# Chapter 19

# Association Network

# [H1] Learning Objectives

# [INSERT NL]

1. Construct an association network model through Poisson regression
2. Check for independence of pairs of two, three, and more than three variables

# [END NL]

# [H1] Key Concepts

# [INSERT BL]

* Independence
* Conditional independence
* Spurious correlations
* Poisson regression
* Association networks
* Residual deviance
* G-squared

**[END BL]**

# [H1] Chapter at a Glance

This chapter examines independence, a fundamental concept in statistical analysis. It describes bivariate independence, multivariate independence, and network models of dependence (associations).

**[H1] Not in Widespread Use**

This chapter focuses on the concept of independence and association networks. Managers and analysts seldom use these concepts. Association networks are not seen often in analysis of healthcare data. Some use occurs in analysis of social networks, but otherwise this tool is not used as often as regression tools. We introduce the concept of independence and association networks in this chapter in anticipation of a much more practical tool discussed in chapter 20: causal networks.

# [H1] Concept of Independence

We introduced the concept of independence in chapter 3 in this book. We have repeatedly referred to it. It is one of the most fundamental concepts in statistics. In fact, statistics may be described as a search for relationship—the absence of independence—among variables. In probabilities, the concept of independence has a very specific meaning. If two events are independent of each other, then the occurrence of one event does not tell us much about the occurrence of the other event. *Independence* means that the presence of one clue does not change the value of another. An example might be prevalence of diabetes and car accidents; knowing the probability of car accidents in a population will not tell us anything about the probability of diabetes. For another example, if a patient has a heart attack, that fact may not change the diagnosis of the next patient. If so, we can consider the illnesses of the two patients independent of each other. This may be a reasonable assumption if the current and the next patient are not related or do not influence each other’s lifestyle.

Independence is a symmetric concept. If *a* is independent of *b*, then *b* is independent of *a*. Verifying independence of one is sufficient to prove the independence of another. Note that for independence to hold, it must hold at every level of the variable. If we are examining the independence between two variables, they should be independent in every combination of the two variables—independent when the variables are present or absent. To stay with our earlier example, if diseases of patients are independent from each other, this should be the case when patients have the disease or when they do not—when the patient has a heart attack or when he does not. In both situations, the probability of the next patient having a heart attack does not change. We typically do not show the absence, as dealing with negatives is cumbersome, but whether shown or not, independence must hold for all levels of the variable.

Here is another example of two variables that should be independent. Suppose we have two clinics, one in San Francisco (SF) and the other in Washington, DC. If I tell you that lots of clinic patients are waiting for healthcare in DC, it may not change the probability of people waiting in San Francisco. It would, if there was a national demand for our clinical service, but in the absence of this national demand we are safe to assume that the two clinics’ waiting times are independent. In probabilistic terms, we can write this as an equation indicating that the conditional probability of waiting in San Francisco, given that there are long waits in Washington, DC, is the same as the unconditional probability of waiting in San Francisco:

**[INSERT EQUATION]**

.

**[END EQUATION]**

As we had also said before, this relationship holds for whether the wait time is long or short in DC:

**[INSERT EQUATION]**

**[END EQUATION]**

It is possible for two events to be dependent, but when conditioned on the occurrence of a third event, they may become independent of each other. For example, we may think that scheduling long shifts will lead to medication errors. Thus we may show ( means not equal to)

**[INSERT EQUATION]**

.

**[END EQUATION]**

At the same time, we may consider that in the population of employees who are not fatigued (even though they have long shifts), the two events are independent of each other:

**[INSERT EQUATION]**

**[END EQUATION]**

This example shows that related events may become independent under certain conditions.

# [H1] Shrinking Universe of Possibilities

One way to verify independence is through restricting the analysis to situations in which a subset of data on the probability of the event does not change. In particular, one could check that the conditional and unconditional probabilities are the same. Mathematically, this condition can be presented as

**[INSERT EQUATION]**

.

**[END EQUATION]**

An example can demonstrate. Suppose we want to check independence, and we have the data in exhibit 19.1. This contingency table shows the counts of various combinations of events. The total sample size is *a* + *b* + *c* + *d*. This is the entire universe of data available to the analyst.

**[INSERT EXHIBIT]**

**Exhibit 19.1** Shrinking the Universe of Possibilities

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Calculation of p(B)** | | | | |  | **Calculation of p(B|A)** | | | | |
|  | | Event A | | Total |  | | Event A | | Total |
| Yes | No | Yes | No |
| Event B | Yes | *a* | *b* | *a* + *b* | Event B | Yes | *a* | *b* | *a* + *b* |
| No | *c* | *d* | *c* + *d* | No | *c* | *d* | *c* + *d* |
| Total | | *a* + *c* | *a* + *d* | *a* + *b* + *c* + *d* | Total | | *a* + *c* | *b* + *d* | *a* + *b* + *c* + *d* |

**[END EXHIBIT]**

By definition, the probability of any event in a sample is the number of times the event occurs divided by the universe of possibilities, the sample size. The probability of event B is the number of times B occurs. B occurs *a* times when A occurs and *b* times when A does not occur, so the count of times B occurs is *a* + *b*; the probability is the number of times B occurs, *a* + *b*, divided by what is possible to occur, *a* + *b* + *c* + *d*: (a + b) ÷ (a + b + c + d). To calculate the conditional probability of B given *a*, the first step is to shrink the sample size to the situations where the condition has been met. The condition is that event *a* has occurred. The sample shrinks to only cases where *a* has occurred. All other events are no longer possible and therefore not part of the universe of possibilities; these situations have been crossed out in the right-hand side of exhibit 19.1. There are *a* + *c* cases where event A has occurred. This is now the new universe of possibilities. We want to know the frequency of event B among these possible cases. Event B occurs *a* times in this smaller sample size. So the conditional probability of B can be calculated as *a* ÷ (*a* + *c*). Think of conditional probability as recalculating the probability in a smaller universe of possibilities. The following shows how the calculation of the probability of B given A changes from the calculation of probability of B:

**[INSERT EQUATION]**

**[END EQUATION]**

In essence, we are recalculating the probability of the event but now in a smaller sample where the possible cases are limited to all the cases that meet the condition. This method of calculating independence lends itself perfectly to calculations using Structured Query Language (SQL). The original probability of B is calculated as

**[LIST FORMAT]**

SELECT Sum(B)/Count(B) AS [Prob of B]

FROM [All Data]

**[END LIST]**

In SQL, you can shrink the universe of possibilities by using the WHERE command.

**[LIST FORMAT]**

SELECT Sum(B)/Count(B) AS [Prob B given A]

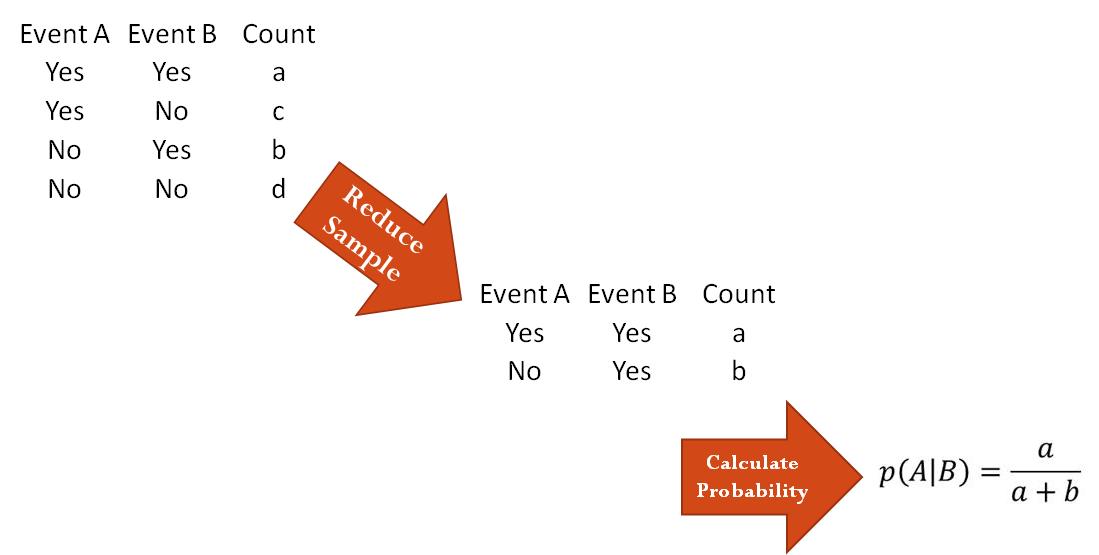
FROM [All Data]

WHERE A is true

**[END LIST]**

Exhibit 19.2 shows how the SQL command reduces the rows of available data and calculates the probability of interest; in this case, calculating the conditional probability of A given B. The initial data were organized in four strata, each representing different combinations of the events of interest. You have seen similar steps organizing data into strata throughout this book. In the next step, the strata where the condition has been met is kept. This reduces four strata to two. Then, in the last step, the conditional probability is calculated from these reduced data. This example shows the relationship between strata and conditioning, a concept that links methods such as stratified covariate balancing (discussed in chapter 16) to association networks (discussed in this chapter) and probability networks (discussed in chapter 20).

**[INSERT EXHIBIT]**

**Exhibit 19.2** Conditioning Through Reduction in Strata

**[END EXHIBIT]**

For example, in exhibit 19.3, 18 cases from a special unit of a hospital prone to medication errors are presented. The question is whether the rate of medication errors is independent of length of work shift.

**[INSERT EXHIBIT]**

**Exhibit 19.3** Medication Errors in 18 Consecutive Cases

|  |  |  |  |
| --- | --- | --- | --- |
| **Case** | **Medication Error** | **Long Shift** | **Fatigue** |
| 1 | No | Yes | Yes |
| 2 | No | Yes | Yes |
| 3 | No | No | Yes |
| 4 | No | No | Yes |
| 5 | Yes | Yes | Yes |
| 6 | Yes | No | Yes |
| 7 | Yes | No | Yes |
| 8 | Yes | Yes | Yes |
| 9 | No | No | No |
| 10 | No | No | No |
| 11 | No | Yes | No |
| 12 | No | No | No |
| 13 | No | No | No |
| 14 | No | No | No |
| 15 | No | No | No |
| 16 | No | No | No |
| 17 | Yes | No | No |
| **18** | **Yes** | **No** | **No** |

**[END EXHIBIT]**

Using the data in exhibit 19.3, the probability of medication error is calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

Now, we can verify independence by comparing the product of the marginal probabilities and the joint probability of the two events, via the equation

**[INSERT EQUATION]**

**[END EQUATION]**

These calculations show that the probability of medication error and the length of the shift depend on each other. Knowing the length of the shift tells us something about the probability of error in that shift. Surprisingly, it reduces error rates, perhaps because there are fewer points of switching from one provider to another. But consider a situation in which we are examining these two events among cases where the provider was fatigued. Now the population of cases we are examining is reduced to cases 1 through 8. As described earlier, conditional probabilities are calculated by restricting the universe of possibilities. In the population of fatigued providers (i.e., cases 1 through 8), there are several cases of working a long shift (i.e., cases 1, 2, 5, and 8). We can use this information to calculate conditional probabilities:

**[INSERT EQUATION]**

*p*(Error|Fatigue) = 0.50 and   
*p*(Error|Fatigue and long shift) = 2 ÷ 4 = 0.50.

**[END EQUATION]**

Among fatigued providers, medication error is independent of the length of work shift. These procedures show how independence can be verified by counting cases in reduced populations. Despite its inherent logic, the approach of comparing probabilities is not a particularly good way to go about verifying independence. This approach is problematic because the two probabilities may be different from each other as a result of small random measurement errors. To truly test for independence, we need a procedure that ignores random differences among the calculated probabilities.

# [H1] Product of Marginal Probabilities

Independence can also be verified by checking that the probability of a joint event is the product of the probability of each event using the formula

**[INSERT EQUATION]**

.

**[END EQUATION]**

If this formula is displayed in terms of counts, we see that it implies that a count of two independent events can be calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

Removing the sample size from both sides of the equation yields

**[INSERT EQUATION]**

.

**[END EQUATION]**

**[H2] Rehospitalization Example**

A numerical example can also demonstrate how counts can be used to verify independence. Consider a hospital administrator who wants to know if the hospice program is reducing 30-day readmissions for heart failure patients. Heart failure patients have many rehospitalizations, and the Centers for Medicare & Medicaid Services does not pay for rehospitalizations that occur within 30 days of the original admission. Hospital administrators would like to promote programs that reduce readmissions. One option available to end-stage heart failure patients is to use the hospice program. While the program is known to increase the patient’s comfort, it is not always clear that it reduces rehospitalization. Even after a patient joins the hospice program, she may suddenly have dyspnea (severe shortness of breath). Sometimes families concerned about dyspnea may decide to rehospitalize their loved ones. Exhibit 19.4 provides a set of hypothetical data for the hospitalization rate among patients in hospice and nonhospice programs.

**[INSERT EXHIBIT]**

**Exhibit 19.4** Joint Probability for 100 Cases

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Rehospitalized in 30 days** | **Not Rehospitalized in 30 Days** | **Total** |
| In hospice | 0.51 (51) | 0.39 (39) | 0.9 (90) |
| Not in hospice | 0.05 (5) | 0.05 (5) | 0.1 (10) |
| Total | 0.56 (56) | 0.44 (44) | 1 (100) |

**[END EXHIBIT]**

Note that exhibit 19.4 provides joint and marginal probabilities by dividing the 100 patients into various categories. Marginal probabilities refer to the probability of any one event; in exhibit 19.1, marginal probabilities are provided in rows and columns labeled Total. Joint probability refers to the probability of two events co-occurring. For example, the joint probability of having both rehospitalization and hospice is 0.51, meaning that out of 100 patients, 51 are both rehospitalized and in hospice.

The total number of patients is referred to as the *sample size* or *universe of possibilities*. In exhibit 19.4, the universe of possibilities is 100 patients, and each patient may have different combinations of hospice and hospitalization. If the analyst wishes to calculate a conditional probability, the total universe of possible patients must be reduced to patients with the condition. Suppose the analyst wants to calculate the conditional probability of rehospitalization given that the patient is already in hospice. In this case, the universe of possibilities is reduced to all patients who are already in hospice. In this reduced universe, the total number of patients in hospice is 90 patients. Among these 90 patients, 51 were rehospitalized. Therefore, the conditional probability of rehospitalization among patients in a hospice program is

**[INSERT EQUATION]**

**[END EQUATION]**

Because exhibit 19.4 provides the joint and marginal probabilities, we can describe these calculations in terms of joint and marginal probabilities:

**[INSERT EQUATION]**

.

**[END EQUATION]**

The point of this example is that you can calculate conditional probabilities from marginal and joint probabilities if you keep in mind how the condition has reduced the universe of possibility.

# [H1] Chi-Square Test of Independence

Chi-square can be used to test independence. Details of how chi-square works were introduced in the introduction to probabilities in chapter 3. Chi-square tests compare observed counts to expected counts under the assumption of independence. The chi-square test statistic, is calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, the observed counts are shown as , the expected counts are shown as , and *c* denotes the degrees of freedom. On the web, several sites provide the probability of observing different chi-square statistics, given the degrees of freedom (calculated as sample size minus one).

The key in using the chi-square test is having a simple way to estimate the expected count. If two events are independent, then the expected count of the combination of two events, is calculated as  
**[INSERT EQUATION]**

**[END EQUATION]**

The formula says that the expected count of two independent events is the product of the count of each event divided by the total number of cases in the sample.

Chi-square testing has several limitations. The chi-square test is sensitive to sample size—therefore, in massive data, even small differences are likely to be statistically significant. For these cases, examining both the statistical significance and the effect size is important. Small dependencies can be ignored. Chi-square testing is also sensitive to the distribution of counts in the contingency table cells. If possible, it is good practice to collapse several categories, each of which having only a few observations, into one cell. Chi-square testing may not be accurate if fewer than five observations fall in the cell.

# [H1] Visual Display of Dependence

Independence among variables can be displayed through networks, where nodes represent the variables and arcs represent direct dependence among them. The independence of two variables can be shown as a function *I* taking the two parameters of *A* and *B* separated by a comma. Thus, we may indicate that two variables are independent by showing them as *I*(*A*,*B*); this is read “A and B are independent.” If the context allows, *I* is dropped and only the two variables are shown with a comma separating them: (*A*, *B*). If two events are independent—neither an association nor a causal relationship exists among them—we show the events as an unconnected network (see exhibit 19.5).

**[INSERT EXHIBIT]**

**Exhibit 19.5** Display of Two Independent Events (A,B)

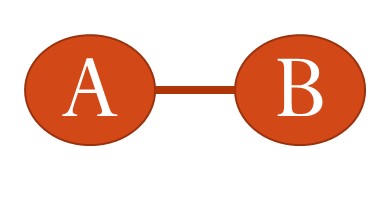
****

**[END EXHIBIT]**

If two events are associated with each other, the analyst does not separate *A* and *B* with a comma and the network is shown with an arc between the two nodes. In exhibit 19.6, *A* and *B* are related events—knowing something about *A* changes the probability of *B*.

**[INSERT EXHIBIT]**

**Exhibit 19.6** Display of Two Dependent Events (AB)

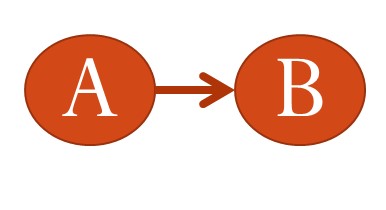
****

**[END EXHIBIT]**

If an arrow is shown between the events, a sequence is assumed, which enhances causal interpretation. The network in exhibit 19.7, for example, says that *A* precedes, is associated with, and may cause *B*.

**[INSERT EXHIBIT]**

**Exhibit 19.7** Display of Causal Relationships



**[END EXHIBIT]**

An association network shows the association between any pair of variables. If a variable is not correlated with any others in the study, then it is not included in the network model. Every variable in a network model is associated with the other variables; some of these associations are direct and shown in the network, and other associations are indirect and through other variables. Exhibit 19.8 is an example of an association network. There are five variables in this network, and pair-wise associations are shown. Every variable is directly associated every other.

**[INSERT EXHIBIT]**

**Exhibit 19.8** Saturated Association Network for Five Variables

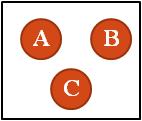
**[END EXHIBIT]**

# [H1] Independence for Three Variables

If we are dealing with three or more variables, the establishment of independence becomes increasingly complex. Now all variables could be independent of one another, pairs of variables can be independent from a third variable, trios of variables could be independent from a fourth variable, and so on. In this section, we focus on how independence can be explored for three variables.

There are at least three ways the variables could be related to each other: complete independence, joint independence, and saturated models. In complete independence, all variables are independent of each other. In exhibit 19.9, we show the three variables *A*, *B*, and *C* without any linkage between them to mark that they are independent from each other. We can also show this as *I*(*A*,*B*,*C*)—the commas between the variables indicate their independence.

**[INSERT EXHIBIT]**

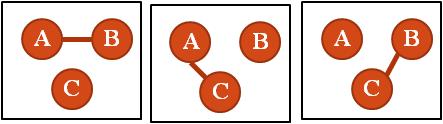
**Exhibit 19.9** A Network Showing Three Independent Variables  


**[END EXHIBIT]**

Joint independence is the situation where two variables are independent of a third (see exhibit 19.10). In these graphs, a bar between two variables indicates that they are dependent on each other (i.e., there is a relationship among them). There are three ways that joint independence could occur among three variables. *A* and *B* could be related but independent of *C*, *A* and *C* could be related but independent of *B*, and *B* and *C* could be related to each other but independent of *A*. These three situations can also be shown as *I*(*AB*,*C*), *I*(*AC*,*B*), and *I*(*BC*,*A*). The saturated model occurs when all variables are dependent on each other (see exhibit 19.11).

**[INSERT EXHIBIT]**

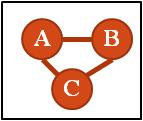
**Exhibit 19.10** Joint Independence in Triple of Variables



**[END EXHIBIT]**

**[INSERT EXHIBIT]**

**Exhibit 19.11** A Saturated Network of Three Dependent Variables

****

**[END EXHIBIT]**

These models should be considered in order of parsimony, with complete independence being most parsimonious and the saturated model being the least parsimonious. In science, parsimonious models are preferred. If data can be modeled with both saturated and joint independence, then the joint independence model is preferred. If data are completely independent, they can also be modeled by any of the other five methods, but the analyst gives preference to the complete independence model.

# [H1] Chi-square Testing for Three Variables

To understand how several variables are independent from each other, we make assumptions about independence, generate expected values from the assumed independence structure, and test the fit of the model to observed data using the chi-square test. With three variables, the chi‑square test is done using the formula

**[INSERT EQUATION]**

.

**[END EQUATION]**

is the count of co-occurrences of the three variables; is the estimated count for the same combination given the assumption of independence, joint independence, or saturated model.

An example can demonstrate how chi-square testing is carried out. This example is organized to demonstrate the ideas behind a test of independence—statisticians rarely follow these procedures in everyday practice. More sophisticated methods (e.g., Poisson regression) can accomplish the analysis described here more quickly and easily. However, this example can demonstrate the basic ideas behind the test and provide the reader with intuition about how tests of independence among multiple variables are carried out. Suppose that we have three categorical variables, *A*, *B*, and *C*, where *A* takes possible values 1 through *I*. *B* takes possible values 1 through *J*. *C* takes possible values 1 through *K*. For example, *A* could be two physicians in our clinic, Ruiz and Smith. *B* could be two nurses in our clinic, Washington and Nguyen. And *C* could be whether a patient has complained about the combined physician–nurse team. It takes on values of Yes and No. If we collect the trio *A*, *B*, and *C* for each unit in a sample of *n* units, the data can be summarized as a three-dimensional table (see left side of exhibit 19.12). Let Y be the count of units having *A* being *i*, *B* being *j*, and *C* being *k*. Then *Y* will provide a count for each cell in the three-dimensional table of *A*, *B*, and *C*. When all variables are categorical, a multidimensional contingency table can be displayed (see left side of exhibit 19.12), or data can be shown in strata (see right side of exhibit 19.12). For example, *Y*Ruiz,Nguyen,Yes is the count of patients who were seen by the clinical team of Ruiz and Nguyen and who had complained. This is one cell value in a larger table. Each cell in this table is one of the observed counts. In exhibit 19.12, we see the distribution of *Y* for different clinical teams and complaint combinations. We see the distribution of the counts for different combinations of *A*, *B*, and C.

**[INSERT EXHIBIT]**

**Exhibit 19.12** Satisfaction Across Teams of Providers

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Contingency Table** | | | | |  | **Data by Strata** | | | |
| A: Physicians | B: Nurses | C:Complaint | | Total | A: Physician | B:Nurse | C:Complaints | Count |
| Yes | No | Ruiz | Washington | Yes | 53 |
| Ruiz, MD | Washington, RN | 53 | 424 | 477 | Ruiz | Washington | No | 424 |
| Nguyen, RN | 11 | 37 | 48 | Ruiz | Nguyen | Yes | 11 |
| Smith, MD | Washington, RN | 0 | 16 | 16 | Ruiz | Nguyen | No | 37 |
| Nguyen, RN | 4 | 139 | 143 | Smith | Washington | Yes | 0 |
| Total | | 68 | 616 | 684 | Smith | Washington | No | 16 |
|  | | | | | | Smith | Nguyen | Yes | 4 |
| Smith | Nguyen | No | 139 |

**[END EXHIBIT]**

In a partial table, one of the variables is held constant. For example, in the left side of exhibit 19.12, the physician, Ruiz, is held constant. All of the data in this partial table are about teams involving Ruiz. The next two rows in exhibit 19.12 indicate another partial table about Dr. Smith.

A marginal table is obtained by summing out one of the variables. So, if we remove the physicians, we would sum up the values for physicians Ruiz and Smith and put them in a new table. For example, the number of positive comments that nurse Nguyen has received is the number she received while working with Dr. Ruiz plus the number she received while working with Dr. Smith. This yields 11 + 4 = 15. We will use + to indicate summation over a subscript; for example, *Yi*,+,*k* indicates summing over the subscript *j* to produce a marginal table involving only subscripts *i* and *k*. In this marginal table, variable *B* is ignored. In *Y*+,+,*k*, both the subscripts *i* and *j* are summed over. We are ignoring the variables *A* and *B*. We see counts associated with *k* levels in *C*.

The association between any two variables can be examined in its own marginal table, if the two variables are independent from each other. We show how this is done for verifying the saturated model, joint independence, and complete independence. In the saturated model, no assumptions of independence are made. The expected count for the saturated model is the value of each cell. By definition, the chi-square statistic is 0, and the degree of freedom is 0 too. The data fit the model perfectly. This should not be surprising, as the saturated model assumes that all combinations of the variables are a separate predictor. At the same time, fitting a saturated model does not reveal any special structure that may exist in the relationships among *A*, *B*, and *C*. To investigate these relationships, one has to pursue different assumptions of independence.

Joint independence indicates that *A* and *B* are related, but *C* is independent of these two variables. The model is tested by creating marginal tables and using the count in these tables to estimate the expected count. A chi-square test can then be used for this model. In three variable models, a statistician can test for joint independence of *AB*, *AC*, and *BC* from their complements. Here we are showing how to test whether *AB* is independent of *C*. The expected count is calculated as   
**[INSERT EQUATION]**

.

**[END EQUATION]**

The chi-square test is then carried out with the following degrees of freedom:  
**[INSERT EQUATION]**

.

**[END EQUATION]**

Finally, under the assumption of complete independence, the expected count can be estimated from the product of the marginal counts:  
**[INSERT EQUATION]**

**[END EQUATION]**

The chi-square test can then be carried out using the following degrees of freedom:

**[INSERT EQUATION]**

.

**[END EQUATION]**

An example can demonstrate the procedures for these tests. Exhibit 19.13 indicates the number of positive and negative comments received by each team. We use the data to identify the relationship among the three variables. We will begin with the test of complete independence of the three variables. We use the formula for expected count to predict the counts in each cell. These values are shown in the column titled “Expected.” The chi-square formula is calculated from the difference of expected and observed values. The chi-square value with 4 degrees of freedom is 425.07, which is statistically significant at alpha levels less than 0.01, and therefore the hypothesis that the relationship among the variables is complete independence is rejected. The data do not fit the assumptions of complete independence.

**[INSERT EXHIBIT]**

**Exhibit 19.13** Test of Independence of Team Members

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **A: Physicians** | **B: Nurses** | **C: Complaint** | **Expected Under Complete Independence Assumption** | **Expected Under Joint Independence Assumptions** | **Observed** |
| Ruiz | Washington | Yes | 38 | 47 | 53 |
| Ruiz | Washington | No | 341 | 430 | 424 |
| Ruiz | Nguyen | Yes | 15 | 5 | 11 |
| Ruiz | Nguyen | No | 132 | 43 | 37 |
| Smith | Washington | Yes | 11 | 2 | 0 |
| Smith | Washington | No | 103 | 14 | 16 |
| Smith | Nguyen | Yes | 4 | 14 | 4 |
| Smith | Nguyen | No | 40 | 129 | 139 |

**[END EXHIBIT]**

In exhibit 19.13, we also report the expected values under the joint independence assumptions. As before, the expected values are calculated from marginal tables, but this time under the assumption of joint independence. Assuming that *A* and *B* are related and independent of *C* produces a chi-square of 19.09, which is statistically significant at two degrees of freedom. The hypothesis that the model fits the data is rejected, but one thing is clear: The fit is better than under the assumption of complete independence. The joint independence model assumes that the impact of the clinical team depends on who is working together—in other words, the combination of the team matters. The point of the example is that it is possible to find out which model fits the data by calculating the expected value using different assumptions. The model that best fits the data has the smallest chi-square value.

# [H1] Spurious Correlation

Correlations can be easily calculated in Excel using the CORREL function. In R, correlations can be calculated using the package called “cor.” Correlation values can indicate the relationship with two variables. Dependent variables are correlated with each other, but the reverse is not always true: lack of correlation does not imply independence.

Misinterpretation of correlations is easy. People often mistakenly think that a correlation of zero means that there is no relationship between the two variables. A nonlinear relationship may still exist. A correlation of zero only says that there is no linear relationship, and it is not reasonable to assume that nonlinear relationships do not exist. If the data were transformed so that the relationship between the two variables was linear, then higher correlations may be observed.

In addition, two variables may be uncorrelated in the population but correlated in another context (conditional correlation). The mere lack of correlation does not mean that, in every subset of data, the two variables are uncorrelated. For example, a history of repeated infections may not be predictive of Alzheimer’s disease—but among patients over 85 years, the relationship may exist. In the presence of another variable, unrelated variables may become correlated.

Correlation between two variables may disappear if a third variable is introduced. In multivariate regression, this phenomenon is known as multicollinearity. Two factors are typically posited as causes of increased mortality risks: chronological age and severity of illness. Both chronological age and severity of illness are highly correlated with days until death. Yet, when severity of illness is used to predict mortality, the correlation between chronological age and mortality may disappear. Risk of mortality depends on the nature of patients’ illnesses, and not aging. If the patients’ severity of illness was adequately measured, their chronological age may no longer be relevant.

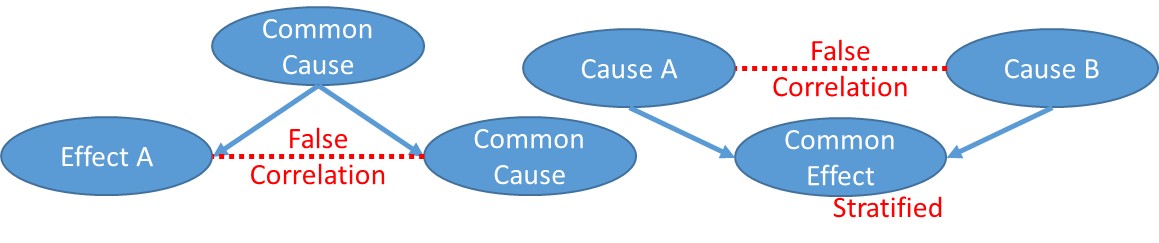
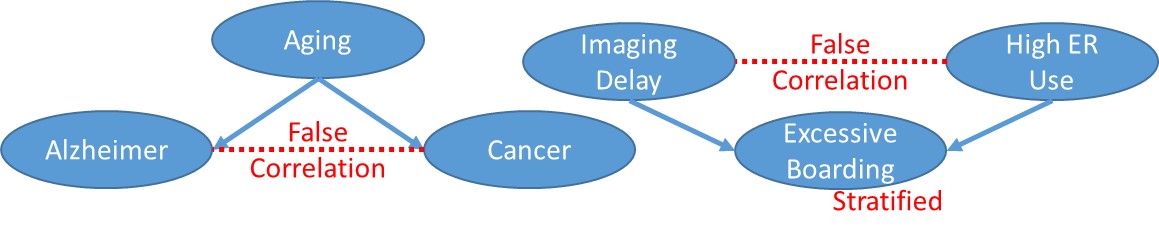
Correlation and causation have a tortured relationship filled with misunderstanding. Of course, correlation does not imply causation. Even strong correlation, of +1 or –1, does not necessarily imply causation. To interpret correlation between *x* and *y* as evidence of strong causation, several other assumptions are necessary, including sequence (*x* must precede *y*), mechanism (there must a way in which changes in *x* lead to changes *y*), and counterfactual (if *x* is removed, the frequency *y* should decline).

Not all correlations are real, meaning that some correlations are artifacts of the relationships among other variables. When two variables have a common cause, the two variables *covary*. Common cause creates an apparent correlation (see the top left side of exhibit 19.14) where no relationships may exist. The two variables are really unrelated even though they co-vary; the relationships exist only when the common cause is present. For example, aging causes both cancer and Alzheimer’s disease (see exhibit 19.14, lower left side). The existence of the correlation between these two diseases is not a sign that they are related but that they have a common cause.

If two variables have a common effect, stratifying the common effect will also create a false correlation between the two variables (see the top right side of exhibit 19.14). Again, this correlation exists when the common effect is stratified and does not exist otherwise. For example, Kheirbek and colleagues (2015) examined causes of excessive boarding time in emergency departments (EDs). They identified 26 different causes. Backup in the hospital’s imaging center may lead to ED delays, as well as longer stays in the hospital. Patients may have to wait an unnecessarily long time for images to be processed. If one examines the data, there is a surprise correlation with excessive boarding. This correlation is spurious (false) because when there is no backup in the image processing center and among patients with excessive boarding time, there is spurious correlation between demand for emergency room and backup of imaging services. When two variables are correlated in one subgroup and uncorrelated in another, it may be a signal that the correlation is spurious.

**[INSERT EXHIBIT]**

**Exhibit 19.14** Examples of Spurious Correlations Among Unrelated Variables



**[END EXHIBIT]**

**[H1] Mutual Information**

One way to understand whether two variables are interrelated (dependent) is to calculate the mutual information across the two variables. The mutual information between two variables, *x* and *y*, is defined as the weighted sum of the log of probability of both events divided by the product of the probability of each event, written in mathematical terms as

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, is the joint probability of observing both *x* and *y* values, is the marginal probability of observing *x*, and is the marginal probability of observing the *y* value. Note that when the joint and product of marginal probabilities are close to each other, the calculated mutual information between the two variables is close to zero. For mutual information, a value closer to zero implies independence, and a value farther away from zero implies dependence. The range of values for mutual information is from zero to plus infinity. If we are to use mutual information to determine independence or dependence, typically a threshold is specified. For example, we can say mutual information smaller than 0.01 is interpreted as independence, and a value larger than 0.01 is interpreted as dependence. This threshold value must be tinkered with to gauge what is appropriate and differs from data set to data set.

Mutual information can also be calculated from counts of occurrences of the *x* and *y* variables:

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, is the number of times *x* and *y* co-occur; is the number of times *x* occurs; is the number times *y* occurs; and *N* is the total sample size. These counts can be easily calculated through SQL code. For *x* and *y* variables having true or false values, the SQL code is provided in exhibit 19.15.

**[INSERT EXHIBIT]**

**Exhibit 19.15** Calculating Mutual Information Using SQL Code

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X** | **Y** | **SQL** | **Count** | **Variable** |
| TRUE | TRUE | SELECT Count(\*) as total  FROM data  WHERE X=True and Y=True | 3,791 |  |
| TRUE | FALSE | SELECT Count(\*) as total  FROM data  WHERE X=True and Y=False | 1,276 |  |
| FALSE | TRUE | SELECT Count(\*) as total  FROM data  WHERE X=False and Y=True | 6,833 |  |
| FALSE | FALSE | SELECT Count(\*) as total  FROM data  WHERE X=False and Y=False | 8,100 |  |
|  |  | SELECT Count(\*) as total  FROM data | 20,000 | *N* |
| TRUE |  | SELECT Count(\*) as total  FROM data  WHERE X=True | 5,067 |  |
| FALSE |  | SELECT Count(\*) as total  FROM data  WHERE X=False | 14,933 |  |
|  | TRUE | SELECT Count(\*) as total  FROM data  WHERE Y=True | 10,624 |  |
|  | FALSE | SELECT Count(\*) as total  FROM data  WHERE Y=False | 9,376 |  |

**[END EXHIBIT]**

The mutual information between *x* and *y* can be calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

The conditional mutual information can also be calculated in the same fashion as the mutual information, but this time all counts are taken where the condition has occurred. Like mutual information, the value of conditional mutual information is in the range 0 to plus infinity, where values closer to zero imply conditional independence and values farther from zero imply conditional dependence. One must also specify a cutoff value to separate conditional independence from conditional dependence.

# [H1] Poisson Regression and Tests of Dependence

There are many ways to test the independence of variables. We have already described the procedure for calculating the difference between observed and expected counts of events under different assumptions of independence. The approach leads to a chi-square test of independence but is cumbersome to carry out when there are many variables. An alternative is to use log-linear Poisson regression. Poisson regression is well suited for an analysis of count of co-occurrences and can easily turn these counts into measures of association between the variables. Poisson regression assumes the response variable *y* has a Poisson distribution; the logarithm of its expected value can be modeled by a linear combination of other variables.

In a Poisson distribution, the probability of observing *k* items can be calculated from the formula

**[INSERT EQUATION]**

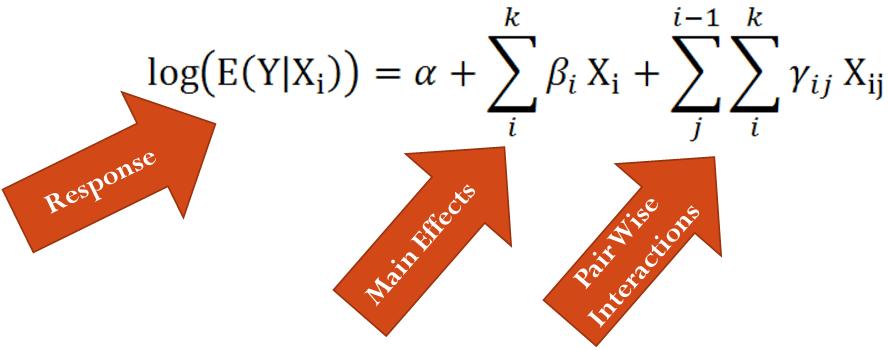
**[END EQUATION]**

In this equation, mu, (the Greek letter indicates the expected value of the probability distribution. *k* is a constant, and it is typically the count of an event; it goes from 0 to infinity or, if Poisson is used as an approximation, a relatively large number. The letter *e* indicates the famous irrational number, with the first few digits 2.718. The expected value of a Poisson distribution is the mean of the distribution, which provides us with an easy way to predict what will happen in the future based on a Poisson distribution.

A count of co-occurring events is likely to have a Poisson distribution when the number of trials is large (as in the case of Big data), each trial is independent, and the expected value of the response variable is assumed to be constant over time. These assumptions are likely to be met when we have count data for combinations of events. First, each combination represents a pattern of occurrences across a large number of events. Second, the probability of any combination is relatively small, especially when the number of variables is large. This is the case in almost any contingency table with more than five variables. There are at least two to the power of five (32) possible combinations; if each combination is equally likely, there will be 1 divided by 32, or 0.03 probability of the event occurring, which is relatively small. So it is likely that count data with five or more variables has a Poisson distribution. Of course, the distribution must be tested empirically to verify that the probability is small.

In a homogenous Poisson regression, the response variable is the log of count of combination of events, and the independent variables are either the main effects of each event or pair-wise combination of events (see exhibit 19.16). When a pair of events has a statistically significant impact on the response variable, these two events occur together frequently and, holding all other variables constant, are dependent on each other.**[INSERT EXHIBIT]**

**Exhibit 19.16** Poisson Regression



**[END EXHIBIT]**

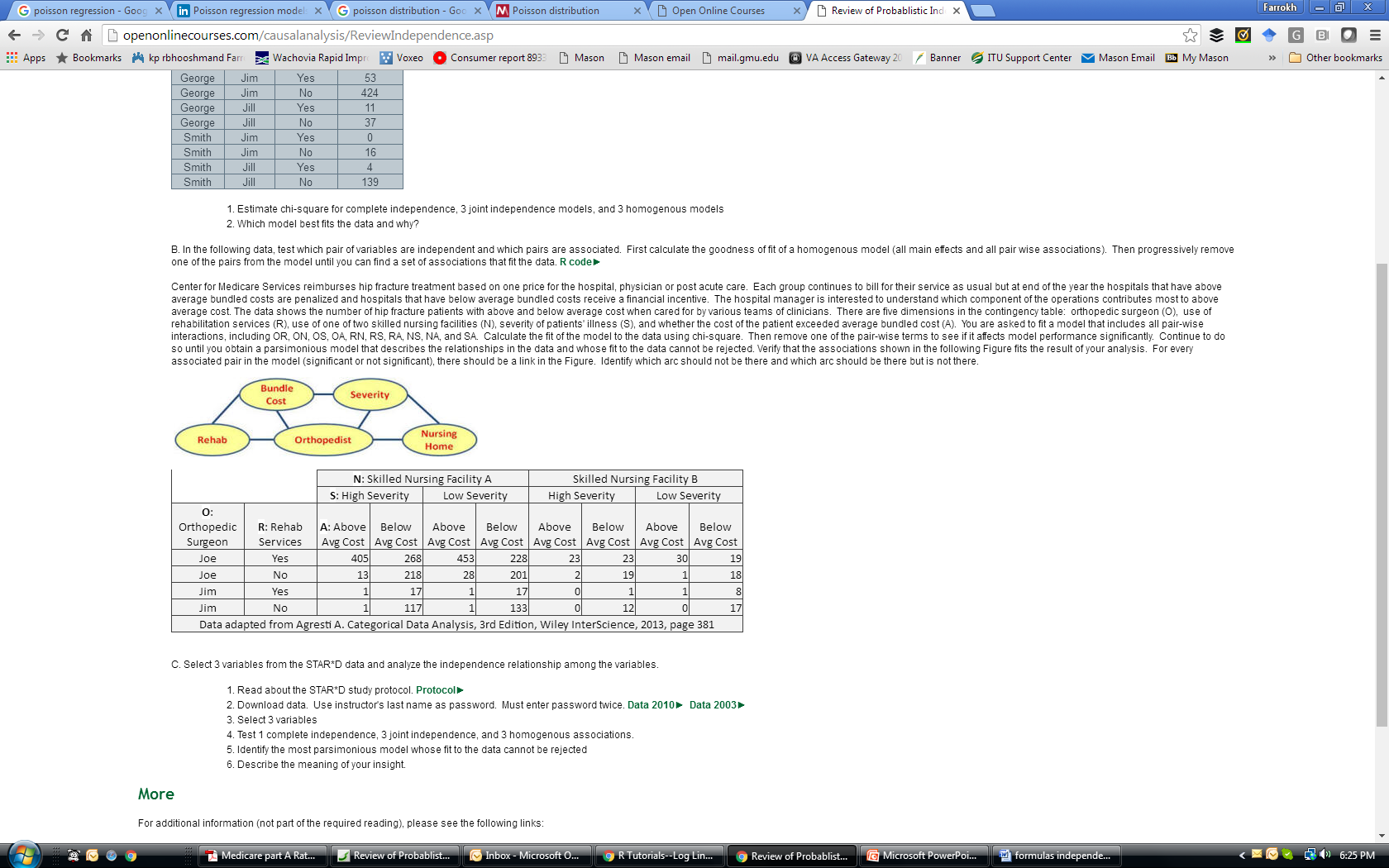
In Poisson regression, when two events have a statistically significant impact on the count of co-occurrences, these two events are occurring frequently together and, holding all other variables constant, are dependent on each other. Poisson regression differs from other regression models in a key manner: It verifies dependencies among all pairs of variables, not just independent variables and a single dependent variable. In Poisson regression, the variation in the count of combinations of variables is explained by associations among pairs of variables. If we were to describe every variable as a node and connect all statistically significant relationships as arcs between the nodes, Poisson regression discovers relationships among all pairs of independent variables. These relationships are not with a single outcome variable. The picture that emerges is a complex network that shows the interrelationship among pairs of independent variables.

When using Poisson regression to construct a network model, often the focus is not on the statistical significance of pair-wise relationships but on goodness of fit. In Poisson regression, the fit between the model and the data is reported using *G-squared*; this statistic is calculated as twice the difference of log likelihood of the saturated model (where every data point has its own parameters) and the model. The simplest model is the null model, in which the intercept is used to predict all data points. If the G-squared is small, it means that the intercept explains the data pretty well. If two models are compared, the model with the higher G-squared is a better fit to the data. G-squared has a chi-square distribution, and therefore the statistical significance of G-squared can be readily tested.

## [H1] Example Construction of Association Network

To demonstrate the use of log linear Poisson modeling, we will explore the data provided in exhibit 19.17. These data record the times a hospital faced above- or below-average bundled payment adjustments. The hospital manager wants to know which of the various areas might have contributed to above-average cost. The data reported are counts of events. There are many observations, but some cells have few counts. To analyze these data and estimate the homogenous association among any pair of the variables, we start with a model in which all homogenous variables are present.

**[INSERT EXHIBIT]**

**Exhibit 19.17** Combined Impact of Orthopedic Surgeons and Nursing Facilities on Cost 

*Note*: Data adapted from Agresti (2013).

**[END EXHIBIT]**

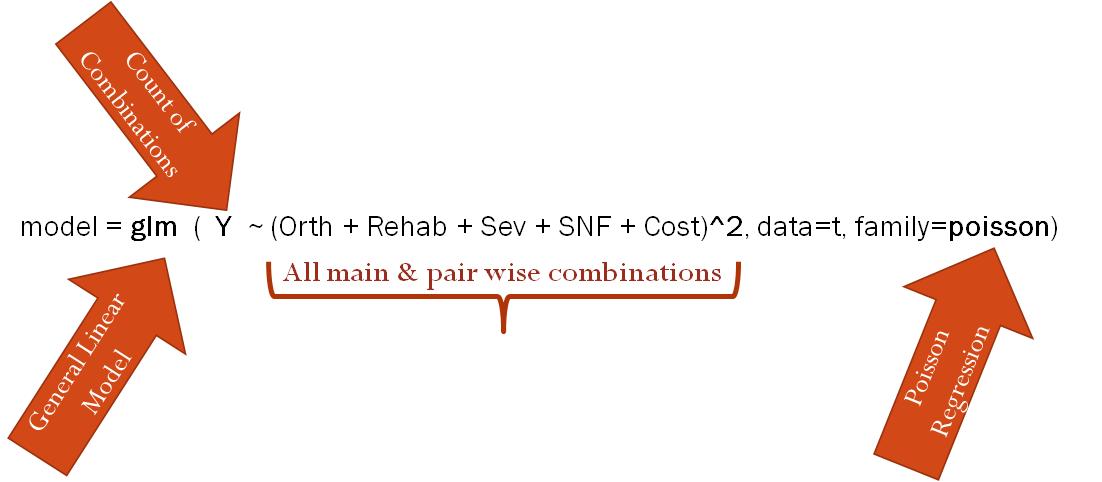
The R code for a generalized linear model using Poisson distribution is given as:

**[INSERT EQUATION]**

model = glm (*y* ~ (O + R+ S + N + A)^2, data = Bundle, family = Poisson).

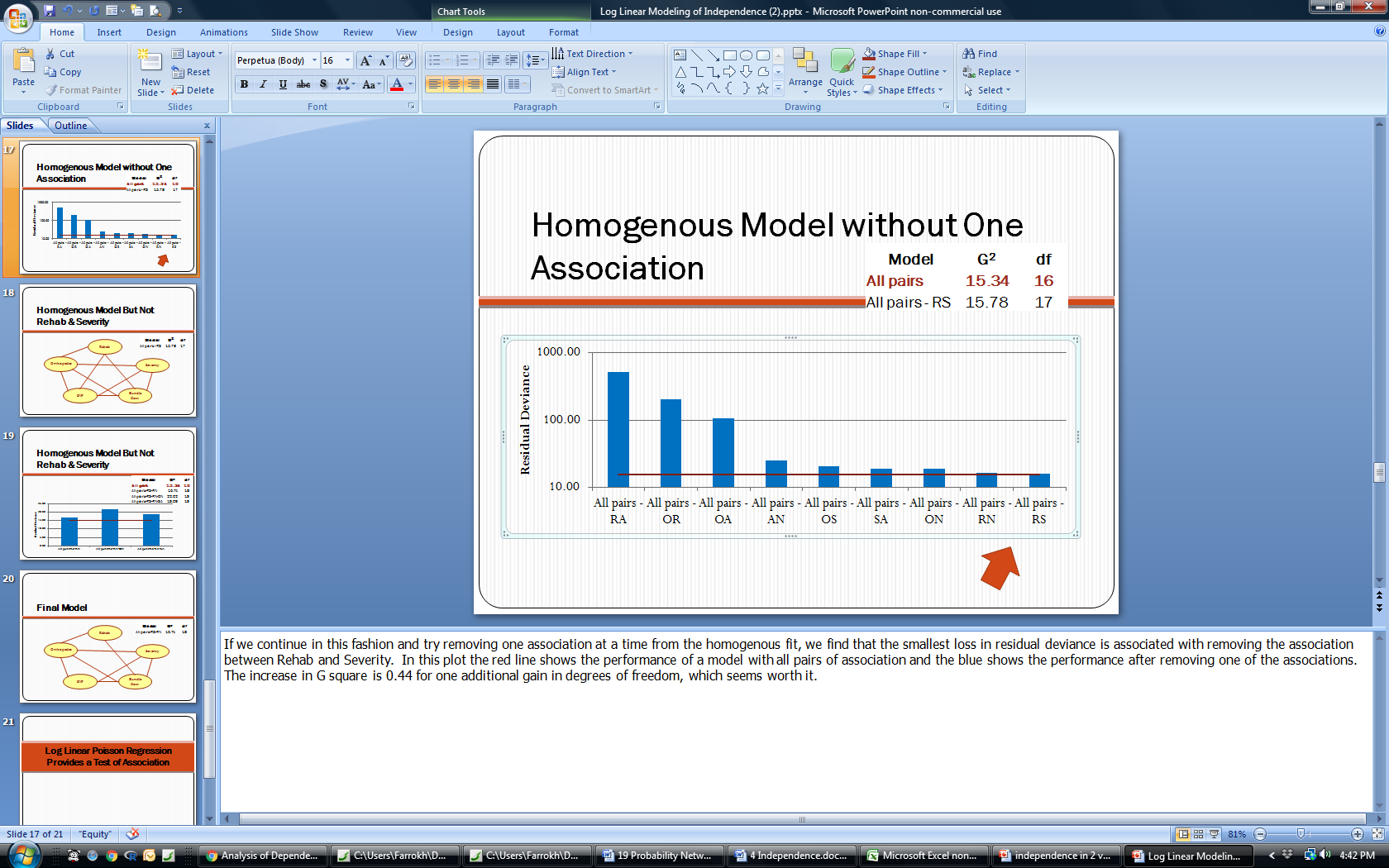
**[END EQUATION]**

In this R code, O stands for orthopedic surgeon, R for rehabilitation facility, N for skilled nursing facility, S for severity level of the patient, and A for whether the patient had above-average cost. We are using the general linear model package with Poisson distribution for the response variable. The variable *y* contains the count of combinations of the variables. This variable is regressed on five variables, as well as pair-wise combinations of these five variables. In this R statement, the use of ^2 means that all pair-wise combinations of the variables are included. This is referred to as the *homogenous model*, as all possible pair-wise associations are present. The homogenous model has a G-squared of 15.34, with 16 degrees of freedom. This is the best performance we can have, and as we drop various pair-wise links in the model, the performance of the model will deteriorate (though, we can hope, not by much). Because all pair-wise associations are included in the homogenous model, the network depiction is one in which all variables are shown to be associated with each other. Exhibit 19.18 shows the relationships assumed in this model, as expected all pairs of variables are related to each other.

If the analyst drops the association between rehabilitation and bundled cost, the G‑squared measure of goodness of fit is 513.47 with 17 degrees of freedom. When we say that we are removing the association between rehabilitation and bundled cost, we mean that the product RA is removed from the R code. Comparing this model with the homogenous model, the exclusion of rehabilitation and bundled cost has led to a significantly worse fit for an additional 1 degree of gain in degrees of freedom. So this is not a good idea. Bottom of Form

If the analyst continues in this fashion and tries removing other associations one at a time from the homogenous fit, then he will find that the smallest loss in residual deviance is associated with removing the association between rehabilitation and severity, RS product. In exhibit 19.18, the red line shows the performance of a model with all pairs of association, and the blue bars show the performance of the model after removing one of the associations. Removing rehab and severity makes sense, as the increase in G-squared is 0.44 for one additional gain in degrees of freedom. This seems like a small cost to pay for a more parsimonious model.

**[INSERT EXHIBIT]**

**Exhibit 19.18** Performance of Models After Removing Specific Relationships

**[END EXHIBIT]**

In exhibit 19.19, we have removed the link between rehabilitation and severity variables (RS). Keep in mind that removal of the link means that the two variables are independent in the context of other variables that are present. We now look at other models, all without the association of rehab and severity, and compare them to the performance of the homogenous model that includes all remaining pair-wise relationships. In exhibit 19.19, we see that the smallest increase in residual deviance occurs when we remove the association between rehabilitation and the skilled nursing facility (RN). All other removals seem to be too large for additional gains in degrees of freedom.

This yields the final log linear Poisson regression for our data with the best possible fit. This model is shown in exhibit 19.20. Note that in this network, there are no arcs between rehabilitation and severity or between rehabilitation and nursing home. Removing these links led to small losses in goodness of fit and resulted in a more parsimonious model. These pairs of variables do not explain a great deal of variation in co-occurrences of the events and therefore can be assumed to be unrelated or, we might want to call them nearly independent. This does not mean that these variables are uncorrelated. Keep in mind that every pair of variables in a network model is either directly or indirectly related to each other. It simply means that, with this new model, and in the context of other variables in the model, these pairs are independent.

**[INSERT EXHIBIT]**

**Exhibit 19.19** Performance of Models After Removing Additional Links

|  |  |  |
| --- | --- | --- |
| **Model** | **G2** | **Degrees of Freedom** |
| **All pairs** | **15.34** | **16** |
| All pairs – RS - RN | 16.74 | 18 |
| All pairs – RS – RN - ON | 22.02 | 19 |
| All pairs – RS – RN - SA | 19.09 | 19 |

*Note*: RS = link between rehabilitation and severity, RN = link between rehabilitation and nursing home,ON = link between orthopedist and nursing home, and SA = link between severity and bundled cost.

**[END EXHIBIT]**

**[INSERT EXHIBIT]**

**Exhibit 19.20** Association Among Variables That Predict Bundled Cost

**[END EXHIBIT]**

# [H1] Summary

Chapter 19 has shown how dependence—or association among variables—can be verified, first in two variables through chi-square, then in three variables through chi-square, and finally in more than three variables through Poisson regression. As the number of variables increases, the use of statistical tools can help in the measurement of dependence. The chapter also warned that associations can be spurious. False correlations can be a function of common causes or a consequence of stratifying common effects.

# [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL code, and other related material are on the course website.

**[H1] References**

Agresti, A. 2013. *Categorical Data Analysis*, 3rd ed. Hoboken: Wiley.

Kheirbek, R. E., S. Beygi, M. Zargoush, F. Alemi, A. W. Smith, R. D. Fletcher, P. N. Seton, and B. A. Hawkins. 2015. “Causal Analysis of Emergency Department Delays.” *Quality Management in Health Care* 24 (3): 162–66.