Further Limitations:

Comparing the sample to the national average obscures the impact of the results as the age, race, and ethnicity of the sample are not representative of the national population. Further, sampling only from community health centers in NYC means that our population for generalization must be limited to community health center patients in NYC. This is an area of important for MOIA? in particular, but should not be extrapolated. The data exists to draw more useful conclusions, which are not accompanied with this study.

Recommendation:

A more well-designed questionnaire could aid in greater insights from the data, particularly regarding discrete variables. For instance, attempting to regress binned data, such as age group 55-64, leaves a enough room for error that either a much larger sample would be needed, or it would be practically useless. The same goes for the number of visits respondents have made in the past year. Ideally, answers that are numeric should be obtained as a numeric answer. Clarification of some answers should be provided by offering alternative answers, such as ‘other’, as in the case of the chronic conditions question. If an option of ‘other unlisted condition’ existed, we could be confident that an answer of ‘None’ certainly means that the respondent has no chronic health conditions, rather than having none of the conditions listed.

Answers of preference are of value in identifying problems experienced by the sample, however we cannot assume that these conditions and concerns can be applied to a wider community writ large without being able to identify the larger population more precisely.

Note: answers of 65+ for age could be distributed as ‘everyone is 65’, ‘everyone is over 80’, or something in between. This amount of ambiguity makes generalizing nearly impossible. Age is obviously an important factor. I could throw out the under 65 respondents and maybe do a simple cohort analysis.

Dimensional issues:

Initially we have 124 participants, answering 20 questions, which sounds great. You must have at least two **observations** for **each variable** to capture ***crude*** variation, so we’re good, right? Sorry. Nearly all of these questions are categorical (allow a particular choice, rather than a numerical-only input). We end up with roughly 100 variables across 124 participants, which means we can’t use it in a direct way. Going back to throwing out the age answers, we can eliminate 2 variables and say ‘this pertains to a group of 65+ year olds’ and still have enough participants to be statistically significant(-2).

There’s a way to take the education data apart and turn those 8 variables into 1 (-7) by using years education as a proxy. Visits to the doctor could be (kinda-sorta) broken down from its groups to numbers, cutting 5 of the 6 variables (-5). Happiness with care could go from 5 categories to 1 (my fault categorizing that) (-4). Length of time with provider could be approximated numerically (-4). This brings the total variables closer to what we need, but I’m sure you see the difficulty. This was a design failure if making significant inferences was the goal.

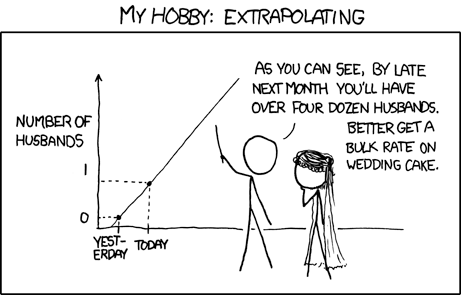
The dimensionality issues are why I did so much cross-category comparison, but can’t do a lot in the way of estimating covariance with a brute-force computation of each pair. Say 85 variables after cleanup:

~~85! / (83! \* 2) = 3570 hand-made (or at least personally inspected) comparisons.~~

Cleaned up to just the questions:

20! / 18!\*2 = 190

That’s 190 considerations for just 124 people. Not only is 190 analyses seriously convoluted, but for such a small sample (relatively), it’s basically useless.

Due to the high dimensions of the data, we have very little opportunity to interpolate, and rather must extrapolate, which is very faulty and can lead to horribly misleading interpretations.

Notes on the statistics:

Measure of Association:

Odds Ratio (factor of effect) *pA*/(1-*pA)*

*pB*/(1-*pB*) = *pA*(1-*pB*)/*pB*(1-*pA*)

Group A v Group B of **sample**

*p* = P(Y=1 | *XA* , X*B* , … , X*K*) for K variables

logit (*p*) = ln (*p*/(1-*p*)) = B0 + BAX*A +* BBXB + … + BKXK

no epsilon for error as probabilistic estimate

B’s come from Maximum Likelihood Estimation – can automate from statsmodels

odds ratio = e^(logit(*pA*)-logit(*pB*))

where *pA* = *pB* +B*variable\_of\_interest*

Dimensionality: