**Poisson Regression**

Version of Thursday, June 01, 2017

Farrokh Alemi, PhD1

Amr Elrafey1

Larry Tang PhD2

1George Mason University, Department of Health Administration and Policy

2George Mason University, Department of Statistics

**Introduction**

In previous chapters we have introduced two different type of regression. Standard regression could be analyzed to find the relationship of continuous variables such as cost to several predictors. Logistic regression could be used to analyze the relationship between a binary variable and several predictors. In this chapter we introduce Poisson regression, which can be used to examine the relationship between count data and several predictors. By count data, we mean the number of times, during a defined time period, in which variables co-occur. Poisson regression is used in several different types of problems commonly faced by health care managers, including:

1. Jacobs et al (2012) studied how introduction of managed care affected use of radiotherapy [[[1]](#endnote-2)]. They counted number of men diagnosed with prostate cancer and treated with radiotherapy. They regressed this count on a number of predictors including participation in managed care to see if managed care influenced the number of patients receiving radiotherapy for prostate cancer. A similar analysis was done for understanding the effect of managed care on adoption of robotic prostatectomy [[[2]](#endnote-3)]. These type of application of Poisson regression may be taught in courses on health administration courses on managed care.
2. Poisson regression could be used to examine changes in adverse events. For example, Gibbons and colleagues (2008) studied how count of suicides during a specific time period could be predicted from use of antidepressants and other predictors [[[3]](#endnote-4)]. Similar analysis could be made for the count of any adverse event and various predictors. The use of Poisson regression for count of adverse events is especially appropriate when adverse events are rare, thereby the assumption of Poisson distribution of the count data is met. These type of application of Poisson regression may be discussed in health administration courses on quality of care.
3. Poisson regression was used to examine the success of social media campaigns [[[4]](#endnote-5)]. In these types of analysis one regresses the count of patients who benefitted from the campaign on several predictors. For example, one might count the number of new patients who show at a clinic after a combination of media campaigns and see which type of campaign led to more new patients. These type of Poisson regressions may be discussed in health administration courses on marketing.
4. Poisson regression can be used to see the factors that affect count of patients turned away from hospitals without care [[[5]](#endnote-6)]. The study could examine if the number of patient denied care was related to severity of the patient’s illness, availability of insurance payments, or other factors. These types of studies could be used to examine non-competitive market behavior such as focusing care on most profitable patients. These types of applications of Poisson regression may be included in health administration courses on strategic planning.

**Poisson distribution**

Poisson distribution gives us a sense of variation that occurs around the count of an event in a fixed time period. Typically events are counted over different time periods. The mean of these counts are assumed to be constant and is shown by the Greek letter lambda: λ. The density function for the count of occurrence of the event can be calculated as:

In this equation, k is the count of the event, is k factorial and is calculated as k\*(k-1)\*(k-2)\* …\*2\*1, is the average of the count of the event across time, and e is a constant that is approximately 2.71828.

The Poisson distribution assumes that:

1. The target event either occurs or does not, only two situations are possible and these situations are mutually exclusive and exhaustive. It is not possible to have the event both occur and not occur; it is also not possible that neither the event occurs nor it does not.
2. The count of the occurrences of the event, is calculated over a specified time period and is relatively small. The probability that the event occurs is proportional of the length of time we are looking for the event. Furthermore, it is reasonable to assume that in an extremely small time period, a fleeting moment, the probability of observing the event is zero. Another way of saying the same thing is that that the probability of more than one event occurring in a short period of time is negligible.
3. The rate of occurrence of the events is constant across trials. The probability of an event occurring in a small time interval does not depend on the length of time since the previous event. If the prior occurrence of an event affects its future occurrence, e.g. when hospitalization raises the probability of subsequent hospitalization, then this assumption is violated.

In the definition of Poisson distribution, we have designed the sample trials so that each time period indicates one trial. We are counting the number of the event in each trial, in this case in each time period. It is also possible to use other measures of trials in the sample, including region/area in which one is looking for the event or fixed samples of a population.

An example of Poisson distribution is the count of adverse events that occur on weekly basis. Suppose that on average we have 1 event occurring weekly. The probability that next week we would have no adverse events is given by:

As number of trials increases and the probability of occurrence drops, Poisson distribution approximates Binomial distribution. Typically a Poisson distribution can be used instead of Binomial distribution if the number of trials (time periods, visits, regions, etc.) is large and the sentinel event has a low probability of occurrence during the trial. In Poisson distribution, the mean and the variance of the distribution are the same number, making it relatively easy to calculate the Poisson density function.

For another example, assume that wrong side surgeries has occurred once in the last 5 years, we now want to know what is the chances that we will not see a wrong side surgery next month. As discussed under the Bernoulli density function in an earlier chapter, the daily probability of wrong side surgery is given as follows:

You can also get to the same probability by using the definition of probability of an event as the number of occurrences of the event divided by the number of samples. There is 1 day in which wrong side surgery has occurred and 5 times 365 days in which we have sampled to see if wrong side surgery might occur. In order to use Poisson distribution, we need to turn this probability into count of wrong side surgery during a month. Since we have 30 days in a month, the total number of wrong side surgeries we expect in a month is 30 times the daily probability yielding 0.016 events. Note that the expected number of wrong side surgeries in a month is lower than 1. Now we can use Poisson distribution to estimate the probability that no wrong side surgery will occur next month:

We are almost certain that no wrong side surgery will occur next month.

Let us try a different example. Suppose that review of incidence reports in the past 3 years has identified 2 types of unauthorized disclosure of patient information: disgruntled employee selling information (done May 11, 2004, and November 22, 2007), and clinician discussing patient information in social settings (done December 5, 2006, and December 27, 2007). If the monthly incidence report indicates a new unauthorized disclosure, what is the most likely cause? The first step is to calculate the time to each type of unauthorized disclosures: 1290 days for disgruntled employees selling information and 387 days for clinicians discussing patient information in social gatherings. This provides an estimated daily probability of the event. Note that this estimate is different than dividing the number of events by 3 years. Such a division would have resulted in the same rates for the 2 causes while one has a 3 times more time between its reoccurrences. The way we have estimated the daily rate reflects the time between events and not just the number of events. This leads to the probability of disclosure due to socialization to be 3 times higher than its probability due to employees selling information.

Once the daily rate has been established, we estimate the daily number of unauthorized disclosures and use this in calculating the probability of no-unauthorized disclosure in next month. The monthly hazard rate is estimated as 1 − (probability of no-unauthorized disclosure). The relative contribution of each hazard rate is calculated as the hazard rate divided by the sum of hazard rates from any source. Here are the calculations for attributable risk (AR), associated with a disgruntled employee selling information:

The socialization of clinicians is much more likely to lead to unauthorized disclosure than disgruntled employee selling the information. This type of information can be used to set priority of where risk reduction efforts should focus on.

**Poisson Regression**

In Poisson regression the variation in count of an event is explain as a function of a number of independent variables [[[6]](#endnote-7)]. The response variable, the variable we are predicting, is the count of the event in a particular time period or number of trials. For example, we might want to count the number of dissatisfied patients in visits to our clinic and we may wish to know if this count is affected by age, gender, diagnoses of the patients. Poisson regression assumes that the dependent variable has a Poisson distribution in which the expected value and standard deviations are equal. If **C**{\displaystyle \mathbf {x} \in \mathbb {R} ^{n}} is a vector of independent variables, T the treatment variable, then the Poisson Regression of outcome Y on the independent variables takes the form:

{\displaystyle \log(\operatorname {E} (Y\mid \mathbf {x} ))=\alpha +\mathbf {\beta } '\mathbf {x} ,}

In Poisson regression, associations among any pair of variables are identified through regressing log of count of combination of variables on main and pair-wise interactions of the variables. If a pair of variables are associated with each other, one would expect a statistically significant relationship for the interaction term.

Poisson regression can be used to describe the data as a network structure. Agresti [[[7]](#endnote-8) ] showed how hierarchical Poisson regression can be used to identify the associations among the variables. First, he found the best fit to data by progressively removing pair-wise interaction terms and re-examined the model’s goodness of fit to the data. A network was drawn by creating a node for each variable. For every association that remained in the best fit model, whether statistically significant or not, Agresti drew an arc between the nodes.

An alternative approach, one that we prefer, is to map to the network only statistically significant pair-wise associations that have causal implications. If we draw an arc between every pair of independent variables whose interaction predicts the response variable, then we have a network structure describing the data. In this chapter we show how to learn a network structure from complex data using Poisson regression.

**A Simulated Example**

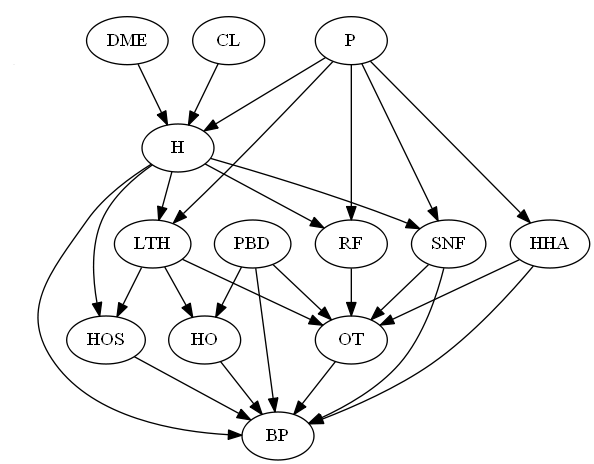
To demonstrate how Poisson Regression can be used to construct a network model, we simulated data based on the network in Figure 1. This network shows the relationship among various sources of cost overrun and whether bundle-payment of the Center for Medicare and Medicaid (CMS) was above or below the total cost. At end of the year, CMS reduces payments made to a hospital for treatment of hip fracture (DRG 469 or 470) and subsequent 90 days post-discharge costs, if these costs were higher than average. The following costs are included:

1. Physicians' services, P
2. Inpatient hospital services, H
3. Long-term care hospital services, LTH
4. Inpatient rehabilitation facility services, RF
5. Skilled nursing facility services, SNF
6. Home health agency services, HHA
7. Hospital outpatient services, HO
8. Outpatient therapy services, OT
9. Clinical laboratory services, CL
10. Durable medical equipment, DME
11. Part B drugs, PBD
12. Hospice, HO
13. Bundled payments, BP

The Centers for Medicare and Medicaid excludes from the calculation of bundled payments unrelated costs, e.g. if the person was hospitalized for cancer, the cost would be excluded. Despite efforts to avoid unrelated costs, the management of services to patients paid in bundled amounts is often difficult. Hospital administrators have to improve the efficiency and quality of their own operations as well care at other organization so that the hospital does not lose funds as a consequence of bundled payments. For a hospital administrator to do so, he needs to understand causes of cost overrun, which could be a function of either pricing of the service, volume of the service, or inappropriate quality of care in any of the services. The network in Figure 1 shows a possible situation where a hospital administrator might find himself, where cost overruns in one area have led to total cost exceeding bundled payments. In these circumstances, Center for Medicare and Medicaid would reduce reimbursement to the hospital creating a financial incentive to find the causes of cost overrun and improve reimbursement.

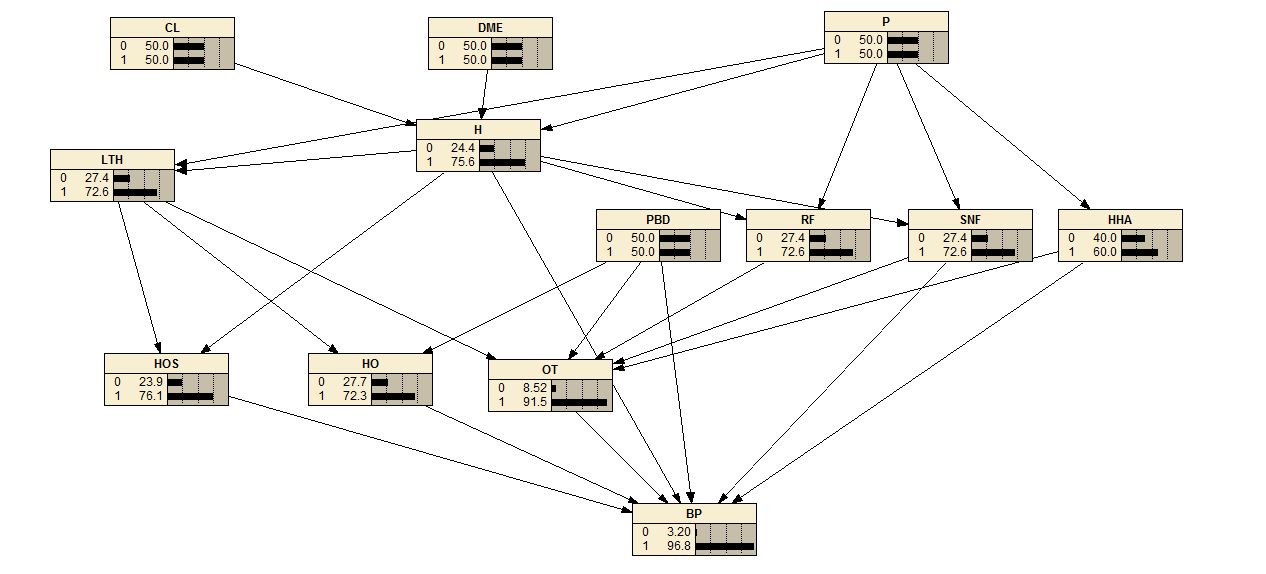
**Figure 1: Network Used to Simulate Total Cost Exceeding CMS’s Bundled Payment**Nodes show cost overrun for CL = Clinical Laboratory, DME = Durable Medical Equipment, H = Hospital, P = Physician, LTH = Long Term Hospital, RF = Rehab Facility, SNF = Skilled Nursing Facility, HHA = Home Health Agency, HO = Hospital Outpatient, OT = Outpatient Therapy, PBD = Part B Drugs, HOS = Hospice,

BP = Bundle Payment



The arcs in the network show causes of cost exceeding bundled payments. Thus cost overrun in one service is shown to affect the probability of cost overrun in total 90 days. The hospital administrator, of course, is not aware of these causal relationships and must discover these relationships from reports on individual patients. In essence, he has to discover the causal network from the cases. We simulated 10,000 cases from network in Figure 1 using the Netica software from Norsys. In order to reduce errors due to Power of the test to detect the effect we made sure that each cause either more than doubled or reduced by more than ½ the probability of the effect. The data used in this analysis are provided online [[[8]](#endnote-9)]. The Netica network used to generate the simulated data is provided below along with the probabilities of each of the nodes.

**Figure 2: Netica Network Used to Generate the Data**



**Checking Assumptions of Poisson Regression**

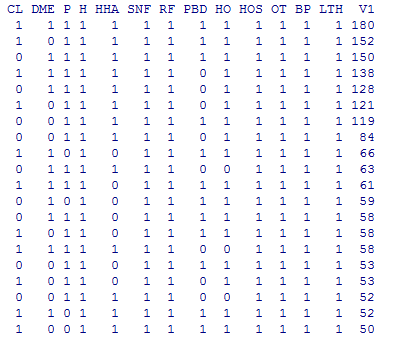
We begin the analysis of the data by checking the assumptions of Poisson regression in our data. Since our data is counts of co-occurrences of events, the assumptions are likely to have been met. Nevertheless, we need to verify that it is so. To accomplish this task we used the function ddply() from the R package Plyr (version 1.8.4) as follows:

|  |
| --- |
|  |

A total of 2,243 combinations of the 13 variables in the simulation were identified. Among these, 1,037 combinations occurred only once in the database. Many (67%) of the theoretically possible combinations of 13 binary variables never occurred. Table 1 provides the top 20 combinations with highest counts. These data suggest that a Poisson process could have generated the count of combinations as the probability of any one of the combinations repeating is exceedingly small, less than 0.01.

**Table 1: 10 Combination of Cost Overruns in Different Services**Cost overruns for CL = Clinical Laboratory, DME = Durable Medical Equipment, H = Hospital, P = Physician, LTH = Long Term Hospital, RF = Rehab Facility, SNF = Skilled Nursing Facility, HHA = Home Health Agency, HO = Hospital Outpatient, OT = Outpatient Therapy, PBD = Part B Drugs, HOS = Hospice,

BP = Bundle Payment



**Analysis of Data Using Poisson Regression**

Given our data, we used a Poisson regression to find all association in the data that predict response variable. Here is the R code for analyzing the data.

|  |
| --- |
| Here is part of the result of the regression in tabular format: |

The Poisson regression identifies a large number of associations that were not in the original network. This is problematic and suggests that Poisson regression may be picking up spurious and non-causal associations.

**Problems in Learning Networks from Poisson Regression**

A glance at our findings should convince us that Poisson regression identifies many spurious associations not in the original data. This result is troubling. A solution can be found if we understand why these spurious associations show up in our findings. Recall that we generated the data using a directed network structure that could be interpreted as representing causes and effects. Every directed arc in the network that was used to generate the data may be considered a cause of change in the variables that followed it. There are at least two ways that Poisson regression may identify non-causal associations: (1) stratifying common effects, and (2) showing correlations due to common causes. A common effect is when two or more causes have the same effect, e.g. in the network X🡪Y🡨Z, X and Z have the common effect Y. If multiple causes have a common effect, then stratifying the common effect will create an association among the causes [[[9]](#endnote-10),[[10]](#endnote-11)]. In our example, stratifying Y will create a correlation among X and Z. Since Poisson regression detects a relationship between two variables by stratifying all other variables it is likely that it will also stratify a common effect and report a non-causal association.

The second type of non-causal associations are due to common causes. A common cause refers to the situation where one variable affects two or more other variables. For example, in this network X🡨Y🡪Z, Y is the common cause of X and Z. A common cause is likely to lead to co-variation among its effects (X and Z are likely to be correlated). Poisson regression may indicate an association even though there is no causal link in the original network that generated the data.

**Sequence of Data**

One way to make Poisson regression more accurate is to use the sequence with which independent variables occur to limit the number of variables in Poisson regression. In smaller Poisson regression, it is less likely that the identified association is non-causal and an artifact of other variables in the analysis. We mentioned earlier that there are two ways that spurious association may show up in Poisson regression: through stratifying common effects and through common source of variability. To help avoid spurious correlations due to stratifying common effects, one can go through a stepwise procedure for Poisson regression. Since all causes occur in time period prior to the effects; then a time-based, stepwise, procedure could avoid discovery of non-causal association due to stratifying of common effects. Variables at time period t are examined first, statistically significant interaction terms are kept, and all other interactions are discarded; thus, preventing these interactions from becoming significant in models constructed in later time periods.

Sequence of events can also be used to prevent non-causal associations due to shared common source of variability. Once the variables that are associated with a variable are identified, then we can, a concept called Markov Blanket [[[11]](#endnote-12)], then we can repeat the Poisson regression with this smaller set and verify that none of the identified variations are due to some common source of variability elsewhere in the network. In particular, a smaller Poisson regression is done with all variables in the Markov Blanket of the effect. If the association remains significant, then it is likely to be causal and real. If not, then the interaction term is ignored in the subsequent time periods and the association is not shown as an arc in the discovered network.

Our solution for how to make Poisson regression more accurate relies on knowing the sequence with which various events occur. In electronic health records, the sequence among the data can be easily discovered using several different methods. Sometimes, the definition of the variables tell us that one variable is measured before another. For example, demographic variables are measured before other covariates since they are features that the patient acquired on birth. For another example, outcomes are usually measured after other covariates or treatment and not before.

Sometimes, the timing of data collection indicates the sequence of the variables. Thus, we may learn sequence among diagnoses of a patient by examining the timing of these events. Still other times, timing of events can be deduced through age at which these events typically occur. Thus, one may deduce that cardiovascular events usually occur in 60s and 70s while Alzheimer occurs in 80s and 90s. The age at which various events are most likely to occur may reveal the sequence among these events. Finally, there are also empirical tests of sequence. These include conditional independence test of colliders which can establish that two or more causes occur before their common effect. Other tests like Probabilistic Contrast model, and Goodman and Kruskal error reduction methods can also be used to establish sequence [[[12]](#endnote-13)].

**Determining the Sequence in the Example Data**

In this example, the determination of sequence is relatively easy. We know that the episodes starts with a hospital admission, during which the physician bill for their service, clinical laboratory tests are ordered and durable medical devices are placed in patient’s hip. So, DME, CL and P are the events that occur in the first time period. Hospital discharge, H, occurs in the second time period. Patients who are discharged from the hospital go to a series of post-acute services. These post-acute services are assumed to occur in time period 3. Discharge to these post-acute services will be primarily to outpatient services. We have assumed that no re-hospitalization occur and there are no cycles in the network. In addition, the way we simulated the data assumes that all causes occur prior to effects and not simultaneously with the effect. The final event is bundle payment which is assumed to occur 90 days after initial hospitalization.

**Stepwise Pass through Data**

The Poisson regression is done through a time-based stepwise process, where variables are entered into the analysis in order of their occurrence. For example, The R code for analysis of variables at Time 1 is given by:

|  |
| --- |
| First, we place the three variables of Time 1 (“CL”, “DME” , “P”) in one data.frame    Then we perform a poisson regression    and the results are as follows    So what these results are saying is that none of the variables of time1 are related to each other. |

The full R code for sequential processing of the data is given below:

|  |
| --- |
| For Time2, since no relations were found in time 1, we have :-    And the model we have is    So poisson regression has accurately uncovered the relations in times 1 amd 2.  For time3, we have    And the model we get is    For time4 we have |

Table 2 provides a summary of the results of Poisson regressions. In this Table, we report only findings of significance for pairwise interactions, main effects were part of the model but of no relevance to our analysis and are not reported here. If an interaction term was not included in the regression model, then the cell is shaded. Some interaction terms that show in initial analysis but not in subsequent analysis are eliminated from all subsequent analysis and therefore will have shaded cells in subsequent analysis. Interaction terms that were excluded because of timing also have shaded cells.

In Table 2, T1 reports the first Poisson regression. This regression includes the interactions among the 3 variables in Time 1: DME, P and CL. Note that none of the interactions among these 3 variables were significant. These interaction terms are excluded from subsequent analysis. At time 2, we add in the Hospital, H, variable. The model in Time 2 includes all interaction terms that were significant at Time 1 (none) and every possible interaction with variables in time 2. In Time 2 several statistically significant relationships are found. Note that H is a common effect of P, DME and CL, and if we had not done the stepwise process, the Poisson regression would have also found new relationships among the variables in Time 1.

**Table 2: Statistically Significant Relationships in Poisson Regression**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Interaction Terms in Poisson Regression** | **T1** | **T2** | **H** | **T3** | **HHA** | **SNF** | **RF** | **PBD** | **LTH** | **T4** | **OT** | **HO** | **HOS** | **T5** | **BP** | **Recovered** | **Original** |
| P:CL | N |  |  |  |  |  |  |  |  |  |  |  |  |  |  | N | N |
| DME:P | N |  |  |  |  |  |  |  |  |  |  |  |  |  |  | N | N |
| DME:CL | N |  |  |  |  |  |  |  |  |  |  |  |  |  |  | N | N |
| P:H |  | Y | Y | Y |  |  |  |  |  | Y |  |  |  | Y |  | Y | Y |
| CL:H |  | Y | Y | Y |  |  |  |  |  | Y |  |  |  | Y |  | Y | Y |
| DME:H |  | Y | Y | Y |  |  |  |  |  | Y |  |  |  | Y |  | Y | Y |
| Other:HHA |  |  |  | Y | N |  |  |  |  |  |  |  |  |  |  | N | N |
| P:HHA |  |  |  | Y | Y |  |  |  |  | Y |  |  |  | Y |  | Y | Y |
| Other:SNF |  |  |  | Y |  | Y | N |  |  |  |  |  |  |  |  | N | N |
| P:SNF |  |  |  | Y |  | Y | Y |  |  | Y |  |  |  | Y |  | Y | Y |
| H:SNF |  |  |  | Y |  | Y | Y |  |  | Y |  |  |  | Y |  | Y | Y |
| Other:RF |  |  |  | Y |  |  | N |  |  | N |  |  |  | N |  | N | N |
| P:RF |  |  |  | Y |  |  | Y |  |  | Y |  |  |  | Y |  | Y | Y |
| H:RF |  |  |  | Y |  |  | Y |  |  | Y |  |  |  | Y |  | Y | Y |
| Other:PBD |  |  |  | Y |  |  |  | N |  | N |  |  |  | N |  | N | N |
| Other:LTH |  |  |  | Y |  |  |  |  | N | N |  |  |  | N |  | N | N |
| P:LTH |  |  |  | Y |  |  |  |  | Y | Y |  |  |  | Y |  | Y | Y |
| H:LTH |  |  |  | Y |  |  |  |  | Y | Y |  |  |  | Y |  | Y | Y |
| Other:OT |  |  |  |  |  |  |  |  |  | Y | N |  |  |  |  | N | N |
| HHA:OT |  |  |  |  |  |  |  |  |  | Y | Y |  |  | Y |  | Y | Y |
| SNF:OT |  |  |  |  |  |  |  |  |  | Y | Y |  |  | Y |  | Y | Y |
| RF:OT |  |  |  |  |  |  |  |  |  | Y | Y |  |  | Y |  | Y | Y |
| PBD:OT |  |  |  |  |  |  |  |  |  | Y | Y |  |  | Y |  | Y | Y |
| LTH:OT |  |  |  |  |  |  |  |  |  | Y | Y |  |  | Y |  | Y | Y |
| Other:HO |  |  |  |  |  |  |  |  |  | Y |  | N |  | N |  | N | N |
| PBD:HO |  |  |  |  |  |  |  |  |  | Y |  | Y |  | Y |  | Y | Y |
| LTH:HO |  |  |  |  |  |  |  |  |  | Y |  |  | N |  |  | N | N |
| Other:HOS |  |  |  |  |  |  |  |  |  | Y |  |  | N |  |  | N | N |
| LTH:HOS |  |  |  |  |  |  |  |  |  | Y |  |  | Y | Y |  | Y | Y |
| H:HOS |  |  |  |  |  |  |  |  |  | Y |  |  | Y | Y |  | Y | Y |
| PBD:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | N | N | N | Y |
| SNF:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | Y | Y | Y |
| HHA:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | Y | Y | Y |
| OT:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | Y | Y | Y |
| HO:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | Y | Y | Y |
| HOS:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | Y | Y | Y |
| H:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | Y | Y | Y |
| Other:BP |  |  |  |  |  |  |  |  |  |  |  |  |  | Y | N | N | N |
| Main effects are not listed in this Table | | | | | | | | | | | | | | | | | |

In Time 1 and 2, the relationships that have a statistically significant relationship with H, are in the Markov Blanket of H. These relationships can be tested by stratifying all other statistically significant relations and examining if the effect remains statistically significant. This is done by conducting a smaller regression involving H and its current Markov Blanket. In this case, this is the same regression that was done for Time 2.

Table 2 shows what happens when we move to Time 3. We include in the model all variables that remained statistically significant at Time 2. In addition, a set of 5 new variables are introduced in Time 3. The interaction among these 5 variables and all other variables are included in the Poisson regression model. The R code for this regression is given below, where we have highlighted the new Time 3 related interaction terms:

|  |
| --- |
| model.time3=glm(V1 ~ CL\*H + DME\*H + P\*H + CL + P + DME + H + LTH + PBD + RF + SNF + HHA + CL\*LTH + CL\*PBD + CL\*RF + CL\*SNF + CL\*HHA + DME\*LTH + DME\*PBD + DME\*RF + DME\*SNF + DME\*HHA + P\*LTH + P\*PBD + P\*RF + P\*SNF + P\*HHA+ H\*LTH + H\*PBD + H\*RF + H\*SNF + H\*HHA  , data=counts.time3,family=poisson) |

The result of this regression is shown in the Appendix and also summarized in Table 2. Note that the analysis has identified a number of new associations that did not exist in the original network, which was used to simulate the data. Figure 2 shows these errors in red. At this point, we do smaller Poisson regressions involving variables in the Markov Blanket of variables introduced in Time 3. Many of the erroneous relationships are removed from the data by these smaller Poisson regressions.

In the next forward step, we include variables that remained significant in Time 3 and add in interactions with variables in Time 4. Note again that the analysis of Time 4 identifies a number of associations that exist in the data but not in the network that simulated the data. In Figure 2 these associations are shown in red.

**Figure 2: Errors that Remain in Different Time periods**Black Lines show Correct Links, Red Lines Incorrect Links

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

In the last forward step, we include Bundled Payment, BP, in the analysis. The analysis includes all pair-wise interactions that were both statistically significant and large in the previous analysis. In addition, it includes all pair-wise interactions among BP and all previous variables. Table 2 shows the summary of findings and Appendix provides the details. Again, Poisson regression of the parents of Markov Blanket of the variables introduced in this time period removes the errors.

The analysis has nearly identified the network that was used to generate the data. To test the accuracy of the analysis we could examine the frequency of discrepancies between discovered and original network. A directed arc could be absent in the original and present in the discovered network, and vice versa. Clearly sequential use of Poisson regression reduced errors.

These data can also be reported using sensitivity and specificity of the discovered network.

**Discussion**

This chapter has demonstrated that Poisson regression can be used to detect network structure and parameters. We have shown how forward and backward passes through the data reduce the errors in associations discovered through Poisson regression. In the forward steps, we showed how associations that exist because of collider effects are prevented to enter the model. In the backward steps, we showed that progressive test of causal impact improves the accuracy of Poisson regression. In this example, the discovered model had high sensitivity and specificity encouraging the use the proposed forward and backward passes.

The procedure used here are similar to procedures proposed by Shojaie and colleagues [[[13]](#endnote-14)]. Like them, we used partial sequence among the variables to discover network structures and parameters. Unlike their approach, we relied on Poisson regression and not logistic or linear regression. Furthermore, we devised a forward and backward pass that removes most of the non-causal associations; i.e. associations that exist in the data but not in the causal model that generated the data. These include associations that exist because of stratifying a common effect or because of co-variation of several effects due to a common cause.

There are a number of limitations in the procedures used in this chapter. This procedure is not always accurate. The accuracy has only been shown in one example. We have not established that Poisson regression will be accurate in every type of network. Additional simulation studies are needed so that one can anticipate when Poisson regression will be effective and why. The chapter generated a large sample size and large effects for impact of each cause. These conditions may not always be present and will distort the findings of Poisson regression.

**References**

1. Jacobs BL, Zhang Y, Skolarus TA, Wei JT, Montie JE, Schroeck FR, Hollenbeck BK.Managed care and the diffusion of intensity-modulated radiotherapy for prostate cancer. Urology. 2012 Dec;80(6):1236-42. [↑](#endnote-ref-2)
2. Zhang Y, Hollenbeck BK, Schroeck FR, Jacobs BL. Managed care and the dissemination of robotic prostatectomy. Surg Innov. 2014 Dec;21(6):566-71 [↑](#endnote-ref-3)
3. Gibbons RD, Segawa E, Karabatsos G, Amatya AK, Bhaumik DK, Brown CH, Kapur K, Marcus SM, Hur K, Mann JJ . Mixed-effects Poisson regression analysis of adverse event reports: the relationship between antidepressants and suicide. Stat Med. 2008 May 20;27(11):1814-33. [↑](#endnote-ref-4)
4. Wilkinson AL, Pedrana AE, El-Hayek C, Vella AM, Asselin J, Batrouney C, Fairley CK, Read TR, Hellard M, Stoové M. The Impact of a Social Marketing Campaign on HIV and Sexually Transmissible Infection Testing Among Men Who Have Sex With Men in Australia. Sex Transm Dis. 2016 Jan;43(1):49-56. [↑](#endnote-ref-5)
5. Lee KH, Lim S, Park J. Expelled uninsured patients in a less-competitive hospital market in Florida, USA. Int J Equity Health. 2016 Jun 4;15:85. [↑](#endnote-ref-6)
6. Preston DL. Poisson regression in epidemiology. In: Armitage P, Colton T, editor. Encyclopedia of Biostatistics. 2. Vol. 6. Chichester: John Wiley & Sons; 2005. pp. 4124–4127. [↑](#endnote-ref-7)
7. Agresti, A. Categorical Data Analysis. Wiley-Interscience, 2002. [↑](#endnote-ref-8)
8. See <http://openonlinecourses.com/causalanalysis/simulated%20bundled%20data.csv> accessed on October 17, 2016 [↑](#endnote-ref-9)
9. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann 1988. [↑](#endnote-ref-10)
10. Pearl J. Fusion, Propagation and Structuring in Belief Networks. Artificial Intelligence 1986. 29 (3): 241–288. [↑](#endnote-ref-11)
11. Pearl J. An introduction to causal inference. Int J Biostat. 2010 Feb 26;6(2):Article 7. [↑](#endnote-ref-12)
12. Alemi F, Zargoush M, Vang J. Using observed sequence to orient causal networks. Health Care Manag Sci. 2016 Jul 30. [↑](#endnote-ref-13)
13. Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika. 2010 Sep;97(3):519-538. Epub 2010 Jul 9. [↑](#endnote-ref-14)