Western Kentucky University

From the SelectedWorks of Matt Bogard

January 2012

Analysis of GMO vs Non-GMO Corn Hybrid Persistence Using Simulated Time Dependent Covariates in SAS

Contact Author

Start Your Own SelectedWorks

Notify Me of New Work

Available at: http://works.bepress.com/matt_bogard/19

Analysis of GMO vs Non-GMO Corn Hybrid Persistence Using Simulated Time Dependent Covariates in SAS

By Matt Bogard

Abstract

Fitting survival models utilizing time dependent covariates with the PHREG procedure in SAS involves the utilization of programming statements within the procedure that involve array processing. Based loosely on the descriptive statistics reported by Ma and Shi, I use the SAS data step loop to simulate a data set consisting of 10,000 corn hybrids including variables for genetic modification, vertical integration, and the number of close substitutes for each hybrid for each time period. This paper demonstrates the use of PROC PHREG to fit a Cox Proportional Hazards model utilizing the simulated time dependent covariates. In addition, I demonstrate the use of the %PUT statement to verify that values of the time dependent covariates are correctly utilized for each period.

In their paper, *GM vs. Non-GM: A Survival Analysis of Hybrid Seed Corn in the US*, Xingliang Ma and Guanming Shi report results from a survival analysis of corn hybrids. Their data set consisted of a stratified sample of U.S. corn farmer seed purchases consisting of 10,245 different hybrids purchased between 2000 and 2007. They considered an event 'exit' to have occurred when a particular hybrid's seed is not reported as being purchased and disappears from the sample. From their abstract they report that market structure variables have impacts on hybrid survival (survival being how long a hybrid stays in the sample, i.e. how many years do farmers choose to purchase that particular hybrid). Particularly they report in the paper that if a hybrid is supplied by a vertically integrated firm (a seed company that is also in the biotech business) that hazard (loosely speaking, the probability that a hybrid will fail or exit the market) decreases. In addition they find that an increase in the number of substitutes for a hybrid significantly lowers hazard. They attribute this to positive information spillovers among farmers. Finally they find that in general that the introduction of a GM trait reduces the hazard of a hybrid. In their analysis they make distinctions between various biotech traits and consider the impact of stacked traits.

They do not directly specify time dependent covariates in their model, but they do interact variables with the logarithm of survival time to capture time-varying effects.

In this post, based loosly on the descriptive statistics reported by Ma and Shi, as well as other details provided in the discussion, I use the SAS data step loop to simulate a data set consisting of 10,000 corn hybrids, including variables for GM traits, degree of vertical integration, and the number of substitutes for each hybrid. In the simulation, I produce a time dependent covariate for the number of subsitutes, indicating the number of subsititutes for each hybrid for each year that it is found in the 'hypothetical' sample. This is not a reflection or critique in any way of Ma and Shi, but an effort to demonstate how to use SAS PROC PHREG to fit a model utilizing hypothetical continuous time dependent covariates. This is particularly interesting because fitting these models requires utilization of programming statements within the PHREG procedure that involve array processing. To say the least, its not as simple and straight forward as doing a basic logistic regression . Allison(1995) provides a really thorough reference for fitting these models in SAS and actually provides code for referencing time-dependent dummy variables.

After fitting the model, I take a random sample from my simulated data and demonstrate using the **%PUT** statement that the correct values of the number of substitutes is correctly referenced for each period 't'. My simualtion is for 5 years of data. Summary statistics and model results are posted below:

Summary Statistics					
Hybrid Types	**N**	**Mean Fail**	**Mean # Substitutes (YR 1)**	**Mean Vertical Integration**	**Observed Mean Survival Time (yrs)**
GMO	7000	0.51	1135	0.09	1.67
Non-GMO	3000	0.62	1051	0.09	1.48
Overall	10000	0.54	1110	0.09	1.61

Model Specification

In survival analysis, what is particularly of interest is modeling the hazard of an event. The hazard function is a conditional density function giving the instantaneous risk that an event will occur at time't', where t ϵ T, a random variable *time to event*.

$h(t) = \lim_{\Delta t \to 0} \left(\frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} \right)$

$= f(t)/S(t)$ where $f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t)}{\Delta t}$ (probability density) and

$S(t) = \Pr(T>t) = 1 - F(t) = \int_t^{\infty} f(x)dx = e^{-\lambda t}$ (survival function)

If we let $h(t) = \lambda$ or λ_0 or $\ln h(t) = \ln(\lambda_0) = \mu$ be the 'baseline' hazard, we can easily extend the model to include covariates as such:

$h(t) = \lambda_0(t) \exp(\boldsymbol{\beta X})$ or $\ln h(t) = \mu + \boldsymbol{\beta X}$ where X is a vector of covariates.

The Cox-Proportional Hazards Model assumes that the proportional hazard for an individual i vs. i' can be written as:

$h_i(t)/ h_{i'}(t) = \lambda_0(t) \exp(\boldsymbol{\beta X_i})/ \lambda_0(t) \exp(\boldsymbol{\beta X_{i'}})$

with the baseline hazard $\lambda_0(t)$ canceling out we have:

$\exp(\beta (\mathbf{X_i} - \mathbf{X_{i'}}))$ or $e^{\zeta i - \zeta i'}$

The key point of the model is that because the baseline hazard term drops out of the equation, we don't have to explicitly specify its functional form, allowing for a more flexible and robust estimation. The model is estimated via partial likelihood, maximizing:

$L_p(\beta) = \prod_1^n \frac{\exp(x_i'\beta)}{\sum_{i' \in R} \exp(x_{i'}'\beta)}$

(for more details see Survival Analysis 2012 and Fox 2002 and 2006).

This estimation can be implemented in SAS using the PHREG procedure as follows:

```
PROC PHREG DATA=TEMP3_HYBRIDS;
   MODEL YEAR*FAIL(0)=GM VERT NSUB_YR;
     ARRAY NSUB{*} NSUB_1-NSUB_5;
     NSUB_YR = NSUB[YEAR];
  HAZARDRATIO NSUB_YR / UNITS = 100;
RUN;
QUIT;
```

The array specification ensures that for each year, the correct value for the number of substitute hybrids is referenced in constructing the partial likelihood function used in the estimation.

Results

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > Chi Sq**	**Hazard Ratio**
GM	1	-0.22803	0.02874	62.9482	<.0001	0.796
VERT	1	-0.32644	0.05421	36.257	<.0001	0.721
NSUB_YR	1	-0.0008187	0.0000215	1448.24	<.0001	0.999

Hazard Ratios for NSUB_YR			
Description	**Point Estimate**	**95% Wald Confidence Limits**	
NSUB_YR Unit=100	0.921	0.918	0.925

It can be seen that all of the estimated co-efficients are negative and highly significant. This has no actual real world implication, as the simulated data was specifically designed to produce results similar to those found in Ma and Shi. If we were to interpret these results, the interpretation would be similar, at least in a mechanical sense to interpreting odds ratios in logistic regression. For instance, the Hazard Ratio for GM is simply the exponentiated co-efficient for GM , exp(-.22803) = .796. This implies that being a GMO hybrid vs. non-GMO changes hazard by 100*(.796-1) = -20.4%. i.e. the hazard of a hybrid failing or its discontinued use by corn growers is 20.4% less for GMO vs. non-GMO hybrids. The hazard of a hybrid produced by a vertically integrated seed company is reduced by 100*(.721-1) = -27.9%. Because the number of substitutes for each hybrid was on average over 1000 for each year, a 1 unit change in the number of competing substitutes isn't likely to have much impact on hazard. The default hazard ratio reported (.999) is based on a 1 unit change and doesn't make sense. Similar to the 'UNITS' statement used in PROC LOGISTIC, the command `HAZARDRATIO NSUB_YR / UNITS = 100;` calculates the estimated hazard for every 100 unit change in the number of substitutions. These results are output separately from the other estimates, and indicate that for every 100 additional substitutes available for a particular hybrid, the hazard is reduced by 100*(.921-1) = -7.9 or 7.9%. As indicated in Ma and Shi, the fact that a decrease in hazard is observed as the number of close substitutes increases may be an indication of spillover information learning or adoption effects assocatiated with particular hybrid technologies. Again, this is simply a rough simulation of data, designed specifically to give these results. I'm not making any claims about real world outcomes based on this data. However, my estimates for GM were very similar to many of estimates for the particular traits reported in Ma and Shi. My estimate for vertical integration is quite a bit smaller, a little less than half in magnitude, and similarly to Ma and Shi, the coefficeint on the number of substitutes was very small. So my simulation probably doesn't fully

capture all of the information provided in their real world sample, but may at least be useful for the demonstration at hand.

To demonstrate that the correct values for the number of substitutes for each hybrid were used in the model for each time 't', the **%PUT** statement can be utilized. To demonstrate this, a random sample stratified by YEAR was generated from the larger simulated data, depicted below.

Obs	YEAR	VERT	GM	NSUB_1	NSUB_2	NSUB_3	NSUB_4	NSUB_5	FAIL	ID
1	1	0	0	1008	.	.	.	.	1	7521
2	1	1	1	326	.	.	.	.	1	1091
3	2	0	1	840	834	.	.	.	1	2371
4	2	1	0	1150	583	.	.	.	1	8545
5	3	0	1	1841	323	1507	.	.	1	3160
6	3	1	1	2096	2143	1997	.	.	1	3115
7	4	0	0	1609	709	2250	1223	.	1	8707
8	4	1	1	2997	2540	580	498	.	1	3362
9	5	0	1	264	2722	2228	2020	2976	0	4016
10	5	1	1	351	1389	2784	2436	551	0	4278

As shown above, for each year that a hybrid 'persists' there is a specific number of competitive substitutes on the market. When the hybrid 'fails' or drops out of the sample, I no longer simulate substitute values. YEAR is the number of years that a hybrid persists on the market, it is in essence 'survival time' while FAIL indicates that a hybrid was no longer being purchased. It is the 'event' or 'censoring' indicator variable in the model. Hybrids that persist 5 years without failing are 'censored' observations. The model can be fit on this sample data set using the code below.

```
DATA TEMP_HYBRID_DEMO;
INPUT YEAR VERT GM  NSUB_1-NSUB_5 FAIL ID;
CARDS;
1     0     0     1008  .     .     .     .     1     7521
1     1     1     326   .     .     .     .     1     1091
2     0     1     840   834   .     .     .     1     2371
2     1     0     1150  583   .     .     .     1     8545
3     0     1     1841  323   1507  .     .     1     3160
3     1     1     2096  2143  1997  .     .     1     3115
4     0     0     1609  709   2250  1223  .     1     8707
4     1     1     2997  2540  580   498   .     1     3362
5     0     1     264   2722  2228  2020  2976  0     4016
5     1     1     351   1389  2784  2436  551   0     4278
;
RUN;
```

```
* FIT MODEL FROM SAMPLE DATA;
PROC PHREG DATA=TEMP_HYBRID_DEMO;
   MODEL YEAR*FAIL(0)=GM VERT NSUB_YR;
     ARRAY NSUB{*} NSUB_1-NSUB_5;
     NSUB_YR = NSUB[YEAR];
     FILE LOG;
     PUT @1 YEAR= @13 ID= @25 FAIL = @35 NSUB_YR= ;
RUN;
QUIT;
```

Because of the nature of the small stratified sample, none of the results were significant. The point of interest here is to demonstrate that PHREG correctly utilizes time dependent covariates. The commands below are key:

```
     FILE LOG;
     PUT @1 YEAR= @13 ID= @25 FAIL = @35 NSUB_YR= ;
```

These commands tell the procedure for each period 't' to 'put' to the SAS log the value of the time dependent covariate referenced by the array statement. The log file is depicted below:

LOG FILE:

```
YEAR=5    ID=4016   FAIL=0   NSUB_YR=2976
YEAR=5    ID=4278   FAIL=0   NSUB_YR=551
YEAR=4    ID=8707   FAIL=1   NSUB_YR=1223
YEAR=4    ID=3362   FAIL=1   NSUB_YR=498
YEAR=3    ID=3160   FAIL=1   NSUB_YR=1507
YEAR=3    ID=3115   FAIL=1   NSUB_YR=1997
YEAR=2    ID=2371   FAIL=1   NSUB_YR=834
YEAR=2    ID=8545   FAIL=1   NSUB_YR=583
YEAR=1    ID=7521   FAIL=1   NSUB_YR=1008
YEAR=1    ID=1091   FAIL=1   NSUB_YR=326
YEAR=4    ID=4016   FAIL=0   NSUB_YR=2020
YEAR=4    ID=4278   FAIL=0   NSUB_YR=2436
YEAR=4    ID=8707   FAIL=1   NSUB_YR=1223
YEAR=4    ID=3362   FAIL=1   NSUB_YR=498
YEAR=3    ID=4016   FAIL=0   NSUB_YR=2228
YEAR=3    ID=4278   FAIL=0   NSUB_YR=2784
YEAR=3    ID=8707   FAIL=1   NSUB_YR=2250
YEAR=3    ID=3362   FAIL=1   NSUB_YR=580
YEAR=3    ID=3160   FAIL=1   NSUB_YR=1507
YEAR=3    ID=3115   FAIL=1   NSUB_YR=1997
YEAR=2    ID=4016   FAIL=0   NSUB_YR=2722
YEAR=2    ID=4278   FAIL=0   NSUB_YR=1389
YEAR=2    ID=8707   FAIL=1   NSUB_YR=709
YEAR=2    ID=3362   FAIL=1   NSUB_YR=2540
YEAR=2    ID=3160   FAIL=1   NSUB_YR=323
```

```
YEAR=2    ID=3115    FAIL=1    NSUB_YR=2143
YEAR=2    ID=2371    FAIL=1    NSUB_YR=834
YEAR=2    ID=8545    FAIL=1    NSUB_YR=583
YEAR=1    ID=4016    FAIL=0    NSUB_YR=264
YEAR=1    ID=4278    FAIL=0    NSUB_YR=351
YEAR=1    ID=8707    FAIL=1    NSUB_YR=1609
YEAR=1    ID=3362    FAIL=1    NSUB_YR=2997
YEAR=1    ID=3160    FAIL=1    NSUB_YR=1841
YEAR=1    ID=3115    FAIL=1    NSUB_YR=2096
YEAR=1    ID=2371    FAIL=1    NSUB_YR=840
YEAR=1    ID=8545    FAIL=1    NSUB_YR=1150
YEAR=1    ID=7521    FAIL=1    NSUB_YR=1008
YEAR=1    ID=1091    FAIL=1    NSUB_YR=326
```

This file is to be read from the bottom up. If we start from the top, we see that the entire sequence of observations in the data set is simply printed from t = 5 to t=1. Reading from the bottom up, and comparing to the actual values in the data set we can see that for each time t, the array values NSUB_YR match that period's value for the time dependent covariate for each observation. For example, in YEAR = 1, the correct value for the number of substitute hybrids for ID =1091 is given by the variable in the data set NSUB_1 = 326. The log indicates that the variable NSUB_YR (which is the variable specified in the MODEL statement of PHREG) is assigned the value specified by the array NSUB[YEAR]for YEAR = 1, which turns out to be 326. (as specified in the log NSUB_YR = 326). For each value of YEAR, only observations in the 'risk set' (individuals that have not yet experienced the event 'FAIL') are considered, as they are the only observations that contribute to the partial likelihood in the estimation. So you will see as we iterate from YEAR to YEAR (reading from the bottom up) only hybrid ID's that have persisted to that point are referenced along with the correct value of the time dependent covariate. You will also notice that in the sample data set where YEAR = 5, in both cases the value of FAIL = 0. These observations are considered 'censored' and do not contribute to the partial likelihood. (again see details in Survival Analysis and Fox 2002 and 2006). As a result, the log file, read from the bottom up stops at YEAR = 4.

The full code for the simulation is given following the references below.

References:

Survival Analysis Using SAS: A Practical Guide. Paul D. Allison. 1995. The SAS Institute.

Survival Analysis. Matt Bogard. Econometric Sense.
http://econometricsense.blogspot.com/2012/04/survival-analysis.html

The Calculation and Interpretation of Odds Ratios. Matt Bogard. Econometric Sense.
http://econometricsense.blogspot.com/2011/03/calculation-and-interpretation-of-odds.html

Cox Proportional-Hazards Regression for Survival Data:
Appendix to An R and S-PLUS Companion to Applied Regression. John Fox Februrary 2002

Introduction to Survival Analysis. Sociology 761 Lecture Notes. John Fox. 2006.

GM vs. Non-GM: A Survival Analysis of Hybrid Seed Corn in the US
University of Wisconsin-Madison
Department of Agricultural & Applied Economics
Staff Paper No. 553 November 2010
By Xingliang Ma and Guanming Shi

SAS SIMULATION AND PHREG CODE

```
 *----------------------------------------------------
 | PROGRAM NAME: HYBRIDS
 | DATE: 4/2/12
 | CREATED BY:MATT BOGARD
 | PROJECT FILE: P:\SAS CODE EXAMPLES
 |(copy)\SURVIVAL ANALYSIS
 *-------------------------------------------------------
 | PURPOSE: SIMULATION AND ANALYSIS OF TIME DEPENDENT
 |COVARIATES  IN SAS PROC PHREG
 | REFERENCE: GM vs. Non-GM: A Survival Analysis of
 |Hybrid Seed Corn in the US
 | University of Wisconsin-Madison
 | Department of Agricultural & Applied Economics
 | Staff Paper No. 553 November 2010
 | By Xingliang Ma and Guanming Shi
 |
 | Survival Analysis Using the SAS System: A Practical
 | Guide by Paul D. Allison
 |
 | SAS Publications order # 55233
 | ISBN 1-55544-279-X
 | Copyright 1995 by SAS Institute Inc., Cary, NC, USA
 |
 |
 *-------------------------------------------------------;




 *-----------------------------------------
 |
 |
 | PART 1: SIMULATE DATA
 |
 |
 *-----------------------------------------;
```

```
 *-----------------------------------
 |  SIMULATE DATA FOR GMO HYBRIDS
 *--------------------------------;

%LET NOBS = 7000;
DATA TEMP1_GMO;
    MAX = 40;
      GM = 1;
      CALL STREAMINIT(123);
      C = 0; D = 3000;

            DO J = 1 TO &NOBS;
                  U2 = RAND("UNIFORM");
                  K = CEIL(MAX*U2);
                  ARRAY NSUB_ (5);
                  DO I = 1 TO 5;
                        U = RAND("UNIFORM");
                        NSUB_(I) = ROUND(C + (D-C)*U);
                                    IF K LE 13 THEN DO;
                                        E = 0; F = 1200;
                              U = RAND("UNIFORM");
                              NSUB_(I)=ROUND(E +(F-E)*U);
                                          YEAR = 1;
                                          FAIL = 1;
                                          DO L = 2 TO 5;
                                                NSUB_(L) = .;
                                          END;

                                    END;
                                    IF K IN (14,15,16,17,18) THEN DO;
                                       E = 300; F = 1500;
                               U = RAND("UNIFORM");
                               NSUB_(I)=ROUND(E+(F-E)*U);
                                        YEAR = 2;
                                          FAIL =1;
                                          DO L = 3 TO 5;
                                                NSUB_(L) = .;
                                          END;

                                    END;
                                    IF K = 19 THEN DO;
                                          YEAR = 3;
                                          FAIL =1;
                                          DO L = 4 TO 5;
                                                NSUB_(L) = .;
                                          END;
                                     END;
                                    IF K  = 20 THEN DO;
                                          YEAR = 4;
                                          FAIL =1;
                                          DO L = 5;
```

```
                                                    NSUB_(L) = .;
                                              END;
                                        END;
                                        IF K = 21 THEN DO;
                                        YEAR = 5;
                                              FAIL = 1;
                                        END;
                                        IF K > 21 THEN DO;
                                        YEAR = 5;
                                        FAIL = 0;
                                        END;
                        END;

                    OUTPUT;

              END;
RUN;

* SIMULATE VERTICAL INTEGRATION DATA;
DATA TEMP1_GMO2;
      SET TEMP1_GMO;
      VERT = 0;
      IF YEAR = 1 THEN VERT = RAND("BINOMIAL",.06,1);
      IF YEAR = 2 THEN VERT = RAND("BINOMIAL", .07,1);
      IF  YEAR >2 THEN VERT = RAND("BINOMIAL",.12,1);
RUN;

 *-------------------------------
 |  SUMMARY STATISTICS
 *-------------------------------;

PROC SORT DATA = TEMP1_GMO2;
BY YEAR;
RUN;

PROC MEANS DATA = TEMP1_GMO2 N MEAN ;
VAR FAIL NSUB_1 VERT GM;
TITLE "BY YR";
BY YEAR;
RUN;

FOOTNOTE "Simulated Data";
PROC MEANS DATA = TEMP1_GMO2 N MEAN ;
VAR FAIL YEAR NSUB_1 VERT GM;
TITLE "GMO Stats";
RUN;

PROC MEANS DATA = TEMP1_GMO2 N MEAN ;
VAR FAIL YEAR NSUB_1 VERT GM;
WHERE FAIL = 1;
TITLE "GMO Survival Stats";
RUN;


 *------------------------------------
 |  SIMULATE DATA FOR NON-GMO HYBRIDS
 *------------------------------------;
```

```
%LET NOBS = 3000;
DATA TEMP2_NGMO;
    MAX = 40;
      GM = 0;
      CALL STREAMINIT(123);
      C = 0; D = 3000;

            DO J = 1 TO &NOBS;
                  U2 = RAND("UNIFORM");
                  K = CEIL(MAX*U2);
            ARRAY NSUB_ (5);
                  DO I = 1 TO 5;
                        U = RAND("UNIFORM");
                        NSUB_(I) = ROUND(C + (D-C)*U);
                                    IF K LE 19 THEN DO;
                                        E = 0; F = 1200;
                                        U = RAND("UNIFORM");
                              NSUB_(I) = ROUND(E + (F-E)*U);
                                         YEAR = 1;
                                         FAIL = 1;
                                          DO L = 2 TO 5;
                                               NSUB_(L) =.;
                                          END;
                                    IF U2 LE .02 THEN VERT =
                                    1;ELSE VERT = 0;
                                    END;


                                    IF K IN (20,21,22) THEN DO;
                                          E = 300; F = 1500;
                                        U = RAND("UNIFORM");
                              NSUB_(I) = ROUND(E + (F-E)*U);
                                         YEAR = 2;
                                         FAIL =1;
                                          DO L = 3 TO 5;
                                               NSUB_(L) =.;
                                          END;
                                   IF U2 LE .05 THEN VERT =   1;ELSE VERT =0;
                                    END;
                                    IF K = 23 THEN DO;
                                          YEAR = 3;
                                          FAIL =1;
                                          DO L = 4 TO 5;
                                               NSUB_(L) =.;
                                          END;
                                    END;
                                    IF K  = 24 THEN DO;
                                          YEAR = 4;
                                          FAIL =1;
                                         DO L = 5;
                                               NSUB_(L) =.;
                                          END;
                                    END;
                                     IF K = 25 THEN DO;
                                    YEAR = 5;
                                          FAIL = 1;
```

```
                                        END;

                                       IF K > 25 THEN DO;
                                       YEAR = 5;
                                        FAIL = 0;
                                       END;
                        END;
                  OUTPUT;
            END;

RUN;


 *-----------------------------------------
 |  SUMMARY STATISTICS
 *-----------------------------------------;

DATA TEMP2_NGMO2;
      SET TEMP2_NGMO;
      VERT = 0;
      IF YEAR = 1 THEN VERT = RAND("BINOMIAL",.06,1);
      IF YEAR = 2 THEN VERT = RAND("BINOMIAL", .07,1);
      IF  YEAR >2 THEN VERT = RAND("BINOMIAL",.12,1);
RUN;

PROC SORT DATA = TEMP2_NGMO2;
BY YEAR;
RUN;

PROC MEANS DATA = TEMP2_NGMO2 N MEAN ;
VAR FAIL NSUB_1 VERT GM;
TITLE "Non-GMO Stats";
BY YEAR;
RUN;

PROC MEANS DATA = TEMP2_NGMO2  N MEAN ;
VAR FAIL NSUB_1 VERT GM;
TITLE "Non-GMO Stats";
RUN;

PROC MEANS DATA = TEMP2_NGMO N MEAN ;
VAR  YEAR NSUB_1 VERT GM;
TITLE "Non-GMO Survival Stats";
WHERE FAIL = 1;
RUN;

 *-----------------------------
 |  COMBINE GMO AND NON GMO DATA
 *-------------------------------;

DATA TEMP3_HYBRIDS;
 SET TEMP1_GMO2 TEMP2_NGMO2;
 ID + 1; /* CREATE OBS ID */
 DROP C D J K MAX U U2 I E F L ;
 RUN;

* SUMMARY STATS;
```

```
PROC MEANS DATA = TEMP3_HYBRIDS N MEAN ;
VAR FAIL NSUB_1 VERT GM;
TITLE "Overall GMO/NonGMO Summary Statistics";
RUN;

PROC MEANS DATA = TEMP3_HYBRIDS N MEAN ;
VAR FAIL YEAR NSUB_1 VERT GM;
TITLE "Overall GMO/NonGMO Survival Statistics";
WHERE FAIL = 1;
RUN;

 *-------------------------------
 |
 |
 | PART 2: ANALYSIS
 |
 |
 *-------------------------------;

 *-------------------------------
 |  MODEL 1: TIME DEPENDENT COVARIATES
 *-------------------------------------;

PROC PHREG DATA=TEMP3_HYBRIDS;
   MODEL YEAR*FAIL(0)=GM VERT NSUB_YR;
     ARRAY NSUB{*} NSUB_1-NSUB_5;
     NSUB_YR = NSUB[YEAR];
  HAZARDRATIO NSUB_YR / UNITS = 100;
RUN;
QUIT;


 *-----------------------------------------
 |  MODEL 2: NON-TIME DEPENDENT COVARIATES
 *-----------------------------------------

PROC PHREG DATA=TEMP3_HYBRIDS;
   MODEL YEAR*FAIL(0)=GM VERT NSUB_1;
RUN;
QUIT;

*--------------------------------------
|  MODEL 3: USE OF PUT STATEMENT TO VALIDATE
|REFERENCING OF TIME DEPENDENT COVARIATES
*--------------------------------------;


* GENERATE RANDOM SAMPLE STRATIFIED ON EVENT GRADUATED;
PROC SORT DATA = TEMP3_HYBRIDS OUT = TEMP4_SRT; /* PRE-SORT BY STRATA */
BY YEAR VERT ;
RUN;

PROC SURVEYSELECT DATA=TEMP4_SRT
      OUT=TEMP5_SRS
      METHOD=SRS
    SEED = 123
```

```
      N=10;
      STRATA YEAR VERT / ALLOC=PROP;
RUN;

PROC PRINT DATA = TEMP5_SRS;RUN;

* FIT MODEL FOR DEMONSTRATION OF PUT STATEMENTS;
PROC PHREG DATA=TEMP5_SRS;
   MODEL YEAR*FAIL(0)=GM VERT NSUB_YR;
     ARRAY NSUB{*} NSUB_1-NSUB_5;
     NSUB_YR = NSUB[YEAR];
     FILE LOG;
    PUT @1 YEAR= @13 ID= @25 FAIL = @35 NSUB_YR= ;
RUN;
QUIT;
```