Logistic Regression Modeling and Worker Retention

Introduction and Methodology

Using the data set specified in Attachment B, we used binary logistic regression models to identify which ***newly hired*** workers were most likely to be retained. In this case the dependent variable was whether newly hired workers were retained in the subsequent quarter after hire (coded either “yes” or “no”), regressed on worker demographic characteristics and employment history. In the data set, workers can appear multiple times in each year and quarter or across quarters as dictated by the number of UI liable employers who paid them from first quarter 2005 (2005Q1) to first quarter 2009 (2009Q1), for a total span of 21 quarters.

Because of the repeated nature of observations (SSNs with employers), we first tested a repeated measures model and compared those results to a “regular” model which did not account for repeated observations. An investigation of the results revealed that there was little or no difference between the two methods. Because of this and the fact that the non-repeated measures model requires considerably less time to process (less than two minutes compared to 30 minutes for 900,000+ records), we proceeded with the former model rather than the latter.

The preliminary model contained the following variables:

1. Whether or not the worker was retained (ui_next_qtr)
2. Worker sex (male, female, unknown)
3. Worker age categories (16 – 19, 20 – 24, 25 – 34, 35 – 44, 45 – 54, 55 – 64, 65+, unknown)
4. Worker wage deciles (determined by year, quarter, and industry)
5. Industry (see attachment for complete list)
6. Quarter in which worker was hired by the employer (to account for seasonal effects)

Results and Discussion

Modeling results are shown in Attachment A. To keep the number of printed pages to a minimum, we chose to show the most important model output statistics.

The first set of output statistics contains the odds ratio estimates for the independent variables used in the model. We will now explain the odds ratios using the wage deciles results as an example. First, note that the reference category for the variable is the lowest wage decile. This means that all odds ratios for other deciles use the lowest decile as the base. For example, the estimated odds ratio for the 10 – 20% decile was 1.981. This means that people who earned wages in the 10 – 20^{th} percentile were 1.981 times as likely to be retained as those in the lowest decile when accounting for the other factors used in the model. Also note that as worker wages increase, so do the odds ratios associated with retention. Those in the highest deciles were six to seven times as likely to be retained as those in the lowest deciles. This was a result we anticipated and we are pursuing the use of other theoretically relevant variables. Suggestions are welcome.

One can also use the printed odds ratios to calculate additional statistics. For example, the male vs. unknown odds ratio estimate was 2.840, while the female vs. unknown odds ratio estimate was 3.410. From these two numbers we can determine an odds ratio estimate of male vs. female by dividing the former result by the latter result (e.g., 2.840/3.410), which yields 0.833. This means that males are 83.3% as likely as females to be retained by their employers when accounting for the other model factors.

Validation

Although the odds ratios provide information on relative risk of worker retention, we can observe how well the model performs by outputting calculated model probabilities to a data set and comparing predicted results to actual results (see Tables 1 and 2 of Attachment A). In these tables, the variable **pp** stands for the predicted probability of retention. For the purposes of our analysis, records with a predicted probability value of greater than 0.5 were classified as "Predicted Still Working", while those with **pp** values less than 0.5 were classified as "Predicted Not Working". Table 1 displays the results for mining, while Table 2 displays the results for construction.

When performing this kind of analysis, two categories of errors are possible. The first is a Type I or false positive (model predicts retention when the worker was not retained). The second is a Type II or false negative (model predicts non-retention when the worker was retained). These errors are quantified in both tables in the lower left and upper right boxes. In the mining table, we see that the Type I or false positive rate was 4.63% while the Type II or false negative rate was 80.49%. This indicates that the model does a much better job of identifying those who will be retained compared to those who will not be retained in mining. The overall accuracy rate in mining is calculated by dividing the number of correctly modeled outcomes by the total number of outcomes (e.g., (4,595 + 49,720)/75,686 = 71.7%). The results for construction (see Table 2) are somewhat different. Here we see that the false positive rate was 22.93%, while the false negative rate was 53.93%. The overall accuracy rate in construction was 63.8%.

Conclusion

The modeling example demonstrates how a potential sampling strategy could be optimized for the ARRA project. While the overall accuracy of the model was good (greater than 70%), variables will be added to see if the false negative rate in particular can be decreased. Future iteration results will be reported to the workgroup as they become available.

Application to Other States' Data Sets

The modeling process and results discussed above serve as a guideline for similar activities pursued in other states. State-specific industry mixes; usage of labor, geography, and other factors could significantly impact not only the relevant variables used, but also the estimated outcome statistics.

Attachment A: Selected Logistic Regression Results

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
agecat 16 - 19 vs Undetermined	0.630	0.609	0.652
agecat 20 - 24 vs Undetermined	0.577	0.558	0.597
agecat 25 - 34 vs Undetermined	0.602	0.582	0.623
agecat 35 - 44 vs Undetermined	0.604	0.583	0.626
agecat 45 - 54 vs Undetermined	0.631	0.609	0.654
agecat 55 - 64 vs Undetermined	0.683	0.656	0.711
agecat 65+ vs Undetermined	0.686	0.650	0.725
wages 10 - 20% Decile vs Lowest 10%	1.981	1.956	2.006
wages 20 - 30% Decile vs Lowest 10%	2.922	2.881	2.963
wages 30 - 40% Decile vs Lowest 10%	3.880	3.818	3.942
wages 40 - 50% Decile vs Lowest 10%	5.112	5.019	5.208
wages 50 - 60% Decile vs Lowest 10%	6.088	5.957	6.222
wages 60 - 70% Decile vs Lowest 10%	7.010	6.829	7.196
wages 70 - 80% Decile vs Lowest 10%	7.007	6.798	7.222
wages 80 - 90% Decile vs Lowest 10%	7.042	6.798	7.294
wages Highest 10% vs Lowest 10%	6.451	6.205	6.707
sex female vs unknown	3.410	3.296	3.527
sex male vs unknown	2.840	2.746	2.936
qtr 1 vs 4	1.332	1.314	1.351
qtr 2 vs 4	1.430	1.413	1.449
qtr 3 vs 4	0.960	0.947	0.972
naics2d accomodation vs wholesale trade	0.354	0.342	0.366
naics2d administration vs wholesale trade	0.266	0.256	0.275
naics2d agriculture vs wholesale trade	0.427	0.403	0.452
naics2d arts vs wholesale trade	0.486	0.463	0.510
naics2d construction vs wholesale trade	0.399	0.386	0.413
naics2d education vs wholesale trade	0.895	0.859	0.932
naics2d finance vs wholesale trade	1.933	1.820	2.053
naics2d health vs wholesale trade	1.112	1.071	1.156
naics2d information vs wholesale trade	0.958	0.904	1.014
naics2d managment vs wholesale trade	0.447	0.382	0.524
naics2d manufacturing vs wholesale trade	0.764	0.733	0.797
naics2d mining vs wholesale trade	0.906	0.874	0.940
naics2d other services vs wholesale trade	0.458	0.440	0.477
naics2d professional vs wholesale trade	0.669	0.641	0.698
naics2d public administration vs wholesale trade	1.238	1.185	1.293
naics2d real estate vs wholesale trade	0.648	0.616	0.681
naics2d retail trade vs wholesale trade	0.524	0.506	0.543
naics2d transportation vs wholesale trade	0.682	0.654	0.711
naics2d unknown vs wholesale trade	1.411	1.149	1.733
naics2d utilities vs wholesale trade	2.128	1.878	2.410

Attachment A: Selected Logistic Regression Results

Table 1 of pp by ui_next_qtr
Controlling for naics2d=mining

pp / Frequency, Percent, Row Pct, Col Pct	ui_next_qtr(ui_next_qtr): not employed in 2nd qtr	employed in 2nd qtr	Total
Pred. Not Working	4595 6.07 65.55 **19.51**	2415 3.19 34.45 **4.63**	7010 9.26
Pred. Still Working	18956 25.05 27.60 **80.49**	49720 65.69 72.40 **95.37**	68676 90.74
Total	23551 31.12	52135 68.88	75686 100.00

Table 2 of pp by ui_next_qtr
Controlling for naics2d=construction

pp / Frequency, Percent, Row Pct, Col Pct	ui_next_qtr(ui_next_qtr): not employed in 2nd qtr	employed in 2nd qtr	Total
Pred. Not Working	29977 19.68 59.97 **46.07**	20013 13.14 40.03 **22.93**	49990 32.81
Pred. Still Working	35097 23.04 34.29 **53.93**	67267 44.15 65.71 **77.07**	102364 67.19
Total	65074 42.71	87280 57.29	152354 100.00

Attachment B: Wyoming's Variable List for Consideration of Retention Models.

The current models we are constructing for the likelihood of being hired in one quarter and retained to the next are based on historic wage records, QCEW micro data to capture industry and employer characteristics and driver's license (among other databases with demographic data) to capture demographics. Once I get the inventories of what the other states have we will test our models with limited data access. For example, we may try to run the models with a limited Wage Records history or an absence of age and gender.

Our current model (still in development) includes the following for every SSN, UI, Year, Qtr record from 2000q1 to 2009q3.

Study	Variable Name	Variable Description
1	ssn	Social Security Number
	year	Year of wages
1	qtr	Quarter of wages
	period	Numerical representation of year and quarter 1900q1 = 1, 1900q2 = 2, etc
1	sex	Gender
1	age	Age in quarter of employment
	ui	Unemployment Insurance Account number
1	naics2d	Two digit NAICS code of employer
1	wages	Wages paid to SSN in quarter
1	ui_qtr_exp	Total quarters of experience the SSN has with the employer
	ui_qtr_poss	Total quarters the SSN could have with employer if continuously employed
	ui_tw	Total wages
	ui_aw	Average quarterly wage of the SSN with the employer
	ui_prev_qtr	Does the SSN appear with the employer in the previous quarter
1	ui_next_qtr	Does the SSN appear with the employer in the next quarter
	ui_qtr_tocome	Total quarters the SSN appears with the employer after this quarter
	naics2d_n_ui	Total number of UI accounts the SSN worked with in the same 2 digit NAICS industry
	naics2d_qtr_exp	Total number of quarters experience the SSN has with the 2 digit NAICS industry
	naics2d_qtr_poss	Total quarters the SSN could have worked in the 2 digit NAICS industry
	naics2d_tw	Total wages in the 2 digit NAICS industry
	naics2d_aw	Average quarterly wage of the SSN with the 2 digit NAICS industry
	naics2d_qtr_tocome	Number of subsequent quarters the SSN will work with the 2 digit NAICS industry
	wy_n_ui	Total number of UI accounts the SSN has ever worked with in Wyoming
	wy_n_naics2d	Total number of 2 digit NAICS industries the SSN has ever worked with in Wyoming
	wy_qtr_exp	Total quarters the SSN has worked in Wyoming
	wy_qtr_poss	Total quarters the SSN could have worked in Wyoming
	wy_tw	Total wages paid to the SSN while working in Wyoming
	wy_aw	Average wage the SSN made in Wyoming
	wy_qtr_tocome	Number of subsequent quarters the SSN will work in Wyoming
	rate_growth_ui_1y	The UI accounts percent growth in number SSNs over the previous 8 quarters
	rate_growth_ui_2y	The UI accounts percent growth in number SSNs over the previous 12 quarters
	rate_growth_naics_1y	The 2 digit NAICS percent growth in number SSNs over the previous 8 quarters
	rate_growth_naics_2y	The 2 digit NAICS percent growth in number SSNs over the previous 12 quarters
OS	wc_hit	Was the SSN a workers compensation claimant in the current quarter
OS	days_lost	Days lost as a result of workers compensation claim
OS	prior_wc	Prior number of times the SSN was a workers compensation claimant

Study 1: Doug Leonard's regression analysis discussed on 2/18/2010 conference call used the indicated variables. The criteria to determine who was included in the model was that ui_qtr_exp = 1 which meant that the SSN had never previously occurred with the UI account. The outcome used was ui_next_qtr which equaled 1 for retained and 0 for not retained.

OS: Indicates the field is not relevant for research under discussion but is captured for other research R&P is conducting.

T Glover, Wyoming Department of Employment, Research Planning. 02/18/2010