[redacted]

Office of Research

07/23/2014

Statistical Analysis of Underwriting Outcomes in [redacted] Exam

cfpb Consumer Financial Protection Bureau

Overview

- Some of the difficulties with underwriting analyses
- My general approach
  - Plain language
  - Not-so-plain language (but still not too technical)
- Different measures of interest, and their pros and cons
- Examples from an actual exam

Some of the problems with underwriting analyses

- We are implicitly concerned with counterfactuals that we can never observe
  - What would have happened if applicants of type A had been type B?
  - What about if applicants of type B had been type A?

- We observe only a binary outcome (denied/approved), but must estimate the probability of denial/approval (think: proxy issues)

- Even with estimates of the disparity, evaluating harm is difficult
  - Can (theoretically) be disparity without harm (e.g. no "marginal" candidates)
  - Can even be harm without disparity (e.g. marginal candidates treated differently, but on average, treated the same)

My (limited) understanding of the "old way"

- Use logit to estimate odds ratios
  - Odds ratios make identifying the *existence* of a disparity very easy
  - But the *magnitude* of disparity is very hard to interpret with odds ratios (and easy to misinterpret); e.g. what does odds ratio of 3 mean?
    - ~~Denial rate for test group is 3 times higher than for control group?~~
    - 25% denial rate for test vs. 10% denial rate for control?
    - 75% denial rate for test vs. 50% denial rate for control?
    - 99% denial rate for test vs. 97% denial rate for control?

- Also, difficult to show how many people harmed, by how much, etc.
  - (sadly, my approach still struggles with this one)

What I did in this case (plain language)

- Ran a slightly different statistical model, then estimated two counterfactuals:
  - Conditional on all the other characteristics, what would the denial rate have been if all the applicants were treated like the "control" group?
  - With the same conditions, what would the denial rates have been if all the applicants were treated like the "test" group?

- For the disparity estimate, took the average of the differences between these estimates <u>across all applicants</u> (avg. treatment effect, or ATE)

- For the harm estimate, took the average we estimate for <u>members of the test group only</u> (avg. treatment effect on the treated, or ATT)

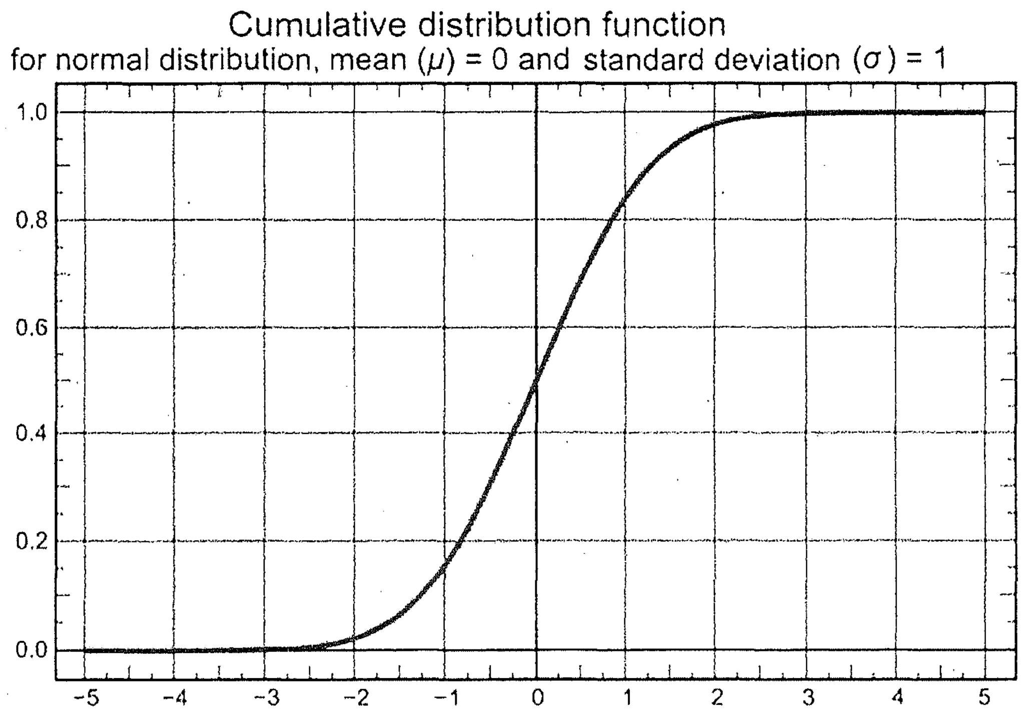
What I did in this case (less plain language)

- Applied a probit regression
  - Similar to linear regression, but no longer fitting a line
  - Formally, a probit regression is based on the normal ("Gaussian") distribution
  - Think of it as a "best fit curve" rather than a best fit line
- Calculated "Average Marginal Effects" from the resulting coefficients to approximate the ATE and ATT
- Employed a number of robustness checks and bias corrections to ensure estimates are conservative
  - Note that I did not present these as the "findings," as they are likely too opaque to be of practical relevance

What does this actually mean? (depicting a probit)

- So let's say we run a probit regression on race and price and we get some model $Decline = \Phi(\beta_0 + \beta_1 * FICO + \beta_2 * Black)$ ..What does this mean?
- Perhaps best described visually:

Pr(denied)



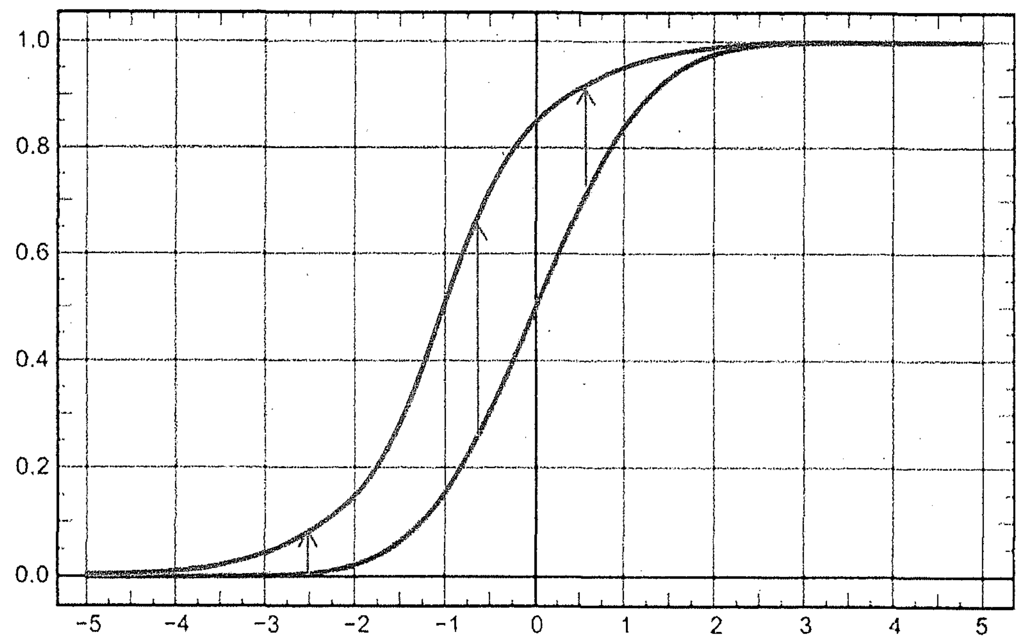
$\Phi(\beta_0 + \beta_1 * FICO\beta_0 + \beta_2 * Black = 0)$

Note: this is only intended to be a conceptual representation

What does this actually mean? (depicting dummy coefficients and marginal effects)

- Being Black here shifts the curve to the left, meaning these applicants are more likely to be denied at any FICO level
- But, *how much* more likely depends on the FICO (which moves us along the curve)
  - Note: this gets more complicated when additional variables are added

Pr(denied)



$\Phi(\beta_0 + \beta_1 * FICO\beta_0 + \beta_2 * Black)$

Black denial prob. (Black=1)

White denial prob. (Black=0)

Example Marginal Effects

Note: this is only intended to be a conceptual represe tion

So, what can we make of this?

- While the coefficients from the probit may be difficult to interpret, we can use them to calculate some useful information, e.g.:
  - Would the (expected) denial rate have been if everyone was Black?
  - Would the (expected) denial rate have been if everyone was White?

- Taking the difference between these gives the marginal effect, from which we can calculate some measures of interest
  - The “Average Treatment Effect” (ATE), a measure of the disparity in the full sample (taking average marginal effect from the whole sample)
  - The “Average Treatment Effect on the Treated” (ATT), a measure of the harm faced by the observed protected applicants (taking the average marginal effect of the protected group)

Two measure of disparity: ATE vs. ATT

- ATE is the average change to the expected probability of denial given the respondents other characteristics
  - Calculated for the entire distribution
  - Uses all the same information used in calculating the coefficient (ensures "unbiasedness")
  - Looks at both sides of disparity (e.g. positive benefit for control group, negative penalty for test group)

- ATT is the average change to the expected probability of denial for "treated" applicants, given these applicants' characteristics
  - Calculated for only the "test" group, not the entire distribution
  - Only captures negative penalty facing test group
  - Not always a reliable measure when calculated this way

Actual estimates from [redacted], with an example interpretation

	Black, 30-yr. conventional	Black, 30-yr. MHP	Hispanic, 30-yr. conventional
Unconditional "control" denial rate	12.1%	13.7%	12.1%
Unconditional "test" denial rate	25.6%	26.1%	26.1%
Unconditional disparity	13.5%	12.4%	11.1%
ATE (conditional disparity)	4.1%	7.1%	5.3%
ATT (conditional disparity, test only)	5.8%	10.5%	7.4%

- On average 30-year conventional applicants would have been denied in an additional 4.1% of the cases if they had been transformed from White to Black
- On average Black 30-year conventional applicants faced denial in 5.8% more cases than they would have had they been White

Some caveats

- An argument can be made that the ATE is as good (perhaps better) measure of harm than the ATT

- In some cases, calculating the ATT in the manner described leads to biased estimates
  - E.g. when characteristics of the groups don't sufficiently overlap
  - Solutions to this problem (e.g. PS weighting/matching) add complexity, and demands more of the data

- Since estimates are of a probability of outcome, very difficult to identify who (if anyone) is actually impacted, let alone the resulting harm

Benefits to this approach

- These marginal effects/ATE/ATT estimates are comparable across products, lenders, etc.
  - An estimated ATT of 5.8% means we expect the test group was denied in 5.8% more cases, regardless of the “odds” of denial
  - *__If__* we determine a ratio with which we are comfortable, we compare relative denial rate disparities across lenders*

- The estimated “harm” is a little more clear here, so long as we are content to deal in averages
  - 5.8% more denials means harm can be estimated at .058 x (number of test applicants) x (amount of harm per applicant)
  - Note that we are estimating this in probability space, so we still don’t know *who* (if anyone) was actually harmed (again, think: proxy)

Some potential ratios of interest

Measure	Interpretation	Example: Black, 30-yr. conv.
ATE/unconditional denial rate for all applicants	Increase in denial rate average applicant would have experienced if moved from control group to test group as % of <u>observed denial rate</u>	30.6%
ATE/conditional denial rate for all	Increase in denial rate average applicant would have experienced if moved from control group to test group as % of <u>estimated denial rate</u>	29.8%
ATT/unconditional denial rate for control	% increase in denial rates for test group relative to <u>control group denial rate</u>	47.9%
ATT/unconditional denial rate for test	% of <u>test group denial rate</u> that remains "unexplained" after controlling for underwriting factors	22.7%
ATT/counterfactual denial rate for test if treated like control	% increase in denial rate for test group over the <u>denial rate they would have expected if they were in control group</u>	26.8%
ATC/unconditional denial rate for control	% increase in denial rate for <u>control group if they were moved to test group</u>	32.2%

Each of these measures has merit, but none is a clear "best measure" of relative disparity

Bonus Slides

Disparity (ATE) vs. Damages (ATT): Why are they different here?

- Disparity here (ATE) is the difference in (conditional) probability of denial between the test and control groups
  - Inference is about the *process* leading to denial
- Damages here (ATT) is the difference between the expected and realized outcomes for harmed group
  - Inference is about the *outcomes* that actually resulted from that process

- Imagine underwriting depends (only) on rolling a die: Whites are denied if they roll a 3 or less Blacks and Hispanics are denied if they roll a 4 or less
  - Disparity is 16.7%, i.e. the *chance* that someone rolls a 4 (and getting a different outcome)
  - Damages only accrue when test group *actually rolls* a 4 (resulting in an outcome different than the control group would have achieved)

Die roll underwriting example, continued

- Take the case below showing rolls and outcomes for 6 applicants from 3 different groups
  - Conditional disparity (~ATE) is still a 16.7 percentage point increase in denials for minorities vs. Whites
  - But the harm (***in bold***, ~ATT) would be higher than disparity for Blacks (33.3 percentage points), and lower than disparity for Hispanics (0%), since they happened to roll more 4s

Whites	Blacks	Hispanics
5 (approved)	6 (approved)	6 (approved)
5 (approved)	5 (approved)	6 (approved)
4 (approved)	***4 (denied)***	3 (denied)
3 (denied)	***4 (denied)***	3 (denied)
3 (denied)	4 (denied)	2 (denied)
1 (denied)	2 (denied)	1 (denied)

Note: this is only intended to be a conceptual representation

Disparity (ATE) vs. Damages (ATT): Some takeaways

- Disparity (ATE) here is an estimate of the *expected* difference in conditional outcomes
  - Good measure of the risk of harm
  - Might be a better measure of "unobserved" harm, e.g.
    - If we think protected group members are harmed by disparate processes in addition to disparate outcomes
    - If we think institution got "lucky" to have only non-marginal applicants
- Damages (ATT) is an estimate of the *actual* conditional differences realized by the test group
  - Good measure of the impact of the disparity
  - Might not capture the "true" problem
    - Could miss the risk of additional (unrealized) harm
    - Could penalize institution for taking on more "marginal" applicants
- "All potential applicants face the same disparity, but the damages vary with the actual applicants"*

*Note: this isn't technically correct, but it isn't too far off, and might be an easy way to conceptualize the difference.