FoGM: Project Milestone

Andres Potapczynski (ap3635)

The current document is ordered in the following manner. First, I set the context of the problem by describing the data set and some key features about it. Here I also re-state the questions that we (the start-up and I) want to answer. Next I show some preliminary models that I ran in order to make sense of the data. The idea is to find the set variables that we should include in the final model and to understand at what level of granularity are those variables meaningful. Then I show a proposal of a graphical model that I then plan to estimate via VI. Finally, I conclude with a pipeline of models that I will be running in the following month as well as some questions where I would like to get some feedback on.

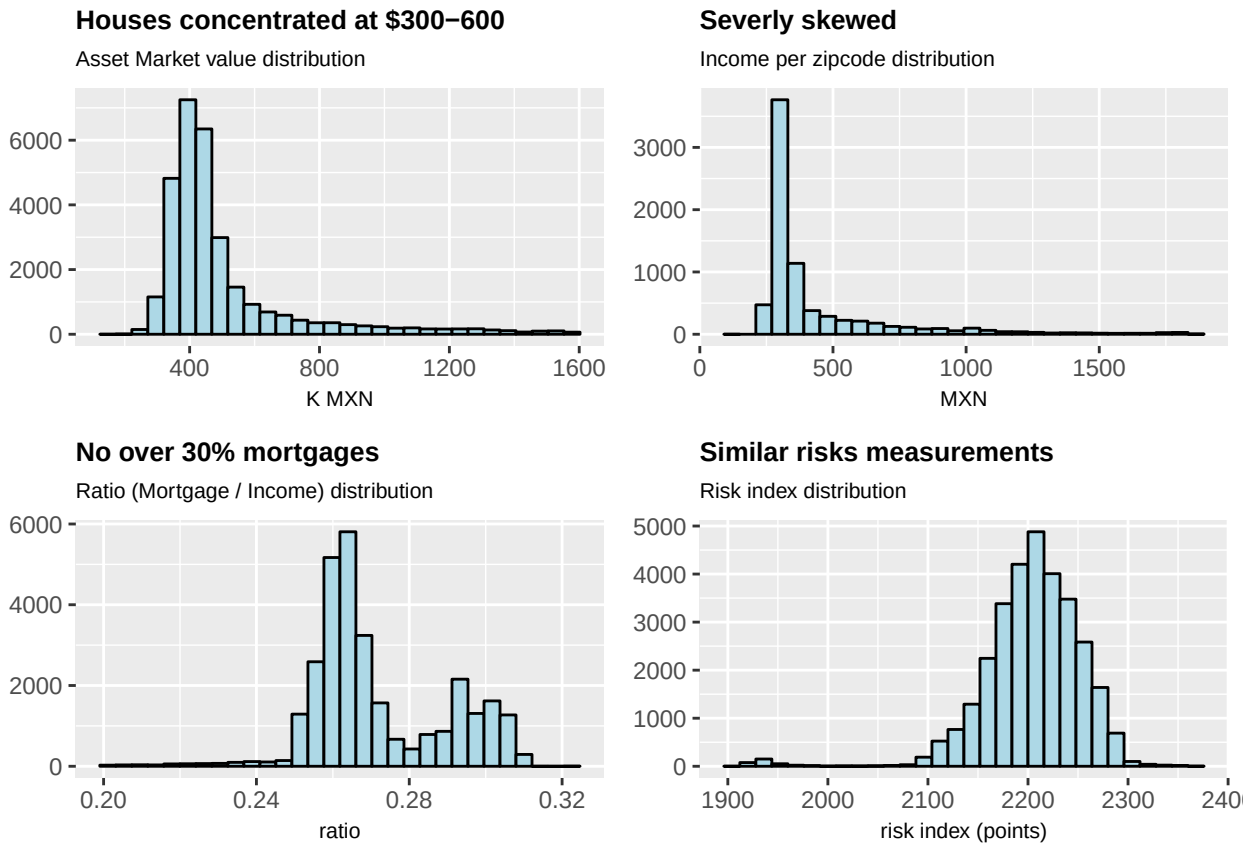
Setting the preamble

The data set was provided by a mortgage lending start-up company in Mexico. This data set consists of **30,499 mortgages with over 90 covariates** (where some of features are simply administrative and thus were not included in the analysis). **The average default rate is 6%**. Also, the data set provided was **their latest report available at August 2018**; if needed, I could incorporate historical information of previous months. Some of the key variables are:

- [**Location**] `state`, `city` and `zip`: These features will allow me to understand if the client's behavior varies by geography. Also, I have this location features for both the house acquired and for the owner's location.
- [**Demographics**] `age`, `sex`, `ratio`, `risk_index`, `client_income` and `credit_score`. The feature `ratio` is the % of the client's income that the monthly payments for its mortgage represent. The `risk_index` is a variable that combines several metrics associated with his likelihood of paying a debt. The `client_income` feature is the monthly income that the person earned at that moment in time when she signed the mortgage. Finally, the ordinal variable `credit_score` evaluates how the client has performed in previous debts (related to `risk_index`).
- [**Asset features**] `vendor name`, `new_used`, `inv` and `appraisal_value`: These features tell us who was the vendor (either a construction company or an individual), if the asset is new or not, then `inv` (made up feature) is a dummy variable that takes the value of 1 if the house bought is in the same state where the owner lives and the last variable is the price of the asset.
- [**Employment**] `employer_name` and `factor_employed`. The first feature conveys who is the employer and the second one information about the client's status of employment.
- [**Payment Records**] `days_pay`, and `y`. The first feature counts the number of days that have passed between the last payment date and the date on which the mortgage started. The next feature is the target variable that is 1 if the mortgage has at least one month without a payment and 0 otherwise.

Note that features like `interest_rate` and `contract_length` are present but have no variance. The current *product offer* in the data base is a 12% interest rate mortgage for 30 years. Thus it is impossible to pose counter factual questions for those variables.

The empirical distribution of some of the previous key variables is



Understanding the data

Personally, the best way to understand a data set is to run different models and see where they fail. I ran the next set of models in `STAN`. Monte Carlo sampling is adequate for this aggregate top-down analysis, however, for the final model the complexity will require us to derive VI updates.

Model 1 (Simple Logistic Regression):

$$\beta \sim N(\mu, \sigma^2)$$
$$y_i \sim Ber(logistic^{-1}(X\beta))$$

where $i = 1, \ldots, N$ where N is the total number of mortgages. Also, X is augmented to include an intercept and the covariates where standardize. The results are the following.

```
## Inference for Stan model: logistic_reg_v01.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##           mean se_mean  sd    2.5%     25%     50%     75%   97.5% n_eff
## alpha     -3.5     0.0 0.1    -3.6    -3.5    -3.5    -3.4    -3.4  1757
## beta[1]   -0.4     0.0 0.0    -0.4    -0.4    -0.4    -0.3    -0.3  1693
## beta[2]   -0.4     0.0 0.0    -0.4    -0.4    -0.4    -0.4    -0.3  1856
## beta[3]    0.1     0.0 0.0     0.0     0.0     0.1     0.1     0.1  2329
## beta[4]    0.1     0.0 0.1     0.0     0.1     0.1     0.2     0.2  2530
## beta[5]    0.0     0.0 0.1    -0.1    -0.1     0.0     0.0     0.1  2060
## beta[6]   -1.4     0.0 0.0    -1.4    -1.4    -1.4    -1.4    -1.3  2052
## beta[7]   -0.2     0.0 0.0    -0.3    -0.2    -0.2    -0.2    -0.1  2338
## beta[8]    0.1     0.0 0.1    -0.1     0.0     0.1     0.1     0.2  2098
## beta[9]   -0.2     0.0 0.0    -0.3    -0.2    -0.2    -0.2    -0.2  2474
## lp__    -5728.5     0.1 2.2 -5733.8 -5729.6 -5728.2 -5726.9 -5725.2   991
##         Rhat
## alpha      1
## beta[1]    1
## beta[2]    1
## beta[3]    1
## beta[4]    1
## beta[5]    1
## beta[6]    1
## beta[7]    1
## beta[8]    1
## beta[9]    1
## lp__       1
##
## Samples were drawn using NUTS(diag_e) at Fri Nov 16 23:13:04 2018.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

The order of the variables is the following: `client_income`, `ratio`, `age`, `sex` (dummy), `new_used` (dummy), `days_pay`, `factor_employed`, `inv` and `risk_index`. At this aggregate level the only relevant variable appears to amounts of days that the client has payed his mortgage `days_pay` and the intercept α.

Model 2 (Location Hierarchical Logistic Regression):

$$\alpha \sim Ga(a_0, b_0)$$
$$\beta \sim Ga(a_0, b_0)$$

where for each zip code

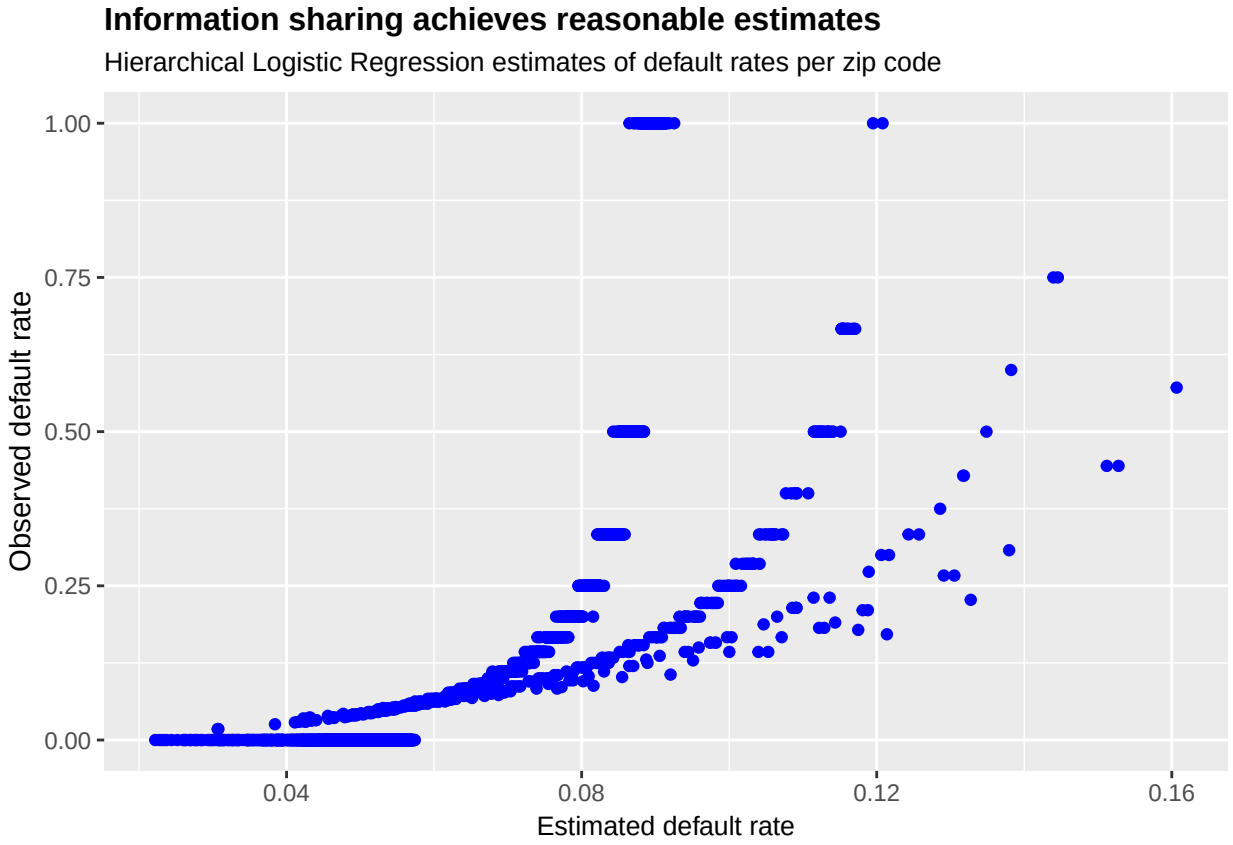
$$\theta_j \sim Beta(\alpha, \beta)$$
$$y_j \sim Bin(n_j, \theta_j)$$

where now y_j represents the accumulated cases of default per zip code. The idea here it to uncover if there are some areas that are more prone to default. The results are summarized in the next plot.

```
## Inference for Stan model: zip_code_v01.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##           mean se_mean    sd     2.5%      25%      50%      75%    97.5%
```

```
## alpha      1.9     0.0   0.2      1.6      1.8      1.9      2.0      2.3
## beta      26.7     0.3   2.4     22.3     25.0     26.7     28.3     31.6
## lp__  -16159.5    41.8 363.4 -16872.6 -16401.7 -16142.5 -15906.8 -15471.1
##       n_eff Rhat
## alpha    79  1.1
## beta     88  1.0
## lp__     76  1.1
##
## Samples were drawn using NUTS(diag_e) at Fri Nov 16 23:17:38 2018.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

where we see, as expected, that the parameters of the shared beta prior lean towards low values of θ_j. Note also how this model yielded a much higher log posterior than the vanilla logistic regression.



Also, the estimates are pooled towards their group mean, yet, there are still some zip codes that have a higher probability of default than others. Thus, one of the main insights is that we should model each spatial unit differently.

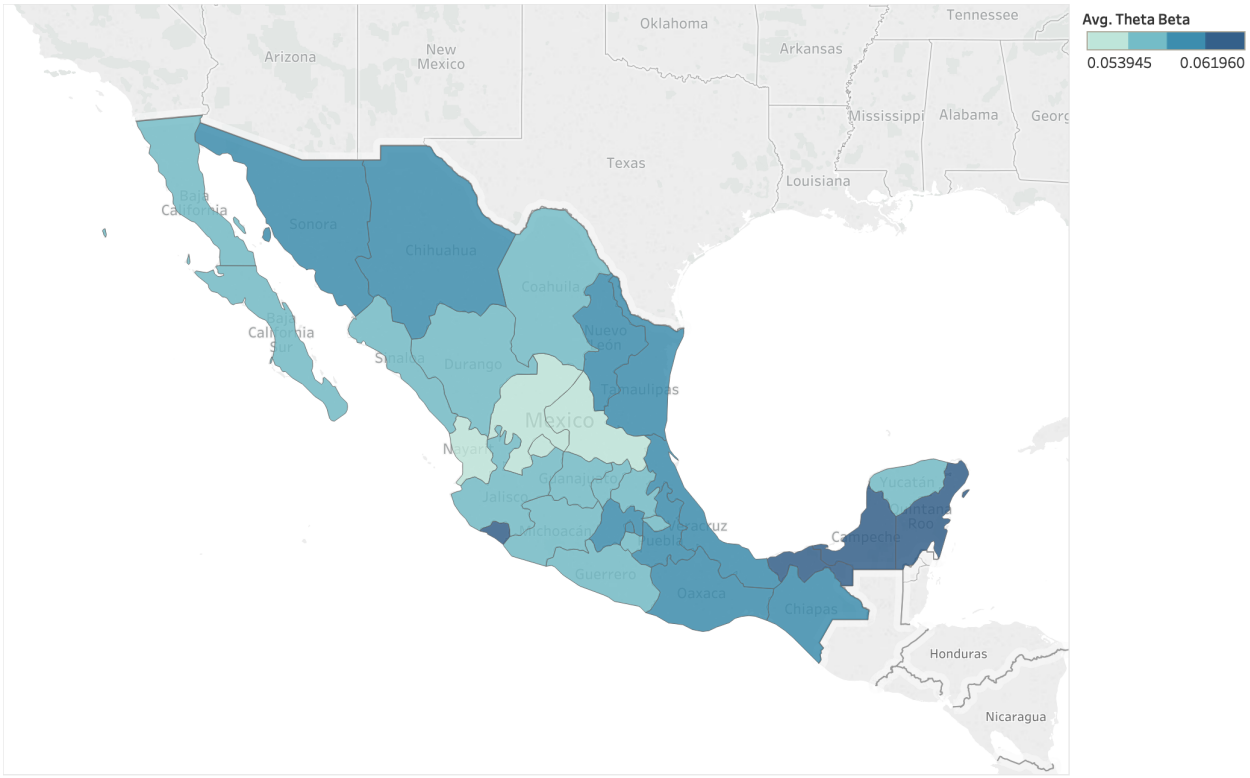


Figure 1: Mexico's States colored by average default estimates

Moreover, the data is consistent with our hypothesis that conflict areas in Mexico are having a collateral effect on the ability of the people to pay their mortgage. The northern states of Mexico are some of the wealthiest states but also the states colored in darker blue are exactly where a drug war is taking place (Sonora, Chihuahua, Nuevo Leon and Tamaulipas). Yet, this map is also consistent with the notion that poor areas tend to default more. For example, the lower part of Mexico (Veracruz, Oaxaca and Chiapas). However, what is surprising is the selection of the darker states (Colima, Campeche and Quintana Roo) since there is not clear explanation of why this is the case.

Questions we want to answer

Below I restate the questions that we want to answer (in order of importance):

- **Why do people default?**: Generally put, this is the question that will enable us to answer all the questions below but, even though it is easy to understand, it is hard to pin down formally. Next are our main hypothesis of why people default. One hypothesis is that the customer might have run into some financial difficulties. Another hypothesis is that the customer might not like the property that she bought and thus abandon it. Yet another hypothesis is that some conditions in the area where the customer lives have made him move away and thus abandon the property. One final hypothesis is that the customer has bought this asset as an investment but it is not interested in keeping it anymore. The answer might be a mixture of all these plausible stories, yet we want to uncover which is more consistent with the data and an appropriate model formulation to address each.
- **Why are some geographical areas more prone to default?**: Since 2010 Mexico started a drug war against different cartels located throughout the country. Still today in 2018, there are some states that are still in conflict. We want to understand what is the collateral damage that this event has on the ability of people paying their mortgages in those areas. Also, as seen in the previous analysis, we have uncover that some states where there is a drug conflict should a higher probability of defaulting.
- **Are there some untruthful vendors?**: We have been told that one of the main reasons that people default is because some vendors enticed the customers into buying some house that they ended up not liking.
- **How likely is someone to default?**: This is a prediction question that is fully covered with logistic regression. However, as seen above, it is needed that we expand our logistic regression model since the off-the-shelf approach is not separating the data satisfactorily.

Proposing models to address these questions

We want to construct a model that captures the trade-offs that each person undergoes every month when deciding whether to pay its mortgage loan or not. It appears that this trade-off boils down to two latent quantities ζ_i and I_i where $i = 1, \ldots, N$ and N is the total number of mortgages in the data base. The parameter ζ_i, which is a number between $[0, 1]$, represents the percentage of the income that this individual is willing to give in order to pay her mortgage. Thus, if the customer does not like her house then we would expect ζ_i to be low but, if customer has all her dependents living inside, then ζ_i should be close to 1. I_i is latent because even though we have an observation of the client's income when he signed the contract, we cannot assure that this stays constant over time (for example, the client might lose its job or acquire other liabilities). By design, this two latent quantities merge together so that

$$\Pr[Y_i = 1|X_i] = logit^{-1}(\zeta_i I_i)$$

hence the center of the model boils down into modeling ζ_i and I_i. There are two main topics to address: (1) what functional forms to give to each latent variables (in order to connect them to the covariates) and (2) how is the information shared across the graphical model. In principle, the functional form of ζ_i could be shared across all individuals but its value would only change depending on the person's covariates. In contrast, I_i should be shared only across similar spatial units. Thus far we are thinking

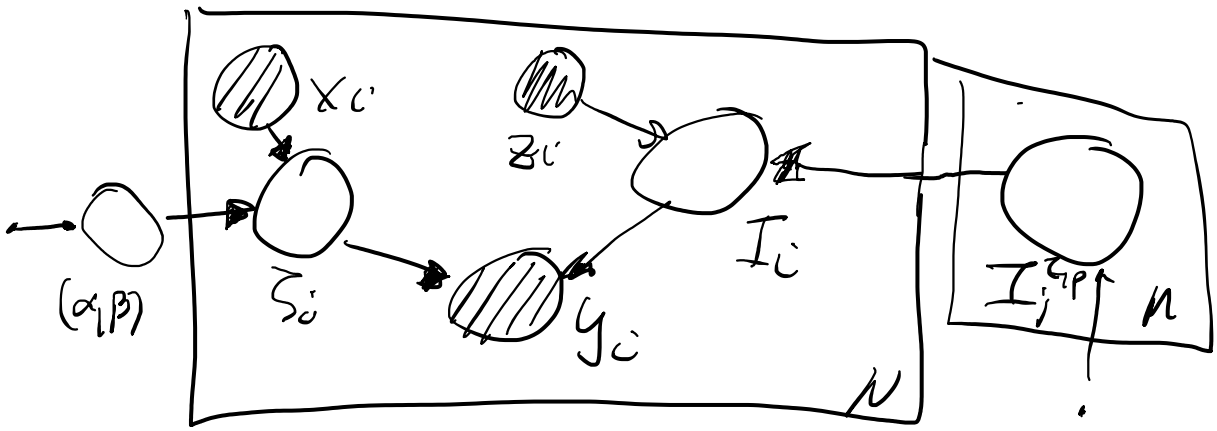
$$\zeta_i = logit^{-1}(\alpha + X\beta)$$

where X contains only covariates related to the person's likelihood to assign more money to its mortgage. For example, `risk_index`, `age` (as a proxy for family dependents), `sex` and `new_used` and `inv`. In terms of the other latent variable

$$I_i = I_j^{zip} + I_i^{job} - M_i$$

where I_j^{zip} relates to the average income level of that zip code, I_i^{job} to the salary that she ears and M_i relates to the mortgage amount she has to pay. Note that this approach will allow us to pair individuals with similar ζ_is or similar I_i and thus understand how default changes between those matched groupds.

The graph for the previous model is (z_i represents the observed assignation of the individual to its zipcode):



Laying out concerns for feedback

Our main concerns are:

- **Do we have the information necessary to address our research questions?** Does the data that I presented in the first section appears suitable to address our questions? Is there anything that we might be asking that is ill-posed and would not be able to be seen from our data? Are we able to draw any type of causal inferences with our data?
- **Are we posing a model that will allow us to derive the insights that we want?** We are afraid that maybe the idea of modelling ζ_i and I_i might be too restrictive as a family of models. Or that it might be really hard to estimate due to its functional form. Related to this, the inference is also complicated since we do not have a conjugate family and thus it will requires to use gradient reparameterizations. We are also worried that we might be complicating our approach more than it is necessary. Should we better try to estimate a simpler model like a linear logistic regression where we shared the parameters according differently (for example base β_{income} we shared across zip codes but the β_{age} we share across all individuals).
- **Would incorporating historical data help us make any causal inferences?** Our questions are centered about knowing why something happened. Furthermore, we have the same data for previous periods. Do you think that any of our research questions might benefit from introducing the historical values of the same covariates? Also, howshould we introduce this historical dependencies in our model?
- **Is there any way to construct a possible embedding?** It would be really interesting to project all the features into a lower dimensional space and see how the data is grouped together. However, it is not evident neither what we could use as interactions nor what would we be making the interactions between.