Loglinear Models and Goodness-of-Fit Statistics for Train Waybill Data

HERBERT LEE
Duke University

KERT VIELE
University of Kentucky

ABSTRACT

Counts of carloads of train shipments are effectively described with loglinear models. This paper presents models of counts by origin, destination, and commodity type. Such models can highlight structures in the data and give useful predictions. In particular, there are definite interactions between origin and destination and between origin and commodity, and these models can capture these relationships. Model selection depends on the choice of goodness-of-fit statistic; this paper addresses several issues relating to this choice.

INTRODUCTION

Roughly 1.7 billion tons of cargo is moved by train every year within the United States. In this paper, we explore a statistical method for modeling data from train waybills. In particular, we focus on the counts of carloads of cargo by commodity type and by origin and destination. This information can be arranged into a large three-dimensional table and is thus suitable for analysis via loglinear models. In addition to describing the data, such models allow us to compare flows of freight between different areas, search the data for unusual flows, and make predictions of future flows. Choosing a good model requires the selection of a goodness-of-fit statistic, and we discuss issues involved in this process. We

Herbert Lee, ISDS, Duke University, Box 90251, Durham, NC 27708. Email: herbie@stat.duke.edu.

also note several challenges that this data set presents, including a lack of symmetry and a large number of zero counts.

DATA

The data we analyze are from the Carload Waybill Sample issued by the Interstate Commerce Commission for the years 1988 through 1992 (ICC 1992). The data are a stratified sample from all waybills for railroads with over 4,500 cars per year of traffic or 5 percent or more of a state's traffic. There are over 1.9 million total records, each of which has 62 fields of information. Here we focus on three fields: the origin of the shipment, the destination, and the type of commodity. Both the origin and destination are classified into 1 of 181 regions (for the continental United States) as defined by the Bureau of Economic Analysis (BEA) although some are missing or unknown. The commodities are classified by Standard Transportation Commodity Codes (STCC), as per the Association of American Railroads. Using the two-digit aggregate codes gives us 37 categories of commodities in this data set (for example, farm products, coal, printed matter, etc.). Each record in the file is a sample shipment, which may consist of multiple carloads of freight. The sample is stratified, and strata were sampled with different frequencies. Thus, to get an estimate of the total count of carloads of commodity with a particular origin and destination, we first multiply the number of carloads in a record by the inverse sampling frequency and then sum these products over all such commodity shipments from the same origin to the same destination. For example, a record of 7 carloads in a stratum that was sampled with frequency 1 in 40 would get a weighted product of 280 carloads. These sums are entered into a large three-dimensional table, which is then ready for analysis.

As an example of the heterogeneity in the data, we spotlight Chicago, Illinois, and Huntington, West Virginia. Chicago is both the origin of the most traffic, as well as the most frequent destination. Over eight million carloads originate from the Chicago region, and these shipments are spread over many different categories of commodities and are well distributed across the country. In contrast, Huntington is in the top 5 regions by origin of total freight (over 3.5 million carloads), but this freight is nearly all coal. It goes to a smaller number of destinations, and much less freight is sent to Huntington in return. In modeling this data set, we need a model flexible enough to work for both general-freight cities like Chicago and for commodity-specific cities such as Huntington.

Unlike many tables, there is no symmetry in the data since commodities (such as coal) are generally shipped along particular routes, with cities either being origins or destinations but not both. Another potential problem is the large number of zero counts. For example, few things besides coal originate from the Huntington area. However, we do note that these zeroes are not structural zeroes. While many of the zeroes are easily predictable, there is no inherent reason any entry is zero. For example, much freight now moves via intermodal transport, meaning that it could go by truck partway and then be transferred to a train at an intermediate location. Thus, the intermediate location would show as the origin with respect to the train shipment even though it is not the true origin of the commodity.

LOGLINEAR MODELS

Data consisting of counts, such as the waybills, are naturally modeled by the Poisson distribution, which takes values on the nonnegative integers. Instead of a standard regression model with an assumption of Gaussian error, we use a Poisson regression model. Such models are often called loglinear because they are a linear model for the mean after logarithms are taken. Here we model the mean of the distribution of counts from origin i to destination j of commodity k by m_{ijk}. The full loglinear model in this context is

$$\begin{aligned} &\log m_{ijk} = \log a_i + \log b_j + \log c_k + \log d_{ij} + \\ &\log e_{ik} + \log f_{jk} + \log g_{ijk}, \qquad (1) \\ &\text{or } m_{ijk} = a_i b_j c_k d_{ij} e_{ik} f_{jk} g_{ijk} \end{aligned}$$

where a_i is a main effect for origin i (and b_j and c_k are analogous), d_{ij} is an interaction effect for when origin i and destination j have cargo flows not proportional to the product of the main effects a_i and

b_j (e and f are analogous), and g_{ijk} is a three-way interaction between origin i, destination j, and commodity k. The actual counts n_{ijk} of commodity k from i to j thus follow a Poisson distribution with mean m_{ijk}:

$$P(n|m) = \prod_{i=0}^{181}\prod_{j=0}^{181}\prod_{k=1}^{37}\frac{(m_{ijk})^{n_{ijk}}e^{-m_{ijk}}}{n_{ijk}!} \tag{2}$$

In practice, not all interaction terms may be necessary, and some may be dropped from the model. Also note that the model in (1) is overspecified (there are more free parameters than degrees of freedom), so some sort of restriction is needed. For example, $b_1 = c_1 = d_{i1} = d_{1j} = e_{i1} = e_{1k} = f_{j1} = f_{1k} = g_{i11} = g_{1j1} = g_{11k} = 1$ for all i, j, k. While loglinear models with only a few interaction terms can be fit directly, most complex models require iterative solutions, the most popular method being iterative proportional fitting (Deming and Stephan 1940). For more background on loglinear models, the reader is referred to one of the many good references on the topic (Agresti 1990; McCullagh and Nelder 1989; Bishop et al. 1975).

Loglinear models are part of the same family of models as gravity models (such as Sen and Smith 1995). Gravity models also contain a term relating the distance between the origin and destination to the rate of flow and so would have a term depending on this distance in equation (1). We have found that train cargo flow is not related to distance, and thus the additional term in the gravity model is unhelpful for our data. In contrast to focusing on modeling the effect of distance, we focus on the complex interaction effects of the covariates.

In this paper, we take the Bayesian approach. The gamma distribution serves as a conjugate prior for all parameters, and the posterior can be easily estimated via Markov chain Monte Carlo. With the full posterior, one can easily get estimates of uncertainty, in addition to simple point estimate. Either an informative prior or a noninformative (improper) prior can be used. Using a noninformative prior leads to posterior mode estimates equal to the maximum likelihood estimates. We use an essentially noninformative prior. More details on Bayesian loglinear models can be found in Gelman et al. (1995), and West (1994) discusses Bayesian loglinear models in the context of gravity models.

ASSESSING GOODNESS-OF-FIT

To compare how well different models fit, we employed cross-validation (see Stone 1974). For this data set, annual counts seemed a natural unit of validation. Thus for each year s =1,…,5, we fit each model under consideration using the other 4 years of data and used the fitted model to predict the counts for year s. These fitted counts were then used to compute a goodness-of-fit statistic $q_{r,s}$ for model r for year s. To get the overall cross-validation score, the goodness-of-fit statistics are summed across all years giving $q_r = \sum_{s=1}^{5} q_{r,s.}$ The rest of this section discusses choices of goodness-of-fit statistics.

Mean square error is an appropriate goodness-of-fit statistic when the variance of observations is the same for all observations (not true for Poisson data) or when one is not interested in adjusting for differing variances, such as when one is most interested in predicting the largest table entries correctly, that is, when nominal error is more important than relative error. This may be the case for train data in that predicting 100 carloads when the truth was 200 (a nominal error of 100, relative error of 100%) is much less of a concern than predicting 100,000 carloads when the truth is 150,000 (nominal error of 50,000, relative error of 50%). Those 50,000 extra carloads could represent a much larger logistical problem than the 100 extra carloads, in which case mean square error would be a useful summary. Equivalent to mean square error is its square root, root mean square error (RMSE), which has the advantage of being on the scale of the data and thus being more interpretable.

Alternatively, one may be more interested in relative error. Statistical theory says that one should adjust for the variance in computing goodness-of-fit. The Pearson chi-squared statistic is

$$X^2 = \sum_i\sum_j\sum_k\frac{(n_{ijk}-\hat{m}_{ijk})^2}{\hat{m}_{ijk}} \tag{3}$$

where n_{ijk} is the actual count and $\hat{m}_{ijk}$ is the predicted count. When the model holds, X^2 is asymptotically distributed as a chi-squared distribution

(see for example, Agresti 1990). The denominator in (3) is the estimated variance of the prediction, and thus X^2 is a measure of relative error. However, for an application such as cargo, it does not make much sense to inflate the error when the prediction is smaller than one. For example, if the model predicts 0.1 carloads in a year, and in truth 2 carloads were observed, the contribution to X^2 would be $(2 - .1)^2/.1 = 36.1$, larger than the nominal error. When routes have hundreds of thousands of cases, a nominal error of 2 carloads is rather insignificant, and its contribution to the total error does not seem like it should be inflated. As a further complication, when this goodness-of-fit statistic is used for predictions, the model might predict a count of zero when the actual count could be nonzero. In that case, X^2 is infinite, and it is impossible to compare models. If a small number were added to each cell of the table, the comparison of models can depend on the size of the value added. To avoid these problems, we modify X^2 so that the denominator is no smaller than one:

$$\tilde{X}^2 = \sum_i \sum_j \sum_k \frac{\left(n_{ijk} - \hat{m}_{ijk}\right)^2}{\max\left\{1, \hat{m}_{ijk}\right\}}$$

The Cressie-Read power-divergence family of goodness-of-fit statistics (Read and Cressie 1988), indexed by a single power parameter λ, is a general family that includes many common measures as special cases, including the Pearson chi-square and the loglikelihood ratio statistic. Most of the members of this family have the common problem of being undefined for prediction either when some entries predicted to be zero are nonzero $(\lambda \geq 0)$, or when there are zero entries that were predicted to be nonzero $(\lambda \leq -1)$. One standard goodness-of-fit measure is in the intermediate power range and thus is directly applicable to prediction with zero entries: the Freeman-Tukey statistic, given here as parameterized in Fienberg (1979):

$$F^2 = 4 \sum_i \sum_j \sum_k \left(\sqrt{n_{ijk}} - \sqrt{\hat{m}_{ijk}}\right)^2$$

F^2, employing the variance-stabilizing transformation for a Poisson distribution, represents a compromise between the mean square error and the Pearson chi-square statistic. Note that while all members of the Cressie-Read family have the same asymptotic chi-square distribution, their distributions may be different for finite samples. In particular, when the data table is sparse (with many zeroes, as with the waybill data), there can be problems with the chi-square approximation for all of the statistics (for example, Koehler 1986).

DATA ANALYSIS

The models under serious consideration were the full model (equation 1), the model without a three-way interaction (*g* of equation 1) but including all two-way interactions, and the three models with no three-way interaction and only two two-way interactions (that is, no *g* and only two of *d, e,* and *f* in equation 1). Models with fewer terms were unable to capture the complexity of the data. Table 1 compares goodness-of-fit statistics for all models with at least two two-way interaction terms. We note that we can not use the unmodified X^2 statistic because during cross-validation, some entries predicted to be zero are instead nonzero, leading to infinite values of X^2.

From the table, we see that the choice of best model does depend on the choice of measure of goodness-of-fit. The full model seems best for reducing absolute error since it has the lowest RMSE (and does fairly consistently for each year of the cross-validation). If relative error is more important, the model using only two-way interactions for origin versus destination and for origin versus commodity performs best. Also of note is the

TABLE 1 Cross Validation Goodness-of-Fit Statistics for the Top Models

Model	RMSE	F^2	$\tilde{X}^2$
Full model	2029.2	6.59e+7	3.71e+9
All 2-way interactions	2118.2	9.28e+7	3.34e+9
No origin-destination interaction	4052.0	2.61e+8	3.81e+9
No origin-commodity interaction	4418.4	3.04e+8	4.03e+9
No destination-commodity interaction	2606.5	1.51e+8	2.86e+9

Note: RMSE is the root mean squared error of prediction, F^2 is the Freeman-Tukey statistic, and $\tilde{X}^2$ is the modified Pearson statistic.

model with all two-way interactions. It is a reasonable compromise model having an RMSE close to that of the full model yet also having the second lowest F^2 and X^2. Thus, this model might be chosen for its robust performance with respect to multiple goodness-of-fit statistics.

Substantively, it is interesting that the other models with only two two-way interactions do not perform as well. It seems clear that any reasonable model must include both an interaction term for origin versus destination and a term for origin versus commodity. An instructive example is that of Huntington. As mentioned earlier, Huntington primarily exports coal and only to a specific set of destinations. Yet Huntington is one of the largest areas in terms of total number of carloads. Thus, any model must be able to account for both the fact that Huntington exports a very large amount of coal but little else as well as the fact that it exports large amounts to a relatively small number of destinations, unlike general shipping hubs like Chicago. In contrast, there are no obvious examples of cities that are destinations for large amounts of particular commodities out of balance with their imports of other commodities, so the removal of the interaction term for destination versus commodity has much less impact on the fit of the model.

CONCLUSIONS

The train waybills data set is interesting both for its information on commodity flows as well as for its statistical challenges. Loglinear models provide an effective method for describing the relationship between cargo volume and origin, destination, and commodity type. The size of the data set[1] is much larger than in a standard statistical problem. While this size is beyond the capabilities of many standard statistical software packages, loglinear models can be programed directly.

Model selection raised a number of statistical issues. In contrast to many data sets used with loglinear regression, the waybills are sorted by year, providing a natural breakdown for cross validation. The choice of goodness-of-fit measures has been discussed. Dealing with the large number of zero counts during cross-validation appears to be a topic not fully addressed in the statistical literature. The analyses of this paper should be seen as a starting point for further work, both methodological and relating to the interpretations of the indicated models.

[1]Approximately 2 million records in the data file, resulting in over 119 million carloads distributed in a three-way table containing over 1.2 million cells.

ACKNOWLEDGMENTS

The authors would like to thank three anonymous referees and the Editor-in-Chief for their valuable suggestions. This work was partially supported by National Science Foundation grants DMS-9971954 and DMS-9873275.

REFERENCES

Agresti, A. 1990. *Categorical Data Analysis*. New York: Wiley.

Bishop, Y.V.V., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Deming, W.E. and F.F. Stephan. 1940. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics* 11:427–44.

Fienberg, S.E. 1979. The Use of Chi-Squared Statistics for Categorical Data Problems. *Journal of the Royal Statistical Society B* 41:54–64.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman & Hall.

Interstate Commerce Commission (ICC). 1992. Carload Waybill Statistics: Territorial Distribution, Traffic and Revenue by Commodity Class. Published on CD-ROM by the Bureau of Transportation Statistics, Washington, DC.

Koehler, K.J. 1986. Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables. *Journal of the American Statistical Association* 81:483–93.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*. London: Chapman & Hall.

Read, T.R.C. and N.A.C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Sen, A. and T.E. Smith. 1995. *Gravity Models of Spatial Interaction Behavior*. Berlin: Springer-Verlag.

Stone, M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society B* 36:111–47.

West, M. 1994. Statistical Inference for Gravity Models in Transportation Flow Forecasting. Technical Report 94–40, Duke University, Institute of Statistics and Decision Sciences, Durham, NC.