Tax Structure Study – Technical Advisory Group

Model Review: Household Tax Burden Model

Date

February 20, 2020

Contact

Research and Fiscal Analysis Division (RFA)
Analyst: Kris Bitney; krisb@dor.wa.gov; (360) 534-1532
Manager: Valerie Torres; ValerieT@dor.wa.gov; (360) 534-1521

Model Purpose

The Household Tax Burden Model simulates household tax burdens under current law or alternative tax policies. The model estimates the following for households: retail sales tax, alcohol beverages taxes, cigarette and tobacco taxes, marijuana taxes, insurance premiums tax, gasoline tax, public utility taxes, and property tax. The model generates estimates within income quantiles and geographic regions. The model also estimates tax burden as a share of household income so we can estimate the degree of proportionality of Washington's tax structure for households. We measure tax burden as the total tax imposed on a household by state and local sources.

Data Sources

The data used in this model includes:
IRS Individual Income Tax Data
County Property Tax Rolls
Bureau of Labor Statistics - Consumer Expenditure Survey
US Census Bureau - American Community Survey

Requirements Model Used to Fulfill

This model will fulfill these requirements in ESHB 1109 (2019):[1]

Sec. 137(c)(vii)(A) With respect to the final report of findings and alternatives submitted by the Washington state tax structure study committee to the legislature under section 138, chapter 7, Laws of 2001 2nd sp. sess.:

(I) Update the data and research that informed the recommendations and other analysis contained in the final report;

(II) Estimate how much revenue all the revenue replacement alternatives recommended in the final report would have generated for the 2017-2019 fiscal biennium if the state had implemented the alternatives on January 1, 2003;

[1] Washington State Legislature (2019). *HB 1109: Making 2019-2021 biennium operating appropriations.* (https://app.leg.wa.gov/billsummary?BillNumber=1109&Year=2019&Initiative=false).

Requirements Model Used to Fulfill, *continued*

(III) Estimate the tax rates necessary to implement all recommended revenue replacement alternatives in order to achieve the revenues generated during the 2017-2019 fiscal biennium as reported by the economic and revenue forecast council;

(IV) For estimates above, estimate the impact on taxpayers, including:

- Tax paid as a share of household income for various income levels, and
- Tax paid as a share of total business revenue for various business activities

(V) Estimate how much revenue all of the revenue replacement alternatives in the report would have generated for the 2017-2019 Biennium if the state had implemented the alternatives on January 1, 2003, excluding any recommendations implemented before the effective date of this act.

Sec. 137(c)(vii)(B):

(II) Estimate how much revenue would have been generated for the 2017-2019 biennium if the 1 percent revenue growth limit on regular property taxes was replaced with a limit based on population growth and inflation if the state had implemented this alternative on January 1, 2003.

Questions for Technical Advisory Group

I greatly appreciate any and all feedback about the model. The following are areas in which your assistance would be particularly helpful.

Adjustments for Misreported Consumption
After estimating expenditures, we make adjustments to account for under- and over-reporting of consumption of certain goods. Survey respondents tend to overestimate their healthy consumption habits and underestimate their consumption habits that are unhealthy or socially undesirable. We made such adjustments in 2002. How can we best measure misreporting and implement adjustments?

Other Adjustments
In the 2002 model, we also adjusted results so that aggregated revenue estimates matched known totals estimated by internal sources. This practice is useful because there are known factors that can bias our estimates, such as the inability to account for cost of living differences, for which we currently cannot account. How can we best make such adjustments?

Adjustments for Underrepresented Populations
Our model assumes the population of federal income tax filers in Washington is representative of Washington's population of households. Because we base our model on IRS Individual Income Tax Data, we may underrepresent the sub-population of Washingtonians who do not file tax returns. As shown in Table A1 in the appendix, the population count in the IRS Individual Income Tax Data is about 8% lower than the April 1 Official Population Estimates reported by the Office of Financial Management. Likewise, the average income in the IRS Individual Income

Tax Data is greater than that reported in the American Community Survey, suggesting the IRS Individual Income Tax Data is less representative of households with below-average income. How can we account for the difference in representation and its potential to affect our results?

CES data to include in sample

We must decide which CES Interview quarters to include in the sample. The BLS conducts the CES Interview survey quarterly, but they ask respondents about their income during the prior year at each interview. For example, an interview that occurs in January 2017 will be part of the 2017 CES dataset but the household's responses will regard their income during the prior 12 months and expenditures during the prior 3 months—i.e. 2016 income and expenditures occurring between October 2016 and January 2017. The IRS Individual Income Tax Data represents income accrued during the 2017 tax year, which for most households is the 2017 calendar year. The alignment of income and expenditures can affect the correlation between the two, and therefore model performance.

Questions from Technical Advisory Group

We will capture at our meeting and record here

Model Technical Description

Summary

A tax microsimulation model applies tax policy rules to microdata that is representative of the population of interest. The use of microdata keeps the model independent of tax policy rules, allowing for flexible simulation of tax policy alternatives.

The Household Tax Burden Model is primarily a microsimulation of Washington households. The model uses IRS Individual Income Tax Data to represent the population of Washington. The model uses the Consumer Expenditure Survey to train statistical models that we can use to estimate Washingtonians' consumer expenditures in about 575 expenditure categories. After computing expenditure estimates for Washington households, we can simulate sales and consumer excise tax burdens of each household under alternative tax policies or current law. Additionally, the Household Tax Burden Model uses a macro-level approach to estimate property tax burdens. We aggregate final estimates by income quantile and geographic region.

Objectives

Objective 1: Estimate and report household tax burdens, measured as the total tax imposed on a household by state and local sources.

Objective 2: Estimate household tax burden as a share of household income, and report the degree of proportionality of Washington's tax structure for households.

Objective 3: Model alternative rates and types of taxation

Table A2 in the appendix contains the output of the 2014 Household Tax Burden Model. It is our objective to replicate that table—with updated results—for each geographic region in which we estimate tax burdens and for each potential tax structure change with first incidence on households.

The model estimates retail sales tax, alcohol beverages taxes, cigarette and tobacco taxes, marijuana taxes, insurance premiums tax, gasoline tax, public utility taxes, and property tax by household. We intend for the model to be able to estimate some taxes not currently implemented in Washington as well. Table 1 lists the taxes we currently intend to model, as well as those we currently do not plan to model due to data limitations.

Table 1: Consumer Taxes to Model

Current Washington Taxes to Model	Alternative Taxes to Model	Taxes Not Modeled
Sales tax	Food	Use Tax
Alcohol	Consumer services	Estate Tax
Cigarettes and other tobacco products	Medical services	Solid Waste Tax
Marijuana	Trade in value	Rental Car Tax
Insurance Premiums	Individual Income Tax	Leasehold Excise Tax
Gasoline		Other taxes
Public utilities		
Property tax		
Real Estate Excise Tax		

Fortunately, the taxes we do not currently plan to model account for only a small fraction of tax receipts. The most significant of them are the use tax and estate tax. According to Department of Revenue data, the use tax and estate tax were about 3% and 1.2% of 2019 tax revenues, respectively. Washington collects most use tax revenues from businesses. Other taxes we do not model each accounted for less than 0.2% of 2019 tax collections. While we do not expect these taxes to substantively alter our findings, it is important that we consider how they might affect the proportionality of our tax system.

About the Data

Consumer Expenditure Survey

Produced by the federal Bureau of Labor Statistics (BLS), the Consumer Expenditure Survey (CES) is a nationally representative survey on expenditures, income, and demographic characteristics of American consumers. The CES defines households as "consumer units", meaning groups who share income and expenditures. The CES program consists of two distinct survey questionnaires—the "Diary" and "Interview" surveys—that BLS administers to non-overlapping samples.

The Diary survey asks households about their expenditures during the most recent week. BLS contacts households two weeks in a row. BLS spreads interviews equally throughout the 52 weeks of the year.

The Interview survey asks households about their spending during the most recent three months. BLS revisits households in this survey sample each quarter for four quarters, producing one year of data. One-fourth of the households contacted each quarter are new to the survey. BLS drops households from the survey after four consecutive quarters in the sample.

IRS Individual Income Tax Data

IRS Individual Income Tax data includes data collected from federal individual tax returns for all Washington residents who filed federal tax returns for the 2017 tax year. Tax returns include form 1040 and supplemental forms and schedules associated with form 1040.

County Property Tax Rolls

The Property Tax Rolls contain data on property valuations for Washington residents. Washington counties provide the data to the Department. The data provided differs by county. This model uses data from 2016 property valuations, which represents tax payments due in 2017. This aligns with the IRS individual income tax data used in the model

Geocoding

This model uses taxpayer income and address in the IRS Individual Income Tax Data to aggregate results by income quantile and geographic region. Regions may include legislative districts for both state chambers, and counties. We must combine low-population regions to comply with privacy laws. The model will also generate statewide estimates.

We will assign taxpayers to regions by geocoding the addresses reported on their tax returns. We can determine in which five digit zip code nearly all tax filers reside. For a majority of tax filers we associate their nine digit zip code or street address with land parcels. We can then use parcel numbers to join the IRS Individual Income Tax Data and property tax datasets.

Data Assumptions and Considerations

Assumptions about the IRS Individual Income Tax Data

First, it's important to note that we assume households filing federal income tax returns are representative of Washington as a whole. Estimates may underrepresent low-income and no-income households who are not required to pay federal income tax or choose not to file. The underrepresentation of low-income households may adversely affect the accuracy of our results, particularly our estimates of the taxes borne by the bottom quantile of income earners.

The unit of observation in the IRS Individual Income Tax Data is a tax return, which does not necessarily correspond to a household as defined by the CES. To account for this, we collapse the IRS Individual Income Tax Data dataset on home address. In doing so, we assume that everyone sharing an address belongs to the same consumer unit. The average family size in the nationally representative CES Interview Q1 2017 data is 2.47 (other quarters are about the same), while the average family size in the collapsed IRS Individual Income Tax Data is 2.64 and the average family size per tax return is 2.03.

Table A1 in the appendix reports population, household, and income figures for Washington from the American Community Survey (ACS), Washington's Office of Financial Management (OFM), and IRS Individual Income Tax Data. Relative to OFM estimates, IRS Individual Income Tax Data underestimates Washington's population and household count by about 8% and 17%, respectively, after collapsing the IRS Individual Income Tax Data on address. Mean household incomes are about 11% higher in the IRS Individual Income Tax Data than in the 2017 ACS.

We also assume we can represent consumer expenditures in each expenditure category as a function of the information found in the IRS Individual Income Tax Data. While aggregate expenditures highly correlate with income, an individual expenditure category may be more difficult to approximate. The relationship between income and consumption of a particular kind of product need not be linear or monotonic. Table 2 lists the variables in common between the Consumer Expenditure Survey and IRS Individual Income Tax Data, which include income and family composition characteristics.

Table 2: Variables in both IRS Individual Income Tax Data and Consumer Expenditure Survey data

Variable Description
Family size
Number of persons under age 18 in household
Number of persons over age 64 in household
Income, interest and dividends
Income, rent/royalties
Income, pensions
Income, wage/salary
Income, social security
Sum of 5+ income categories
State of residence

Both the IRS Individual Income Tax Data and the CES include some income types that are not common to both datasets. The income types we list in Table 2 are those with income definitions that are approximately the same between the two datasets. When we sum income categories, we can include additional income categories that do not have a 1:1 relationship between variables in the two datasets but that lead to similar definitions of total income in aggregate.

When relating the datasets using a statistical model, we are limited to the use of income categories the datasets have in common, individually or in aggregate. This is a limitation because, for example, our statistical model does not model account for capital gains as a source of income. This may affect the accuracy of our estimates for some high income households. The limitation does not affect our reporting. We can report results based on the full Adjusted Gross Income (AGI) information available in the IRS Individual Income Tax Data.

We have precise information about the ages of household members, but it's not yet clear how that information is best incorporated.

The primary alternative to IRS Individual Income Tax Data is the American Community Survey (ACS), which the Department of Revenue used for the Household Tax Burden Model in prior years. The ACS uses frequency weights designed to make it representative of Washington households. It has the relative advantage of containing more information that we could use to predict expenditures, such as education and employment data. It also uses a more comparable and interpretable household definition. Its primary disadvantage is its granularity: while the ACS is statistically representative of Washington's population, we can use the IRS Individual Income Tax Data to produce geographically local estimates because it contains observations on every tax return in the state.

Assumptions about the Consumer Expenditure Survey (CES) dataset
We must decide which CES quarters to include in the sample. The BLS conducts the CES quarterly, but they ask respondents about their income during the prior year at each interview. For example, an interview that occurs in January 2017 will be part of the 2017 CES dataset but the household's responses will regard their income during the prior 12 months and expenditures during the prior 3 months—i.e. 2016 income and expenditures occurring between October 2016 and January 2017. The IRS Individual Income Tax Data represents income accrued during the 2017 tax year, which for most households is the 2017 calendar year. The alignment of income and expenditures can affect the correlation between the two, and therefore model performance.

We aggregate households to their annual expenditures and income. Since the BLS interviews households each quarter for four quarters, the income they report at their fourth interview represents their income during their time in the sample. Unfortunately, some households participate in less than four interviews. One way to implement this approach is to drop survey participants that did not participate in all four quarters of interviews, potentially introducing systematic bias in the sample. Instead, we intend to interpolate consumer expenditures across the four quarters by assigning each household's missing quarters with the same household's mean expenditures from non-missing quarters. We will use households' most recent report incomes. While this will also introduce bias in the sample, we hope that bias will be less than any bias from systematically missing values. We will treat the Diary survey analogously, so that we annualize all expenditure estimates.

Another consideration with the CES dataset is top-coding and bottom-coding. The BLS censors the ranges of some variables to protect respondents' privacy.

Assumptions about connection between Consumer Expenditure Survey and IRS Individual Income Tax Data
We assume the consumption habits of Washingtonians are similar to those of other Americans, and that we can model those consumption habits adequately based on the information available in the nationally representative Consumer Expenditure Survey and the federal tax returns of Washingtonians.

Assumptions about connection between IRS Individual Income Tax Data and property tax rolls datasets
To estimate property tax burdens, we join IRS Individual Income Tax Data and property tax rolls datasets using physical and mailing addresses. There is a time difference between the two datasets. Tax authorities assessed the property values in calendar year 2016, but we have tax filers' addresses as of spring 2018, when they filed

their TY2017 tax returns. In associating the datasets, we are assuming housing mobility will not meaningfully affect our results.

Methodology

Property Tax Estimation

Assigning property tax burdens to households is difficult because we do not know which households own or rent their residence. After matching property assessment data to IRS Individual Income Tax Data through our geocoding procedure, we will randomly assign home ownership based on characteristics of the distribution of home ownership in the American Community Survey five-year estimates. The distribution will be discrete and conditional on income quantile, geographic region, and possibly additional factors.

Unlike our approach to consumer expenditure estimation, this is not a microsimulation model. With this method, we can estimate mean household property tax burdens within the income quantiles and geographic regions on which we condition home ownership. The degree of granularity achievable using the American Community survey limits the granularity for which we can estimate property tax burdens. Because we are randomly assigning home ownership, our synthetic home ownership variable will be uncorrelated with variables other than income quantile and geographic region. Likewise, the synthetic variable cannot provide new information about income quantiles or geographic regions that are more granular than those used in the original tabulation. We will produce estimates for nineteen geographic regions in Washington.

Table 3: Statewide home ownership rate by reported household income decile

Decile	Mean Household Income	Home Ownership Rate
0	$8,299.67	33.2%
1	$22,020.60	41.6%
2	$34,267.46	48.5%
3	$46,413.49	54.0%
4	$59,233.33	60.8%
5	$73,830.91	66.8%
6	$91,172.93	71.4%
7	$113,796.90	76.4%
8	$148,978.60	82.3%
9	$286,909.40	87.2%

Table 3 displays the Washington home ownership rate within statewide income deciles. We have created analogous tables for each of the geographic regions.

Note that the estimates for Sec. 137(c)(vii)(B), which refers to a change in the property tax revenue growth limit, will be statewide estimates only.

Expenditure Modeling

In order to assign consumer expenditure estimates to Washington households, we must relate the Consumer Expenditure Survey and IRS Individual Income Tax Data in such a way that we assign Washington households reasonable estimates of their consumer expenditures. Specifically, we use the Consumer Expenditure Survey to train statistical models that output estimates for each of the 575 consumer expenditure categories, using inputs available in both the Consumer Expenditure Survey and IRS Individual Income Tax Data. The Department of Revenue used a matching technique in the 2002 tax structure study. For 2020, we have evaluated the comparative performance of multiple statistical models using the method of cross-validation.

We examined the following model classes:

- Baseline model (as used in 2016 or 2002)
- K-Nearest Neighbors (KNN) with dimensionality reduction
- Ordinary Least Squares (OLS)
- Lasso with polynomial features
- Gradient Boosted regression ensemble
- Ensemble of Regressor Chains (ERC) with Lasso base estimator

The following sections describe these model classes, discuss their theoretical advantages and disadvantages,, describe the cross-validation process that is used to select the final model, and review the results of the cross-validation process.

Baseline model

The 2016 implementation of the Household Tax Burden Model provides a useful baseline. We can evaluate other models based on their performance relative to this model. Alternatively or additionally, we could use as a baseline the 2002 version of the model. Both versions use matching algorithms.

The 2016 model

For each Washington household, we match exactly on three indicator variables and a categorical variable, and then randomly select a nearby neighbor whose income is within a specified range. If we do not find a match, we loosen the matching rules and repeat the process. In each round, we grow the acceptable income differences. After the second or third round, we may no longer require exact matches on the indicator or categorical variables. We implement the matching algorithm "with replacement". We use the multi-round process to maximize match similarity while ensuring we match every Washington household to a CES household.

The 2002 model

The original version of the model is similar to the 2016 version, with a few substantive differences. Primarily, it repeats the procedure multiple times and averages the results. And in addition to adjustments for over- and under-reporting of consumption, the 2002 model includes adjustments to the results to ensure aggregated estimates are similar to known statewide figures:

> *"To account for discrepancies between reported consumption levels in some categories and actual levels implied by tax collections, the amount of consumption reported in the CEX was adjusted. These discrepancies exist, for example, in the reported consumption for items such as alcoholic beverages and tobacco products. In addition, other expenditure categories are also underreported. Based on BLS publication that compares reported survey expenditures with independent estimates, the amount of spending was adjusted. In addition, some further adjustments were made so that aggregate tax revenue from households match estimates of revenue for the specific revenue sources such as alcohol taxes, tobacco taxes, and the gasoline tax."*[2]

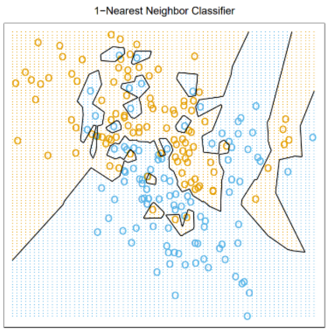
A final notable difference between the 2002 and 2016 models is the use of linear regression for imputation of missing values in 2002 model.

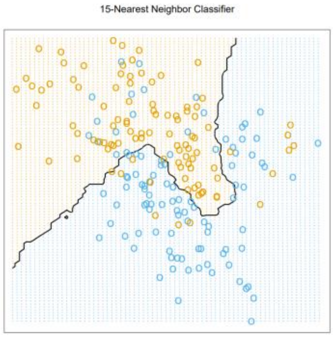
[2] Peterson, R. (2002). Washington Excise Tax Microsimulation Model 2002. Washington House of Representatives, Office of Program Research.

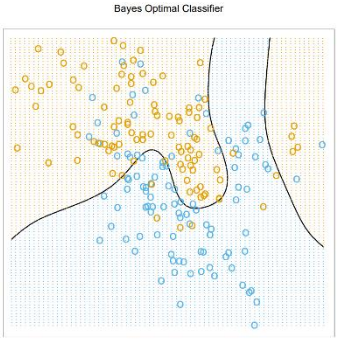
Matching Estimators

One-to-one (i.e. nearest neighbor) matching is one example of an effective model that can suffer from high variance because it attempts to fit a function that goes through every training point. This is typically addressed by averaging the outcome variable of multiple nearby neighbors in the feature space—an approach known as "K nearest neighbors". By averaging the target variables of the K nearest training points, rather than just the one nearest point, KNN smooths the function toward the local conditional expectation function. The KNN modification can improve performance substantially.

Below are Figures 2.2, 2.3, and 2.5 from Hastie et al., 2009.







The images depict a nearest neighbor classifier for a binary target and two features. We know the Bayes Optimal Classifier, the optimal decision boundary, because the data are simulated and we therefore know the data generation process a priori. Although this is a classification example, not a regression example, it provides an analogous illustration of how KNN matching can reduce overfitting relative to one nearest neighbor matching.[3]

The "curse of dimensionality" problem can hinder matching performance: without increasing the sample size, adding more features can make it more difficult to find close matches even when the additional features are informative. In this respect, the curse of dimensionality limits matching estimators' ability to reduce bias when estimators need many informative features to explain the dependent variable. We use "principal components analysis" (PCA), a dimensionality reduction technique, to help address this possibility.

Simply put, we can start with one set of variables and use PCA to generate a smaller number of variables that is almost as informative as the original, larger set. If we are limited in the number of dimensions we can accommodate, PCA helps us get the most bang for our buck. After standardizing the data and performing principal components analysis, we can conduct KNN matching using the transformed features.

In more technical terms, principal components analysis is a matrix decomposition technique. PCA uses a singular value decomposition to produce an optimal linearly transformed feature space, wherein we can use each possible ordered reduced rank subset of the transformed features to reconstruct the variance matrix of the original feature space with minimum square error. The transformed features—the "principal components"—are mutually orthogonal and can often be interpreted as latent variables.

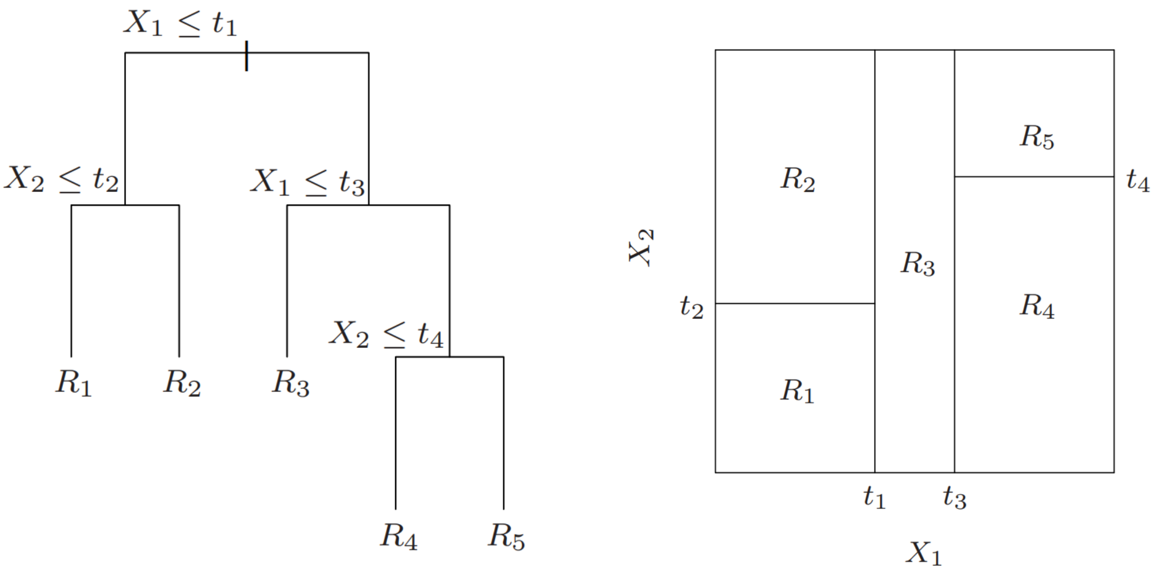
[3] Here, "regression" means a setting with a continuous dependent variable as opposed to a discrete, categorical dependent variable. Since the matching process doesn't depend on the target variable, KNN is the same in either context.

Linear Models
Ordinary Least Squares (linear regression) is a relatively simple additive model of linear relationships. The method fits the line (or hyperplane) that minimizes the mean square error. When the true conditional expectation function $E[Y|X]$ is linear, this is the best we can do.

Lasso modifies the OLS loss function by adding a regularization term, which effectively automates variable selection and prevents overfitting. Saturating the OLS model with cubic polynomial terms would increase the number of model parameters we have to estimate by as much as hundreds or thousands, depending on our initial variable selection, reducing the residual degrees of freedom by that same number. We would need a much bigger sample size to accommodate that. Using Lasso, we can include polynomial terms and interactions while controlling the effective degrees of freedom used for parameter estimation.

Regression Trees and Gradient Boosting
A decision tree partitions the feature space into a set of rectangular regions, then fits a constant in each region. The constant that minimizes mean square error is the mean of the target variable for the observations in the region. The decision tree partitions the feature space through a series of binary variable splits implemented in a greedy, stagewise fashion. At each stage, the tree partitioning algorithm chooses a variable and split point to minimize a loss criterion. The tree's terminal nodes, or "leaf nodes", define the final rectangular regions.

Below is part of Figure 9.2 from Hastie et al., 2009. The images are an example of a simple binary tree and the resulting partition of its two-dimensional feature space.



The benefits of tree-based methods are numerous—they are robust to uninformative features, are computationally efficient, are invariant to monotonic feature transformations, and more. While decision trees can have low bias when grown deeply, they have notoriously high variance and are not great predictors on their own. Their positive qualities make them an ideal base estimator for ensemble methods.

Gradient Boosted regression is a tree-based ensemble method that produces an additive model with a regression tree basis. It is a sum of regression trees. The algorithm is a forward stagewise procedure. At each stage, the Gradient Boosting algorithm first a new tree to the residual of the existing model. In other words, each new tree attempts to correct the error left over by the trees before it.

Ensemble of Regressor chains

Ensemble of Regressor Chains leverages dependencies among target variables in a multiple-output regression problem to potentially improve performance. Since we have hundreds of expenditure categories to predict, we might benefit by using information about their relationships.

The Ensemble of Regressor Chains (ERC) method fits a series of models in which future models use the outputs from all previous models to augment their base inputs.[4] The first model uses no output variables as inputs, the second model uses only the output the first model predicted, and so on. Because the augmented feature space of any single model depends on the order of the chain, ERC applies an ensemble approach that estimates several randomly-ordered regressor chains and averages their final estimates.

One drawback of this approach is noise. Because the method adds many estimated variables to the input space, it is very possible that the resulting increase in variance (from use of estimates) will outweigh the bias reductions we obtain from the additional information. We have many times more targets than standard input variables, so it would be very important for us to use a base estimator that is robust to noisy inputs (the “bet on sparsity” principle). We use Lasso regression as a base estimator.

Model Candidate Discussion

The following criteria guided selection of expenditure model candidates:

- Functional form
- Implementation feasibility
- Use of available information
- Bias-variance tradeoff

Functional form

We can describe function form as the general shapes of the “true” functions we are trying to approximate. For example, it may be the case that we can model tobacco expenditures as a linear function of one or more of the features available to us. That would be nice because it would be simple. However, we might find it best to model tobacco expenditures as a quadratic function or something more complex. In such a case, we might prefer the flexibility of matching.

It’s difficult to visualize multi-dimensional hyperspace, so functional forms are hard to guess just by looking at graphs. Economists sometimes identify functional forms using subject matter expertise and statistical tests like RESET. Those approaches make most sense when the analyst’s goal is to estimate structural (i.e. causal) equations. Instead, we will compare the empirical performance of model classes that differ in their functional representations.

[4] Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., & Vlahavas, I. (2016). Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, *104*(1), 55-98.

Linear models like OLS and Lasso are able to learn functional forms that are linear combinations of the features. These models describe additive, linear relationships. They are useful for modeling continuous polynomial functions, for example. It is possible to model discontinuities in linear functions when we already know the discontinuities' locations. An OLS model works well for approximating simple models, while Lasso is better suited for higher order polynomial terms with many interactions.

In contrast, matching methods and methods based on the use of decision trees are unable to reproduce continuous functions or linear combinations. They instead partition the feature space discontinuously, which allows them to approximate complex (or simple), nonlinear functions.

When we need to estimate multiple target variables, Ensemble of Regressor Chains allows the target variables to influence the functional forms of the regressions in which they are not the target of interest.

Implementation Feasibility

It's important that models are computationally feasible and are relatively easy to implement in SAS. Because we have about 575 output variables to estimate, we will have to compute each model at least that many times. In cross-validation, it's not unusual to try tens, hundreds, or thousands of hyperparameter combinations for each model class.

Use of available information and bias-variance tradeoff

In statistics, the "bias-variance tradeoff" is the common situation in which models must incur higher variance in order to reduce bias, or vice versa. We can decompose the mean square error of a regression model into the sum of variance and bias squared. Therefore, the only way to improve a regression model is to either reduce bias or reduce variance. Note the term "bias" here refers to the difference between the true value and the expected value of the model output.

Hastie et al., 2009 suggests a good way to understand generalization error (also called "expected prediction error") is by focusing on in-sample prediction error. We can describe in-sample prediction error as the sum of model training error and "optimism". The optimism term exists because we could draw multiple observations from the same population *at the same training points* and their target variable values need not coincide. A function drawn through every point in the training data sample will fit that sample perfectly but can perform worse with seemingly-identical samples. The optimism term is a function of the covariance between output estimates and their true values. "The harder we fit the data, the greater $Cov(\hat{y}, y)$ will be, thereby increasing optimism" (Hastie et al., 2009). The same general principle applies to extra-sample prediction error, where we cannot assume feature values will coincide with training points.

One-nearest-neighbor matching is one example of model that can suffer from high variance because it attempts to fit a function to every training point. We can mitigate this limitation by using K Nearest Neighbors. However, the "curse of dimensionality" limits the number of dimensions a matching method can model. We use "principal components analysis", a dimensionality reduction technique, to help maximize the amount of information we can use. A further potential drawback of matching is that the matching process doesn't depend on the target variable, so it treats every input variable as though it is equally important.

Lasso and Gradient Boosting are robust to uninformative variables. Lasso regression uses "regularization" to filter noisy variables and prevent overfitting. Decision trees incorporate only the most informative variable currently available at each step in the iterative tree "growing" process. A decision tree never includes variables that are never informative. Decision trees have notoriously high variance, but ensemble techniques can reduce it considerably.

Ensemble of Regressor Chains is a meta-model, in which we can use Lasso as a base estimator and garner its benefits. The purpose of ERC is to incorporate information from target variables in situations where we estimate multiple related outputs. We might expect to reduce bias by incorporating the additional information. The ERC approach can also have high variance than using the base estimator alone. Ideally, we improve model performance by reducing bias more than we increase variance.

Cross Validation and Model Selection

We will use 5-fold cross validation (CV) to compare models. In K-fold cross validation, we divide the sample into K groups. In each iteration, we fit a model on K-1 groups of the training data and use it to predict values for the Kth group in each consumer expenditure category. We compare the predicted values for the Kth group to the true expenditure values, and record the difference between the predicted and true values. We repeat this for K iterations. We leave each observation out of the training sample exactly once. We assess model performance as the average of the K error estimates.

Table 4 charts the role of each group during each iteration of 5-fold cross validation. In the first of the five iterations, we use groups 2-4 to train the statistical model, while we leave group 1 out to ensure it remains statistically independent. Since Group 1 is statistically independent of the model, we are able to use it use as a test-case to see how well the model would perform on new data. We repeat this process five times—once for each group—to assess how well the model would estimate consumer expenditures on five different independent samples. We average the five results to get a final expected value. The final average helps us gauge how well the model would perform if used to estimate expenditures for a sixth group, the IRS Individual Income Tax Data.

Table 4: Role of each group during each of the 5 iterations in 5-Fold cross validation

K	Group 1	Group 2	Group 3	Group 4	Group 5
1	Validation	Train	Train	Train	Train
2	Train	Validation	Train	Train	Train
3	Train	Train	Validation	Train	Train
4	Train	Train	Train	Validation	Train
5	Train	Train	Train	Train	Validation

K-fold cross-validation produces consistent estimates of expected prediction error: the expectation of the test error of a model across training samples drawn from the population of interest. In other words, expected prediction error treats the training sample as a random variable.

We select hyperparameters using Grid Search. In Grid Search, we define a grid of possible hyperparameters and estimate the cross validation score for every permutation.

We measure error as mean square error, a metric with desirable statistical properties for continuous dependent variables. We will average error estimates across all expenditure categories for each model and choose the model that performs best in aggregate.

Cross Validation Results

We have early cross validation results.

After evaluating their comparative performance, we found that Ordinary Least Squares, Lasso, Ensemble of Regressor Chains with Lasso as a base estimator, and K-Nearest Neighbors performed equally well—within a margin of error. In contrast, Gradient-boosted regression and the matching methods used for the household tax burden model in 2002 and 2016 had higher average error. We report the mean square error (MSE) of the best formulation of each model class in Table 5 for both the Interview and Diary portion of the Consumer Expenditure Survey. Figures 1 and 2 display the results as graphs.

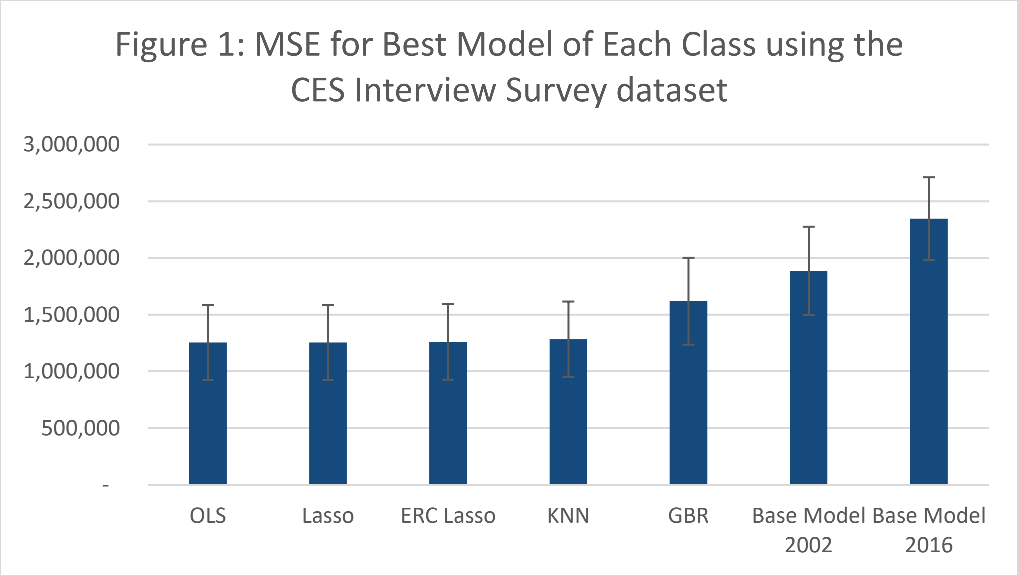
You can interpret MSE as the squared difference between predicted and true values of an estimate, averaged over a sample. We have 575 dependent variables, one for each of the 324 expenditure categories in the CES Interview survey, and one for each of the 251 expenditure categories in the CES Diary survey. We trained each model on each of the expenditure categories within a survey, and calculated the MSE for each expenditure category. We then average those MSE estimates. We do this five times, once for each fold in the K-fold cross validation process, to produce five cross validation scores. We report the average of the five scores in Table 5 for each model.

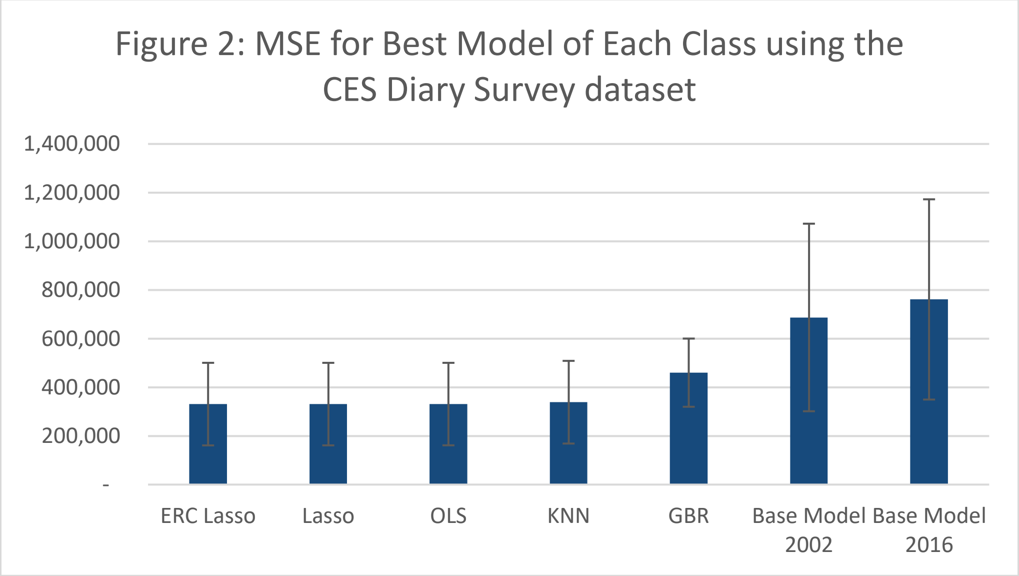
The models that produced these results use a consolidated income variable that is equal to the sum of all of the income categories common to both the Consumer Expenditure Survey and the Federal Tax Information data. When we use a separate variable for each type of income—e.g. wage income, social security income—model performance is unchanged (or worse, the cases of the 2002 and 2016 benchmark models).

Notably, K-Nearest Neighbors had the same performance with and without use of principal components analysis. This is important because we used a matching technique in the 2002 report. If we use K-Nearest Neighbors—a matching technique—for the current study, our model will be more consistent with the methodology used for the 2002 report.

Table 5: Cross Validation Score across All Expenditure Categories

	CES Interview Survey		CES Diary Survey	
Model	Mean Square Error	Standard Deviation	Mean Square Error	Standard Deviation
OLS	1,254,977	331,737	330,919	169,461
Lasso	1,255,710	332,141	330,822	169,581
ERC Lasso	1,261,106	333,814	330,796	169,687
KNN	1,284,811	331,480	338,625	169,911
GBR	1,619,245	383,290	460,096	140,186
Base Model 2002	1,886,204	389,839	686,847	385,450
Base Model 2016	2,346,774	364,293	761,013	411,505





Tax Estimation

Washington primarily taxes households through the retail sales tax, other consumer excise taxes, and property tax. We will model the major consumer taxes currently implemented in Washington, as well as the likely impact of implementing certain consumer taxes the State of Washington does not currently impose. With the exception of property taxes, we base all of the taxes we model on consumer expenditure estimates we derive from the Consumer Expenditure Survey.

Forecasting

The base year of our Washington consumer expenditure estimates—the combined Federal Tax Information and Consumer Expenditure survey datasets—is 2017. We must model tax revenues during the 2017-19 Biennium. We plan to inflate the results using estimates and forecasts published by the Washington State Economic and Revenue Forecast Council (ERFC) and IHS Markit. Proposed sources are listed in Table 6.

Table 6: Sources of consumption change estimates and forecasts

Tax	Forecast Source
Sales tax	ERFC Table 3-11
Alcohol	ERFC Table 3-11
Cigarettes and other tobacco products	ERFC Table 3-11
Marijuana	ERFC Table 3-11
Insurance Premiums	ERFC Table 3-11
Gasoline	IHS Markit Baseline Tables
Public utilities	ERFC Table 3-11
Property tax	ERFC Table 3-11
Real Estate Excise Tax	ERFC Table 3-11
Food	IHS Markit Baseline Tables
Consumer services	IHS Markit Baseline Tables
Medical services	IHS Markit Baseline Tables
Trade in value	Undetermined

Assumptions about economic behavior

Our model is economically naïve, in that our model does not account for changes in economic behaviors that we would expect households to exhibit in response to policy changes. It is not possible for us to measure the extent to which this affects our results.

First, we assume that between-state differences in cost of living and economic policy will not substantively affect our results. Cost of living differences may imply differences in the prices of goods, which can lead to consumer behaviors that differ from those of Washingtonians. For example, a household might spend a higher proportion of their budget on housing if rent prices are relatively high in the area in which they reside. Likewise, they might purchase “more” housing if rent prices are low, even while spending a lower share of their income. The actual effect of a change in prices depends on the price elasticities of the goods in question.

Tax policy can influence economic behaviors for households near a decision margin by altering prices. Exogenous price changes from a tax policy can alter the tradeoffs households observe when making their decisions. Behavioral deviations do not necessarily depend on the level of taxation, but rather on the extent to which a tax policy distorts which choices are optimal from consumers’ perspectives.

Government bodies often implement tax policies with the express intention of altering consumer behavior. This includes both taxes and tax expenditures, though taxes are our focus. Such “Pigovian taxes” reduce negative market externalities by including the externalities in the prices of goods. For example, Washington currently imposes additional excise taxes on tobacco products to reduce consumer demand. Similarly, some Washingtonians have suggested Washington should use a carbon tax to increase the price of carbon, leading businesses to reduce carbon emissions. Because Pigovian tax rates vary substantially across state and local jurisdictions, we might expect that consumer behaviors vary for goods subject to these taxes. This would be a source of bias in our model.

Likewise, we are directed to assume the party on whom the state initially imposes a tax incurs the full burden of the tax. In the real world economy, the group of entities who ultimately bear a tax burden is broader than the group on whom the state formally imposes the tax. For example, an excise tax increase on a specific consumer good might lead the producers of the good to decrease prices. Alternatively, an increase in taxes on firms might lead the firms to increase the prices of the goods they produce. In either example, buyers and sellers ultimately share the tax burden even when the initial tax incidence falls wholly on one group or the other. The degree to which parties share a tax depends on the “price elasticity of demand” and the “price elasticity of supply”—i.e. the economic relationships between parties and goods who are directly or indirectly connected to the tax. Our household tax burden model cannot account for the nuances in tax incidence we might encounter in real-world tax implementation.

Expected Outputs

The primary model output is a table for each of the geographic regions for which we compute estimates. We use each table to describe the mean household tax burden for each major consumer tax, as well as the mean total household tax burden, by income quantile. The tables also describe tax burden as a share of income within each income quantile. We measure tax burden as the total tax imposed on a household from state sources.

Table A2 in the appendix contains the output of the 2014 Household Tax Burden Model. It is our objective to replicate that table—with updated results—for each geographic region and statewide in which we estimate tax burdens.

Resource Links

Data Sources

Consumer Expenditure Surveys Public-use Microdata Getting Started Guide

American Community Survey Public Use Microdata Technical Documentation

Office of Financial Management April 1 Official Population Estimates

IRS Prior Year Forms and Instructions

Washington State Economic and Revenue Forecast Council

IHS Markit

Modeling and Methodology

Washington Excise Tax Microsimulation Model 2002

Elements of Statistical Learning

Multi-target regression via input space expansion: treating targets as inputs

Table A1: Comparison of IRS Individual Income Tax Data, *American Community Survey (ACS), and Office of Financial Management (OFM) datasets across key variables. Estimates from the IRS Individual Income Tax Data are likely to change marginally as we revise our data preparation procedure.*

Data	Population (Persons)	Number of Households	Mean Family Size	Mean Household Income	Total Income	Mean Household Wage Income	Total Wage Income
ACS household 2013-2017	6,847,968	3,025,516	2.49	$88,516	$244b	$64,668	$178b
ACS person 2013-2017	7,169,967	-	-	-	$241b	-	$182b
ACS household 2017	7,091,096	3,103,263	2.5	$93,728	$266b	$71,254	$202b
ACS person 2017	7,405,743	-	-	-	$272b	-	$206b
OFM 2017	7,310,300	3,083,371	-	-	-	-	-
IRS Individual Income Tax Data collapsed 2017	6,759,307	2,563,022	2.64	$116,208	$298b	$78,889	$202b
IRS Individual Income Tax Data tax units 2017	6,759,307	3,335,675	2.03	$89,290	$298b	$60,615	$202b

Table A2: Output of the 2014 Household Tax Burden Model

Tax Burden on Households

Major State and Local Taxes

Current Law

Household Income	$0	$15,000	$25,000	$35,000	$45,000	$55,000	$70,000	$85,000	$105,000	over
	$15,000	$25,000	$35,000	$45,000	$55,000	$70,000	$85,000	$105,000	$140,000	$140,000+
Retail Sales Tax	$905	$1,170	$1,453	$1,690	$1,988	$2,340	$2,729	$3,217	$3,832	$5,908
Alcoholic Beverages Taxes	$55	$63	$78	$93	$98	$119	$125	$143	$167	$239
Cigarette & Tobacco Taxes	$156	$177	$194	$197	$204	$211	$198	$193	$166	$119
Insurance Premiums Tax	$22	$36	$47	$55	$67	$74	$83	$94	$107	$138
Gasoline Tax	$132	$184	$233	$273	$311	$349	$392	$427	$467	$498
Public Utility Taxes	$113	$140	$157	$171	$186	$199	$217	$236	$258	$324
Property Tax	$804	$1,089	$1,307	$1,524	$1,896	$2,230	$2,678	$3,102	$3,824	$6,130
Total Tax	$2,187	$2,859	$3,469	$4,003	$4,749	$5,522	$6,422	$7,413	$8,821	$13,354
Tax as % of Income	26.5%	14.2%	11.5%	10.0%	9.5%	8.9%	8.3%	7.8%	7.3%	5.9%