

# ACCELERATED FAILURE TIME MODEL

---

Dr. Aric LaBarr

Institute for Advanced Analytics

MSA Class of 2014

# MODEL STRUCTURE

---

# Accelerated Failure Time Model

- The accelerated failure time (AFT) model is a regression that relates covariates (independent variables) to the event time  $T$ .
- The AFT model is a parametric model – depends on knowledge of the underlying distribution of the data.

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i}$$

# Accelerated Failure Time Model

- We can transform this model into a linear regression model by taking the natural log of both sides of the equation:

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i}$$

- The equation now becomes:

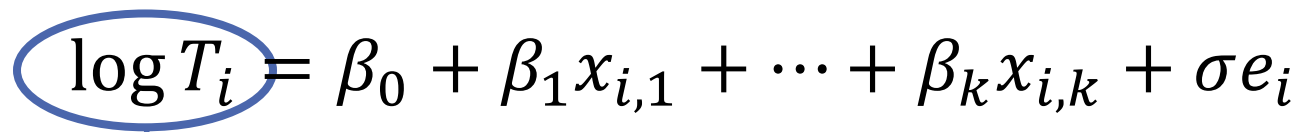
$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i$$

# Accelerated Failure Time Model

- We can transform this model into a linear regression model by taking the natural log of both sides of the equation:

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i}$$

- The equation now becomes:


$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i$$

Ensures positive predictions of  $T$

# Accelerated Failure Time Model

- We can transform this model into a linear regression model by taking the natural log of both sides of the equation:

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i}$$

- The equation now becomes:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i$$

Covariates used to predict  $T$

# Accelerated Failure Time Model

- We can transform this model into a linear regression model by taking the natural log of both sides of the equation:

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i}$$

- The equation now becomes:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i$$

Variance of the disturbances

# Accelerated Failure Time Model

- We can transform this model into a linear regression model by taking the natural log of both sides of the equation:

$$T_i = e^{\beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i}$$

- The equation now becomes:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_k x_{i,k} + \sigma e_i$$

Errors in the model





# Accelerated Failure Time Model

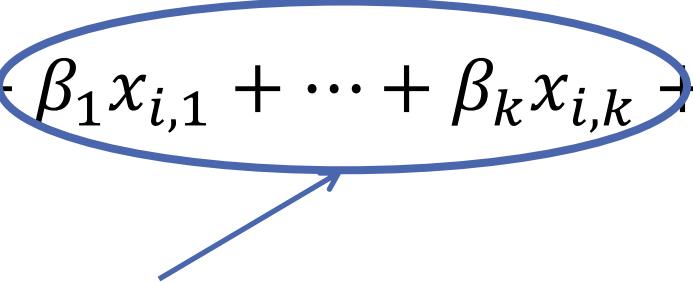
$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i$$

Errors in the model



- The errors in the AFT model can follow many different distributions.
- Assumptions:
  - Constant Mean
  - Constant Variance ( $\sigma$ )
  - Independence across observations

# Accelerated Failure Time Model

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i$$


Covariates used to predict  $T$

- If there is no censoring in the data, traditional OLS could estimate the parameters.
- If there is censoring, maximum likelihood estimation could estimate the parameters.

# Accelerated Failure Time Model

```
proc lifereg data=Survival.Loyalty;
  model Tenure*censored(1) = Income Age Loyalty /
                                dist=lnormal;
run;
```

# AFT Model Parameter Interpretation

- If a parameter estimate is positive, increases in that variable increase the expected survival time.
- If a parameter estimate is negative, increases in that variable decrease expected survival times.
- $100 \times (e^{\beta} - 1)$  is the % increase in the expected survival time for each one-unit increase in the variable.

# Recidivism Parameter Interpretation

| Variable          | $\beta$ Estimate | $100(e^{\beta} - 1)$ |
|-------------------|------------------|----------------------|
| Financial Aid     | 0.3319           | 39.36%               |
| Age at Release    | 0.0333           | 3.39%                |
| Marital Status    | 0.5609           | 75.22%               |
| Prior Convictions | -0.0743          | -7.16%               |



# ERROR DISTRIBUTIONS

---

# Alternative Distributions

- The distribution of the error term determines the distribution of  $T$ .

| Distribution of $e$          | Distribution of $T$ |
|------------------------------|---------------------|
| Extreme Value (1 parameter)  | Exponential         |
| Extreme Value (2 parameters) | Weibull             |
| Normal                       | Log-Normal          |
| Logistic                     | Log-Logistic        |
| Log-Gamma                    | Gamma               |



# Exponential Model

- Survival Function:

$$S(t) = e^{-\frac{t}{\lambda}}$$

- Hazard Function:

$$h(t) = \frac{1}{\lambda}$$

# Exponential Model

- Accelerated Failure Time Model:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i$$

- Proportional Hazards Model:

$$\log h(t) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_k x_k$$

# Exponential Model

- Accelerated Failure Time Model:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i$$

- Proportional Hazards Model:

$$\tilde{\beta}_j = -\beta_j$$

$$\log h(t) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_k x_k$$

# Exponential Model

- There are restrictions of the exponential model.
- In the exponential model  $\sigma = 1$ .
- To evaluate if the data follows an exponential distribution, we can test if  $\sigma = 1$ :
  - Lagrange Multiplier Statistic
  - Null Hypothesis:  $\sigma = 1$

# Exponential Model

```
proc lifereg data=Survival.Recid;  
  model Week*arrest(0) = fin age race wexp mar paro prio  
                        / dist=exponential;  
run;
```

# Weibull Model

- Survival Function:

$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^{\delta}}$$

- Hazard Function:

$$h(t) = \frac{\delta}{\lambda} \left(\frac{t}{\lambda}\right)^{\delta-1}$$

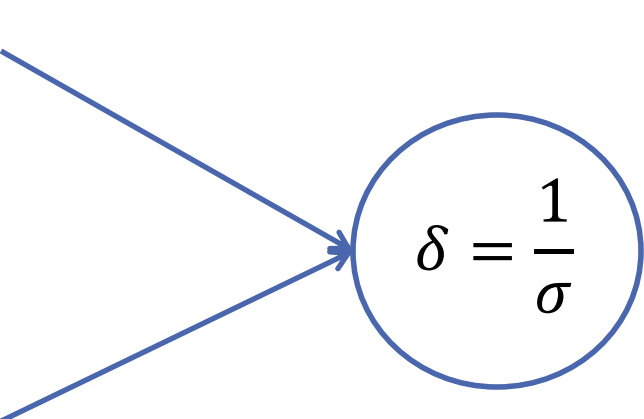
# Weibull Model

- Survival Function:

$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^{\delta}}$$

- Hazard Function:

$$h(t) = \frac{\delta}{\lambda} \left(\frac{t}{\lambda}\right)^{\delta-1}$$


$$\delta = \frac{1}{\sigma}$$

# Weibull Model

- Accelerated Failure Time Model:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i$$

- Proportional Hazards Model:

$$\log h(t) = \alpha \log t + \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_k x_k$$



# Weibull Model

- Accelerated Failure Time Model:

$$\log T_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \sigma e_i$$

- Proportional Hazards Model:  $\alpha = \frac{1}{\sigma} - 1 \quad \tilde{\beta}_j = \frac{-\beta_j}{\sigma}$

$$\log h(t) = \alpha \log t + \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_k x_k$$

# Weibull Model

```
proc lifereg data=Survival.Recid;  
  model Week*arrest(0) = fin age race wexp mar paro prio  
                        / dist=weibull;  
run;
```

# Weibull Model

- One of the most popular models due to simplicity.
- Equals the exponential model when  $\delta = 1$ .
- Weibull model (and its special case – the exponential model) is the only model that is both an AFT model and a proportional hazards model.
- Survival function is easy to manipulate.

$$S(t) = e^{-\left(\frac{t}{\lambda}\right)^\delta} = \exp \left\{ -\left( t e^{-\beta x_i} \right)^\delta \right\}$$

# Weibull Model

```

proc lifereg data=Survival.Recid outest=Beta;
  model Week*arrest(0) = fin age prio / dist=weibull;
run;

data _null_;
  set Beta;
  call symput('b_Int', Intercept);
  call symput('sigma', _SCALE_);
  call symput('b_fin', fin);
  call symput('b_age', age);
  call symput('b_prio', prio);
run;

data Recid2;
  set Survival.Recid;
  if fin = 1 then delete;
  Survival = exp(-(week*exp(-(&b_Int + &b_fin*fin +
    &b_age*age + &b_prio*prio))**(&sigma)));
  Old_t = (-log(Survival))**(&sigma)*exp(&b_Int +
    &b_fin*fin + &b_age*age + &b_prio*prio);
  New_t = (-log(Survival))**(&sigma)*exp(&b_Int +
    &b_fin + &b_age*age + &b_prio*prio);
  Difference = New_t - Old_t;
run;

proc means data=Recid2 mean median min max;
  var Difference;
run;

```

# Gamma Model

- PROC LIFEREG estimates the Gamma model with the DIST=GAMMA option.
- There are two types of gamma distributions – the standardized and generalized gamma.
- The generalized gamma distribution takes a wide variety of shapes including the following distributions as special cases.
  - Exponential
  - Weibull
  - Log-Normal
  - Standardized Gamma



# GOODNESS-OF-FIT TESTS

---

# Goodness-of-Fit Tests

- Since these models are nested within the generalized gamma, we can use the likelihood ratio test.
- Likelihood Ratio Test:

$$\text{LRT} = -2(\log L_{\text{Nested}} - \log L_{\text{Full}})$$



# Goodness-of-Fit Tests

- Here are the log-likelihood values for the models we can compare:

| Log-Likelihood Value | Implied Distribution |
|----------------------|----------------------|
| -325.83              | Exponential          |
| -319.38              | Weibull              |
| -322.69              | Log-Normal           |
| -319.46              | Standard Gamma       |
| -319.38              | Generalized Gamma    |

# Goodness-of-Fit Tests

- Here are the likelihood ratio test values for the comparisons to the generalized gamma:

| LRT   | P-value | Comparison                            |
|-------|---------|---------------------------------------|
| 12.90 | 0.0016  | Exponential vs.<br>Generalized Gamma  |
| 0.00  | 1.00    | Weibull vs. Generalized<br>Gamma      |
| 6.62  | 0.0101  | Log-Normal vs.<br>Generalized Gamma   |
| 0.16  | 0.6892  | Stand. Gamma vs.<br>Generalized Gamma |

# Goodness-of-Fit Tests

```
data GOF;  
  Exp = -325.83;  
  Weib = -319.38;  
  LNorm = -322.69;  
  SGam = -319.46;  
  GGam = -319.38;  
  
  LRT1 = -2*(Exp - GGam);  
  LRT2 = -2*(Weib - GGam);  
  LRT3 = -2*(LNorm - GGam);  
  LRT4 = -2*(SGam - GGam);  
  
  P_Value1 = 1 - probchi(LRT1,2);  
  P_Value2 = 1 - probchi(LRT2,1);  
  P_Value3 = 1 - probchi(LRT3,1);  
  P_Value4 = 1 - probchi(LRT4,1);  
  
run;  
  
proc print data=GOF;  
  var LRT1-LRT4 P_Value1-P_Value4;  
run;
```

# Graphically Evaluating Model Fit

- We can also use graphical diagnostics to evaluate the fit of the data to distributional assumptions.
  - Exponential:  $t$  is linearly related to  $-\log S(t)$
  - Weibull:  $\log t$  is linearly related to  $\log(-\log S(t))$
- SAS provides these plots.

# Graphically Evaluating Model Fit

- Patterns exist in the log-logistic and log-Normal distributions as well.
  - Log-logistic:  $\log t$  is linearly related to  $\log \left( \frac{S(t)}{1-S(t)} \right)$
  - Log-Normal:  $\log t$  is linearly related to  $\Phi^{-1}(1 - S(t))$
- SAS does **not** give these plots through options.
- We have to create them ourselves.

# Graphically Evaluating Model Fit

```
proc lifetest data=Survival.RECID method=life
              plots=(s,ls,lls) outsurv=Pred_Surv
              width=1;
  time week*arrest(0);
run;

data Pred_Surv;
  set Pred_Surv;
  s = survival;
  logit = log((1-s)/s);
  lnorm = probit(1-s);
  lweek = log(week);
run;

proc sgplot data=Pred_Surv;
  series y=logit x=lweek;
run;

proc sgplot data=Pred_Surv;
  series y=lnorm x=lweek;
run;
```

