The PSPMCM SAS® macro

A simulated example

1 – Description of the dataset

1.1 – General description

The dataset SIMULTEX is constituted with 500 individuals (one record per individual), with the following variables:

- IDENT: identification number;
- UNCURED: the TRUE (simulated) uncured status (0 for cured, 1 for uncured). This variable is unobserved in real datasets
- SURVTIME: failure or censoring time;.
- STATUS: censoring indicator (censored=0, uncensored=1);
- VARX: name of the first explanatory variable. Values are 0 or 1;
- VARY: name of the second explanatory variable, with values 0, 1 or 2.
- VARY1 and VARY2 are the dummy variables created for VARY such as:
  If VARY=0 then VARY1=0 and VARY2=0,
  If VARY=1 then VARY1=1 and VARY2=0,
  If VARY=2 then VARY1=0 and VARY2=1.

Tables 1gives the repartition of the different explanatory variables:

Table of varX par varY					
Frequency Percent Row percent Col percent	varX	varY 0	 1	 2	Total
	0	91 18.20 30.33 65.94	86 17.20 28.67 58.11	123 24.60 41.00 57.48	300 60.00
	1	47 9.40 23.50 34.06	62 12.40 31.00 41.89	91 18.20 45.50 42.52	200 40.00
	Total	138 27.60	148 29.60	214 42.80	500 100.00

Table 1: repartition of VARX and VARY in the simulated dataset

Individuals are followed-up for a maximum of 25 years. Information about the following-up length is given below:

Analysis Variable : SURVTIME						
STATUS	Nb obs.	25e centile	Median	75e centile	90e centile	Maximum
0	221	3.26	8.65	16.22	20.94	23.91
1	279	2.58	4.34	6.33	8.33	17.95

Table 2: following-up length

Over the period, 279 event of interest are observed, so that the censoring is 44.20 %.

1.2 – Effects of explanatory variables on the cure fraction and on the survival of uncured individuals

The data are generated so that VARX has an effect on the cured fraction (higher cured fraction within individuals with VARX=0 than within those with VARX=1), but not on the survival of uncured individuals. Conversely VARY has no effect on the cured fraction, but influences the survival of uncured individuals. Compared to the baseline hazard (uncured individuals with VARY=0), the hazard of failure for uncured individuals with VARY=1 is reduced, and increased for uncured individuals with VARY=2.

The TRUE (simulated) proportion of cured individuals given VARX and VARY is as follows:

varX	varY	Nb obs.	Nb Cured	Cured Fraction (%)
0	0	91	35	38.46
	1	86	30	34.88
	2	123	41	33.33
1	0	47	9	19.15
	1	62	12	19.35
	2	91	25	27.47

Table 3: simulated cured fractions given VARX and VARY

The TRUE (Simulated) effect of VARX and VARX on the survival of uncured individuals can be estimated by fitting the standard Cox PH hazards model to (simulated) uncured individuals (after checking for proportional hazards assumption).

Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
varX	1	0.14667	0.12257	1.4318	0.2315	1.158	0.911	1.472
varY1	1	-0.62591	0.16903	13.7116	0.0002	0.535	0.384	0.745
varY2	1	0.65052	0.14997	18.8141	<.0001	1.917	1.428	2.571

Table 4: parameter estimates for the latency part (uncured fraction).

The variable UNCURED in not observed in real datasets. Hence the only information we have is whether individuals have experienced the event of interest (STATUS=1) or if they are right censored (STATUS=0). Obviously, if STATUS=1 then individuals are uncured.

2 – Analysis of the simulated dataset

The Kaplan-Meier estimate of the survival function for time to failure for time to failure for the overall sample is shown in figure 1:

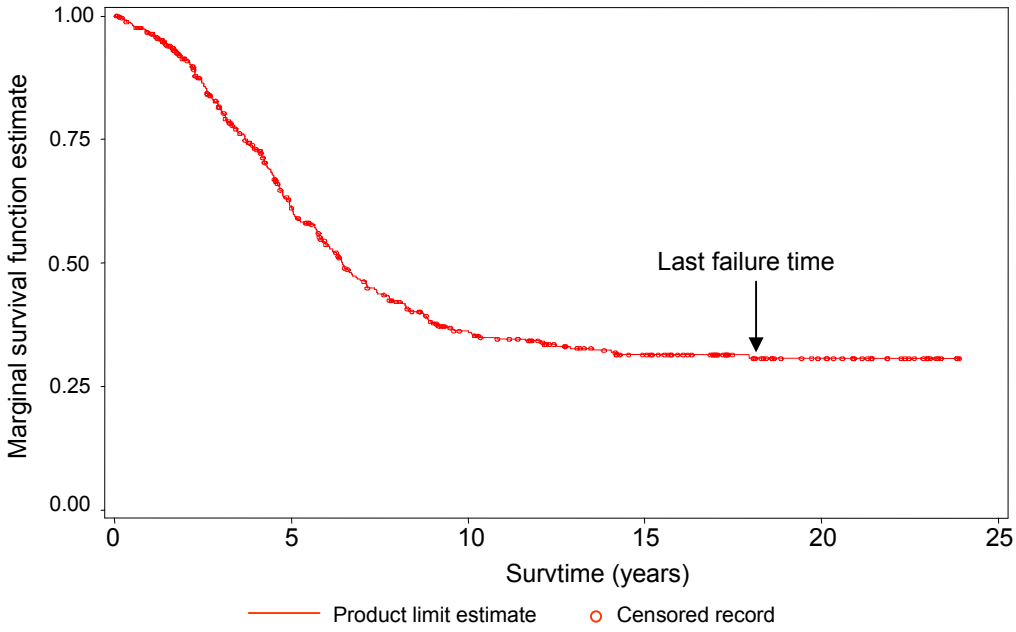

Figure 1 : Kaplan-Meier estimates of the survival function for the overall sample.

From figure 1, it can be seen that the survival function estimates levels off at nonzero proportion (approximately 30%) after a long follow-up and that no further event of interest occurs after 18 years.

2.1 – Analysis with the standard Cox's proportional hazards model

The standard Cox's proportional hazards model is first applied to the dataset is first fitted with Proc PHREG.
The statement is:

```
proc phreg data=simultex;
model survtime*status(0)= varX varY1 varY2/rl;
run;
```

Results for parameters estimates are shown bellow.

Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
varX	1	0.34035	0.12060	7.9640	0.0048	1.405	1.110	1.780
varY1	1	-0.18671	0.16390	1.2978	0.2546	0.830	0.602	1.144
varY2	1	0.27537	0.14617	3.5492	0.0596	1.317	0.989	1.754

Table 5: estimates from the standard Cox PH model.

From table 5, it is concluded that individuals with VARX=1 are about 1.4 time more likely to undergo the event of interest than individuals with VARX=0 (relative Risk= 1.405, p=0.0048). Or equivalently (taking the reciprocal), the hazard for those with VARX=0 is about 71 percent of the hazard for those with VARX=1.

The parameter for VARY1 is not significant at the 5% level (p=0.2546). This means that the failure rate is not increased for individuals with VARY=1, compared to the baseline group (VARY=0). There is however some evidence of higher hazard for individuals with VARY=2, compared to those with VARY=0, but the estimates for VARY2 does not reach the 5 % statistical significance level (p=0.0596).

By using the standard Cox PH model, we make the hypothesis that, if complete follow-up was possible for all individuals, each would eventually experience the event of interest. This hypothesis, however, may not hold for the present dataset. From the Kaplan Meier estimate of the survival (figure 1) there are strong evidences that some individuals may be cured, and the use of standard survival models is questionable.

2.2 – Analysis with mixture cure models.

2.2.1 – Notations

Let *U* be the indicator denoting an individual is susceptible (*U*=1) or non susceptible (*U*=0) to the event of interest and *T* is a nonnegative random variable denoting the failure time of interest, defined only when *U*=1. The mixture cure model is given as follows:

$$S(t,x,z) = \pi(z)\, S(t \,|\, U=1,x) + (1-\pi(z))$$

where $S(t\,|\,x,\,z)$ is the unconditional survival function of *T* for the entire population.
$S(t \mid U=1,\ x)=P(T> t \mid U=1,\ x)$ is the survival function for susceptible individuals given a covariate vector $x=(x_1,\ldots,x_p)'$, and $\pi(z)=P(U=1|z)$ is the probability of being susceptible given a covariate vector $z=(z_1,\ldots,z_p)'$ which may include the same covariates as x. The survival function of cured individuals can be set to one for all finite values of *t* because they will never experience the event of interest.

2.2.2 – Analysis with the PSPMCM SAS macro

2.2.2.1 – The Cox PH mixture cure model

The PSPMCM SAS macro is invocated with the following statement:

```
%pspmcm(DATA= simultex, ID= ident, CENSCOD= status, TIME= survtime,
         VAR= varX(IS, 1) varY1(IS, 0) varY2(IS,0),
         INCPART= logit,
         SURVPART= Cox,
         TAIL= zero, SU0MET= pl,
         MAXITER= 200, CONVCRIT= 1e-5, ALPHA= 0.05,
         FAST= Y,BOOTSTRAP= Y, NSAMPLE= 2000, BOOTMET= ALL,
         GESTIMATE= Y,
         BASELINE=Y,
         SPLOT=Y);
run;
```

- The dataset name is `DATA=simultex;`

- The identification number of individuals is `ID=ident`;
- The censoring indicator is `CENSCOD=status`, the time to failure (or censoring) is `TIME=survtime`.
- The name of the explanatory variables is specified in the `VAR=` parameter, with the information of whether is as to be included in the incidence (I) and/or latency (S) part of the mixture cure model. The values of explanatory variables for with the marginal and conditional survival functions estimates, i.e. S(t) and S(t|U=1) are plotted are specified after the comma. From the above statement `VAR= varX(IS, 1) varY1(IS, 0) varY2(IS,0)`, means that VARX, VARY1 and VARY2 are included in both incidence and latency and that the estimates of the survival functions for uncured individuals and with VARX=1 and VARY1=0 and VARY2=0 will be plotted.
- The incidence part of the mixture cure model is modelled by a logistic regression model (`INCPART=logit`).
- The latency part of the mixture cure model is modelled by the use of semiparametric Cox's PH model (`SURVPART=Cox`).
- The default zero tail constraint (i.e. S(t|U=1)=0 for t> τ_k where τ_k is the last failure time) is imposed to the estimate of S(t|U=1), (`TAIL=zero`) and the product limit estimate is used to compute the estimate or S(t|U=1) (`SU0MET=pl`).
- The maximum number of iterations to perform is the default one (`MAXITER=200`). The convergence criterion is set to 10^{-5} (`CONVCRIT= 1e-5`) and the significance level used for the confidence limits is 5% (`ALPHA= 0.05 `).
- Results are displayed in the output windows when convergence is attained (`FAST=Y `).
- The computation of bootstrap confidence intervals for parameters estimates is requested (`BOOTSTRAP=Y`) and 2000 replicates will be created by resampling with replacement from the original dataset (`NSAMPLE=2000`). All the bootstrap methods for confidence intervals computation, i.e. percentile, bias corrected, normalized bias corrected, hybrid, bias corrected accelerated CI, as well as jacknife after bootstrap are requested (`BOOTMET= ALL`).
- Q-Q plots and histograms of parameters estimates distribution will be plotted (`BASELINE=Y`).
- The estimates of the baseline survival function and of parameters estimates will be stored in a BASELINE dataset (`BASELINE=Y`).
- Finally the plot of the marginal and conditional survival functions are requested (`SPLOT=Y`). They will be displayed for values specified in the `VAR=` parameter.

Convergence is attained in 16 iterations.

Output datasets.

The following dataset are created by the macro:

- LIKELIHOOD stores information about convergence and likelihood.
- ESTIMATES contains the point estimates for parameters. The prefix 'L' indicates that the estimate is for the incidence part, while the prefix "S" indicates that the estimate is for the survival part.
- FAST_INCI_EST and FAST_INCI_OD are created if the parameter FAST=Y is specified. They store the estimates and standard errors and the corresponding odds ratios estimates that have been displayed when convergence was attained.
- FAST_SURV_EST does the same for the latency part.
- PROBCURE_SIMULTEX stores the estimated probability of been uncured which is 1 for uncensored individuals. For censored individuals it is given by:

$$P(U_i = 1 | x_i, z_i, t_i) = \frac{\pi_i(z_i)\, S(t_i | U = 1, x_i)}{1 - \pi_i(z_i) + \pi_i(z_i)\, S(t_i | U = 1, x_i)}$$

- BASELINE_SIMULTEX: is created if the parameter BASELINE=Y is specified. Stores the estimate for the conditional baseline survival function, as well as parameters estimates.
- BASELINE_T_SIMULTEX: is created if the parameters BASELINE=Y and BOOTSTRAP=Y are specified. Stores the estimates for parameters and the conditional baseline survival function from the bootstrap replicates.

- BOOTCI: is created if the parameter BOOTSTRAP=Y is specified. Stores the bootstrap confidence intervals for parameters estimates for the incidence. The BOOT_OR and BOOT_RR dataset store the corresponding bootstrap confidence intervals for the Odds Ratios and the Hazards Ratios.
- BOOTDIST_T_SIMULTEX is created if the parameter BOOTSTRAP=Y is specified. Stores the parameters estimates and some information about convergence for each bootstrap replicate.
- JACKDIST_T_SIMULTEX: is created if the parameter BOOTMET=JACK or BOOTMET=ALL is specified. Stores the parameters estimates and some information about convergence for each jackknife replicate.

Results for the Simultex dataset example are shown in output 1.

RESULTS FOR DATA simultex
Convergence and Log_likelihood
tail completion method : ZERO
SuO estimation : PL

N_Obs	N_failed	N_iter	Converg	LogL_ LOGIT	LogL_ coxPH	LogTotal
500	279	16	4.58793E-06	-307.531	-1300.68	-1608.21

FAST RESULTS FOR THE LOGIT PART (simultex)
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
Intercept	1	0.5477	0.1918	8.1584	0.0043
varX	1	0.4603	0.2032	5.1342	0.0235
varY1	1	0.1545	0.2572	0.3607	0.5481
varY2	1	0.0536	0.2349	0.0521	0.8195

Odds Ratio Estimates

Effect	Odds Ratio Estimate	Lower 95% Confidence Limit for Odds Ratio	Upper 95% Confidence Limit for Odds Ratio
varX	1.585	1.064	2.360
varY1	1.167	0.705	1.932
varY2	1.055	0.666	1.672

FAST RESULTS FOR THE SURVIVAL PART (simultex)
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Hazard Ratio	95% Lower Confidence Limit for Hazard Ratio	95% Upper Confidence Limit for Hazard Ratio
varX	1	0.22329	0.12245	3.3251	0.0682	1.250	0.983	1.589
varY1	1	-0.60548	0.16991	12.6985	0.0004	0.546	0.391	0.762
varY2	1	0.67403	0.15012	20.1596	<.0001	1.962	1.462	2.633

Output 1 : parameters estimates for the Simultex dataset using the Logistic Cox PH mixture cure model (PSPMCM macro).

The standard errors of estimates and the corresponding confidence intervals are based on the inverted Hessian matrix computed on the last maximum likelihood iteration, when convergence is attained. They may be underestimated, especially for the survival part of the model.

From the results of the incidence part, it is concluded that the effect of VARX on the cured fraction is significant, and that VARY as no effect. From the results of the latency part it is concluded that VARX does not influence the survival of uncured individuals, and conversely, that the hazard of failure is increased for uncured individuals with VARY=2 and decreased for uncured individuals with VARY=1.

The cured fraction is computed using the PROBCURE_SIMULTEX dataset. Results are shown in table 6.

varX	varY	N Obs	Cured Fraction (%)	
			estimates	true values
0	0	91	37.79	38.46
	1	86	34.34	34.88
	2	123	33.71	33.33
1	0	47	24.51	19.15
	1	62	22.15	19.35
	2	91	27.99	27.47

Table 6 : cured fraction estimates from the logistic/Cox PH mixture cure model for the Simultex dataset.

The estimated cured fractions are in close agreement with the true (simulated) values.

The marginal and conditional survival functions estimates from the logistic Cox PH mixture cure model are shown in figure 2 and figure 3.

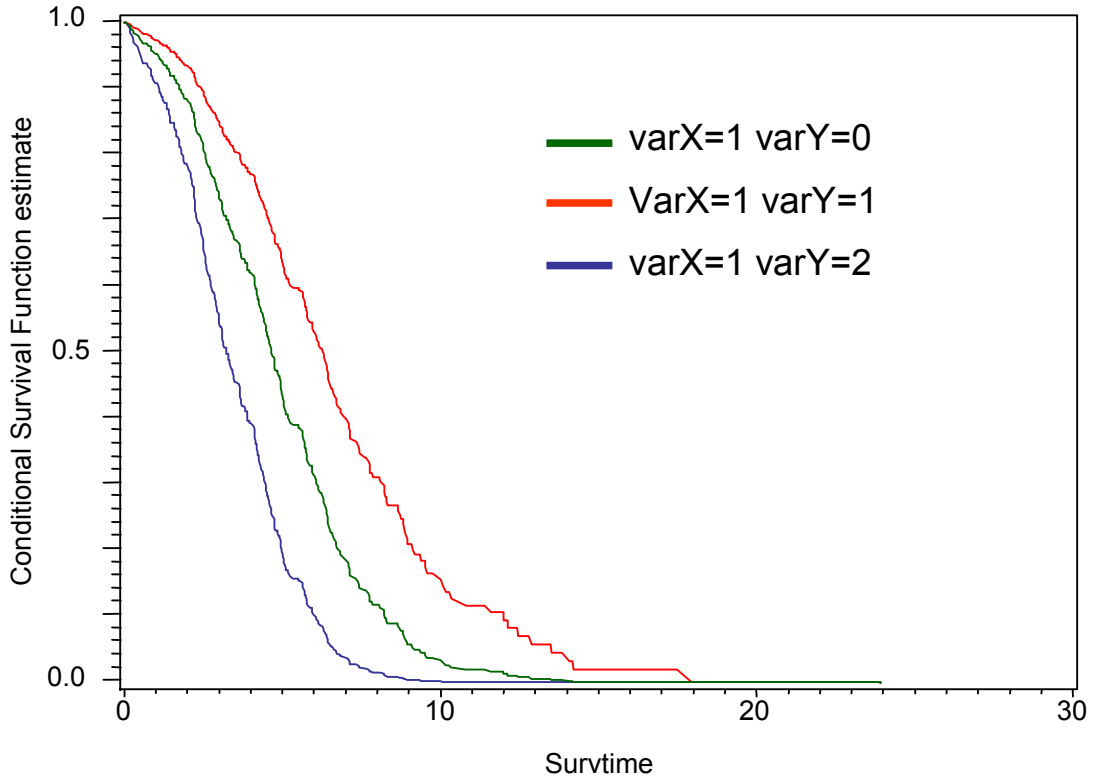

Figure 2 : conditional survival function estimates (survival estimates for uncured individuals) from the logistic Cox PH mixture cure model for Simultex dataset

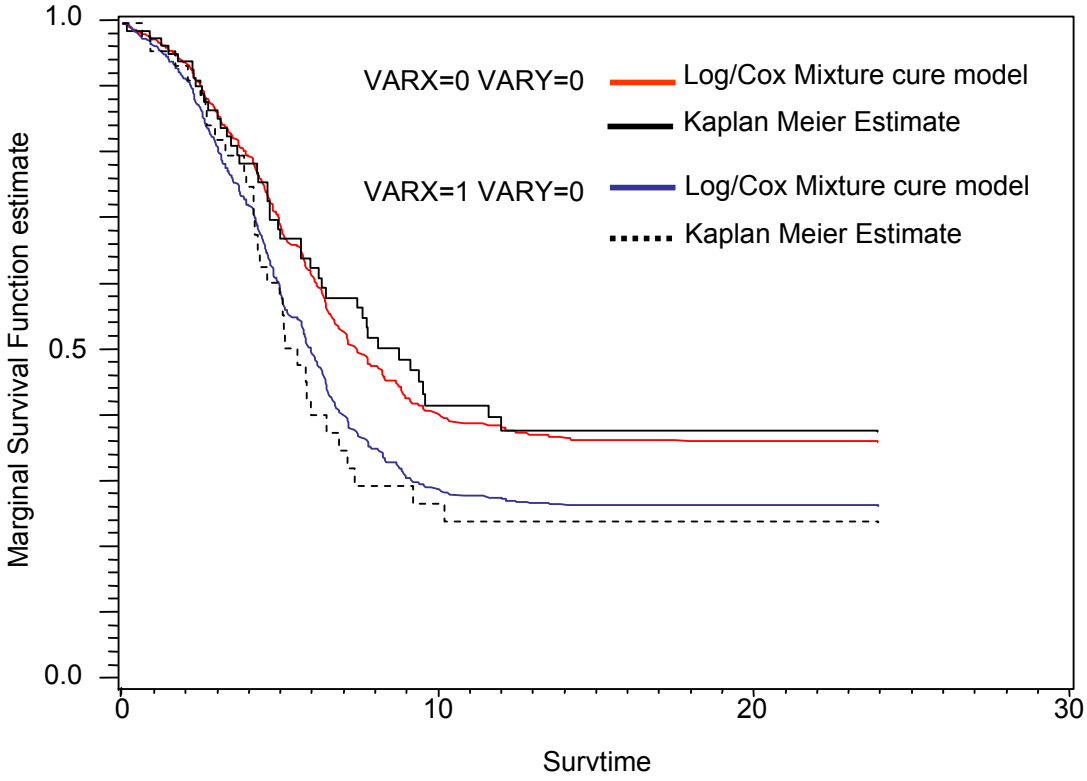

Figure 3 **:** marginal survival function estimates from the logistic Cox PH mixture cure model for Simultex dataset

The bootstrap confidence interval for parameters estimates are given in output 2. The corresponding odds Ratios and Hazard Ratios are shown in output 3.

BOOTSTRAP CONFIDENCE INTERVAL FOR PARAMETERS ESTIMATES
Data set= simultex
(confidence level=95 %, 2000 bootstrap resamples)

Variable	Observed Statistic	Method=BC Lower Confidence Limit	Method=BC Upper Confidence Limit	Method=BCA Lower Confidence Limit	Method=BCA Upper Confidence Limit	Method= PCTL Lower percentile	Method= PCTL Upper percentile	Method=HYB Lower percentile
L_Int	0.54766	0.06588	0.97221	0.05815	0.96716	0.09028	1.01763	0.07769
L_VARX	0.46049	-0.01111	0.93934	-0.01375	0.93561	-0.00032	0.94835	-0.02738
L_VARY1	0.15428	-0.44614	0.84873	-0.44737	0.84873	-0.49847	0.80505	-0.49648
L_VARY2	0.05363	-0.45812	0.62285	-0.45647	0.63424	-0.50815	0.57347	-0.46621
S_VARX	0.22329	-0.04811	0.46969	-0.05453	0.46699	-0.04230	0.47233	-0.02575
S_VARY1	-0.60548	-0.96809	-0.25052	-0.96809	-0.25052	-0.97022	-0.25432	-0.95664
S_VARY2	0.67403	0.36079	0.99630	0.36079	0.99630	0.36501	1.00194	0.34612

Variable	Observed Statistic	Method=HYB Upper percentile	Method=BOOTN Lower Confidence Limit	Method=BOOTN Upper Confidence Limit	Method=JACK Lower Confidence Limit	Method=JACK Upper Confidence Limit
L_Int	0.54766	1.00505	0.07978	0.98792	0.08458	0.99338
L_VARX	0.46049	0.92129	-0.01319	0.91561	-0.01288	0.91455
L_VARY1	0.15428	0.80703	-0.48157	0.81785	-0.47857	0.79699
L_VARY2	0.05363	0.61541	-0.46950	0.59261	-0.47435	0.58873
S_VARX	0.22329	0.48888	-0.03765	0.48096	-0.03708	0.47999
S_VARY1	-0.60548	-0.24074	-0.96244	-0.24870	-0.96840	-0.25181
S_VARY2	0.67403	0.98305	0.34891	0.98793	0.34944	0.98471

Output 2 : bootstrap confidence intervals for paramaters estimates from the logistic/ Cox PH mixture cure model for the Simultex dataset.

ODDS RATIO FOR THE LOGISTIC PART
Data set= simultex
(confidence level=95 %, 2000 bootstrap resamples)

Variable	Odds Ratio	PCTL Lower OR	PCTL Upper OR	BOOTN Lower OR	BOOTN Upper OR	HYB Lower OR	HYB Upper OR
L_VARX	1.58485	0.99968	2.58146	0.98690	2.49829	0.97299	2.51253
L_VARY1	1.16682	0.60746	2.23680	0.61781	2.26563	0.60867	2.24124
L_VARY2	1.05509	0.60161	1.77441	0.62531	1.80870	0.62738	1.85042

Variable	Odds Ratio	JACK Lower OR	JACK Upper OR	BC Lower OR	BC Upper OR	BCA Lower OR	BCA Upper OR
L_VARX	1.58485	0.98720	2.49566	0.98895	2.55828	0.98634	2.54876
L_VARY1	1.16682	0.61967	2.21885	0.64010	2.33668	0.63931	2.33668
L_VARY2	1.05509	0.62229	1.80171	0.63247	1.86424	0.63352	1.88560

HAZARD RATIO FOR THE SURVIVAL PART
Data set= simultex
(confidence level=95 %, 2000 bootstrap resamples)

Variable	Hazard Ratio	PCTL Lower RR	PCTL Upper RR	BOOTN LowerRR	BOOTN Upper RR	HYB Lower RR	HYB Upper RR
S_VARX	1.25018	0.95858	1.60372	0.96305	1.61763	0.97458	1.63049
S_VARY1	0.54581	0.37900	0.77544	0.38196	0.77981	0.38418	0.78604
S_VARY2	1.96213	1.44053	2.72355	1.41753	2.68566	1.41358	2.67260

Variable	Hazard Ratio	JACK Lower RR	JACK Upper RR	BC Lower RR	BC Upper RR	BCA Lower RR	BCA Upper RR
S_VARX	1.25018	0.96360	1.61607	0.95303	1.59950	0.94693	1.59518
S_VARY1	0.54581	0.37969	0.77739	0.37981	0.77840	0.37981	0.77840
S_VARY2	1.96213	1.41828	2.67703	1.43447	2.70825	1.43447	2.70825

Output 3 : bootstrap confidence intervals for odds ratios and hazard ratios from the logistic/ Cox PH mixture cure model for the Simultex dataset.

2.2.2.2 – The Logistic Weibull mixture cure model

The Logistci/ Weibull mixture cure model is also applied to the Simultex dataset with the following statement :

```
%pspmcm(DATA= simultex, ID= ident, CENSCOD= status, TIME= survtime,
         VAR= varX(IS, 1) varY1(IS, 0) varY2(IS,0),
         INCPART= logit,
         SURVPART= Weibull,
         BASELINE=Y,
         SPLOT=Y);
run;
```

Results are shown in output 4.

results for simultex
distribution WEIBULL, link=LOGIT

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
L_int	0.5474	0.2282	500	2.40	0.0168	0.05	0.09902	0.9958	-0.00009
L_VARX	0.4592	0.2348	500	1.96	0.0510	0.05	-0.00209	0.9204	-0.00003
L_VARY1	0.1568	0.3217	500	0.49	0.6262	0.05	-0.4752	0.7887	-0.00002
L_VARY2	0.05540	0.2690	500	0.21	0.8369	0.05	-0.4732	0.5840	-0.00002
_scale	1.8164	0.07236	500	25.10	<.0001	0.05	1.6743	1.9586	-0.00012
S_VARX	0.2176	0.1329	500	1.64	0.1023	0.05	-0.04358	0.4787	-9.94E-6
S_VARY1	-0.6115	0.1868	500	-3.27	0.0011	0.05	-0.9785	-0.2445	0.000047
S_VARY2	0.6427	0.1594	500	4.03	<.0001	0.05	0.3295	0.9560	-6.11E-6
_shape	0.5215	0.02540	500	20.53	<.0001	0.05	0.4716	0.5714	-0.00052

Odd's ratio estimates for data=simultex
distribution WEIBULL, alpha=0.05

Parameter	OR	Lower	Upper	Pr > |t|
L_VARX	1.58275	0.9979	2.5103	0.0510
L_VARY1	1.16973	0.6218	2.2006	0.6262
L_VARY2	1.05696	0.6230	1.7931	0.8369

Relative risk estimates for data=simultex
distribution WEIBULL,alpha=0.05

Parameter	RR	Lower	Upper	Pr > |t|
S_VARX	1.24303	0.9574	1.6139	0.1023
S_VARY1	0.54254	0.3759	0.7831	0.0011
S_VARY2	1.90167	1.3902	2.6013	<.0001

Output 4 : Results from the Logistic/ Weibull mixture cure model for the Simultex dataset.

The marginal and conditional survival functions estimates from the logistic/ Weibull mixture cure model are shown in figure 4 and figure 5.

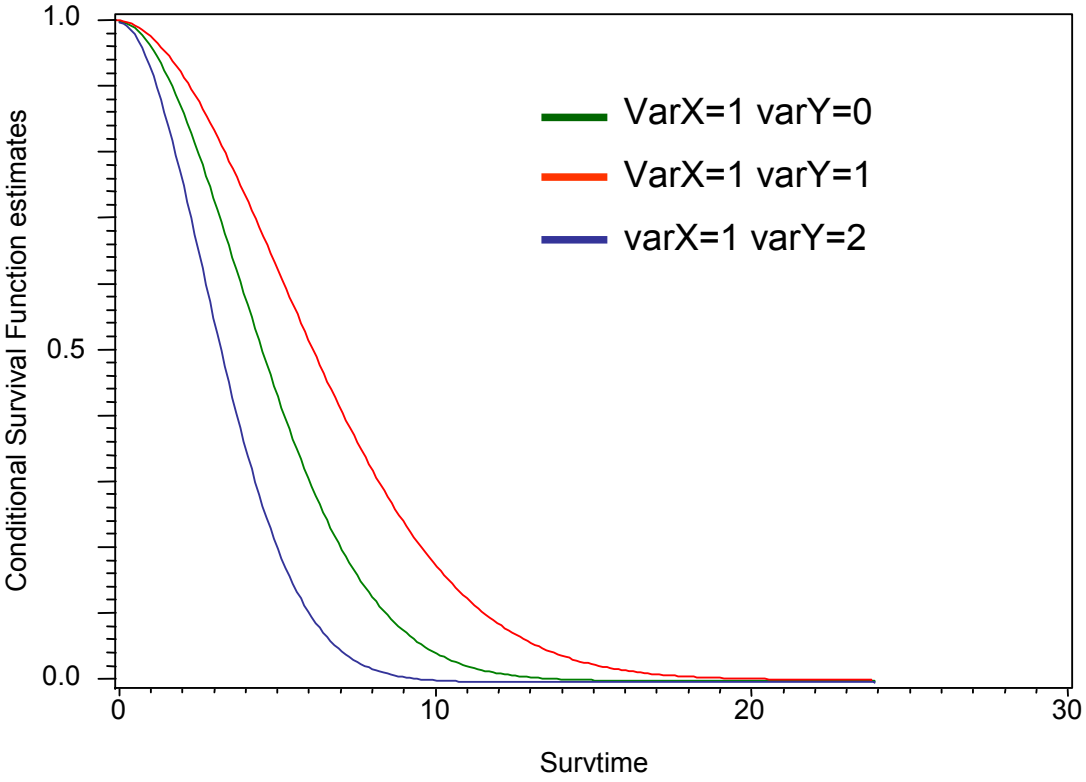

Figure 4: conditional survival function estimates (survival estimates for uncured individuals) from the logistic Weibull mixture cure model for the Simultex dataset.

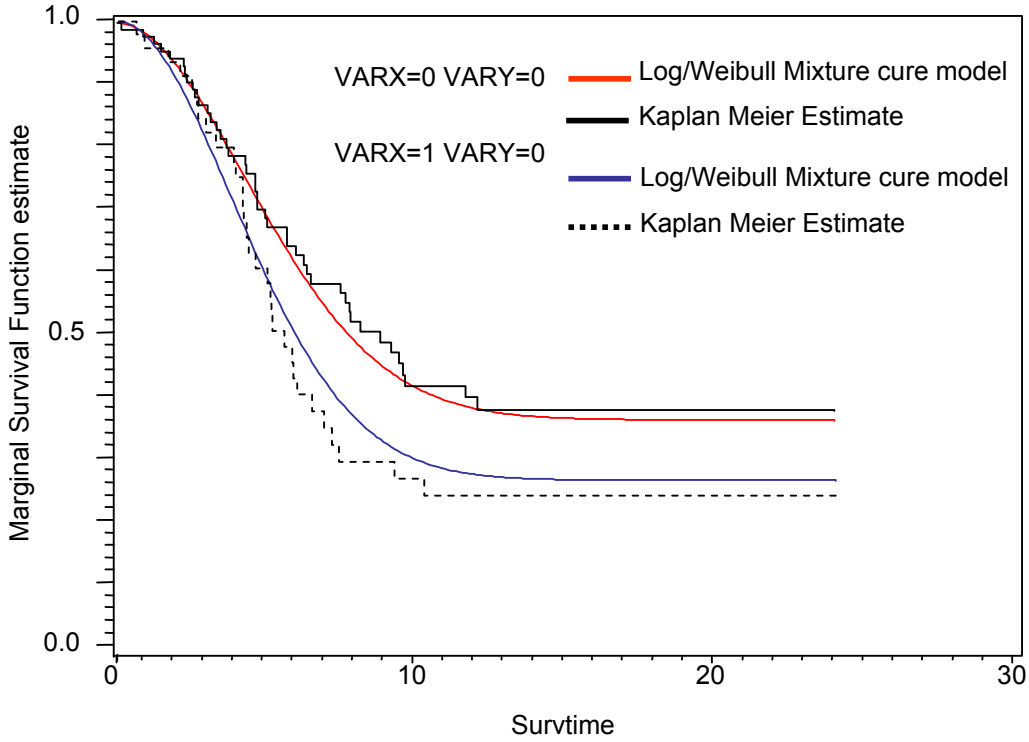

Figure 5 : marginal survival function estimates from the logistic Weibull mixture cure model for the Simultex dataset

The cured fraction estimates from the logistic/ Weibull mixture cure model are shown in table 7.

varX	varY	N Obs	Cured Fraction (%)	
			Estimates	True values
0	0	91	36.65	38.46
	1	86	33.09	34.88
	2	123	35.37	33.33
1	0	47	26.76	19.15
	1	62	23.81	19.35
	2	91	25.69	27.47

Table 7 : cured fraction estimates for the Logistic / Weibull mixture cure model for the Simultex dataset.