

Applying Logistics Regression to Forecast Annual Organizational Retirements

Alan Dunham, Greybeard Solutions®, LLC

ABSTRACT

This paper briefly discusses the labor economics research about employee decisions to retire from corporations and government organizations, and then shows how that reasoning can be used for practical forecasting by applying logistics regression. A test dataset is generated and stochastically perturbed for the forecasting exercise. The SAS® PROC LOGISTICS procedure is used to forecast the probabilities of annual individual retirements. Given the binomial nature of the retirement decision, the probabilities can be summed to obtain annual total retirement forecasts. Theoretical underpinnings, test data, code, and results for the test dataset are provided.

INTRODUCTION

Under a changing employment landscape, various employee cohorts transition in and out of corporate employee rolls at rates that vary significantly year-over-year. The unforeseen transitions also affect productivity and timely product delivery. All of those transitions affect productivity, recruiting, and, ultimately, retirement costs.

The final transition from the perspective of corporations is the individual employee retirement decision. Unfortunately, so-called 'baby-boomers' (born between 1946 and 1964) are retiring at rates that confound the simplistic approach of projecting forward using last year's retirement rates. Nearly 6-in-10 boomers now expect to retire at age 65 or later, including 26% of boomers that anticipate retirement at age 70 or later. By 2014, 23 percent of men and about 15 percent of women ages 65 and older were in the labor force, and these levels are projected to rise further by 2022, to 27 percent for men and 20 percent for women.

Complicating issues include early out programs, corporate staff reductions, and corporate mergers and buyouts. Typically, corporations and government organizations use a simplistic projection such as last year's retirement rate, or an average of the past two or three years, or a weighted average such as exponential smoothing. However, there is a much better way to capture the exogenous impacts of uncontrollable events on individual retirement decisions.

Retirement forecasting is a difficult topic to tackle for public papers, because the needed demographic datasets often are constrained by privacy needs. The goal of this paper was to provide deductive logic to build a retirement decision model, show how to simulate demographic data, and then show an example application of SAS logistic regression to support a forecast. The actual simulated predictor variables generated by the code in this paper all seemed nicely conditioned, but the dataset as a whole turned out to be more chaotic than real life, which prevented complete forecast satisfaction. However, the availability of end-to-end complete code examples should prove useful to the SAS community.

DEVELOPING A RETIREMENT MODEL

CAPITALIZING ON DEDUCTIVE RESEARCH TO SHORTEN THE MODELING CYCLE

The current formula for acceptable model building generally involves applying multiple modeling methods to the data to see which gives better quantitative results. Where much is known and understood about a topic, a model can be developed using deductive reasoning rather than relying solely on mining data in a search for empirical correlation. We know a lot through previous research about factors that may influence individual retirement decisions. Reviewing that literature is beyond the scope of this paper, but the interested reader is referred to some selected papers shown in Other References below.

Taken collectively, previous research shows that individual decisions to retire from a given corporation or government organization may be affected by (not necessarily an exhaustive list):

- Age
- Years of employment
- Retirement plan eligibility
- Gender
- Health
- Accumulated wealth
- Alternative employment opportunities
- Ethnicity
- Generational attitudes and differences
- Early out offers
- Job satisfaction
- Opportunities for advancement
- Education
- Occupation
- Marriage
- Work conditions
- Wage rate
- Health of family members
- Reductions in numbers of corporate employees

A retirement decision model ideally would be able to incorporate all the above potential predictor variables, but the reality is that often we can directly measure only some variables such as wages, gender, age, occupation, and years of employment. Some variables, such as work conditions, are difficult to measure. Other variables, such as health and wealth, are difficult to obtain due to privacy constraints. We may be able to resort to a few substitute 'proxy' metrics to try to capture the impact of predictor variables such as health, wealth, and alternative employment opportunities.

CHOOSING A MODELING TECHNIQUE

First, some science: the decisions by all employees about whether to retire from a corporation or government organization becomes a Binomial distribution where each yes/no decision is a Bernoulli trial. The binomial distribution frequently is used to model the number of successes (retirements) in a sample of size n drawn with replacement from a population of size N . If the sampling is carried out without replacement, the draws are not independent (people are retiring so N goes down) and so the resulting distribution is a hypergeometric distribution, not binomial. However, for N (the number of employees) much larger than n (the number of retirees), the binomial distribution remains a good approximation.

First, why not use our old friend, linear regression? The answer is that, for binary outcomes, linear regression easily can provide out-of-range probabilities (less than 0 and greater than 1), and because of the variance structure, the fitted equation coefficient parameters are unstable (wide variance) and can be highly affected by outliers. Figure 1 (Sonawane: August 30, 2016) for example, could show probability of retirement versus an index of individual wealth or some other continuous predictor variable. In Figure 1, a fitted linear regression line shows some predicted probability values less than zero, which is out-of-range.

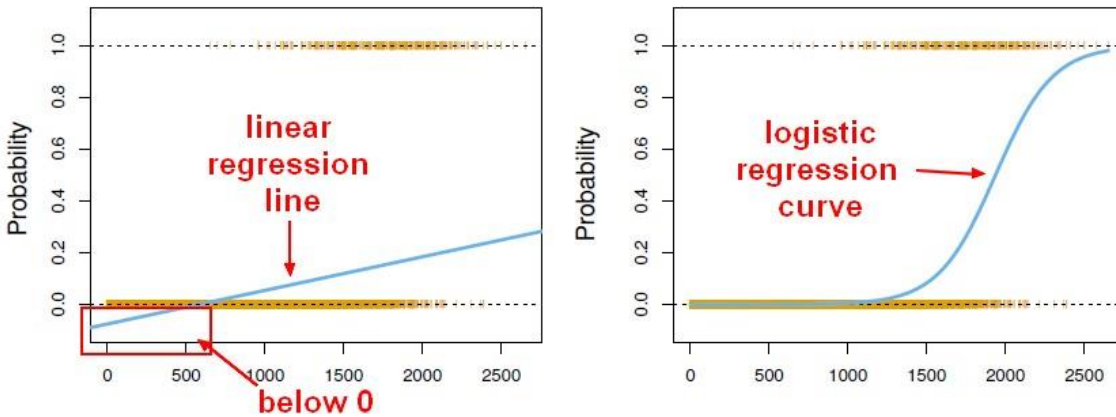


Figure 1 Probability vs Hypothetical Index of Wealth

The usual candidate modeling techniques for a binary choice model include logistics regression, decision trees, machine learning, neural nets, clustering, and others. We can expend a lot of client resources applying our data science skills to select a technique through model comparison. Here we focus on logistics regression because there is previous successful applications for this topic, plus logistics regression has some preferred scientific benefits in this matter:

- It is easier to interpret and understand the effects of predictor variables by computing closed-form results for any arbitrary combination of predictor variables.
- When classes are not well-separated, logistics regression tends to be less susceptible to overfitting. For example, the effects of gender, occupation, and wealth typically do not result in sharply defined separations between retire and not retire.
- Fitted logistics regression parameters permit more fine-grained calculation of confidence intervals for any combination of predictor variable values. Other techniques, such as decision trees, provide a patchwork of confidence intervals across the decision space.
- Once the logistics regression model has been fitted, because it is a binomial distribution, forecasts of numbers of retirees is a straightforward operation of summing the probabilities of individual probabilities of retirement.
- It is reasonably simple to 'age' the workforce and obtain retirement forecasts for one or more following years beyond the current year.
- Under some reasonable conditions of constant time interval observations, serial or autocorrelation does not cause complications for logistics regression.

SPECIFYING THE LOGISTICS REGRESSION RETIREMENT DEMONSTRATION MODEL

There are many excellent books and papers describing the math and derivation of logistics regression methodology. The reader is referred to the classic and eminently readable Logistics Regression Using SAS (Allison 2012). Another classic reference is Applied Logistic Regression (Hosmer, Lemeshow, Sturdivant: 2013). Additional papers are cited in the references.

Here we proceed directly to the model equation to be fitted, shown as Figure 2:

$$\ln(p/(1-p)) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots \beta_d * X_d$$

Figure 2 General Logistic Regression Model Equation

where p is the probability that an individual will retire, X's are the predictor variables, and β's are the coefficients to be obtained in this demonstration with PROC LOGISTIC.

This model has the excellent property that the probabilities that individuals will retire can be obtained easily by simple calculations using the estimated b coefficients. SAS will do those calculations for you in PROC LOGISTIC, or you can use a spreadsheet. Figure 3 shows the general formula, where Y=1 means that the individual chose to retire:

$$p(Y = 1|X_1 \dots X_d) = \frac{\exp[\beta_0 + \beta_1 X_1 \dots + \beta_d X_d]}{1 + \exp[\beta_0 + \beta_1 X_1 \dots + \beta_d X_d]}$$

Figure 3 Probability Equation

For demonstration purposes, we will create a retirement forecasting model for a hypothetical corporation using these predictor variables (the X's above):

- Years – Years of employment with current employer
- Generation (other common definitions in literature may vary a little from these)
 - 1 for pre-World War II (age 72 or older)
 - 2 for Baby Boomer (ages 53 – 71)
 - 3 for Gen X (ages 37 – 52)
 - 4 for Millennial (ages 18 – 36)
- ITOccupation – employed in some sort of information technology occupation
- Age – in years
- Gender – 1 if female, 0 if male
- U6 – unemployment rate for the year including discouraged workers
 - Used as a proxy for alternative job opportunities
- LagU6 – U6 the previous year (U6 lagged one year)
- SandP500 – Standard and Poor 500 index for the year
 - Used as a wealth proxy
- LagSandP500 – SandP500 last year (Lagged one year)
- Available_Sick_Days – a health proxy variable

THE DATA

We need individual demographic data over time for our approach to work,. Both corporations and government organizations may permit inside employee analysts to access the needed data, but generally are loath to release individual demographic data for external research. This is an understandable restriction to preserve individual privacy. Fortunately, we can use SAS to create reasonable simulated individual demographic data for our purposes. This section provides detail on how that is done, which includes some interesting code logic. However, this section about simulating the data can be skipped without losing information about the logistic regression retirement forecast.

The observations for this demonstration are simulated using the code described below, which can be obtained by email from the author. A reader who uses SAS to run the code as is will obtain exactly the same dataset, although the designated library (sesug17) will have to be changed.

The code creates one observation per employee per year until retirement. The first code segment shown starts the Data step with two loops, the outer for year, and the inner to generate 100 employee observations for each year. The number 100 is arbitrary, and should provide enough observations to allow for variability among the predictors. The array _temporary_ sets up the proportions for the four

generations of interest (McNulty:2013). Gender is created with the uniform random number generator, rounded to the nearest single digit integer (0,1), which should result in an even split between Males and Females.

```
***Create a logistics regression teaching dataset;
***First, the predictor variables are generated;
***Next, the Retired variable is created as 1(retired) or 0(not retired);
***The time interval is year, but it could be quarter or month depending on
client needs;
data sesug17.LogReg_Retire_Dataset (drop = i x);
    *** use streaminit so the stream of random nummbers can be reproduced
    if desired;
    call streaminit(1969);
    *** set up a temporary array of probabilities for age distribution:
    WW2, Boomer, Genx, Millenial;
    array p[4] _temporary_ (.03, .31, .31, .35);
    do Year = 2007 to 2017;
        do i = 1 to 100;
            ***Identity numbers for observations often prove useful;
            IDNumber = 1000 + i;
            ***The gender random number call is from the uniform
            distribution;
            ***The round function will round to the nearest single digit
            integer, which is either 0 or 1 here;
            Gender = round(rand("Uniform"), 1);
```

The following code segment is an interesting use of the _temporary_ array and the uniform random number generator to produce the right proportions of generations in the workforce. Note the method of ensuring the correct age interval for each generation, such as Baby Boomers are limited to age 53 plus random multiple of 18 years, for a maximum age of 71. Also, Available_Sick_Days is set to increase to a maximum of 30 days for Millenials under the notion that younger employees will be healthier and thus accumulate more sick leave. This may not be correct for a given organization, but is an assumption for the purposes of this example.

```
***Use adjusted random number call to create reasonable
proportions of employee generations.;
Generation = rand("Table", of p[*]);
    if Generation = 1 then do;
        Age = int(ranuni(0)*18+72);
        Available_Sick_Days = int(ranuni(0)*7);
    end;
    if Generation = 2 then do;
        Age = int(ranuni(0)*18+53);
        Available_Sick_Days = int(ranuni(0)*14);
    end;
    if Generation = 3 then do;
        Age = int(ranuni(0)*15+37);
        Available_Sick_Days = int(ranuni(0)*21);
    end;
    if Generation = 4 then do;
        Age = int(ranuni(0)*18+18);
        Available_Sick_Days = int(ranuni(0)*30);
    end;
```

The following code creates Information Technology employees as 15% of the employees to demonstrate the concept that some types of employees might have good job opportunities that would motivate toward

earlier retirement if eligible. Another rationale for earlier retirement might be that some occupations impose more wear and tear on health, but that is not represented in this example.

```
***Set 15% of employees to be in an IT occupation;
x = ranuni(0);
  if x <= .15 then ITOccupation = 1;
  else ITOccupation = 0;
```

Years of employment are set in a different range for each generation (derived from Bureau of Labor Statistics: September 2016):

```
*** Create years of employment with the current organization as a
function of generation;
  if Generation = 1 then Years = int(ranuni(0)*10.3);
  if Generation = 2 then Years = int(ranuni(0)*10.1);
  if Generation = 3 then Years = int(ranuni(0)*6.4);
  if Generation = 4 then Years = int(ranuni(0)*3.9);
```

Figure 4 shows the results of the Generation and Gender data simulation.

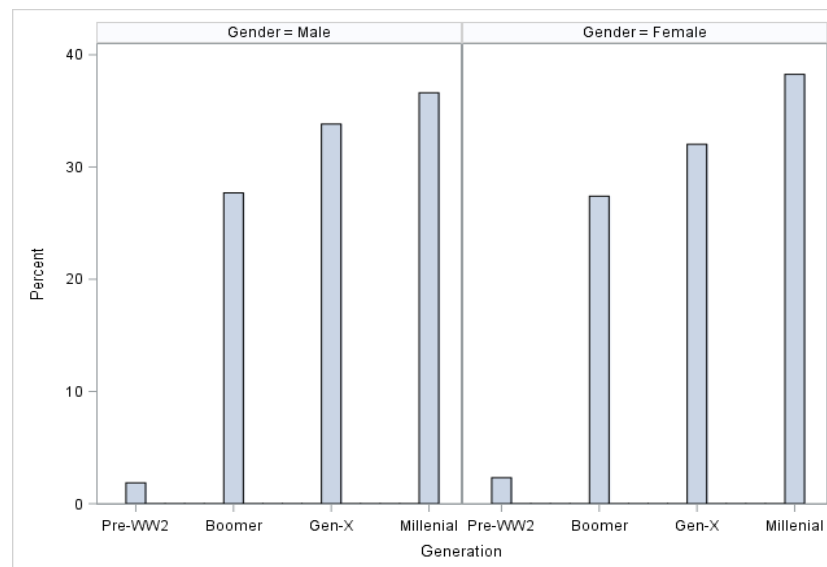


Figure 4 Histogram of Gender vs Generation from PROC SGPanel

Wealth and job opportunity proxy variables are taken from historical stock market results (Federal Reserve Bank: September 2017) and employment data (Bureau of Labor Statistics: September 14, 2017). An Early Out variable created for 2008, at the time of significant business turbulence:

```
** Enter S&P 500 and U6 historical data for Sept of each year;
  if Year = 2007 then do; SandP_500 = 1471.6; U6 = 8.4;
Early_out = 0;end;
  if Year = 2008 then do; SandP_500 = 1251.7; U6 = 11.0;
Early_out = 0;
  Early_out = 1;end;
  if Year = 2009 then do; SandP_500 = 1049.34; U6 = 16.7;
Early_out = 0;end;
  if Year = 2010 then do; SandP_500 = 1121.9; U6 = 16.8;
Early_out = 0;end;
  if Year = 2011 then do; SandP_500 = 1162.3; U6 = 16.4;
Early_out = 0;end;
```

```

        if Year = 2012 then do; SandP_500 = 1436.6; U6 = 14.8;
Early_out = 0;end;
        if Year = 2013 then do; SandP_500 = 1683.4; U6 = 13.7;
Early_out = 0;end;
        if Year = 2014 then do; SandP_500 = 1985.5; U6 = 11.8;
Early_out = 0;end;
        if Year = 2015 then do; SandP_500 = 1961.1; U6 = 10.0;
Early_out = 0;end;
        if Year = 2016 then do; SandP_500 = 2159.0; U6 = 9.7;
Early_out = 0;end;
        if Year = 2017 then do; SandP_500 = 2496.5; U6 = 8.6;
Early_out = 0;end;          output;
    end;
end;
run;

```

Since decisions to retire may be influenced by changing economic conditions with some lead time, we introduce some simple lagged and year-over-year change variables. Note the use of the “by” variable construction for which the data must be sorted. SAS keeps track of the first and last occurrences of a Year value, and these can be used to control processing. To use the “BY” construct we have to ensure the data are properly sorted:

```

***sort data to ensure order by ID and Year;
proc sort data= sesug17.LogReg_Retire_Dataset
    out= sesug17.LogReg_Retire_Dataset_Sort;
    by IDNumber Year;
run;

```

Next the lags and differences are created using a form of the DOW loop (Dorfman, Shajenko: 2012). Notice the use of the lag and dif functions, the implicit counter _N_, the “BY” construct, and the “first.” designator, which tracks the first instance of an ID Number: .

```

*** Add lagged and change (delta)variables to the dataset;
data sesug17.LogReg_Retire_Dataset_lag (drop= i j k l);
do _n_ = 1 by 1 until (eof);
    set sesug17.LogReg_Retire_Dataset_Sort;
        by IDNumber Year;
            i = lag(U6);
            j = dif(U6);
            k = lag(SandP_500);
            l = dif(SandP_500);

    LagU6= i;
    U6_Delta = j;
    LagSandP_500 = k;
    SandP_500_Delta =l;
    if (first.IDNumber) then do;
        LagU6 = .;
        U6_Delta = .;
        LagSandP_500 = .;
        SandP_500_Delta =.;
    end;
    output;
end;
run;

```

Finally, the actual dependent variable, Retire, is created using the method shown in (Wicklin:2013, p. 227). Note that 30% of the data are sent to a separate dataset with a Validation flag:

```

***Create the dependent variable and final datasets;
data sesug17.LogReg_Retire_Dataset_Final
    sesug17.LogReg_Retire_Dataset_Valid;
set sesug17.LogReg_Retire_Dataset_lag;
do i = 1 until (eof);
    eta = (.01 + .5*Age
        - .001*Gender
        + .005*Years_Corp
        - .05*Available_Sick_Days
        + .02*ITOccupation
        + .0005*SandP_500
        + .0007*LagSandP_500
        + .0009*SandP_500_Delta
        - .005*U6
        - .007*LagU6
        - .009*U6_Delta
        - .02*Early_Out) / 1000;
    mu = exp(eta) / (1+exp(eta));
    Retire = rand("Bernoulli", mu);
    if Age < 60 then Retire = 0;
    *** partition 30% of the data for validation;
    Validation = 0;
    x = ranuni(0);
    if x <=.3 then do;
        Validation = 1;
        output sesug17.LogReg_Retire_Dataset_Valid;
        continue;
    end;
    output sesug17.LogReg_Retire_Dataset_Final;
end;
run;

```

The simulated data created by the above code is for demonstration purposes only. The code examples are offered to show one way to use SAS code for the purposes shown, but do not produce real data.

FITTING THE RETIREMENT LOGISTIC REGRESSION MODEL

MODEL FIT USING THE MAIN MODEL DEVELOPMENT DATASET

The following code (run in SAS 9.4) fits a logistic regression curve to estimate the coefficients (β 's) for the data we created above. The `outest` dataset contains the coefficients. The `out` dataset contains the input dataset plus the estimated probabilities of retirement, and the upper and lower 95% confidence intervals.

```

proc logistic data= sesug17.LogReg_Retire_Dataset_Final
    outest = sesug17.LogReg_Retire_Dataset_outest;
model Retire (event= "1")=
    Gender
    Generation
    Age
    Available_Sick_Days
    ITOccupation
    Years_Corp
    SandP_500
    U6
    Early_out
    LagU6

```

```

        U6_Delta
        LagSandP_500
        SandP_500_Delta
    /lackfit;
    output out = sesug17.LogReg_Retire_Dataset_Out
    predicted = RetireProb lower = Lower95 upper = Upper95;
run;

```

Some of the PROC LOGISTIC procedure output is displayed below. The Response Profile shows that after removing 30% of the observations to a validation dataset, and after ignoring records that were missing (due to the lag for the first observation for each ID Number), there were 719 records, of which 70 were retirees.

The Null Hypothesis test tells us that the fitted equation is better than no equation at all at the .0001 confidence level. The Hosmer-Lemeshow Goodness-of-Fit Test (Allison:2012, pp66-68) confirms.

The Maximum Likelihood Estimates of the Parameters (β 's) do not look very good under the Wald Chi-Square test, except for Age.

Other than Age, the Odds Ratio estimates cannot be taken seriously due to the lack of statistical significance. However, as an illustration, if Early_Out were significant, it would indicate that, for someone who was present during the early out offer in 2008 and eligible to retire, their odds of retiring would have been over three times the odds of someone else retiring.

Figure 5 Selected Output from PROC LOGISTIC

Response Profile		
Ordered Value	Retire	Total Frequency
1	0	649
2	1	70

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	261.3970	11	<.0001
Score	209.3702	11	<.0001
Wald	72.2470	11	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-27.1023	18.4115	2.1669	0.1410
Gender	1	0.4003	0.3679	1.1839	0.2766
Age	1	0.2542	0.0328	60.1268	<.0001
Available_Sick_Days	1	0.0493	0.0428	1.3277	0.2492
ITOccupation	1	0.1793	0.5776	0.0963	0.7563
Years_Corp	1	0.0819	0.0621	1.7369	0.1875
SandP_500	1	0.0002880	0.00197	0.0214	0.8837

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
U6	1	0.2938	0.6296	0.2178	0.6407
Early_out	1	1.1595	2.9109	0.1587	0.6904
LagU6	1	0.1434	0.1879	0.5829	0.4452
LagSandP_500	1	0.00186	0.00425	0.1918	0.6614

Odds Ratio Estimates		
Effect	Point Estimate	95% Wald Confidence Limits
Gender	1.492	0.7263.069
Age	1.289	1.2091.375
Available_Sick_Days	1.051	0.9661.143
ITOccupation	1.196	0.3863.711
Years_Corp	1.085	0.9611.226
SandP_500	1.000	0.9961.004
U6	1.342	0.3914.608
Early_out	3.188	0.011957.974
LagU6	1.154	0.7991.668
LagSandP_500	1.002	0.9941.010

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.0327	8	0.3395

MODEL FIT USING THE VALIDITY TEST DATASET

The same model equation is used to fit against the validity test dataset (271 observations:35 Retired), so the code will not be displayed again. The results are very similar.

The overall equation chi-square test was statistically significant at the .0001 level. Also, the Hosmer-Lemeshow Goodness-of-Fit test was statistically significant.

The maximum likelihood estimates for the coefficients also are not significant except for Age.

Interestingly, the validity test model odds ratio estimate for Early_Out is a lot smaller (.016 vs 3.19), which is an example of why generally we ignore variables that are not statistically significant.

Figure 6 Selected PROC Logistic Output Using The Validity Dataset

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gender	0.771	0.279	2.131
Age	1.261	1.155	1.375
Available_Sick_Days	1.000	0.890	1.124
ITOccupation	1.415	0.383	5.227
Years_Corp	1.037	0.881	1.220

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SandP_500	1.004	0.995	1.014
U6	0.690	0.128	3.723
Early_out	0.016	<0.001	28.107
LagU6	0.467	0.219	0.997
LagSandP_500	0.987	0.973	1.001

COMPARING THE DEVELOPMENT AND THE VALIDITY TEST MODELS

Figure 7 shows selected PROC LOGISTIC outputs from the two model fitting runs. The overall magnitude of the -2 Log L criterion increases with number of observations, so it cannot be used to compare the two models. . Based on Aikake's Information Criterion (AIC) and the Schwarz Criterion (SC) (Allison:2012, pp. 22-23), the validity data run is a worse fit. However, with most of the predictor variables not statistically significant in both model runs, this comparison does not lead to any suggested model tuning.

Figure 7 Validity Model Fit Statistics

Model Fit Statistics Using Validity Data		
Criterion	Intercept Only	Intercept and Covariates
AIC	213.258	127.602
SC	216.896	167.624
-2 Log L	211.258	105.602

Model Fit Statistics Using Development Data		
Criterion	Intercept Only	Intercept and Covariates
AIC	461.064	223.905
SC	465.641	274.262
-2 Log L	459.064	201.905

CONCLUSION

This paper has shown end-to-end how to simulate data and apply PROC LOGISTIC to the simulated data for a specific problem. The empirical results using the simulated data are a bit problematic, preventing carry-through to an actual forecast demonstration by aggregating probabilities. One path for improvement would be to simulate employee data to follow individual employees over time, entering and leaving corporations and retiring. Also, it is likely that the parameters used in generating the dependent variable can be improved. The SAS code in this paper is self-contained and no additional data are needed. The code may be obtained by emailing the author.

REFERENCES

RECOMMENDED READING

- Allison, Paul D. 2012. Logistic Regression Using SAS: Theory and Application, Second Edition. Cary, NC. SAS Institute, Inc.
- Wicklin, Rick. 2013. Simulating Data with SAS. Cary, NC. SAS Institute, Inc.

ADDITIONAL REFERENCES

Bamberger, Peter A; Bacharach, Samuel B. "Predicting Retirement Upon Eligibility: An Embeddedness Perspective." 2012. Cornell University, Ithaca, NY. Available at <https://www.ilr.cornell.edu/sites/ilr.cornell.edu/files/Predicting%20Retirement%20Upon%20Eligibility%20An%20Embeddedness%20Perspective.pdf> .

Bureau of Labor Statistics. "Economic News Release: Table 1. Median years of tenure." September 22, 2016. Available at <https://www.bls.gov/news.release/tenure.t01.htm> .

Bureau of Labor Statistics. "(seas) U-6 Total unemployed, plus all marginally attached workers plus total employed part time." September 14, 2017. Available at <https://data.bls.gov/timeseries/LNS13327709> .

Carvalho, Carlos M. "Section 6: Model Selection, Logistic Regression, and More..." University of Texas McCombs School of Business. <http://faculty.mcombs.utexas.edu/calos.carvalho/teaching/> .

Chinsinglee. "Logistic Regression versus Decision Trees." September 28, 2016. Available at <https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/> .

Dorfman, Paul M., Shajenko, Lessia S. "How to DOW." 2012. Available at <http://support.sas.com/resources/papers/proceedings12/156-2012.pdf> .

Federal Reserve Bank of St. Louis. "FRED Economic Data: S&P 500." September 2017. Available at <https://fred.stlouisfed.org/series/SP500> .

Hosmer, David W., Lemeshow, Stanley, Sturdivant, Rodney X. 2013. Applied Logistics Regression, Third Edition. Hoboken, NJ. John Wiley & Sons, Inc.

McAllaster, LTC Douglas L. 1999. "Forecasting Army Officer Attrition with Logistic Regression Using the SAS System." Office of Deputy Chief of Staff for Personnel, Headquarters, US Army. Available at <http://analytics.ncsu.edu/sesug/1999/072.pdf> .

McNulty, Mary Ann and West, Elizabeth. "Gen X and Y Join Boomers." BTNGroup. September, 2013. Available at http://www.businesstravelnews.com/uploadedFiles/White_Papers/BTN_TTrans_GenX_0913-FINAL2.pdf .

Meeker, Mary. "Internet Trends 2015 – Code Conference." KPCB. May 27, 2015. Available at http://kpcbweb2.s3.amazonaws.com/files/90/Internet_Trends_2015.pdf .

Sonawane, Sayali. "Linear regression output as probabilities." August 30, 2016. Available at <https://stackoverflow.com/questions/12146914/what-is-the-difference-between-linear-regression-and-logistic-regression> .

CONTACT INFORMATION

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

Alan D. Dunham
Greybeard Solutions®, LLC
alan@greybeard-solutions.com

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.