Paper 257-28

Logistic Regression Modeling - JMP Start™ Your Analysis with a Tree

Pippa Simpson, Jeff M Gossett, James G Parker, Renée A Hall
University of Arkansas for Medical Sciences, Little Rock, AR

ABSTRACT

In regression, the decision about which variables to include and in which form they should be included in the model can be very difficult. Screening variables can be very tedious; perhaps that is the reason why many models seen, for example, in nutrition, only include main effects. However, because of its uses in screening, a tree can JMPstart your regression model analysis.

All types of variables can be included in a tree, including variables with missing values and variables that are highly interrelated. This enables consideration of the form of the variables to be included. Because of the tree methodology, cutpoints for variables that best optimize a function are given, so it is possible to consider new variables generated from the old variables. Trees are also useful for exploring the interaction of variables. For example, if a variable appears on one side of a tree and not on the other, it suggests that there is indeed an effect of interaction.

Using a large nutrition dataset, we will show how using regression alone can lead to misleading conclusions whereas the use of tree analysis in conjunction with logistic regression can enable building an appropriate model.

INTRODUCTION

With large datasets a statistician is often asked to build a model for predictive reasons or for assessing the impact of certain variables upon the outcome. In both linear and logistic regression the decision about which variables to include and in which form they should be included in the model can be very difficult. Many interactions in a model can make it unstable, particularly when dummy variables are included for categorical variables. A linear relationship may not exist for continuous variables; hence a decision about possibly categorizing the variable should be made. Moreover, it is not always clear that all categories of a categorical variable should be retained in an analysis. This is why a tree can JMPstart your regression model analysis.

All types of variables can be included in a tree, including variables with missing values and variables that are highly interrelated. This enables consideration of the form of the variables to be included. The tree methodology results in cutpoints for variables that best optimize a function; so it is possible to consider new variables generated from the old variables. In addition, trees are useful for exploring the interaction of variables. For example, if a variable appears on one side of a tree and not on the other, it suggests that there is indeed an effect of interaction. Outlier groups can be detected using trees, especially when coupled with other methodologies as described by Friedman. (*2*) Another advantage of tree methodology is that expert opinion can be easily incorporated because of the intuitive and interactive nature of the Partition® platform in JMP®, Version 5 software.

Usually, p-values do not come from a tree by default. Moreover, in general it is desirable to use some variables as continuous and others as not. In logistic regression, odds ratios can be interpreted as risk, and in linear regression the slope parameters give us useful information. In addition, regression is a more traditional methodology. Therefore, they work well in tandem.

We show this approach using a large nutrition dataset where we have both continuous and dichotomous outcomes of interest. We will show how regression, on its own, may lead to misleading or incomplete conclusions.

AIMS AND OBJECTIVES

The objective of this paper is to show the value of the recursive partitioning methodology in JMP software in developing models. Specifically, we will develop a logistic model for the large national nutritional NHANES 1999-2000 dataset, where obesity is the outcome. Diet, demographics, income status, federal aid, and obesity risk factors, as well as their interactions are considered in building the model. Our aims follow.
Aim 1: *To show that analyzing the data only using logistic modeling limits the possible variables considered.*
Aim 2: *To show how screening the data with JMP recursive partitioning can be incorporated into building a model.*
Aim 3: *To develop the model using JMP and SAS® software, validate the results, and interpret the results. .*
Aim 4: *To develop the model using the weights in SUDAAN and validate the model.*

METHODS

DATA

NHANES 1999-2000 is the eighth in a series of national examination studies conducted in the United States by the National Center for Health Statistics (NCHS) beginning in 1960. Beginning in 1999, NHANES became a continuous survey of a representative sample of the U.S. civilian, non-institutionalized, household population of all ages every year. Each year, approximately 6000 people in 5 locations are expected to be examined and/or interviewed. The sampling plan is a complex, stratified, multistage, probability cluster design. The survey takes place both at the sample person's home and in the Mobile Examination Centers (MEC) using computer assisted interviewing techniques. A blood sample and other laboratory samples are also collected at the MEC. We took a subset of the study, which included white and black adults (20 years and older) of non-Hispanic origin with complete data only. This resulted in a dataset of 2177 records.

Variables include the following.

- Demographics such as *age* (20 to 84 and 85+), *sex* (Male, Female), *race* (Black, White), *edu*cation (1=less than High School(HS), 2=HS or General Education Diploma (GED) , 3=more than HS or GED).

- Poverty index ratio*, PIR*, is the ratio of household income to the corresponding poverty level income, which incorporates household size. The distribution of PIR is highly skewed and ranges from 1 – 103. The average PIR for our sample was 3.8, and the median was 2.7. A PIR value less than 1.84 indicates eligibility for some government assistance.

- Nutrition information is extensive in NHANES, but processing is necessary to derive measures of diet quality. We consider the 10 component sub-scores of the healthy eating index (HEI).*(1)* Each ranges from 0 –10 with 10 being the best score. The sub-scores measure adherance to USDA recommendations for food pyramid servings, sodium intakes, percentage of calories from saturated fat and total fat, and diet variety. A score of 5 on a pyramid serving sub-score indicates that 50% of the recommended servings were consumed.

- Risk factors include information on smoking, *smoke* (1=Current , 2=Former, and 3=Never , including those smoking fewer than 100 cigarettes in lifetime) and drinking of alcoholic beverages, *drinker* (yes, no).

- *Overweight* is classified using the CDC classification based on the body mass index (BMI). The BMI is the ratio of body mass to height (kg/m^2). Adults with a BMI of 25 or larger are classified as overweight. We use two categories normal (0) vs. overweight/obese (1).

We took a 50% random sample of the data (n=1089), and reserved the remaining subset for validation purposes.

SOFTWARE

NHANES data were collected using a complex survey sample. Jackknife type I replicate weights are provided for variance estimation. The NHANES analytic guidelines recommend using SUDAAN for analyses to obtain valid variance estimates. Variance estimates and derivative p-values from SAS and JMP analyses should be considered exploratory. They are not considered valid and may be misleading, as Gossett et al. showed.*(3)* For logistic regression analyses, we use the RLOGIST procedure in SAS callable SUDAAN. The RLOGIST procedure is relatively crude in that there are no model building capabilities (like stepwise). It also has limited model fit diagnostics.

Data management, including creation of derivative variables (e.g., HEI scores) was performed using SAS/BASE® software, Version 8 of the SAS System for Windows. Repetitive calculations, particularly the calls to SAS callable SUDAAN, were automated using SAS/MACRO® software

NHANES Analytic Guidelines (*5*) highly recommend the plotting of data prior to analyses, and JMP is ideal for exploratory analyses. The Distribution® Platform of JMP software was used for exploratory analyses (distributions of PIR, sampling weights, other predictors) while the Partition platform was used to interactively build classification trees for overweight.

In the Partition platform in JMP software, Version 5, recursive partitioning is used. A tree is created using the relationship between the independent variables ***X*** (factor columns) and the independent *Y* values. The tree is structured as a series of questions. The answers to these questions give the branches or path taken on the tree. The endpoint is a set of hierarchical rules that segment the data into groups. The factor columns (***X***'s) can be either continuous or categorical (nominal or ordinal). If an ***X*** is continuous, then a *cutting value* creates the splits (partitions). The sample is divided into values below and above this cutting value. On the other hand, if the *X* is categorical, then the sample is divided into two sets. Essentially, the questions relate to whether one variable is in a set or not. The split is chosen to maximize the difference in the responses of Y between the two branches of the split. The investigator has to decide how large a tree should be.

Unweighted as well as simple weighted analyses (using measurement weights in NHANES, called *wtmec2yr*) were performed using JMP software and SAS software. These weights do not account for the clustering in the NHANES sample design. For Sudaan, the weighted logistic regression used the jackknife weights (known as JACKWGTS WTMREP01-WTMREP52 in NHANES) that are recommended (*5*). The NHANES survey oversampled various target populations; hence, the weights vary over several orders of magnitude and can play an important role in modeling.

ANALYSIS

As described previously, overweight is the dependent variable in the models with the independent variables as defined above. The odds ratios represent the risk of being overweight in individuals with various risk factors. Classification trees are used in selecting the risk factors in the modeling process.

LOGISTIC REGRESSION (AIMS 1, 3, 4)

Logistic regression differs from linear regression in how stepwise regression works. In linear regression, adding variables to a model will improve the fit. This is not necessarily true in logistic regression. Therefore, for logistic regression, Hosmer and Lemeshow suggest a technique whereby each variable is modeled separately and variables for the next step are considered if p<0.2. Then pairs are selected from the candidates and modeled with their interaction. Again selection is based on p<0.2 and the need for hierarchy. This process is continued until the level of interactions desired is reached.

Two way interactions

	SAS weighted	Sudaan	SAS unweighted
Effect		**p-value**	
age*drinker	0.1041	0.1708	0.005
age*meat	0.0058	0.0129	
cholesterol*education	0.0576	0.1049	
cholesterol*smoke	0.0974		
drinker*smoke	0.0918	0.1268	0.0040
fat*race	0.0965	0.0546	
fat*smoke	0.0743	0.1279	
Grain*fat	0.0115	0.0467	0.1508
meat*agegroup	0.1225		0.1102
smoke*education	0.0339	0.0083	0.0294
smoke*race	0.1400	0.0313	0.0753
Variety*Cholesterol	0.1829		
Variety*grain	0.0585	0.1033	
Variety*vegetable	0.0825	0.1235	
Vegetable*cholesterol	0.1351		
Vegetable*fat	0.0452	0.0574	
Vegetable*meat	0.0320	0.0442	

Using SAS software for unweighted and simple weighted analyses and SUDAAN for complex, weighted analyses with the Jackknife weights, we found that there were 17 possible two-way interactions:

Only those which coincide with the significant results for the weighted analyses are included; all the unweighted "significant" variables are not given. These included demographic interactions with risk factors such as age and drinker, sub-scores such as age and the meat score, both risk factors, smoke and drink; risk factors and subscales such as smoke and total fat score; and pairs of HEI subscales such as vegetable score and meat score. These last types of interactive effects on obesity are not surprising. For example, a low vegetable score and high meat score is a risk for overweight status. Given the fact that the measurement weights only partially account for the design, it is not surprising that differences in the "significant" variables were seen between the SUDAAN analysis and the weighted and unweighted analysis using SAS software.

Because of the repetitive nature of the selection process we wrote a macro both for SAS software and SUDAAN to implement the stepwise analysis. A sample of the code for SUDAAN is shown below.

```
%macro logit1(x,n);
proc rlogist data=work.nhanes
   design=jackknife;
   JACKWGTS WTMREP01-WTMREP52/
```

```
        ADJJACK=.980769;
    subgroup &x;
    levels  &n;
    model overwt = &x;
  run;
  %mend logit1;
```

We then incorporated all the significant interactions with their main terms in a stepwise regression and obtained the following results. With Hosmer and Lemeshow's (*4*)approach, we found that with a higher variety score the risk of obesity is lessened. There was a significant interaction effect between age and the meat score with obesity increasing with a higher age and meat score interaction. When you are a never or current smoker and drink the risk of obesity is lower when compared to being a smoker and drinker. Moreover, a higher education has a more positive effect with smoking. When there are continuous effects it is usual to include the main effects. Neither meat nor age, as a main effect, is significant. This finding raises the question whether there is a subpopulation for which the meat score does affect overweight status. With any general linear modeling, one of the questions you should always ask is whether the independent variables are truly being included in the correct form; that is, is the dependency a linear form or some other form. The interaction terms may truly be significant or reflect a non-linearity in the dependence. In any case, they can be difficult to explain to a non-statistician, especially when the main effects are non-significant. Hosmer and Lemeshow suggest various approaches for investigating non-linearity. A tree approach, such as the recursive partitioning in the Partition platform of JMP software enables investigation of potential non-linearity.

It is interesting to note that stepwise logistic regression gives a different model again. We took all variables, where there were no interactions. We first set a "loose" requirement with slstay=0.15 and slentry=0.15 and then reduced these levels to 0.05. The variables of interest for unweighted analysis would be meat score with the risk of being overweight increasing and variety score with the risk decreasing. Also of interest is smoking status (current smoker versus never) and drinking status (drink versus not) where the risk decreases. For weighted analysis it would be age, meat score, variety score and sex, with risk increasing with age, and meat score, decreasing with variety score and being female versus male a current smoker versus never, and a drinker versus non-drinker.

The code for the weighted analysis is given below.

```
  Title 'weighted logistic with main factors
     only';
  proc logistic descending;
     class sex smoke edu drinker;
     model overwt = age pir Sfat_sc fat_sc  sex
     Meat_sc Dairy_sc veg_sc chol_sc fruit_sc
     grain_sc var_scor sodi_sc r_race edu smoke
     edu drinker
        /lackfit plcl selection=stepwise
        slstay=0.15 slentry=0.15;
     weight wtmec2yr / norm;
  run;
```

Note that you need a weight statement and that we definitely need to specify smoke and education as a class variable since they have three categories.

The recurrent theme here is that smoking, age, meat score or variety score, and potentially education and drinking play a role in overweight status.

TREE MODEL (AIMS 2, 3)

In the tree model, we incorporated all variables described above including the PIR and corresponding grouped. We did this so we could investigate whether there were "better" ways of grouping each variable to see its effect.

We found that the tree had different variables used for splitting in different parts of the tree. When this happens, the implication is that there is an interaction causing a variable to be significant only for a subset. With the interactive nature of the algorithm in the Partition platform of JMP software, we were able to investigate whether a variable, which was not quite "best", would do a good job nevertheless. This meant that we could investigate whether the interaction indicated was real or not. We found that in no case could we keep the symmetry in the tree.

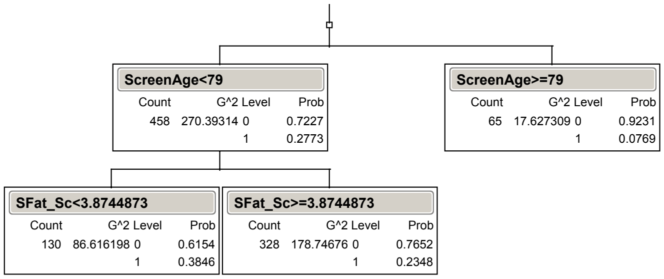
When we observed that the objective function of a candidate was close to the best candidate, we used expert knowledge and investigated whether there might be value in substituting a near best. We found that this was the case when smoking was considered and we show the two results.

Thus, we showed that the following interaction terms should be considered in a logistic model.

In the unweighted tree we saw sex was the best first split. Then for females, a PIR of < 3.38 was the best first split. A value of 3.38 corresponds fairly closely to what would be deemed a good income. Females with lower income had a high rate of being overweight (48%) versus those with a higher income (28%). The best first split for males was age, where those above 79 (8%), had less obesity than those below (28%).

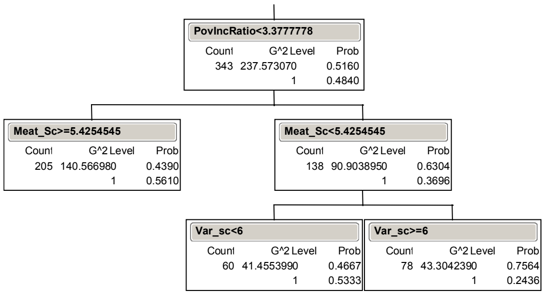
The tree is shown below, in parts.

Males



Using expert knowledge, we decided to use the split on saturated fat with a G^2 score of 10.06, although the best candidate was in fact smoking with a G^2 score of 10.36.

Continuing on with the *Females*: we click on the "candidates" button to produce a list of possible splits. We see that females with Incomes greater than 3.38 times the poverty income have obesity of 28% versus 48.4% for those with lower incomes (relative to the poverty income for their family sizes). For those with a lower PIR we get the following sub-tree.

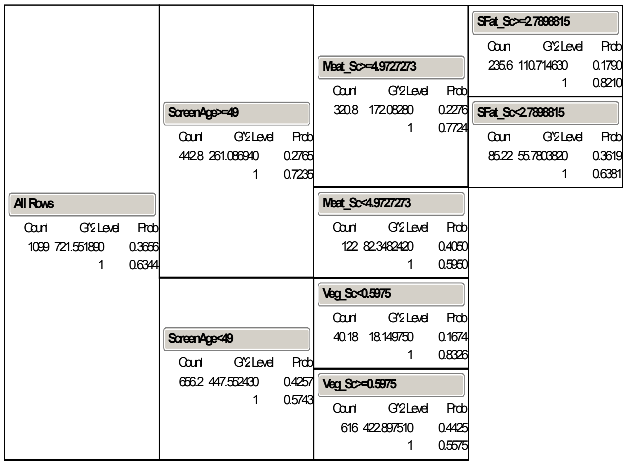


Note that the meat score and variety score play a role.

For a PIR > 3.8 in females, age, total fa,t and dairy score play a role. In all but total fat, the cutpoint chosen is around 5. In nutrition papers, 5-6 has been taken as a cutpoint, with below 5 representing inadequate scores.

WEIGHTED TREE

The weighted tree had a different profile. Age (< 49 and >=49), meat score, vegetable score, and saturated fat score played a role. When the vegetable score was low in those under 49(< 0.6), 84% were obese. In those at least 49, if the meat score was >=5, 77% were overweight compared to 50% with a lower meat score.



In both the logistic and tree modeling we found that incorporating weights gave different results to the unweighted analysis.

LOGISTIC MODELING & TREE INFORMATION

From both the weighted and unweighted trees, the variables suggested are the meat score, possibly as a categorized variable of two values, <5 or ≥5, a saturated fat score with cutpoint at 3, or total fat at 8, age cut at 50 and 80, and possibly a vegetable score cut at 0.6. It should be noted that the vegetable and fruit scores are on average much lower than the other scores, reflecting the way we all eat.

We used the cutpoints in investigating models. We found that whether we used the cutpoints or kept the full HEI subscales we got essentially the same variables in the model.

Using the trees, we developed models 1a and 1b. In model 1a, we used a cutpoint for age. In models 1b and 1c, we used the cutpoints for age and for the HEI sub scores suggested by the trees. In model 1b, we included information form the unweighted tree showing that income played a role.

In model 1a we found that if you are female, variety in diet is protective; if you are over 50, a high meat score indicates a likelihood of being overweight and if you are over 80 you are protected from being overweight. We had interactions without main effects and this is not troublesome because the interactions only involve one continuous variable and an indicator variable. This means that we have simultaneously fit several models to different subpopulations, described by being under 50, 50 to 80, or over 80 and female or male. These models could then be reported separately for the 6 subgroups.

In model 1b, we found that we had essentially the same model as model1a except we used a meat score of at least 5 instead. Also, the main effects of drinking and smoking as well as their interactions seemed important with drinking being protective as well as being a current smoker vs never.

In model1c, we found that a PIR≥1.85 and <3.3 or being male put you at risk for overweight status, as did being over 50 with a meat score higher than 5. As before, drinking and being a current or never smoker was protective.

Using the Hosmer logistic information (*4*) and the cutpoints, we obtained the easily interpretable model 2. If you are over 80 or are a current or never smoker and drink, your risk of obesity is less. With increased variety in diet, your risk of obesity is less, but if you are over 50, an increased meat score is indicative of obesity. Here we have three models fit simultaneously to the three age groups described in model 1.

SUDAAN VALIDATION

Sudaan is designed for complex sample surveys. It has no stepwise procedures and therefore does not lend itself easily to exploratory analysis. We used the 4 models developed using a combination of the tree model and the Hosmer-Lemeshow method. (*4*) We found all models reasonably stable with misclassification rates around 35%in the working dataset and increasing to 39% in the validation set.

Model 1b gave good results.

		95% CI for OR	
Variables	Odds Ratio	Lower	Upper
Intercept	2.33	1.47	3.68
SEX			
Male	1.5	1.06	2.12
Female	1	1	1
SMOKE			
Current	0.53	0.29	0.96
Former	0.67	0.41	1.09
Never	1	1	1
DRINKER			
Yes	0.47	0.29	0.76
No	1	1	1
DRINK and SMOKE			
Smoke+Drink	1.78	0.66	4.76
Former smoker, Drink	2.09	1.17	3.74
Never Smoked, Drink	1	1	1
Smoke, No Drink	1	1	1
Former Smoke, No Drink	1	1	1
Never Smoke, No Drink	1	1	1
AGE≥50, MEAT≥5			
Over 50 and Meat ≥ 5	1.75	1.03	
Over 50 and Meat < 5	0.86	0.55	
Under 50 and meat ≥ 5	0.79	0.48	
Under 50 and meat << 5	1	1	

CONCLUSION

The advantages of using the tree analysis are as follows.

- A suitable cutpoint for age was found. It is not reasonable to expect age to affect any health in a continuously increasing/decreasing way. Many health models use cutpoints to either categorize age or to create piecewise linear functions. In this application, the results are easily interpretable. Moreover, it was not necessary to include non-significant main effect terms.

- For interventions on overweight status, it is of interest to identify the best subpopulations to target. Trees aid this process.

The limitations of the tree approach follow.

- There is no capability as yet to incorporate complex survey weights.
- Many researchers uninitiated to trees are willing to accept the results without a p-value. Although p-values can be obtained by resampling or by the use of other statistical procedures, the heavy reliance on p-values has to be overcome.

In general, it should be realized that trees are an exploratory technique. Some form of confirmatory analysis or validation procedure is mandatory with the use of this tool. Nevertheless, trees are a welcome addition to our descriptive arsenal of summary statistics, graphs, and models.

REFERENCES

Basiotis PP, Carlson A, Gerrior SA., Juan, WY, & Lino, M (2002). *The Healthy Eating Index: 1999-200*0. U.S. Department of Agriculture, Center for Nutrition Policy and Promotion. CNPP-12. <http://www.usda.gov/cnpp/Pubs/HEI/HEI99-00report.pdf> (December 16, 2002)

Friedman J, Hastie T, and Tibshirani R (1998). Additive Logistic Regression: a Statistical View of Boosting. <http://www-stat.stanford.edu//~jhf/#reports/boost.ps> (December 16, 2000)

Gossett JM, Simpson PM, Parker JG, and Simon WL; "How complex can complex survey analysis be with SAS?" *Proceedings of the 27th Annual SAS User's Group International Conference* April 2002. <http://www2.sas.com/proceedings/sugi27/p266-27.pdf> (December 16, 2002)

Hosmer DW and Lemeshow S (2000). *Applied Logistic Regression*, New York, NY: John Wiley & Sons, Inc.

NHANES 1999-2000 Addendum to the NHANES III Analytic Guidelines. <http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf> (December 16, 2002)

ACKNOWLEDGMENTS

Work for this paper has been supported in part by ARS/USDA Project #6251-53000-003-00D.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the author via the following information.

Pippa Simpson
University of Arkansas for Medical Sciences
Department of Pediatrics
Section of Biostatistics
800 Marshall St
South Campus Slot 512-43
Little Rock, AR 72202
Work Phone: 501-364-6631
Fax: 501-364-1552
Email: simpsonpippam@uams.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.