# Instabilities using Cox PH for forecasting or stress testing loan portfolios

Joseph L Breeden
Prescient Models LLC
breeden@prescientmodels.com

Anthony Bellotti
Imperial College
a.bellotti@imperial.ac.uk

Aleh Yablonski
Belarusian State University
yablonski@bsu.by

August 16, 2015

**Abstract**

Taking a cue from the literature on Age-Period-Cohort models, we explore possible instabilities in applying Cox PH in the context of modeling credit risk on loan portfolios. Although Cox PH performs well at recovering known parameters for test data sets in one or two dimensions, when applied to all three dimensions origination date (vintage), performance date (time), and age of the loan (age), numerical experiments reveal errors in recovering the initial parameters. By retracing the original derivation of the partial likelihood function used to estimate parameters in Cox PH, we demonstrate that the linear dependency between age, vintage, and time prevents the original derivation from providing unique answers. The end result is that Cox PH can become unstable when origination scoring factors and macroeconomic factors are included simultaneously alongside the assumed hazard function.

Keywords: Cox Proportional Hazards Models; Age-Period-Cohort Models

## 1 Introduction

Survival models [7] are an essential element in understanding the dynamics of processes in many fields. For loan portfolios, they can quantify the time to default as a function of the age of the loan, time to attrition, time to recovery, and many others. As has been observed in previous portfolio analyses [2, 4, 3], loans within a homogenous segment will share a characteristic survival function, capturing the cumulative probability of surviving the specified event to a given time. The corresponding hazard function is the probability of an event like default occurring at a given time conditioned on not having occurred prior to then.

Cox [6] recognized that not every "unit" being tested for failure would have identical properties, so the hazard function could be scaled according to the unique risks of the unit. For lending, the analogy is clear. We expect a common hazard function for default probability also known as a lifecycle function or loss timing function, but individual loans could have higher or lower risk of default throughout their lifecycles based upon their attributes (credit bureau attributes, debt-to-income ratio, loan-to-value for the collateral, etc.)

Cox's solution was the proportional hazard model, or Cox model, as shown here.

$$\lambda(a, z) = e^{zb}\lambda_0(a) \tag{1}$$

where $\lambda(a, z)$ is the probability of an event, $\lambda_0(a)$ is the hazard function, $z$ are a set of factors and $b$ are the coefficients estimated to scale the first relative to the hazard function given the input factors. Expressions of the Cox model typically refer generically to time of an event. We instead refer to the age $a$ at which an event occurs in order to avoid confusion with macroeconomic effects versus calendar date.

In the language of lending, a 'score' is created, $zb$, where the estimation of the coefficients is adjusted for the timing of the event. Did the event occur earlier or later than one might expect given the hazard function, and can that divergence be explained with available factors? Cox's derivation involved the use of another innovation, the partial likelihood function. Since all loans would be subject to the same hazard function, this hazard function drops out of the parameter estimation for the scoring coefficients. The consequence is that one can estimate the score without even estimating the hazard function. However, to make forecasts of actual probabilities, both are needed but can be separately estimated. The most common techniques for measuring the hazard function are the KaplanMeier [10], NelsonAalen [11, 1] , or Breslow [5] estimators.

In the current research, we explore the use of Cox proportional hazards (Cox PH) models in the context of stress testing. In addition to the usual lifecycle and scoring factors, we want to add macroeconomic factors such that we can create scenario-based forecasts of the future probability of default, prepayment, etc. However, introducing a third dimension in the Cox PH framework adds a complexity not present in the original application. We explore the implications of applying Cox PH in stress testing, both analytically and numerically, to demonstrate that it suffers the same model specification error known in Age-Period-Cohort models. Unique solutions may be obtained from Cox Ph because of specific assumptions made regarding the specification errors.

## 2 Conceptual Basis

In this section, we review the derivation of Cox PH in the context of stress testing in order to highlight some intrinsic problems.

Suppose we observe $n$ independent individuals with identically distributed failure times, which for lending could be loan default. Some of them are observed to fail and the rest be censored.

Denote the distinct failure ages by $a(1) < a(2) < ... < a(k)$ . The set of individuals at risk at age $a$ is called the risk set at age $a$ and denoted $R(a)$. This consists of those individuals whose failure or censoring age is at least $a$. Censoring can occur if something other than the event being studied causes the individual to disappear from the data set. For example, when studying default, attrition (loan pay-off) is the most likely cause of censoring.

Writing the log of Equation 1 we get

$$log(\lambda(a, z)) = log(\lambda_0(a)) + zb \tag{2}$$

For stress testing, we will want to including scoring and macroeconomic factors. Assume $zb = G(v)\beta + H(t)\gamma$, where $G(v)$ are a set of scoring factors observed at loan origination date $v$, and $H(t)$ are a set of economic factors observed at calendar date $t$. $\beta$, and $\gamma$ are vectors of unknown parameters. $G(v) = (G_1(v), G_2(v), ..., G_q(v))$ and $H(t) = (H_1(t), H_2(t), ..., H_r(t))$ are known vector valued functions (basis functions). With these assignments, we obtain

$$log(\lambda(a, v, t)) = log(\lambda_0(a)) + G(v)\beta + H(t)\gamma \tag{3}$$

For notational symmetry, we will express the log of the baseline hazard function as $F(a)\alpha$ where $F(a) = (F_1(a), F_2(a), ..., F_p(a))$ is a vector-valued set of basis functions and $\alpha$, are vectors of unknown parameters. This is a parameterization of the hazard function, but in discrete time, which is the case for loan performance data, the basis functions to express the hazard function could be as simple as the set of observable ages with estimated coefficients $\alpha$ for each.

The model then becomes

$$log(\lambda(a, v, t)) = F(a)\alpha + G(v)\beta + H(t)\gamma \tag{4}$$

These notational changes do not affect Cox's proposed estimation. $G(v)\beta + H(t)\gamma$ can be estimated with Cox's partial likelihood function and $F(a)\alpha$ is still just the hazard function as estimated before.

We can assume that linear functions such as $a + bv$ and $c + dt$ can be represented as linear combinations of respective entries $G(v)$ and $H(t)$. If not we will include them into the set of basic functions. For account $i$ originated at $v_i$ whose failure occurs at age $a_i$ and calendar date $t_i = a_i + v_i$, the partial likelihood is

$$L(\beta, \gamma) = \prod_{i=1}^{k} \frac{exp(G(v_i)\beta + H(t_i)\gamma)}{\sum_{j \in R(a_i)} exp(G(v_j)\beta + H(t_j)\gamma)} \tag{5}$$

$$L(\beta, \gamma) = \prod_{i=1}^{k} \frac{exp(G(v_i)\beta + H(a_i + v_i)\gamma)}{\sum_{j \in R(a_i)} exp(G(v_j)\beta + H(a_j + v_j)\gamma)} \tag{6}$$

If some functions $sG_j$ or $H_i$ are constants, then we cannot estimate the respective coefficients $\beta$ or $\gamma$. Equation 2 can have only a single constant term.

In Cox's original derivation, he points out importantly that estimation of the baseline hazard function $\lambda_0(a)$ assumes that all $z = 0$, so the constant term must be part of the hazard function.

In addition, suppose that $G(v)$ and $H(t)$ contain linear functions, say $G_1(v) = b + cv$ and $H_1(t) = d + et$. Denote $G^*(v) = (G_2(v), ..., G_q(v))$ and $H^*(t) = (H_2(t), ..., H_r(t))$. Also let $\beta^*$ and $\gamma^*$ be vectors $\beta$ and *gamma* without the first, linear component. Then

$$L(\beta,\gamma) = \prod_{i=1}^{k} \frac{exp(G^*(v_i)\beta^* + H^*(a_i+v_i)\gamma^* + G_1(v_i)\beta_1 + H_1(a_i+v_i)\gamma_1)}{\sum_{j\in R(a_i)} exp(G^*(v_j)\beta^* + H^*(a_i+v_j)\gamma^* + G_1(v_i)\beta_1 + H_1(a_i+v_j)\gamma_1)} \tag{7}$$

$$L(\beta,\gamma) = \prod_{i=1}^{k} \frac{exp(G^*(v_i)\beta^* + H^*(a_i+v_i)\gamma^* + b\beta_1 + d\gamma_1 + v_i(c\beta_1+e\gamma_1) + ea_i\gamma_1)}{\sum_{j\in R(a_i)} exp(G^*(v_j)\beta^* + H^*(a_i+v_j)\gamma^* + b\beta_1 + d\gamma_1 + v_j(c\beta_1+e\gamma_1) + ea_i\gamma_1)} \tag{8}$$

Canceling terms we get

$$L(\beta,\gamma) = \prod_{i=1}^{k} \frac{exp(G^*(v_i)\beta^* + H^*(a_i+v_i)\gamma^* + v_i(c\beta_1+e\gamma_1)}{\sum_{j\in R(a_i)} exp(G^*(v_j)\beta^* + H^*(a_i+v_j)\gamma^* + v_j(c\beta_1+e\gamma_1)} \tag{9}$$

Therefore, coefficients $\beta_1$ and $\gamma_1$ cannot be estimated separately. Only their linear combination is estimable. In the general case we can write $G(v) = \hat{G}(v)T$ and $H(t) = \hat{H}(t)S$ , where $\hat{G}(v) = \hat{G}_1(v), \hat{G}_2(v), ..., \hat{G}_q(v)$ , $H(t) = \hat{H}_1(t), \hat{H}_2(t), ..., \hat{H}_r(t)$ with $\hat{G}_1(v) = a + bv$ and $\hat{H}_1(t) = c + dt$. $T$ and $S$ are nonsingular $qq$ and $rr$ matrices respectively. Then we have

$$\lambda(a,v) = exp(F(a)\alpha + \hat{G}(v)\hat{\beta} + \hat{H}(t)\hat{\gamma}), \tag{10}$$

where $\hat{\beta} = T\beta$ and $\hat{\gamma} = T\gamma$ . Therefore coefficients $\hat{\beta}_1$ and $\hat{\gamma}_1$ are unidentifiable. It means that not all initial coefficients $\beta$ and $\gamma$ can be estimated uniquely.

We can also see this result from the perspective of the APC literature. Previous research [9] has shown that Equation 4 can have only one constant and two linear terms because of the relationship $t = v + a$. Since Cox states that $z = 0$ for the estimation of the hazard function, the hazard function will have whatever overall constant and age-based trend are in the data. That leaves a single estimable trend for the $v$ and $t$ functions. In fact, because of the constraint $z = 0$, we know that the linear trends must sum to zero.

In this way, the Cox PH estimation provides a unique solution, but subject to a set of constraints. From the APC literature, we know that only one constant term is estimable and we cannot estimate three independent trends among the three dimensions $a$, $v$, and $t$. The constraints implicit within Cox PH appear to be the following.

1. The constant term is captured in the hazard function. Any estimated constants in $zb$ will net to zero.

2. One linear trend versus age is incorporated in the hazard function.

3. Only the sum of the trends in calendar date $t$ and origination date $v$ are estimable, not the separate trends in $v$ and $t$.

We will see in numerical experiments that the third constraint is not always a true representation of the process that generated the data set to be modeled. In-sample, any such weaknesses will not be apparent in the residuals or confidence intervals, but during forecasting, any misspecification of the trend can lead to non-stationarity.

# 3 Numerical Experiments

To demonstrate the above concepts, we conducted a series of numerical experiments. The goal was to generate data sets from known parameters and then use Cox PH to estimate the parameters from which the data was estimated. We want to identify systematic divergences between estimated and original parameters to within estimation error. Note that Cox PH always minimizes the error. The question at hand is whether Cox PH can uniquely estimate the original parameters.

## 3.1 Data Generation

To create simulated data, we use a form of the Cox model where $t$ is the age of account $i$, $X_i$ is the idiosyncratic credit risk of individual $i$, $G(v)$ is the net credit risk for the vintage cohort, $H(t)$ is the impact versus calendar date such as economic impacts.

$$\lambda(a, z_i(v,t), \beta) = \lambda_0(t)e^{z_i(t)\beta} = \lambda_0(a)\exp(\beta_1 X_i + \beta_2 G(v) + \beta_3 H(t)) \tag{11}$$

For estimation with Cox PH, we create a data set with columns $X_i$, $G(v)$, $v$, $H(t)$, $st$ the survival time expressed as the age of the account at event, and *delta* the event indicator, either 1 for default or 0 for censored. Censored accounts are those which survive at least to the end of the observation window. We assume $X_i \in N(0, 1)$.

For the simulations we start by assuming the baseline hazard function takes the following form.

$$\lambda_0(a) = 0.002\frac{1 + 2\phi(log(a), \mu = 1.8, \sigma = 1)}{0.4} \tag{12}$$

This form was chosen to create a hazard function consistent with our experience in retail lending. Figure 1 shows the resulting function.

The credit risk functions are generated via an AR(1) process, $y(v)$. The final function is generated by rescaling to a common range by dividing by the estimated standard deviation, $\sigma_y$.

$$y(v) = y(v-1) + \epsilon \tag{13}$$

$$y(0) = 0 \tag{14}$$

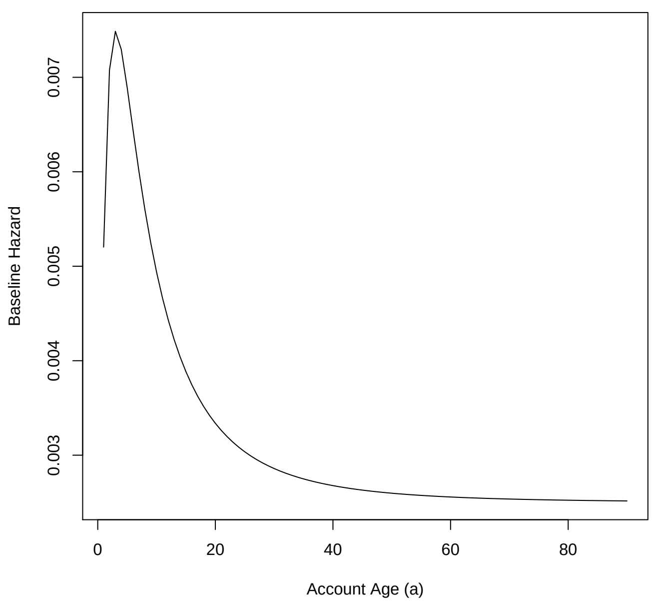


Figure 1: The baseline hazard function from Equation 12 used in generating test data.

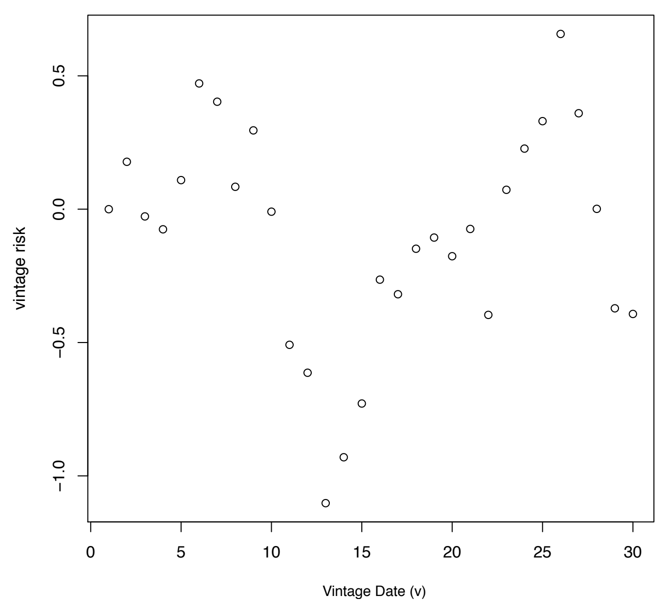


Figure 2: An example of one of the AR(1) generated vintage functions.

$$G(v) = y(v)/\sigma_y \tag{15}$$

The environment function $H(t)$ is generated by assuming a cyclical base function with additive noise. Because of the importance of linear trends to the current analysis, we test three different environment functions. The first is half a cycle, where a polynomial fit would show a strong linear component. The second is a full cycle which has no trend in the cycle but may exhibit a small linear trend depending upon the additive noise. The third is with two full cycles, where given the same noise level will exhibit less linear trend statistically.

The environment function for the full economic cycle is generated as

$$H(t) = sin(2\pi t) + 0.1\epsilon \tag{16}$$

Examples of half and full cycle environment functions are shown in Figure 3.

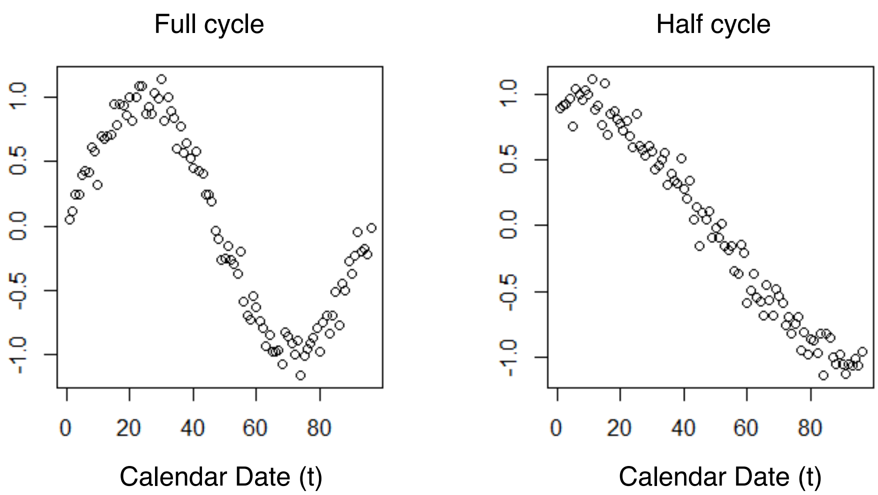


Figure 3: An example of one of the environment functions simulating either a full (left) or half (right) an economic cycle.

Given these functions and the loan-level idiosyncratic factors $X_i$, we generate the probability of default each month for each loan. Default thresholds are randomly generated for each loan $i$, $p_i \in [0, 1]$. The age at which the loan exceeds that threshold is recorded as the survival age. Loans that are still below the threshold by the end of the simulation time interval are recorded as right censored.

## 3.2 Cox PH parameter estimation

With four available components (hazard function versus $a$, credit risk versus $v$, environment versus $t$, and loan-level idiosyncratic effect $X_i$) we tested a range

| Model | $X_i$ | $G(v)$ | $H(t)$ |
|---|---|---|---|
| 1 $X_i$ | 0, 2, 0 | - | - |
| 2 $X_i$ & $G(v)$ | 0, 2, 0 | 2, 0, 0 | - |
| 3 $X_i$ & $H(t)$ | 0, 2, 0 | - | 1, 51, 25 |
| 4 $X_i$ & $G(v)$ & $H(t)$ | 0, 2, 0 | 4, 2, 1 | 1, 46, 24 |

Table 1: Results of 100 simulations of 10,000 accounts using the half-cycle environment function $H(t)$ as shown in Figure 3. The number of parameter failures are shown for the $X_i$, $G(v)$, and $H(t)$ components. Each table cell has the number of low estimates and high estimates at the 1% significance level, and the overall number of Bonferroni-corrected failures at the 1% level.

of combinations to see which would be estimated most easily by the Cox model. The hazard function $\lambda_0(a)$ and idiosyncratic effects $X_i$ were included in all simulations, but we tested including or excluding the time and vintage components.

For each simulated data set, we tested four possible models.

1. $\lambda(a, z_i(v,t), \beta) = \lambda_0(a)\exp(\beta_1 X_i)$
2. $\lambda(a, z_i(v,t), \beta) = \lambda_0(a)\exp(\beta_1 X_i + \beta_2 G(v))$
3. $\lambda(a, z_i(v,t), \beta) = \lambda_0(a)\exp(\beta_1 X_i + \beta_3 H(t))$
4. $\lambda(a, z_i(v,t), \beta) = \lambda_0(a)\exp(\beta_1 X_i + \beta_2 G(v) + \beta_3 H(t))$

Although the hazard function $\lambda_0(a)$ is implicit in the Cox PH model, it is not directly estimated. For our tests, we focused on the estimates of $\beta$ as provided by Cox PH. For each generated data set, we estimated the coefficients for the four different models. The null hypothesis that the estimation was equal to the true value was tested at the 1% confidence level.

The columns below show the test results separately for the parameters $\beta_1$, $\beta_2$, and $\beta_3$ of $X_i$, $G(v)$, and $H(t)$ respectively. In each case, we record three values. When the estimate was low, high, and the overall number of failures using the Bonferroni correction [8] so that the overall test result is also significant to the 1% level.

Table 1 shows the results for 100 simulations using half a cycle in the environment function, Table 2 has the results for the full-cycle environment function, and Table 3 has the results for an environment function with two full economic cycles.

## 3.3 Interpretation

Some features of the tabulated results are obvious. The data with half an economic cycle creates problems for parameter estimation. When we attempted to estimate the coefficient for the environment function used to generate the data, we failed half the time. When the data generation captures a full economic cycle, the failure rate drops to single digits. Expanding to two economic cycles drops the failure rate even further.

| Model | $X_i$ | $G(v)$ | $H(t)$ |
|---|---|---|---|
| 1 $X_i$ | 1, 0, 0 | - | - |
| 2 $X_i$ & $G(v)$ | 1, 1, 0 | 10, 9, 5 | - |
| 3 $X_i$ & $H(t)$ | 0, 1, 0 | - | 1, 14, 2 |
| 4 $X_i$ & $G(v)$ & $H(t)$ | 0, 1, 0 | 3, 1, 0 | 0, 3, 0 |

Table 2: Results of 100 simulations of 10,000 accounts using the full-cycle environment function $H(t)$ as shown in Figure 3. The number of parameter failures are shown for the $X_i$, $G(v)$, and $H(t)$ components. Each table cell has the number of low estimates and high estimates at the 1% significance level, and the overall number of Bonferroni-corrected failures at the 1% level.

| Model | $X_i$ | $G(v)$ | $H(t)$ |
|---|---|---|---|
| 1 $X_i$ | 1, 0, 0 | - | - |
| 2 $X_i$ & $G(v)$ | 0, 2, 0 | 4, 3, 0 | - |
| 3 $X_i$ & $H(t)$ | 0, 1, 0 | - | 1, 4, 1 |
| 4 $X_i$ & $G(v)$ & $H(t)$ | 0, 2, 0 | 1, 0, 0 | 1, 1, 0 |

Table 3: Results of 100 simulations of 10,000 accounts using the double-cycle environment function $H(t)$ which duplicates the structure shown in Figure 3. The number of parameter failures are shown for the $X_i$, $G(v)$, and $H(t)$ components. Each table cell has the number of low estimates and high estimates at the 1% significance level, and the overall number of Bonferroni-corrected failures at the 1% level.

With any data set, estimating the idiosyncratic factors $X_i$ is always reliable. Only the estimation of systematic effects is problematic.

All of these results are consistent with the earlier discussion about estimability. The idiosyncratic effects are not functions of time or vintage, so no conflict arises. When the generated data includes a strong linear trend, as with the half-cycle function, accurate estimation is not possible, because the algorithm can only accurately estimate the sum of linear trends in $v$ and $t$. As the data set is extended through more economic cycles, the linear component diminishes toward zero, but is never exactly so, due to the added noise.

# 4 Conclusions

The Cox proportional hazards model is not flawed, in the sense that it performs perfectly well in the situations for which it was originally designed. The question of this paper was whether it could perform well in the case of stress testing loan portfolios where we can have factors of origination date (vintage) and calendar date in addition to the usual hazard function with age.

In fact, it was obvious from the outset that some problems must arise, because the Cox model in the dimensions of age, vintage, and time is easily mapped to the Age-Period-Cohort model for which the problems are well studied. Moreover, in discrete time, the Cox PH model is identical to a generalized linear model, which is the same estimator used in APC, except that APC imposes certain constraints in the design matrix in order to achieve estimable parameters.

In the APC literature, the problem of the specification error in linear trend is solved by imposing problem-specific constraints. The real question here was to identify the constraints by which the Cox PH estimator is able to provide unique answers. In addition to the obvious constraint around the constant term, we find that the estimator is only able to correctly determine the sum of the linear trends in vintage and time. Beyond that, the coefficients estimated will depend delicately on the specific input factors, just as happens with colinear factors in a multiple regression equation. Note however that vintage and time factors are not colinear in the traditional sense. They are along different dimensions. It is only because $t = v + a$ that the "colinearity" occurs.

The numerical experiments confirm the difficulty of estimating the original coefficients when linear trends are present. In all cases the Cox PH estimator provided coefficients, just not always the original values because of the implicit constraints did not align with the generation of the data.

The obvious conclusion of this work is that the linear specification error in APC models is not exclusive to those models. Any model that includes factors in age, vintage, and time will have an embedded specification error. Obtaining a unique solution is dependent upon having constraints. When those constraints are not explicitly known, we cannot be certain if the obtained stress test model is either under or over sensitive to economic factors, as is possible in the presence of the linear trend specification problem.

We also learn from APC models that the nonlinear components of the age, vintage, and time functions are fully estimable. For stress test models, this means that if we have enough history, at least one full economic cycle but preferably more, then the desired macroeconomic sensitivities are confined in the nonlinear components. The numerical experiments here confirm that given enough data, all these methods almost always provide effective solutions, because the linear trends are approximately zero.

## References

[1] O.O. Aalen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701.

[2] G Andreeva, J Ansell, and J Crook. Modelling profitability using survival combination scores. *European Journal of Operational Research*, 183(3):1537–1549, 2007.

[3] T. Bellotti and J.N. Crook. Credit scoring with macroeconomic variables using survival analysis. 60:1699–1707, 2009.

[4] Joseph L. Breeden, Lyn C. Thomas, and John McDonald III. Stress testing retail loan portfolios with dual-time dynamics. *Journal of Risk Model Validation*, 2(2):43 – 62, 2008.

[5] N. E. Breslow. Discussion following regression models and life tables by D. R. Cox. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2):187.

[6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187.

[7] D. R. Cox and D. O. Oakes. *Analysis of Survival Data*. Chapman and Hall, London, 1984.

[8] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52.

[9] T. Holford. The estimation of age, period, and cohort effects for vital rates. *Biometrics*, 39(39):311.

[10] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457.

[11] W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945.