hat{\theta}_i) - a(\check{\theta}_i) + a(\hat{\theta}_i))}{\phi}}$$

The sign on the square root is taken as the sign of $y_i - \hat{\mu}_i$.

Anscombe residuals

$$\frac{h(y_i) - h(\hat{y}_i)}{\dot{h}(\hat{y}_i)\sqrt{V(\hat{y}_i)}}$$

where $\dot{h}(y) = V(y)^{-1/3}$

Checking the link Plot $g(\hat{\mu}_i) + \dot{g}(\hat{\mu}_i)(y_i - \hat{\mu}_i)$ against $x_i'\beta$. It should be a straight line.

Added Variable Partial residual plots Added value plot is similar to classic linear case. For the partial residual plot, plot $(y_i - \hat{y}_i)\dot{g}(\hat{y}_i) + x_{ij}\hat{\beta}_j$ against x_{ij}.

Leverage Leverage is based on diagonal elements of the hat matrix, which for GLM is defined by

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

For 1 variable regression: it can be simplified as $h_{ii} = \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_j-\bar{x})^2}$

3.3 Model Estimation

3.3.1 Maximum Likelihood

The MLE of β and θ are derived by maximizing the log-likelihood, defined as

$$l(\beta, \phi) = \sum_{i=1}^{n} \ln f(y_i; \beta, \phi) = \sum_{i=1}^{n} \{\ln c(y_i, \phi) + \frac{y_i\theta_i - a(\theta_i)}{\phi}\}$$

which assumes independent exponential family responses y_i.

Consider the MLE for β_j. To find the maximum, $l(\beta, \phi)$ is differentiated with respect to β_j:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}$$

where

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - a'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}, \frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i} x_{ij}$$

Here $\eta_i = x_i'\beta$ and x_{ij} is component i of x_j. Setting $\partial l / \partial \beta_j = 0$ yields the first order conditions for likelihood maximization:

$$\sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i} x_{ij}(y_i - \mu_i) = 0 \iff X'D(y - \mu) = 0$$

where D is the diagonal matrix with diagonal entries $\partial \theta_i / \partial \eta_i = \{g'(\mu_i)V(\mu_i)\}^{-1}$.

Hence, for β, the resulting formula is

$$X'WG(y - \mu) = 0$$

where

- $\mu = g^{-1}(X'\beta)$.
- W is the diagonal matrix with entries $(g'(\mu_i)^2 V(\mu_i))^{-1}$.
- G is the diagonal matrix with entries $g'(\mu_i)$.

If we use a Taylor series approximation, then $g(y_i)g(\mu_i) + g'(\mu_i)(y_i - \mu_i) \implies g(y)g(\mu) + g'(\mu)(y - \mu)$ where $g(y)$ is a vector with entries $g(y_i)$ and similarly for $g(\mu)$. We can then conclude that

$$\hat{\beta} \approx (X'WX)^{-1} X'Wg(y)$$

3.3.2 Fisher Scoring

Usually, $X'WG(y - \mu) = 0$ is difficult to solve except for cases such as the normal with the identity link. Recall, Newton assumes that the first and second derivatives of the function to be maximized can be easily evaluated at each point. Using these derivatives, a quadratic approximation is made to the function at this point and it is the quadratic that is maximized. The resulting maximizer is then used to derive a new quadratic which in turn is maximized. This sequence of approximate maximizers is often found to converge quickly to the actual maximum.

To simplify this discussion, suppose ϕ is known and so write $l(\beta)$ to denote the log-likelihood as a function of the unknown parameter vector β. If β contains a single parameter then the quadratic Taylor series approximation at any point β is

$$l(\beta + \delta) \approx l(\beta) + l'(\beta)\delta + \frac{\delta^2}{2} l''(\beta)$$

Differentiating the right hand side as a function of δ and equating to zero yields

$$l'(\beta) + \delta l''(\beta) = 0 \implies \delta = -\{l''(\beta)\}^{-1} l'(\beta)$$

With β given and δ as specified, a higher point thus appears to be $\beta - l'(\beta)/l''(\beta)$. Denoting $\beta^{(m)}$ as the value for β at the iteration m, the update equation is

$$\beta^{(m+1)} = \beta^{(m)} - \{l''(\beta^{(m)})\}^{-1} l'(\beta^{(m)})$$

For a maximum $l''(\beta) < 0$. Iteration of this equation, each time revising β, leads to a sequence which, as stated above, often rapidly converges. This procedure is called Newton-Raphson iteration. Note, $-l''(\beta)$ is non-negative definite to achieve maximum and $l'(\beta)$ is called score vector.

Well, as for Fisher scoring, we can replace $l''(\beta)$ by its expectation $E\{l''(\beta)\}$. The matrix $-E\{l''(\beta)\}$ is called the "Fisher information" matrix. For GLMs the information matrix is

$$E\{l''(\beta)\} = \phi^{-1} X'WX$$

Hence we can write

$$\beta^{(m+1)} = \beta^{(m)} + (X'WX)^{-1} X'WG(y - \mu)$$

This equation often is rewritten as the following

$$\beta^{(m+1)} = (X'WX)^{-1} X'W\{X\beta^{(m)} + G(y - \mu)\}$$

where the expression in the curly brackets on the right is called the "local dependent variable". The local dependent variable is computed by replacing μ by the estimate $\mu^{(m)}$ where $g(\mu^{(m)}) = X\beta^{(m)}$. Computing $\beta^{(m+1)}$, requires $V(\mu)$, the variance function of the distribution and $g'(\mu)$, the derivative of the link. The dispersion parameter ϕ os mpg required.

The inverse of the Fisher information matrix is approximately, for late n, the covariance matrix of $\hat{\beta}$. Also MLEs are asymptotically unbiased and, again for large samples, approximately normal. Hence

$$\hat{\beta} \sim N\{\beta, \phi(X'WX)^{-1}\}$$

3.3.3 Quasi-Likelihood

For every exponential family response, $Var(y) = \phi V(\mu)$. However, many combinations of ϕ and $V(\mu)$ do not correspond to an exponential family response. Hence for such combinations of and $V(\mu)$ the following expression

$$\phi^{-1}X'D(y-\mu)=\sum_{i=1}^{n}\frac{\partial\mu_i}{\partial\beta}\frac{y_i-\mu_i}{\phi V(\mu_i)}$$

is not the derivative of an exponential family log-likelihood. A solution to this problem is to maximize the "quasi-likelihood". A quasi-likelihood $Q(\beta)$ is any function of β which has derivatives $Q'(\beta)$ as the same as above equation. With quasi-likelihood estimation. It is $Q(\beta)$ which is maximized with respect to β. Obviously, the quasi-likelihood estimate of β coincides with $\hat{\beta}$ in those cases where corresponds to the derivative of a likelihood. Similar to the proper maximum likelihood situation, the covariance matrix of $\hat{\beta}$ is determined from the second derivatives of $Q(\beta)$.

Hence, we can see that quasi-likelihood produces the same β estimation but in some cases, it inflates the standard error of the estimators.

In case of Poisson, we can use $\sqrt{\phi}$ as a multiplier for the standard error of β, where ϕ can be estimated as $\sqrt{\frac{\Delta}{n-p-1}}$.

3.3.4 collinearity

We can assess the importance of including one or more candidate explanatory variables in the model with the t and F tests. It is also good practice to check for collinearity among the explanatory variables themselves. Collinearity occurs when one explanatory variable is a linear combination of other explanatory variables. High correlation among the explanatory variables results in high variances of the corresponding regression parameter estimators. The variance inflation factor is useful statistic in measuring the degree of collinearity present among the explanatory variables. The VIF for the nth explanatory variable is found by regressing this variable against the other explanatory variables in the model. The $VIF_j=(1-R_j^2)^{-1}$, where R_j^2 is the coefficient of determination for this linear model.

4 Time Series with Constant Variance

4.1 Use Time Series to Model Trends

Learning Objectives

- Estimation, data analysis and forecasting
- Forecast errors and confidence intervals[2]

4.1.1 Background

Notation

Let's denote a time series of length n by $\{x_t : t = 1, \cdots, n\} = \{x_t\}$ and $\hat{x}_{t+k|t}$ is a forecast made at time t for a future value at time $t + k$. A forecast is a predicted future value, and the number of time steps into the future is the lead time (k).

Models

- Additive decomposition model: $x_t = m_t + s_t + z_t$ where, at time t, x_t is the observed series, m_t is the trend, s_t is the seasonal effect, and z_t is an error term that is, in general, a sequence of correlated random variables with mean zero.
- Multiplicative model: $x_t = m_t \cdot s_t + z_t$. If the random variation is modelled by a multiplicative factor and the variable is positive, an additive decomposition model for $\log(x_t)$ can be used: $\log(x_t) = m_t + s_t + z_t$. Some care is required when the exponential function is applied to the predicted mean of $\log(x_t)$ to obtain a prediction for the mean value x_t, as the effect is usually to bias the predictions. If the random series z_t are normally distributed with mean 0 and variance σ^2, then the predicted mean value at time t is given by $\hat{x}_t = e^{m_t+s_t}e^{\frac{1}{2}\sigma^2}$.

Estimating trends and seasonal effects

There are various ways to estimate the trend m_t at time t, but a relatively simple procedure is to calculate a moving average entered on x_t. A moving average is an average of a specified number of time series values around each value in the time series, with the exception of the first few and last few terms. For example, we use

[2] Assume you know R. Here we only discuss univariate time-series.

the following for monthly moving average:

$$\hat{m}_t = \frac{\frac{1}{2}x_{t-6} + x_{t-5} + \cdots + x_{t-1} + x_t + x_{t+1} + \cdots + x_{t+5} + \frac{1}{2}x_{t+6}}{12}$$

In particular, this formula is derived from average two moving averages, that is, the average from Feb to Jan next year, and Jan to Dec. This process is called centering. Hence we can obtain the seasonal affection by removing $\hat{m}_t$ (subtract if additive model; divide if multiplicative model). In other words, $\hat{s}_t = x_t - \hat{m}_t$ for additive and $\hat{s}_t = x_t/\hat{m}_t$ for multiplicative. Then we will average $\hat{s}_t$ for each month to obtain tentative seasonal adjustment factors. Then this effect will be removed from the seasonal factors to reach final seasonal adjustment factors.

Smoothing

The centred moving average is an example of a smoothing procedure that is applied retrospectively to a time series with objective of identifying an underlying signal or trend.

A second smoothing algorithm offered by R is stl. This uses a locally weighted regression technique known as loess. The regression, which can be a line or higher polynomial, is referred to as local because it uses only some relatively small number of points on either side of the point at which the smoothed estimate is required. The weighting recedes the influence of outlying points and is an example of robust regression. The term filtering is also used for smoothing, particularly in the engineering literature.

Correlation

The mean function of a time series model is

$$\mu(t) = E(x_t)$$

and, in general, is a function of t. The expectation in this definition is an average taken across the ensemble of all the possible time series that might have been produced by the time series model.

If the mean function is constant, we say that the time series model is stationary in the mean where the sample estimate of the population mean, μ, is the sample mean, $\bar{x}$:

$$\bar{x} = \sum_{t=1}^{n} x_t/n$$

Above equation does rely on an assumption that a sufficiently long time series characterizes the hypothetical model. Such models are known as ergodic. A time series model that is stationary in the mean is ergodic in the mean if the time average for a single time series tends to the ensemble mean as the length of the time series increases:

$$\lim_{n\to\infty} \frac{\sum x_t}{n} = \mu$$

Consider a time series model that is stationary in the mean and the variance. The variables may be correlated, and the model is second order stationary if the correlation between variables depends only on the number of time steps separating them. The number of time steps between the variables is known as the lag. A correlation of a variable with itself at different times is known as autocorrelation or serial correlation. If a time series model is second-order stationary, we can define an auto covariance function ($acvf$), γ_k, as a function of the lag k:

$$\gamma_k = E[(x_t - \mu)(x_{t+k} - \mu)]$$

. The lag autocorrelation function (acf), ρ_k, is defined by

$$\rho_k = \frac{\gamma_k}{\sigma^2}$$

. The sample acvf, c_k, is calculated as

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

and the sample acf is defined as

$$r_k = \frac{c_k}{c_0}$$

By default, the acf function from R produces a plot of r_k against k, which is called the correlogram. Dashed lines are drawn at $y = -\frac{1}{n} \pm \frac{2}{\sqrt{n}}$. Assume that $\rho_k = 0$ for all k, these are bounds for the values of r_k at 5% significant. In addition, the linear effect of one variable on another is proportionate to the square of the correlation. An autocorrelation of 0.1 indicates only a 1% linear effect.

Cross-correlation

Suppose we have time series models for variables x and y that are stationary in the mean and the variance. The variables may each be serially correlated, and correlated with each other at different time lags. The combined model is second-order stationary if all these correlations depend only on the lag, and then we can define the cross covariance function (ccvf), $\gamma_k(x,y)$, as a function of the lag, k:

$$\gamma_k(x,y) = E[(x_{t+k} - \mu_x)(y_t - \mu_y)]$$

The cross-correlation function (ccf), $\rho_k(x,y)$, is defined by

$$\rho_k(x,y) = \frac{\gamma_k(x,y)}{\sigma_x \sigma_y}$$

Note $\gamma_k(x,y) = \gamma_{-k}(y,x)$ and similarly for $\rho_k(x,y)$. We estimate ccvc and ccc from the sample using $c_k(x,y)$ and $r_k(x,y)$ respectively, which are defined by

$$c_k(x,y) = \frac{\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(y_t - \bar{y})}{n}$$

$$r_k(x,y) = \frac{c_k(x,y)}{\sqrt{c_0(x,x)c_0(y,y)}}$$

4.1.2 Bass Model

The Bass formula for the number of people, N_t, who have bought a product at time t depends on three parameters: the total number of people who eventually buy the product, m; the coefficient of innovation, p; and the coefficient of imitation, q. The Bass formula is

$$N_{t+1} = N_t + p(m - N_t) + qN_t(m - N_t)/m$$

According to the model, the increase in sales, $N_{t+1} - N_t$, over the next time period is equal to the sum of a fixed proportion p and a time varying proportion $q\frac{N_t}{m}$, of people who will eventually buy the product but have not yet done so. The rationale for the model is that initial sales will be to people who are interested in the novelty of the product, whereas later sales will be to people who are drawn to the product after seeing their friends and acquaintances use it.

Let $f(t), F(t)$ and $h(t)$ be the density, cumulative distribution function and hazard of the distribution of time until purchase. The definition of the hazard is

$$h(t) = \frac{f(t)}{1 - F(t)}$$

The interpretation of the hazard is that if it is multiplied by a small time increment it gives the probability that a random purchaser who has not yet made the purchase will do so in the next small time increment. Then the continuous time model of the Bass formula can be expressed in terms of the hazard:

$$h(t) = p + qF(t)$$

The solution of this equation will be

$$F(t) = \frac{1 - e^{-(p+q)t}}{1 + (q/p)e^{-(p+q)t}}$$

Two special cases of the distribution are the exponential distribution and logistic distribution, which arise when $q = 0$ and $p = 0$. Hence, we can get the solution for the Bass formula is just

$$N_t = m(\frac{1 - e^{-(p+q)t}}{1 + (q/p)e^{-(p+q)t}}$$

Also note to maximize $S(t) = mf(t)$, we can get the peak sales in the continuous-time model occur at

$$t_{peak} = \frac{\ln q - \ln p}{p + q}$$

4.1.3 Exponential Smoothing and Holt-Winters Methods

Exponential Smoothing

Our objective is to predict some future value x_{n+k} given a past history $\{x_1, x_2, \cdots, x_n\}$ of observations up to time n. A typical application is forecasting sales of a well-established product in a stable market. The model is

$$x_t = \mu_t + w_t$$

where μ_t is the non-stationary mean of the process at time t and w_t are independent random derivations with a mean of 0 and a standard deviation σ. Given that there is no systematic trend, an intuitively reasonable estimate of the time at time t is given by a weighted average of our observation at time t and our estimate of the time at time $t - 1$:

$$a_t = \alpha x_t + (1 - \alpha)a_{t-1}, 0 < \alpha < 1 \qquad (3)$$

The a_t is the exponentially weighted moving average (EWMA) at time t. The value of α determines the amount of smoothing, and it is referred to as the smoothing parameter. Since we have assumed that there is no systematic trend and that there are no seasonal effects, forecasts made at time n for any lead time are just the estimated mean at time n. The forecasting equation is

$$\hat{x}_{n+k|n} = a_n, k = 1, 2, \cdots$$

Furthermore, we can rewrite Equation (3) as

$$a_t = \alpha(x_t - a_{t-1}) + a_{t-1}$$

By repeated back substitution we obtain

$$a_t = \alpha x_t + \sum_{k=1}^{\infty} \alpha(1-\alpha)^k x_{t-k}$$

When written in this form, we see that a_t is a linear combination of the current and past observations, with more weight given to the more recent observations. The one-step-ahead prediction errors, e_t, are given by

$$e_t = x_t - \hat{x}_{t|t-1} = x_t - a_{t-1}$$

By minimizing the same of squared one-step-ahead prediction errors, we can obtain a value for the smoothing parameter α.

Holt-Winters Method

We will refer to the change in level from one time period to the next as the slope and seasonally adjusted mean as level. The Holt-Winters method generalizes Equation (3) and the additive seasonal form of their updating equations for a series $\{x_t\}$ with period p is

$$\begin{cases} a_t = \alpha(x_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \\ b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \\ s_t = \gamma(x_t - a_t) + (1-\gamma)s_{t-p} \end{cases}$$

where a_t, b_t and s_t are the estimated level, slope and seasonal effect at time t, and α, β and γ are the smoothing parameters. The first updating equation takes a weighted average of our latest observation, with our existing estimate of the appropriate seasonal effect subtracted, and our forecast of the level made one time step ago. The one-step-ahead forecast of the level is the sum of the estimated of the level and slope at the time of forecast. The second equation takes a weighted average of our previous estimate and latest estimate of the slope, which is the difference in the estimated level at time t and the estimated level at time $t-1$. Finally, we have another estimate of the seasonal effect, from the difference between the observation and the estimate of the level, and we take a weighted average of this and the last estimate of the seasonal effect for this season, which was made at time $t-p$.

The forecasting equation for x_{n+k} made after the observation at time n is

$$\hat{x}_{n+k|n} = a_n + kb_n + s_{n+k-p}, k \leq p$$

where a_n is the estimated level and b_n is the estimated slope, so $a_n + kb_n$ is the expected level at time $n+k$ and s_{n+k-p} is the exponentially weighted estimate of the seasonal effect made at time $n = k-p$.

The Holt-Winters algorithm with multiplicative seasonal is

$$\begin{cases} a_n = \alpha(\frac{x_n}{s_{n-p}}) + (1-\alpha)(a_{n-1} + b_{n-1}) \\ b_n = \beta(a_n - a_{n-1}) + (1-\beta)b_{n-1} \\ s_n = \gamma(\frac{x_n}{a_n}) + (1-\gamma)s_{n-p} \end{cases}$$

The forecasting equation for x_{n+k} made after the observation at time n becomes

$$\hat{x}_{n+k|n} = (a_n + kb_n)s_{n+k-p}$$

4.1.4 White Noise and Random Walks

A residual error is the difference between the observed value and the model predicted value at time t. If we suppose the model is defined for the variable y_t and $\hat{y}_t$ is the value predicted by the model, the residual error x_t is

$$x_t = y_t - \hat{y}_t$$

. We know that features of the historical series, such as the trend or seasonal variation, are reflected in the correlogram. Thus if a model has accounted for all the serial correlation in the data, the residual series would exhibit no obvious patterns. Hence we can get the following definition.

Definition 4.1. *A time series $\{w_t : t = 1, 2, \cdots, n\}$ is discrete white noise if the variables $w_1, w_2, \cdots, w_n$ are independent and identically distributed with a mean of zero. This implies that the variables all have the same variance σ^2 and $Cor(w_i, w_j) = 0, \forall i \neq j$. If, in addition, the variables also follow a normal distribution the series is called Gaussian white noise.*

Note the second-order properties of a white noise series are assumed by the definition. Simulated white noise data will not have autocorrelations that are exactly zero because sampling variation.

Definition 4.2. *Let $\{x_t\}$ be a time series. Then $\{x_t\}$ is a random walk if*

$$x_t = x_{t-1} + w_t$$

where $\{w_t\}$ is a white noise series. Substituting $x_{t-1} = x_{t-2} + w_{t-1}$ and then substituting for x_{t-2}, followed by x_{t-3} and so on gives:

$$x_t = w_t + w_{t-1} + w_{t-2} + \cdots$$

In practice, the series above will not be infinite but will start at some time $t = 1$. Hence

$$x_t = w_1 + w_2 + \cdots + w_t$$

Back substitution is used to define more complex time series models and also to derive second-order properties.

Definition 4.3. *The backward shift operator $\mathbf{B}$ is defined by*

$$\mathbf{B}x_t = x_{t-1}$$

The backward shift operator is sometimes called the 'lag operator'. By repeatedly applying $\mathbf{B}$, it follows that

$$\mathbf{B}^n x_t = x_{t-n}$$

The second-order properties of a random walk follow as

$$\begin{cases} \mu_x = 0 \\ \gamma_k(t) = Cov(x_t, x_{t+k}) = t\sigma^2 \end{cases}$$

The covariance is a function of time, so the process is non-stationary. In particular, the variance is $t\sigma^2$ and so it increases without limit as t increases. It follows that a random walk is only suitable for short term predictions.

The time-varying autocorrelation function for $k > 0$ is:

$$\rho_k(t) = \frac{Cov(x_t, x_{t+k})}{\sqrt{Var(x_t)Var(x_{t+k})}} = -\frac{t\sigma^2}{\sqrt{t\sigma^2(t+k)\sigma^2}} = \frac{1}{\sqrt{1+k/t}}$$

so that, for large t with k considerably less than t, ρ_k is nearly 1. Hence, the correlogram for a random walk is characterized by positive autocorrelations that decay very slowly down from unity.

Definition 4.4. *The difference operator ∇ is defined by*

$$\nabla x_t = x_t - x_{t-1}$$

Note that $\nabla x_t = (1 - \mathbf{B})x_t$, so that ∇ can be expressed in terms of the backward shift operator $\mathbf{B}$. In general, higher order differencing can be expressed as

$$\nabla^n = (1 - \mathbf{B})^n$$

Note differencing can be very useful for "filtering".

A random walk with drift is a process with $x_t = x_{t-1} + \delta + w_t$. The mean is $\mu(t) = t\delta$ and the autocorrelations are the same as for a random walk without drift. The drift has no effect on variance or autocorrelation.

4.2 Model relationships of current and past values of a statistic/metric

Learning Objectives

- Estimation, data analysis and forecasting
- Forecast errors and confidence intervals

4.2.1 Autoregressive Models

Definition 4.5. *The series $\{x_t\}$ is an autoregressive process of order p, abbreviated to AR(p), if*

$$x_t = \alpha x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + w_t$$

where $\{w_t\}$ is white noise and the α_i are the model parameters with $\alpha_p \neq 0$ for order p process. Above equation can be expressed as a polynomial of order p in terms of the backward shift operator:

$$\theta_p(\mathbf{B})x_t = (1 - \alpha_1\mathbf{B} - \alpha_2\mathbf{B}^2 - \cdots - \alpha_p\mathbf{B}^p)x_t = w_t$$

The equation $\theta_p(\mathbf{B}) = 0$, where $\mathbf{B}$ is formally treated as a number, is called the characteristic equation. The roots of the characteristic equation must all exceed unity in absolute value for the process to be stationary.

Properties of an $AR(1)$ model

The $AR(1)$ process is given by

$$x_t = \alpha x_{t-1} + w_t$$

where $\{x_t\}$ is a white noise series with mean zero and variance σ^2. It can be shown that the second-order properties follow as

$$\begin{cases} \mu_x = 0 \\ \gamma_k = \alpha^k \sigma^2/(1-\alpha^2) \end{cases}$$

The autocorrelation function follows as

$$\rho_k = \alpha^k, (k \geq 0)$$

where $|\alpha| < 1$. Thus the correlogram decays to zero more rapidly for small α.

The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags. In general the partial autocorrelation at lag k is the kith coefficient of a fitted $AR(k)$ model; if the underlying process is $AR(p)$, then the coefficients α_k will be zero for all $k > p$. Thus an $AR(p)$ process has a correlogram of partial autocorrelations that is zero after lag p.

To forecast values in an $AR(p)$ model, express them in terms of the known values recursively, omitting the w_t term.

Definition 4.6. *A time series model $\{x_t\}$ is strictly stationary if the joint statistical distribution $x_{t_1}, \cdots, x_{t_n}$ is the same as the joint distribution of $x_{t_1+m}, \cdots, x_{t_n+m}$ for all $t_1, \cdots, t_n$ and m, so that the distribution is unchanged after an arbitrary time shift. Note that strict stationarity implies that the mean and variance are constant in time and that the auto-covariance $Cov(x_t, x_s)$ only depends on lag $k = |t-s|$ and can be written $\gamma(k)$. If a series is not strictly stationary but the mean and variance are constant in time and auto covariance only depends on the lag, then the series is called second-order stationary.*

4.2.2 Moving Average Models

Definition 4.7. *A moving average (MA) process of order q is a linear combination of the current white noise term and the q most recent past white noise terms and is defined by*

$$x_t = w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q} \tag{4}$$

where $\{w_t\}$ is white noise with zero mean and variance σ_w^2. Above equation can be rewritten in terms of the backward shift operator $\mathbf{B}$

$$x_t = (1 + \beta_1\mathbf{B} + \beta_2\mathbf{B} + \cdots + \beta_q\mathbf{B}^q)w_t = \phi_q(\mathbf{B})w_t$$

where ϕ_q is a polynomial of order q. Because MA process consist of a finite sum of stationary white noise terms, they are stationary and hence have a time-invariant mean and auto covariance.

The mean is zero and variance is $\sigma_w^2(1 + \beta_1^2 + \cdots + \beta_q^2)$. The autocorrelation function, for $k \geq 0$, is given by

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \sum_{i=0}^{q-k} \beta_i\beta_{i+k} / \sum_{i=0}^{q} \beta_i^2 & k = 1, \cdots, q \\ 0 & k > q \end{cases}$$

where β_0 is unity.

An MA process is invertible if it can be expressed as a stationary autoregressive process of infinite order without an error term. In general, an $MA(q)$ process is invertible when the roots of $\phi_q(\mathbf{B})$ all exceed unity in absolute value. The autocovariance function only identifies a unique $MA(q)$ process if the condition that the process be invertible is imposed.

A description of the conditional sum of squares algorithm for fitting an $MA(q)$ process follows. For any choice of parameters, the same of squared residuals can be calculated iteratively by rearranging Equation (4) and replacing the errors, w_t, with their estimates, which are denoted by $\hat{w}_t$:

$$S(\hat{\beta}_1, \cdots, \hat{\beta}_q) = \sum_{t=1}^{n} \hat{w}_t^2 = \sum_{t=1}^{n} \{x_t - (\hat{\beta}_1\hat{w}_{t-1} + \cdots + \hat{\beta}_q\hat{w}_{t-q})\}^2$$

conditional on $\hat{w}_0, \cdots, \hat{w}_{t-q}$ being taken as 0 to start the iteration. A numerical search is used to find the parameter values that minimize this conditional sum of squares.

To forecast terms in an $MA(q)$ process apply the model recursively to known residuals. $x_{n+1|n} = \beta_1\hat{w}_n + \beta_2\hat{w}_{n-1} + \cdots + \beta_q\hat{w}_{n-q+1}$.

4.2.3 Mixed Models: The ARMA process

Definition 4.8. *A time series $\{x_t\}$ follows an autoregressive moving average (ARMA) process of order (p, q), denoted $ARMA(p, q)$ when*

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + w_t + \beta_1 w_{t-1} + \beta_2 w_{t-2} + \cdots + \beta_q w_{t-q}$$

where $\{w_t\}$ is white noise. Above equation may be represented in terms of the backward shift operator $\mathbf{B}$ and rearranged in the more concise polynomial form

$$\theta_p(\mathbf{B})x_t = \phi_q(\mathbf{B})w_t$$

Parameter parsimony. When fitting to data, an ARMA model will often be more parameter efficient than a single MA or AR model.

Parameter redundancy. When θ and ϕ share a common factor, a stationary model can be simplified.

Note $\rho_k = \alpha\rho_{k-1}$ for $AMRA(1,1)$ process.

4.3 Understand Forecast Produced by ARIMA

Definition 4.9. *A series $\{x_t\}$ is integrated of order d, denoted as $I(d)$, if the dth difference of $\{x_t\}$ is white noise $\{w_t\}$. Since $\nabla^d \equiv (1-\mathbf{B})^d$, where $\mathbf{B}$ is the backward shift operator, a series $\{x_t\}$ is integrated of order d if*

$$(1-\mathbf{B})^d x_t = w_t$$

Definition 4.10. *A time series $\{x_t\}$ follows an $ARIMA(p,d,q)$ process if the dth differences of the $\{x_t\}$ series are an $ARMA(p,q)$ process. If we introduce $y_t = (1-\mathbf{B})^d x_t$, then $\theta_p(\mathbf{B})y_t = \phi_q(\mathbf{B})w_t$. We can now substitute for y_t to obtain the more succinct form of an $ARIMA(p,d,q)$ process as*

$$\theta_p(\mathbf{B})(1-\mathbf{B})^d x_t = \phi_q(\mathbf{B})w_t$$

where θ_p and ϕ_q are polynomials of orders p and q, respectively.

Definition 4.11. *A seasonal ARIMA model uses differencing at a lag equal to the number of seasons to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lag s. The seasonal $ARIMA(p,d,q)(P,D,Q)_s$ model can be most succinctly expressed using the backward shift operator*

$$\Theta_P(\mathbf{B}^s)\theta_p(\mathbf{B})(1-\mathbf{B}^s)^D(1-\mathbf{B})^d x_t = \Phi_Q(\mathbf{B}^s)\phi_q(\mathbf{B})w_t$$

where Θ_P, θ_p, Φ_Q and ϕ_q are polynomials of orders P, p, Q and q, respectively. In general, the model is non-stationary, although if $D = d = 0$ and the roots of the characteristic equation all exceed unity in absolute value, the resulting model would be stationary.