

Forecasting Army Officer Attrition with Logistic Regression Using The SAS® System

by LTC Douglas L. McAllaster

formerly of, Office of Deputy Chief of Staff for Personnel
Headquarters, US Army, Pentagon, Washington DC

ABSTRACT:

This paper explains how Army personnel analysts model and forecast officer attrition with logistic regression using SAS System software. This statistical application applies a simple logistic regression model (only one predictor variable, fiscal year) to predict a binary response, whether officers stay or leave the Army. However, the complete application is large and complex since we run this logistic regression against more than one-hundred and fifty homogeneous BY-groups (officer cohorts). Furthermore, the application is enhanced with a semi-automatic capability to identify and exclude inappropriate outliers through the usage of SAS Macro Language. Thus, this application demonstrates the power SAS Software brings (1) to summarize vast amounts of data (fourteen years of legacy data - over a million observations), (2) to model binomial outcomes and forecast trends using logistic regression, and (3) to repeat the statistical analysis code using macro language.

DEFINE PROBLEM:

Army manpower analysts must quantify officer attrition rates and trends. Personnel planners use these attrition rates to (1) determine accession requirements (recruiting needs) (2) to determine promotion requirements and (3) to discern attrition trends.

First, analysts use the annual attrition rates in a Markov-chain probability model to determine the annual officer accession requirement. Then HQ US Army issues officer recruiting missions to the Reserve Officer Training Corps (ROTC), Officer Candidate School (OCS), and Warrant Officer Candidate School (WOCS) recruitment. Officer recruitment is the only method the Army has to replenish officer separations. Thus, this attrition analysis is a critical component in the Army's officer accession program, which is our only method to replenish officer separations.

Second, personnel analysts use these attrition rates to develop annual promotion plans by grade. That is, analysts must forecast the separations out of a grade in order to determine the number of officers it must promote into that grade and, thus, to staff the officer corps at authorized grades.

Third, senior leaders are especially concerned to identify and remedy deleterious attrition trends. In particular, analysts need to identify any trends showing increasing attrition among "due course" officers, those who have not been passed over. Senior Army leaders can then use this statistical ammunition as justification to request improvements to remuneration from Congress, such as: pay raises, better retirement programs, and other improvements to quality of life benefit programs.

EXPLANATORY VARIABLES:

Personnel analysts agree that there are three critical officer attrition explanatory variables: (1) grade (GRD:LT, CPT, MAJ, LTC, or COL), (2) years of service (YOS:1-30), and (3) number of times not selected for promotion (XNS:0,1,2+). Grade and YOS are obviously explanatory since these two variables determine officers' base pay.

Grade and XNS are key explanatory variables for attrition rates due to Army officer separation policy. Company grade officers must leave service after having been passed over (not selected) for promotion a second time. Majors who have been passed over for promotion to LTC must separate (retire) from active duty at 20 years of service. Thus, grade and XNS are key predictor variables.

Years of service (YOS) is the third key explanatory variable. Most officers must serve four or five years in fulfillment of their active duty service obligation (ADSO). Officers who continue after this period are increasingly likely to remain on active duty until retirement eligibility. Lieutenant Colonels may serve to 28 YOS and Colonels to 30 YOS.

HOMOGENEOUS BY-GROUPS:

The analyst must decide whether to model these variables as continuous or categorical. We choose to stratify our population into homogeneous groups based on the above three key categorical explanatory variables. That is, we separate the officer corps into homogeneous BY-groups as defined by GRADE, YOS, and XNS. We then analyze each homogeneous BY-group in order to determine the attrition trend for that group over time.

The appendix has a table of the data for one such homogeneous BY-group: Captains (GRD=3), having nine years of service (YOS=9), and who have not been passed over (XNS=0). There are 160+ such homogeneous BY-groups. The first task the analyst faces is getting the data. The following paragraphs explain how we get our data into a single analysis data set.

DATA WAREHOUSE:

The Army Personnel Command (PERSCOM) maintains a data warehouse called the Total Army Personnel Data Base (TAPDB). This data base has monthly legacy files at the individual record level. That is, each legacy file has a record for every officer on active duty at the end of each month. The officer's record has his key personnel data: SSN, grade data, passed-over data, and service data (date entered into service as well as months of service to date).

These files are readily accessible on IBM MVS mainframe tapes as far back as fourteen years. The key variables are fairly well maintained on the TAPDB. However, some of the files have erroneous years of service data which must be corrected using other service data columns.

JOINING RECORDS:

Data analysts created a data mart of data on local PCs by downloading key columns of annual legacy data from the PERSCOM mainframe data warehouse. Since each officer record has a unique key (SSN), analysts can readily join (or match merge) one end of fiscal year (EFY) file to the subsequent EFY file and create a binary outcome (0=stay and 1=leave), along with other key data, for each individual officer who started the FY. This binomial outcome is the response variable of interest.

SUMMARY DATA:

We further refine our data mart into a single analysis data set by summing over the binary outcome variable to achieve a data set in the "events-trials" format, within each BY-group. PROC MEANS with a CLASS statement quickly summarizes these 1.4 million records into about 2300 observations of grouped summary data. That is, we have about 160 BY-groups, each having about 14 FYs of data. Variable "leave" has the "events" = number of officers who left. Variable "start" has the "trials" = number of officers who started each FY. Thus, PROC MEANS summarizes our data legacy data into a single analysis data set but which is composed of separate homogeneous BY-group cohorts, officers having the same GRD, YOS, and XNS variables. The appendix includes an actual table of attrition data for a typical analysis cohort BY-group. The table shows attrition data on the BY-group: GRD=3, YOS=9, and XNS=0. Each BY-group has up to fourteen points, one for each legacy FY. Variable YearID identifies the fiscal year. We calculate variable O_Rate as the actual observed percentage which left service, leave/start.

BY-GROUP ANALYSIS:

Fortunately, SAS PROCs are well equipped to handle data which are grouped in this manner with the BY statement. PROC MEANS creates a sorted the output data set and we simply include a BY statement on our call to PROC GENMOD. The procedure then processes each BY-group separately, fitting a separate logistical regression line to each BY-group of data.

LOGISTIC REGRESSION:

Logistic regression is the appropriate statistical model for this binomial response, the rate at which officers leave the service. Our model is a simple logistic regression model. We define the dependent variable using the events/trials format and name the explanatory variable FY on the right hand side of the MODEL statement equation: leave/start = FY. Options on the MODEL statement specify that we want PROC GENMOD to fit a logistic regression. That is, we specify that the link function is the logit and that the distribution is binomial. The BY statement applies the logistic regression model to each homogeneous BY-group (GRD YOS XNS) separately.

MODEL PREDICTION:

The logistic regression model fits a line to these data (using maximum likelihood) and gives a predicted rate for each point. The table of data in the appendix includes variable Pred as the model predicted attrition rate. PROC GENMOD will also calculate a Wald confidence interval on this predicted rate. Note that the procedure also calculates a predicted rate even for the current FY, which has no observed (leave / start) data. This final point is the predicted rate which analysts use to forecast near term losses (FY99).

As for the attrition trend, the sign of the coefficient on the explanatory variable, FY, indicates whether the attrition is increasing or decreasing. The default output to the PROC GENMOD for our example BY-group is in the appendix. This BY-group has increasing attrition, as indicated by the positive coefficient, 0.1191.

ATTRITION PLOTS:

The analyst can then use PROC GPLOT to display plots of the actual attrition rates (O_Rate) and the predicted rate (Pred) over time. These plots show the percent of a cohort which left service (dependent-response variable on the vertical y-axis) per FY (independent-explanatory-predictor variable on the horizontal x-axis) over the past fourteen FYs. These useful plots display our fit of predicted rates to the observed attrition rates and the trend on officer attrition. A plot of our example BY-group is included.

ARMY DRAWDOWN:

The Army underwent a drastic draw-down after the end of the Cold War. Fiscal years 1992 through 1995 were primary years during which the Army encouraged officers to leave the service, through both voluntary and involuntary separation programs. The TAPDB does not capture either the forced or voluntary losses. The primary involuntary program was a board which selected officers, who were already retirement eligible (lieutenant colonels and colonels), to leave the service.

As for voluntary programs, the Army offered two primary incentives for officers to leave the service, an early retirement for majors and a separation bonus for captains. In total, a couple thousand majors, who were within a few years of retirement eligibility, volunteered to retire early at a reduced annuity. Likewise, a few thousand Captains volunteered to receive a separation

bonus payment (either a lump sum or annuity - not a retirement).

IDENTIFY OUTLIERS:

These drawdown programs created many attrition rate "outliers." These are data points which are due to special causes and which we want to exclude from our analysis. Even the casual observer will notice some of these obvious outlier points on the included attrition plots. One also can see the dramatic effect of these points on the fitted regression line.

Making a decision whether to include or exclude individual suspect outliers based on regression plots (observed verses predicted rate) would be extremely tedious. An analyst would have to visually examine suspect points on each of the 160+ regression plots and then make a decision whether they should be excluded. Furthermore, he would also have the cumbersome data processing task of identifying and excluding these outlier observations. This would be a labor intensive process, especially when dealing with more than 2300 attrition rate observations, about 14 observations for each of the 160+ BY-groups.

DIAGNOSTIC STATISTICS:

Fortunately, most SAS/STAT regression procedures calculate a statistic called the standardized residual (Pearson) Chi-square (StResChi). An observation having a StResChi larger than three is certainly suspect as an outlier, since the StResChi's standard deviation is unity. The analyst may be able to assert that all points lying more than some number of standard deviations (say 10) off the regression line should be excluded from the analysis. The large number of BY-groups in this analysis leads us to desire an automated procedure to identify outliers on each of the 160+ separate logistic regressions.

PROC GENMOD OUTPUT:

PROC GENMOD is particularly suited to an automated outlier identification analysis because it can create an output data set which includes the key diagnostic statistic, StResChi. The output data set can also include other useful observation statistics: the predicted value and the confidence interval (upper and lower bounds on the predicted value). These data are called "observation statistics" since each observation has a corresponding statistic.

In this sense, PROC GENMOD was a Beta procedure for the output delivery system (ODS) now released in Version 7 of The SAS System. That is, PROC GENMOD (V6.12) has a MAKE statement which permits the user to choose variables (columns) to include in an output data set.

OUTLIER EXCLUSION:

We've seen that the observation diagnostic statistic StResChi enables us to identify outliers in an automated fashion. However, we still have the data processing task of excluding these outliers from subsequent analysis.

Since the observations statistics output data set has the same number and order of observations as the input original data, a simple DATA step with MERGE statement easily combines our original input data set with the output data set (observation statistics).

This post-logistic regression DATA step with MERGE statement combines the input and output data sets into a single combined in-out data set. Since this combined data set has all the original data and a StResChi column, we can subset our original data to include only those observations having an acceptable StResChi. Then, we can then re-run the logistic regression against the combined in-out data set from which outliers have been excluded.

We repeat this process against subsequent subsets of the original data until we have identified and excluded all inappropriate outliers.

REPETITION STRATEGY:

Thus, our strategy for excluding outliers is a three-fold process. First, we run a logistic regression against the original data. Second, we exclude inappropriate outliers from the original data, creating a subset for subsequently analysis. Third, we re-run the logistic regression against this subset (the original data from which outliers have been excluded). We repeat the last two steps of the process until we have removed all inappropriate outliers and have thus come to an acceptable fit of the observed data (less outliers) and a predicted value.

MACRO STRATEGY:

This repetition of code (DATA and PROC steps) leads us to use SAS Macro Language to semi-automate our outlier exclusions. An appropriate a macro strategy

enables us to repeat these DATA and PROC steps efficiently, that is, without repeated copying and pasting of code.

SAS Macro Language is a powerful extension which enables us to use structured programming logic to efficiently repeat PROC and DATA steps. We house our some DATA steps and PROC GENMOD inside a macro. We then call this macro to repeat the logistic regression several times against subsequent subsets.

The structured program proceeds as follows: we define a macro which includes pre-processing DATA steps, PROC GENMOD (the logistic regression), and post-processing DATA steps. The program then submits the code as follows: #1 prepare original data (one time) for analysis; #2 call the macro to pre-process the data, run the logistic regression, and to post-process the output data set; #3 exclude outliers from the most recent output data set; and #4 return to #2 to re-run the logistic regression against the most recent subset from which outliers have been excluded.

Note that step #3 is not automatic. The analyst must examine the output from each run, establish criteria each by which to exclude outliers (based primarily on StResChi), and then insert "hard-code" to exclude those outliers from the data set by sub setting on acceptable StResChi values. This process requires judgment on each of the 160+ BY-groups. Fortunately, we can eliminate most outliers by using the same StResChi criteria for all 160+ BY- groups.

The appendix includes our work-horse macro and the brief program which calls the macro twice.

MACRO PARAMETERS:

We want to save the results of each logistic regression to a different permanent data set. So, we use macro parameters to assign a unique name to each succeeding output data set. Thus, macro parameters enable us to avoid over-writing subsequent output data sets.

1-2 KNOCK-OUT PUNCH:

This combination of BY-group analysis and SAS Macro Language enables a very compact but powerful analysis framework. Recall that each PROC GENMOD has BY statement which analyzes each BY-group separately, 160+ separate logistic regressions. Then, remember that we submit the PROC GENMOD several times: 1st against the original data and then several times against

subsequent data sets from which outliers have been excluded. Thus, we conquer this mountain of data with a small but powerful smart bomb. Or better, we mine this large vein of ore with a small and precise shovel.

OUTLIER CAUTIONS:

A logistic regression is the proper statistical model for this analysis. However, excluding outliers can induce some problems into the analytical framework, especially for logistic regression. Remember that a logistic model is non-linear, the regression line is an S-shape curve starting at zero and going up to unity (or from unity down to zero). A curvilinear regression is much more susceptible to fluctuations than a linear regression model since the logistic regression line can change at two inflection points.

The basic caution is this: excluding a data point as an outlier can drastically affect the curvature of the regression line. In particular, boundary (beginning or ending) observations can potentially exert a significant influence on the fitted line when there is no nearby data to attenuate the curvature. In other words, orphan observations at the ends of the data horizon may exert an inordinate influence on the curve.

In our case, if we exclude drawdown years as outliers, and if the post-drawdown rate is different than the pre-drawdown rate, then the trend on recent data will not be attenuated by the missing drawdown data. In other words, missing data (excluded outliers) just prior to the end of the observed data horizon can give an undue influence to recent data point. This can cause accentuated curvature at the end of the data horizon.

We submit a few solutions to this "outlier-exclusion or "induced-inflection" problem.

FLAT LOGISTIC REGRESSION:

The above logistic (non-linear) complication may lead to the following overly-conservative logistic regression model. One might perform a logistic regression with no dependent variable, i.e., no variable on the right side of the model statement. Essentially, this is a flat horizontal regression line (simple mean) on the predicted rate. This model still permits identification and exclusion of outliers in the usual way but it does not permit conclusions on whether the attrition rate is getting better or worse since we get no coefficient on the FY variable - it's not there! Nevertheless, this may provide the best

prediction for a long-term mean and deserves serious consideration.

LINEAR MODEL:

We could perform linear regression rather than a logistic regression since a linear regression is much not susceptible to induced inflection points. (Furthermore, the problem of non-normally distributed error terms is a theoretical prohibition, without much deleterious significance from a practical viewpoint.) However, the linear model provides much less power for identifying outliers than the logistic regression. Hence, we prefer the logistic regression since it permits us to exploit power of having over a million observations in our crucial effort to identify and exclude outliers.

COMPRESS X-AXIS:

Our solution to the "induced-inflection" problem is to compress missing values out of the analysis data set altogether. This prevents the model from inducing an inflection point at a missing data point, since missing data are not in the analysis data set at all. From another viewpoint, this technique prevents boundary observations from over influencing the curve, since we eliminate orphans by keeping all points adjacent. We accomplish this technique by re-scaling the horizontal x-axis for each BY-group so that there are no missing values on the horizontal axis.

This solution is sufficiently conservative in that the model results are not overly sensitive to our outlier exclusion decisions. Thus, our results are stable, not fluctuating wildly from run to run, depending on our outlier criteria. Furthermore, this model still gives us great power to detect outliers and enables us to assert the existence of attrition trends. Finally, we are comfortable considering the scale of the x-axis as primarily ordinal-numeric scaled, not interval-numeric scaled.

FORECAST LOSSES:

After settling on an acceptable model fit, we apply predicted attrition rates to forecast officer losses for the coming FY. We multiply each BY-group predicted attrition rate against each FY99 starting population to calculate a forecast on the number of officers we expect to separate per BY-GROUP and round the result to integer losses. Then we simply add losses by grade to get aggregate predicted losses for that grade.

CONCLUSION:

This analytical effort is significant. We started with 14 years of legacy data. We summarized it to a single analysis data set having only 2300+ event/trials observations. Then, we ran a separate logistic regression against each of the 160+ homogeneous BY-groups in that analysis data set. Finally, we used SAS macro language to automate our identification and exclusion of outliers from the analysis data set in order to achieve an acceptable model fit of the data. Thus, Army officer personnel analysts have a statistically robust method for estimating mean attrition rates in order to predicting officer separations. The SAS System has the all the data management, statistical power, and convenient features for the Army to draw substantial and important conclusions on officer attrition.

REFERENCES:

Stokes, Maura E., Davis, Charles S., Koch, Gary G., *Categorical Data Analysis Using the SAS® System*, Cary, NC: SAS Institute Inc., 1995. 499pp.

Agresti, Alan, *An Introduction to Categorical Data Analysis*, New York: John Wiley & Sons, 1996. 290pp.

SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

AUTHOR CONTACT:

The author accepted reassignment to the Army Logistic Management College (Systems Engineering Department) at Fort Lee Virginia, just south of Richmond and adjacent to the Petersburg National Battlefield Park. He is often jogging in the park (happy to be outside the pentagon) but you can leave email: McAllasterD @ Lee.Army.Mil

```

* _
SAS code by LTC Doug McAllaster for SESUG'99 Statistics section paper
abstract: run logistic regression on one BY-group
input   : logitlib.save1 and library.formats.catalog
process : subset to one BY-group for SESUG'99 paper
          pre-process input data set for regression
          run simple logistic regression
          post-process regression output data set
          merge original input data with logistic output data set (obstats)
          print results of regression
          plot observed rate (o_rate) and predicted rate on same plot
          exclude outliers from output data set based on StResChi,
          re-run logistic on subset which excludes outliers
output  : logit&sfx
-;

*===== define macro to: =====
1. pre -process regression input data for
2. fit logistic regression using proc genmod
3. post-process regression output data
4. print & plot regression results
5. export      regression results to dbf
=====;

%macro genmod(sfx);

*-1st: pre-process input data-;

*-sort data into proper chronological order-;
proc sort data=logitin;
  by   grd yos xns yearid; run;

*-re-scale x-axis:
  so that year12 gives a simple integer x-axis w/unity between each point
  both   yearid and year12 have integers in range from 1-15
  but    yearid is fixed and is formatted to print '85'-'99' for its FY
  whereas year12 varies from run to run depending omitted outliers-;
data logitin1;
  set logitin;
  year12=_n_; run;

*-2nd, fit logistic regression model to data-;
proc genmod data=logitin1;
  by   grd yos xns;
  model leave / start = year12 / link=logit dist=binomial obstats residuals;
  make 'obstats' out=logitout (keep=pred streschi) noprint; run;

*-3rd, post-process output data:
  merge input data (logitin1) with output obstats data (logitout)
  both input output data sets have same number and order of observations-;
data logit&sfx;
  retain runid &sfx;
  retain grd yos xns yearid year12 start leave obsvd pred streschi;
  merge logitin1 logitout;
  *-calculate actual observed attrition rate for plot-;
  if start then obsvd=leave/start; run;

*-4th, print and plot results of logistic regression-;
proc print data=logit&sfx uniform; run;

```

```

*-rotate data for proc plot-;
proc transpose
  data=logit&sfx (keep=grd yos xns year12 obsvd pred)
  out=logitlib.plot&sfx (keep=grd yos xns year12 col1 _name_);
  by  grd yos xns year12;
  var  obsvd pred; run;

proc datasets library=logitlib nolist;
  modify plot&sfx;
  label col1='attrition rate'; run;

*-plot observed (obsvd) and predicted values (pred)-;
proc plot data=logitlib.plot&sfx;
  format col1 percent6.2;
  plot col1 * year12 = _name_ /
    haxis=0 to 15
    vaxis=0 to .3 by .05; run;

*-keep dbf file so excel can print table with nice grid lines-;
proc dbf data=logit&sfx db3=logit&sfx;
  format obsvd pred streschi 8.5; run;

%mend;

*===start program===;

* libname logitlib 'c:\sasdata\loss\logit';
  libname logitlib 'd:\sasdata\loss\logit';
* libname library 'c:\sasdata\loss\logit';
  libname library 'd:\sasdata\loss\logit';

*-subset to a single BY-group for SESUG'99 paper proceedings-;
data logitin;
  keep  grd yos xns yearid start leave;
  set  logitlib.save1;
  *-keep only this single BY-group-;
  if  grd=3 and xns=0 and yos=9;
  *-for FY99, yearid is missing,
    so assign value to it-;
  if  not yearid then yearid=15; run;

*-run model against all original data-;
%genmod(1)

*-exclude outliers based on StResChi-;
data logitin;
  keep  grd yos xns yearid start leave;
  set  logit1;
  if  streschi > 10 then delete; run;

*-run model against a subset of original data,
  which excludes outliers-;
%genmod(2)

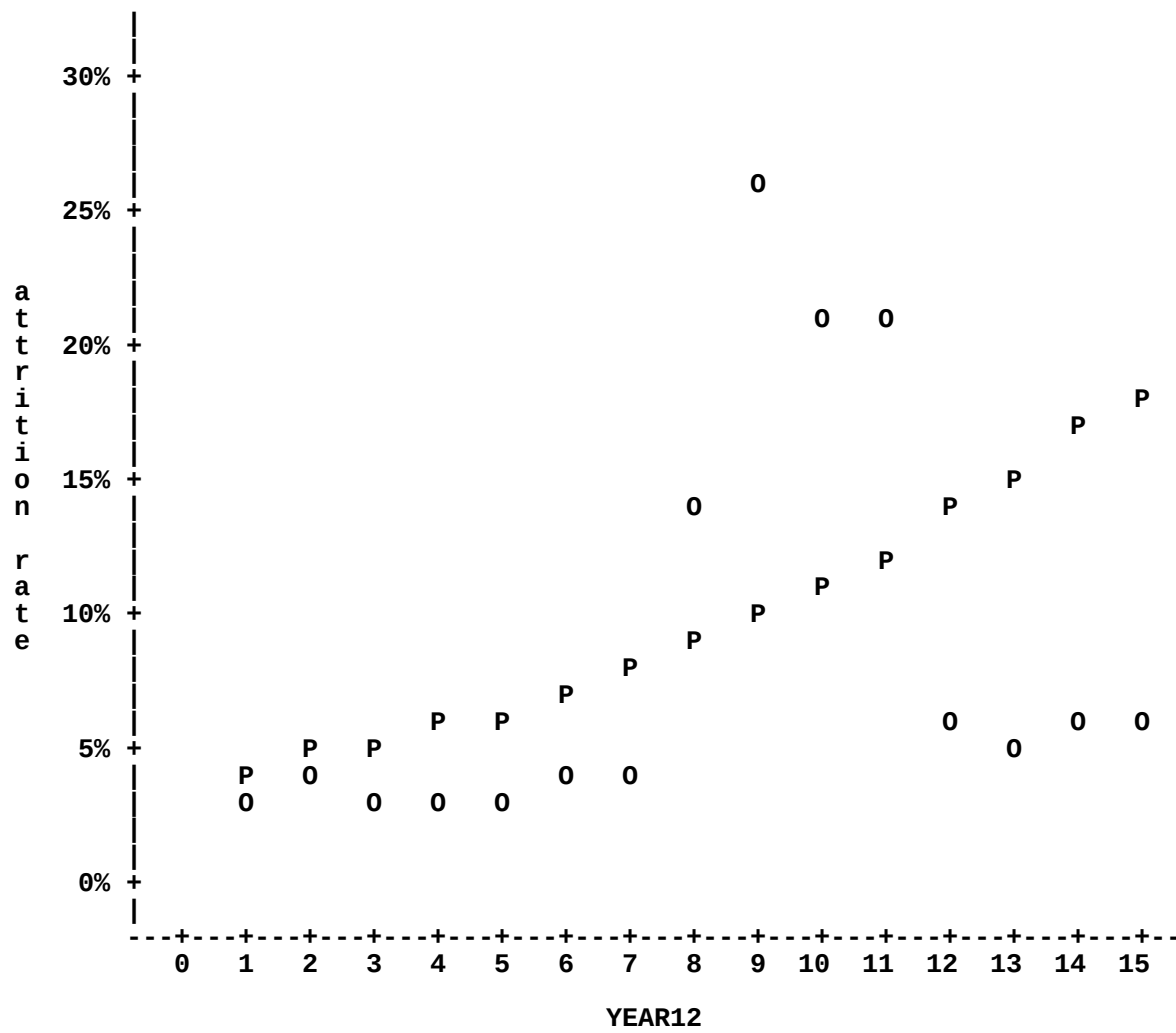
```

Model Information

Criteria For Assessing Goodness Of Fit

Analysis Of Parameter Estimates

OBS	RUN ID	GRID	YOS	XNS	YEAR	YEAR	START	LEAVE	O	PRE	STRESSCH
					I	1			2		
1	1	CPT	9	0	85	1	2782	85	0.03055	0.04062	-2.9508
2	1	CPT	9	0	86	2	2957	112	0.03788	0.04552	-2.1781
3	1	CPT	9	0	87	3	3293	99	0.03006	0.05099	-5.9492
4	1	CPT	9	0	88	4	3362	111	0.03302	0.05707	-6.4817
5	1	CPT	9	0	89	5	3252	101	0.03106	0.06383	-8.1283
6	1	CPT	9	0	90	6	3288	132	0.04015	0.07133	-7.3229
7	1	CPT	9	0	91	7	3097	139	0.04488	0.07963	-7.4617
8	1	CPT	9	0	92	8	3147	450	0.14299	0.08881	11.1476
9	1	CPT	9	0	93	9	3086	803	0.26021	0.09894	31.3953
10	1	CPT	9	0	94	10	2800	589	0.21036	0.11008	17.8403
11	1	CPT	9	0	95	11	2686	552	0.20551	0.12230	14.0749
12	1	CPT	9	0	96	12	2339	145	0.06199	0.13567	-11.3113
13	1	CPT	9	0	97	13	2262	114	0.05040	0.15026	-14.9582
14	1	CPT	9	0	98	14	2439	148	0.06068	0.16611	-16.9359
15	1	CPT	9	0	99	15	.	.	0.06068	0.18328	.



--