Two-Stage Employment Classification

Aditya Retnanto
Department of Computer Science, Harris School of Public Policy
University of Chicago
aretnanto@uchicago.edu

April 21, 2023

1 Introduction

1.1 Background

Granular data that captures a household's economic well-being is needed to assist policy-makers and government agencies provide faster assistance to their constituents. A mechanism to do so is the Survey of Income and Program Participation (SIPP), a multi-year survey that follows multiple households in the United States [1]. The program is constrained by its data collection method: the annual survey asks participants to provide information about the previous year. When joining data from the prior year's survey to the current year's, there is a noticeable difference between the last month of a year and the first month of a year. This is known as *Seam Bias*, where, for each survey period, researchers have noticed that responders tend to overreport the final month's values [6].

1.2 Data Set

We will examine data from the 2018 cohort where 3 waves of interviews have already been conducted. Similar to Ham et Al, we will limit the scope to women between the ages of 16 to 55 [5]. The data extraction and filtration script are appended in the document. Specific features are discussed in the following sections of this paper.

1.3 Literature Review

Prior to 2014, SIPP was conducted three times a year. In an effort to reduce costs, the survey has henceforth been administered once annually. This may have had the unintended consequence of creating a more pronounced seam bias.

Methods to counteract seam bias have previously been studied. Grogger (2003) avoids interacting with seam bias by dropping the first three months and retaining the last month of a wave [4]. This approach constrains policy makers, because they are only offered insights every 4 months. Ham et. al proposes including a "dummy-variable" feature, a binary indicator, for seam month and then using that to perform regression [5]. We did not encounter supervised learning approaches in the literature that account for seam bias. It will be difficult to benchmark our approach. However, utilizing machine learning to augment instrumental variable is not novel. Xu et. al utilizes deep learning methods to obtain non-linear effects on treatment and instruments [8].

1.4 Approach

We would like to predict whether or not a survey respondent is employed. In the dummy-variable approach, the coefficient of the predicted weight corresponding to the binary indicator represents the effect of belonging in a biased month. However, this assumes that the effects of seam bias is constant over each unit and over time. This also assumes that in non-biased months, the responders are reporting with the same amount of 'noise' each time. We aim to capture this uncertainty by utilizing a two-stage approach, first classifying whether a sample is from a seam month. Using a

Relevent Vector Machine (RVM), we can obtain both a binary classification label and probabilities of belonging to that certain label. We then use probability as features for the linear regression to classify employment. Constraining the second stage to a least squares approach maintains the output's interpretability.

2 Model

2.1 Motivation

We define $y \in \mathbb{R}^n$ as a vector of labels (whether a person is employed or not), $X \in \mathbb{R}^{nxp}$ as the design matrix. Additionally $y_i = \{-1, 1\}$ for $i = 1 \dots n$, a binary label. The true model is denoted as:

$$y = Xw + \epsilon$$

where ϵ is noise. We would like to to estimate w by $\hat{w}$ to get a predicted $\hat{y}$. To evaluate the model we introduce an error function.

$$e = y - \hat{y} \tag{1}$$

We can finding an optimal set of weights that minimizes e over a loss function. Substituting in to the error term yields:

$$e = X(w - \hat{w}) + \epsilon$$

Clearly the error term is not solely due to the modeling error but is also a function of the noise term. To proceed with the methods taught in class, an assumption that needs to be made is that $\epsilon \sim N(0, \sigma^2 I)$. However, this is not the case for our problem. Let's reframe the approach using a panel data approach where we look at a unit's features over time.

$$y_{it} = X_{it}w + \epsilon_{it} \tag{2}$$

where i is an indicator for the a unique survey responder, t is the the survey month. The assumption that $\epsilon \sim N(0, \sigma^2 I)$ no longer holds. For t that corresponds to interview months our noise term would have different parameters of it's distribution and can impact the model. Additionally noise for each individual can differ. One can create models for each individual; however, there are associated costs to do so. Approaching the problem in a solely deterministic way has its limitations. Using a stochastic lens, the model can then be redefined as

$$y_{it} = X_{it}w + \epsilon_{it} + d_{it} \tag{3}$$

where d_{it} is the likelihood of a data point being in a seam month.

2.2 Relevent Vector Machines

Intuitively, we postulate that there is a clear seperation between values during seam and non-seam months, yet the effects are different per unit over time. We would like the model to be robust to outliers, and hence would like to create a margin between the two sets of data. The method introduced in class are Support Vector Machines (SVM). However, as the effects of seam bias is not constant per unit, we can not make a definitive claim that a data point has seam bias. RVMs augment the SVM as it provides an additional probabilistic output [7]. We will show a brief derivation of the classifier. As shown in class we can use the kernel trick, and obtain the following:

$$\hat{d} = K\alpha \tag{4}$$

Where $\alpha = (K + \lambda I)^{-1}y$. The kernel used for RVM is Gaussian and is denoted below:

$$K_{i,j} = k(x_i, x_j) = e^{-\frac{||z_i - z_j||_2^2}{2\sigma^2}} \tag{5}$$

x_i and x_j are vectors in the X matrix from equation (3). The derivation onwards is developed by Tipping [7]. Parameter derivation in the RVM is out of the scope of this class. The Bayesian approach allows us to bypass the need for a step-size (used in SVM). Rather than minimize an error function, we can iteratively maximize the posterior probabilities. This is analogous to gradient descent in the case of SVM as the optimization is also not closed form. We utilize the parameters used by Tipping [7], where μ is the posterior mean and Δ is the posterior covariance.

$$\Delta = \sigma^{-2}K^T K + A \tag{6}$$

$$\mu = \sigma^{-2}\Delta^T K y \tag{7}$$

A is composed of Diag(a) a vector normally distributed priors of α. Σ. For each iteration we can calculate a new set of a_i such that:

$$a_i^{(new)} = \frac{1 - a_i^{(old)}\Delta_{ii}}{\mu_i} \tag{8}$$

Additionally for the parameter σ, we can calculate

$$(\sigma^2)^{(new)} = \frac{||d - K\mu||_2}{N - \Sigma_i(1 - a_i\Delta_{ii})}$$

Let s be a stop criteria, when

$$||a^{(new)} - a^{(old)}||_2 < s$$

we stop the iterative optimization. Once a converges, the final Δ, σ and μ is retained. To predict future samples we can obtain both it's predicted label and confidence.

$$\hat{d}_{new} = \mu^T \phi(z_{new})$$

$$\sigma^2_{new} = \sigma + \phi(z_{new})^T \Delta \phi(z_{new})$$

Additionally, we set a $a_{threshold}$ such that if $a > a_{threshold}$, we "prune" that feature. Tipping states that this is equivalent to being certain that the weight associated with this is 0 [7]. For this project, implementation is attached within the zip file. While, the code is original, implementation of the RVM is guided by psuedocode from [3].

2.3 Second Stage

After obtaining the a vector of $\hat{d}$, we append this to the classic least squares classification such that

$$y_{it} = Xw + w_{p+1}\hat{d} + \epsilon \tag{9}$$

3 Experimentation

3.1 First Stage RVM

We would like to predict equation (4). The features of this model will be the *Mean Monthly Sum of Income per Unit and Wave, Difference of Income between Waves* to predict *Seam Month*. In this stage of the experiment we will not do a train and test split. There is a heavy class imbalance. Additionally when applying (5), to all training samples the process took 32 mins 15 s. The iterative method to maximize posterior probabilities will be computationally taxing to compute; hence it would be wise to drop samples. Since most of the samples are non-seam months, we chose to retain months 11, 12, and 1. This reduces the feature matrix to 3809×2 kernalization to 1 min 34s. Next we evaluate the training performance of both the least squares approach and the RVM. We define a sign function such that

$$\hat{d} >= 0.5, 1$$

$$\hat{d} < 0.5, 0$$

In the least squares approach, we notice that the output is skewed towards 0. The algorithm plays it safe by assuming that the label has no seam bias. In RVM $\hat{d}$ corresponds to a probability of that sample being a member of that class. Using the above sign function the algorithm predicts 28 instances of seam bias, all of which are correct. Although the performance of the RVM is marginally better than least squares, the output can be utilized for the next stage model to signify a weighted impact of being in a biased month. We denote d_{RVM}, d_{LS}, d_{KRR} and d_{Dummy} each corresponding to the prediction of based on full data set. Intuitively, the RVM provides the uncertainty of seam bias presenting within a class, Least Squares provides the projection of bias explained in the data, KRR maps it to a distribution and then projects bias onto the mapped space, and the dummy variable represents the seam month.

3.2 Second Stage Least Squares

Here we can finally predict $\hat{y}$, the employment status of an individual. We will predict (3) with varying $\hat{d}$ obtained from different regression. The intuition is that a higher classification accuracy means seam bias is accounted for within the design matrix rather than the residuals. One important thing to note is that the employment labels are imbalanced. One work around is to up-sample the dataset, where we make extra copies of the under sampled label. In this case, it would be the unemployed label. We can then randomly split the data into train, test split and calculate the average error rate. The results are shown in Table 1.

Table 1: Classification error of regression

First Stage	Error
Dummy Variable	49.17%
Relevance Vector Machine	30.03%
Kernal Ridge Regression	49.31%
Least Squares	49.31%

4 Conclusion

In this project, we have shown that adding the uncertainty of a sample improves the classification of employment by 20% compared to the alternative approaches. These results are preliminary and more rigorous analysis is required. Nevertheless, there are many avenues for improvement and future iteration. Additional knowledge in Bayesian statistics would also aid in the tuning of parameters to detect seam bias. Domain knowledge of the SIPP data set would greatly aid the design of the feature matrix. With additional features, we could potentially improve the classification accuracy of the model. One challenge was that with the 2014 survey redesign, respondents in previous years may have been able to more accurately recall their prior months' responses, as opposed to respondents post 2014, who were surveyed once annually, and likely struggled more to recall their previous responses. The RVM still assumes a uniform standard deviation for the bias, yet through iterative optimization of its posterior, we obtain a distribution for the estimate of a seam bias.

References

[1] https://www.census.gov/programs-surveys/sipp/about/sipp-content-information.html

[2] J. Q Candela. Learning with Uncertainty – Gaussian Processes and Relevance Vector Machines.

[3] T. Fletcher. Relevance Vector Machines Explained.

[4] J. Grogger. Welfare Transitions in the 1990s: The Economy, Welfare Policy, and the EITC. In *Journal of Policy Analysis and Management*, 23(4): 671-695, 2004.

[5] J. C. Ham, X. Li and L. Shore-Sheppard. Correcting for Seam Bias when Estimating Discrete Variable Models, with an Application to Analyzing the Employment Dynamics of Disadvantaged Women in the SIPP 1. 2007.

[6] L. J. Rips, F. G. Conrad, S. S. Fricker, Straightening the Seam Effect in Panel Surveys, *Public Opinion Quarterly*, 67(4): 522–554, 2003.

[7] M. E. Tipping. The Relevance Vector Machine. In S. A. Solla, T. K. Leen , and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 652-658. MIT Press, 2000.

[8] L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, A. Gretton . Learning Deep Features in Instrumental Variable Regression.