Dataset Quality Assessment for Business Analytics Use

Dr. James G. Wendelberger
Los Alamos

Abstract

Electronic datasets are received for use in company business products and services. These datasets do not always contain the data that is expected. Assessing dataset quality is important and desirable. An assessment process is defined which breaks down the assessment into five steps. Step one is a check of the reasonableness of the file for a "big picture" view of the dataset. It confirms basic file information to verify that the file and format of the contents are as expected. The next steps test individual data values, individual variable distributions, multivariate variable distributions and the likelihood that the data is correct based upon external information, constraints or assumptions about the data. Here we present various examples of each of these assessments. Passing the data quality assessment tests results in acceptance of the received dataset. Failure(s) of the data quality assessment test(s) results in either data corrections/imputations or in the extreme case a new dataset request to the vendor. If a new dataset request is needed, then the specific assessment test failure(s) may be are disclosed to the vendor.

Key Words: Dataset, Quality, Assessment, Testing, Business Analytics, Vendor

1. Introduction

A company receives data from many sources. This data is not always what was expected and possibly is with flaws. These flaws may include missing or aberrant values. It is the purpose of the data assessment system to identify these flaws. The classic company case of this is the aberrant data detection system that uses the inverse Beta distribution(see Appendix A) for suspect data value detection. This system uses the high level brand market (national or state level) share information to determine if low level (census tract or block group) values of brand registration data are consistent with these high level brand shares. It is desirable to have similar systems that detect potential data quality issues for other types of data at the company.

Data other than the new vehicle registration data may include: other automotive data, such as, new vehicle sales by dealership or lead data for potential buyers; demographic data, such as, population, age, gender, neighborhood cluster codes; or economic data, such as, income, employment or wealth data.

It is desired to provide a process to apply to specific data that a company receives. This process is described next.

2. Process

At a very high level the process to assess data is summarized as: define the data and define the data detection mechanism. In the registration data aberrant data detection system the data is registration data by brand at a very low level of geography, such as, census tracts or block groups, and the data detection mechanism is testing the fidelity of the data to the national levels of market share.

The data definition is that of interest for the purpose at hand. It may consist of automotive, demographic, economic or other types of data. Once the data is identified one must create an aberrant data detection system for the specific data. This is done in multiple parts. The first part is a single value by variable detection. The second part looks within a specific variable for unusual or extreme issues in each data variable. The third part looks for extreme aggregates of multiple variables. And finally, the forth part looks for aberrations related to external or aggregate data knowledge, information or constraints.

2.1 Step 1: Reasonableness Checks

Even before beginning the well-defined process below one should check the dataset for obvious issues that may indicate problems. Is the dataset digital size that which is expected? Does the dataset have the proper file extension? Does a look at the data reveal characters, numbers and/or images or is it a corrupt jumble of uncommon symbols? Is there an identifier for an important aspect of the data, such as, geographic unit (GU) identifier, dealer name, vehicle identification number (VIN), etc.? After looking at the data, or part of the data, in a computer file software editor does it appear to be as expected? There are many of these "common sense" tests that should be conducted before proceeding to the rest of the well-defined assessment process below. In particular, does the dataset match the specifications that were provided to the vendor?

2.2 Step 2: Single Value Process Checks

Single value detection looks at each individual data value and asks the question: is this value possible or reasonable? So for example, negative values of sales, may or may not be possible or reasonable or buyers of age less than 5 years old may be impossible or unreasonable. Another common check answers the question: are non-numeric data supplied for numeric fields? To accomplish this step requires the user to specify individual value tests that will be performed on each value in the data set. The users obligation here is to supply the tests to be undertaken. It is suggested that the user simply select from a common set of tests. This common set includes: character or numeric and within range tests. For the within range variable test the user must supply for each data value the allowed range. Using population as an example, one would expect the population of a low level piece of geography to be between the values of zero and the population of the country containing the low level pieces of geography.

2.3 Step 3 Single Variable Process Checks

Single variable detection is used by variable for each variable in the data set. An example of this includes the overall data distribution for the variable in question. If for example we have a continuous variable such as income and it only takes on the two values: zero and one, then we may have discovered a data problem. In a similar fashion if the data variable may take on one of 50 values, say in the cluster coding of a neighborhood and we find that a variable takes on 2,000 values then there may be a data issue of concern or problem. The user's obligation in this case is to select relevant data distribution tests and associated violation parameters. For example, a lead in a specific timeframe may result in a closed sale or in no sale. This variable may expect at most three values: sale, no sale or missing. These may be indicated by the values: 1, 0 and 9 for example. The test would be to look at the distribution for this variable and see that it contains at most three values and whether or not they are: 0, 1 and 9? The user would need to supply the knowledge that we are looking for a discrete distribution, with at most

three values, which are numeric and take on the values 0, 1 and 9. As with the prior single value detection the user may select from a group of tests. This group will include: discrete or continuous, number of values anticipated and compliance with the exact values allowed. A very common test here is to identify variables which have the same value for all observations. This may (or may not) indicate a dataset problem, if for example the variable is population of a census tract. Or, it may not indicate a problem, if for example the variable is an indicator of census tracts with both males and females as members of their population or the total country population for one country of consideration.

2.4 Step 4 Multivariate Distribution Process Checks

The third data check is to look for extremes when considering all of the data as one large multivariate distribution. One of the most common tests here might be for a variable which has the value “missing” for all variables of a specific row, vehicle or consumer. Or the value zero for all. The zero value may or may not make sense in the context of the data problem or issue under consideration. The missing value may or may not add information to the data. One may also test here whether the “missing" value indicator(s) is properly used in the data set. Or was, zero or some other value used to indicate missing. The user has much latitude here and may check for example if a geographic unit that has very large sales for one brand and zero sales for another competitive brand exist. This may or may not indicate a possible data problem. Specific test of this type include: looking for missing or zero rows or other unexpected or unexplained patterns in the entire data set.

Another part of the fourth data check is to identify a (possibly unique) identify, if one exists, for the data rows and/or columns. Are the identifiers as expected? If they should be unique then are they unique? Are the variable names as expected? Is the data in the structure that you anticipate? That is if, for example, there are dealership images embedded in the data, then are the images where they should be located and uncorrupted? Are there the right number of variables and observations? Questions such as these should be confirmed when looking at the data as a whole.

2.5 Step 5 External Information Probabilistic Process Checks

The fifth type of data check scheme requires the most involvement of creativity and ingenuity on the part of the user. This check uses the entire data set and other external information to check for data reasonability. The inverse Beta detection test mentioned earlier (Appendix A) is of this type. Other examples are: checking that totals of aggregations of some variables equal other sums or total values. For example, do the age cohorts sum to the population amount for each GU?

It has been common practice to evaluate both univariate descriptive statistics on variables and cross tabulations of all bivariate combinations in the dataset. If the assessment is done properly then this common evaluation will be part of the assessment process described above.

3. Evaluation of Assessment results

Passing all assessment tests does provide a degree of assurance that the dataset is what it is purported to be. However, even a dataset that passes a rigorous assessment should still be continually scrutinized during use.

Evaluation of the assessment will provide pass and fail grades for each test as well as, in the case of failure, specific issues to investigate. Failure of a few tests on a few limited observations may only be

indicative of real data and not of full dataset failure. On the other hand failure of any of the assessment tests requires investigation as to the root cause. Often what looks like a minor failure turns out to be the result of catastrophic data problems or issues of concern. Think of these detected problems as they may be “the tip of the iceberg.”

There are also probabilistic data test results, such as, Beta I. These probabilistic results identify unlikely, but not impossible events. These tests are based upon a statistical model and structure that may only approximate the actual data structure. The model may assume independence, unbiasedness or have other assumptions that may only hold approximately. Thus the identified unlikely events should be investigated to determine if they are simple the result of Type 1 error or if indeed they have identified a true underlying problem with the dataset.

When evaluating probabilistic assessment results it is important to be aware of and account for multiple testing corrections to probability assessments. Due to the large number of tests that may be made as part of this assessment process one would expect to encounter false positives by chance. The number of false positives identified can be reduced by proper accounting of the multiple comparisons. That is the more tests that are done the more likely that the Type 1 error will identify perfectly normal data as aberrant. To adjust for this the Type 1 error of each individual test should be reduced accordingly. The most common and one of the easiest ways to do this is by using the Bonferroni correction to the alpha level.

4. Relevant Data Practices

Given the data detection system there are other data practices that should be adhered to. One of these practices is data time stamping. When was the data set and individual data pieces received? When was the data set uploaded into the file or software system?

Data repair is not encouraged and is a very risky business. If undertaken, then at a very minimum it should be disclosed to the users of the data and to the users of any of the data products that may result from its use. This may present legal hazards and should only be done in consultation with appropriate legal advice, warnings and consent of the company legal team.

The failure of any of the tests undertaken in the process should be investigated. It is common that failure of one of these tests actually is indicative of a much larger data problem or issue.

In the probabilistic data checks, such as Beta I, there may be false positives. As all violations should be investigated these should be identified and then investigated in a case by case manner. False positives are a necessary aspect of any reasonable probabilistic test of this kind.

In general data set problems or issues are not statistical estimation matters. They may be identified by statistical tests and are then "fixed" by proper data collection or assembly methods

One common data set problem is that missing values are sometimes replaced by zeroes. This is a practice that should be avoided. Missing values should be recorded as missing and not as zeroes. For example someone’s age may be missing. This may not mean they are zero years old.

There are also ways of imputing problamatic or missing data. This is a practice that is fraught with trappings. Not the least of which is legal liability. The practice may induce spurious relationships,

invalidate model probabilities and/or infect parameter estimates. Imputing data values is a process that is best left for a professional statistician.

5. Conclusion

The overall assessment process is: accept delivery of the data set, construct the diagnostic tests described in the Process section (2) above, perform the tests on the dataset, and investigate violations of any of the tests. If violations do not exist or are identified as acceptable then utilize the data for company products and processes. If violations are not acceptable then the dataset needs to be returned and the issues need to be fixed by the data supplier.

Appendix A: Aberrant Automotive Retail Registration Data Detection

1.0 Summary

A retail automotive registration data aberration detection method is described. The procedure identifies retail automotive data, called aberrant data, which may have not properly had fleet data removed, data which may have been influenced by the proximity to the retail network outlet locations, data which may have been otherwise influenced by consumer behavior, or data that is extreme due simply to unknown and unmeasured random variations in data influencing factors (lurking variables). Aberrant data may represent a registration-clumping situation. The method has, as a null hypothesis, a binomial model for the data. The incomplete beta function is used to compute the attained significance level for the observed data. An algorithm to implement the method is described. The method does not identify the cause of the aberration. The method identifies particular geographic units and make/segment/brand data for further scrutiny.

2.0 The Aberrant Retail Registration Data Detection Method

Step 1: The Make/Model Number per Geographic Unit

Determine the retail segment number of registrations for each make/model for each geographic unit in the national data set and define this as the **make/model number per geographic unit**. Call the make/model number per geographic unit $a_{m,i}$ for each make/model m and each geographic unit i. The make/model types are numbered from 1 to M.

Step 2: The Total Number per Geographic Unit

Aggregate the make/model numbers per geographic unit for each geographic unit in the national data set and define this as the **total number per geographic unit**. Call the total number per geographic unit t_i for each geographic unit i. For each i, calculate the total number per geographic unit as: $t_i = \sum_{m=1}^{M} a_{m,i}$. The geographic units are numbered from 1 to I.

Step 3: The Make/Model Market Share Standard

At the National level, calculate the average retail segment market share for each make/model and define this as the **make/model market share standard**. Call the make/model market share standard ms_m for each make/model m. For each m, calculate the make/model market share standard as: $ms_m = \sum_{i=1}^{I} a_{m,i} \Big/ \sum_{i=1}^{I} t_i$.

Step 4: The Complimentary Make/Model Market Share Standard

At the National level, calculate the average retail segment market share for each make/model and define as the **complimentary make/model standard share**. Call the complimentary make/model market share standard cms_m for each make/model m. For each m, calculate the complimentary make/model market share standard as: $cms_m = 1 - ms_m$.

Step 5: Determine The Total Number Of Statistical Tests

Each make/model number for each geographic unit will be tested against the standard. The total number of statistical tests is the number of geographic units, I, times the number of make/models, M, or $T = M \cdot I$. If only a subset of the data is to be tested then only those make/models and geographic units should be used to determine the number of statistical tests.

Step 6: Set the Overall Significance Level

The overall statistical level is often set at 5%. This means that if the null hypothesis is true that we are willing to reject the null in favor of the alternative 5% of the time. The overall significance level may not be set equal to zero because if set to zero it will maximize the probability of not rejecting the null hypothesis when in fact the alternative is true. For this reason, a small but acceptable value is selected. A value of 5% is often used in the automotive network analysis field.

Step 7: Determine The Individual Test Significance Level

The individual test significance level, α', is computed as follows: $\alpha' = 5\%/T$. The individual test significance level is used to determine the aberrant data. Attained significance levels less than the individual test significance levels are classified as aberrant. Why do we need to modify the individual test significance level? Why is it different from the overall significance? If we do not modify the individual test significance accordingly then it becomes very likely, higher than 5%, that we will find significance where there is none. For this reason each individual significance test uses an individual test significance level, smaller than 5%, to achieve the desired overall significance level of 5%. This is called a multiple comparison correction or modification. This particular correction is called the Bonferoni correction. The Bonferoni correction modifies the overall significance level by dividing it by the total number of statistical tests to yield the individual test significance level of $\alpha' = 5\%/T$.

Step 8: Compute The Attained Significance Level For Each Make/Model Number Per Geographic Unit

The attained significance level for each make/model number per geographic unit, $\alpha_{attained}$, is computed. It is computed by determining the probability of seeing the event we see, $a_{m,i}$, or a less likely event under the binomial hypothesis. This probability is called the attained significance level (for each make/model number per geographic unit). The binomial hypothesis is that the data are generated from a binomial process with probability ms_m and number of trials t_i. The attained significance level is computed as follows:

Step 8.1 Utilizing the binomial hypothesis, define the probability, $p_{m,i}(a)$, of observing a number, a, of make/model m in geographic unit i.

$$p_{m,i}(a) \equiv \binom{t_i}{a} (ms_m)^a (cms_m)^{t_i - a}$$

Step 8.2 Computation of $p_{m,i}(a)$ is via the incomplete beta function. The incomplete beta function may be found in suites of mathematical subroutines, such as, IMSL.

$$p_{m,i}(a) = \begin{cases} 1 - I_{ms_m}(1, t_i), & \text{for } a = 0 \\ I_{ms_m}(a, t_i - a + 1) - I_{ms_m}(a+1, t_i - a), & \text{for } a = 1, \ldots, t_i - 1 \\ I_{ms_m}(t_i, 1), & \text{for } a = t_i \end{cases}$$

Where

.0

$$I_x(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \int_0^x t^{a-1}(1-t)^{b-1}\, dt, \ (0 \le x \le 1), \ \Gamma(a) = (a-1)!.$$

Step 8.3 Compute the probability of the observed data value $a_{m,i}$ under the binomial hypothesis, $p^*_{m,i}$. This is done using the incomplete beta function as defined in the computation of $p_{m,i}(a)$ as in step 8.2.

$$p^*_{m,i} \equiv \binom{t_i}{a_{m,i}} (ms_m)^{a_{m,i}} (cms_m)^{t_i - a_{m,i}} = p_{m,i}(a_{m,i})$$

Step 8.4 Compute the attained significance level for make/model m and geographic unit i. The probabilities for the observed data and all less likely data, $\{a \mid a = 0, \ldots, t_i \wedge p_{m,i}(a) \le p^*_{m,i}\}$, are theoretically added together.

$$\alpha_{attained} = \sum_{\{a \mid a=0,\ldots,t_i \ \wedge\ p_{m,i}(a) \le p^*_{m,i}\}} p_{m,i}(a)$$

The computation does not proceed directly but rather makes use of the cumulative nature of the inverse binomial and the unimodal nature of the binomial distribution as follows.

Step 8.4.1 Determine the first tail probability, $tail_{m,i,1}(a_{m,i})$. Compute

$$tail_{m,i,1}(a_{m,i}) = \min\left(1 - I_{ms_m}(a_{m,i}, t_i - a_{m,i} + 1) + p_{m,i}(a_{m,i}), I_{ms_m}(a_{m,i}, t_i - a_{m,i} + 1)\right)$$

Step 8.4.2 Determine which values are required for the second tail. If $tail_{m,i,1}\left(a_{m,i}\right)=1-I_{ms_m}\left(a_{m,i},t_i-a_{m,i}+1\right)+p_{m,i}\left(a_{m,i}\right)$ then the second tail requires larger values and the first tail is for smaller values. If $tail_{m,i,1}\left(a_{m,i}\right)=I_{ms_m}\left(a_{m,i},t_i-a_{m,i}+1\right)$ then the second tail requires smaller values and the first tail is for larger values.

Step 8.4.3 Determine the start of the second tail. If the larger values are required for the second tail then find the smallest integer y in the sequence $a_{m,i}+1,a_{m,i}+2,\ldots,n$ for which $p_{m,i}\left(y\right)\le p_{m,i}\left(a_{m,i}\right)$. If such a value of y exists then call this value y^*. If no such y exists then note that y^* does not exist. If the smaller values are required for the second tail then find the largest integer y in the sequence $a_{m,i}-1,a_{m,i}-2,\ldots,0$ for which $p_{m,i}\left(y\right)\le p_{m,i}\left(a_{m,i}\right)$. If such a value of y exists then call this value y^*. If no such y exists then note that y^* does not exist.

Step 8.4.4 Use the incomplete beta function to compute the second tail probability, $tail_{m,i,2}\left(a_{m,i}\right)$. If the value of y^* does not exist then set $tail_{m,i,2}\left(a_{m,i}\right)=0$. If y^* exists and the larger values are required for the second tail then set $tail_{m,i,2}\left(a_{m,i}\right)=I_{ms_m}\left(y^*,n-y^*+1\right)$. If y^* exists and the smaller values are required for the second tail then set $tail_{m,i,2}\left(a_{m,i}\right)=1-I_{ms_m}\left(y^*+1,n-y^*\right)$.

Step 8.4.5 The attained probability is: $\alpha_{attained}=\min\left(tail_{m,i,1}\left(a_{m,i}\right)+tail_{m,i,2}\left(a_{m,i}\right),1\right)$.

Step 9: Determine the Aberrant Set of Data

Compare the attained significance level for each make/model number per geographic unit to the individual test significance level. Classify the make/model number per geographic unit as aberrant whenever $\alpha_{attained}\le\alpha'$. Alternatively, one can assign the attained significance level to the corresponding data and sort the data set by the attained significance level. Then scrutinize the data starting with the smallest attained significance levels. In this case an arbitrary cutoff point of attained significance can be set to determine the aberrant set of data. Aberrant data may represent a registration-clumping situation.

Step 10: Analyze the Data Characterized as Aberrant

Analyze the data in an attempt to determine causes for the unlikely data observed as aberrant. Possible causes are that the data may not have had fleet data properly removed, data may have been influenced by the proximity to the retail network outlet locations, data may have been otherwise influenced by consumer behavior, or data may be extreme due simply to unknown and unmeasured random variations in data influencing factors (lurking variables).

3.0 Notes and Comments

The computation of the binomial probabilities directly may cause overflow problems. For this reason the incomplete beta function is used. The incomplete beta function subroutine in IMSL seems adequate for the needs of this exercise. However, it occasionally produces inaccurate results. The inaccuracies observed were for very small probabilities ($< 10^{-10}$) and may not be large enough to change the classification of aberrant data for reasonable size overall significance levels. Occasionally, negative values were returned for the probabilities. It has been suggested that when negative values are returned that they be set equal to zero to help minimize the inaccuracy. Any software package that accurately computes the binomial probabilities may be used.

Attached here is a table for some unidentified make/model and geographic unit data with the attained significance levels. The values in this table may be used for checking any new code.

Make/Model Registrations	Total Registrations	Market Share Standard	Attained Significance
47	51	0.111	2.12854642E-40
11	14	0.258	5.47306114E-05
16	32	0.223	8.47722882E-04
13	15	0.459	1.49102345E-03
9	10	0.459	7.45505195E-03
8	8	0.161	4.51447246E-07
0	44	0.161	7.15055991E-04
1	48	0.161	4.79224172E-03
8	10	0.228	2.09084445E-04
1	51	0.228	6.67657342E-05
11	15	0.189	7.02301360E-06
46	104	0.132	9.35440956E-15
9	11	0.308	7.17794546E-04
31	40	0.364	1.50610384E-07
12	18	0.364	1.20991345E-02
12	12	0.0359	4.58282615E-18
15	32	0.237	5.37735730E-03
9	13	0.237	6.49024791E-04
17	35	0.237	1.24302629E-03
15	31	0.237	2.58492486E-03
26	54	0.237	8.57581863E-05