Mortgage Loan Prepayment Risk - Survival Analysis

Yuya Ogawa

2024-07-15

Introduction

In this paper, I demonstrate that the interest rate gap—defined as the difference between the current mortgage rate and the average long-run fixed rate—is indeed a strong factor influencing mortgage loan prepayment. From a central banking perspective, this paper contributes to the understanding of the path-dependency effects of monetary policy, as confirmed in prior literature. My goal is to show how to perform the analysis in R using survival analysis and other tools while communicating my findings effectively. Using mortgage data from Freddie Mac, I model prepayment risk by applying the Cox proportional hazards model. I then compare the effect of a permanent shock in interest rates on mortgage prepayment risk. Finally, I will also conduct logistic regression and linear probability regression with additional control variables to further support my initial findings.

Data Description

The data comes from Freddie Mac. Follow the instructions and code I have provided to first load and prepare the data for analysis. After running the provided code, you should have two main datasets: LOAN_surv.csv and pp_df_rate.csv. We will also use unemployment data from FRED[https://fred.stlouisfed.org/series/UNRATE].

First load the main data we use for analysis.

```
#Set the libraries we will use in the analysis.
library(survival)
```

```
## Warning: package 'survival' was built under R version 4.3.3
```

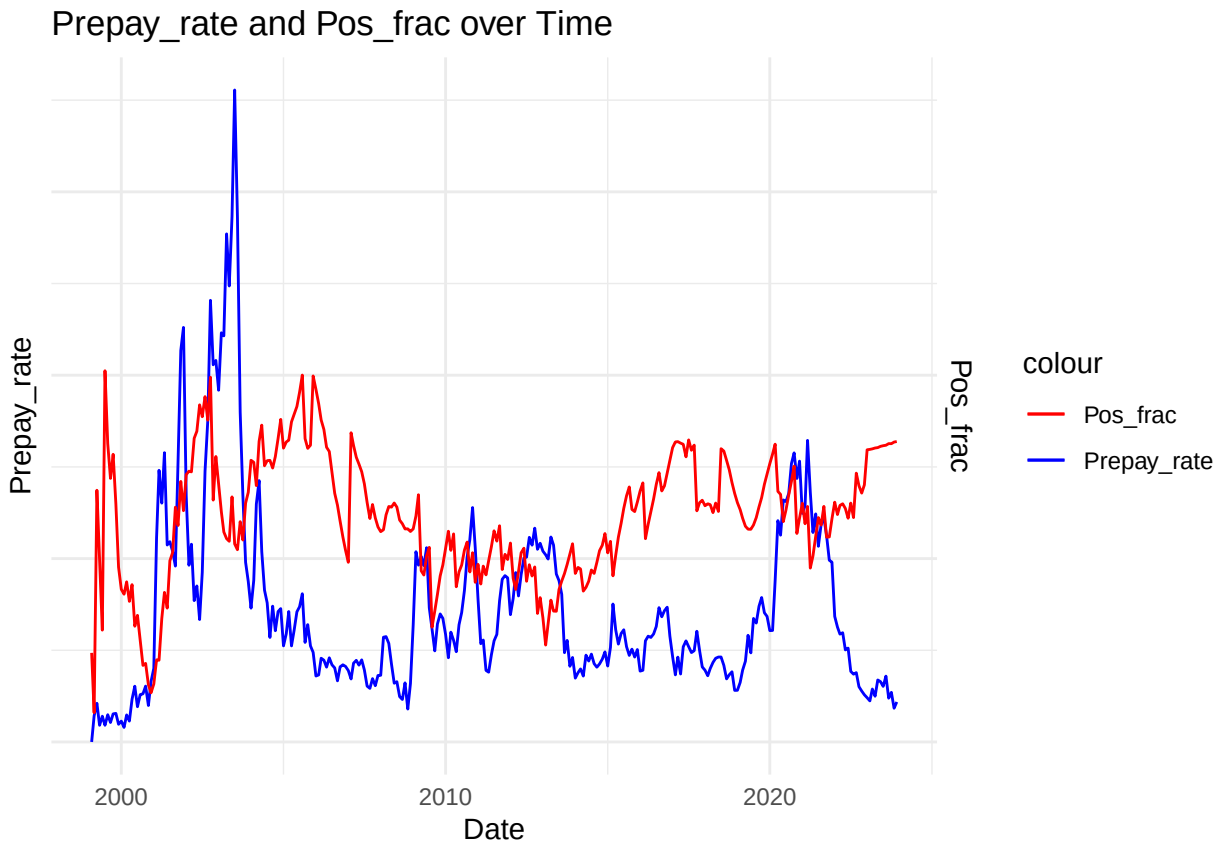
```
library(ranger)
library(ggplot2)
library(dplyr)
library(ggfortify)
library(lubridate)
library(plm)
library(fpp3)
library(survminer)


# Load the data
data = read.csv('LOAN_surv.csv')
temp = read.csv('UNRATE.csv')
# Convert the DATE so that it has the same key I can use when merging.
temp <-
  temp %>% mutate(DATE = as.Date(DATE)) %>% mutate(YearMonth = format(DATE, "%Y%m"))
# Merge data and temp on the specified keys
data <- merge(data, temp, by.x = "MONTHLY_REPORTING_PERIOD", by.y = "YearMonth", all = FALSE)
```

Then let us load another dataset that we use for visualization.

```
#
temp1 = read.csv('pp_df_rate.csv')
temp2 = read.csv('UNRATE.csv')
temp2 <-
  temp2 %>% mutate(DATE = as.Date(DATE)) %>% mutate(YearMonth = format(DATE, "%Y%m"))
timeseries_rate <- merge(temp1, temp2, by.x = "MONTHLY_REPORTING_PERIOD", by.y = "YearMonth", all = FALS
timeseries_rate = timeseries_rate %>% mutate(Date = as.Date(DATE)) %>% as_tsibble(index = Date)
```

The following graph shows the prepayment rate on the left (the fraction of loans that are prepaid) and the positive gap rate on the right (the fraction of loans with a positive rate gap). This illustrates that the positive gap (the difference between the current mortgage rate and the long-run fixed rate) tends to move in the opposite direction of the prepayment rate. As Berger et al. (2018)[https://www.aeaweb.org/articles?id=10.1257/aer.20181857] argued, this is due to refinancing incentives. When a consumer finances a loan with a fixed rate for the long term, they face interest rate risk. In the case of low interest rates, consumers have an incentive to prepay to avoid paying a higher interest rate relative to the current rate.

```
#
ggplot(timeseries_rate, aes(x = Date)) +
  geom_line(aes(y = Prepay_rate*10, color = "Prepay_rate")) +
  geom_line(aes(y = Pos_frac*2-80, color = "Pos_frac")) +  # Rescale the pos_frac for comparison
  scale_y_continuous(
    name = "Prepay_rate",
    sec.axis = sec_axis(~./10, name = "Pos_frac")  # Adjusting the scale back to original for secondary
  ) +
  labs(title = "Prepay_rate and Pos_frac over Time",
       x = "Date") +
  scale_color_manual(values = c("Prepay_rate" = "blue", "Pos_frac" = "red")) +
  theme_minimal()+
  theme(
    axis.text.y = element_blank()
  )
```



Introduction to Survival Analysis

Before we start the analysis, here is a brief introduction to survival analysis. Although there are excellent textbooks available for learning survival analysis, I will outline what it is, why it can be applied to loan termination, and how it can be estimated.

First, why is survival analysis relevant? Consider that you are trying to estimate the probability of loan termination due to prepayment, but your data also include cases where loan termination occurs for reasons other than prepayment, such as default or maturity. For these cases, you do not directly observe termination by prepayment. In such situations, we refer to our main variable of interest as "censored."

Mathematically, let T_i represent the time when loan i is terminated due to prepayment (failure time), and C represent the time when loan i is terminated for reasons other than prepayment (censoring time), such as maturity or default. We only observe the time $\min(T_i, C_i)$. To address this issue, you might consider using only the loans where you directly observe T_i by excluding those where the failure time is not observed. However, this approach could lead to selection bias and result in biased estimates. To avoid this, we need to incorporate all loan data, without excluding any, by using survival analysis! So the question is: how?

Kaplan-Meier Survival Curve

In survival analysis, we are often interested in finding what is called a survival curve, which displays the probability of survival (or the probability that loans are not prepaid) on the y-axis and the age of the loan on the x-axis. In other words, the survival function as a function of time represents the probability of survival beyond time t. Mathematically, $S(t) = P(T > t)$. The estimate of S(t) using sample data is called the Kaplan-Meier estimate. To come up with the estimate of $S(t)$, assume we have a sample size of N and let $K = \{t_i\}_{i \leq N}$ be an ordered sequence of failure time for all of the loans in our sample. It is obvious that

$$P(T > t_{n+1}) = P(T > t_{n+1} | T > t_n) P(T > t_n)$$

. Sequentially, you have that

$$P(T > t_n) = \prod_{k=1}^{N} P(T > t_{k+1} | T > t_k) P(T > t_1)$$

Now the question is to come up with the estimate of $P(T > t_{k+1}|T > t_k)$ for each k. A natural candidate is to use the percentage of loans that survived past t_k. If you let N_{t_k} be the number of loans "alive" right before t_k and D_{t_k} be the number of loans that died (or prepaid) at the time of t_k, then the estimate is $\hat{P}(T > t_{k+1}|T > t_k) = \frac{N_{t_k} - D_{t_k}}{N_{t_k}}$. So our sample estimate of the survival curve is

$$S(t) = \prod_{k=1}^{N} \frac{N_{t_k} - D_{t_k}}{N_{t_k}}$$

.

Cox Proportional Hazard Model

Consider a function defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

This is called the Hazard function which captures the instantaneous rate of death at a given point at time t. The numerator captures the instantaneous probability of death given that you survive until time t, but the denominator represents the change in time. This function is important because of the following property:

$$h(t) = \frac{f(t)}{S(t)}$$

where $f(t)$ is the density function corresponding to the probability of death, and $S(t)$ is the survival curve. Simple calculation also shows that indeed $h(t) = -[\frac{S'(t)}{S(t)}]$, which implies

$$S(t) = \exp[-\int_0^t h(\tau) d\tau]$$

With this, I will now introduce the method I will use in my methodology which is called the Cox Proportional Hazard model. This model is special because of its semi-parametric nature. To see this, we first assume the form of our hazard model to be:

$$h(t|x) = h_0(t) \exp(\sum_n \beta_n x_n)$$

This is semi-parametric because we do not assume any functional form of $h_0(t)$ which captures the time variant component of hazard rate. To estimate the parameter of our interest $\{\beta_i\}_i$, we use maximum likelihood estimation. A more detailed discussion on this can be readily found elsewhere.

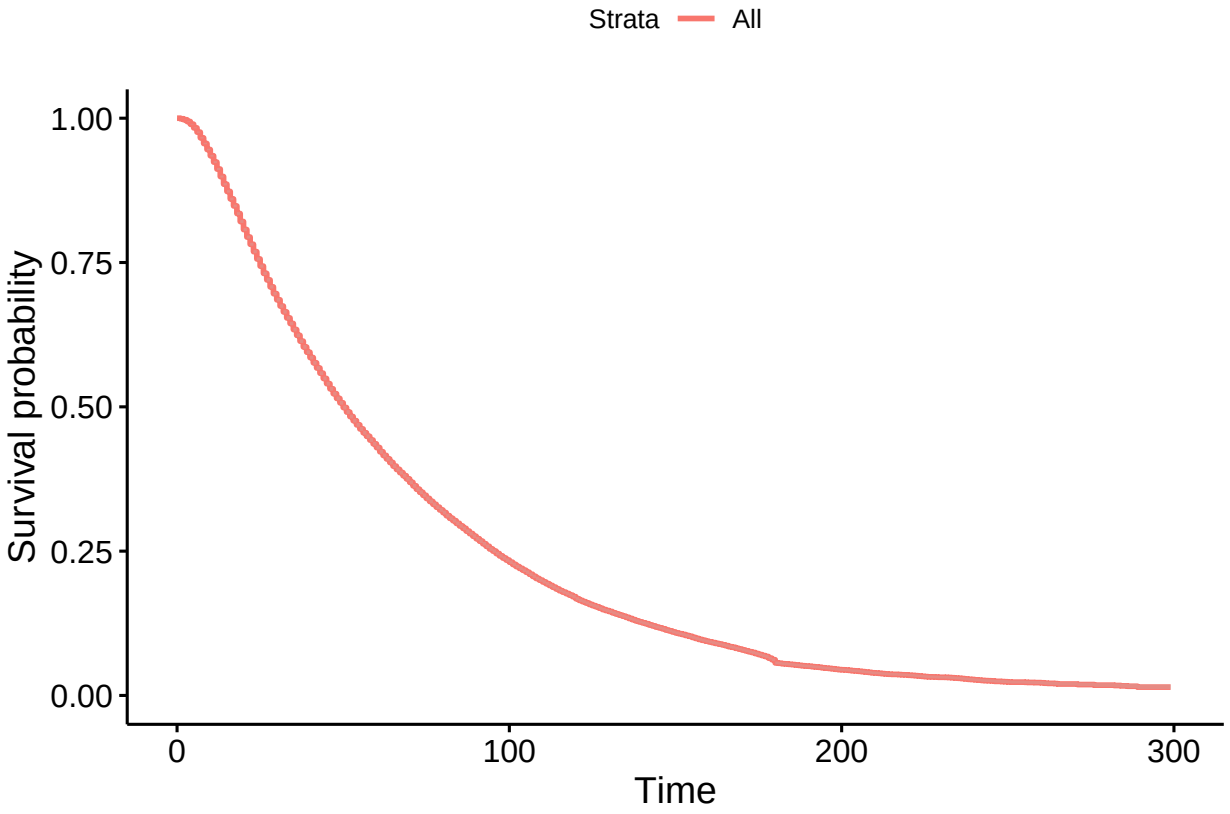
Method & Result

Comparison using Kaplan-Meier

In this section, I will plot the survival curve for loans with different characteristics using the Kaplan-Meier estimator. Finally, I will estimate the survival curve with the Cox Proportional Hazards model, incorporating various variables of interest such as the interest rate gap, unemployment, credit score, etc.

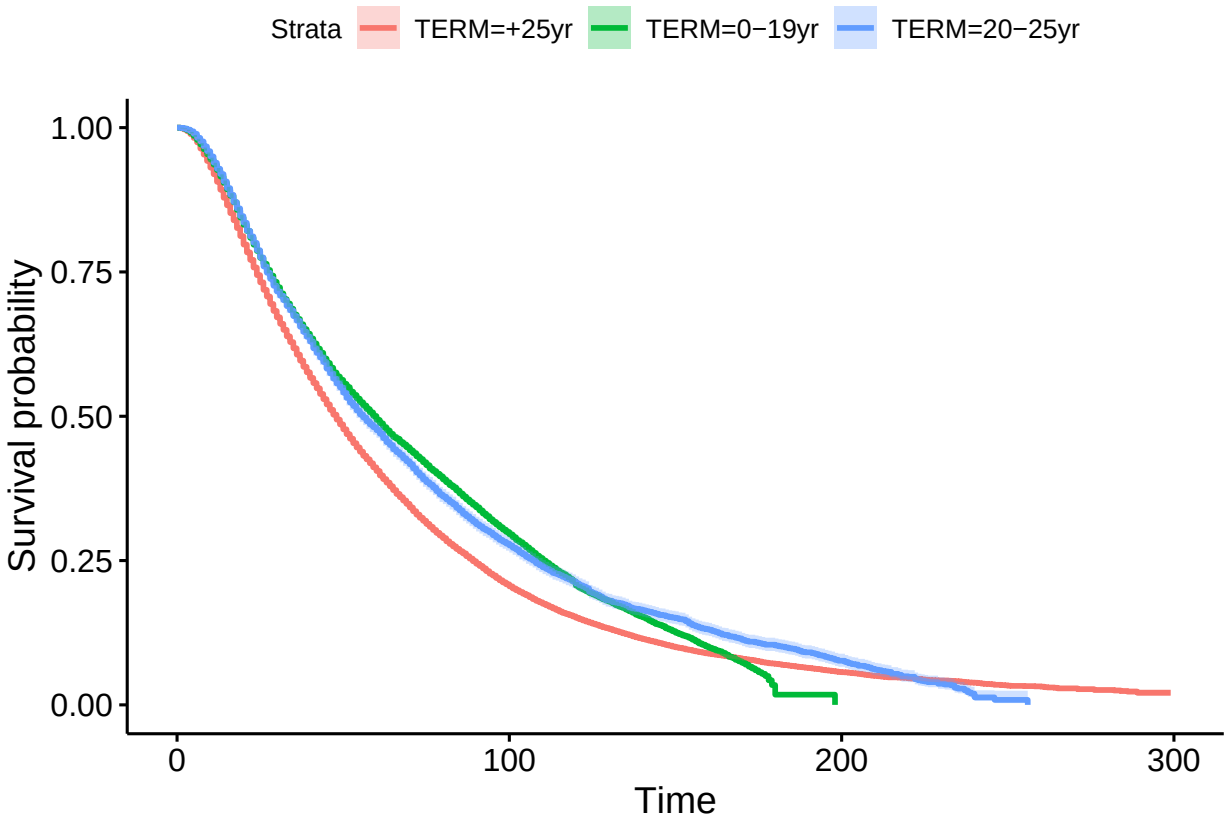
First, let us plot the Kaplan-Meier survival curve by pooling all the loans together. At this point, it is important to ensure that the data is formatted in terms of start time and stop time.

```r
KapM <- survfit(Surv(Start_time, Stop_time, Prepayment_status) ~ 1, data)
ggsurvplot(KapM,
           data = data,
           conf.int = TRUE,
           risk.table = FALSE,
           censor = FALSE)
```



It shows the survival curve, but this is not very meaningful for two reasons: first, all loans with different terms are pooled together, and second, there is no comparison among different loan characteristics. Therefore, I will plot the survival curves for loans with different characteristics to provide a more meaningful analysis.

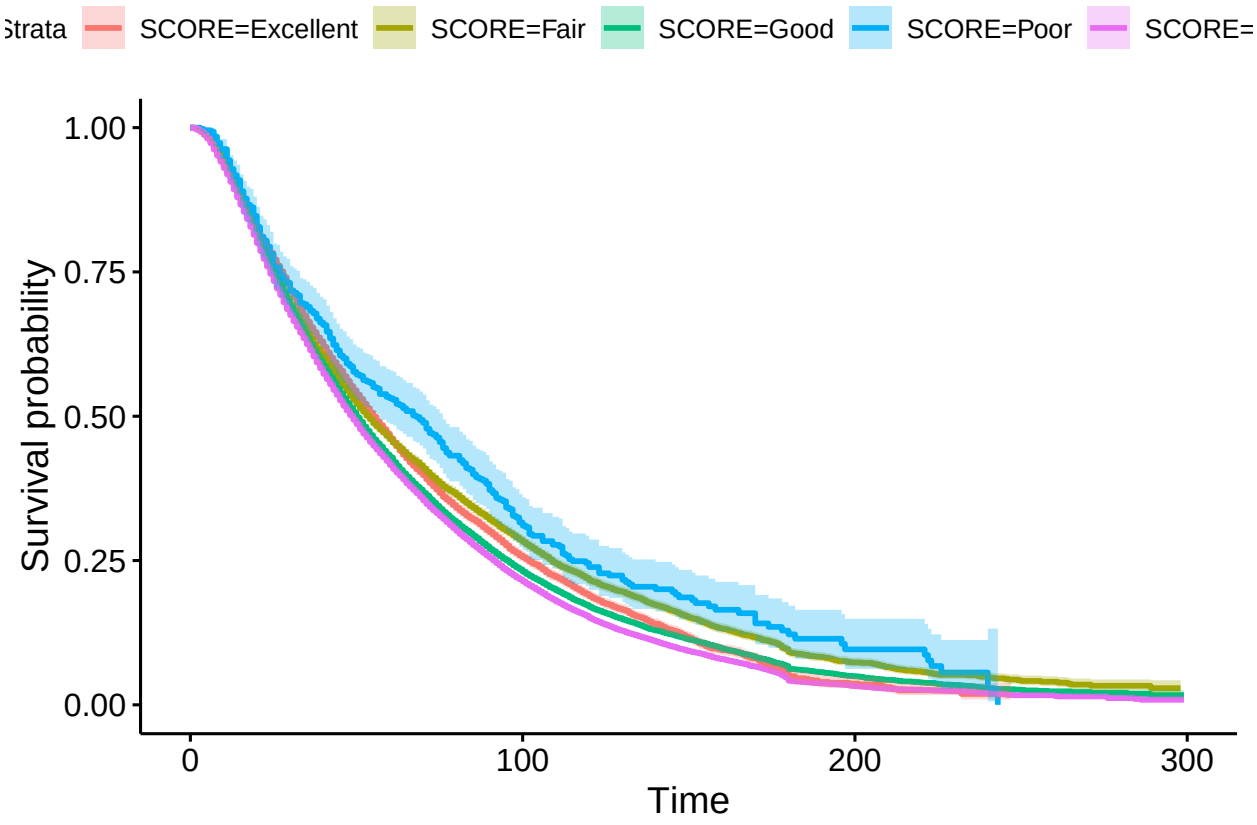
```r
KapM1 <- survfit(Surv(Start_time, Stop_time, Prepayment_status) ~ TERM, data)
ggsurvplot(KapM1,
           data = data,              # Provide the data here
           conf.int = TRUE,         # Do not show confidence interval
           risk.table = FALSE,       # Do not show risk table
           censor = FALSE)
```



As you can see, loans with different terms exhibit different structures, and these differences appear to be significant.

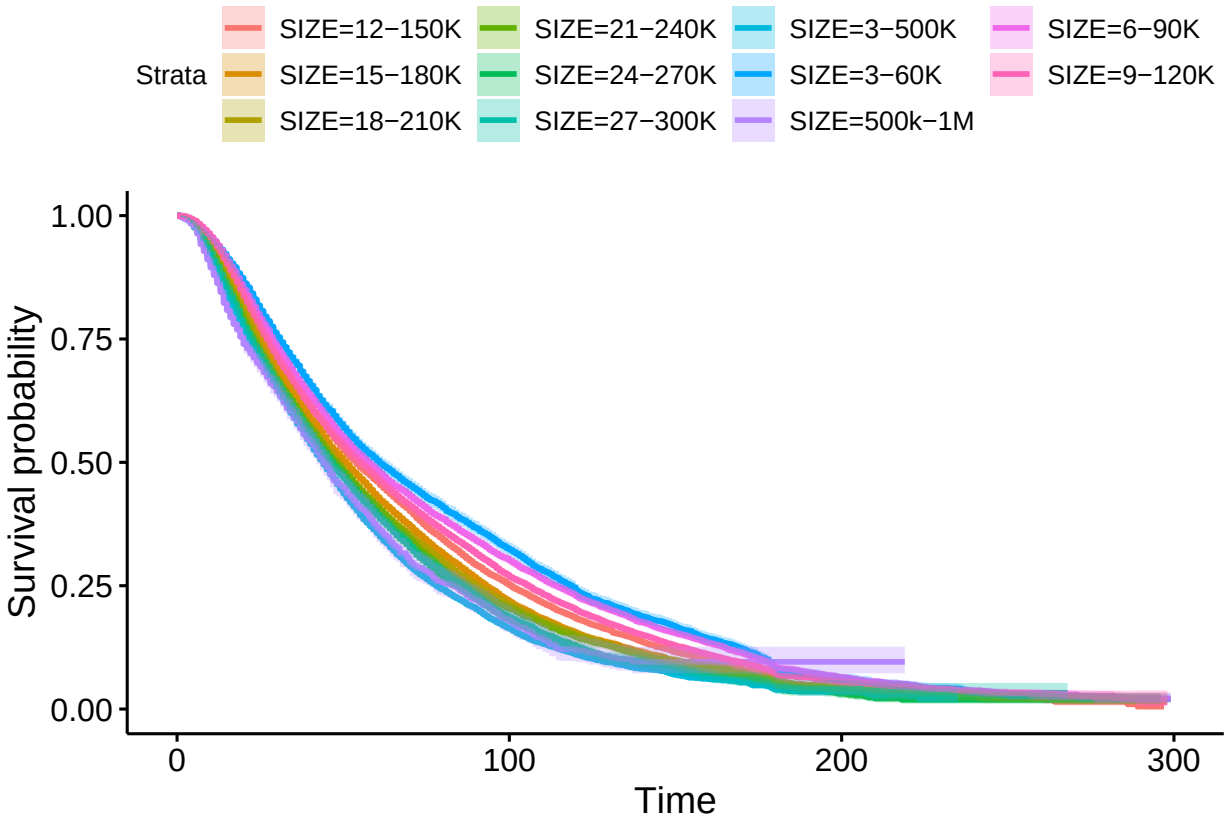
Next, let us plot the survival curves by credit score. Since some scores have missing values, we will exclude those cases. This does not affect the analysis, as the Kaplan-Meier curve is estimated independently for each score type. It seems that credit score does not have a strong impact on survival—at least without controlling for other factors.

```
filtered_data <- data %>% filter(SCORE != "")

KapM2 <- survfit(Surv(Start_time, Stop_time, Prepayment_status) ~ SCORE, filtered_data)

ggsurvplot(KapM2,
           data = filtered_data,
           conf.int = TRUE,
           risk.table = FALSE,
           censor = FALSE)
```



Now let us compare the loan size measured in terms of financed amount.

```
filtered_data <- data %>% filter(SIZE != "")
KapM3 <- survfit(Surv(Start_time, Stop_time, Prepayment_status) ~ SIZE, filtered_data)
ggsurvplot(KapM3,
           data = filtered_data,
           conf.int = TRUE,
           risk.table = FALSE,
           censor = FALSE)
```



Interest rate gap effect: Cox model

Our main interest is the interest rate gap, defined as the difference between the current rate and the long-run fixed rate, which is time-variant. To estimate the effect of time-dependent variables, we use the Cox Proportional Hazards model, which can handle time variance.

Now, let us use the interest rate gap along with our control variables to estimate the Cox Proportional Hazards model. The control variables include credit score, interest rate gap, unemployment rate, the initial mortgage amount to house price value ratio, loan size, debt-to-income ratio, and the original loan term. The estimates are shown below.

```
cox1 <-
  coxph(Surv(Start_time, Stop_time, Prepayment_status) ~ CREDIT_SCORE +
          Rate_gap + UNRATE + ORIGINAL_LOAN.TO.VALUE + ORIGINAL_UPB + ORIGINAL_DEBT.TO.INCOME +
          ORIGINAL_LOAN_TERM, data)
summary(cox1)
```

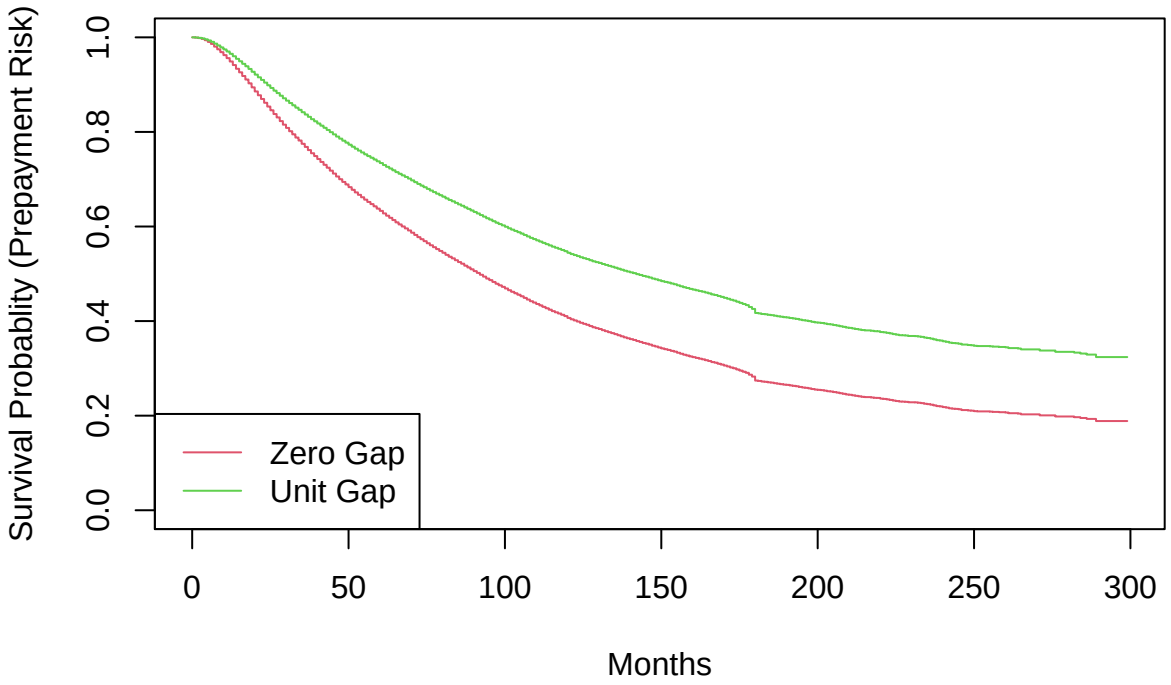
```
## Call:
## coxph(formula = Surv(Start_time, Stop_time, Prepayment_status) ~
##     CREDIT_SCORE + Rate_gap + UNRATE + ORIGINAL_LOAN.TO.VALUE +
##         ORIGINAL_UPB + ORIGINAL_DEBT.TO.INCOME + ORIGINAL_LOAN_TERM,
##     data = data)
##
##   n= 5231342, number of events= 72548
##    (7108 observations deleted due to missingness)
##
##                              coef  exp(coef)   se(coef)       z Pr(>|z|)
## CREDIT_SCORE             6.341e-04  1.001e+00  7.131e-05   8.892  < 2e-16 ***
## Rate_gap                -3.918e-01  6.759e-01  5.220e-03 -75.056  < 2e-16 ***
```

```
## UNRATE                   8.966e-02  1.094e+00  1.580e-03  56.763  < 2e-16 ***
## ORIGINAL_LOAN.TO.VALUE  -1.627e-03  9.984e-01  2.095e-04  -7.765 8.19e-15 ***
## ORIGINAL_UPB             1.723e-06  1.000e+00  3.363e-08  51.220  < 2e-16 ***
## ORIGINAL_DEBT.TO.INCOME -2.590e-04  9.997e-01  1.430e-05 -18.114  < 2e-16 ***
## ORIGINAL_LOAN_TERM      -1.001e-03  9.990e-01  5.513e-05 -18.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                         exp(coef) exp(-coef) lower .95 upper .95
## CREDIT_SCORE               1.0006     0.9994    1.0005    1.0008
## Rate_gap                   0.6759     1.4796    0.6690    0.6828
## UNRATE                     1.0938     0.9142    1.0904    1.0972
## ORIGINAL_LOAN.TO.VALUE     0.9984     1.0016    0.9980    0.9988
## ORIGINAL_UPB               1.0000     1.0000    1.0000    1.0000
## ORIGINAL_DEBT.TO.INCOME    0.9997     1.0003    0.9997    0.9998
## ORIGINAL_LOAN_TERM         0.9990     1.0010    0.9989    0.9991
##
## Concordance= 0.634  (se = 0.001 )
## Likelihood ratio test= 10714  on 7 df,   p=<2e-16
## Wald test            = 11174  on 7 df,   p=<2e-16
## Score (logrank) test = 11185  on 7 df,   p=<2e-16
```

Indeed, all of the variables are significant. What is notable here is the magnitude of their effects. Specifically, a higher rate gap, loan-to-value ratio, and debt-to-income ratio correspond to a lower probability of prepayment. When the rate gap is positive, it means the current rate is higher than the contracted fixed rate, thus reducing the incentive for mortgage owners to prepay (or refinance).

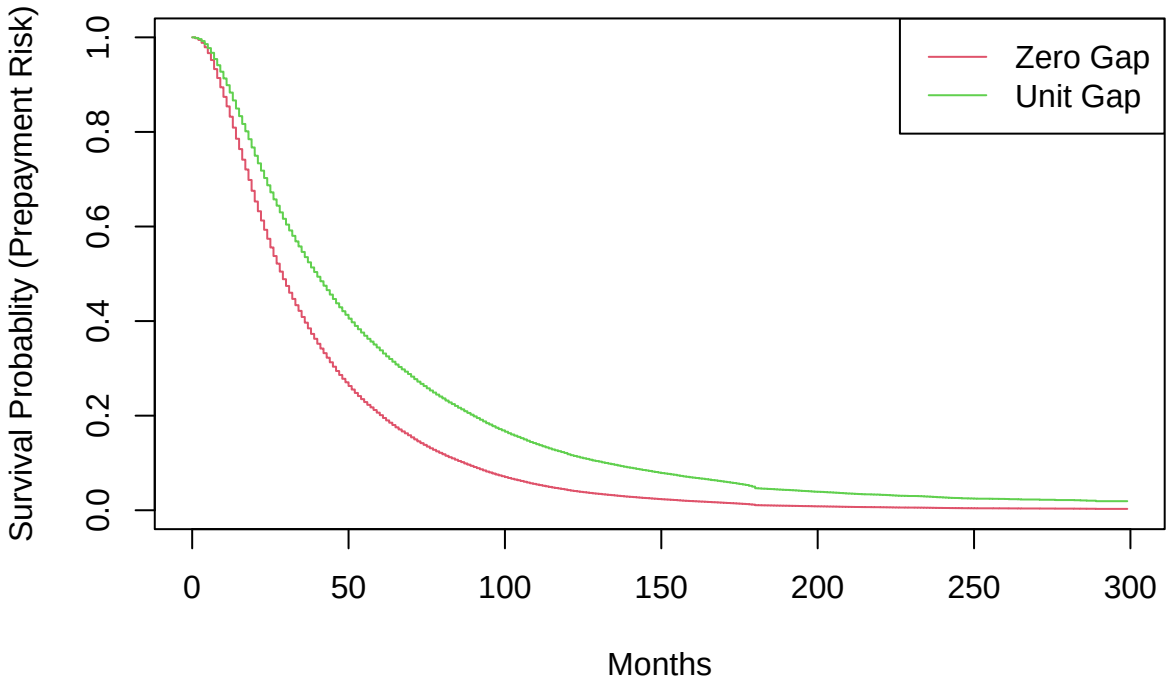
Now that we have modeled the hazard function using our variables of interest, let us examine the impact of a permanent shock in the rate gap on the survival curve. Specifically, we will plot the survival curve for a loan financed by a household with a credit score of 700, a loan-to-value ratio of 50%, an average loan size, an average debt-to-income ratio, and a loan term of 360 months, under the theoretical assumption of a zero unemployment rate.

```
newdata1 <- data.frame(
  CREDIT_SCORE = rep(700,2), Rate_gap = c(0, 1),
  UNRATE = rep(0,2), ORIGINAL_LOAN.TO.VALUE = rep(50,2),
  ORIGINAL_UPB = rep(mean(data$ORIGINAL_UPB),2),
  ORIGINAL_DEBT.TO.INCOME = rep(mean(data$ORIGINAL_DEBT.TO.INCOME),2),
  ORIGINAL_LOAN_TERM = rep(360,2)
)

survplots <- survfit(cox1, newdata1)

plot(survplots, xlab = "Months",
     ylab = "Survival Probablity (Prepayment Risk)", col = 2:5)

legend("bottomleft", c('Zero Gap','Unit Gap'), col = 2:5, lty = 1)
```



The results show that the rate gap is indeed significant. When the fixed rate is higher than the current rate, there is less incentive for prepayment, which causes the survival curve to shift upward. However, the question now is: "Does a recession have any amplifying effect on the rate gap?" As documented in the literature, economic fluctuations do significantly impact the effects of monetary policy. To investigate this, we will set the unemployment rate to 14% and examine its impact.

```
newdata1 <- data.frame(
  CREDIT_SCORE = rep(700,2), Rate_gap = c(0, 1),
  UNRATE = rep(14,2), ORIGINAL_LOAN.TO.VALUE = rep(50,2),
  ORIGINAL_UPB = rep(mean(data$ORIGINAL_UPB),2),
  ORIGINAL_DEBT.TO.INCOME = rep(mean(data$ORIGINAL_DEBT.TO.INCOME),2),
  ORIGINAL_LOAN_TERM = rep(360,2)
)

survplots <- survfit(cox1, newdata1)

plot(survplots, xlab = "Months",
     ylab = "Survival Probablity (Prepayment Risk)", col = 2:5)

legend("topright", c('Zero Gap','Unit Gap'), col = 2:5, lty = 1)
```



It appears that the entire curve has become steeper, but it is difficult to determine whether the gap between the two survival curves has widened. To formally test this, I will include an interaction term, UNRATE * Rate_gap, in the model. If the coefficient on the interaction term is positive, it would indicate that a recession has a diminishing effect on the sensitivity to the interest rate gap.

```
cox2 <-
  coxph(Surv(Start_time, Stop_time, Prepayment_status) ~ CREDIT_SCORE +
          Rate_gap + UNRATE + UNRATE*Rate_gap + ORIGINAL_LOAN.TO.VALUE +
          ORIGINAL_UPB + ORIGINAL_DEBT.TO.INCOME +
          ORIGINAL_LOAN_TERM, data)
summary(cox2)
```

```
## Call:
## coxph(formula = Surv(Start_time, Stop_time, Prepayment_status) ~
##     CREDIT_SCORE + Rate_gap + UNRATE + UNRATE * Rate_gap + ORIGINAL_LOAN.TO.VALUE +
##         ORIGINAL_UPB + ORIGINAL_DEBT.TO.INCOME + ORIGINAL_LOAN_TERM,
##     data = data)
##
##   n= 5231342, number of events= 72548
##    (7108 observations deleted due to missingness)
##
##                               coef  exp(coef)   se(coef)       z Pr(>|z|)
## CREDIT_SCORE             6.444e-04  1.001e+00  7.132e-05   9.036  < 2e-16 ***
## Rate_gap                -5.292e-01  5.891e-01  1.276e-02 -41.468  < 2e-16 ***
## UNRATE                   9.402e-02  1.099e+00  1.614e-03  58.248  < 2e-16 ***
## ORIGINAL_LOAN.TO.VALUE  -1.588e-03  9.984e-01  2.095e-04  -7.580 3.44e-14 ***
## ORIGINAL_UPB             1.709e-06  1.000e+00  3.363e-08  50.813  < 2e-16 ***
## ORIGINAL_DEBT.TO.INCOME -2.657e-04  9.997e-01  1.431e-05 -18.568  < 2e-16 ***
## ORIGINAL_LOAN_TERM      -9.898e-04  9.990e-01  5.510e-05 -17.963  < 2e-16 ***
## Rate_gap:UNRATE          2.298e-02  1.023e+00  1.957e-03  11.744  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                           exp(coef) exp(-coef) lower .95 upper .95
## CREDIT_SCORE                 1.0006     0.9994    1.0005    1.0008
## Rate_gap                     0.5891     1.6976    0.5745    0.6040
## UNRATE                       1.0986     0.9103    1.0951    1.1021
## ORIGINAL_LOAN.TO.VALUE       0.9984     1.0016    0.9980    0.9988
## ORIGINAL_UPB                 1.0000     1.0000    1.0000    1.0000
## ORIGINAL_DEBT.TO.INCOME      0.9997     1.0003    0.9997    0.9998
## ORIGINAL_LOAN_TERM           0.9990     1.0010    0.9989    0.9991
## Rate_gap:UNRATE              1.0233     0.9773    1.0193    1.0272
##
## Concordance= 0.634  (se = 0.001 )
## Likelihood ratio test= 10853  on 8 df,   p=<2e-16
## Wald test            = 11176  on 8 df,   p=<2e-16
## Score (logrank) test = 11202  on 8 df,   p=<2e-16
```

As you can see, the coefficient on the interaction term is positive and statistically significant. This indicates that the effect of the interest rate gap depends on economic conditions and diminishes during periods of higher unemployment (or amplifies during economic booms).

Formal Test by regression

Having demonstrated that, from a survival analysis perspective, the interest rate gap is significant and affects the survival curve accordingly, I will now use a simple probit model and a linear probability model to conduct a formal test, following the approach outlined by Berger et al. (2018)[https://www.aeaweb.org/articles?id=10.1257/aer.20181857].

Below, I run a pooled probit regression without controlling for time-variant heterogeneous factors. I include state fixed effects to account for state-specific heterogeneous factors. The results are indeed significant. Time-variant homogeneous factors are partially controlled for using the loan age variable, which progresses linearly over time.

```
probit1 <- glm(Prepayment_status ~ CREDIT_SCORE + Rate_gap + UNRATE +
                 ORIGINAL_LOAN.TO.VALUE + ORIGINAL_UPB +
                 ORIGINAL_DEBT.TO.INCOME +
                 ORIGINAL_LOAN_TERM + PROPERTY_STATE + LOAN_AGE,
               family =binomial(link = "probit"), data = data)

summary(probit1)
```

```
##
## Call:
## glm(formula = Prepayment_status ~ CREDIT_SCORE + Rate_gap + UNRATE +
##     ORIGINAL_LOAN.TO.VALUE + ORIGINAL_UPB + ORIGINAL_DEBT.TO.INCOME +
##     ORIGINAL_LOAN_TERM + PROPERTY_STATE + LOAN_AGE, family = binomial(link = "probit"),
##     data = data)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.530e+00  3.844e-02 -65.828  < 2e-16 ***
## CREDIT_SCORE              1.527e-04  2.801e-05   5.450 5.04e-08 ***
## Rate_gap                 -1.566e-01  2.079e-03 -75.307  < 2e-16 ***
## UNRATE                    3.932e-02  6.333e-04  62.076  < 2e-16 ***
## ORIGINAL_LOAN.TO.VALUE   -5.642e-04  8.513e-05  -6.627 3.42e-11 ***
## ORIGINAL_UPB              6.784e-07  1.484e-08  45.698  < 2e-16 ***
## ORIGINAL_DEBT.TO.INCOME -9.166e-05  5.447e-06 -16.828  < 2e-16 ***
## ORIGINAL_LOAN_TERM      -4.328e-04  2.170e-05 -19.944  < 2e-16 ***
```

```
## PROPERTY_STATEAL       -2.027e-02  3.360e-02  -0.603 0.546213
## PROPERTY_STATEAR       -2.290e-02  3.590e-02  -0.638 0.523449
## PROPERTY_STATEAZ        6.426e-02  3.167e-02   2.029 0.042435 *
## PROPERTY_STATECA        3.715e-02  3.069e-02   1.210 0.226097
## PROPERTY_STATECO        1.286e-01  3.167e-02   4.063 4.85e-05 ***
## PROPERTY_STATECT       -5.424e-02  3.364e-02  -1.612 0.106922
## PROPERTY_STATEDC       -1.244e-02  4.405e-02  -0.282 0.777564
## PROPERTY_STATEDE       -3.139e-03  3.919e-02  -0.080 0.936145
## PROPERTY_STATEFL       -7.203e-02  3.099e-02  -2.325 0.020094 *
## PROPERTY_STATEGA       -3.989e-02  3.154e-02  -1.265 0.205916
## PROPERTY_STATEGU       -5.185e-01  1.187e-01  -4.367 1.26e-05 ***
## PROPERTY_STATEHI       -7.956e-02  3.993e-02  -1.992 0.046331 *
## PROPERTY_STATEIA        6.702e-02  3.371e-02   1.988 0.046827 *
## PROPERTY_STATEID        1.088e-01  3.550e-02   3.065 0.002178 **
## PROPERTY_STATEIL        4.240e-02  3.104e-02   1.366 0.171864
## PROPERTY_STATEIN        3.714e-02  3.180e-02   1.168 0.242854
## PROPERTY_STATEKS        5.190e-02  3.392e-02   1.530 0.126032
## PROPERTY_STATEKY        3.125e-02  3.268e-02   0.956 0.338922
## PROPERTY_STATELA       -3.678e-02  3.421e-02  -1.075 0.282347
## PROPERTY_STATEMA        2.846e-02  3.179e-02   0.895 0.370613
## PROPERTY_STATEMD       -2.174e-02  3.191e-02  -0.681 0.495631
## PROPERTY_STATEME        2.215e-02  3.801e-02   0.583 0.560118
## PROPERTY_STATEMI        3.201e-02  3.127e-02   1.024 0.305957
## PROPERTY_STATEMN        3.745e-02  3.159e-02   1.186 0.235756
## PROPERTY_STATEMO        6.433e-02  3.197e-02   2.012 0.044228 *
## PROPERTY_STATEMS       -3.653e-02  3.917e-02  -0.932 0.351113
## PROPERTY_STATEMT        3.562e-02  3.821e-02   0.932 0.351252
## PROPERTY_STATENC       -2.938e-05  3.142e-02  -0.001 0.999254
## PROPERTY_STATEND        9.640e-02  4.368e-02   2.207 0.027337 *
## PROPERTY_STATENE        8.024e-02  3.581e-02   2.240 0.025067 *
## PROPERTY_STATENH        4.272e-02  3.568e-02   1.197 0.231116
## PROPERTY_STATENJ       -3.675e-02  3.166e-02  -1.161 0.245707
## PROPERTY_STATENM       -5.709e-02  3.644e-02  -1.567 0.117158
## PROPERTY_STATENV       -4.280e-03  3.409e-02  -0.126 0.900083
## PROPERTY_STATENY       -1.302e-01  3.131e-02  -4.159 3.19e-05 ***
## PROPERTY_STATEOH        1.048e-02  3.128e-02   0.335 0.737523
## PROPERTY_STATEOK       -1.097e-02  3.444e-02  -0.319 0.750044
## PROPERTY_STATEOR        5.309e-02  3.224e-02   1.647 0.099598 .
## PROPERTY_STATEPA       -1.982e-02  3.141e-02  -0.631 0.527889
## PROPERTY_STATEPR       -3.035e-01  4.552e-02  -6.667 2.61e-11 ***
## PROPERTY_STATERI        1.334e-02  3.961e-02   0.337 0.736348
## PROPERTY_STATESC       -5.659e-03  3.265e-02  -0.173 0.862401
## PROPERTY_STATESD        1.009e-01  4.519e-02   2.234 0.025509 *
## PROPERTY_STATETN        2.990e-02  3.238e-02   0.923 0.355785
## PROPERTY_STATETX       -1.072e-02  3.096e-02  -0.346 0.729261
## PROPERTY_STATEUT        1.103e-01  3.278e-02   3.365 0.000766 ***
## PROPERTY_STATEVA       -2.124e-03  3.159e-02  -0.067 0.946392
## PROPERTY_STATEVI       -1.110e-01  2.267e-01  -0.490 0.624269
## PROPERTY_STATEVT        1.249e-02  3.895e-02   0.321 0.748507
## PROPERTY_STATEWA        4.407e-02  3.153e-02   1.398 0.162140
## PROPERTY_STATEWI        1.249e-01  3.178e-02   3.929 8.55e-05 ***
## PROPERTY_STATEWV       -6.006e-02  4.050e-02  -1.483 0.138084
## PROPERTY_STATEWY        1.156e-01  4.570e-02   2.529 0.011429 *
## LOAN_AGE                1.568e-04  4.088e-05   3.835 0.000126 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 764831  on 5231341  degrees of freedom
## Residual deviance: 751675  on 5231280  degrees of freedom
##   (7108 observations deleted due to missingness)
## AIC: 751799
##
## Number of Fisher Scoring iterations: 7
```

Below, I use a linear probability model for convenience in interpretation. I include the same variables as in the previous analysis.

```
lpm <- lm(Prepayment_status ~ CREDIT_SCORE + Rate_gap + UNRATE + ORIGINAL_LOAN.TO.VALUE + ORIGINAL_UPB +

summary(lpm)
```

```
##
## Call:
## lm(formula = Prepayment_status ~ CREDIT_SCORE + Rate_gap + UNRATE +
##     ORIGINAL_LOAN.TO.VALUE + ORIGINAL_UPB + ORIGINAL_DEBT.TO.INCOME +
##     ORIGINAL_LOAN_TERM + PROPERTY_STATE + LOAN_AGE, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05270 -0.01758 -0.01321 -0.00947  1.00695
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.800e-03  1.331e-03   2.104 0.035337 *
## CREDIT_SCORE             4.911e-06  9.667e-07   5.080 3.77e-07 ***
## Rate_gap                -5.697e-03  7.706e-05 -73.929  < 2e-16 ***
## UNRATE                   1.503e-03  2.408e-05  62.407  < 2e-16 ***
## ORIGINAL_LOAN.TO.VALUE  -1.892e-05  2.869e-06  -6.596 4.22e-11 ***
## ORIGINAL_UPB             2.544e-08  5.483e-10  46.397  < 2e-16 ***
## ORIGINAL_DEBT.TO.INCOME -2.926e-06  1.737e-07 -16.845  < 2e-16 ***
## ORIGINAL_LOAN_TERM      -1.673e-05  7.518e-07 -22.252  < 2e-16 ***
## PROPERTY_STATEAL        -5.226e-04  1.161e-03  -0.450 0.652635
## PROPERTY_STATEAR        -5.621e-04  1.232e-03  -0.456 0.648146
## PROPERTY_STATEAZ         2.414e-03  1.104e-03   2.187 0.028753 *
## PROPERTY_STATECA         1.509e-03  1.066e-03   1.415 0.156940
## PROPERTY_STATECO         5.086e-03  1.110e-03   4.581 4.63e-06 ***
## PROPERTY_STATECT        -1.650e-03  1.158e-03  -1.426 0.154001
## PROPERTY_STATEDC        -4.558e-04  1.567e-03  -0.291 0.771146
## PROPERTY_STATEDE         6.503e-05  1.357e-03   0.048 0.961779
## PROPERTY_STATEFL        -2.223e-03  1.073e-03  -2.072 0.038310 *
## PROPERTY_STATEGA        -1.167e-03  1.091e-03  -1.070 0.284735
## PROPERTY_STATEGU        -1.165e-02  2.711e-03  -4.298 1.73e-05 ***
## PROPERTY_STATEHI        -2.753e-03  1.376e-03  -2.000 0.045453 *
## PROPERTY_STATEIA         2.515e-03  1.175e-03   2.140 0.032376 *
## PROPERTY_STATEID         4.124e-03  1.261e-03   3.269 0.001079 **
## PROPERTY_STATEIL         1.624e-03  1.078e-03   1.506 0.132040
## PROPERTY_STATEIN         1.524e-03  1.104e-03   1.380 0.167607
```

```
## PROPERTY_STATEKS         1.974e-03  1.183e-03   1.668 0.095362 .
## PROPERTY_STATEKY         1.275e-03  1.134e-03   1.124 0.260899
## PROPERTY_STATELA        -1.056e-03  1.178e-03  -0.897 0.369943
## PROPERTY_STATEMA         1.051e-03  1.107e-03   0.949 0.342637
## PROPERTY_STATEMD        -7.547e-04  1.107e-03  -0.682 0.495345
## PROPERTY_STATEME         8.311e-04  1.324e-03   0.628 0.530052
## PROPERTY_STATEMI         1.350e-03  1.085e-03   1.244 0.213432
## PROPERTY_STATEMN         1.384e-03  1.097e-03   1.261 0.207215
## PROPERTY_STATEMO         2.381e-03  1.114e-03   2.139 0.032466 *
## PROPERTY_STATEMS        -1.033e-03  1.340e-03  -0.771 0.440433
## PROPERTY_STATEMT         1.371e-03  1.343e-03   1.021 0.307241
## PROPERTY_STATENC         8.287e-05  1.090e-03   0.076 0.939388
## PROPERTY_STATEND         3.529e-03  1.558e-03   2.266 0.023461 *
## PROPERTY_STATENE         2.924e-03  1.256e-03   2.327 0.019945 *
## PROPERTY_STATENH         1.573e-03  1.253e-03   1.256 0.209156
## PROPERTY_STATENJ        -1.201e-03  1.098e-03  -1.094 0.273960
## PROPERTY_STATENM        -1.712e-03  1.244e-03  -1.376 0.168699
## PROPERTY_STATENV        -9.836e-05  1.183e-03  -0.083 0.933715
## PROPERTY_STATENY        -4.079e-03  1.081e-03  -3.772 0.000162 ***
## PROPERTY_STATEOH         6.078e-04  1.084e-03   0.561 0.575060
## PROPERTY_STATEOK        -2.300e-04  1.191e-03  -0.193 0.846861
## PROPERTY_STATEOR         1.969e-03  1.125e-03   1.750 0.080035 .
## PROPERTY_STATEPA        -4.545e-04  1.087e-03  -0.418 0.675997
## PROPERTY_STATEPR        -7.639e-03  1.377e-03  -5.549 2.88e-08 ***
## PROPERTY_STATERI         4.372e-04  1.390e-03   0.314 0.753169
## PROPERTY_STATESC        -8.357e-06  1.132e-03  -0.007 0.994108
## PROPERTY_STATESD         3.685e-03  1.608e-03   2.292 0.021897 *
## PROPERTY_STATETN         1.180e-03  1.125e-03   1.049 0.293987
## PROPERTY_STATETX        -2.192e-04  1.074e-03  -0.204 0.838292
## PROPERTY_STATEUT         4.257e-03  1.152e-03   3.694 0.000221 ***
## PROPERTY_STATEVA        -2.447e-05  1.097e-03  -0.022 0.982203
## PROPERTY_STATEVI        -4.094e-03  7.724e-03  -0.530 0.596048
## PROPERTY_STATEVT         5.845e-04  1.356e-03   0.431 0.666441
## PROPERTY_STATEWA         1.650e-03  1.098e-03   1.503 0.132934
## PROPERTY_STATEWI         4.699e-03  1.110e-03   4.231 2.33e-05 ***
## PROPERTY_STATEWV        -1.473e-03  1.360e-03  -1.083 0.278844
## PROPERTY_STATEWY         4.444e-03  1.668e-03   2.664 0.007722 **
## LOAN_AGE                 6.150e-06  1.500e-06   4.099 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1168 on 5231280 degrees of freedom
##   (7108 observations deleted due to missingness)
## Multiple R-squared:  0.002586,   Adjusted R-squared:  0.002574
## F-statistic: 222.3 on 61 and 5231280 DF,  p-value: < 2.2e-16
```

Here, I use the fixed effects within transformation with a linear probability model to control for loan-specific time-invariant heterogeneity. This approach accounts for factors such as the financial illiteracy of mortgage owners, which are unobservable but likely constant over the sample period. Note that due to the within transformation, time-invariant variables cannot be directly estimated. To address this, I included an interaction term between credit score and loan age to examine the effect of credit score over time. The results remain significant.

```
pdata <- pdata.frame(data, index = c("LOAN_SEQUENCE_NUMBER", "DATE"))
```

```r
lpm_fx <- plm(Prepayment_status ~ LOAN_AGE*CREDIT_SCORE + Rate_gap + UNRATE + LOAN_AGE, data=pdata, mod

summary(lpm_fx)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = Prepayment_status ~ LOAN_AGE * CREDIT_SCORE + Rate_gap +
##     UNRATE + LOAN_AGE, data = pdata, model = "within")
##
## Unbalanced Panel: n = 95835, T = 1-298, N = 5231534
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -0.50180879 -0.02385668 -0.01111899  0.00070372  1.00244694
##
## Coefficients:
##                           Estimate  Std. Error  t-value  Pr(>|t|)
## LOAN_AGE              -6.6346e-04  2.2411e-05  -29.604 < 2.2e-16 ***
## Rate_gap              -2.5995e-02  2.4996e-04 -104.000 < 2.2e-16 ***
## UNRATE                 2.6056e-03  2.9657e-05   87.858 < 2.2e-16 ***
## LOAN_AGE:CREDIT_SCORE  1.0931e-06  3.0191e-08   36.207 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    69378
## Residual Sum of Squares: 68504
## R-Squared:      0.012608
## Adj. R-Squared: -0.0058177
## F-statistic: 16394.6 on 4 and 5135695 DF, p-value: < 2.22e-16
```

Conclusion

In this paper, I demonstrated the importance of considering the interest rate gap by using survival analysis. Additionally, I showed that economic fluctuations amplify the effect of the rate gap on prepayment in a pro-cyclical manner. Finally, formal tests using the probit model and the linear probability fixed effects model confirmed that the results are robust.