

SESUG Paper RIV-97-2017

Exploring JMP® Modeling Functionality Using Consumer Expenditure Data

Joshua Klick, Bureau of Labor Statistics

ABSTRACT

The Consumer Expenditure (CE) Surveys collect information on demographic, housing, income, and expenditure characteristics of households in the US. The purpose of this presentation is to explore JMP® modeling techniques that explain consumer unit (CU) expenditures of the CE Public Use Micro Data from the CE Interview survey. An initial linear model is constructed based on an exploratory analysis of the data. The second and third models utilize stepwise regression based on the Bayesian Information Criterion and the Max K-Fold RSquare. The fourth model implements Bayesian model averaging via an interface to the R Bayesian Adaptive Sampling (BAS) package. Last, a fifth model partitions data based on the relationship between predictors and response variables. This paper concludes with a summary of model performance.

INTRODUCTION

The U.S. Consumer Price Index processes CE data to calculate expenditure weights for indexes used by a number of government agencies as measure of consumer price change. The CE data provide a means to define expenditures representative of populations such as the urban population represented by the CPI-U, the wage earner population represented by the CPI-W, and the elderly population represented by the experimental CPI-E. Numerous products and papers describe expenditures represented by various populations; however, limited analysis has been conducted to date that evaluates factors that explain consumer unit expenditures. The Fit Model platform provides a flexible and dynamic application to conduct a multivariate analysis that can determine explanatory factors of consumer unit expenditures.

Source data for this paper is the 2015 CE Public-Use Interview data. A household is interviewed once every three months over the course of a year. Approximately 24,000 CUs were interviewed in 2015. The response variable of analysis is the total expenditures per current quarter; previous quarter expenditures were excluded from this proof-of-concept analysis. To simplify the analysis, a subset of factor variables were selected, as described in Table 1 below.

Subgroup	Variable Descriptions
Demographic	Family size, Age of reference person, Highest education level, # of members younger than 18 and # over 64, Occupation grouping of reference person and spouse, Race, Family type
Income	Composition of earners, Imputed total income before taxes, Income class before taxes
Housing	Housing expense (shelter + utilities + household operations + furnishings), Shelter expense (owned + rented + other dwelling), Rental equivalence, Housing tenure, Total # of bathrooms, # of bedrooms
Other	# of owned vehicles, # of leased vehicles

Table 1. Subgroups of Explanatory Variable Descriptions

Eligibility for modeling purposes is based on non-zero expenditures for the response variable, total expenditures this quarter, and for the housing and shelter explanatory variables; a value of 1 was added to these expenditure records to eliminate missing values. Approximately 16,000 records were eligible for analysis. The occupation category was replaced with three categorical levels as follows: wage earner occupations, urban non-wage earner occupations, and non-urban.

Example of JMP Scripting Language (JSL) code to set up data for analysis by interfacing with SAS® and the open data table below:

```
SAS submit("
LIBNAME CE 'xxx';
DATA CE.FMLY_EXAMPLE;
    SET CE.FMLI15: (KEEP =
        QINTRVMO NEWID BLS_URBN FAM_SIZE AGE_REF EDUC_REF HIGH_EDU
        PERSLT18 PERSOT64 REF_RACE TOTEXPCQ EARNCOMP FINCBTXM INCLASS
        INC_RANK OCCUCOD1 OCCUCOD2 VEHQ VEHQL BATHRMQ HLFBATHQ BEDROOMQ
        CUTENURE RENTEQVX HOUSCQ SHELTCQ);
    /* Non Zero Edit */
    TOTEXPCQ = TOTEXPCQ +1;
    HOUSCQ = HOUSCQ+1;
    SHELTCQ = SHELTCQ+1;
    WHERE TOTEXPCQ > 0 AND HOUSCQ > 0 AND SHELTCQ > 0;
    IF OCCUCOD1 in
        ('04', '05', '08', '10', '11', '12', '13', '14', '15', '17')
        THEN CPIWX_FLAG_REF = '1';
    ELSE IF OCCUCOD1 in
        ('01', '02', '03', '06', '07', '09') THEN CPIWX_FLAG_REF = '2';
    ELSE CPIWX_FLAG_REF = '0';
    IF OCCUCOD2 in
        ('04', '05', '08', '10', '11', '12', '13', '14', '15', '17')
        THEN CPIWX_FLAG_SPO = '1';
    ELSE IF OCCUCOD2 in
        ('01', '02', '03', '06', '07', '09') THEN CPIWX_FLAG_SPO = '2';
    ELSE CPIWX_FLAG_SPO = '0';
    DROP OCCUCOD1 OCCUCOD2;
RUN;");
Open("xxx.sas7bdat",
    Use labels For Var Names())
```

EXPLORATORY DATA ANALYSIS

A JMP Journal is displayed as Figure 1 as an example of organizing analysis and notes within JMP. Analysis was conducted using JMP version 13.1.

1. Analysis of CE Expenditure Models

[JSL: Interface with SAS- FMLY File Extract and subset](#)

[Open fmly2 data](#)

Exploratory Data Analysis & Models

- EDA 1: Distribution of TOTEXPCU, Dashboard- compare to log
- EDA 2: Boxplot of log(TOTEXPCU) to subset of covariates

Models

- Model 1: Initial Linear Model- OLS
- Models 2 & 3: Forward Stepwise Minimum BIC and Max K-Fold R2
- Model 4: Interface w R BAS package to product Marginal Posterior Inclusion Probabilities and Posterior Probabilities

[JSL: Interface with R & BAS package](#)

- Model 5: Predictive Partitioning

Model Summary

[Open Model Summary data](#)

- Parallel Plot
- Multiple Correspondence Analysis

Figure 1. JMP Journal of CE EDA and Models

Exploratory Data Analysis (EDA) consists of evaluating the distribution of variables via the Distribution platform. The distribution of the response variable total current expenditures is right skewed as displayed in Figure 2. To improve modeling, the response variable is log transformed to normalize the distribution and diminish the impact of the outlier record to the right. The log transformed variable can be created dynamically within most of JMP's platform displays.

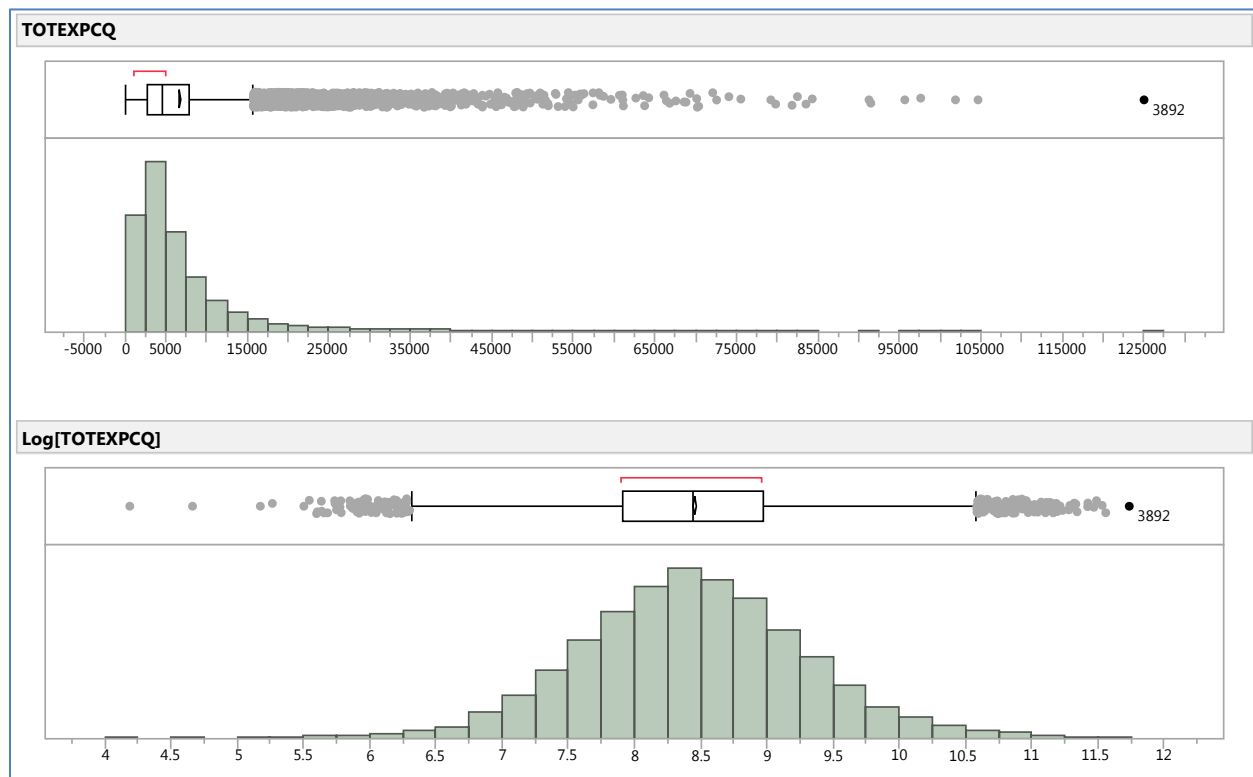


Figure 2. Distribution of Total Current Quarter Expenditures

Further distribution analysis of the response compared to a subset of explanatory variables is displayed in Figure 3. Categorical and numerical values can be selected as subsets of data, which are dynamically highlighted within the distributions for the other displayed variables allowing for interactive data analysis. As an example, when the rented category is selected for the CUTENURE variable, then the representing renter records distributions are highlighted for the other variables- indicating that renters are primarily single or other CUs as indicated by the FAM_TYPE variable and their respective expenditures appear normally distributed.

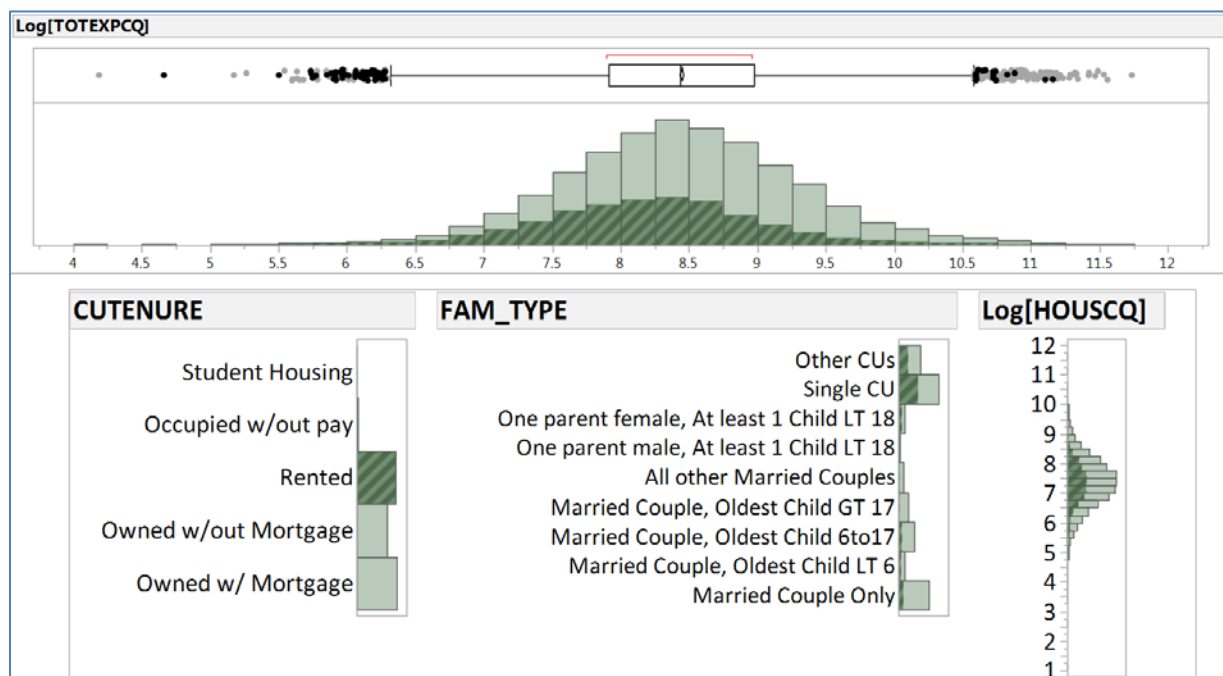


Figure 3. Distribution of Response and Explanatory Variables

MULTIVARIATE MODELS

For each of the models the log of total current quarter expenditures serves as the response variable. Explanatory variables are selected based on the model processing criteria. Each model is summarized, and then relevant content is displayed. Default model settings of the Fit Model and Partition platforms are used unless indicated otherwise.

Models one to three are evaluated using the Fit Model platform. Model 1 displayed in Figure 4 is a Standard Least Squares Personality regression, where variables are selected based on the LogWorth transformation, which adjusts p-values for graphing. If the LogWorth value is greater than 2, then the variable is significant at the .01 level, as summarized by $(-\log_{10}(.01) = 2)$. To measure collinearity among terms, the Variance Inflation Factor is evaluated across parameter estimates, where $VIF_i = (1/(1 - R_i^2))$. Variables can be Removed, Added, or Edited dynamically within the Effect Summary display. LogWorth values less than 2 are excluded from analysis for Models 1 to 3.

The performance of the models will be evaluated by comparing: 1) the number of variables, 2) the adjusted R Square, 3) the root mean square of the log-transformed response variable of total current expenditures, and 4) the coverage probability based on the proportion of observations that fall within the 5% and 95% confidence intervals to evaluate if the model is over-fitting the data. Model metrics 1-3 are contained within the Summary of Fit display. The coverage probability was calculated for Models 1-3 and 5 by saving the Individual Confidence Intervals, and then creating a new flag variable based on the actual response value falling in between the confidence interval values. For the Model 4, prediction intervals are evaluated instead of confidence intervals.

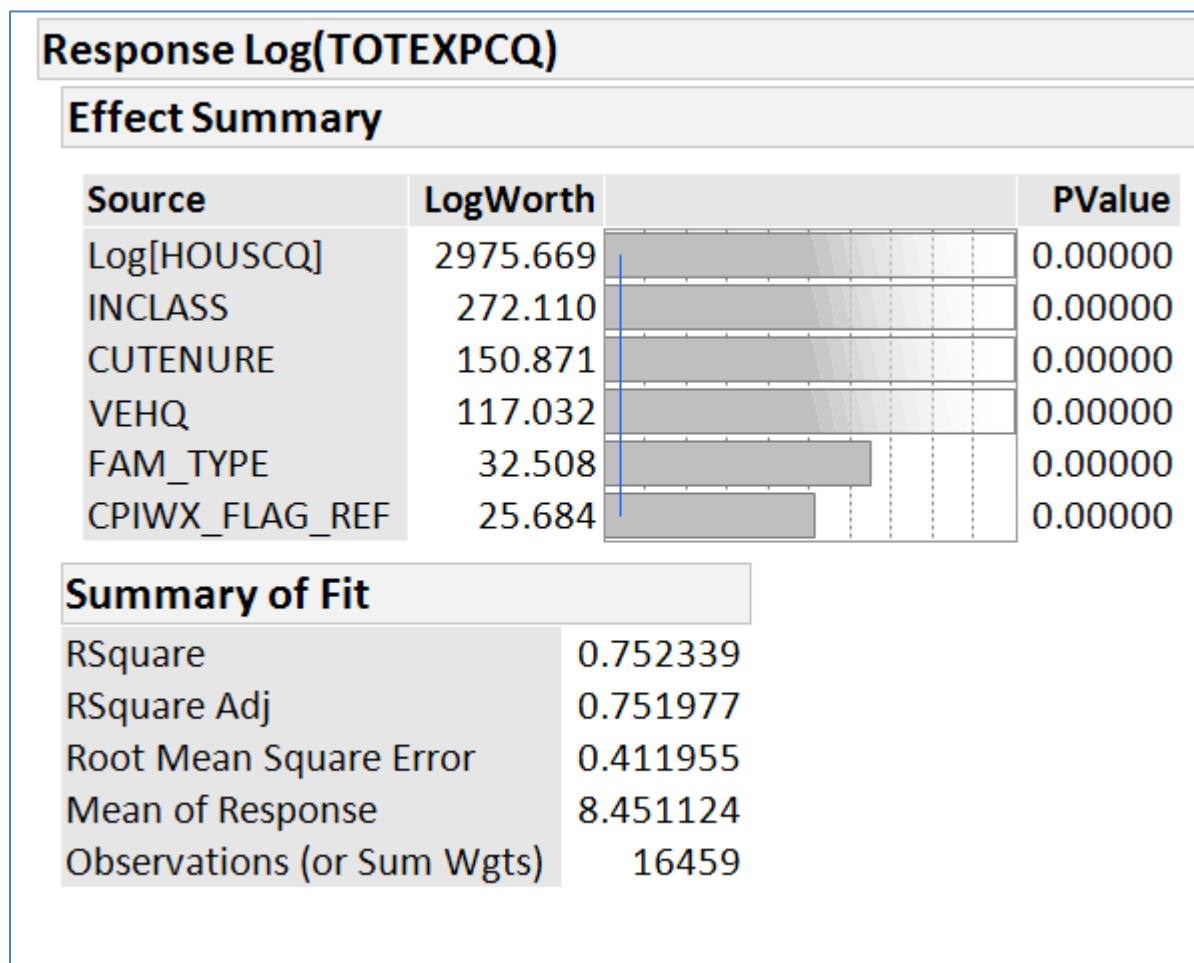


Figure 4. Intuitive Model Summary- OLS

Models 2 and 3 Stepwise Personality displayed in Figures 5 and 6 include a wider subset of the explanatory variables of analysis because stepwise regression will evaluate all of the relevant variables for analysis based on the Stopping Rule criteria. For Model 2, the Forward stepping Stopping Rule is Minimum BIC, which selects variables to determine the best model based on a maximum likelihood estimation where $BIC = 2\text{LogLikelihood} + k/n(n)$, where k is the number of parameters in the model and n is the number of observations. One of the default settings worth noting is the Rules option that is set to Combine, which allows for parsing effects of categorical variables where there are greater than two levels; this prevents selecting non-significant interaction parameter terms. In contrast, the Whole Effects option enters all levels of the categorical variable for selection within the model.

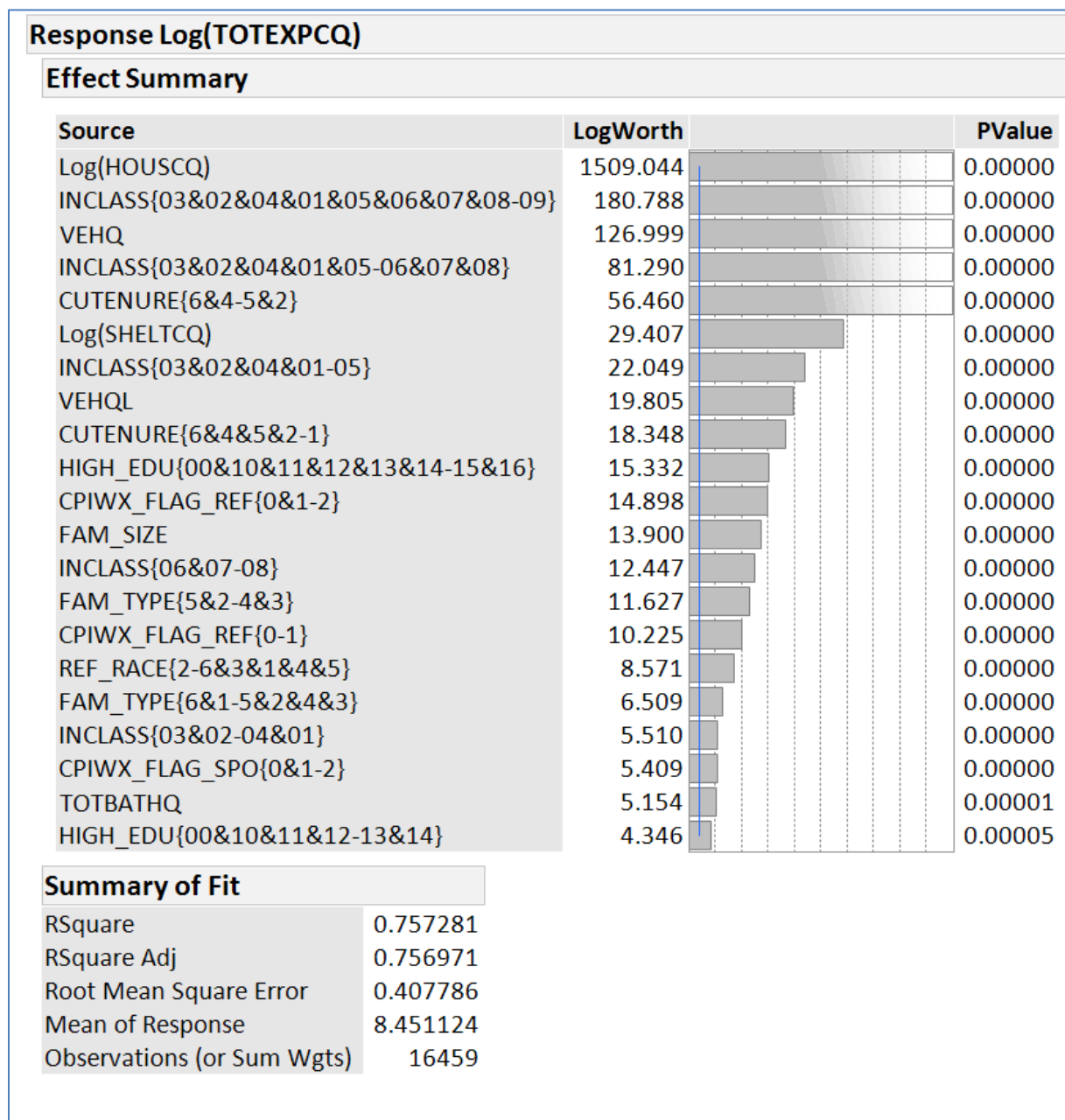


Figure 5. Forward Stepwise Fit- Minimum BIC Model Summary

Model 3 is based on the same Forward Stepwise process, however the Stopping Rule is based on the Max K-Fold RSquare criteria. Folds of data are randomly divided into a number of subsets defined by the user as validation sets, and the remaining data are used as the training set. The number of subsets also defines the number of models evaluated, where model selection is based on the best RSquare results. The number of folds selected for this analysis is based on the default value of 5.

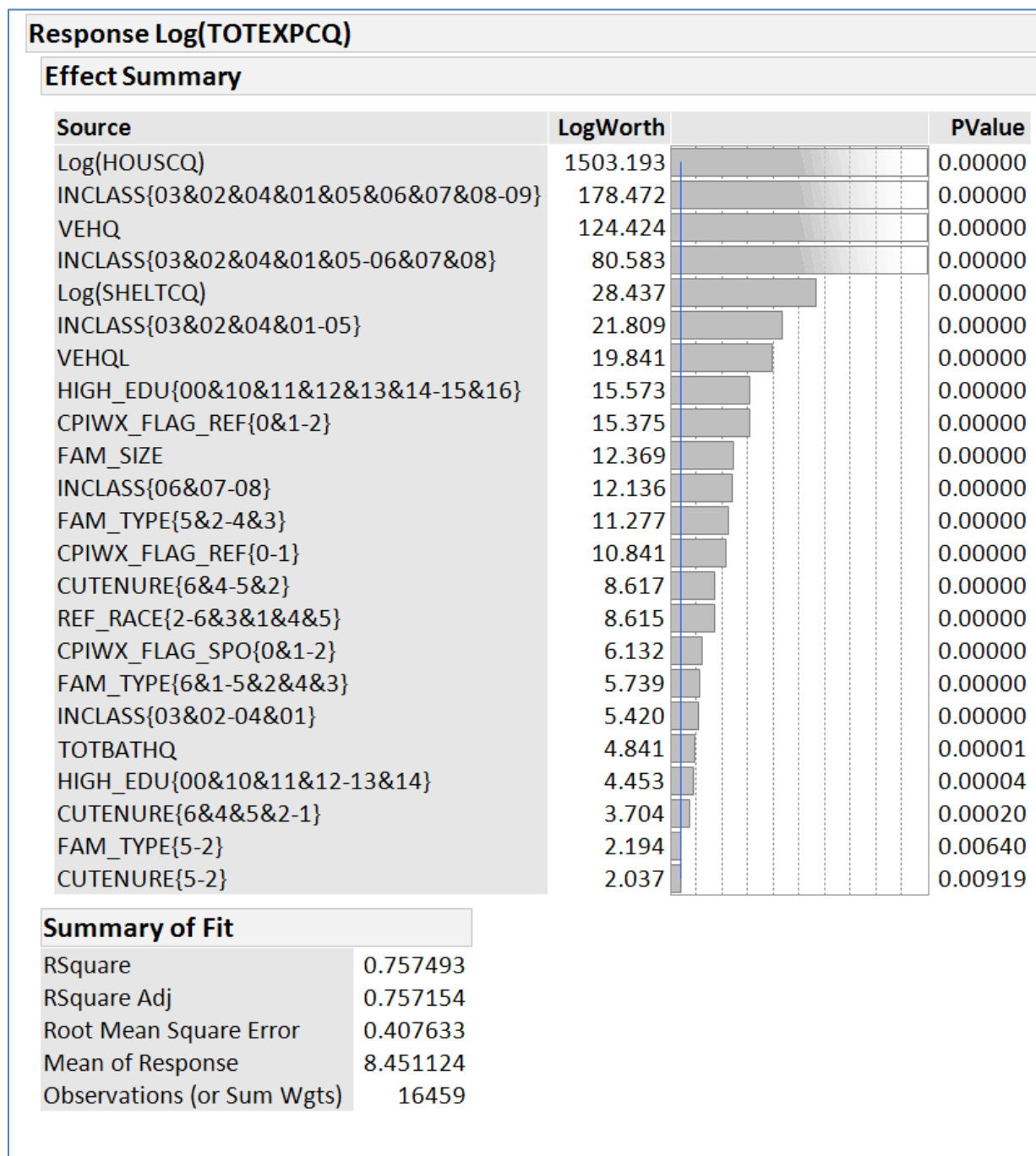


Figure 6. Forward Stepwise Max K-Fold RSquare Model Summary

Model 4 is constructed by JMP interfacing R to process the BAS package `bas.lm` functionality, which conducts Bayesian Model Averaging for variable selection from a posterior distribution of models. One of the benefits of this approach is that the model prior distribution can be defined for coefficients and models; for this paper the prior is BIC and model prior is uniform. JSL to process BAS and to calculate the root mean square error and coverage probability based on prediction intervals is displayed below:

```
dt = Open("C:\Users\Klick_J\Desktop\Temp\JMP SESUG
Materials\intrvw15\intrvw15\fmly2.jmp");
R Init ( );
```

```

R Send (dt);
R Submit("\[
library(BAS)
library(stats)
library(dplyr)
# Fit the BAS model
CUEXP_bma = bas.lm(log(TOTEXPCQ) ~ CUTENURE + VEHQ + log(HOUSCQ) + INCLASS +
VEHQL + log(SHELTCQ) + TOTBATHQ + CHILDAGE + HIGH_EDU + CPIWX_FLAG_REF +
CPIWX_FLAG_SPO + FAM_TYPE + FAM_SIZE + REF_RACE,
               data = dt,
               prior = "BIC",
               modelprior = uniform()
               )
CUEXP_bma
options(width = 100)
summary(CUEXP_bma)

# Root Mean Square Error
pred.HPM = predict(CUEXP_bma, newdata = dt, estimator = 'HPM', se.fit=TRUE)
pred.HPM.rmse2 = sqrt(mean((pred.HPM$fit - log(dt$TOTEXPCQ))^2))
cat('HPM_RMSE_logTOTEXPCQ\n')
pred.HPM.rmse2
# Coverage Probability based on Confidence Interval
out = as.data.frame(cbind(exp(confint(pred.HPM)), TOTEXPCQ = dt$TOTEXPCQ ))
colnames(out)[1:2] = c('lwr', 'upr')
out %>% summarize(COVERAGE_CI = sum(TOTEXPCQ >= lwr & TOTEXPCQ <=upr) /
n())
]\");

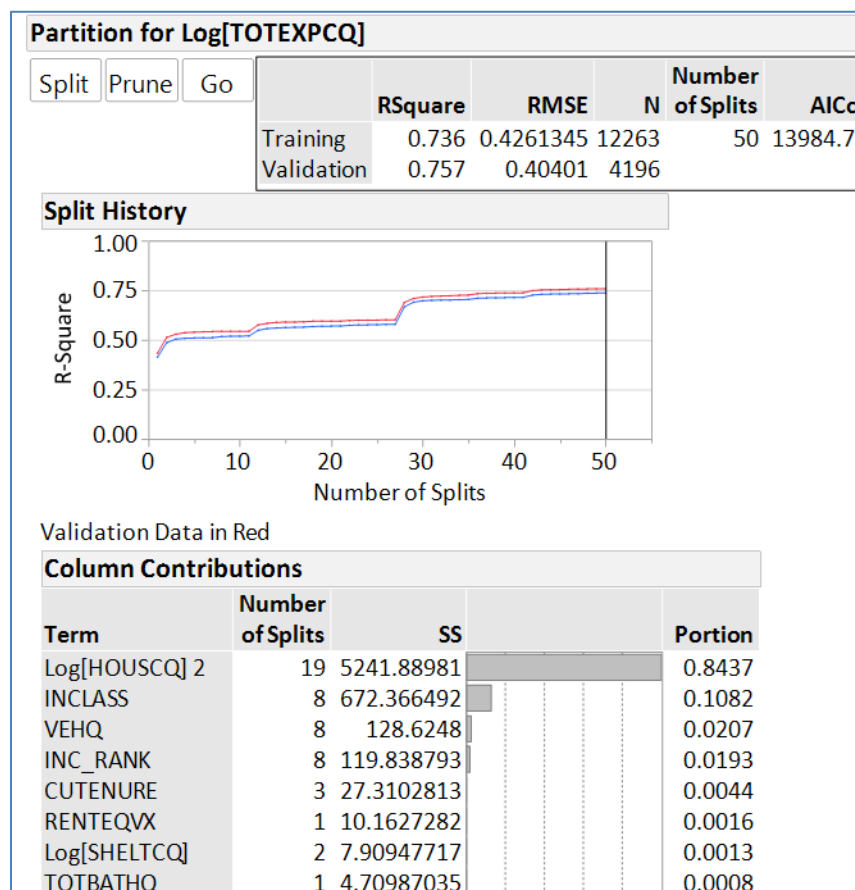
```

Output from the above JSL is accessible from the JMP log. A portion of the log is displayed in Figure 7 below, which could be saved to a JMP data table for further processing.

	P(B != 0 Y)	model 1	model 2
Intercept	1.000000000	1.0000	1.000000e+00
...			
CUTENURE2	1.000000000	1.0000	1.000000e+00
REF_RACE6	0.007508804	0.0000	0.000000e+00
BF	NA	1.0000	5.242055e-01
PostProbs	NA	0.1112	5.830000e-02
R2	NA	0.7575	7.573000e-01
dim	NA	24.0000	2.300000e+01
logmarg	NA	-65232.5723	-6.523322e+04
HPM_RMSE_TOTEXPCQ			
[1]	4993.983		
HPM_RMSE_logTOTEXPCQ			
[1]	0.4073613		
COVERAGE_CI			
1	0.9563765		

Figure 7. Interface with R BAS Model Summary

Model 5 Predictive Partitioning recursively partitions based on the relationship between predictors and the response, where splits occur to best predict the response variable. When a validation set is selected, the GO processes splits until the validation RSquare is better than the next 10 splits from the training set. A 25% validation set was selected for the model below.

**Figure 8. Predictive Partition Model Summary (25% Validation)**

MODEL SUMMARY

Performance metrics of the five models above are summarized as a Parallel Plot and Multiple Correspondence plot in Figures 9 and 10 below. The Parallel Plot connects line segments of the model to visualize relationships across the model metrics. Similarly, the Multiple Correspondence plot uses principal components computed as linear functions to identify relationships between models.

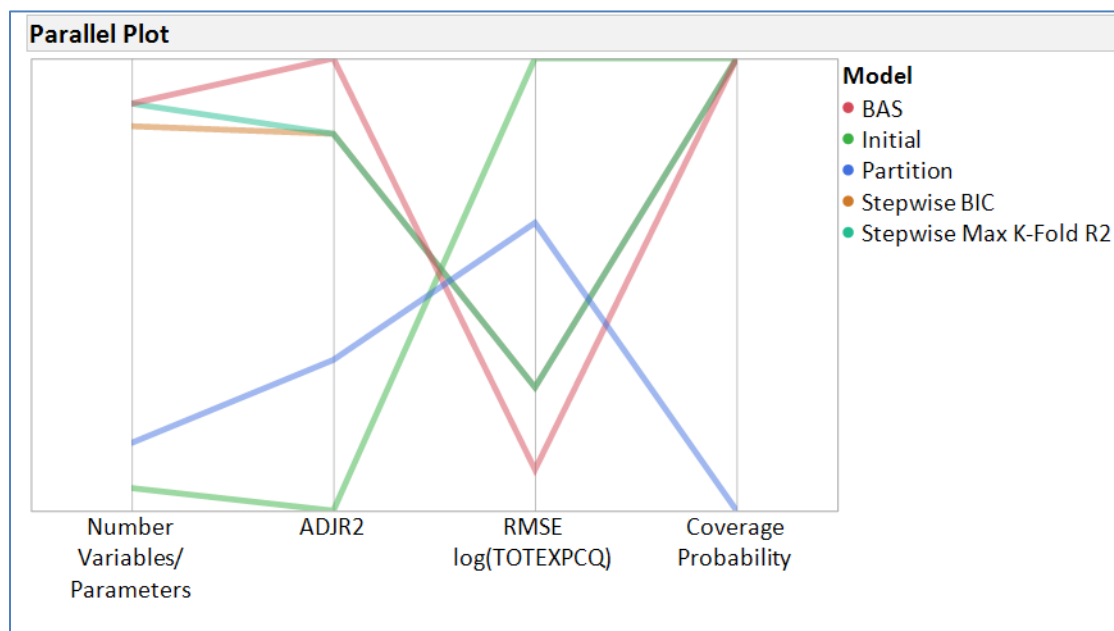


Figure 9. Parallel Plot of Model Summary

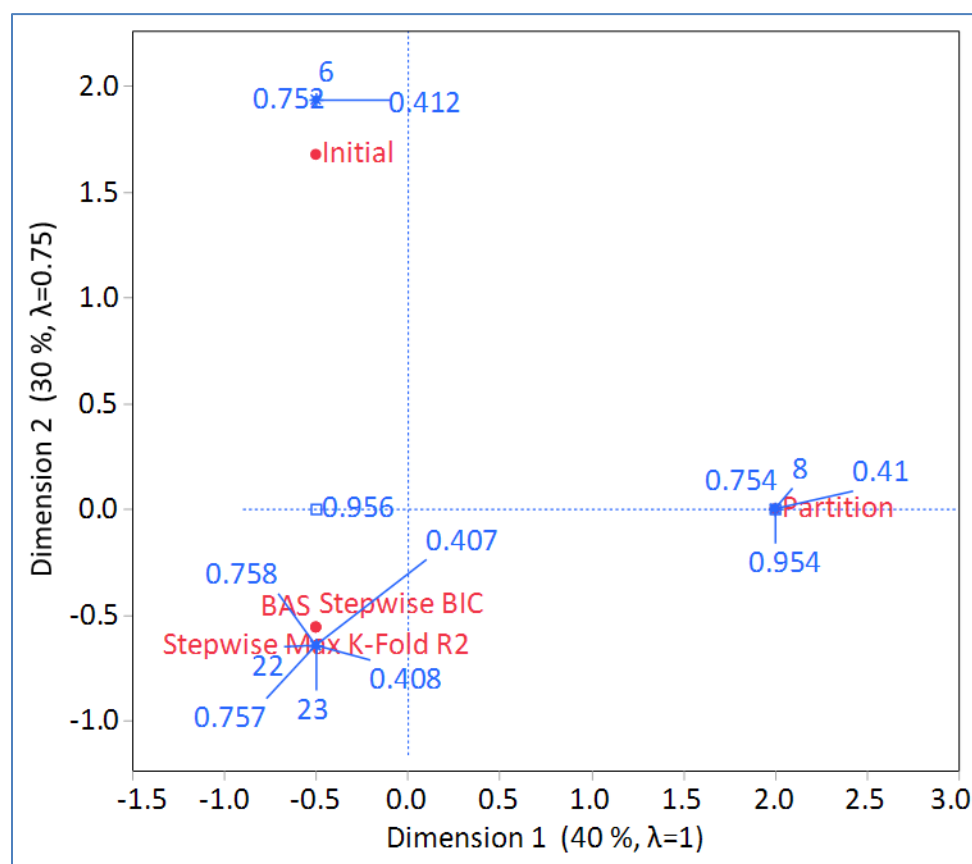


Figure 10. Correspondence Plot of Model Summary

CONCLUSION

Model choice depends on an acceptable level of complexity and prioritizing summary output based on user preferences. One analyst may prefer a less complex model with an acceptable level of error. Another analyst may prefer a more complex model with a better fit that reduces error and does not include potential non-significant interaction terms. Yet another analyst may prefer Bayesian regression to define priors. Performance of any given model may therefore be less important than describing the rationale for choosing that model and any of its potential weaknesses.

REFERENCES

Consumer Expenditure Public Use Micro Data. Available at https://www.bls.gov/cex/pumd_data.htm

Clyde, Merlise. "Using the Bayesian Adaptive Sampling (BAS) Package for Bayesian Model Averaging and Variable Selection." 2017. Available at <https://cran.r-project.org/web/packages/BAS/vignettes/BAS-vignette.html>

"JMP 13 Help Manual." 2016. Available at www.jmp.com.

ACKNOWLEDGMENTS

I would like to thank Melvin Alexander for his steadfast support and assistance preparing this paper. I would also like to thank Ronald Johnson for reviewing and editing this paper.

RECOMMENDED READING

- *Using the JMP9 R Interface to Perform Bayesian Analysis: Why, How, and What*, by Dave LeBlond

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Joshua Klick
Bureau of Labor Statistics, Division of Consumer Prices and Price Indexes
2 Massachusetts AVE NE
Washington, DC 20212
E-mail: klick.joshua@bls.gov
Web: www.bls.gov/cpi

JMP, 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. Other brand and product names are trademarks of their respective companies.

DISCLAIMER

All views expressed in this paper are those of the author and do not necessarily reflect the views of or policies of the U.S. Bureau of Labor Statistics. This paper describes a work in progress and does not reflect any models used by BLS.