Mortgage Defaults: Inferential Statistics

Data Wrangling

First, the data must be structured to to create a dataframe with a single observation for each mortgage, so the data are grouped by mortgage ID number and a dataframe is created of loan characteristics, including: loan-to-value ratio at origination, FICO score at origination, interest rate at origination, type of borrower (investor vs non-investor), and type of property (single family home, condo, or urban development).

Assumptions

Unfortunately, standard statistical tests for differences in rates of default cannot be performed due to the presence of right censoring in the data. The assumption would need to be made that probability of default is unrelated to whether an observation is right censored. This would be an unrealistic assumption as mortgages that span longer periods of time tend to have a lower probability of default and also have a higher probability of being censored. This is a type of selection bias and would invalidate the tests. Therefore, tests will be restricted to characteristics of the loans known at time of origin.

Furthermore, the independence assumption is crucial to the validity of statistical inference and is complicated by the fact that the dataset is a time series spanning 15 years (60 quarters). Assuming independence across time would likely be implausible even under normal circumstances. However, our dataset includes the financial crisis which would significantly skew the results based on the time the loan was originated. To attempt to deal with this complication, tests will only be performed between mortgages originated in the same quarter. Despite losing statistical power by limiting our sample sizes for tests, the dataset still provides enough observations to easily meet the minimum 30-40 observations and at least 5 successes and 5 failures for each group to enable the use of the Central Limit Theorem.

Statistical Tests

First differences in characteristics of loans at origination will be explored for investors vs non-investors. T-tests will be performed to test whether there is a statistically significant difference in the mean of FICO scores and loan-to-value ratios for investors vs non-investors at time of origination.

Investors vs Non-Investors

FICO score at origination

Null: Investors and non-investors have the same mean FICO score at origination

Alternative: Investors and non-investors do not have the same mean FICO score

Mean for investors = 700.2 and for non-investors = 662.8
The test statistic is 15.9 and the p-value is 2.6 * 10^-51 so the null hypothesis of equal means is rejected.

Loan-to-value ratio at origination

Null: Investors and non-investors have the same mean LTV ratio at origination
Alternative: Investors and non-investors do not have the same mean LTV ratio at origination

Mean for investors = 77.71 and for non-investors = 79.98
The test statistic is 6.1 and the p-value is 1.8 * 10^-9 so the null hypothesis of equal means is rejected.

Single Family vs Condos

FICO score at origination

Null: Single family and condo mortgages have the same mean FICO score at origination
Alternative: Single family and condo mortgages do not have the same mean FICO score

Mean for single family = 659.7 and for condos = 689.7
The test statistic is -9.1 and the p-value is 1.8 * 10^-18 so the null hypothesis of equal means is rejected.

Loan-to-value ratio at origination

Null: Single family and condo mortgages have the same mean LTV ratio at origination
Alternative: Single family and condo mortgages do not have the same mean LTV ratio at origination

Mean for single family = 79.8 and for condos = 79.7
The test statistic is .14 and the p-value is 0.89 so the null hypothesis of equal means fails to be rejected.