Demographic Income Estimation in Practice

Ross Gayler - r.gayler@gmail.com

1 September 2017

Need to know someone's income?

ASK THEM!

- Mix of income types is idiosyncratic
- Quantum of income types is idiosyncratic
- Unlikely to have access to idiosyncratic causal factors

Ways to ask:

- Ask: application form
- Look: applicant provides 3rd party documents
- Scrape: direct access to 3rd party information

- Definitional issues:
  - What counts as income? What counts as permanent income? What time scale?

For when you can't ask

- May not be able to ask/look
  - e.g. marketing (if legal)
- May need a second opinion
  - e.g. verification of asked details
- Estimate income based on independent (non-financial) data
- Typically a geodemographic model, based on e.g.
  - Occupation
  - Education
  - Age
  - Gender
  - Residential address

A concrete example

National Consumer Credit Protection Act (2009) - Australia

ASIC Regulatory Guide RG209

- Responsible lending
  - Prevent lending that is "unsuitable" for the customer
  - Unsuitable if (amongst other things), it is likely that the customer is unable to meet payment obligations
- Obliged to assess whether the offer is "not unsuitable"
  - (Amongst other things) make reasonable enquiries as to the customer's financial situation
  - Take reasonable steps to verify the financial situation
- Customer income is an important part of financial situation

Assessing income under NCCP

- Make reasonable inquiries:
  - Ask: application form filled by applicant
  - Look: e.g. pay slips supplied by applicant
  - Scrape: e.g. transactions from bank account
  - Evidential strength: Scrape > Look > Ask
  - Cost/convenience/coverage: Ask > Look > Scrape
- Take reasonable steps to verify:
  - What counts as verification?
    - Needs access to relevant independent data
    - Verification most needed when initial evidence weak
    - Stronger evidence preferred as verification
  - Consistency: "lender should have reasonably known"

Scalability

- RG 209: Reasonable inquiries and verification are scalable
- What is reasonable depends on: (non-exhaustive)
  - Impact on customer
  - Complexity of the credit product
  - Capacity of customer to understand
  - Prior knowledge about customer
  - **Inconsistent customer information**
- More amber flags ==> more lender effort expected

Lender pragmatics

- Customer stated income is mostly quite accurate
  - Verification generally doesn't change result
- Stronger income evidence is not readily available for all
- Customers dislike extra information requirements
  - Extra time and inconvenience
- Lenders have a range of verification methods of increasing cost and inconvenience
- They want to target those methods to applicants of increasing regulatory risk
- For the very lowest levels of regulatory risk some lenders want to ask for income and check that for plausibility

Decison making approach

- The real concern is if the lender thinks the income is higher than it actually is
- The lender wants to compare the claimed income to some threshold based on expected income for that customer
- If the claimed income is sufficiently high relative to expected income the lender decides the claimed income is sufficiently implausible to require further verification
- Needs a defensible estimate of the customer's income

Geodemographic estimates of income

- Difficult to get accurate estimates
  - Geodemographic predictors don't capture idiosyncratic effects
- Lower and upper quartiles of actual income are typically within a factor of two of the predicted income
  - 50% of actuals within a factor of two of predicted income
  - 50% of actual incomes even more distant
- A point estimate of income is very unsafe to use
  - Users tend to take the point estimate as accurate
- Much safer and more useful to estimate the distribution of income conditional on the predictors

Using an estimated distribution

- Threshold can be a quantile from the estimated distribution
  - e.g. require extra verification if claimed income above 80th percentile of estimated income distribution
- Threshold can vary as a function of regulatory risk factors
  - e.g. for larger loan amount threshold might be at the 50% percentile
  - Obviously possible to set the threshold at zero for sufficient regulatory risk so that all customers get further verification
- If distribution of estimates is well calibrated then it is obvious what proportion of customers fall over the threshold

Shape of income distributions

- Income distributions are **strongly** skewed
  - There are some people with much higher incomes than other people with the same predictor values
- Distributions are normal on a logarithmic scale
- Distributions may be asymmetric, e.g. distribution for females may be compressed above the median relative to males
- Tails generally heavier than a normal on a logarithmic scale
  - More people at the upper and lower extremes
  - Fatness of tails may vary, e.g. distribution for managers may be fatter tailed than distribution for clerical workers
- The tails are not always symmetric
  - Upper tail for sales workers may be fatter than lower tail

Details of modelling

Importance of design

- Many data decisions to be made
- Many modelling decisions to be made
- Must be evaluated in the context of use
- Actual choices may not make vast differences
- Must be defensible to a regulator
- Important to document the design process
    - Why you made the decisions you did

How to model shape of distribution

- Model log of income
  - Much better behaved and easier to plot
  - Modelling quantiles is invariant to the transform
- Use analytic distribution with shape/location parameters
  - Ideally need to be able to model the shape effects such as asymmetry and fat tails
- Use quantile regression (currently preferred method)
  - Like ordinary regression with extra parameter (τ) for the target quantile
  - Estimate a separate regression equation for each of a range of τ values e.g. {0.1, 0.2, … 0.9}
  - The set of predicted quantiles at a given set of predictors approximates the distribution and can be interpolated

Data source

- Prefer to use lender data
  - Represents data the model will see in practice
  - Represents population the model will see in practice
  - Application form may need to be modified
- Could use official statistics (microdata surveys)
  - Temporary measure if lender data not available
  - Treat income as verified (no incentive to inflate)
  - Data definitions unlikely to correspond exactly to application form or cover all the same variables
  - Selection bias - sampled to represent entire population
    - Need some way to calibrate the model to the lender's population

Income and employment types

- Ideally, use verified income
- Income types:
  salary, wages, pensions, benefits, investments, business profit
  - Salary/wages more suited for geodemographic models
  - Other income types likely have higher regulatory risk
  - Different models/approaches for different income types
- Employment types: full-time, part-time, casual, self-employed
  - Likely to restrict model to full-time and part-time workers
  - May be better to model hourly rate and multiply by hours
  - Different models/approaches for employment types

Income definitions

- How to treat wage/salary components: "standard" income, overtime, commission, bonus, salary sacrificing, in kind
  - Official surveys tend to have detailed definitions and record the separate components
  - Lender application forms try to minimise the complexity
- Time scale of recording/estimation
  - Official surveys tend to be for short periods (1-4wks)
    - Figures are more volatile (can be negative)
  - Lender application forms try to get values
    - Ask applicant for typical/average figure
    - Tend to allow time period to be whatever is most convenient for the applicant (year, month, etc.)
- Application question design is hard

Predictors

- Typical predictors are:
  - Occupation
  - Age
  - Functions of residential address, e.g.
    - Metropolitan / urban / rural
    - Small-area income estimates
    - Small area wealth estimates
    - Small area dwelling value estimates
    - Small area rental estimates
  - Education/qualifications
  - Gender
- Specific combinations of values can be rare

Occupation

- Typically the strongest predictor
- Resolution: need lots of possible occupations
  - "Manager": manager of local shop and CEO of multinational
  - "Professional": neurosurgeon and suburban accountant
  - Modelling with rare categories is hard
- Reliability: Will the occupation always map to the same label?
  - Free text is hard to map to categories
- Usability: How hard is it for the customer to fill in?
- Hard to trade off resolution / reliability / usability
- Application question design is hard

Age

- Expected income typically starts low, increases steadily until late 30s, remains relatively constant to mid 50s, decreases slowly as age increases
- Trends are expected to be relatively smooth, so better modelled as a smoothed continuous predictor
  - Especially useful for smaller groups where the trend is over a larger neighbourhood
  - Extreme age groups are likely to have different regulatory treatment but worth including those groups in model building to avoid edge effects in smoothed terms
- Worth looking for interactions with occupation
  - The relationship of income to age may differ between occupations

Residential address

- Need to map address to values of predictors that are related to income, e.g.:
  - Metropolitan / urban / rural
  - Small-area income estimates
  - Small area wealth estimates
  - Small area dwelling value estimates
  - Small area rental estimates
- Unlike occupation, this is relatively easy to get high resolution while being automated
- Worth looking for interactions with residential status
  - Relationship to income may differ between owner/renter/living with parents

Qualifications

- Higher level of education/qualification tends to be associated with higher income
  - Some of the effect mediated via occupation
- Relatively easy to capture reliably to the resolution that's reasonable
- Worth looking for interactions with occupation
  - Relationship to income may differ between occupations
  - Can be difficult to model interactions because range of qualifications can be limited within occupations

Gender

- Gender effects may exist although typically smaller after controlling for occupation and hours worked
- May see compression of the range of upper quantiles of income for women relative to men
- Modelling in interactions with occupation may be difficult because some occupations are very gender-biased

Discrimination

- Use of some predictors (e.g. age, gender) may raise questions of illegal discrimination.
- Discrimination legislation is intended to avoid adverse outcomes
  - In this example the model is used to target verification actions, not make lending decisions
    - Is requiring a stronger verification action an adverse outcome or benefiting the applicant?
    - Financial regulators may give examples of considering "discriminatory" attributes in guidance (e.g. an 18 year old applicant claiming $100k income should have stronger verification)
- Ultimately a question for the lender to take legal advice and make a commercial decision

Modelling

- Quantile regression is just like ordinary regression with the extra quantile parameter (τ)
  - Standard modelling practice carries over
- A case can be made for interaction terms
  - But low sample sizes rapidly become a problem
- Out of sample validation and calibration always important
  - Also needed for regulatory inquiries
  - Needed to ensure appropriate threshold setting
- Quantile estimates are independent
  - Consistent ordering guaranteed at predictor centroid
  - Far from the centroid the estimates can cross
    - Can be detected and flagged in the data sample

Future directions (given a free hand)

- Multi-level models
  - Given interactions and multiple levels of grouping of occupations and residential addresses some levels of predictor will have very few observations, in those cases we want to partially fall back to more reliance on the simpler predictors
- Bayesian models
  - Probably needed for a principled approach to multi-level
  - Should provide a better handle on uncertainty of predictions
- Modelling with parametric distributions
  - Guarantees consistent ordering of quantile estimates
  - Wouldn't do this without understanding consistency
  - Probably needs to be Bayesian