WC Imputation

Introduction

Worker's Compensation (WC) is intended to provide a safety net for workers who are injured or become sick on the job, by compensating for lost wages due to work-related injury or illness events, and medical coverage for these same events. These benefits are distributed through private carriers, state funds, and self-insured firms, with each state having its own reporting requirements and practices.

The Current Population Survey (CPS) provides WC micro-data in its March Supplement with its “WC_YN” and "WC_VAL" variables: these variables are the total number of WC (from all sources) recipients as well as their overall weighted-sum benefits respectively. Both WC benefits and WC recipients were underreported in the 2015 CPS compared to the administrative totals from the Social Security Administration's (SSA) 2016 Annual Statistical Supplement, and the National Acadamy of Social Insurance (NASI). More specifically, CPS WC benefits were underreported by 52.5 billion, with individuals only reporting 9.8 billion in claimed benefits, and recipients were underreported by 3 million, with 955,469 individuals receiving WC benefits. CPS underreporting typically occurs with government benefit programs, although this is rather extreme.

We augmented the total number of recipients and their dollar amount of benefits from CPS 2015 March Supplement to match the 2014 WC national reports provided by SSA and NASI. We used the 2015 CPS because the WC_YN and WC_VAL variables report the respondents' worker's compensation and recipiency for the calendar year 2014, rather than 2015.

In addition to matching WC recipient numbers and total benefits, we tried to maintain the original CPS micro-data distribution by utilizing current distribution and WC eligibility rules. However, since CPS data is insufficient in many ways (and partially SSA and NASI administrative data), a number of assumptions are made in order to augment the data reasonably. This report details those assumptions and explains our imputation procedure.

SSA obtains the annual WC benefits claimed amount on their annual statistical supplement from NASI. The total number of benefits claimed in 2014 is an estimate of the amount of benefits paid from all states, and all types of insurers (private carriers, state funds, and self-insurers). Since WC administrative data is not readily available, NASI gathers the benefits paid amount through a series of questionnaires sent to each state, data purchased from A.M. Best (a private company that specializes in gathering insurance data), and data from the National Council on Compensation Insurance (NCCI), called *Annual Statistical Bulletin*. According to NASI[1], this benefit estimate is complete, whereas the

[1] Chris McLaren, (202) 243 – 7280, cmclaren@nasi.org; Senior Researcher at National Academy of Social Insurance.

recipient totals that they entirely obtain from NCCI reports only include the 37 states where NCCI is licensed. Moreover, this recipient total does not include claims from workers who are injured at self-insured firms. Even though these WC totals are not state level specific, we use them since they seem to be the best estimate for WC administrative totals available; thus, we do not impute on the state level, rather on the national level using the NASI data.

The reported total of WC recipients/claims from NCCI is usually behind about 4 years. In our case, using 2016 reports, we had data on the total number of WC claims from 1995 to 2013 (see Table 1 in Appendix); thus, we lacked the 2014 target number of WC claims/recipients. We used an Auto-Regressive Moving Average Model (ARMA), with a Kalman Filter for smoothing, on the recipient data from 1995 to 2013 to impute 2014 recipient amount [2].

CPS micro-data and WC targets

In the Source of income – worker's compensation section of CPS March Supplement, the CPS contains the total worker's compensation benefit amount for each respondent in the "WC_VAL" variable and whether or not each respondent received any worker's compensation in its "WC_YN" variable.

Targets for imputation

The targets for imputation come from SSA and NASI data. The country level Worker's Compensation benefit and recipient totals come from NASI's 2016 "Worker's Compensation: Benefits, Coverage, and Costs" report. We get the total benefits claimed amount from Table 1 in the report, and the WC recipient total from Table 18. NASI's table 18 contains data gathered from NCCI's Annual Statistical Bulletin. Table 18 reports the total recipients/claims amount reported per 100,000 insured workers; SSA gives the total number of insured workers in their Annual Statistical Supplement, which was 129 million in 2014. Using this, we converted the per 100,000 recipients total from the NASI report, to the national recipient total. This gives us the administrative targets for WC_YN, and WC_VAL.

We use calendar year administrative totals because the CPS WC compensation is reported according to calendar year. In the calendar year 2014, administrative data suggests that approximately 3.9 million individuals claimed roughly 62.3 billion dollars in WC benefits. The annual average benefit for each recipient is approximately $15,833.

Below we provide the CPS WC benefit and recipient totals, and the corresponding SSA/NASI administrative benefit and recipient totals before our imputation.

Country	CPS benefits (annually)	Admin benefits (annually)	CPS total recipients	Admin total recipients

[2] The code for the ARMA model is found in arma.ipynb, which is found in the same file as the WC imputation script.

USA	$9,835,193,487	$62,300,000,000	955,469	3,934,656

Imputation Procedure

We follow a similar two-step procedure as the SSI imputation for augmentation, except we doing a national level imputation instead of a state level imputation. First, we add up the individual weights for each CPS respondent in the recipient pool to see if we reach the administrative level for the US. If we don't, then we augment by including the most likely recipients from the non-recipient pool, using probabilities derived from both a logistic regression and a Random Forest classifier model, until we match administrative totals for the US. Second, we obtain an adjust ratio for the benefit amount that allows us to match the administrative dollar benefit totals for the US.

Step I: Recipient Imputation w/ Probabilities from Two Models

Model I:

We propose a basic logistic regression model for analyzing the likelihood of being a WC recipient. Following WC eligibility rules, we use the available corresponding CPS information to create independent variables for which industry individuals work in, how old they are, whether they retired or left a job for health reasons, whether they have a health problem or a disability which prevents them from work or which limits the kind or amount of work they do, family disability income, whether dependents have private health insurance, and gender. These are all statistically significant independent variables for determining WC eligibility, and important variables listed in the Random Forest's feature importances.[3]

WC eligibility rules that we include:

- Eligible individuals must have received a work related injury or illness. The variables dealing with disability and illnesses, although not necessarily work related, serve as a proxy for this rule.
- Although not necessarily an eligibility rule, different industries may be riskier than others, thus increasing the likelihood of work-related injury or illness. Our industry indicators can capture this heterogeneity.

Below we give our proposed logistic regression model for predicting the likelihood of receiving WC.

[3] The code and results for the feature importances can be found in Rf_probs.py

$$\begin{gathered}
WC\_YN = \alpha + \mathit{Armed\ Forces} * \beta_1 + \mathit{Construction} * \beta_2 + \mathit{Educational} \text{ and } \mathit{health\ services} * \beta_3 \\
+ \mathit{Financial\ activities} * \beta_4 + \mathit{Information} * \beta_5 + \mathit{Leisure} \text{ and } \mathit{hospitality} * \beta_6 \\
+ \mathit{Manufacturing} * \beta_7 + \mathit{Mining} * \beta_8 + \mathit{Other\ services} * \beta_9 \\
+ \mathit{Professional} \text{ and } \mathit{business\ services} * \beta_{10} + \mathit{Public\ administration} * \beta_{11} + \mathit{Transportation} \text{ and } \mathit{utilities} * \beta_{12} \\
+ \mathit{Wholesale} \text{ and } \mathit{retail\ trade} * \beta_{13} + \mathit{age\_squared} * \beta_{14} + \mathit{dis\_cs} * \beta_{15} + \mathit{dis\_hp} * \beta_{16} \\
+ \mathit{finc\_dis} * \beta_{17} + \mathit{cov\_hi} * \beta_{18} + \mathit{a\_sex} * \beta_{19} + \varepsilon
\end{gathered}$$

This yielded a pseudo r-squared value of .25.

We then run the model on the CPS dataset. After, we use the fitted coefficients to produce a vector of probabilities for WC recipients. We then rank all recipients according to their fitted probability. We then aggregate the recipient weights, and add extra non-recipients by likelihood until the weights reach administrative US level.

Model II:

We use a Random Forest Classifier (RFC) model to determine WC recipient likelihood. Random Forests performed much better than the logistic regression model, with an accuracy/score of .99, compared to the logistic regression's pseudo r-squared of 0.25.

To train the RFC model we used all of the CPS variables except those that approximately identified those receiving WC benefits. To create feasible variables for the training, we converted all variables containing categorical strings into numerical categorical variables, and created proxy variables for many columns with missing data (Not in Universe, None, etc.).

After training the Random Forest on a training set (80% of the data), we computed the probability that each CPS respondent received WC compensation. Then, we ranked the probabilities as we did above, and imputed recipients until the CPS recipient US totals matched the administrative US totals [4].

Step II: Benefit imputation

For each imputed/augmented recipient, we assign the average benefit amount. We then calculate the new total benefits claimed for the US, and compare these totals with SSA/NASI administrative total benefits claimed. We calculate the adjustment ratio by dividing administrative benefits by the new total benefit amount. We use this adjustment ratio to augment individual's benefits to match the US administrative total.

[4] The code and score results for the Random Forest Classifier model can be found in C-TAM's github documentation

Appendix

Table 1: Number of Workers' Compensation Claims per 100,000 Insured Workers: Private Carriers in 37 Jurisdictions, 1995-2013. We used Total (including medical only).

Policy Period	Total (including medical only)	Medical Only	Medical Only Claims as Percent of Total	Temporary Total	Temporary Total Claims as Percent of Total	Permanent Partial	Permanent Partial Claims as Percent of Total
1995	7,377	5,689	*77.1%*	1,217	*16.5%*	459	*6.2%*
1996	6,837	5,281	*77.2%*	1,124	*16.4%*	419	*6.1%*
1997	6,725	5,230	*77.8%*	1,070	*15.9%*	414	*6.2%*
1998	6,474	5,035	*77.8%*	977	*15.1%*	452	*7.0%*
1999	6,446	5,047	*78.3%*	927	*14.4%*	461	*7.2%*
2000	6,003	4,685	*78.0%*	870	*14.5%*	437	*7.3%*
2001	5,510	4,277	*77.6%*	799	*14.5%*	423	*7.7%*
2002	5,239	4,036	*77.0%*	770	*14.7%*	422	*8.1%*
2003	4,901	3,747	*76.5%*	725	*14.8%*	423	*8.6%*
2004	4,728	3,635	*76.9%*	702	*14.8%*	385	*8.1%*
2005	4,571	3,514	*76.9%*	667	*14.6%*	383	*8.4%*
2006	4,376	3,351	*76.6%*	638	*14.6%*	381	*8.7%*
2007	4,076	3,107	*76.2%*	587	*14.4%*	375	*9.2%*
2008	3,615	2,730	*75.5%*	515	*14.2%*	363	*10.0%*
2009	3,452	2,659	*77.0%*	521	*15.1%*	357	*10.3%*
2010	3,486	2,616	*75.0%*	519	*14.9%*	347	*10.0%*
2011	3,411	2,563	*75.1%*	509	*14.9%*	335	*9.8%*
2012	3,279	2,466	*75.2%*	500	*15.2%*	308	*9.4%*
2013	3,202	2,398	*74.9%*	492	*15.4%*	307	*9.6%*
Percent change, 1995-2013	**-56.6**	**-57.8**		**-59.6**		**-33.1**	

Source: National Council of Compensation Insurance, 1997-2016, Exhibit XII, Annual Statistical Bulletin.

Table 2: Adjustment ratios of benefits for US

Country	Imputed	Admin	adjust ratio
USA	56997866388	62300000000	1.093

Table 3: Administrative and CPS totals after augmentation

Country	post augment CPS total	post augment CPS total infant	Admin total benefits	Admin total recipients

	benefits (annual)	recipients	(annual)	
USA	$62,300,000,000	3,935,000	$62,300,000,000	3,934,656

Table 4: Administrative and CPS totals after augmentation using Random Forest probabilities

Country	post augment CPS total benefits (annual)	post augment CPS total recipients	Admin total benefits (annual)	Admin total recipients
USA	$62,299,999,999	3,934,103	$62,300,00,0000	3,934,656