AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Counting Processes for Retail Default Modeling

Nicholas M. Kiefer and C. Erik Larson

CREATES Research Paper 2015-17

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Counting Processes for Retail Default Modeling

Nicholas M. Kiefer[1]
&
C. Erik Larson[2]

Final: April, 2015

[1]Departments of Economics and Statistical Science, Cornell University, Ithaca, NY, 14853, US and University of Aarhus, Denmark. email: nicholas.kiefer@cornell.edu.
Kiefer acknowledges support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation
[2]Promontory Financial Group, DC, 20006 US. email: elarson@promontory.com

Abstract

Counting processes provide a very flexible framework for modeling discrete events occurring over time. Estimation and interpretation is easy, and links to more familiar approaches are at hand. The key is to think of data as "event histories," a record of times of switching between states in a discrete state space. In a simple case, the states could be default/non-default; in other models relevant for credit modeling the states could be credit scores or payment status (30 dpd, 60 dpd, etc.). Here we focus on the use of stochastic counting processes for mortgage default modeling, using data on high LTV mortgages. Borrowers seeking to finance more than 80% of a house's value with a mortgage usually either purchase mortgage insurance, allowing a first mortgage greater than 80% from many lenders, or use second mortgages. Are there differences in performance between loans financed by these different methods? We address this question in the counting process framework. In fact, MI is associated with lower default rates for both fixed rate and adjustable rate first mortgages.
JEL Classification: C51, C52, C58, C33, C35
Keywords: Econometrics, Aalen Estimator, Duration Modeling, Mortgage Insurance, Loan-to-Value

1 Introduction

Counting process analysis, also called event-history analysis and life-table analysis, is a general and flexible framework for studying sequences of events. In this setup the state space is discrete, for example credit rating categories or in a simpler case default/nondefault. Time is continuous. Transitions between states can occur at any time and the rate or intensity at which these transitions occur is the quantity being measured. Conceptually, the unit of observation is the stochastic process of events through time for a particular agent or subject. In the binary event case the approach is an alternative way of looking at classical survival analysis. There, attention is focused on the hazard rate, which is exactly the transition rate modeled in counting processes.

Survival analysis itself, in the form of life tables, has a long history of applications. Oakes (2000) gives an historical review of survival analysis. He and Andersen et al (1985) note the contribution of John Gaunt, an early demographer known as the father of demography, whose 1662 book dealt with life tables. The earliest applications were actuarial and demographic. The most recent rapid development of methods was in medical applications, and subsequently economics. Introductions to the statistical literature on survival analysis may be found in texts by Kalbfleisch and Prentice (1980), Lawless (1982) and Cox and Oakes (1984). A short introduction is given by Freedman (2008).

Two features of transition data are that the data are censored and that parametric models are widely suspect as too restrictive. Both of these features are easily addressed in the counting process framework.

In the survival application, censoring means that some of the observations have not failed, so the total lifetime is unobserved. In the more general counting process setup censoring occurs because only a portion of the stochastic process is observed, often ending with an interval of time not ending with a transition. Again, incomplete spell lengths complicated matters. Initial work focused on parametric models and techniques for handling censoring. An early result in the nonparametric analysis of censored data was the Kaplan Meier (1958) product-limit estimator.

The presence of covariates in medical and economic applications also complicates matters. In duration models the effect of covariates is more than the familiar mean shift in linear models. Parametrization thus becomes increasingly important and sometimes suspect. The highly influential partial-likelihood model by Cox (1972) allowed for parametric modeling of the effects of covariates without requiring a parametric assumption for the actual distribution of durations. The model uses an ingenious approach based on ranks and allows for censoring.

The classical approach to survival or duration analysis regards durations as random variables and develops models and methods of analysis from this point of view. Kiefer (1988) and Lancaster (1992) review parametric and nonparametric methods using the random variable approach with a view toward economic applications. The counting process viewpoint in contrast regards the stochastic process of events over time as the unit of observation, and focuses on modeling

these processes. It turns out that this approach leads to natural generalizations of the models used for survival analysis and a unified treatment of estimation and inference procedures. Generalizations include application to recurring events in univariate "survival" models (repeated deaths? - perhaps missed or late payments in a credit application) and extension to multivariate counting processses addressing the problem of competing risks. As to inference, Andersen and Gill (1982) use the counting process approach to establish the asymptotic properties of the Cox estimator and Johansen (1983) uses the counting process formulation to give a maximum-likelihood justification for the Cox estimator. Generally, the asymptotic distribution results are easily obtained using martingale arguments. The counting process approach was suggested by Aalen (1978) and further developed by Andersen (1982). Intuition can be gained by thinking of counting processes as Poisson processes. In general, counting processes amount to Poisson processes on a distorted time scale (the interarrival times need not be exponential). The state of the art is described in Aalen, O., Borgan, O. & Gjessing, H. (2008). The applications there are mostly from healthcare and medicine. The progression of a disease through stages, resulting perhaps in recovery, perhaps in death, is much like the progression of the risk status of a risky investment. Counting processes also occur in physical applications, see Gardiner (2009) and Van Kampen (1992),

Our purpose is to give a complete guide to using the counting process approach for loan-level default analysis. To this end, sections 2-5 define the counting process, give a succinct discussion of the principal results on specification and estimation of counting processes, and describe at a high level how these results are obtained. In the long section 6 the reader is taken step by step through an application to residential mortgage defaults, with emphasis on the difference between fixed and adjustable rate mortgages and a concentration on the effect of mortgage insurance on default rates. Our results demonstrate in the context of a very flexible specification that high loan-to-value mortgages financed using mortgage insurance have lower delinquency rates than those financed using second mortgages.

2 Counting Processes

The key to the counting process approach is to think of data as "event histories," a record of the times of switching between states in a discrete state space. The focus is not on distributions (exponential, weibull, ...) but on the entire stochastic process. The point is to estimate the transition rates between states. Transitions are "events" in the language of counting processes. With 2 states this approach is a small generalization of the usual model for survival. The approach uses the theory of martingales and stochastic integration.

2.1 Counts, Rates and Information

A counting process $N = \{N(t)\}_{(t\geq 0}$ is a stochastic process which counts the occurrences of transitions into a target state as time t evolves. It is often mathematically convenient to normalize to $t \in [0, 1]$. We might think of the cumulative count of missed or late payments over time, or perhaps the incidence of overdrafts on a deposit account. The $N(t)$ are nonnegative integers which have jumps of size +1 at random times.

The observable stochastic process $N(t)$ is generated by an intensity process $\lambda(t)$ which is exactly the hazard function occurring in the ordinary survival model. Let $dN(t)$ be the increment $N(t) - N(t-)$to $N(t)$ where $t-$ is the "instant" before t. Hence $dN(t) \in \{0, 1\}$. Then

$$\lambda(t) = \Pr(dN(t) = 1|\mathcal{F}_{t-}) \tag{1}$$

where $\mathcal{F}_{t-}$ is the information available immediately prior to t.

A multivariate counting process $N = \{N_1(t), N_2(t), ..., N_k(t)\}$ is a stochastic process with k components, which counts the occurrences of transitions into each of k states as time t evolves. Again, the counts $N_j(t)$ are nonnegative integers which have jumps of size +1 at random times. An economic example is the evolution of credit ratings for assets over time (Kiefer, N. M. & Larson, C. E. (2007)). In the multivariate case we make the further restriction that any time, only one component can jump. The observable stochastic process N is similarly generated by an intensity process $\lambda = \{\lambda_1(t), \lambda_2(t), ..., \lambda_k(t)\}$ where the $\lambda_j(t)$ are exactly the origin and destination-specific hazard functions occurring in the competing risk model. Let $dN_j(t)$ be the increment to $N_j(t)$ ($dN_j(t) \in \{0, 1\}$). Then$\lambda_j(t) = \Pr(dN_j(t) = 1|\mathcal{F}_{t-})$ is the multivariate intensity process.

2.2 The Multiplicative Intensity Model

The multiplicative intensity model is intuitive and widely used. The key idea is that the failure probability for a given asset, which can depend on covariates and duration at risk, does not depend on the number of other assets at risk. The intensity function for this model is

$$\lambda_j(t) = \alpha_j(t)Y_j(t) \tag{2}$$

where $\alpha_j(t)$ is a deterministic function and $Y_j(t)$ is a *predictable* stochastic process ($\mathcal{F}_{t-}-$ measurable); its value at t is known at $t - .$ $Y_j(t)$ is interpreted as the number at risk for this transition. In the pure survival case for a single individual, where the event - death - can only occur once, $Y(t) = 1 - N(t-)$, illustrating the property of predictability. Turning to a practical specification, let the transition rate for observation i be

$$\lambda_{ji}(t) = \alpha_j(t, X_i(t), \theta)Y_{ji}(t) \tag{3}$$

where $X_i(t)$ is a vector of predictable time-varying covariates, θ is a vector of parameters. Nonparametric estimation is also practical, in which case θ is a

high dimensional parameter. $Y_{ji}(t)$ is a predictable indicator function ($Y_{ji}(t) \in \{0,1\}$) indicating whether observation i is at risk of transition into state j.

2.3 Censoring

Censoring is a feature of duration data that makes the use of ordinary regression techniques typically inappropriate. Censoring occurs when the time to default is unknown, although elapsed time at risk is known. For example, the asset might not have defaulted at the end of the observation period. Or, the asset might disappear from the sample for other reasons, such as prepayment. The treatment of censoring is easy and elegant in the multiplicative intensity model. Let $N_i^u(t)$ be the uncensored stochastic process of counts (a vector). $N_i^u(t)$ is not observed and has intensity

$$\lambda_{ji}^u(t) = \alpha_j(t, X_i(t), \theta) Y_{ji}^u(t) \tag{4}$$

where $Y_{ji}^u(t)$ is the unobserved predictable indicator sequence.

Censoring is determined by a predictable process $C_i(t)$. This can be deterministic or random and can depend on information up to but not including time t ($\mathcal{F}_{t-}-$ measurable). This formulation allows censoring at fixed times, for example the sample ends at the end of the observation period and those histories are subject to analysis. Alternatively, completely random censoring is allowed. A generalization allows censoring to depend on the observed $\mathcal{F}_{t-}$, which is usually the sample path to date (including the regressor path). Thus the censored observations are subject to the same intensity to default as the other observations – they are not discernibly different from other loans before they censor. What the assumption of predictibility does not allow is censoring depending on information available to the bank or borrower but not the analyst. For example, suppose the loan leaves the portfolio and is therefore censored because the bank, anticipating problems based on private information, decides to sell the loan. Then, that kind of censoring is informative about the loan's quality. That should be modeled as a competing risk, not as ordinary censoring.

Define $C_i(t) = 1$ for uncensored times, 0 for censored. Then the counting process with censoring is

$$N_i(t) = \int_0^t C_i(v) dN_i^u(v) \tag{5}$$

The at-risk indicator for the censored process is $Y_i(t) = Y_i^u(t) C_i(t)$. The intensity for the censored process is

$$\lambda_{ji}(t) = \alpha_j(t, X_i(t), \theta) Y_{ji}(t) \tag{6}$$

just as for the uncensored process.

3 Martingales

A martingale is a stochastic process – a sequence of random variables evolving over time – that, while quite general, has properties allowing simple and unified

treatment of estimation and inference for counting processes. A martingale has the property that the expectation of a future value given the current and past values is equal to the current value. The formalities can be skipped by those only interested in the applications. Formally, a continuous time stochastic process $X(t)$ is a martingale with respect to its past if $E(|X(t)|) < \infty$ and $E(X(t)|\{X(u\}, u \leq s) = E(X(t)|\{X(s\}) = X(s)$.See especially Jacobsen, M. (1989).

Note that $dN(t)$ is a (vector) 0-1 random variable (for dt small) with $E(dN(t))|\mathcal{F}_{t-}) = \lambda(t)dt$. Define the stochastic process $M(t)$ by $M(0) = 0$ and

$$dM(t) = dN(t) - \lambda(t)dt \tag{7}$$

Then $E(dM(t)|\mathcal{F}_{t-}) = 0$ and so upon integrating we have the stochastic process

$$\begin{aligned} M(t) &= N(t) - \int_0^t \lambda(u)du \\ &= N(t) - \Lambda(t) \end{aligned} \tag{8}$$

defining the function $\Lambda(t)$ as the integrated hazards. This process is a *martingale.* More generally, and relevant for our consideration of censoring and ultimately of regressors, $X(t)$ is a martingale with respect to the history $\mathcal{F}_{\rm s}$ if it is *adapted* to the history (i.e. $X(t)$ is $\mathcal{F}_t-$ measurable and $E(X(t)|\mathcal{F}_{\rm s}) = X(s)$. $\Lambda(t)$ is called the *compensator* in the language of stochastic processes. This yields the Doob-Meyer decomposition of the process $N(t) : N(t) = \Lambda(t) + M(t)$. Indexing by sample size, martingale central limit theorems give conditions implying that $M^n(t)$ converges to a gaussian martingale. Gaussian martingales have continuous sample paths and deterministic conditional variance processes and so are easy to work with. The variance process for $dM(t)$ is

$$\begin{aligned} V(dM_j(t)|\mathcal{F} &= V(dN_j(t)|\mathcal{F}_{t-}) \\ &= \lambda_j(t)dt(1 - \lambda_j(t)dt) \approx \lambda_j(t)dt \end{aligned} \tag{9}$$

So,

$$V(M(t)|\mathcal{F}_{t-}) \approx \Lambda(t). \tag{10}$$

Studying the properties of estimators will involve integration with respect to this martingale. For a predictable process $H(t)$ define a new process by the stochastic integral

$$M_H(t) = \int_0^t H(u)dM(u) \tag{11}$$

$M_H(t)$ is also a martingale (since H is deterministic and $E(dM) = 0$) and, since $Var\{H(t)dM(t)|\mathcal{F}_{t-}\} = H^2(t)V(d(M(t))$, $M_H(t)$ has variance process

$$V(M_H(t)) = \int_0^t H^2(u)V(dM(u)) \tag{12}$$

4 Parametric Models

Many familiar statistical applications rely on parametric models. These are models that depend on a fixed, and small (when competently applied) number of parameters. This formulation allows easily interpretable models and models that can be estimated with smaller data sets than truly nonparametric models. However, parametric models can lead to error if the chosen parametrization is not a realistic description of the data. For parametric models we specify a functional form for the hazard functions. In

$$\lambda_j(t) = \alpha_j(t)Y_j(t) \tag{13}$$

we might choose a model for $\alpha_j(t)$. For example the Weibull $\alpha_j(t,\gamma) = \gamma t^{\gamma-1}$. Let T_i be the survival times (times to transition or censoring) and let $d_i = 1$ if the observation is not censored (0 if censored).

We consider estimation by maximum likelihood. The likelihood function is the joint density of the data regarded as a function of the unknown parameters. The values of the parameters that maximize this function are the maximum likelihood estimators. Subject to conditions that hold in our models, the maximum likelihood estimator is consistent (tends toward the true parameter values as the sample size increases), asymptotically normally distributed with a variance that can be easily estimated, and efficient (minimum variance among consistent and asymptotically normally distributed estimators). The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \alpha(t_i,\theta)^{d_i} \exp\{-\int_0^{t_i} \alpha(t_i,\theta)\} \tag{14}$$

with log

$$\ln L(\theta) = l(\theta) = \sum_{i=1}^{n} d_i \ln \alpha(t_i,\theta) - \sum_{i=1}^{n} \int_0^{t_i} \alpha(t_i,\theta) \tag{15}$$

This can be simplified to

$$l(\theta) = \int_0^1 \ln \alpha(t,\theta) dN(u) - \int_0^1 \alpha(t,\theta)Y(u)du \tag{16}$$

and the MLE $\hat{\theta}$ is the solution to $\partial l(\theta)/\partial\theta = 0$ where

$$\partial l(\theta)/\partial\theta = \int_0^1 \partial \ln \alpha(t,\theta)/\partial\theta dN(u) - \int_0^1 \partial\alpha(t,\theta)/\partial\theta Y(u)du \tag{17}$$

The asymptotic distribution of the MLE can be derived using the martingale representation

$$\partial l(\theta)/\partial\theta = \int_0^1 \frac{\partial\alpha(t_i,\theta)/\partial\theta}{\alpha(t_i,\theta)} dM(u) \tag{18}$$

a stochastic integral against the stochastic process dM. It can be shown that $\sqrt{n}(\hat{\theta}-\theta) \to N(0, -I(\theta)^{-1})$ where $I(\theta) = \partial^2 l(\theta)/\partial\theta^2$ as usual. The loglikelihood

function for the multi-state (competing risks) model is straightforward

$$l(\theta) = \sum_{j=1}^{k} \int_0^1 \ln \alpha_k(t,\theta) dN_j(u) - \sum_{j=1}^{k} \int_0^1 \alpha(t,\theta) Y_j(u) du \tag{19}$$

with FOC defining the MLE again a stochastic integral involving dM_j.

When there are covariates $X_i(t)$ for each observation i varying over time we cannot simply sum up the counts $N(t)$ and "at risks" $Y(t)$. Instead we focus on a particular transition (drop j) and write

$$\lambda_i(t, X_i(t), \theta) = \alpha(t, X_i(t), \theta) Y_i(t) \tag{20}$$

for the intensity for the $i-th$ observation in the general case. Here $Y_i(t) \in \{0,1\}$ indicating whether or not the $i-th$ observation has been censored as of time t (0 is censored). This can be analyzed directly, but it is much more convenient to use

$$\alpha(t, X_i(t), \theta, \beta) = \alpha(t,\theta) g(X_i(t)\beta) \tag{21}$$

where the hazard and the effect of covariates are separated. Further, the covariates enter in an index form, i.e., only through the linear combination $X_i(t)\beta$. In a parametric specifiction the function $g : R \to R^+$ is specified, perhaps up to parameters. It can be estimated nonparametrically. Even more convenient (and in frequent use) is

$$\alpha(t,\theta) \exp\{X_i(t)\beta\} \tag{22}$$

Specializing to this case (similar results hold for the general model) we find the loglikelihood

$$\begin{aligned} l(\theta,\beta) = &\sum_{i=1}^{n} \int_0^1 (\ln(\alpha(u,\theta) + X_i(u)\beta) dN(u) \\ &- \sum_{i=1}^{n} \int_0^1 \alpha(t,\theta) \exp\{X_i(t)\beta\} Y_i(u) du \end{aligned} \tag{23}$$

The FOC defining the MLE $(\hat{\theta}, \hat{\beta})$ are

$$\partial l(\theta)/\partial\theta = \sum_{i=1}^{n} \int_0^1 \frac{\partial\alpha(t,\theta)/\partial\theta}{\alpha(t,\theta)} dM_i(u) = 0 \tag{24}$$

and

$$\partial l(\beta)/\partial\beta = \sum_{i=1}^{n} \int_0^1 X_i(u) dM_i(u) = 0 \tag{25}$$

Note that the compensator component of $dM(t) = dN(t) - d\Lambda(t)$ depends on both θ and β (and X).

The usual results on the properties of MLEs hold, asymptotic normality, asymptotic variance $-I(\theta,\beta)$, which can be evaluated at the MLEs. Let

$$A(t) = \int_0^t \alpha(u) du \tag{26}$$

be the integrated intensity (differing from $\Lambda(t)$ by the at-risk factor $Y(t)$). Then following Aalen, note that since $\alpha(t)Y(t)$ is the compensator for $N(t)$, an estimator $\hat{A}(t)$ can be defined:

$$\hat{A}(t) = \int_0^t J(u)/Y(u)dN(u) \tag{27}$$

where $J(t)$ is an indicator with $J(t) = 0$ if $Y(t) = 0$ and $0/0$ is treated as 0. This is the *Aalen* or *Nelson-Aalen* estimator. It is a natural estimator, namely the cumulative sum of the inverse of the number at risk at each observed failure time. Let $\{t_f\}$ be the failure times. Then

$$\hat{A}(t) = \sum_{i|t_i \leq t}^{n} 1/Y(t_i). \tag{28}$$

$J(t)$ is included to allow observations to move out and back in to the risk set before failure or censoring. This is rare in applications. For simplicity, assume it doesn't occur. Then

$$\hat{A}(t) - A(t) = \int_0^t (Y(u))^{-1} dM(u) \tag{29}$$

a mean-zero martingale. Since $Y(t)$ is a predictable process we also have

$$V(\hat{A}(t) - A(t)) = \int_0^t (Y(u))^{-2} dN(u) \tag{30}$$

The Nelson-Aalen estimator has been derived as a maximum likelihood estimator by Johanssen (1983). It has the usual (asymptotic) optimality properties associated with regular MLEs, i.e., minimum variance among CUAN estimators. It is perhaps somewhat easier to interpret the associated survivor function

$$\hat{S}(t) = \exp\{-\hat{A}(t)\}. \tag{31}$$

5 Semi-Parametric Models

Semi-parametric models provide a compromise between parametric and fully nonparametric models. Typically, the parametric part is the part we understand well and is of particular interest. For example, the effect of interest rates on mortgage default might be parametrized. The nonparametric part is one that we do not know much about. Often, our major concern is to avoid restrictive parametrizations which might bias our inference on the parameters of primary focus. For example, the baseline hazard rate might be specified nonparametrically. Our general parametric model is

$$\lambda_i(t, X_i(t), \theta) = \alpha(t, X_i(t), \theta) Y_i(t) \tag{32}$$

To make it semiparametric, separate the baseline hazard from the effect of regressors

$$\alpha(t, X_i(t), \beta) = \alpha(t) g(X_i(t), \beta) \tag{33}$$

The parameters are now $\alpha(t)$ (the nonparametric part) and β, the parametric effect of regressors. It simplifies matters to adopt an index specification, so that $X_i(t)$ affects g through the value of the index $X_i(t)\beta$

$$\alpha(t, X_i(t), \beta) = \alpha(t) g(X_i(t)\beta) \tag{34}$$

and simplifying further

$$\alpha(t, X_i(t), \beta) = \alpha(t) \exp\{X_i(t)\beta\} \tag{35}$$

This is now Cox-regression generalized to allow time-varying covariates. The hazard function is

$$\lambda(t, X_i(t), \beta) = \alpha(t) \exp\{X_i(t)\beta\} Y_i(t) \tag{36}$$

Using the same logic as in the fixed-covariate case, the probability that observation i is the first failure, at time t_1, is

$$\frac{\lambda(t_1, X_i(t_1), \beta)}{\sum_{j=1}^{n} \lambda(t_1, X_j(t_1), \beta)} = \frac{\exp\{X_i(t_1)\beta\} Y_i(t_1)}{\sum_{j=1}^{n} \exp\{X_j(t_1)\beta\} Y_j(t_1)} \tag{37}$$

Note that typically $Y_j(t_1) = 1$ for all j.

This forms the basis for the log partial likelihood function $l(\beta, 1)$ where

$$l(\beta, t) = \sum_{i=1}^{n} \int_0^t X_i(u)\beta dN_i(u) - \int_0^t \ln(\sum_{i=1}^{n} \exp\{X_i(u)\beta\} Y_i(u)) \sum_{i=1}^{n} dN_i(u) \tag{38}$$

and the estimator $\hat{\beta}$ is defined by $\partial l(\hat{\beta}, 1)/\partial\beta = 0$. $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with mean zero and variance $-nI(\beta)^{-1}$. This result can be obtained using martingale arguments.

The associated "baseline intensity" can be estimated by

$$\hat{A}(t) = \int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{\sum_{i=1}^{n} \exp\{X_i(u)\beta\} Y_i(u)} \tag{39}$$

which reduces to the Nelson-Aalen estimator if there are no regressors.

6 An Application: Analyzing the Default Risk of High LTV Mortgages

During the housing bubble that led up to the US and global financial crisis of 2008, many US borrowers who lacked a 20% down payment used second mortgages obtained at origination ("piggyback" loans) as a way of avoiding private mortgage insurance on a first lien with a higher than 80% loan-to-value (LTV) ratio at origination. In a typical piggyback transaction, a borrower would take out a first mortgage for 80% of the home's value, a second for 10%, and make a 10% down payment. Mortgages with LTVs greater than 80% have traditionally been considered riskier than 20% down mortgages; in fact, Fannie Mae and Freddie Mac have held policies precluding the purchase or securitization of high LTV mortgages unless accompanied by mortgage insurance sufficient to bring the LTV down to 80%. For this reason piggybacks effectively competed with MI in the eyes of borrowers.

First mortgages with a piggyback second were the most prevalent alternative to the use of mortgage insurance over the decade preceding the financial crisis. We analyze the relative default risk empirically in an analysis of the loan-level performance of a sample of 5,676,428 individual residential mortgages originated from 2003 to 2007.[1] The data, provided by Genworth and First American CoreLogic, included several borrower and loan-level characteristics. Serious delinquency was evaluated using a definition corresponding to a loan having ever been 90 or more days past due (or worse) at any given time.

There are a number of important caveats to consider when attempting to extend the following analytical results to the overall population of mortgages. First, and most importantly, the analysis focuses exclusively on loans with <20% down payment (>80% Loan-to-Value), which is only a portion of the first-lien origination market. Loans with LTV in excess of 80% represented approximately 20% of the overall market at the time. Second, the database does not cover 100% of the loan market, as not all servicers are CoreLogic customers. The coverage over the study period is over 60% of loans originated. This reduces both the number of piggyback and insured loans in the dataset, relative to the population. However, the missing servicers were mainly large diversified national-level players, and there is no reason to think that their omission should have a systematic selectivity bias on the representativeness of mortgage types in our dataset. Third, combined loan-to-value (CLTV) is not reported on all loans in the CoreLogic dataset. The definition of a "loan with a piggyback" is a first lien loan with LTV=80 and with reported CLTV >80. This definition serves to reduce the number of piggybacks potentially included in the study, while not reducing insured loans. Finally, certain exclusions had already been applied to the dataset by Genworth. These included excluding records with missing FICO

[1] At the request of Genworth Financial, one of the authors (Larson) conducted an independent study (the 2011 Promontory Insured Loan Study) to assess the relative default performance of piggyback and insured loans. The results in this paper are based on that study.

at origination.

To limit and ensure the comparability of our analysis, loans were also excluded if the following conditions were found: 1) Regional information was missing; 2) The combined loan-to-value (CLTV) was greater than 105%; 3) The mortgage use categorization was that of 'Non Insured, Sold'; or 4) A mismatch existed between the origination date in the dataset and the origination date as calculated from the performance history. These exclusions resulted in a dataset containing 5,492,097 observations.

6.1 Summary Statistics

We begin by illustrating the performance differences though descriptive analysis of severe (ever 90 days-past-due) delinquency rates and through comparison of vintage cumulative delinquency curves.

Table 1 presents the lifetime cumulative delinquency rates corresponding to our performance definition (ever 90 days past due or worse). In all years except for 2003, the calculated piggyback delinquency rates are higher than the insured delinquency rates. The overall bad rate on the analysis dataset was 19.44% for insured loans and 29.09% for piggyback loans.

Table 1: Delinquency Rates by Origination Year

Origination Year	2003	2004	2005	2006	2007	2003-2007
Insured	12.10%	16.15%	20.49%	24.34%	27.75%	19.44%
Non-Insured with Piggback	9.40%	16.18%	27.47%	36.73%	34.80%	29.09%

Table 2 illustrates how delinquency rates increase with Combined Loan-to-Value (CLTV). For the insured mortgages, the CLTV value is the same as the LTV of the first lien; for non-insured mortgages, the CLTV represents the combined LTV of both the first and second (piggyback) liens.

Table 2: Delinquency Rates by CLTV

Combined LTV at Origination	80-85	85-90	90-95	95-100
Insured	16.14%	17.29%	17.57%	21.97%
Non-Insured with Piggback	30.90%	29.77%	21.80%	33.47%

As expected, increasing FICO scores are associated with lower delinquency rates, with piggyback loans having higher delinquency rates in all FICO score bands, as documented in Table 3.

Table 3: Delinquency Rates by FICO Score

Origination FICO	350-619	620-659	660-699	700-719	720-739	740-759	760+
Insured	34.56%	24.29%	18.53%	15.25%	12.47%	9.90%	7.04%
Non-Insured with Piggback	50.05%	46.35%	37.34%	32.83%	28.11%	22.74%	15.77%

Table 4 shows little difference in severe delinquency rates between purchase and refinance purposes for insured loans, while non-insured (with piggyback) loans supporting refinance are significantly riskier than loans supporting a new purchase. These patterns run against the traditional thinking that a loan supporting a new purchase is riskier than one supporting a refinance; however one may need to control for other factors to see the expected relationship in these data.

Table 4: Delinquency by Loan Purpose

Loan Purpose	Purchase	Refinance
Insured	19.76%	18.66%
Non-Insured with Piggyback	26.42%	38.00%

Table 5 illustrates that low documentation loans are more risky than full-documentation loans for both insured and non-insured loans.

Table 5: Delinquency by Documentation Level

Documentation Level	Full	Low
Insured	17.56%	24.70%
Non-Insured with Piggyback	21.07%	33.67%

And finally, Table 6 illustrates the lower delinquency rates for adjustable rate mortgages that are insured, compared to those that are non-insured. The difference is much smaller for fixed rate loans.

Table 6: Delinquency by Rate Type

Rate Type	Fixed Rate	Adjustable Rate
Insured	19.33%	22.45%
Non-Insured with Piggyback	20.15%	41.96%

6.2 Vintage Curves

Vintage curves provide summaries of the performance of insured and piggyback loans. To construct our vintage curves, we plot the cumulative monthly severe delinquency rate over time for loans originated in a given year. For each vintage, we present curves for sub-segments of insured and piggyback loans. We segment using origination FICO (<=620 is SubPrime, >620 Prime) and CLTV (less than or equal to 90% and greater than 90%). The early vintages (2003 through 2005) have 72 months of performance. Vintages 2006 and 2007 have 60 and 48 months of performance, respectively. As shown in Figures 1 and 2, below, for the 2007 vintage, piggyback loans have significantly accelerated and higher lifetime cumulative delinquency. Appendix A presents additional curves.

Figure 1

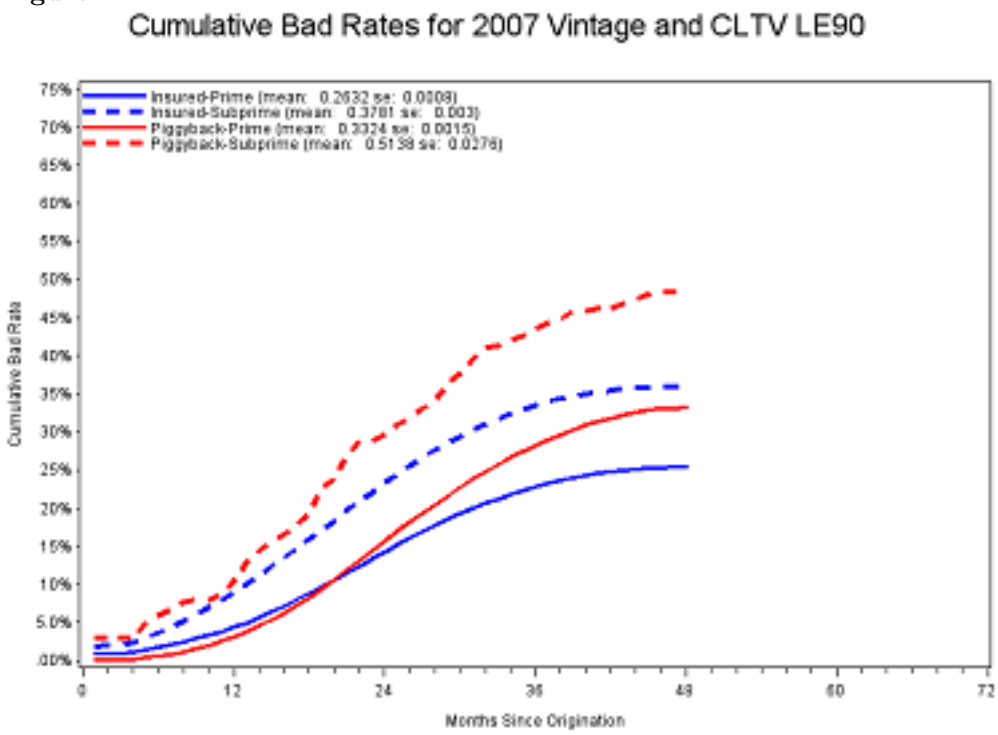
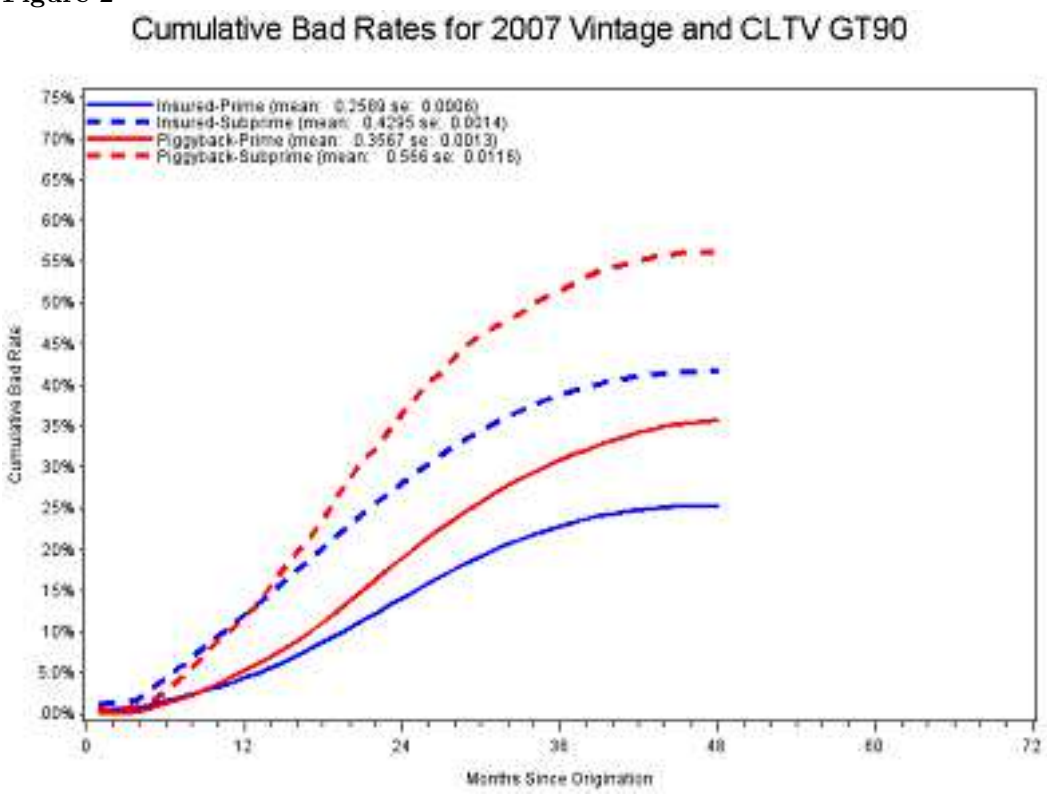

Cumulative Bad Rates for 2007 Vintage and CLTV LE90
Insured-Prime (mean: 0.2632 se: 0.0009)
Insured-Subprime (mean: 0.3781 se: 0.003)
Piggyback-Prime (mean: 0.3324 se: 0.0015)
Piggyback-Subprime (mean: 0.5138 se: 0.0276)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
.00%
0
12
24
36
48
60
72
Months Since Origination

Figure 2



The tables and the vintage curve analysis are both strongly suggestive of differing performance characteristics for insured and non-insured (with piggyback) mortgages. However, it is undoubtedly the case that other risk factors, whose level and impact may differ for insured and non-insured (with piggyback) groups, should be controlled for before any conclusions are drawn or stylized facts established.

For instance, while the vintage curves generally illustrate that non-insured loans with piggyback seconds may have cumulative long-term delinquency rates that are higher than their insured counterparts, the vintage curves do at times cross, with insured loan cumulative severe delinquency rates often being greater during the first 12, and in some instances, first 48 months. This occurs even with vintage curves that attempt to control – albeit weakly – for factors such as origination FICO and CLTV. One potential explanation for this reversal in risk is that differences in payments between the two mortgage types may significantly impact the observed delinquency.

In our dataset, and in the population, insured mortgages overwhelmingly have fixed-rate payment structures, while non-insured (with piggyback) mortgages are almost evenly split between fixed- rate and adjustable-rate payment structures. Since initial rate levels of adjustable-rates loans are usually significantly below those carrying a fixed-rate, and because they remain so for months

or years before any ARM reset, the initial payments for the fixed rate loans are likely to be significantly higher than the adjustable rate loans. Consequently, it would not be surprising if the higher initial payments of fixed rate mortgages (controlling for CLTV) were associated with an initial higher risk of delinquency for insured, predominantly fixed rate, mortgages.

An obvious takeaway is that it will be important to control simultaneously for a potentially large number of risk factors, and to do so in a way that is sensitive to the time varying impact that such factors may have over the life of the mortgage. The dataset allows for one to control for such effects, but the latter requires an appropriate framework to be utilized. We make use of the counting process approach.

6.3 Estimation Results

A Cox Proportional Hazard (PH) Model is used to investigate and quantify the relative performance of piggyback and insured loans while controlling for loan-level factors that are commonly thought to be important in describing loan performance.

The Survival Analysis Modeling Dataset We based our estimates of the stratified proportional hazard model on a modeling dataset consisting of a randomly selected subsample of 538,500 mortgage lifetimes, selected from the parent sample of 5,676,428 individual residential mortgages originated from 2003 to 2007 and provided by Genworth and First American CoreLogic. Summary information is given in table 7.

Table 7: Counts and Dispositions of Observations in the Modeling Dataset

Rate Type	Type	Default	Paid Off	Paying	Total by Rate Type
All Rate Types	Insured	83,641	144,807	203,240	538,500
	Non-insured w/ Piggyback	31,198	33,323	42,291	
Fixed Rate	Insured	73,764	126,260	188,923	452,026
	Non-insured w/ Piggyback	12,774	21,275	29,030	
Adjustable Rate	Insured	9,877	18,547	14,317	86,474
	Non-insured w/ Piggyback	18,424	12,048	13,261	

Appendix B contains additional summary information on loans characteristics in the modeling dataset.

Estimation of Nonparametric (Empirical) Survival Curves Rather than proceeding directly to the estimation of a stratified proportional hazards model, it will be useful to first consider the empirical survival distribution curves for default that are implied by the sample data. To this end, we have constructed

smoothed estimates of the empirical survival function using the method of Kaplan and Meier (1958.) Figures 3 and 4 show the empirical, or non-parametric, estimated default survival curves for insured and non-insured (with piggyback) mortgage loans, computed for subsamples defined by whether the loans were of fixed rate or adjustable rate type. These curves, as do all the estimates presented in this section, focus exclusively on the risk of default, and treat the competing risk of payoff as a censoring event. This approach is a conventional and meaningful way to present results for a risk of interest (here, default) when competing risks are present.

Figure 3. Empirical Survival Curve Estimate, Fixed Rate Loans

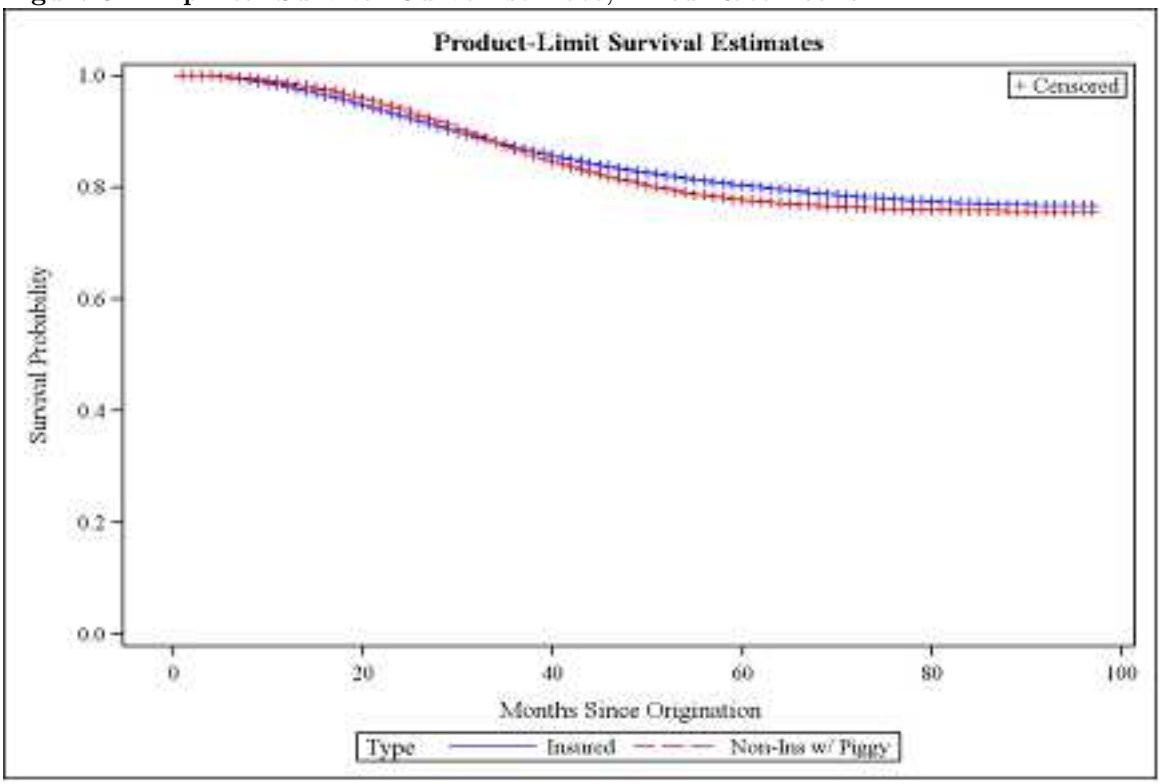
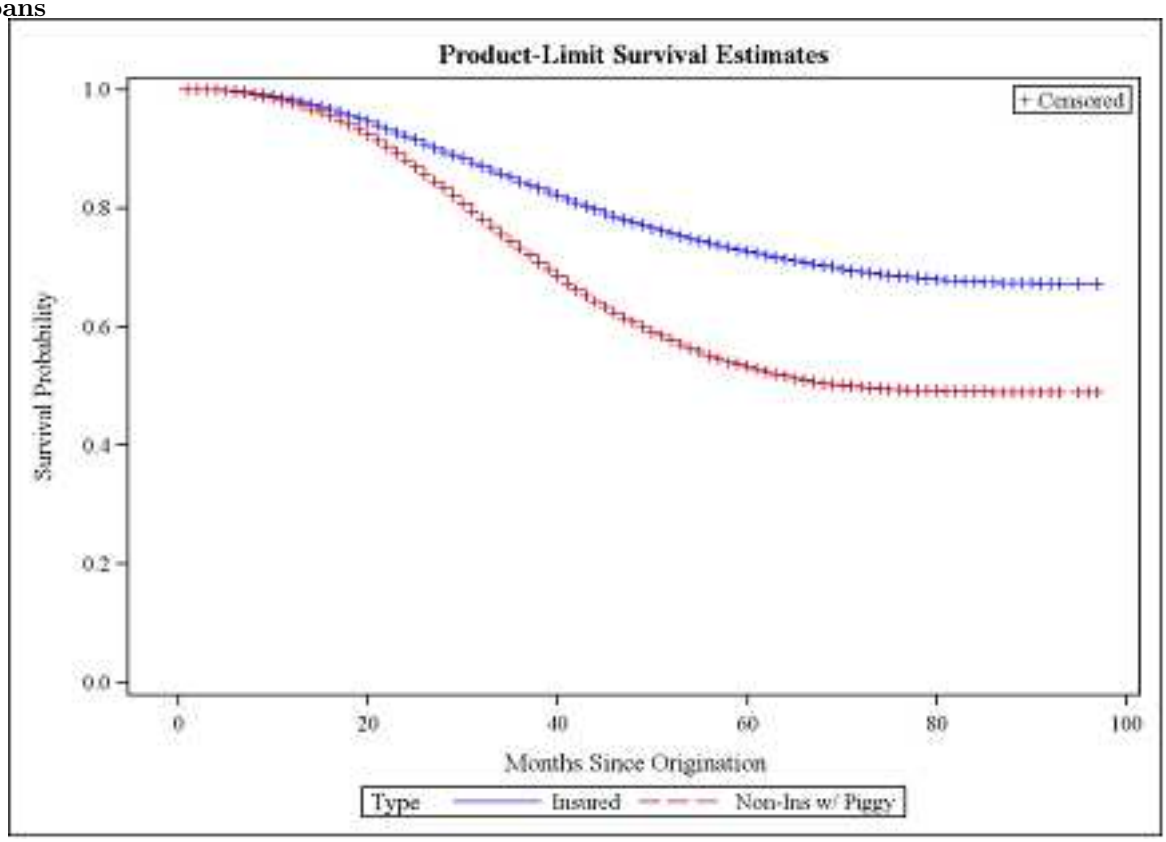


Figure 4. Empirical Survival Curve Estimate, Adjustable Rate Loans



Note that even in the empirical survival curves, the long-term higher default risk associated with non-insured loans having piggyback second liens is easy to identify. This is particularly true for the adjustable rate loans, where the survival proportion for the uninsured mortgages ultimately drops well below that of the insured loans.

Estimation of a Stratified Proportional Hazards Model We are now ready to turn to the estimation of the stratified Cox proportional hazards model. We specify a model in which we include additional covariates and in which we estimate separate stratified models for subsets of our sample, with loans grouped by rate type. Part of the rationale for estimating different models for different rate types (fixed vs. adjustable) is that borrower behavior in response to changes in economic conditions is likely to be very different across these products. Furthermore, differences in mortgage product types or borrower underwriting practices may exist that are unobservable in our data, but which

may result in different magnitudes of the estimated covariate coefficients or in different baseline hazard and survival estimates.

Covariates The covariates in our model include several zero-one categorical (or dummy) variables. For each of these variables, a case that has one of the characteristics is coded as a one, and cases without the characteristic are coded as a zero. These variables include the following

1. Documentation level (low or full documentation, with full documentation = 1);
2. Loan purpose (purchase or refinance, with purchase = 1), and
3. Occupancy status (Owner-occupied or not, with owner-occupied = 1).

The model also includes four continuous variables measured at the time of loan origination:

1. Combined Loan-to-Value;
2. FICO score at origination;
3. Original Interest Rate, and
4. Original Payment, a constructed variable equal to Original Loan Balance X Initial Interest Rate.

Finally, the model includes four time-varying covariates:

1. Interest Rate Differential(t) = Original Interest Rate - Market Interest Rate(t)
2. Change in Payment(t) = [Original Interest Rate - Market Interest Rate(t)] x Original Balance
3. Change in Value(t) = (Original Value) x [%Change in Case-Shiller Index(t)], and
4. Unemployment Rate(t)

The seasonally adjusted civilian unemployment rate and Case-Shiller Index data were matched to each loan based upon MSA/CBSA if available; otherwise a state or national level measure was used, respectively. The market interest rate data was obtained from Freddie Mac, and it was matched based upon the rate type of the loan. Fixed rate loans were matched to the monthly average of the average weekly 30-year rate; adjustable rate loans were matched to the monthly average of the average weekly 1-year rate.

Parameter Estimates Table 8 presents estimation results for the fixed rate and adjustable rate loan group models. Recall that each estimated rate type model has been stratified across insured and non-insured mortgage classes. As a result, we have two sets of parameter estimates, with a given parameter set applying equally to both strata within a given rate group.

The estimated coefficients have signs that are consistent with expectations (recall that due to the proportional hazard specification, a positive parameter indicates that the hazard of default is increasing with the covariate value). Estimated standard errors are extremely small and not reported (all estimates would be judged "significant" at the 0.0001 level by the conventional "test", except the one indicated).

Table 8: Cox Stratified Proportional Hazards Model Parameter Estimates

Loan Type	Fixed Rate	Adjustable Rate
Documentation Level (1=Low)	0.37310	0.76391
Loan Purpose (1=Purchase)	-0.05802	-0.22628
Occupancy Status (1=Owner-Occupied)	-0.14402	-0.38135
Combined LTV at Origination	0.02400	0.03127
FICO Score at Origination	-0.00880	-0.00589
Original Interest Rate	0.21298	-0.12347
Original Payment (Original Int. Rate*Original Balance)	-0.00478	0.01213
Rate Differential (Original Int. Rate - Market Int. Rate)	0.15648	0.09901
Change in Payment (Original Int. Rate - Market Int. Rate)*Original Balance	0.04650	-0.00108**
Change in Value (Original Value)*(%Change in Case Shiller Index)	0.04439	0.02643
Unemployment Rate	0.16021	0.18988

*Note: **Estimate not significantly different from zero. All other estimates are significant at the 0.0001 level.*

Low documentation, non-owner-occupied, high CLTV, and low FICO loans are of greater default risk than loans with the opposite characteristics. Somewhat surprisingly, loans supporting refinancing are of greater risk than loans supporting a new purchase – a result seen in the simple descriptive statistics for this period. The coefficients on the time varying covariates measuring the rate differential between original and current market rates, the change in payment and the change in value are also positive. The greater the difference between the original interest rate and the current market rate, or the greater the different between the original home value and the current implied market value (i.e., the absolute value of potential equity loss), the greater the default risk. Similarly, the higher the current level of unemployment in the MSA or state when the property is located, the higher the default risk. All these impacts are similar across both fixed rate and adjustable rate mortgage groups.

In contrast, when we consider the impact of the level of the original interest rate or the level of the original payment, the signs of the coefficient estimates are reversed between fixed and adjustable rate groups. However, the sign differences make sense: for fixed rate loans, holding original balance constant, higher original interest rates mean higher fixed payments and higher default risk. For adjustable rate loans, the higher original rate probably implies that the risk of a payment shock when the original rate adjusts to market rates is lowered, along with default risk.

Baseline Survival Curve Estimates To illustrate the differences between insured and non-insured loans, it is useful to compare the implied baseline survivor functions for the strata corresponding to our estimated set of models[2]. Figures 4 and 5 shows the implied baseline survival curves resulting from our stratified Cox PH model; estimates reflect the survival probability at month t, evaluated at the mean value covariates across the sample population. Effectively, these baseline survival curve estimates illustrate the fundamental differences in performance between insured and non-insured loan groups, controlling simultaneously and equally for all the effects we have been able to attribute to covariates.

[2]The baseline hazards and survival functions are estimated as arbitrary functions of time through implementation of a restricted maximum likelihood estimation of the $\alpha c(t)$ function, in which the covariates for explanatory variables are restricted to their previously estimated values.

Figure 5. Parametric Baseline Survival Curve Estimates, Fixed Rate Loans

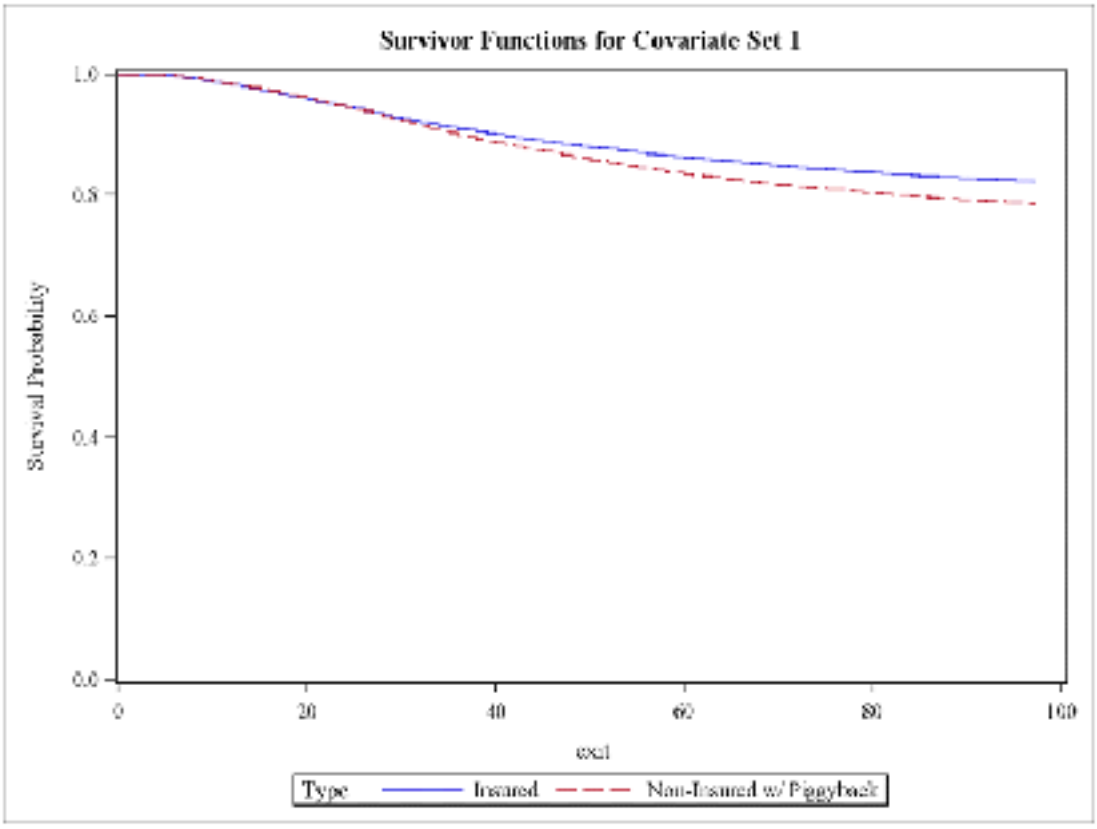
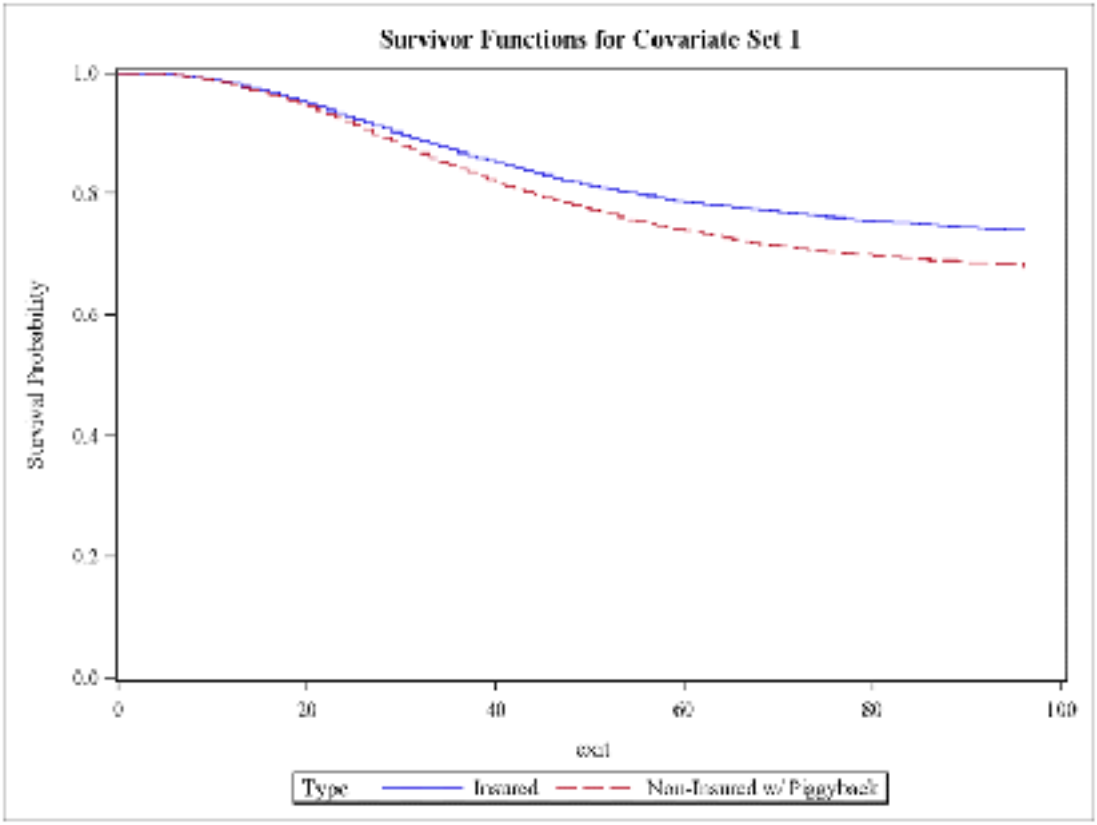


Figure 6. Parametric Baseline Survival Curve Estimates, Adjustable Rate Loans



In these curves, the higher default risk associated with the non-insured (with piggyback) loans is very clear – at times even more so than in the empirical survival curves (which did not control for the effect of covariates). For both fixed rate and adjustable rate mortgages, controlling for the impact of covariates results in implied baseline (strata specific) survival curve estimates in which insured loans continue to demonstrate lower extreme delinquency and default risk than non-insured (with piggyback) loans.

Tables 9 and 10 respectively present the estimated numerical baseline survival rates and cumulative default rates, by strata, for selected months-since-origination. Overall, across both fixed and adjustable rate loans, the proportion of non-insured loans surviving to 72 months was .798, compared to .833 for insured loans. Significantly, as shown in Table 10, this difference implies that the baseline cumulative default rate of non-insured loans is 20.98% percent higher than that of insured loans.

Table 9. Estimated Baseline Survival Rates, S(t)

Proportion Surviving to Selected Months

Rate Type	Type	Months					
		12	24	36	48	60	72
All	Insured	0.983	0.943	0.903	0.873	0.851	0.833
	Non-Insured w/ Piggyback	0.983	0.942	0.890	0.851	0.820	0.798
	Percent Difference (Non-Insured relative to Insured)	*0.04%*	*-0.13%*	*-1.44%*	*-2.52%*	*-3.65%*	*-4.20%*
Fixed Rate	Insured	0.983	0.946	0.910	0.884	0.863	0.846
	Non-Insured w/ Piggyback	0.983	0.946	0.900	0.865	0.835	0.815
	Percent Difference (Non-Insured relative to Insured)	*0.08%*	*0.04%*	*-1.13%*	*-2.15%*	*-3.22%*	*-3.66%*
Adj. Rate	Insured	0.983	0.930	0.869	0.820	0.788	0.767
	Non-Insured w/ Piggyback	0.981	0.920	0.841	0.782	0.740	0.710
	Percent Difference (Non-Insured relative to Insured)	*-0.19%*	*-0.99%*	*-3.16%*	*-4.62%*	*-6.10%*	*-7.32%*

Table 10: Estimated Baseline Cumulative Default Rates, F(t)

Cumulative Proportion Defaulting by Selected Months

Rate Type	Type	Months					
		12	24	36	48	60	72
All	Insured	0.017	0.057	0.097	0.127	0.149	0.167
	Non-Insured w/ Piggyback	0.017	0.058	0.110	0.149	0.180	0.202
	Percent Difference (Non-Insured relative to Insured)	*-2.15%*	*2.09%*	*13.47%*	*17.40%*	*20.79%*	*20.98%*
Fixed Rate	Insured	0.017	0.054	0.090	0.116	0.137	0.154
	Non-Insured w/ Piggyback	0.017	0.054	0.100	0.135	0.165	0.185
	Percent Difference (Non-Insured relative to Insured)	*-4.60%*	*-0.65%*	*11.38%*	*16.32%*	*20.23%*	*20.10%*
Adj. Rate	Insured	0.017	0.070	0.131	0.180	0.212	0.233
	Non-Insured w/ Piggyback	0.019	0.080	0.159	0.218	0.260	0.290
	Percent Difference (Non-Insured relative to Insured)	*10.78%*	*13.11%*	*20.99%*	*21.08%*	*22.66%*	*24.02%*

6.4 Diagnostics: Evaluating the Proportional Hazards Assumption

The assumption of the proportional relationship between hazards and covariates that is implied by the Cox model specification should be subjected to an empirical assessment. To perform such an assessment, it is increasingly common to construct residuals along the lines proposed by Schoenfeld (1982). Instead of a single residual for each individual observation, Schoenfeld's method results in constructing separate residuals for each covariate, for each individual loan, using only those loans that defaulted (were not censored.)

Since the Schoenfeld residuals are, in principle, independent of time, a plot that shows a non-random pattern against time is evidence of violation of the proportional hazards assumption. Appendix C provides plots of the estimated, scaled Schoenfeld Residuals against rank time. The minimal departures from a general, random zero-slope pattern vs. time provide reasonable support for the proportional hazards specification used in our analysis.

7 Conclusions

We propose analyzing defaults at the loan level using an approach based on statistical counting processes. After describing the method, we consider the default experience in a sample of home mortgages. The analysis generally confirms that by controlling for various factors, mortgages with piggyback second lien loans have historically experienced higher lifetime rates of severe delinquency than insured mortgages. This conclusion is supported by descriptive tables, graphical vintage curve analysis and by the results from conducting an analysis using statistical methods of survival analysis based on counting processes.Our results are based on an analysis of a large sample of high loan-to-value mortgages originated over the period 2003 through 2007.

We present the results from estimation from both simple and extended versions of stratified Cox proportional hazards models, the latter estimated across and by US census region. Risk factor parameter estimates are generally in line with expectations as to sign, although variability in the magnitude of estimates exists across regions. We also compare the implied baseline survival curves from the estimated models to smoothed Kaplan-Meier estimates of the empirical survival function. Our modeling approach allows us to produce separate baseline survival estimates for insured and non-insured (with piggyback) mortgages. These baseline curves have been controlled for the impact of risk factors on performance in a way that cannot accomplished by simple tabular or graphical analysis of empirical data

Overall, our analysis supports the assertion that the historical performance of first lien MI-insured loans has been associated with lower rates of extreme delinquency or default, when compared to non-insured first lien loans accompanied by a piggyback second lien, and when controlling for various risk factors. Our results rely on a very flexible statistical specification allowing time-varing

covariates. Our approach is likely to be useful and is certainly feasible in many analyses of loan-level retail default data.

8 References

Aalen, O., Borgan, O. & Gjessing, H. (2008), Survival and Event History Analysis: A Process Point of View, Springer.

Aalen, O. O. (1978), 'Nonparametric inference for a family of counting processes', Annals of Statistics (6).

Andersen, P. K. (1982), 'On the application of the theory of counting processes in the statistical analysis of censored survival data', Lecture Notes-Monograph Series 2, pp. 1– 13.

Andersen, P. K., Borgan, ., Gill, R. D. & Keiding, N. (1996), 'Statistical models based on counting processes. corr. 4th printing'.

Andersen, P. K., Borgan, O., Hjort, N. L., Arjas, E., Stene, J. & Aalen, O. (1985), 'Counting process models for life history data: A review [with discussion and reply]', Scandinavian Journal of Statistics 12:2, 97–158.

Andersen, P. K. & Gill, R. D. (1982), 'Cox's regression model for counting processes: A large sample study', The Annals of Statistics 10(4), pp. 1100–1120. URL: http://www.jstor.org/stable/2240714

Cox ,D.R. (1972) "Regression Models and Life Tables." *Journal of the Royal Statistical Society*, Series B, #34. pp 187-220.

Cox, D. R. (1979), 'A note on the graphical analysis of survival data', Biometrika 66(1), pp. 188–190. URL: http://www.jstor.org/stable/2335265

Cox, D.R., D. Oakes (1984), *Analysis of Survival Data*, London, UK: Chapman and Hall.

Devine, T. J. & Kiefer, N. M., *Empirical Labor Economics: The Search Approach*, Oxford University Press, 1991

Efron, B. (1977), 'The e?ciency of cox's likelihood function for censored data', Journal of the American Statistical Association 72(**??**), pp. 557–565. URL: http://www.jstor.org/stable/2286217

Freedman, D. A. (2008), 'Survival analysis: A primer', The American Statistician (**??**), 110– 119.

Gardiner, C. W. (2009). *Stochastic methods : a handbook for the natural and social sciences.* 4th ed. Berlin: Springer.Hosmer, D.W., Jr.; Lemeshow, S. (1999) *Applied Survival Analysis: Regression Modeling of Time to Event Data*, New York, NY: John Wiley & Sons.

Jacobsen, M. (1989), 'Right censoring and martingale methods for failure time data', The Annals of Statistics 17(3), pp. 1133–1156.
URL: http://www.jstor.org/stable/2241714

Johansen, S. (1983), 'An extension of cox's regression model', International Statistical Review / Revue Internationale de Statistique 51(2), pp. 165–174. URL: http://www.jstor.org/stable/1402746

Kalbfleisch, J.D.; Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, New York, NY: John Wiley & Sons.

Kampen, N. G. van. (1992). *Stochastic processes in physics and chemistry.* Rev. and enlarged ed. Amsterdam: North-Holland.

Kaplan, E.L. ; Meier, P. (1958) "Nonparametric Estimation from Incomplete

Observations." *Journal of the American Statistical Association*, 53, pp. 457-481.

Kiefer, N.M. (1988) "Economic Duration Data and Hazard Functions." *Journal of Economic Literature*, 26(2), pp. 646-679. URL: http://www.jstor.org/stable/2726365

Kiefer, N. M. & Larson, C. E. (2007) "A simulation estimator for testing the time homogeneity of credit rating

transitions." *Journal of Empirical Finance, 14*, 818-835

Lancaster, T. (1992), The Econometric Analysis of Transition Data, Cambridge University Press (Econometric Society Monographs). Lawless, J. F. (1982), Statistical Models and Methods for Lifetime Data, Wiley.

Lawless, J.E. (1982), *Statistical Models and Methods for Lifetime Data*, New York, NY: John Wiley & Sons.

Neumann, G. R. (1999.), Search models and duration data, in M. H. Pesaran & P. Schmidt, eds, 'Handbook of Applied Econometrics Volume II: Microeconomics', Blackwell Publishing.

Oakes, D. (2000), 'Survival analysis', Journal of the American Statistical Association 95(**??**), pp. 282–285. URL: http://www.jstor.org/stable/2669547

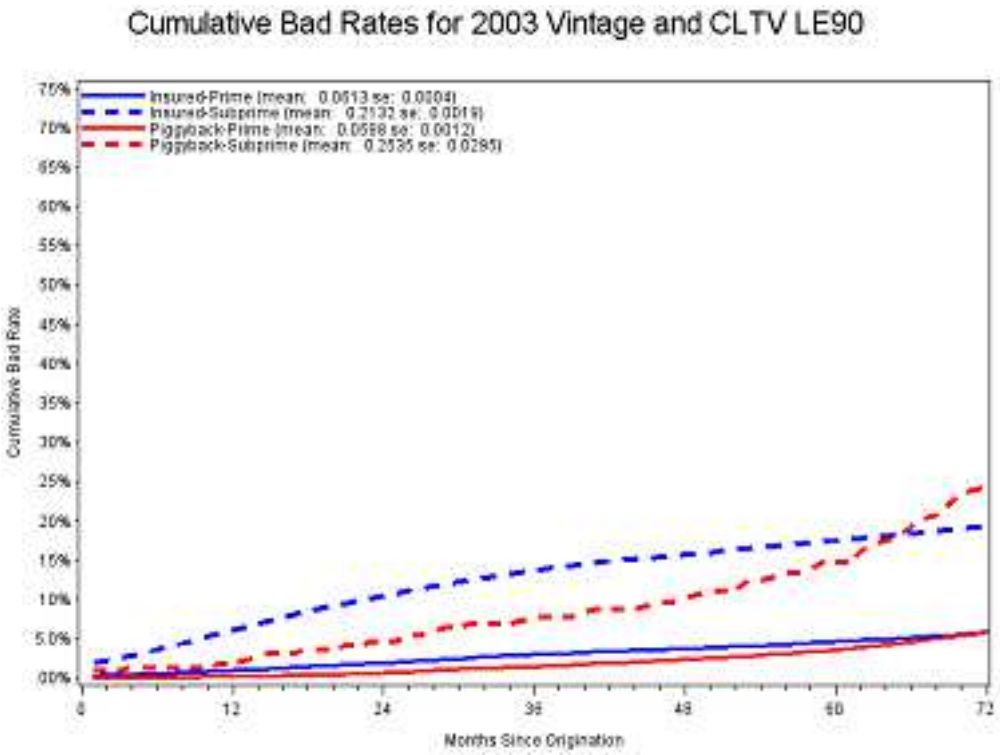
Schoenfeld, D. (1982) "Partial residuals for the proportional hazards regression model." *Biometrika,* 69, pp. 239-241.

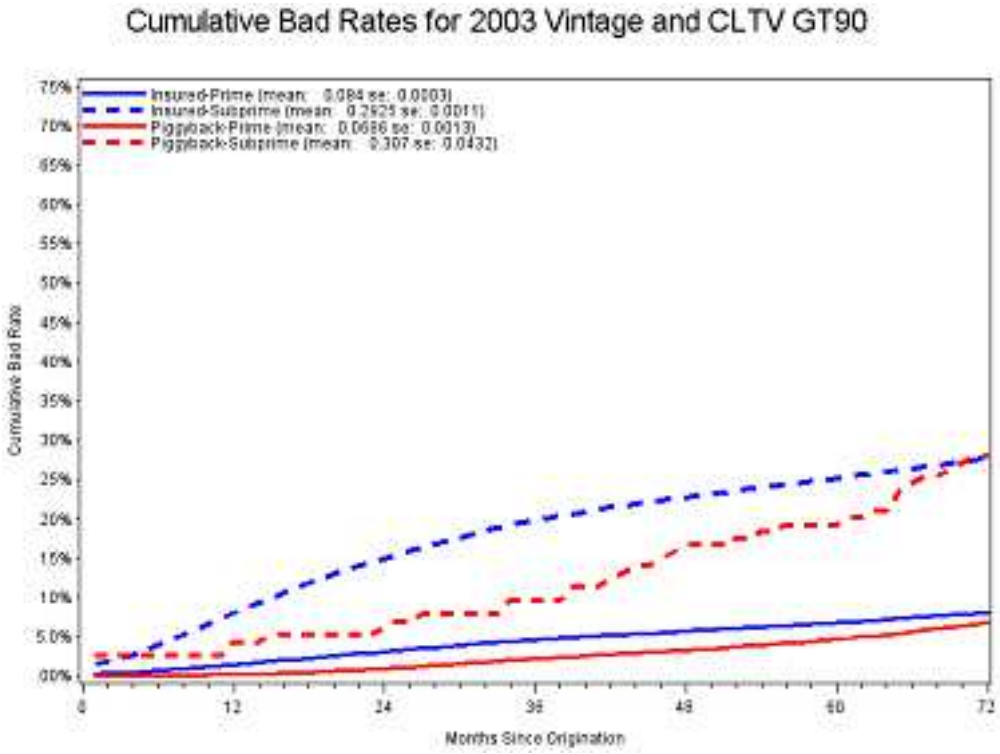
Therneau, T. M.; Grambsch, P. M. (2000) Modeling Survival Data: Extending the Cox Model. New York: Springer-Verlag.

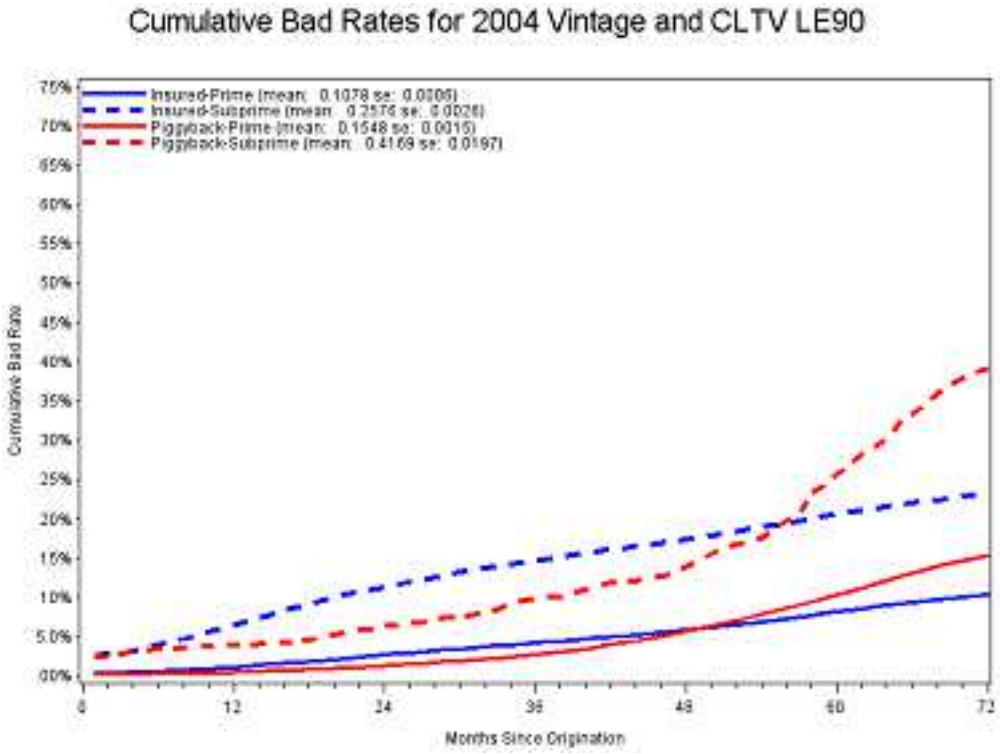
Zucker, D. M. & Karr, A. F. (1990), 'Nonparametric survival analysis with time-dependent covariate e?ects: A penalized partial likelihood approach', The Annals of Statistics 18(1), pp. 329–353. URL: http://www.jstor.org/stable/2241546

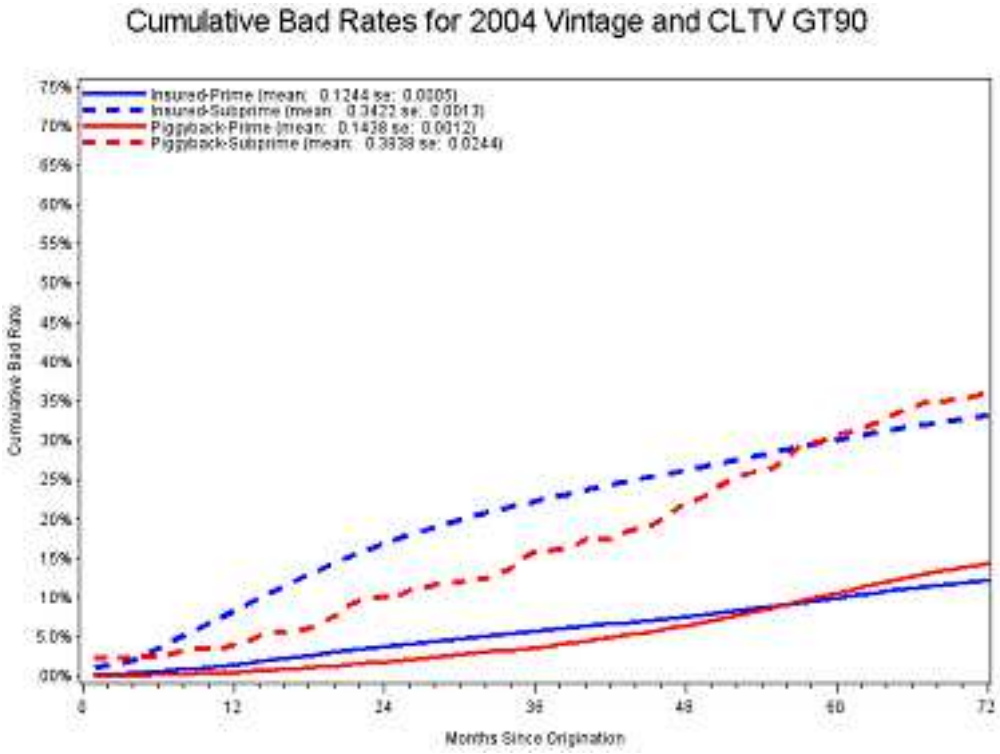
9

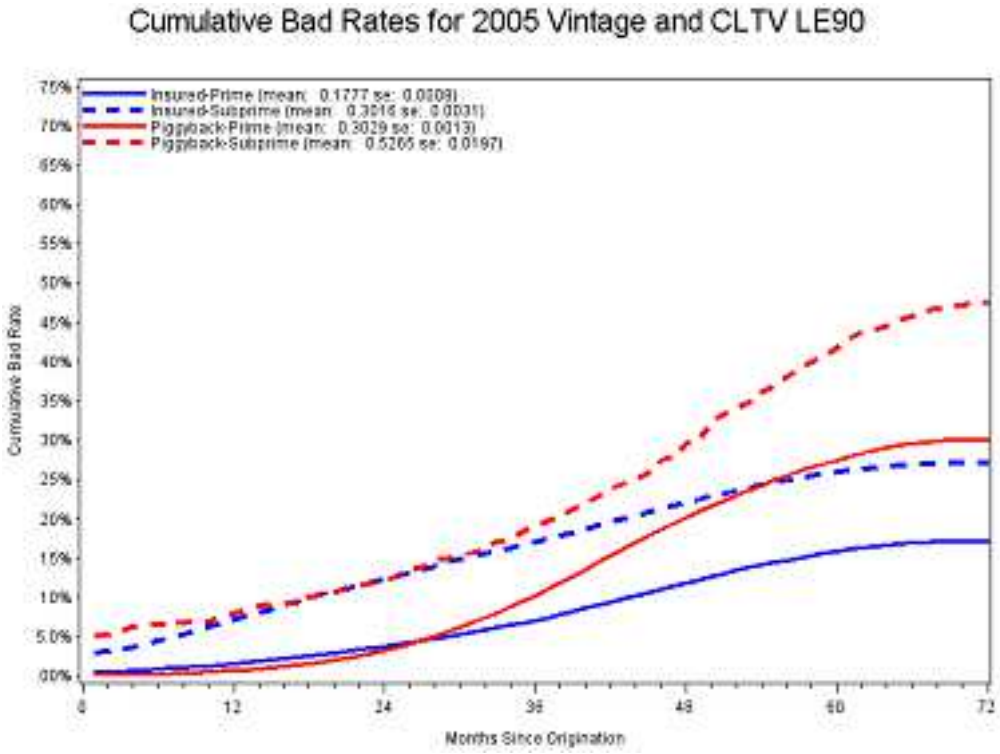
Appendix A: Vintage Curves



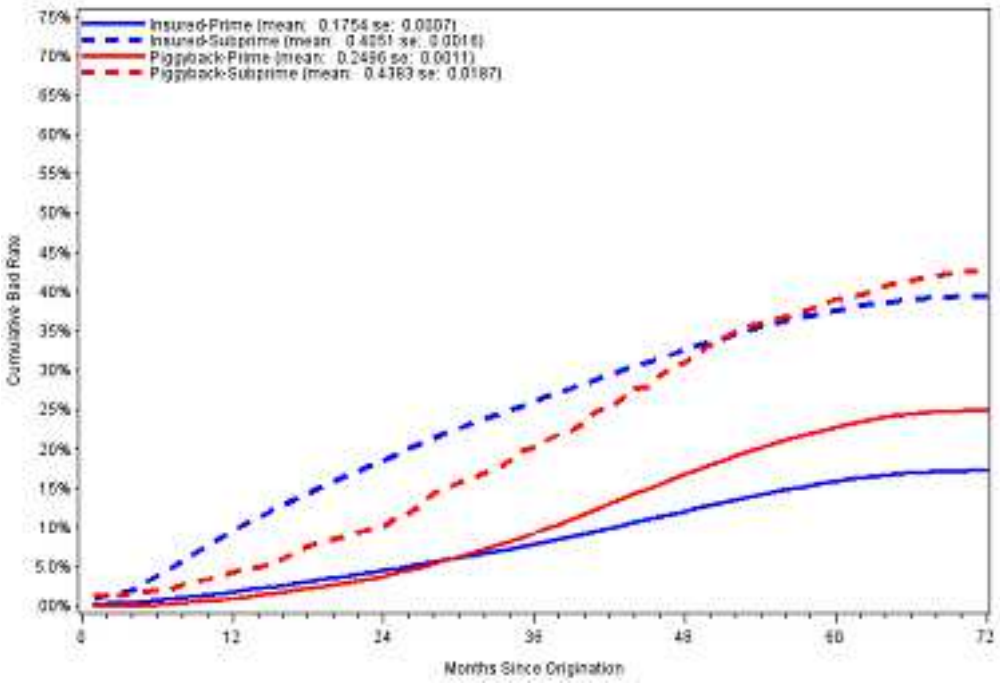

Cumulative Bad Rates for 2003 Vintage and CLTV GT90
Insured-Prime (mean: 0.084 se: 0.0003)
Insured-Subprime (mean: 0.2925 se: 0.0011)
Piggback-Prime (mean: 0.0686 se: 0.0013)
Piggback-Subprime (mean: 0.307 se: 0.0432)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
0.0%
0
12
24
36
48
60
72
Months Since Origination


Cumulative Bad Rates for 2004 Vintage and CLTV LE90
Insured-Prime (mean: 0.1078 se: 0.0005)
Insured-Subprime (mean: 0.2576 se: 0.0026)
Piggback-Prime (mean: 0.1548 se: 0.0015)
Piggback-Subprime (mean: 0.4169 se: 0.0197)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
.00%
0
12
24
36
48
60
72
Months Since Origination

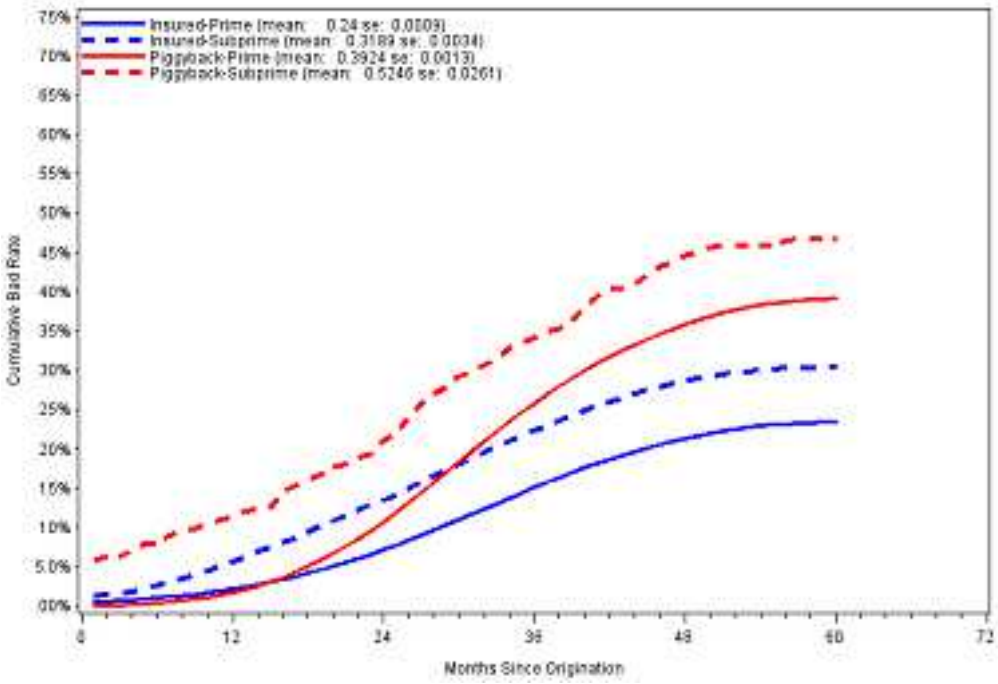

Cumulative Bad Rates for 2004 Vintage and CLTV GT90
Insured-Prime (mean: 0.1244 se: 0.0005)
Insured-Subprime (mean: 0.3422 se: 0.0013)
Piggback-Prime (mean: 0.1438 se: 0.0012)
Piggback-Subprime (mean: 0.3938 se: 0.0244)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
0.0%
0
12
24
36
48
60
72
Months Since Origination

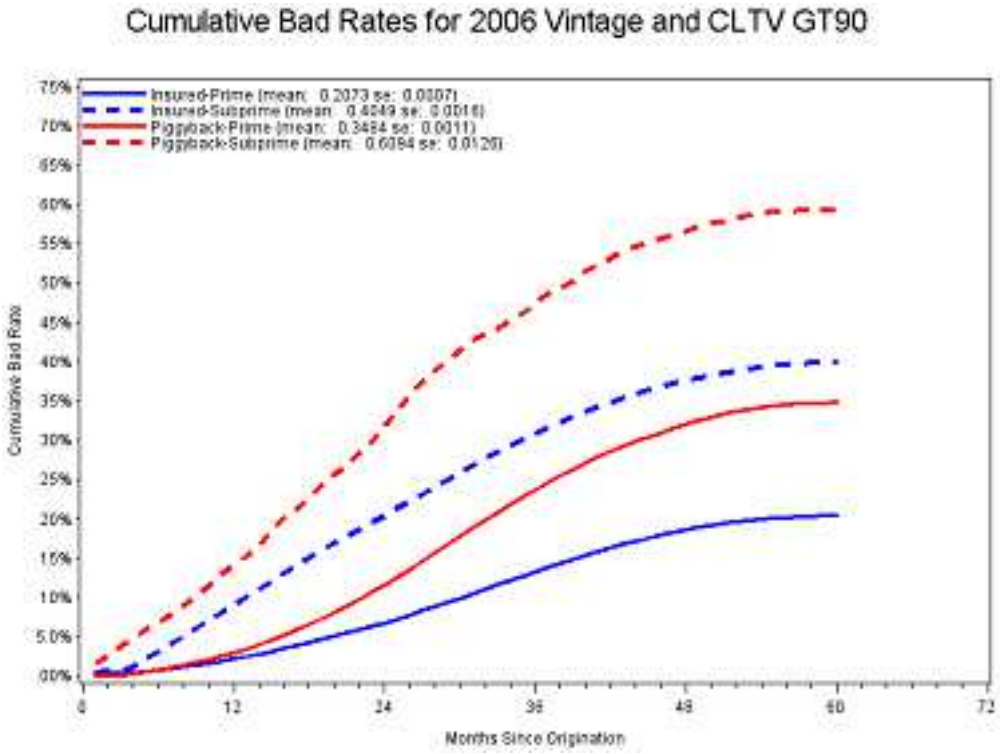

Cumulative Bad Rates for 2005 Vintage and CLTV LE90
Insured-Prime (mean: 0.1777 se: 0.0009)
Insured-Subprime (mean: 0.3016 se: 0.0031)
Piggback-Prime (mean: 0.3029 se: 0.0013)
Piggback-Subprime (mean: 0.5265 se: 0.0197)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
0.0%
0
12
24
36
48
60
72
Months Since Origination

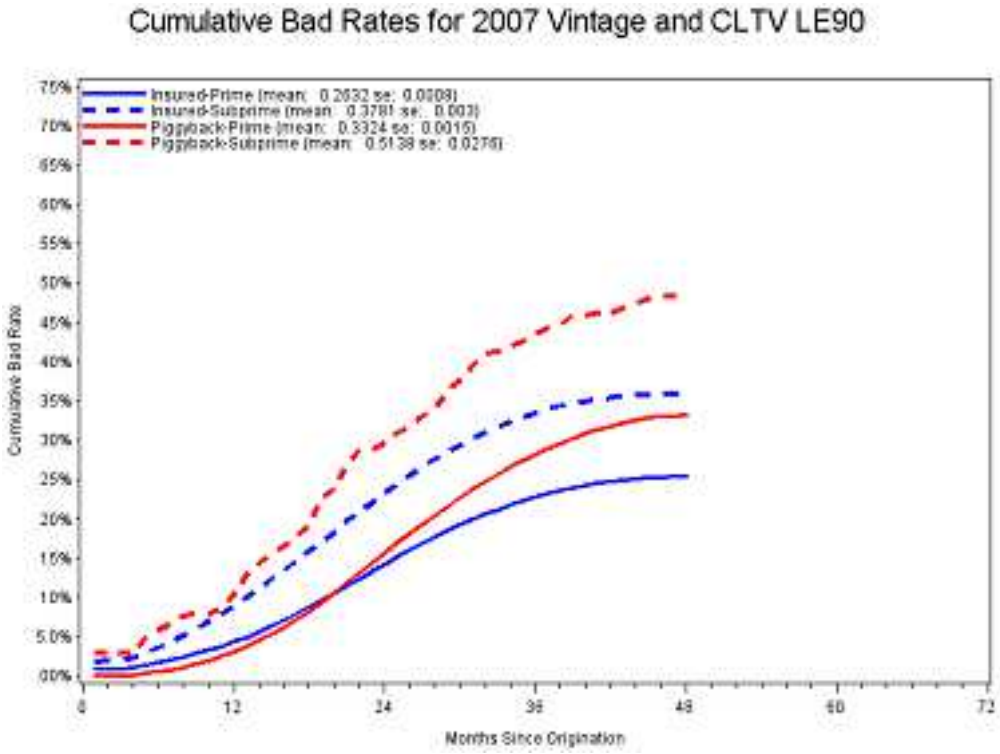
Cumulative Bad Rates for 2005 Vintage and CLTV GT90

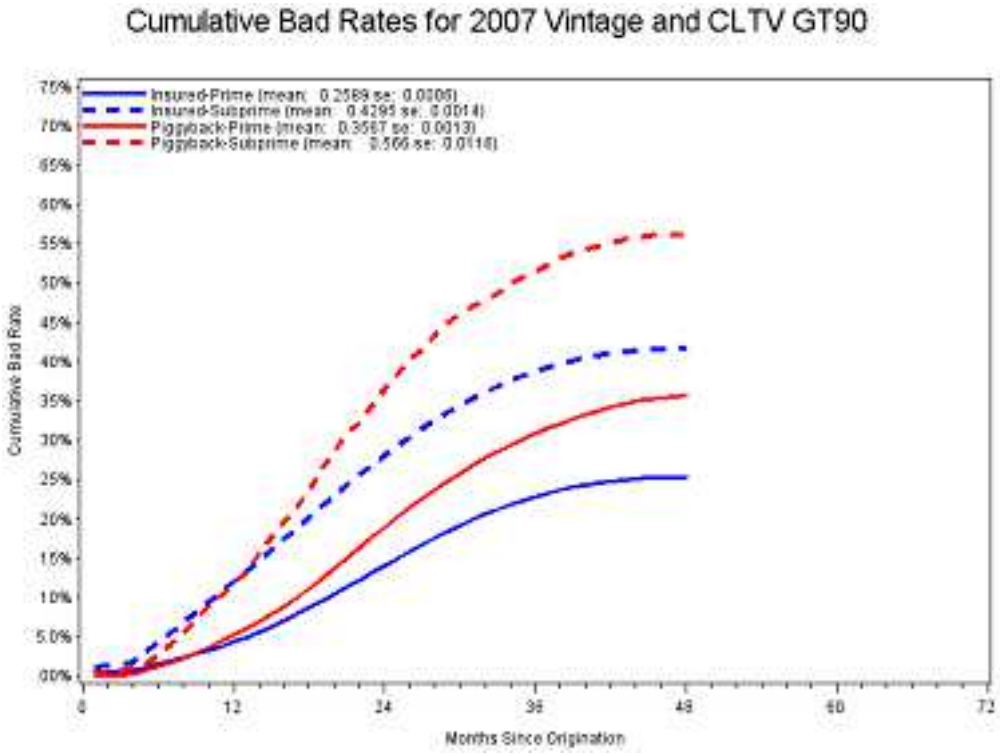


Cumulative Bad Rates for 2006 Vintage and CLTV LE90

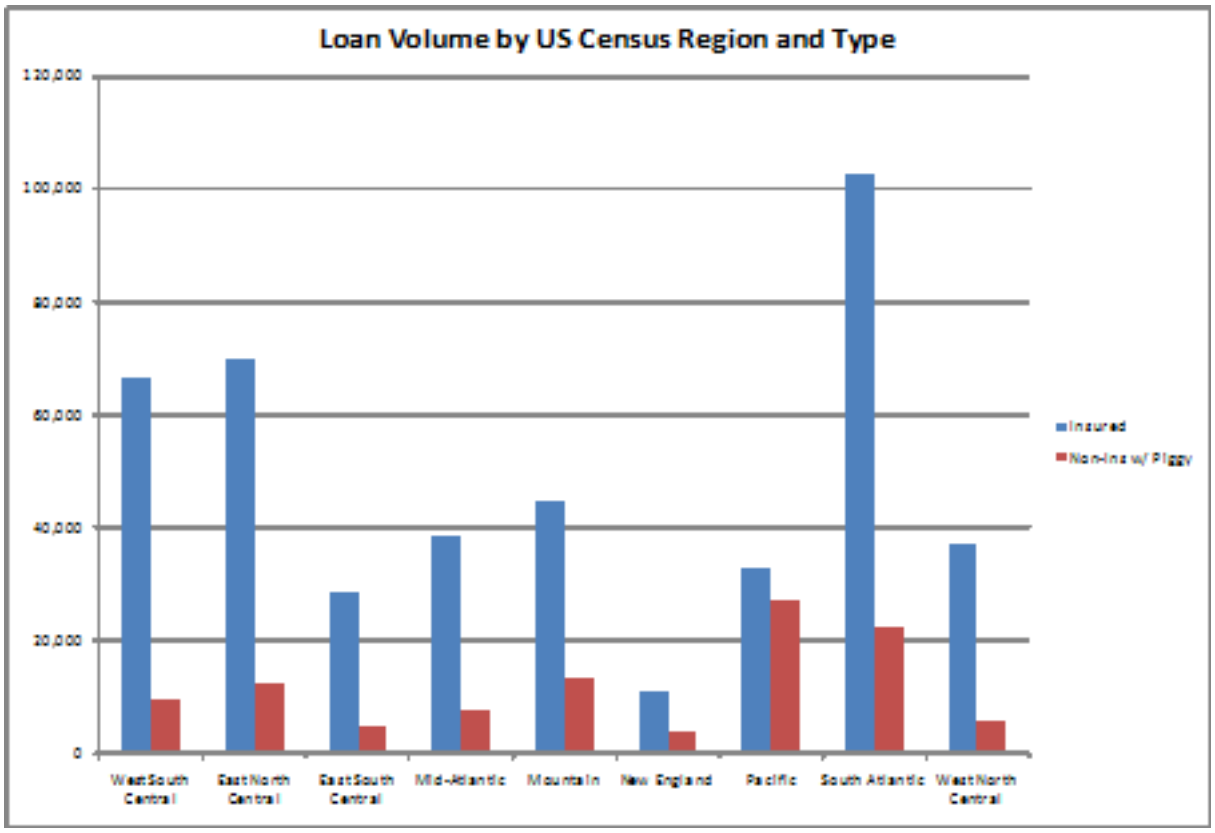


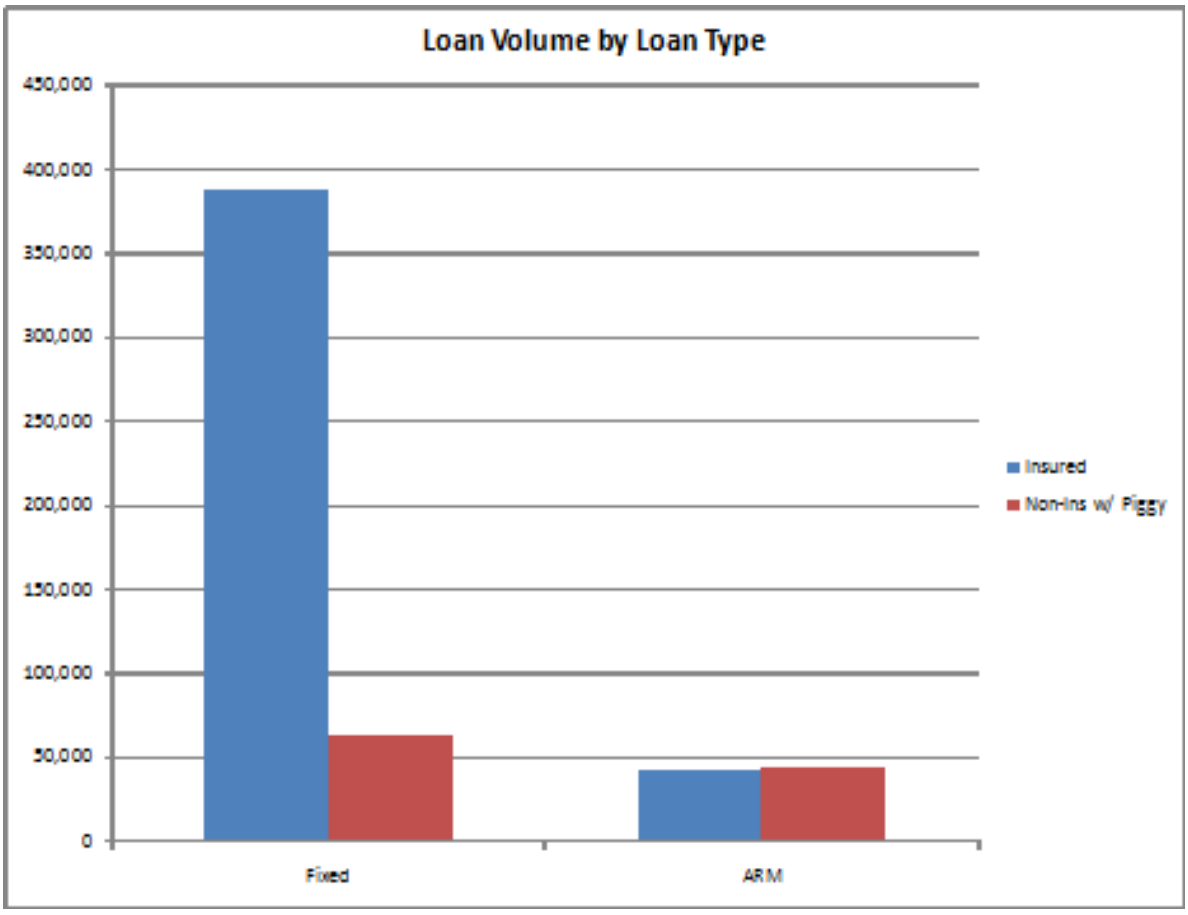

Cumulative Bad Rates for 2006 Vintage and CLTV GT90
Insured-Prime (mean: 0.2073 se: 0.0007)
Insured-Subprime (mean: 0.4049 se: 0.0016)
Piggback-Prime (mean: 0.3484 se: 0.0011)
Piggback-Subprime (mean: 0.6094 se: 0.0126)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
0.0%
0
12
24
36
48
60
72
Months Since Origination

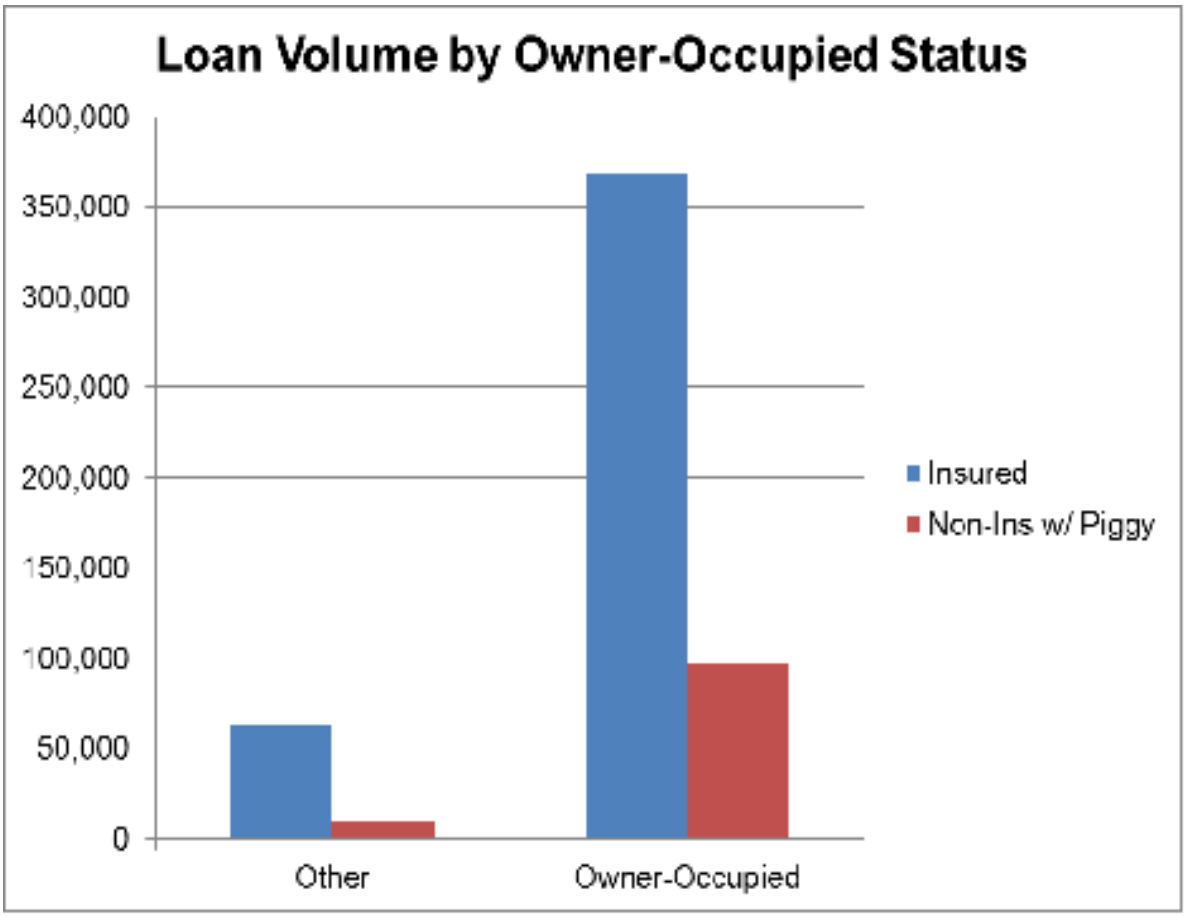

Cumulative Bad Rates for 2007 Vintage and CLTV LE90
Insured-Prime (mean: 0.2632 se: 0.0008)
Insured-Subprime (mean: 0.3781 se: 0.003)
Piggyback-Prime (mean: 0.3324 se: 0.0015)
Piggyback-Subprime (mean: 0.5138 se: 0.0276)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
0.0%
0
12
24
36
48
60
72
Months Since Origination

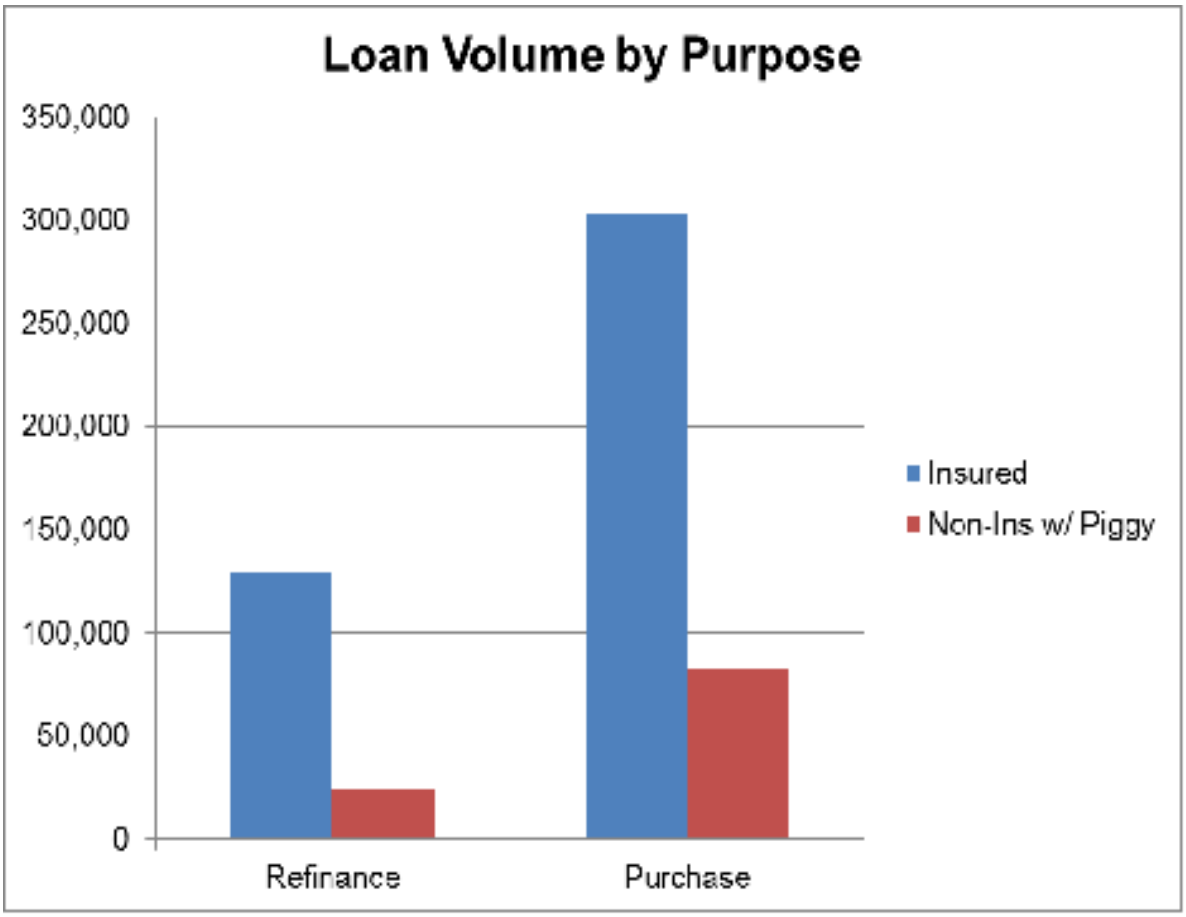

Cumulative Bad Rates for 2007 Vintage and CLTV GT90
Insured-Prime (mean: 0.2569 se: 0.0005)
Insured-Subprime (mean: 0.4295 se: 0.0014)
Piggback-Prime (mean: 0.3567 se: 0.0013)
Piggback-Subprime (mean: 0.566 se: 0.0116)
Cumulative Bad Rate
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5.0%
0.0%
0
12
24
36
48
60
72
Months Since Origination

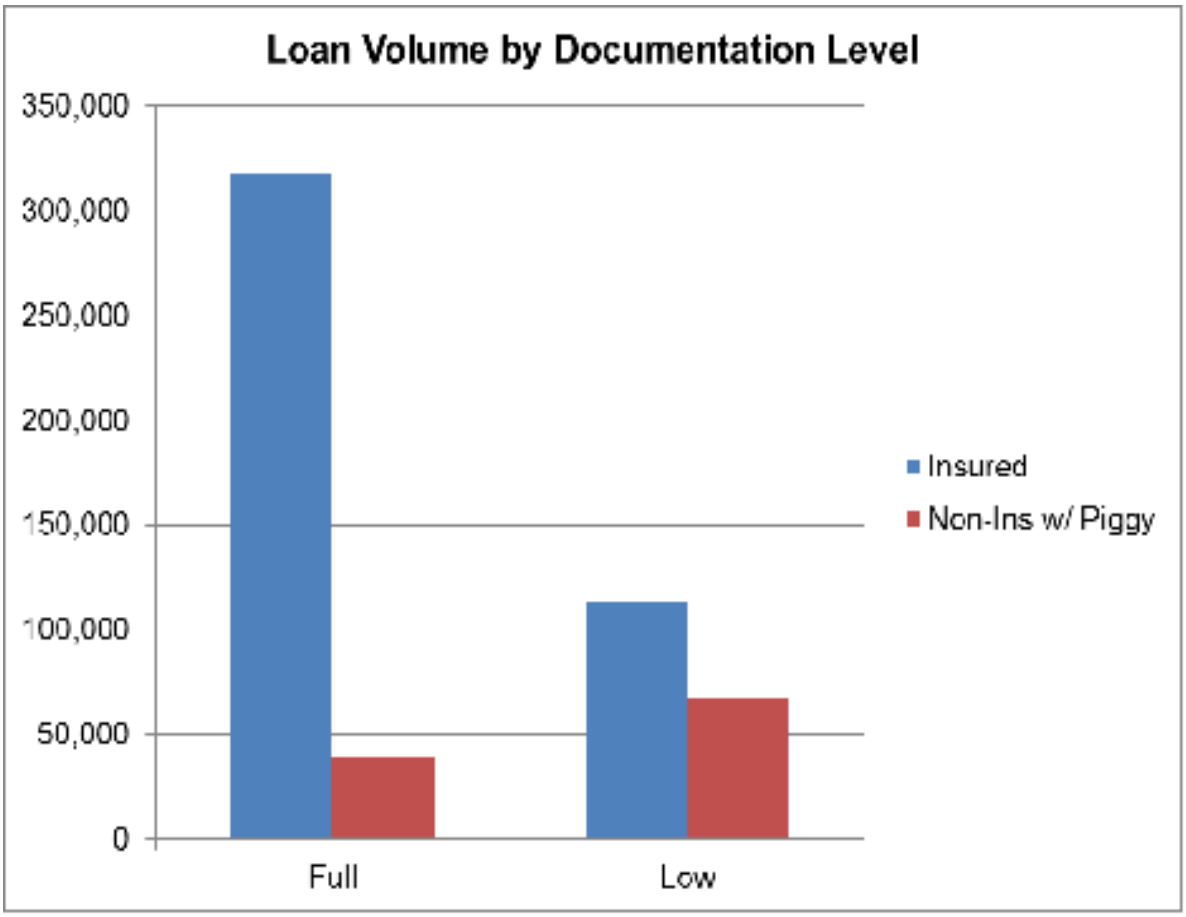
10 Appendix B: Modeling Dataset Summary

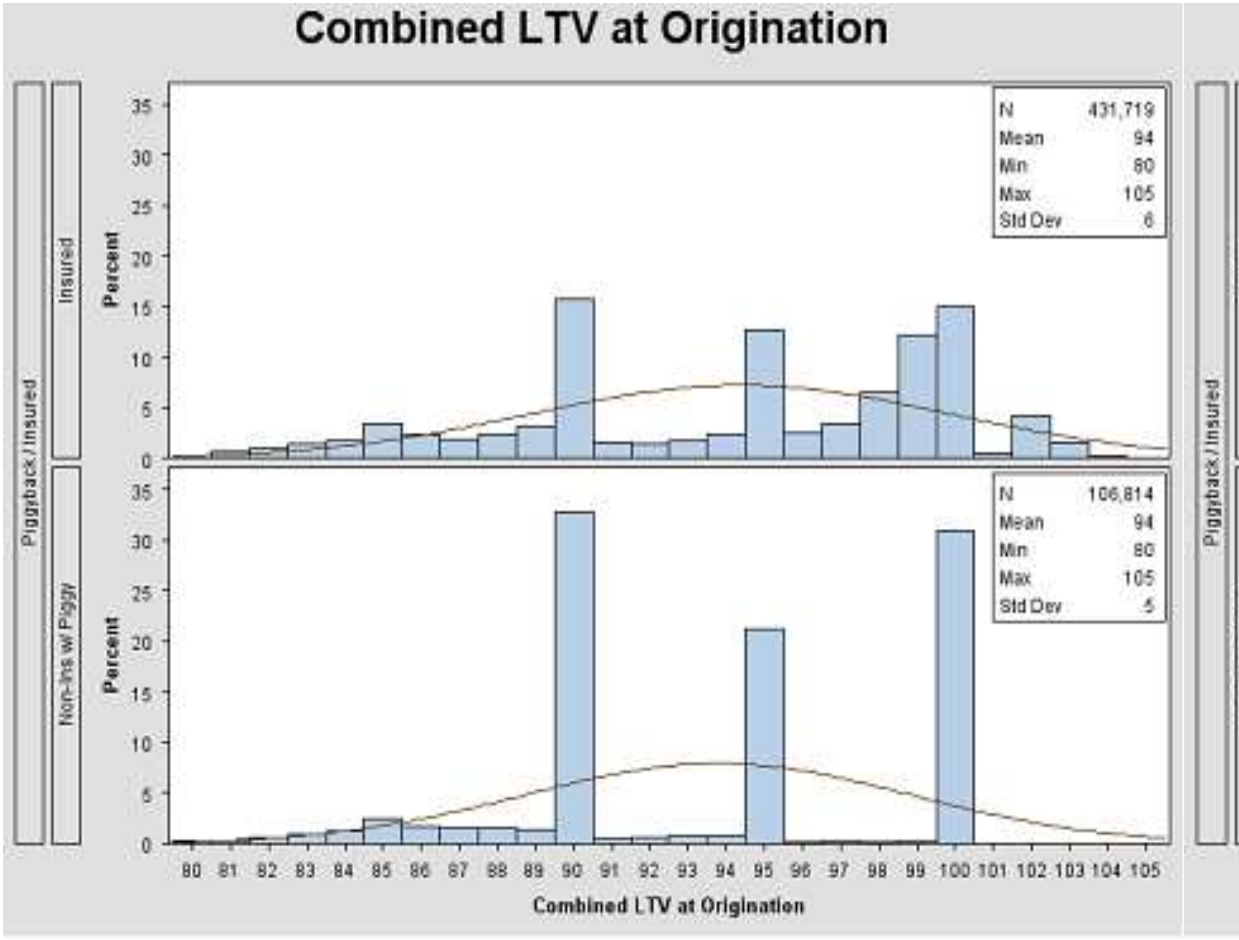



Loan Volume by Loan Type
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Fixed
ARM
Insured
Non-Ins w/ Piggy


Loan Volume by Owner-Occupied Status
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Other
Owner-Occupied
Insured
Non-Ins w/ Piggy


Loan Volume by Purpose
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Refinance
Purchase
Insured
Non-Ins w/ Piggy


Loan Volume by Documentation Level
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Full
Low
Insured
Non-Ins w/ Piggy

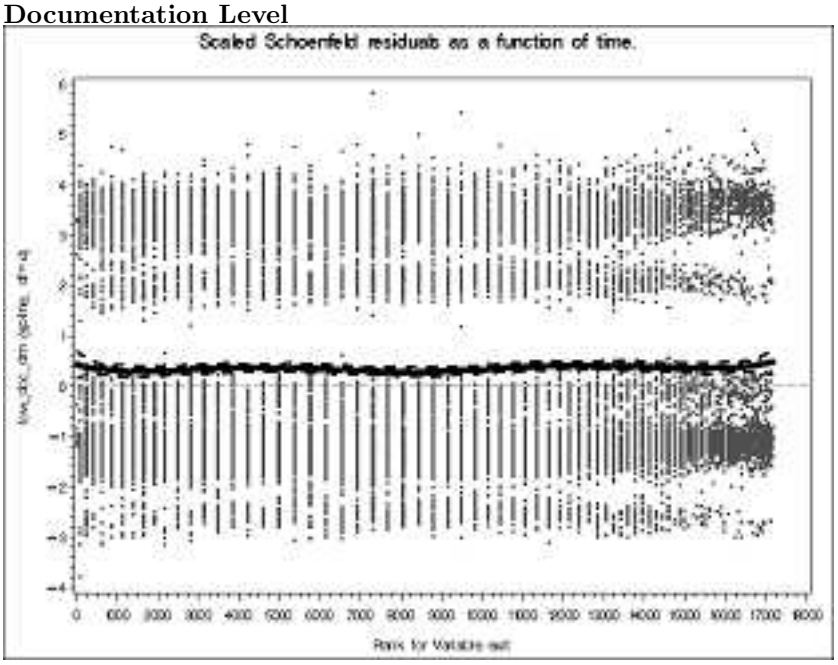

Combined LTV at Origination
Piggyback / Insured
Insured
Non-Ins w/ Piggy
Percent
N 431,719
Mean 94
Min 80
Max 105
Std Dev 6
N 106,814
Mean 94
Min 80
Max 105
Std Dev 5
80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105
Combined LTV at Origination

11 Appendix C: Scaled Schoenfeld Residual Plots

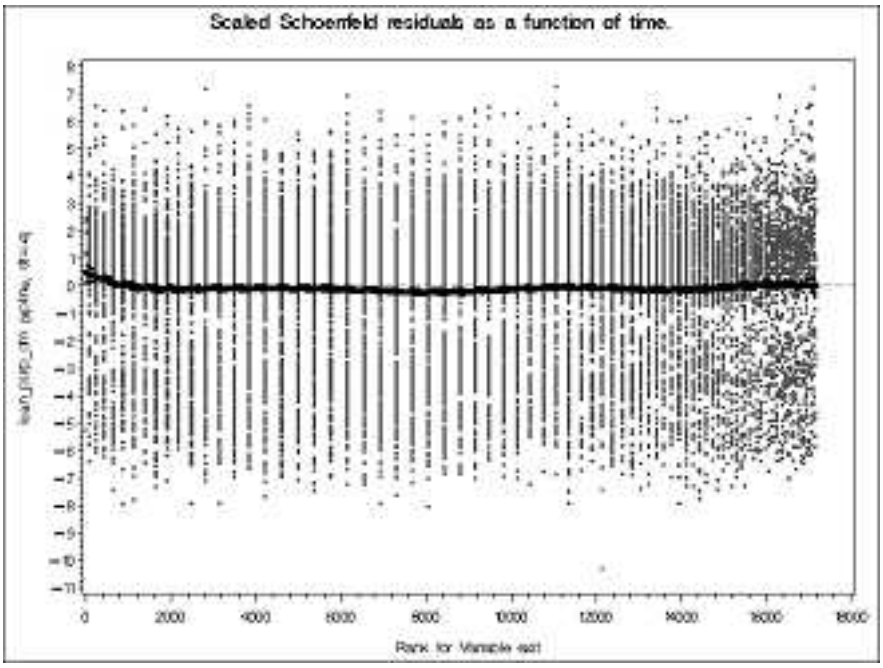
The Schoenfeld residual, r_{ik} is the covariate value, X_{ik}, for the i^{th} loan which actually defaulted at time t, minus the expected value of the covariate for the risk set at time t (i.e., a weighted-average of the covariate, weighted by each loan's likelihood of defaulting at t).

Because they will vary in size and distribution, the Schoenfeld residuals are usually scaled before being analyzed. The k-dimensional vector ofScaled Schoenfeld Residuals, SR, for the i^{th} loan is defined as: SR= β + D*Cov(β)*$\mathbf{r}_i'$, where β=the estimated Cox model coefficient vector, D= the number of loans defaulting, and $\mathbf{r}_i$= the vector of Schoenfeld residuals for loan i.

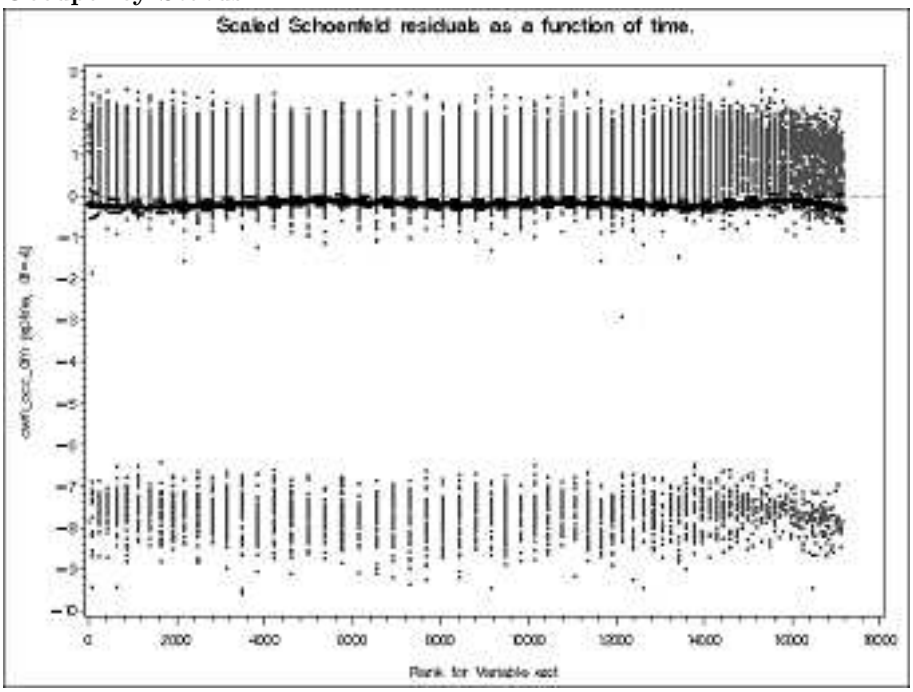
Plots for Fixed-Rate Loans, by Covariate
Documentation Level



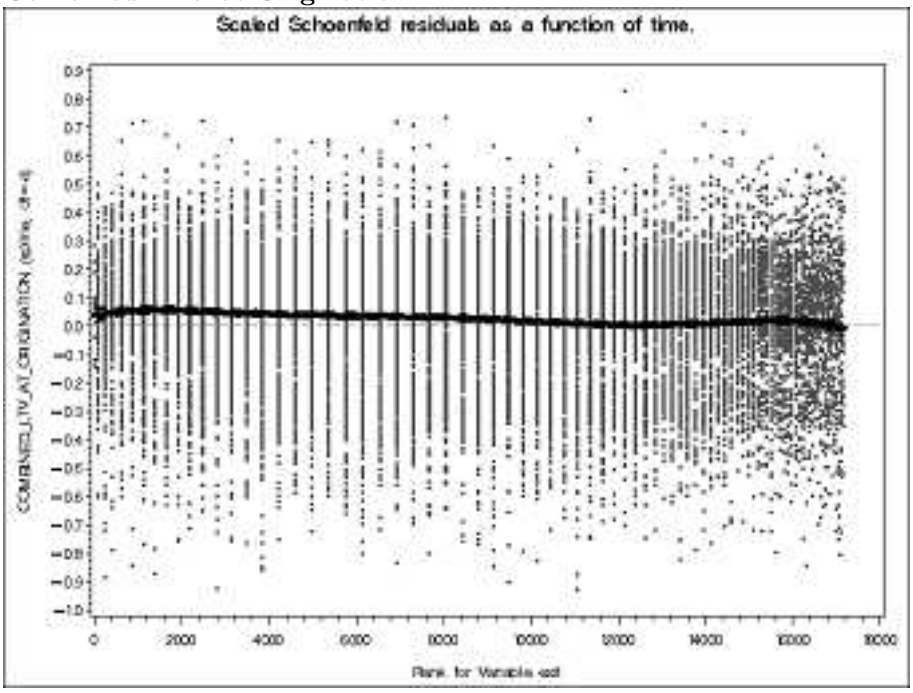
Loan Purpose


Scaled Schoenfeld residuals as a function of time.
Rank for Variable rad

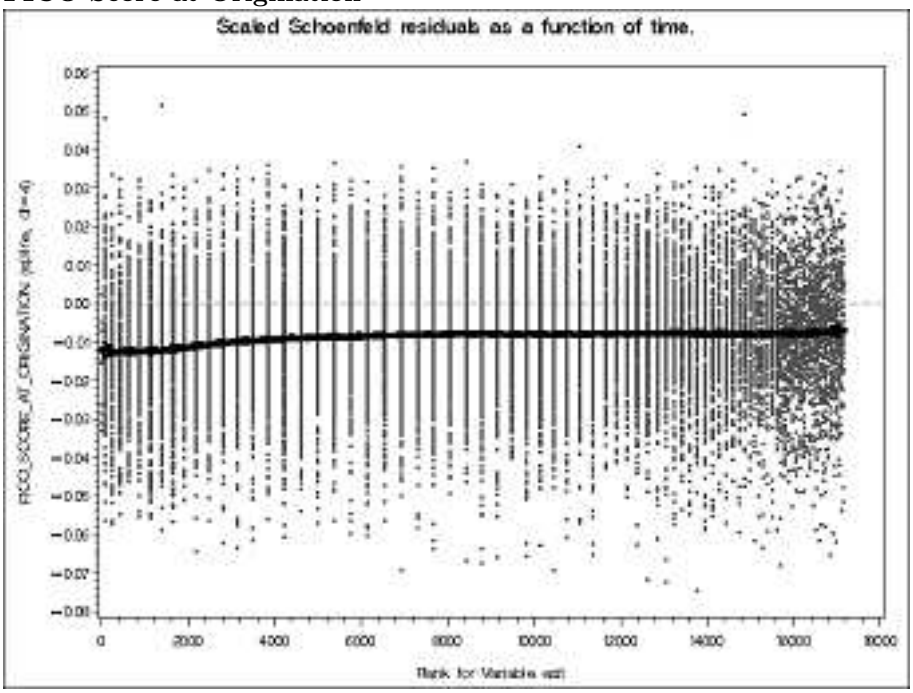
Occupancy Status



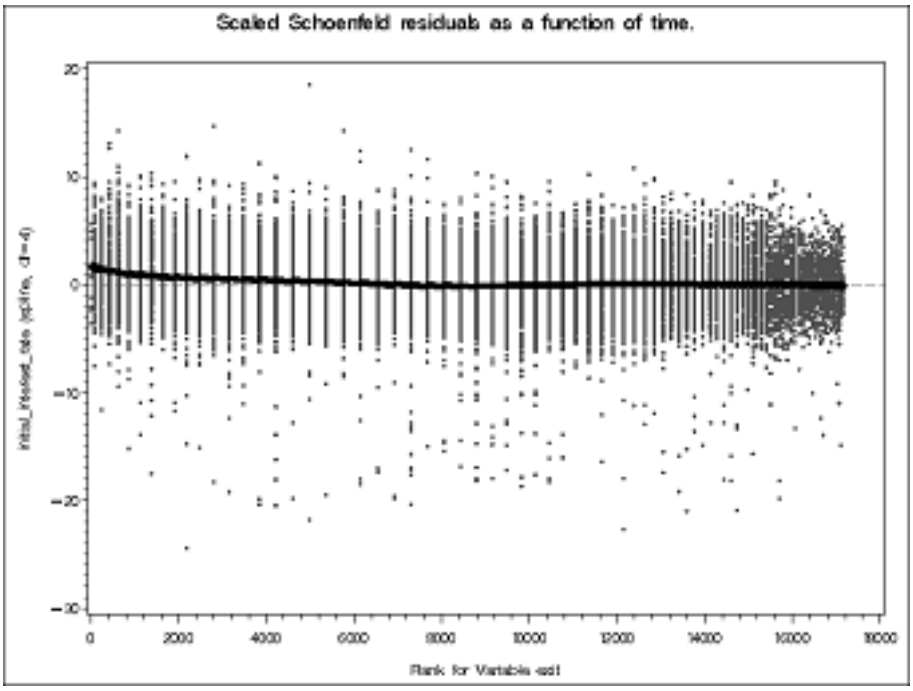
Combined LTV at Origination



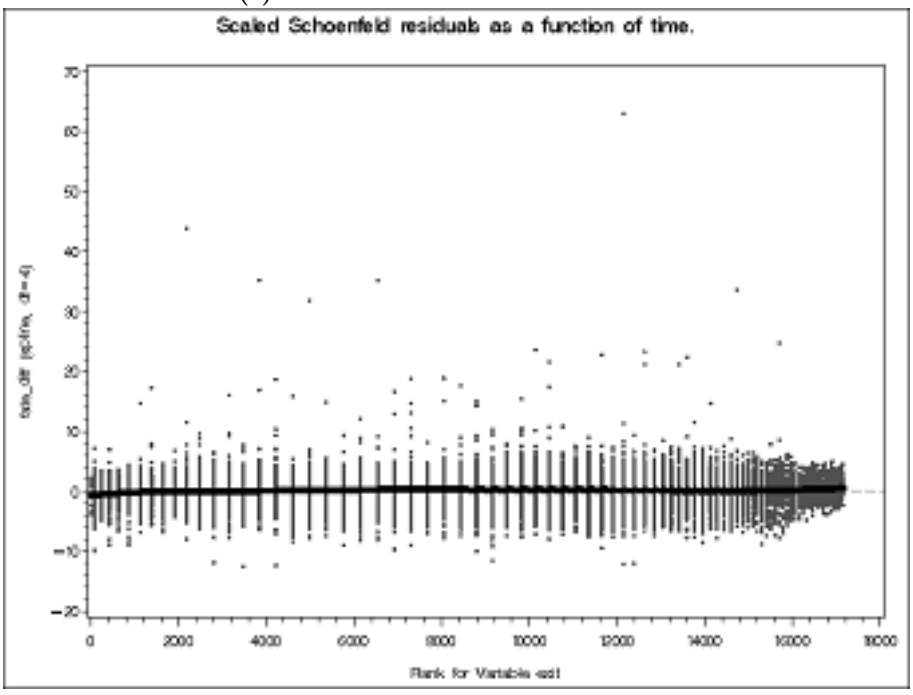
FICO Score at Origination



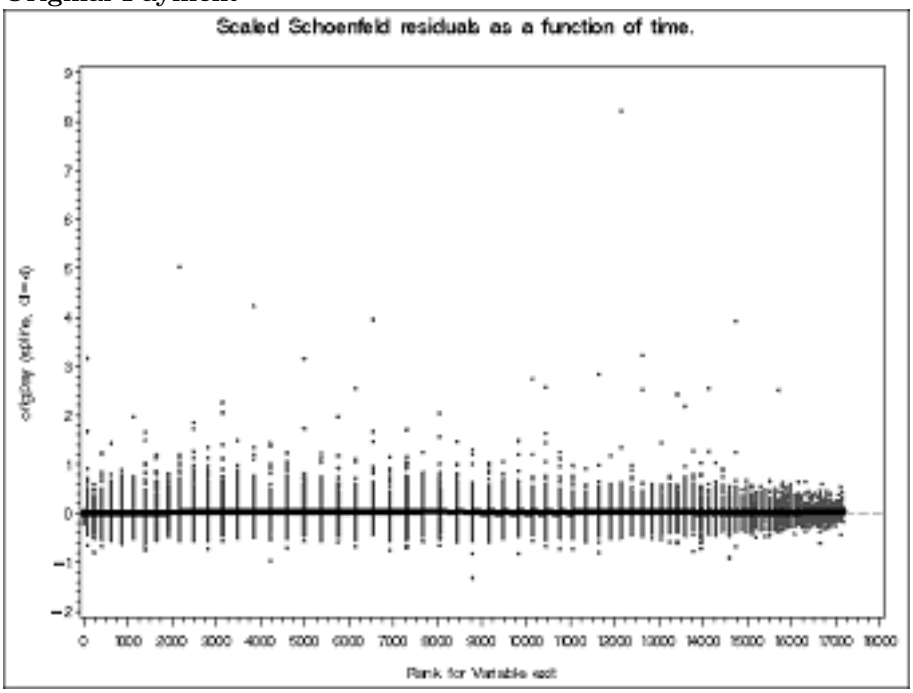
Original Interest Rate



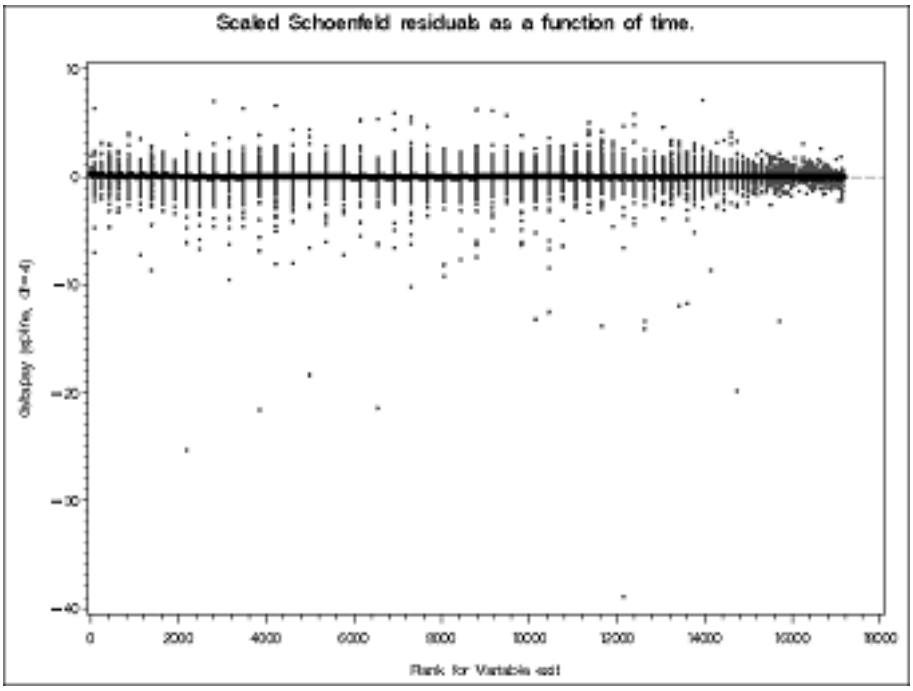
Rate Differential (t)



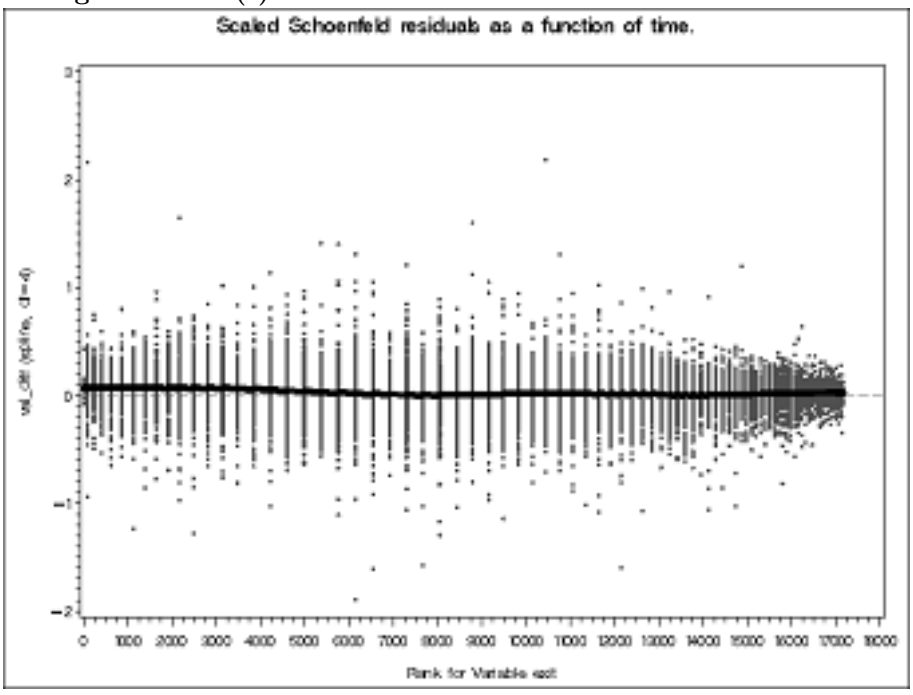
Original Payment



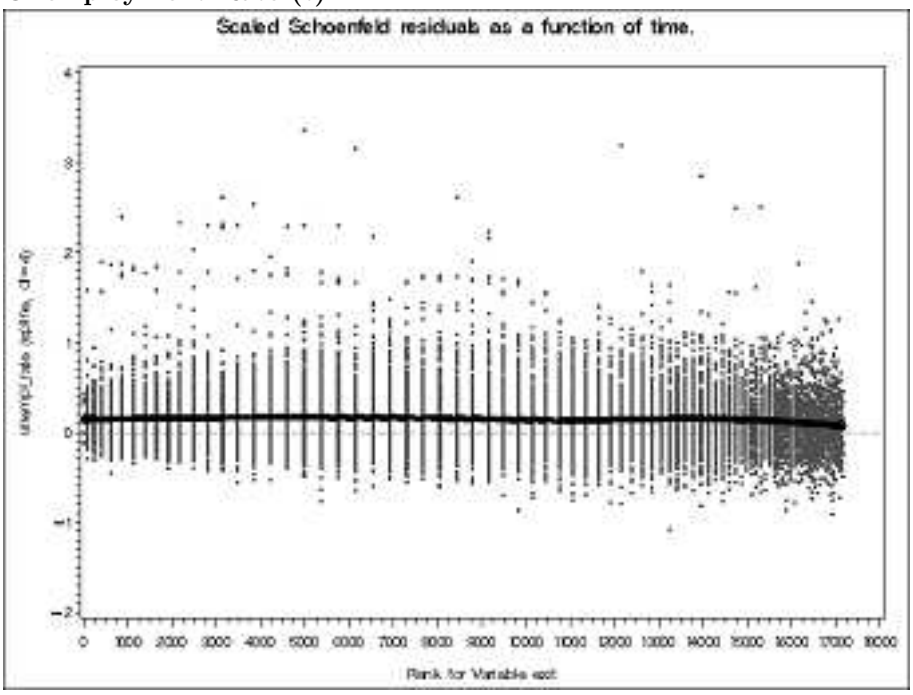
Change in Payment (t)



Change in Value (t)

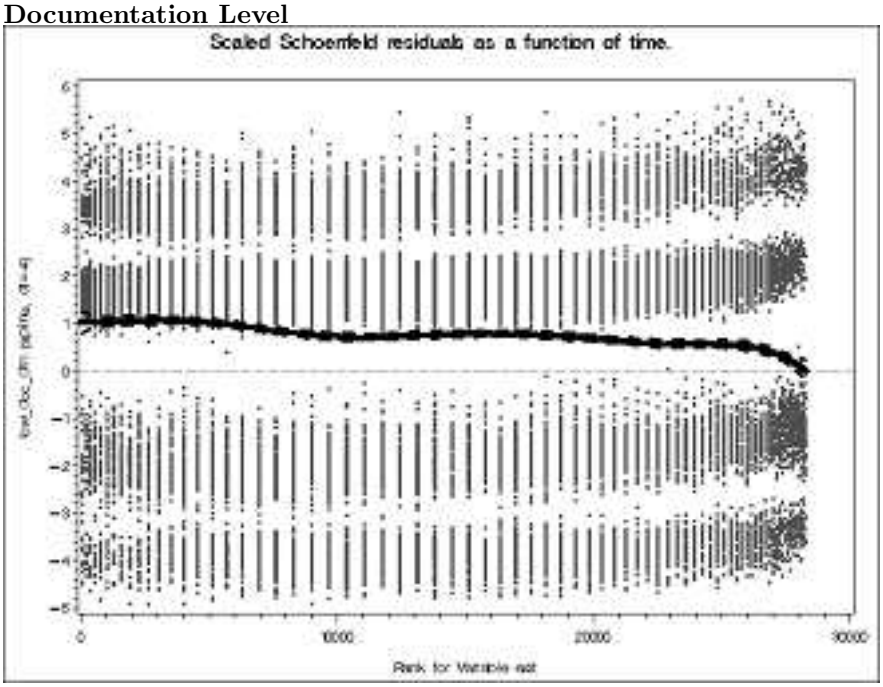


Unemployment Rate (t)

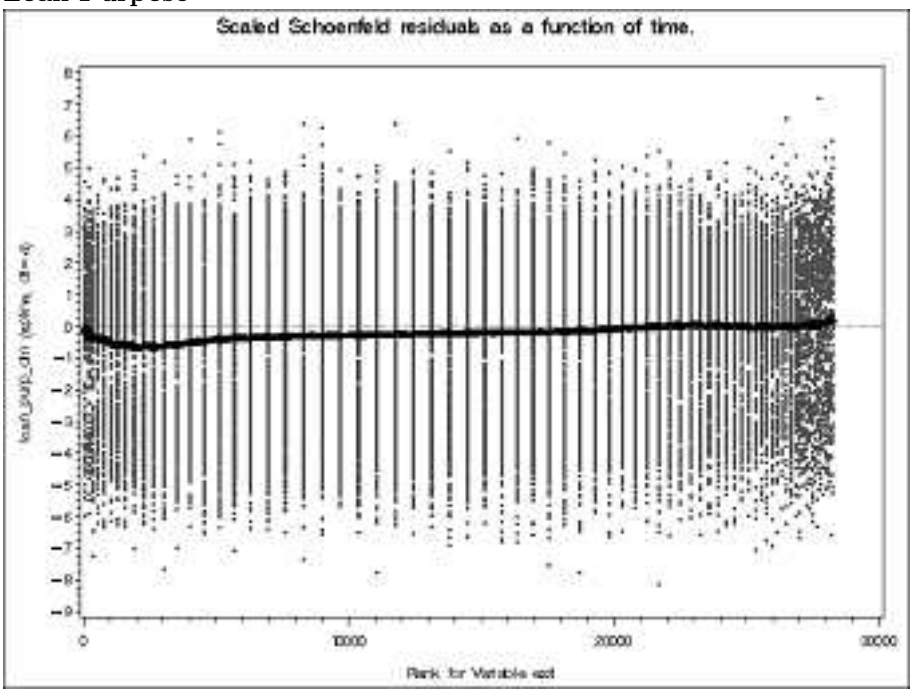


Plots for Adjustable-Rate Loans, by Covariate

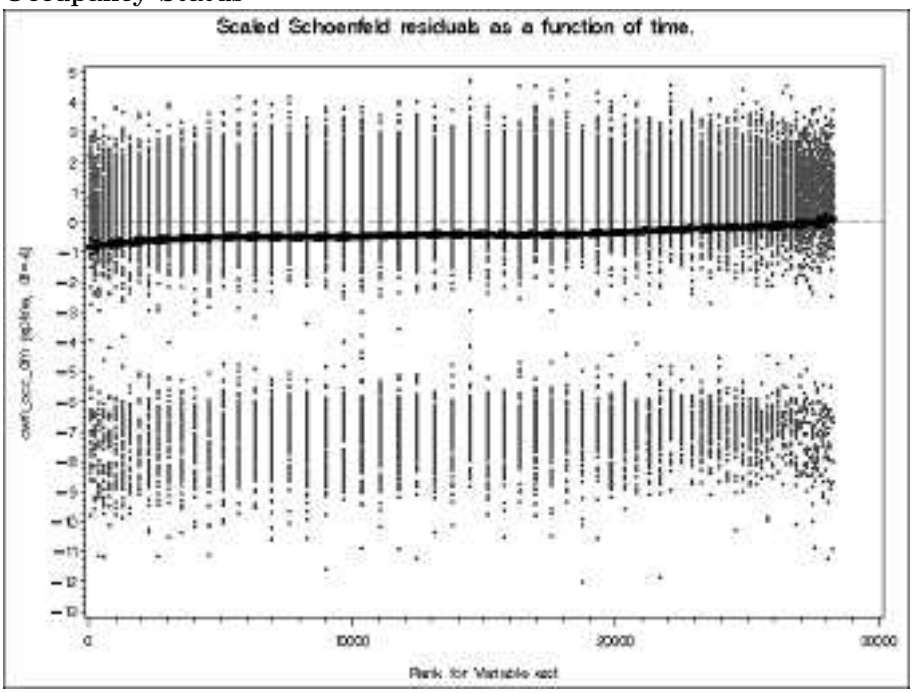
Documentation Level



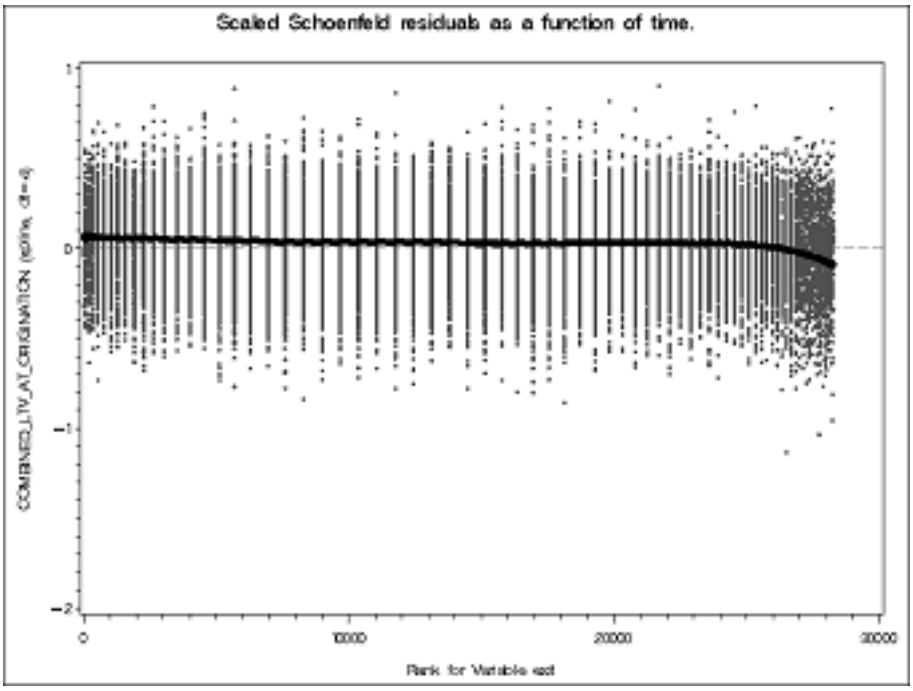
Loan Purpose



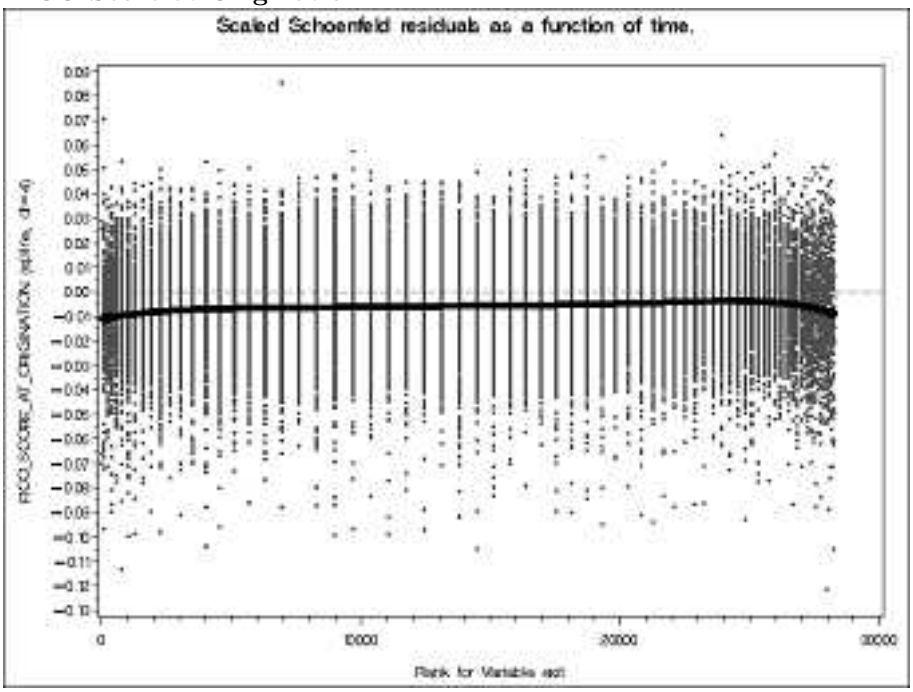
Occupancy Status



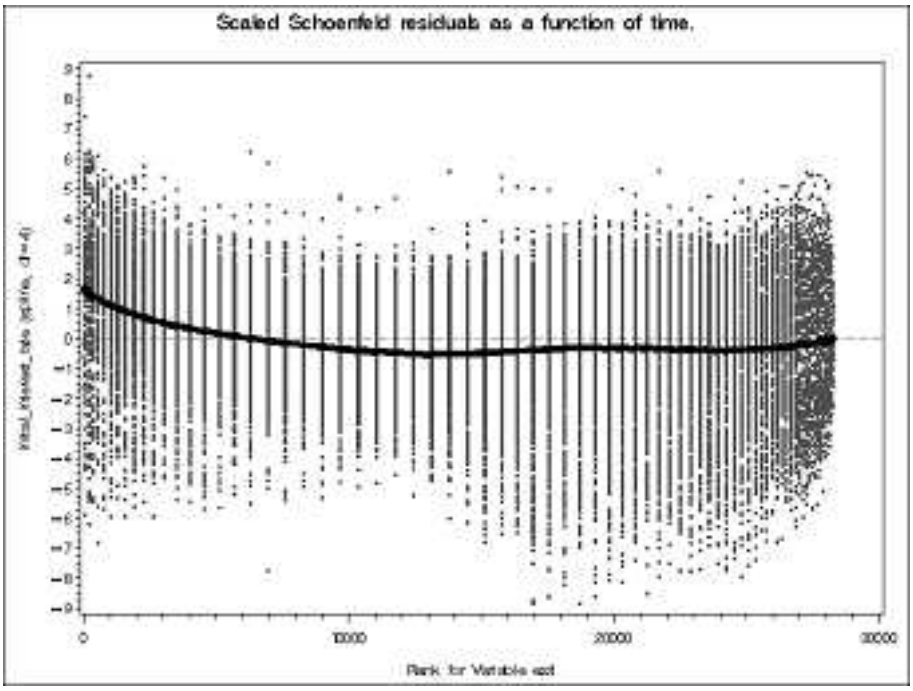
Combined LTV at Origination


Scaled Schoenfeld residuals as a function of time.
COMBINED_LTV_AT_ORIGINATION (spline, df=4)
1
0
-1
-2
0
10000
20000
30000
Rank for Variable est

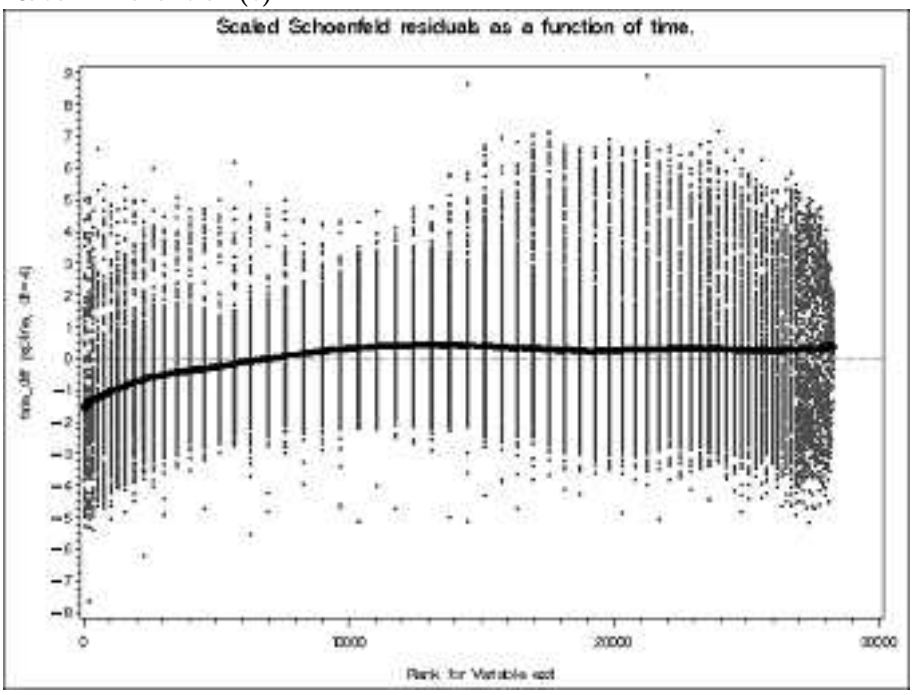
FICO Score at Origination



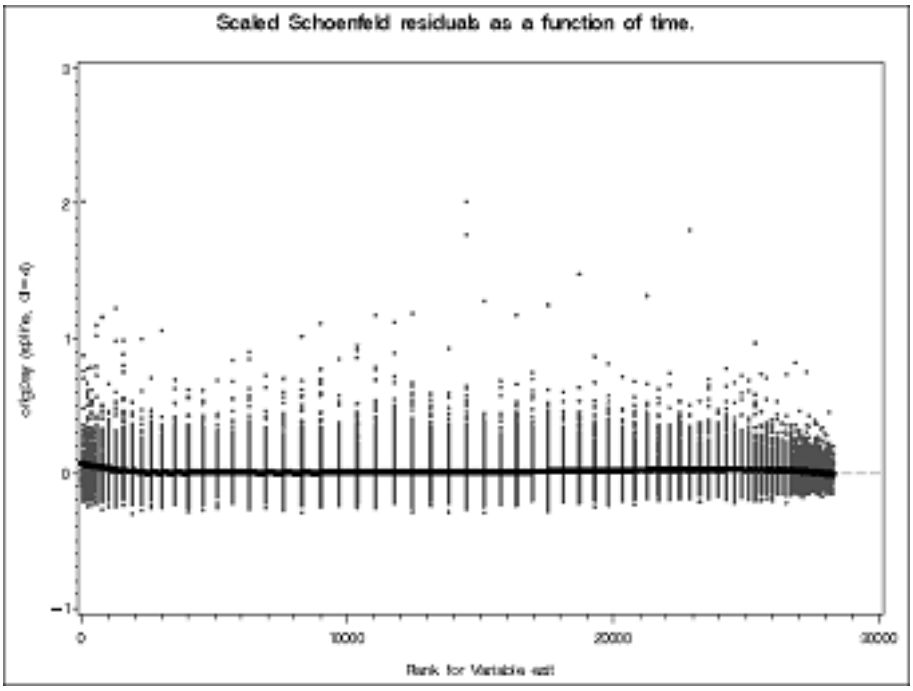
Original Interest Rate


Scaled Schoenfeld residuals as a function of time.
0
10000
20000
30000
Rank for Variable est

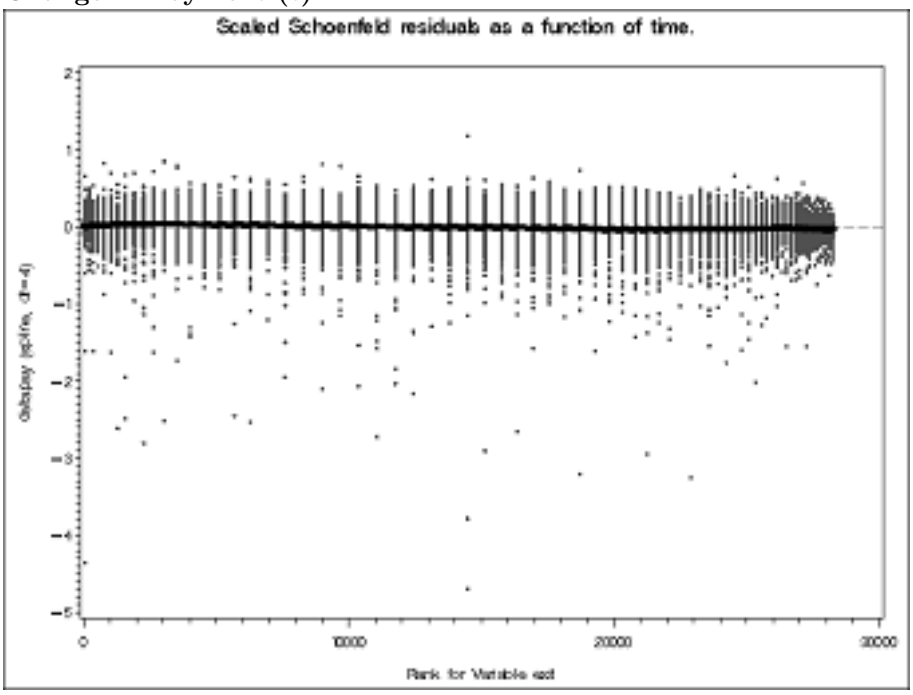
Rate Differential (t)



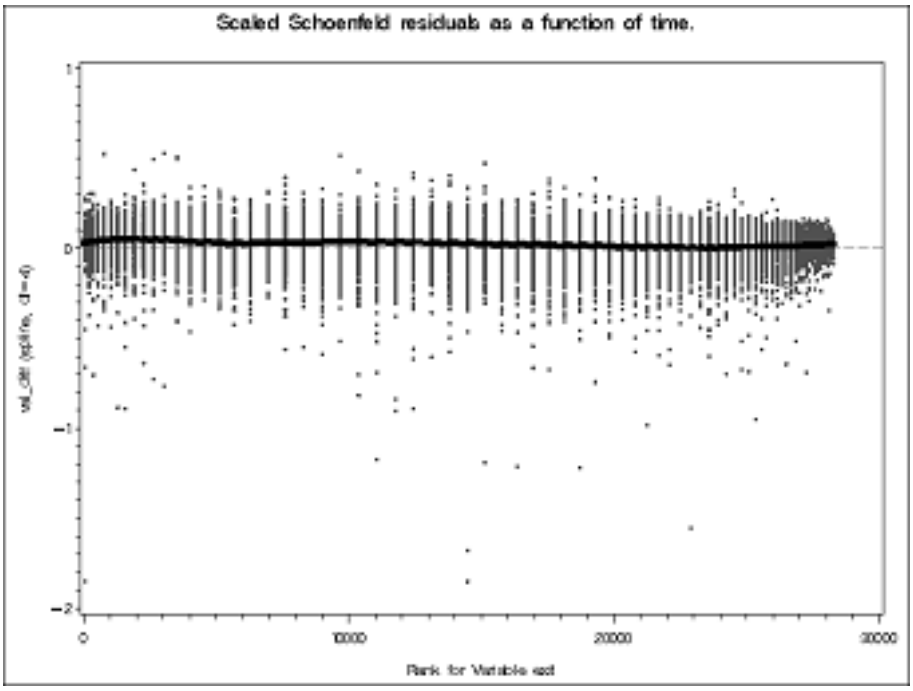
Original Payment


Scaled Schoenfeld residuals as a function of time.
3
2
1
0
-1
0
10000
20000
30000
Rank for Variable exit

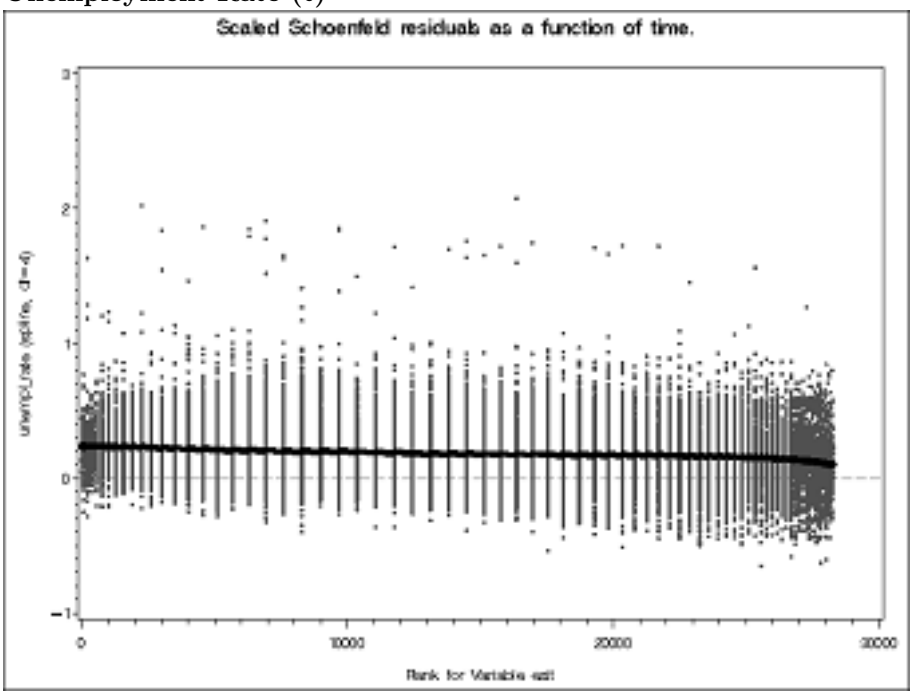
Change in Payment (t)



Change in Value (t)


Scaled Schoenfeld residuals as a function of time.
1
0
−1
−2
0
10000
20000
30000
Rank for Variable ast

Unemployment Rate (t)



Research Papers 2013

2014-58: Anders Bredahl Kock and Haihan Tang: Inference in High-dimensional Dynamic Panel Data Models

2015-01 Tom Engsted, Simon J. Hviid and Thomas Q. Pedersen: Explosive bubbles in house prices? Evidence from the OECD countries

2015-02: Tim Bollerslev, Andrew J. Patton and Wenjing Wang: Daily House Price Indices: Construction, Modeling, and Longer-Run Predictions

2015-03: Christian M. Hafner, Sebastien Laurent and Francesco Violante: Weak diffusion limits of dynamic conditional correlation models

2015-04: Maria Eugenia Sanin, Maria Mansanet-Bataller and Francesco Violante: Understanding volatility dynamics in the EU-ETS market

2015-05: Peter Christoffersen and Xuhui (Nick) Pan: Equity Portfolio Management Using Option Price Information

2015-06: Peter Christoffersen and Xuhui (Nick) Pan: Oil Volatility Risk and Expected Stock Returns

2015-07: Peter Christoffersen, Bruno Feunou and Yoontae Jeon: Option Valuation with Observable Volatility and Jump Dynamics

2015-08: Alfonso Irarrazabal and Juan Carlos Parra-Alvarez: Time-varying disaster risk models: An empirical assessment of the Rietz-Barro hypothesis

2015-09: Daniela Osterrieder, Daniel Ventosa-Santaulària and Eduardo Vera-Valdés: Unbalanced Regressions and the Predictive Equation

2015-10: Laurent Callot, Mehmet Caner, Anders Bredahl Kock and Juan Andres Riquelme: Sharp Threshold Detection Based on Sup-norm Error rates in High-dimensional Models

2015-11: Arianna Agosto, Giuseppe Cavaliere, Dennis Kristensen and Anders Rahbek: Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)

2015-12: Tommaso Proietti, Martyna Marczak and Gianluigi Mazzi: EuroMInd-D: A Density Estimate of Monthly Gross Domestic Product for the Euro Area

2015-13: Michel van der Wel, Sait R. Ozturk and Dick van Dijk: Dynamic Factor Models for the Volatility Surface

2015-14: Tim Bollerslev, Andrew J. Patton and Rogier Quaedvlieg: Exploiting the Errors: A Simple Approach for Improved Volatility Forecasting

2015-15: Hossein Asgharian, Charlotte Christiansen and Ai Jun Hou: Effects of Macroeconomic Uncertainty upon the Stock and Bond Markets

2015-16: Markku Lanne, Mika Meitz and Pentti Saikkonen: Identification and estimation of non-Gaussian structural vector autoregressions

2015-17: Nicholas M. Kiefer and C. Erik Larson: Counting Processes for Retail Default Modeling