Mortgage Interest Rate Determinants

The application of extremely large datasets in modern business is increasing in importance at an ever increasing rate, and real estate finance is no exception. There are trillions of dollars of outstanding loans at any given point in time which affects, either directly or indirectly, everyone in the financial system. Large datasets can help us understand key trends, risks, and how certain loan/borrower characteristics are related to one another. Ultimately, data analytics should help us make more informed decisions.

This assignment will provide you with hands-on experience in working with a portion of the large mortgage dataset available through Freddie Mac. This extensive collection of data is comprised of two parts, mortgage originations and mortgage performance. We will work with originations data, which includes roughly 53.6 million mortgages originated between January 1, 1999 and March 31, 2024.

To make this a workable project, I have randomly selected a much smaller sample of 75,750 observations. This subsample includes 3,000 origination records per year for the 1999 – 2023 period, and 750 records from the first three months of 2024. This is small enough to allow for analysis using Microsoft Excel, and large enough to give you the "big data" experience.

When performing a statistical analysis, it is good practice to spend some time looking at the data source. To aid you in familiarizing yourself with the Freddie Mac data, I have posted a document on Course Site (Single-Family Loan-Level Dataset General User Guide) that explains the data file layout and matches variable numbers with variable names. Note that I added a variable (oyear) to the Freddie Mac Data that indicates in which year the loan was originated. Further, pay close attention to Freddie Mac's arbitrary use of numbers such as "9", "99", "999", or "9999" to represent data not being available for a particular field for a given observation. If you don't erase such numbers, your computations will be incorrect. This is common practice with large datasets.

Deliverables:
You will need to turn in an analysis that includes the following five elements on or before May 10, 2025. **This is to be your own work** (e.g. no working with other past- or present-students.)

1) Create a table of simple summary statistics showing the mean, median, standard deviation, minimum, maximum, and count[†] for the following variables: Original Interest Rate (Rate)*, Original Unpaid Principal Balance (UPB)[†], Original Loan Term (Term)[†], Original Loan-to-Value (LTV)*, Original Debt-to-Income Ratio (DTI)*, Borrower Credit Score (FICO)[†], Number of Borrowers (Borrowers)*, First Time Home Buyer (First)[‡,]**, and Prepayment Penalty (Penalty)[‡,]***. [10 points]

2) Create a table showing simple (i.e. unconditional) pairwise correlations between all variables from #1 above**. Pay attention to the direction and strength of the relations. [10 points]

3) Create a table showing simple averages for each of the following variables, by origination year: Original Interest Rate*, Loans used for a Home Purchase (Purchase)[‡,]**, Original Unpaid Principal Balance (UPB)[†], Original Loan-to-Value (LTV)*, Original Debt-to-Income Ratio (DTI)*, Borrower Credit Score (FICO)[†], Loans with Credit Score below 680 (< 680)[‡,]**, and Loans with a Single Borrower (Single)[‡,]**. Also, include a totals row with the overall

averages. While some of these variables already exist, you will have to generate others using information available to you in the dataset. [20 points]

4) Create a table that lists all the lenders who sold loans to Freddie Mac (Seller), sorted in descending order based on the number of loans. The banks should be numbered (e.g., 1. XYZ Bank, 2. ABC MORTGAGE COMPANY, LLC, and so forth). Include columns showing the frequency, percent of total, and cumulative percent for each lender as well. There should be five columns ordered as follows: Number†, Seller, Frequency†, Percent**, Cumulative**. Only group by identical lender names in your tabulations. Resist the urge to combine highly similar names. [20 points]

5) Using only data from years 2013 – 2024, perform a regression analysis with the original interest rate as the dependent variable and the following explanatory variables: Original Unpaid Principal Balance (UPB), Original Loan Term (Term), Original Loan-to-Value (LTV), Debt-to-Income Ratio (DTI), and Borrower Credit Score (FICO).

   Also, include individual year indicator control variables for years 2014 – 2024 (i.e. for each year create a variable that is populated by a 1 if the observation was originated during that calendar year, and zero otherwise. The 2013 observations will serve as the base year, so we don't include an indicator variable for it.) For reference, the complete regression equation is presented below. This is a fairly standard representation of a regression. Note that there are 11 separate year dummies, variables 6 – 16, that are shown in condensed form in the equation. Present the regression results in a table. Don't alter the formatting, just use the default Excel output. [30 points]

$$\text{Rate}_i = \alpha + \beta_1\, UPB_i + \beta_2 Term_i + \beta_3 LTV_i + \beta_4 DTI_i + \beta_5 FICO_i + \beta_{6-16} Year\ Dummies_{6-16,i} + \varepsilon_i$$

Final Discussion:
Discuss the analysis in depth. To help you get started, consider the following questions: What are the numbers telling you? Are there trends, magnitudes, or relations that stand out? Are the coefficient signs in the regression table consistent with those observed in the simple correlation table? What about the statistical significance? Is there anything you did or did not expect? Is there anything strange that you think would merit further empirical investigation and why? Etc. [10 points]

Grading:
Point allocations for the assignment are based on accuracy, professional presentation/neatness, and the quality of your comments in the discussion. The assignment is worth 100 points; 80 for accuracy, 10 for a clean presentation, 10 for the quality of your comments. A breakdown of the accuracy points for each individual table is shown in [brackets] in the instructions above. I will deduct one point for each number that is incorrect, up to the maximum shown in [brackets] for a given table.

Resources:
As noted above, resource materials are available to you on Course Site. They can be found under the heading "Data Analytics Project." I have also posted a paper by University of Chicago law professor Alan O. Sykes, "An Introduction to Regression Analysis", which provides helpful background on regression that is not overly technical.

You may need to install the Data Analysis TookPak in Excel if you haven't already done so.
[Excel → File → Options → Add-Ins → Manage Add-Ins → Analysis Tool Pack → OK]
[Excel → Tools → Excel Add-Ins → Analysis ToolPak]

† Do not display any decimal places
* Display one decimal place
** Display two decimal places
*** Display three decimal places
‡ Simply create an indicator (also sometimes called a "dummy") variable that equals 1 if a specific condition is met, and 0 otherwise.