DATA ANALYTICS

Unit 5: Advanced Techniques

Swati Pratap Jagdale
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5: Confounding Variables

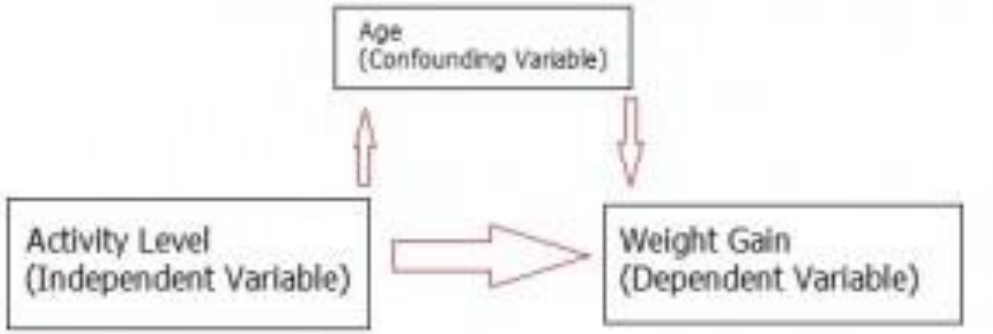
Swati Pratap Jagdale

Department of Computer Science and Engineering

Confounding Variables

- **What is a Confounding Variable?**

- A confounding variable is an “extra” variable that you didn’t account for. They can ruin an experiment and give you useless results. They can suggest there is correlation when in fact there isn’t.
- They can even introduce bias. That’s why it is important to know what one is, and how to avoid getting them into your experiment in the first place.



A confounding variable can have a hidden effect on your experiment’s outcome.

Confounding Variables

- In an experiment, the independent variable typically has an effect on your dependent variable.
- For example, if you are researching whether lack of exercise leads to weight gain, then
  lack of exercise -- independent variable
  weight gain -- dependent variable.
- Confounding variables are any other variable (like age) that also has an effect on your dependent variable (weight gain). They are like extra independent variables that are having a hidden effect on your dependent variables. A confounding variable can also be related to the independent variable (with increase in age, there may be a decrease in the duration or intensity of exercise).

- Confounding variables can cause two major problems:
  - Increase variance
  - Introduce bias.

Confounding Variables

Example:

- You test 200 volunteers (100 men and 100 women). You find that lack of exercise leads to weight gain.
- One problem with your experiment is that is lacks any control variables. For example, the use of placebos, or random assignment to groups.
- So you really can't say for sure whether lack of exercise leads to weight gain. One confounding variable is how much people eat. It's also possible that men eat more than women; this could also make sex a confounding variable.

- A poor study design like this could lead to bias.
- For example, if all of the women in the study were middle-aged, and all of the men were aged 16, age would have a direct effect on weight gain. That makes age a confounding variable.

Confounding Bias

Confounding Bias

- Bias is usually a result of errors in data collection or measurement.
- However, one definition of bias is "***…the tendency of a statistic to overestimate or underestimate a parameter***", so in this sense, confounding is a type of bias.

- Confounding bias is the result of having confounding variables in your model. It has a direction, depending on if it over- or underestimates the effects of your model:
- Positive confounding is when the observed association is biased away from the null. In other words, it overestimates the effect.
- Negative confounding is when the observed association is biased toward the null. In other words, it underestimates the effect.

Confounding Bias

How to Reduce Confounding Variables?

- Make sure you identify all of the possible confounding variables in your study.
- Make a list of everything you can think of and one by one, consider whether those listed items might influence the outcome of your study. Usually, someone has done a similar study before you. So check the academic databases for ideas about what to include on your list.

- Once you have figured out the variables, techniques to reduce the effect of those confounding variables:
  - Bias can be eliminated with random samples.
  - Introduce control variables to control for confounding variables. For example, you could control for age by only measuring 30 year olds.
  - Within subjects designs test the same subjects each time. Anything could happen to the test subject in the “between” period so this doesn’t make for perfect immunity from confounding variables.
  - Counterbalancing can be used if you have paired designs. In counterbalancing, half of the group is measured under condition 1 and half is measured under condition 2.

Terminology and identifying a confounding variable

Synonyms for Confounding Variables and Omitted Variable Bias

- Confounding variables or confounders, and lurking variables.

- A confounding variable is closely related to both the independent and dependent variables in a study. An independent variable represents the suppose *cause*, while the dependent variable is the supposed *effect*.
- A confounding variable is a third variable that influences both the independent and dependent variables. Failing to account for confounding variables can cause you to wrongly estimate the relationship between your independent and dependent variables.

- How do we identify a confounding variable?
  1. There must be three or more variables in the study
     two variables => one is the cause the other is the effect)
  2. The variable we suspect is a confounding variable, changes systematically with at least one of the variables we are measuring (either independent or dependent)
  3. Identify extraneous variables that relate to subjects (age, gender, etc.), the environment in which the study is conducted (weather, location, etc.) and to the two variables we are explicitly measuring to test for systematic changes to identify a confounding variable.

http://dissertation.laerd.com/extraneous-and-confounding-variables-p3.php

Hidden variable vs confounding variable

A hidden variable could be connecting two variables that are spuriously correlated (temperature/ season that connects the two spuriously correlated variables: ice-cream sales and number of robberies)

A confounding variable is related to two variables that are not spuriously correlated (hypothyroidism -> causes increase in weight due to lower metabolic rate; but lower metabolic rate implies lower energy levels; this lowers the ability of one to exercise which in turn leads to increase in weight)

Lack of exercise -> increase in weight;
the two are connected not spuriously

Confounding Bias

Omitting confounding variables from your regression model can bias the coefficient estimates.

- When you are assessing the effects of the independent variables in the regression output, this bias can produce the following problems:
  - Overestimate the strength of an effect.
  - Underestimate the strength of an effect.
  - Change the sign of an effect.
  - Mask an effect that actually exist

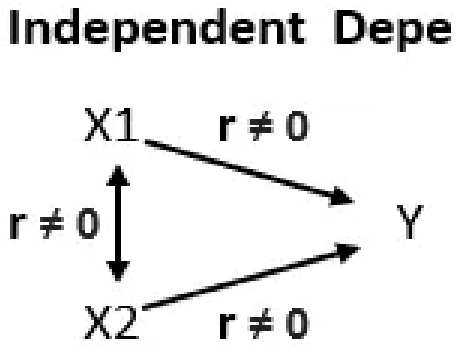
What Conditions Cause Omitted Variable Bias?

- How does this bias occur? How can variables you leave out of the model affect the variables that you include in the model?

- For omitted variable bias to occur, the following two conditions must exist:
- The omitted variable must correlate with the dependent variable.
- The omitted variable must correlate with at least one independent variable that is in the regression model.

Confounding Bias

- There must be non-zero correlations (r) on all three sides of the triangle.
- This correlation structure causes confounding variables that are not in the model to bias the estimates that appear in your regression results. For example, removing either X variable will bias the other X variable.



- The amount of bias depends on the strength of these correlations.
- Strong correlations produce greater bias.
- If the relationships are weak, the bias might not be severe.
- And, if the omitted variable is not correlated with another independent variable at all, excluding it does not produce bias.

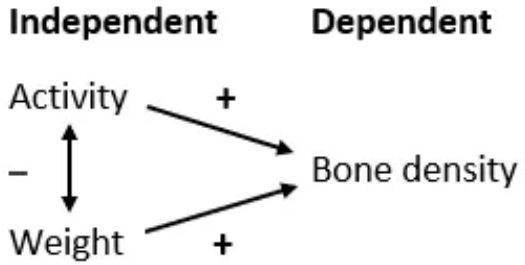
Confounding Bias

- **Example of How Confounding Variables Can Produce Bias**

Example:

- In a biomechanics lab, One study assessed the effects of physical activity on bone density.
- They measured various characteristics including the subjects' activity levels, their weights, and bone densities among many others.
- Theories about how our bodies build bone suggest that there should be a positive correlation between activity level and bone density. In other words, higher activity produces greater bone density.
- Simple regression analysis to determine whether there is a relationship between activity and bone density... there was no relationship at all!

- They included activity level as the only independent variable, but it turns out there is another variable that correlates with both activity and bone density—the subject's weight.

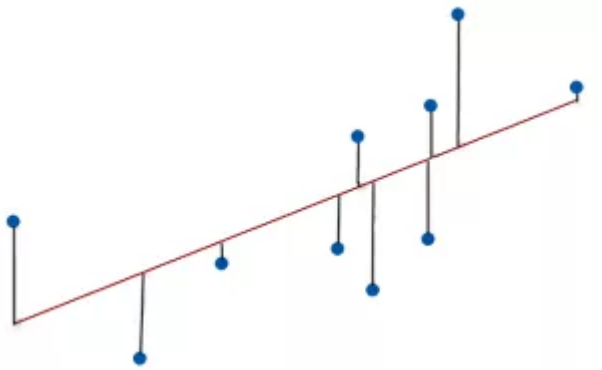


The diagram shows the signs of the correlations between the variables.

Confounding Bias

- Correlations, Residuals, and OLS Assumptions

- When you satisfy the ordinary least squares (OLS) assumptions, the Gauss-Markov theorem states that your estimates will be unbiased and have minimum variance.



Residuals = Observed value – Fitted value

Omitted variable bias occurs because the residuals violate one of the assumptions.

Confounding Bias

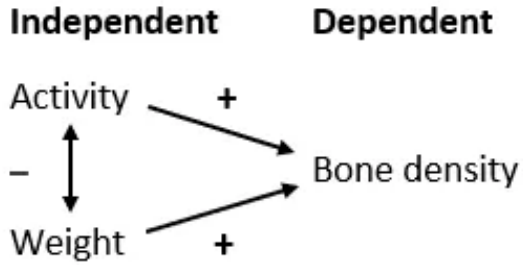
- Consider, regression model with two significant independent variables, X1 and X2. These independent variables correlate with each other and the dependent variable—which are the requirements for omitted variable bias.

- Now, imagine that we take variable X2 out of the model. It is the confounding variable. Here's what happens:

- The model fits the data less well because we have removed a significant explanatory variable. Consequently, the gap between the observed values and the fitted values increases. These gaps are the residuals.
- The degree to which each residual increases depends on the relationship between X2 and the dependent variable. Consequently, the residuals correlate with X2.
- X1 correlates with X2, and X2 correlates with the residuals. Ergo, variable X1 correlates with the residuals.
- This condition violates the ordinary least squares assumption that independent variables in the model do not correlate with the residuals. Violations of this assumption produce biased estimates.

Confounding Bias

	Included and Omitted: Negative Correlation	Included and Omitted: Positive Correlation
Included and Dependent: Negative Correlation	Positive bias: coefficient is overestimated.	Negative bias: coefficient is underestimated.
Included and Dependent: Positive Correlation	Negative bias: coefficient is underestimated.	Positive bias: coefficient is overestimated.



The table summarizes these relationships and the direction of bias.

Included (activity) and omitted (weight) are negatively correlated.
The included variable (weight) and the dependent variable (bone density) have a positive relationship, implies the result has a negative bias.

References

https://www.statisticshowto.com/experimental-design/confounding-variable/

https://statisticsbyjim.com/regression/confounding-variables-bias/

THANK YOU

Swati Pratap Jagdale
Department of Computer Science
swatigambhire@pes.edu

DATA ANALYTICS

Unit 5:Introduction to Stochastic models and Markov processes (first order)

Bharathi R

Department of Computer Science and Engineering

Introduction Stochastic Process

- Stochastic models are powerful tools which can be used for solving problems which are dynamic in nature, that is, the values of the random variables change with time.

- Stochastic process is defined as a collection of random variables
  $\{X_n, n \geq 0\}$ indexed by time (however, index can be other than time).

- The value (cash flow) that the random variable X_n can take is called the **state of the stochastic process at time n**.

- The set of all possible values the random variable can take is called the **state space**.

1.Poisson Process: Examples

Generally we would like to count the number of events that occur over a period of time. Following are few examples of counting process:

1. Retail stores would like to predict footfall (number of customers visiting the store) over a period
of time.
2. Call centres would like to predict the number of calls they receive over a period of time.
3. Number of customer arrivals at banks, airports, restaurants, and any service centres.
4. Demand for spare parts of capital equipment caused due to failure of parts over a period of time.
5. Number of insurance claims received at an insurance company.

1.Poisson Process

Homogeneous Poisson Process (HPP) is a stochastic counting process $N(t)$ with the following properties:

$N(0) = 0$, that is the number of events by time t = 0 is zero.

$N(t)$ has independent increments. That is if $t_0 < t_1 < t_2 < \ldots < < t_n$, then $N(t_1) - N(t_0)$, $N(t_2) - N(t_1)$, …, $N(t_n) - N(t_{n-1})$ are independent.

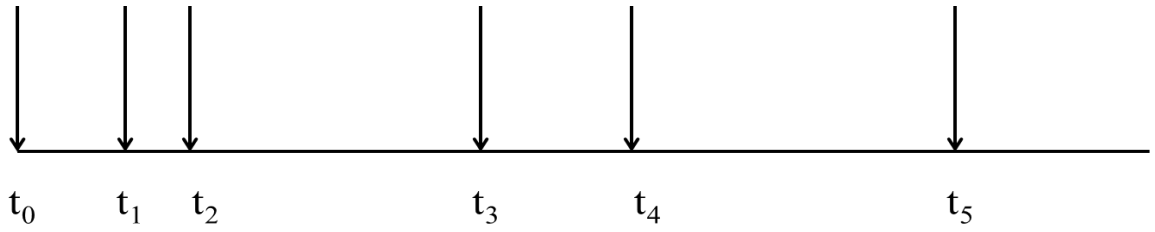


Figure above shows a Poisson process of events in which $(t_1 - t_0)$, $(t_2 - t_1)$, $(t_3 - t_2)$ are time between events.

1. Poisson Process

The number of events by time t, N(t), follows a Poisson distribution, that is

$$P[N(t)=n]=\frac{e^{-\lambda t}\times(\lambda t)^{n}}{n!}$$

Cumulative distribution of number of events by time t in a Poisson process is given by

$$P[N(t)\le n]=\sum_{i=0}^{n}P[N(t)=i]=\sum_{i=0}^{n}\frac{e^{-\lambda t}\times(\lambda t)^{i}}{i!}$$

The mean, $E[N(t)]$, and variance, $\mathrm{Var}[N(t)]$, of a Poisson process $N(t)$ are given by

$E[N(t)] = \lambda t$
$\mathrm{Var}[N(t)] = \lambda t$

In the case of Poisson process, the time between events follows an exponential distribution with parameter λ, that is the time between events have a density function $f(t) = \lambda e^{-\lambda t}$ and cumulative distribution function $F(t) = 1 - e^{-\lambda t}$.

DATA ANALYTICS

Poisson Process: Example

Johny Sparewala (JS) is a supplier of aircraft flight control system spares based out of Mumbai, India. The demand for hydraulic pumps used in the flight control system follows a Poisson process. Sample data (50 cases) on time between demands (measured in number of days) for hydraulic pumps are shown in Table 16.1

TABLE 16.1 Time between demands (in days) for hydraulic pumps

104	90	45	32	12	6	30	23	58	118
80	12	216	71	29	188	15	88	88	94
63	125	108	42	77	65	18	25	30	16
92	114	151	10	26	182	175	189	14	11
83	418	21	19	73	31	175	14	226	8

Example Continued

(a) Calculate the expected number of demand for hydraulic pump spares for next two years.

(b) Johny Sparewala would like to ensure that the demand for spares over next two years is met in at least 90% of the cases from the spares stocked (called fill rate) since lead time to manufacture a part is more than 2 years. Calculate the inventory of spares that would give at least 90% fill rate.

Solution

(a) To calculate the expected number of demand for spares for two years, we have to estimate the parameter λ of the Poisson distribution. The maximum likelihood estimate of λ is given by

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = 0.0125$$

where X_i is the time between failure of i^{th} case and $\frac{1}{n}\sum_{i=1} X_i$ is the mean time between failure.

The expected number of demand for spares, $E[N(t)]$, for 2 years (2×365 days) is given by

$$E[N(t)] = E[N(2\times 365)] = \hat{\lambda} \times t = 0.0125 \times 2 \times 365 = 9.125$$

Solution

(b) To ensure that the demand for spares is met 90% of the time, we have to calculate smallest k such that

$$\sum_{i=0}^{k} \frac{e^{-\hat{\lambda}t} \times (\hat{\lambda}t)^i}{i!} \geq 0.90$$

Table 16.2 shows density and cumulative distribution function values of Poisson process for different values of k.

TABLE 16.2 Poisson density and distribution function for different values of k

k	Poisson Density	Cumulative	k	Poisson Density	Cumulative
0	0.0001	0.0001	11	0.0996	0.7907
1	0.0010	0.0011	12	0.0758	0.8665
2	0.0045	0.0056	13	0.0532	0.9197
3	0.0138	0.0194	14	0.0347	0.9543
4	0.0315	0.0509	15	0.0211	0.9754
5	0.0574	0.1083	16	0.0120	0.9875
6	0.0873	0.1956	17	0.0065	0.9939
7	0.1138	0.3095	18	0.0033	0.9972
8	0.1298	0.4393	19	0.0016	0.9988
9	0.1316	0.5709	20	0.0007	0.9995
10	0.1201	0.6911	21	0.0003	0.9998

Solution

Smallest value of k for which the cumulative probability is greater than 0.90 is 13. That is, JS should stock 13 spares to ensure that they meet demand for spares in 90% of the cases over a two-year period. The probability density function of Poisson distribution with mean 9.125 is shown in Figure 16.2.

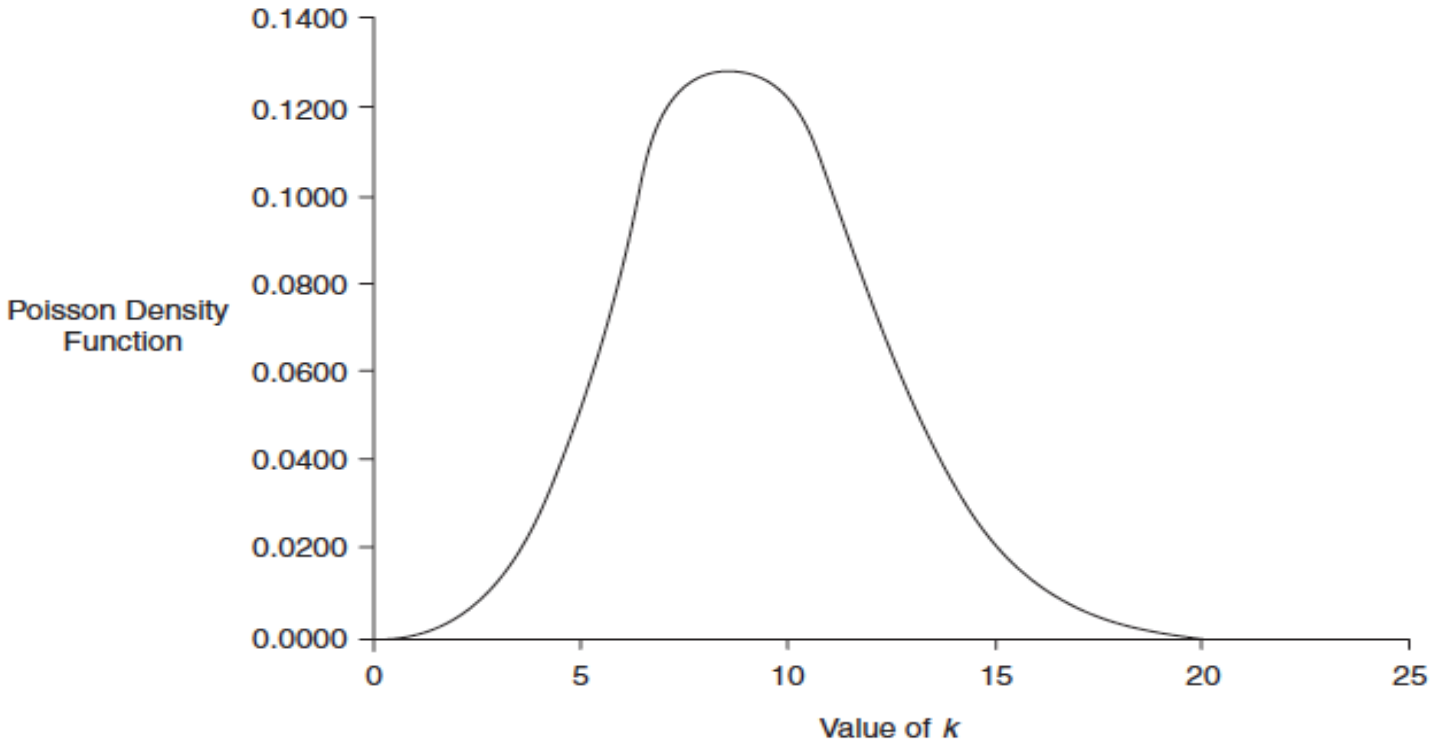


FIGURE 16.2 Poisson process density function.

2. Compound Poisson Process

Compound Poisson process is a stochastic process $X(t)$ where the arrival of events follows a Poisson process and each arrival is associated with another independent and identically distributed random variable Y_i.

Compound Poisson process $X(t)$ is a continuous-time stochastic process defined as

$$X(t) = \sum_{k=1}^{N(t)} Y_k$$

where $N(t)$ is a Poisson process with mean λt and Y_i are independent and identically distributed random variables with mean $E(Y_i)$ and variance $\text{Var}(Y_i)$.

The mean and variance of the compound Poisson process $X(t)$ are given by (Ross, 2010)

$$E[X(t)] = \mu_{X(t)} = \lambda t \times E(Y_i) \tag{16.6}$$

$$\text{Var}[X(t)] = \sigma^2_{X(t)} = \lambda t \times E(Y_i^2) = \lambda t \times (VAR(Y_i) + [E(Y_i)]^2) \tag{16.7}$$

For large t, we can show that the compound Poisson process follows an approximate normal distribution with mean $\mu_{X(t)}$ and standard deviation $\sigma_{X(t)}$.

Compound Poisson Process: Example

Customers arrive at an average rate of 12 per hour to withdraw money from an ATM and the arrivals follow a Poisson process. The money withdrawn are independent and identically distributed with mean and variance INR 4200 and 2,50,000, respectively. If the ATM has INR 6,00,000 cash, what is the probability that it will run out of cash in 10 hours?

Solution

Solution:

The mean and standard deviation of the compound Poisson process $X(t)$ can be calculated as described below:

Mean of compound Poisson process is

$$\mu_{X(t)} = \lambda t \times E(Y_i) = 12 \times 10 \times 4200 = 5,04,000$$

Variance of compound Poisson process is

$$\sigma^2_{X(t)} = \lambda t \times (\text{Var}(Y_i) + [E(Y_i)]^2) = 12 \times 10 \times (250000 + 4200^2) = 21468 \times 10^5$$

Standard deviation of compound Poisson process is

$$\sigma_{X(t)} = \sqrt{\sigma^2_{X(t)}} = \sqrt{21468 \times 10^5} = 46333.57$$

Probability that the cash withdrawal will exceed INR 6,00,000 is given by

$$P(X(t) \geq 6,00,000) = P\left(Z \geq \frac{6,00,000 - 504000}{46333.57}\right) = P(Z \geq 2.0719) = 0.0191$$

That is, there is approximately 2% chance that the ATM will run out of cash in 10 hours.

3. Markov Chains

The condition $P\left[X_{n+1}=j \mid X_0=i_0, X_1=i_1,\ldots, X_n=i\right]=P\left[X_{n+1}=j \mid X_n=i\right]$ is called Markov property named after the Russian mathematician *A A Markov*.

If the state space *S* is discrete then the stochastic process

$$\{X_n,\ n = 0, 1, 2, \ldots\}$$

that satisfies the condition is called a Markov chain

One Step Transition Probabilities of Markov Chains

Let $\{X_n,\ n = 0, 1, 2, \ldots\}$ be a Markov chain with state space *S*. Then the conditional probability

$$P\left[X_{n+1}=j \mid X_0=i_0, X_1=i_1,\ldots, X_n=i\right]=P\left[X_{n+1}=j \mid X_n=i\right]=P_{ij}$$

is called the one-step transition probability. P_{ij} gives conditional probability of moving from state *i* to stage *j* in one period.

One-Step Transition Probabilities of Markov Chain

P_{ij} gives conditional probability of moving from state i to stage j in one period.
One-step transition probabilities between all states in the state space are expressed in the form of one-step transition probability matrix as shown below

$$\mathbf{P} = \mathbf{P_{ij}} =$$

	1	2	…	n
1	P_{11}	P_{12}	…	P_{1n}
2	P_{21}	P_{22}	…	P_{2n}
…	…	…	…	…
n	P_{n1}	P_{n2}	…	P_{nn}

m-step transition probability

An *m*-step transition probability in a Markov chain is given by

$$P_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i)$$

The m-step transition probability $P_{ij}^{(m)}$ can be written as

$$P_{ij}^{(m)} = \sum_{r=1}^{n} P_{ir}^{k} \times P_{rj}^{(m-k)}, \quad 0 < k < m$$

Estimation of One-Step Transition Probabilities of Markov Chain

Transition probabilities of a Markov chain are estimated using maximum likelihood estimate (MLE) from the transition data (Anderson and Goodman, 1957).

The MLE estimate of the transition probability P_{ij} (probability of moving from state i to state j in one step) is given by

$$\hat{P}_{ij} = \frac{N_{ij}}{\sum_{k=1}^{m} N_{ik}}$$

where N_{ij} is number of cases in which $X_n = i$ (state at time n is i) and $X_{n+1} = j$ (state at time $n + 1$ is j).

Hypothesis Tests for Markov Chain: Anderson Goodman Test

The null and alternative hypotheses to check whether the sequence of random variables follows a Markov chain is stated below

H_0: The sequences of transitions (X_1, X_2, …, X_n) are independent (zero-order Markov chain)

H_A: The sequences of transitions (X_1, X_2, …, X_n) are dependent (first-order Markov chain)

The corresponding test statistic is

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right)$$

where

O_{ij} = Observed number of transitions from state i to state j in one period.

E_{ij} = Expected number of transitions from state i to state j assuming independence.

DATA ANALYTICS

Testing Time Homogeneity of Transition Matrices: Likelihood Ratio Test

Anderson and Goodman (1957) suggested a likelihood ratio test for checking whether the transition probability matrices are time homogeneous. The null and alternative hypotheses of the likelihood ratio tests are

$$H_0: P_{ij}(t) = P_{ij},\ t = 1, 2, 3, 4, \text{and } 5$$
$$H_A: P_{ij}(t) \neq P_{ij},\ t = 1, 2, 3, 4, \text{and } 5$$

The test statistic is a likelihood test ratio statistic and is given by (Anderson and Goodman, 1957):

$$\lambda = \prod_t \prod_{i,j} \left[\frac{\hat{P}_{ij}}{\hat{P}_{ij}(t)} \right]^{n_{ij}(t)}$$

Summary

1.Most problems in analytics are dynamic in nature and thus require collection of random variables to model the problem.

2. Stochastic process is a collection of random variables usually indexed by time t and used while modelling problems that are not independent and identically distributed.

3. Poisson process is a counting process that is used in decision-making scenarios such as capacity planning and spare parts demand forecasting. Compound Poisson process can be used to study problems such as cash replenishments at ATMs, total insurance claims, etc.

4. Markov chain is one of the most powerful models in analytics with applications across industry sectors. Google's PageRank algorithm is based on Markov chain.

5. Asset availability, market share, customer retention probability, and customer lifetime value are few applications of Markov chain in analytics.

References

Text Book:

Chapter 16.1 -16.4.4 “Business Analytics,

The Science of Data-Driven Decision Making”, by U. Dinesh Kumar, Wiley 2017