# Survival Model and Attrition Analysis

March 2012

Customer Knowledge and Innovation
**Charles Chen, Ph.D**

# Contents

## Background

- Conventional statistical methods are very successful in predicting customers to have an event of interest given target time window. However, they could be challenged by the question: when is the event of interest most likely to occur given a customer?

  *Or how to estimate the following survivor function – S(•)?*

$$Prob(\textbf{event}=\text{'Y'}|_{\textbf{time}}) = S(\textbf{time}, \textbf{covariates})$$

- The goal of this study is, through estimating *S(•),* to show:
  - How to understand parametric and semi-parametric approaches
  - How to employ parametric and semi-parametric approaches to estimate survival function
  - How to use SAS to conduct them
  - How to evaluate the estimations

# What Can Survival Analysis Do?

- **What is Survival Analysis?**
  - Model time to event
  - Unlike linear regression, survival analysis can have a dichotomous (binary) outcome
  - Unlike logistic regression or decision tree, survival analysis analyzes the time to an event
  - Why is that important?
    - Able to account for censoring and time-dependent covariates
    - Can compare survival between 2+ groups
    - Assess relationship between covariates and survival time
    - Capable of answering “who/when are most likely to have an event?”
- **When to use Survival Analysis?**
  - Example:
    - Time to cancellation of products or services (attrition)
    - Time in acquiring add-on products or upgrading
    - Re-deactivation rate after retention treatment
    - etc.
  - When one believes that 1+ explanatory variable(s) explains the differences in time to an event
  - Especially when follow-up is incomplete

# Conventional Modeling vs. Survival Analysis

- Conventional modeling techniques are hard to handle two common features of marketing data, i.e. censoring and time-dependent
- Survival analysis encompasses a wide variety of methods for analyzing the timing of events

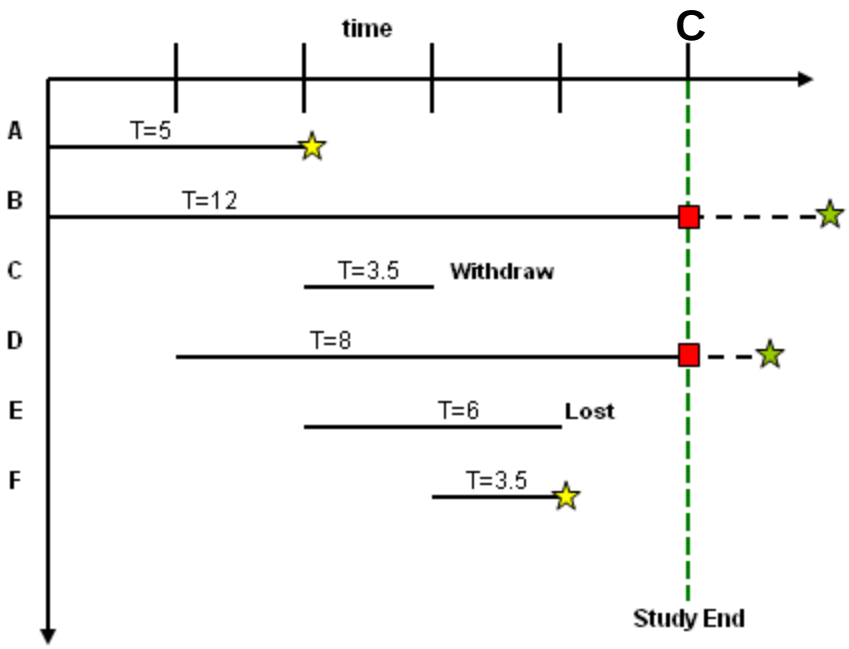
| Technique | Predictors | Outcome Variables | Censoring Permitted |
|---|---|---|---|
| Linear Regression | Categorical or Continuos | Normally Distributed | No |
| Logistic Regression | Categorical or Continuos | Binary or Ordinal or Nominal | No |
| Decision Tree | Categorical or Continuos | Binary or Ordinal or Nominal | No |
| Time Series | Time<br>Categorical or Continuos | Normally Distributed | No |
| Survival Analysis | Time<br>Categorical or Continuos | Binary or Ordinal or Nominal | Yes |

| Techinque | Mathematical Model | Yields |
|---|---|---|
| Linear Regression | $y = kx + b$ | Linear Change |
| Linear Regression | $log(p/(1-p)) = kx + b$ | Odds Ratio |
| Decision Tree | CART or CHAID<br>ID3, C4.5 and C5.0 | Information Gain |
| Time Series | $x(t) = kx(t-i) + b(t)$ | Linear Change |
| Survival Analysis | $h(t) = h0(t) \exp(kx + b)$ | Hazard Rate |

Date: March 2012

# Types of Censoring Schemes

- **Right censoring:**
  - occurs when all you know about an observation on a variable T is that it is less than some value c
- The main reason of right censoring occurring are as follows:
  - Termination of the study
  - Failure due to a cause that is not the event of interest
  - Loss to follow-up
  - We know that subject survived at least to time t
- **Other Types of Censoring**
  - Left censoring – a time of event is only known to be before a certain time.
  - Interval censoring – a data point is somewhere on an interval between two values




The solid line represents an observed period at risk, while the yellow star represents an observed event

The broken line represents an unobserved period at risk ; the filled red box represents the censoring time; and the green star represents an unobserved event

# Approach to Survival Analysis

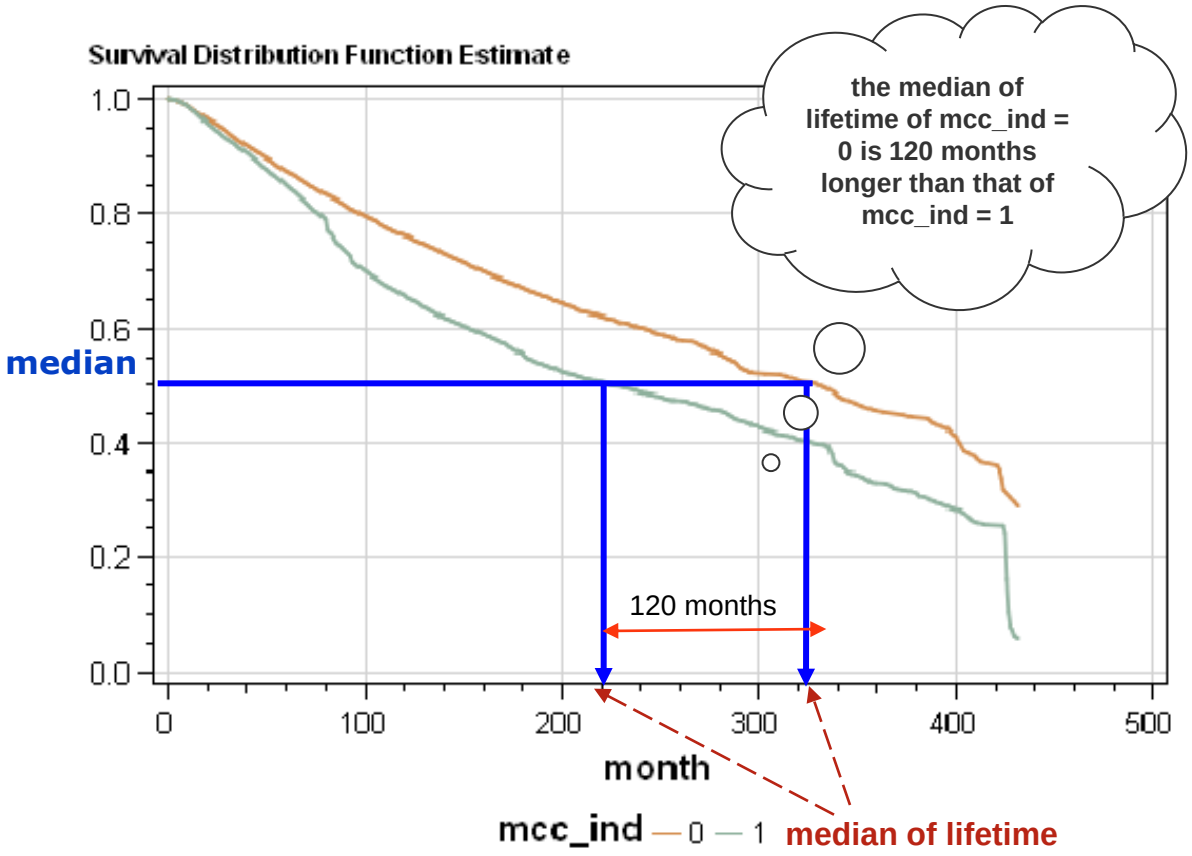
- Like other statistics we have studied we can do any of the following with survival analysis:
  - Descriptive statistics
  - Univariate statistics
  - Multivariate statistics

# Approach to Survival Analysis Contd.

- **Descriptive statistics:**
  - How to describe life time?
    - Mean or Median of survival?
    - What test would you use to compare statistics of survival between 2 cohorts?
  - Average hazard rate
    - Total # of failures divided by observed survival time
    - An incidence rate, with a higher values indicating more events per time
- **Univariate statistics:**
  - Univariate method: Kaplan-Meier survival curves:
    - aka. product-limit formula
    - Accounts for censoring
    - Does not account for confounding or effect modification by other covariates

# Approach to Survival Analysis Contd.

- **Example (Kaplan-Meier curve):** A plot of the Kaplan–Meier estimate of the survival function is a series of horizontal steps of declining magnitude which approaches the true survival function for that population




**Question:**

1. What are the medians of lifetime of 2 types of customers (mcc_ind=0 and 1)?

2. Are their survival distributions significant different?.

| Test of Equality over Strata | | |
|---|---|---|
| **Test** | **Chi-Square** | **Pr >Chi-Square** |
| Log-Rank | 243.7972 | <.0001 |
| Wilcoxon | 254.0723 | <.0001 |
| -2Log(LR) | 241.2043 | <.0001 |

# Approach to Survival Analysis Contd.

- **Comparing Multiple Kaplan-Meier curves**
  - Multiple pair-wise comparisons produce cumulative Type I error – multiple comparison problem
  - Instead, compare all curves at once
    - analogous to using ANOVA to compare > 2 cohorts
    - Then use judicious pair-wise testing
    - Multivariate statistics
- **Limit of Kaplan-Meier Curves**
  - What happens when you have several covariates that you believe contribute to survival?
  - Can use stratified K-M curves – for more than 2 covariates
  - Need another approach – **Model With Covariates** -- for many covariates

# Model With Covariates

- **Three Types of Survival Models**
  - If we model the survival time process without assuming a statistical distribution, this is called ***non-parametric*** survival analysis
  - If we model the survival time process in a regression model and assume that a distribution applies to the error structure, we call this ***parametric*** survival analysis
  - If we model the survival time process in a regression model and assume proportional hazard exists, we call this ***semi-parametric*** survival analysis

# Model With Covariates Contd.

## Proportional Hazards Model

It is to assume that the effect of the covariates is to increase or decrease the hazard by a proportionate amount at all durations. Thus

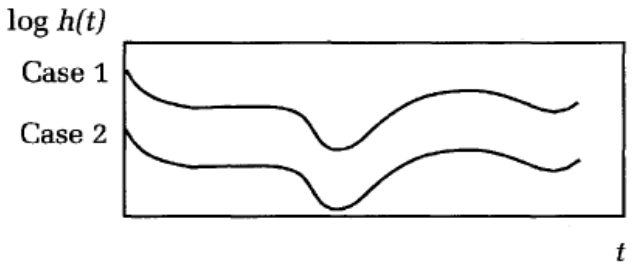
$$\lambda(t, x) = \lambda_0(t)e^{x'\beta}$$

where $\lambda_0(t)$ is baseline hazard, $\exp\{x'\beta\}$ is the relative risk associated with covariate vector x. So,

$$\ln\left(\frac{\lambda(t, x)}{\lambda_0(t)}\right) = x'\beta = \sum_{i=1}^{n} \beta_i x_i$$

Then the survivor functions can be derived as $S(t, x) = S_0(t)^{e^{x'\beta}}$

Parallel Hazard Functions from Proportional Hazards Model can graphed as follows:



# Model With Covariates Contd.

## Proportional Hazards Model Contd.

### Two Common Tests for Examining Proportional Assumption

- Test the interaction of covariates with time

  The covariates should be time-dependent if the test shows the interactions significantly exist, which means the proportional assumption is violated

- Conduct ***Schoenfeld residuals*** Test
  - One popular assessment of proportional hazards is based on Schoenfeld residuals, which ought to show no association with time if proportionality holds. *(Schoenfeld D. Residuals for the proportional hazards regression model. Biometrika, 1982, 69(1):239-241)*

# Model With Covariates Contd.

- **Parametric Survival Model**
  - We consider briefly the analysis of survival data when one is willing to assume a parametric form for the distribution of survival time
  - Survival distributions within the AFT class are the Exponential, Weibull, Standard Gamma, Log-normal, Generalized Gamma and Log-logistic
  - AFT model describes a relationship between the survivor functions of any two individuals. If $S_i(t)$ is the survivor function for individual *i*, then for any other individual *j*, the AFT model holds that

$$S_i(t) = S_j(\phi_{ij} t) \quad \text{for all t}$$

where $\phi_{ij}$ is a constant that is specific to the pair *(i,j).* This model says, in effect, that what makes one individual different from another is the rate at which they age

# Model With Covariates Contd.

## Parametric Survival Model Contd.

- Let T denote a continuous non-negative random variable representing survival time, then a family of survival distributions can be expressed as follows:

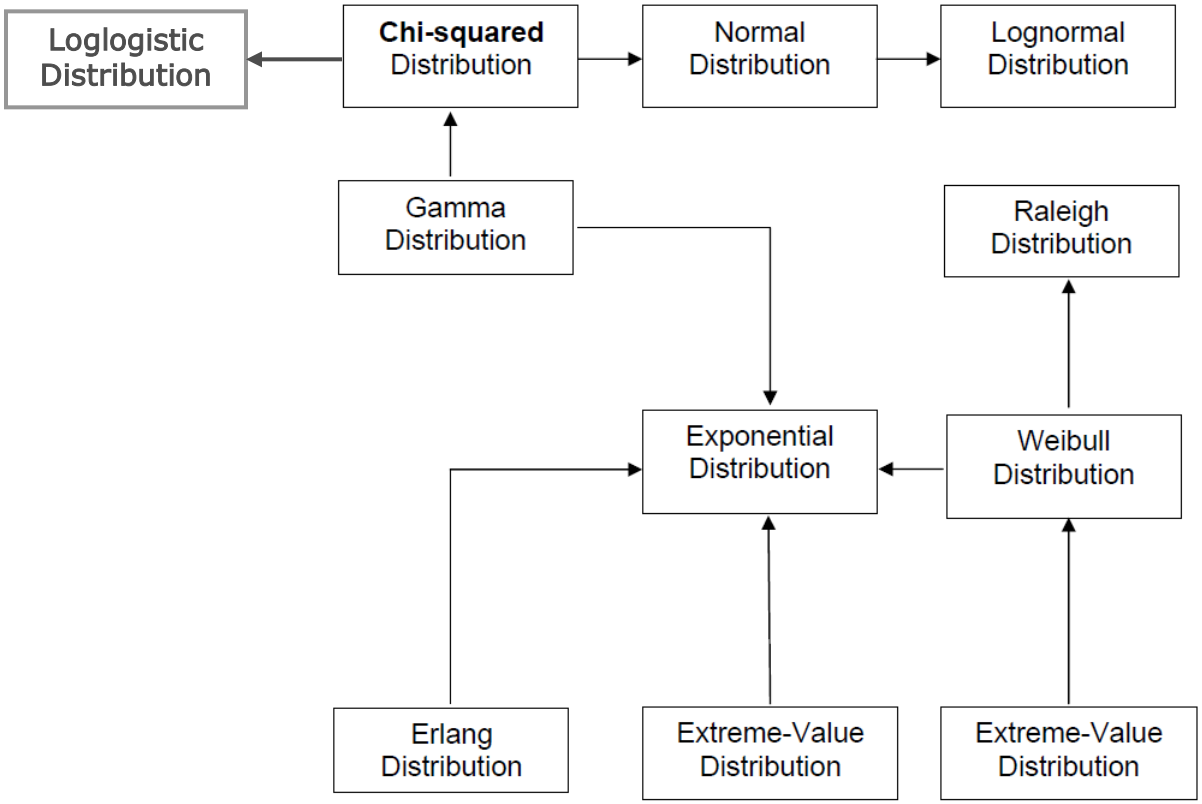
$$\log T_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \sigma W$$

where ***W* is a random disturbance term** with a standard distribution in $(-\infty, \infty)$ and $\sigma$, $\beta_i$ are parameters to be estimated

- A baseline hazard function may change over time
- A linear function of a set of *k* fixed covariates give the relative risk when they are exponentiated
- Parametric approach produces estimates of parametric regression models with censored survival data using the method of ***maximum likelihood***

# Model With Covariates Contd.

- **Parametric Survival Model Contd.**
  - The relationships between various distributions are shown below where the direction of each arrow represents going from the general to a special case



# Model With Covariates Contd.

- **Goodness-of-Fit Tests**
  - There are three common Statistics methods for model comparisons
    - Log-Likelihoods
    - AIC
    - Likelihood-Ratio Statistic

# Model With Covariates Contd.

- **Goodness-of-Fit Tests**
  - **Graphical Methods**
    - Exponential Distribution:
      The plot of - log S(t) versus t should yield a straight line with an origin at 0
    - Weibull Distribution
      The plot of log[-logS(t)] versus log t should be a straight line
    - Log-Normal Distribution
      The plot of $\Phi^{-1}(1-S(t))$ versus log t should be a straight line, where $\Phi(\bullet)$ is the c.d.f
    - Log-Logistic Distribution
      The plot of $\log[(1-S(t))/S(t)]$ versus log t should be a straight line
  - **Cox-Snell Residuals Plot (Collett 1994)**
    - Cox-Snell Residual is defined as:

$$e_i = -\log S(t_i \mid x_i)$$

    - where $t_i$ is the observed event time or censoring time for individual *i*, $x_i$ is the vector of covariate values for individual *i*, and *S(t)* is the estimated probability of surviving to time *t* based on the fitted model.

# Examples:
## Application of Semi-Parametric Survival Model

- **Formulate the Business Problem**
  - Rank the current TD type-A customers by their likelihood to have attrition given a point in time within next 12 months
- **Time Framework**
  - from Dec2009 to Nov2010
- **Population**
  - All customers who are open and active as of Oct2009 except seasonal accounts
  - 10K eligible customers for modeling
  - N customers are flagged as attritors in terms of attrition definition
  - m% overall attrition rate
- **Target** (involuntary attrition is excluded)

| | days since last active date | blockCode | blockDate | attrition | tenure |
|---|---|---|---|---|---|
| case 1 | <=30 | not null | not null | Y | blockDate - open_date |
| case 2 | >30 | null or not null | null or not null | Y | last_active_date - open_date |

**Note**: All examples in this presentation are based on a fake dataset.

# Examples:
## Application of Semi-Parametric Survival Model Contd.

- Model customer data with Cox proportional hazard model using SAS as follows:

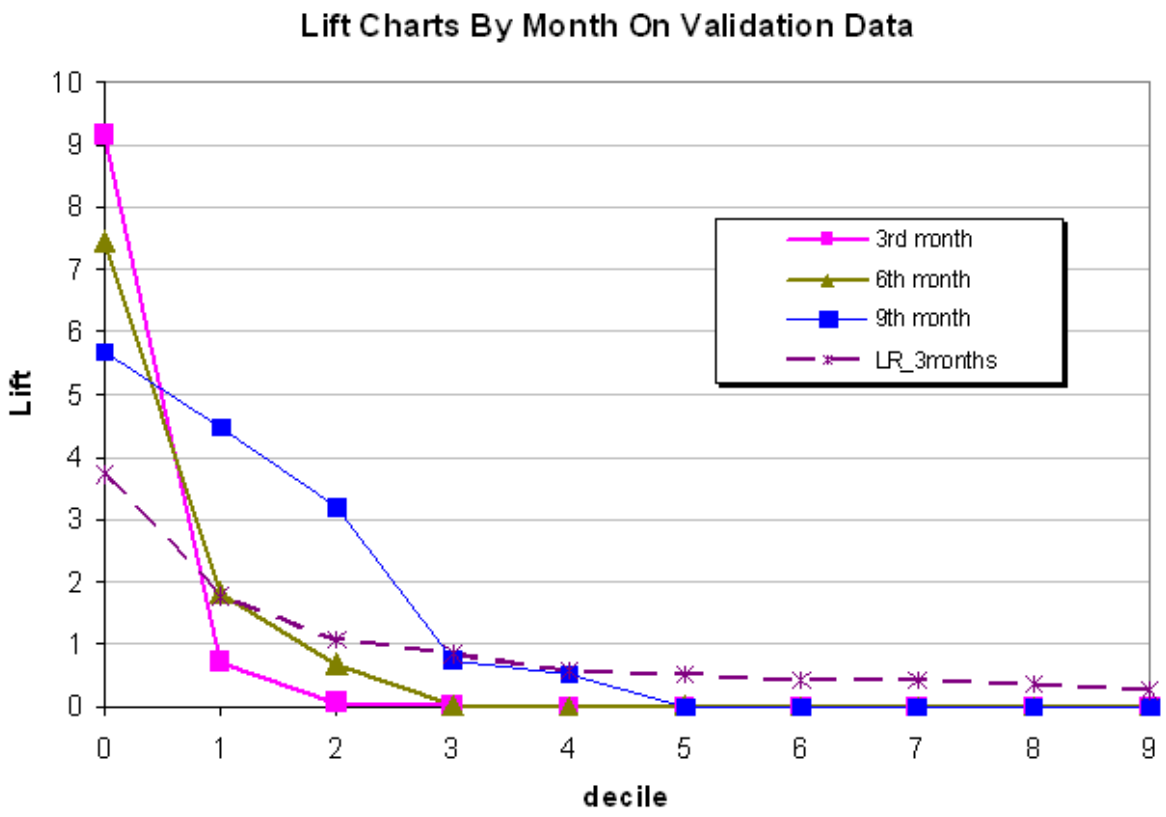
```
proc phreg data=TDM.smpl_typeA_attri_data;
model month*attrition(0)=var1 - var31 /ties=efron ;
baseline out=a survival=s logsurv=ls loglogs=lls;
run;
```

- The syntax of the model statement is MODEL time < *censor ( list ) > = effects < /options > ;
- That is, our time scale is time since Oct2009 (measured in completed months).

# Examples:
## Application of Semi-Parametric Survival Model Contd.

- **Lift Charts**

  The lift charts illustrate the performance of survival model is better than that of logistic regression for modeling this Attrition data



# Examples:
## Application of Semi-Parametric Survival Model Contd.

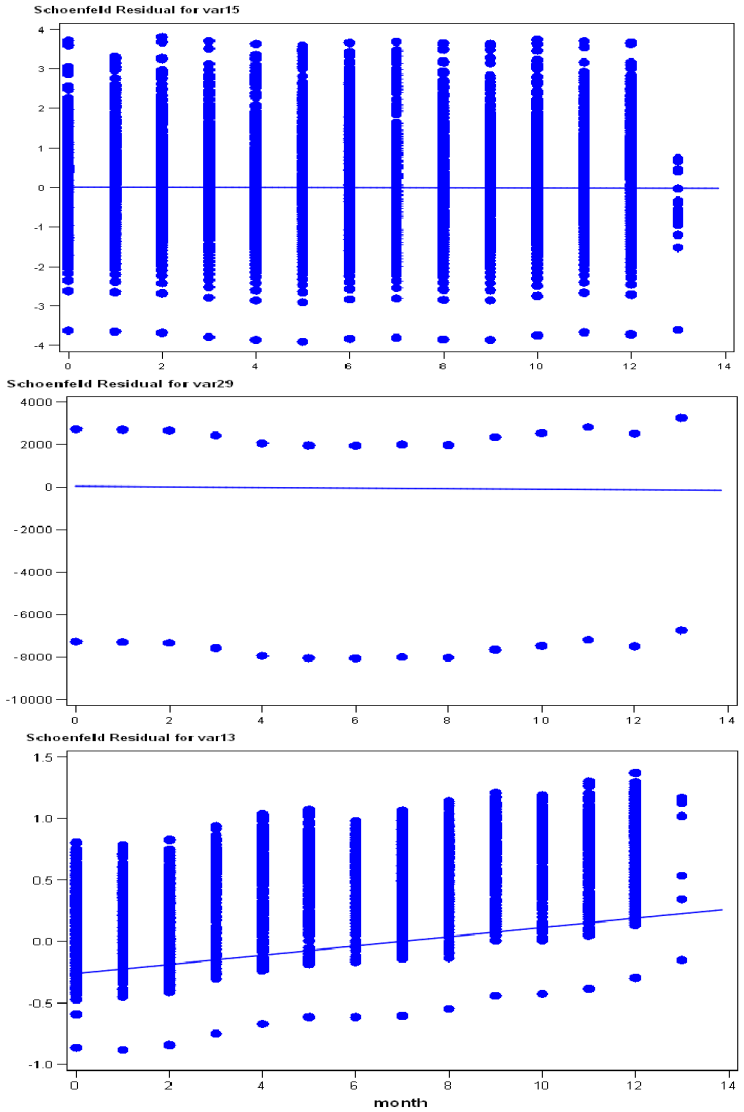
- **Conduct the Tests Using SAS**

```
proc phreg data=TDM.smpl_typeA_attri_data;
model month*attrition(0)= var1-var31 time*var1-
  time*var31/ties=efron;
output out=b ressch=ressch1-ressch31;
test_proportionality: test time*var1-time*var31;
run;
```

The test shows that most of interactions of covariates with time are insignificant at alpha=0.05 level (e.g. p=0.57 and 0.43 for var15*time and var29*time), but a couple of them not. For instance, p<.0001 for var13*time

# Examples:
## Application of Semi-Parametric Survival Model Contd.

- **Schoenfeld Residuals Test**
  - As an example, for ***var15***, its residual has a fairly random scatter, and the OLS regression of the residual on month shows the p-values is 0.5953. That indicates no significant trend exists.
  - For the ***var29*** residuals shows the p-values is 0.1847 and is not very informative , which is typical of graphs for dichotomous covariates
  - The Schoenfeld Residuals test demonstrate there is no evidence of the proportional hazard assumption being violated for those variables
  - For ***var13***, there appears to be a slight tendency for the residuals to increase with time since entering study. The p-value for var13 was 0.02, suggesting that there may be some departure from proportionality for that variable

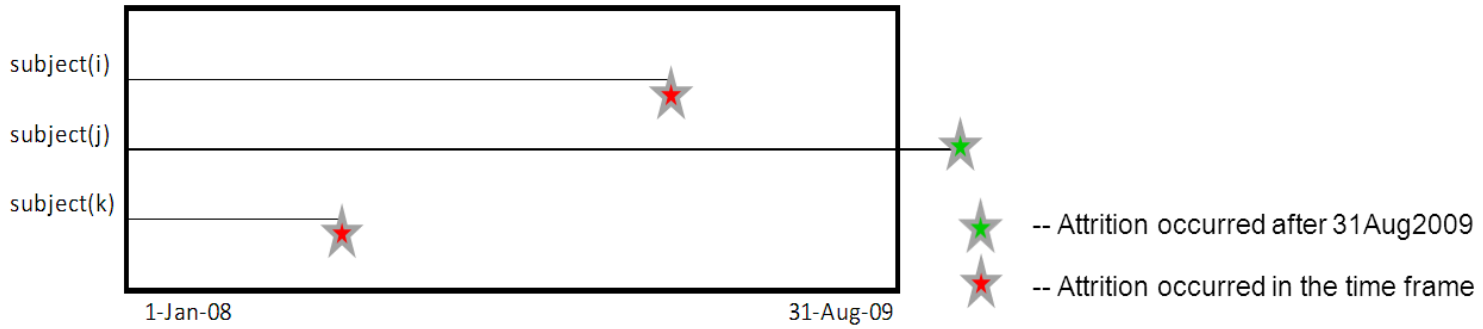


# Examples:
## Application of Parametric Survival Model

- **Objectives**
  - The example will show how to develop parametric survival model using SAS based on TD type-B customer attrition data
  - This analysis will help TD business units better understand attrition risk and attrition hazard by predicting “who will attrite” and most importantly “when will they attrite”
  - The findings from this study can be used to optimize customer retention and/or treatment resources in TD attrition reduction efforts

# Examples:
## Application of Parametric Survival Model Contd.

- **Attrition Definition**
  - TD type-B customer attrition is defined as an type-B customer account that is closed certain number of days (at least 120 days) before maturity
  - The attrition in this study only refers to customer initiated attrition
- **Exclusions**
  - Involuntary attrition are excluded
  - All records with repeat attritions are excluded
  - Mortgage closed within one month after opened are excluded
- **Granularity**
  - This study examines type-B customer attrition at account level
- **Time Frame For Modeling:**
  - 01Jan2008 is the origin of time, and 31Aug2009 is the observation termination time

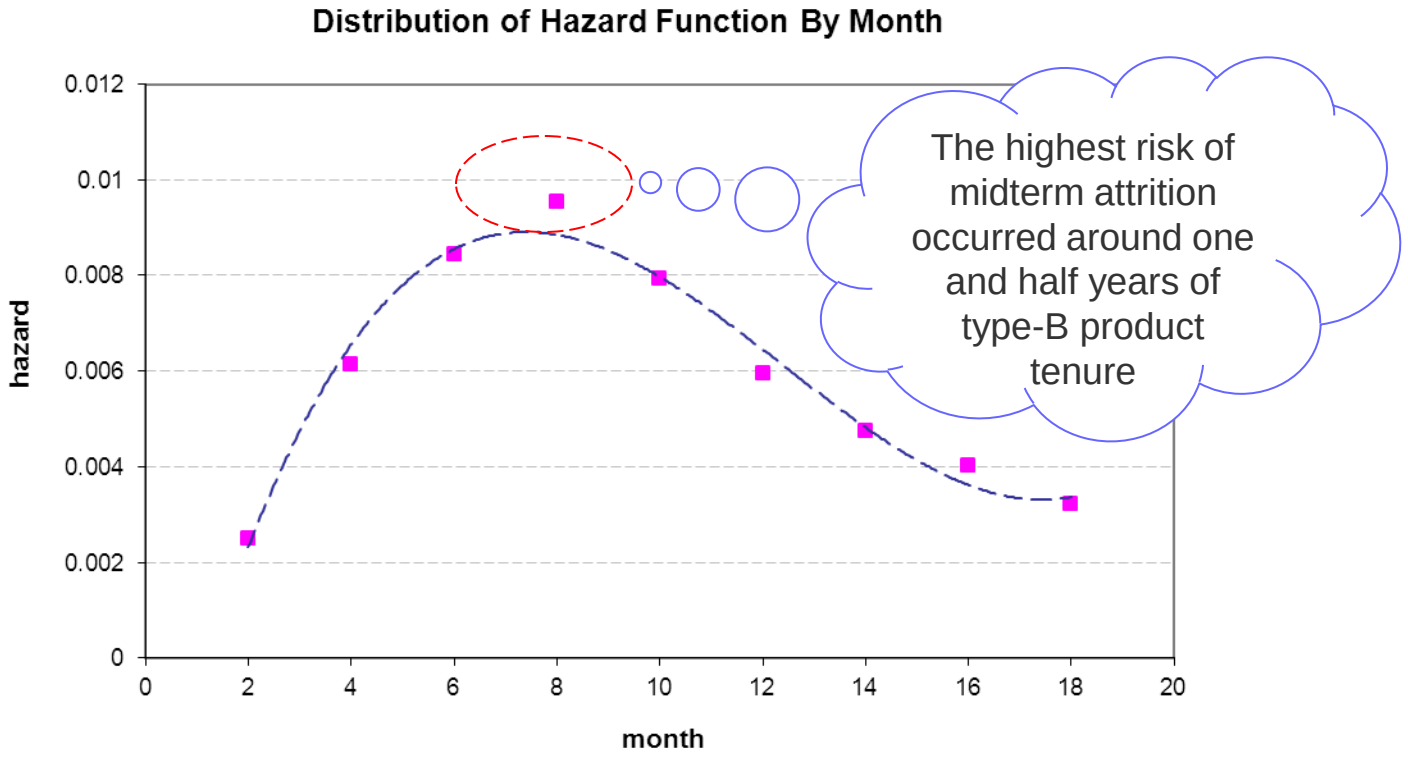


# Examples:
## Application of Parametric Survival Model Contd.

- **Population**
  - Population For Modeling:
    - All type-B customer accounts are active as of 01Jan2008 except those attrite involuntarily in the following months
    - 200K type-B customer accounts are eligible for modeling
    - M accounts are flagged as attritors
    - n% average attrition rate over the 20 months study time window (01Jan2008 to 31Aug2009 )

# Examples:
## Application of Parametric Survival Model Contd.

- **Attrition Hazard Function Estimation**
  - The purpose of estimation is to gain knowledge of hazard characteristics. E.g., when is the most risky time of account tenure for the attrition?
  - The scatter plot below shows that the shape of hazard function approaches to a Log-Logistic distribution.

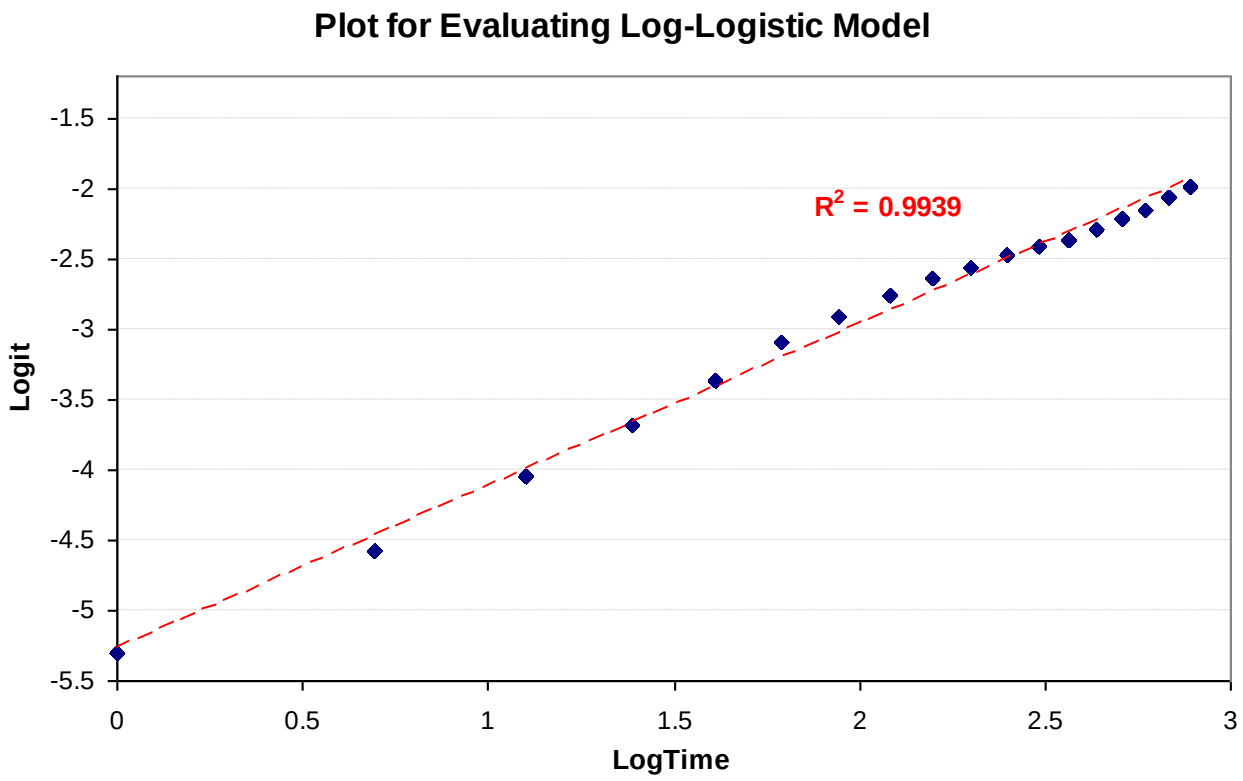


# Examples:
## Application of Parametric Survival Model Contd.

- **Variables For Modeling**
  - There are 20 variables in the modeling dataset
  - 11 categorical variables (X1 – X11) with levels ranging from 2 to 3
  - 9 numeric variables (X12 – X20)

# Examples
## Application of Semi-Parametric Survival Model Contd.

- **Model Type Exploration**
  - The following scatter plot indicates the Log-Logistic model. However, we'll try multiple distributions and select the champion for the final model type



# Examples:
## Application of Parametric Survival Model Contd.

- **PROC LIFEREG -- Parametric Survival Model**

```
proc lifereg data=TDM.smpl_typeB_attri_data;
model time*attrition(0)=&catvars &numvars/dist=&distr;
output out=a cdf=f;
run;
```

Notes:

1. &distr refers to Exponential, Weibull, LogNormal, LogLosgitic, Gamma
2. The performance of each model with different distribution is evaluated by AIC and Cox-Snell residuals plot

# Examples:
## Application of Parametric Survival Model Contd.

- **Evaluation of Model Specification**

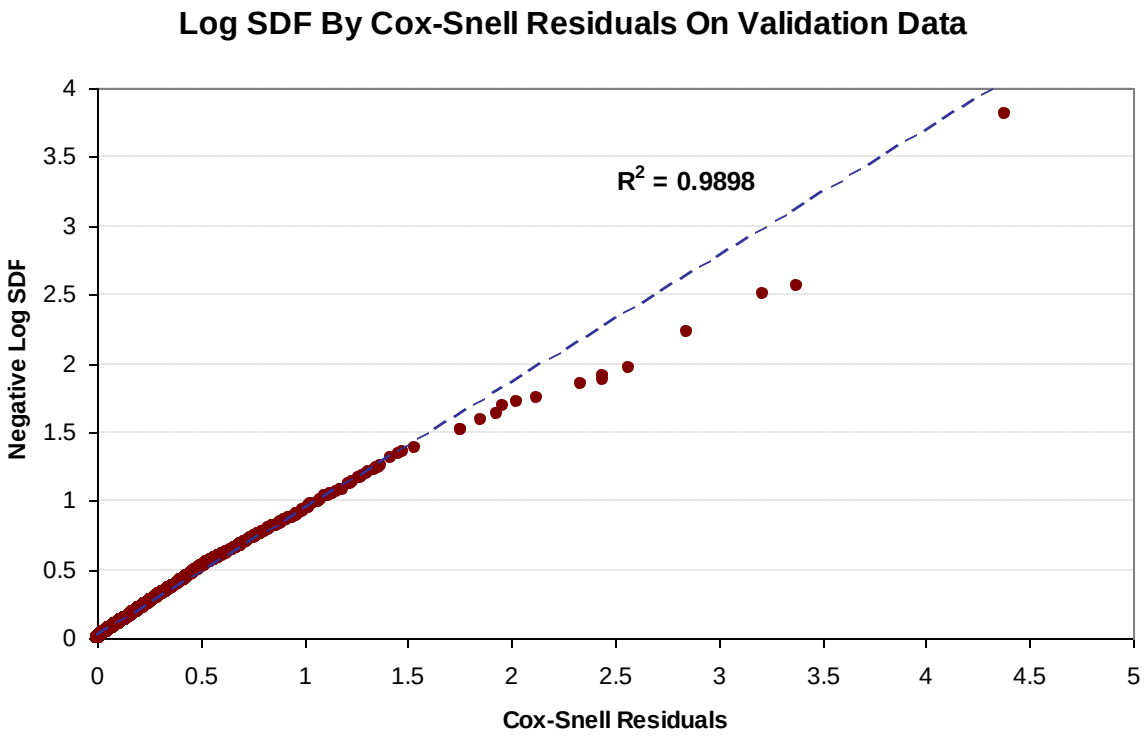
| AIC and Log Likelihood By Model Distribution | | |
|---|---|---|
| ***Log Likelihood*** | ***Distribution*** | ***AIC*** |
| -151746.90 | Exponential | 303495.81 |
| -149094.71 | Weibull | 298193.42 |
| -148662.78 | Lognormal | 297329.56 |
| -252226.57 | Logistic | 504457.13 |
| **-148331.69** | **LLogistic** | **296667.39** |
| -162677.70 | Gamma | 325361.40 |
| | Gamma | >298193 |

**Champion!!**

Note: The Scale is 0.654503 for the champion model

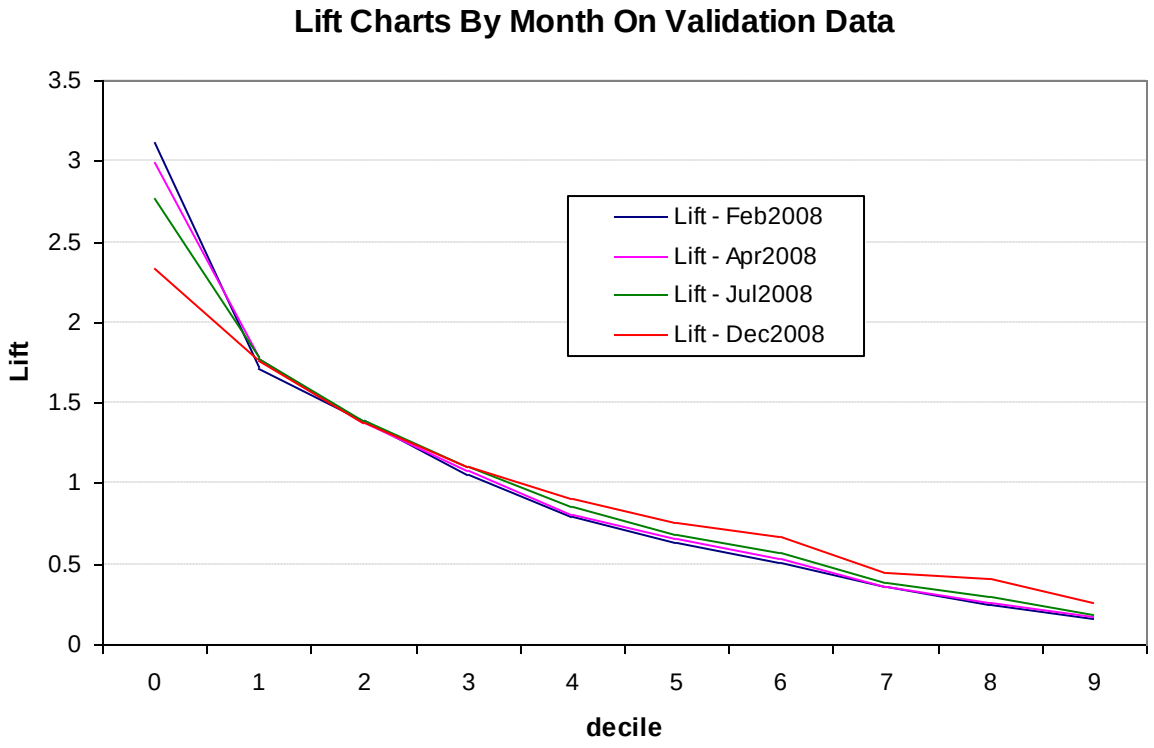
# Examples:
## Application of Parametric Survival Model Contd.

- **Cox-Snell Residuals Plot**
  - The following scatter plot demonstrates the Log-Logistic model fits the type-B customer attrition data nicely



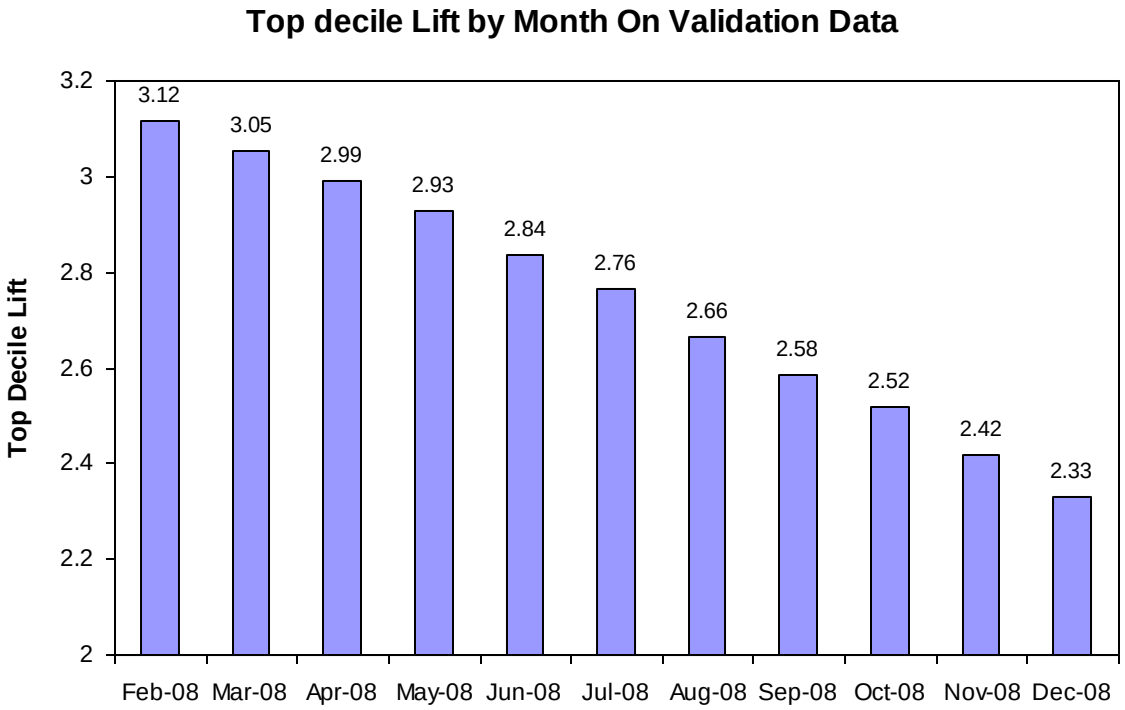
# Examples:
## Application of Parametric Survival Model Contd.

- **Model Performance Validation**
  - The lift decreases monotonically across deciles, which indicates the model has strong predictive power to rank type-B product customers by the probability of attrition



# Examples:
## Application of Parametric Survival Model Contd.

- **Model Performance Validation Contd.**
  - The top decile lift decreases monotonically over month, which is as expected. It means that the power of model rank ordering keeps decaying along with time



# Examples:
## Application of Parametric Survival Model Contd.

- **LIFEREG Procedure Versus PHREG**
  - Estimates of parametric regression models with censored survival data using the method of maximum likelihood
  - Accommodates left censoring and interval censoring, while PHREG only allows right censoring
  - Can be used to test certain hypotheses about the shape of the hazard function, while PHREG only gives you nonparametric estimates of the survivor function, which can be difficult to interpret
  - More efficient estimates (with smaller standard errors) than PHREG if the shape of the survival distribution is known
  - Possible to perform likelihood-ratio goodness-of-fit tests for many of the other probability distributions due to the availability of the generalized gamma distribution
  - Does not handle time-dependent covariates

# Conclusions

- Introduced parametric and semi- parametric survival model approaches, and showed how to conduct and evaluate them using SAS
- Demonstrated Survival analysis is very powerful statistical tool to predict time-to-event in database marketing
- Discovered the insight of attrition risk and attrition hazard over the time of tenure, which is hard for conventional models to do
- Overall, this study is helpful in customizing marketing communications and customer treatment programs to optimally time their marketing intervention efforts