Notes 19: Multicollinearity and other data failures

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

In the last class, we considered the ways in which assumption 2 could fail.

1. Random sampling
2. No perfect multicollinearity
3. We observe all $\{(y_i, x_{1i}, \ldots, x_{ki}) : i = 1, \ldots, n\}$.

We discussed a failure of the third condition: measurement error, which consisted in not observing the variables of the model, but rather a noisy measurement of it. Today we examine other common data failures.

1 Sample selection

Sample selection is the most important of all the data failures. In fact, it is nearly as important as endogeneity. It is also quite common, and it is very subtle and hard to catch.

- **Sample selection:** _____

The idea is that the sample should be random **from the population**. Sometimes the data is not random at all, because the observations are hand-picked. However, this is rare. Good surveys with many observations strive to collect a random sample.

The problem is that very often the sample is not drawn from the population, but rather from a subset. For example, suppose you want to talk about the effect of race in wages, because you want to investigate if there is racism in the labor market. Well, if you only survey employed people, you are not collecting a random sample from the population of people in the labor force. The argument is subtle, but you need to understand it. Suppose that there are 70% whites, and 30% minorities in the labor force, but 80% of the employed workers are white, and 20% are minority. If the employed workers were representative of

the labor force, everything would be fine. However, this is not the case, after all there are proportionally more unemployed minorities than whites. Now, suppose that you compared the employed workers and saw that for the same qualifications, whites and minorities earned the same. Can you conclude that there is no racism in the labor market? No. It could be that although salaries are the same, whites with the same qualifications receive more offers than minorities. However, you could never find out about this if your sample only has employed workers.

When you decide to use a data set for your question, you must investigate it for possible sample selection. Sometimes the selection is explicit, as in the example above. It is easy to see obvious reasons of selection by reading the codebooks and finding out how the data was collected. Typically the population interviewed may be selected for accessibility. For example, the survey could have been conducted at a university, or at a company, or at a given city or state. Sometimes the survey is explicitly defined in a population: pregnant women, men 40-50 years of age, retired people.

There are more subtle forms of selection which are equally important. The most important source of such selection is the manner the survey was conducted. Was it an online survey? The sample will be selected for people that have access to the internet. Was it a phone survey conducted in business hours? The sample will be selected for stay-at-home parents, work-at-home people, retirees, and unemployed. Make sure to think about the reasons why someone would answer a survey. For example, surveys about consumer, worker, and student satisfaction tend to be answered by those with extreme opinions, people that want to either rave or complain about something in the hope of influencing it. This is not a random sample of the population of consumers, workers or students, it is a sample skewed towards the extremes.

What should you do if your sample is selected? When you cannot find a better data set, then perhaps it is best to recognize that you can't answer your original question, but you may be able to answer a more restrictive and still interesting question. In the example of racism in the labor force, perhaps you should restrict your study to be about racism in salaries of employed workers. If your sample is of students, perhaps you should answer a question about students.

The problem is harder when the source of selection is not obvious. In such cases, you must go back to the roots and think what is the omitted variable. As with all the other problems, this too can be modeled in the omitted variable language. The reason for the selection is the omitted variable. Is it a confounder? For example, suppose that you would like to know the effect of the number of exams in the students' satisfaction. Your sample is comprised of course evaluations. We know that the sample is selected due to the extremity of opinions, those that care to have a say about the course. The extremity is obviously associated with the number of exams and with the evaluations. Is it accounted by the

covariates? Well, it depends on the covariates. The fact is that it is very likely that your results will be biased.

If you suspect selection, but you can't modify your question, nor do much to solve it, you have to be upfront about it so the reader is warned. You should make an effort to explain what is the expected type of issue that may arise, and the likely direction of the bias in your case.

2 Missing data

Sometimes data sets have missing information. For example, some respondents may fail to report how much they smoked altogether. No lying, no mistake, just don't report.

- Which of the conditions above may fail because of this problem? _____

_____

The fact is that it is not ethically possible to force somebody to give away information (for research, I mean. The IRS, for example, can most definitely force you to give information). One can provide incentives, as long as no coercion is implied, but ultimately we may end up with a lot of observations about whom we know the answer to most questions, but not all.

We must examine the reasons why we have missing data in order to see if this is a serious or an innocuous problem. Here are some common possibilities:

- _____

  this is the silliest of reasons, but it does happen sometimes. Some data sets fall out of favor with the institution or person that produces it, and may not be taken care. Old data sets in paper form which were manually transformed into computerized data sometimes have big mistakes and omissions.

- _____

  This happens when a survey participant actively decided not to answer a question. This can happen because of several reasons, including shame, guilt, embarrassment, difficulty of the question, laziness of the person, identity protection, etc.

- _____

  This happens in two ways. The first is when the subject abandons the survey while filling it out. If the survey is too long, too difficult, too boring, or too prodding, a

person may decide to stop answering after a certain point. The second way is when a survey follows the same people over time (this is called a longitudinal survey), some respondents may participate for some periods, but then disappear from the remainder. This could be due to them not willing to participate anymore, moving away, or even dying.

- _____

  This is not exactly missing data. Sometimes part of the information is censored for identity protection. For example, when a survey has income, it is very common for it to be capped at $500,000, and sometimes even less. It means that if an observation makes 2 million dollars per year, it will show as $500,000. Since in most places there are very few people that make more than this, it may be possible to figure out who is the respondent by looking at the income and a few other answers (number of children, marital status, age, occupation, etc.) So, to protect them, the income is capped. The censoring doesn't let you know the person's income, but it does let you know that it is $500,000 or more, which is more than in the previous cases.

Missing data is not a big problem when the information is missing at random.

- **Missing-at-random:** _____

  _____

When this is the case, simply remove the observations which lack the information you need. This is not a big deal, unless you have to remove a large part of your sample, and end up with few observations. Poor data entry and sometimes survey abandonment generate missing-at-random information. Though survey abandonment is never due to truly random reasons, depending on the scientific question you are trying to answer, they may not matter. For example, if a large proportion of the sample abandoned the survey at some point because it was too long or boring, if persistence under boredom in a low stakes activity (such as the survey itself) is not a confounder, you can treat the missing information as if it was missing-at-random.

The problem is that very often the missing information is not random at all. For example, if a data set with smoking mothers is missing the smoking variable for a large proportion of the mothers, it is a good idea to investigate why this is so. It could be innocent, for example if the smoking question was not in the questionnaire given to these mothers. However, it could be due to selective answering. If the mothers that don't answer are exactly those that feel guilty, then if you simply eliminate those mothers from the sample, your remaining sample won't be random at all! It will be a very selected part

of the population, formed by those women that don't smoke, and those that smoke but don't feel guilty.

So, from the example above, you can see the problem when data is not missing-at-random. If you eliminate the observations with missing information, your remaining sample may be selected! From the first part of the class, we know that sample selection causes bias in the estimates. Hence, eliminating those observations is not a good solution at all.

What should you do? It depends on the reason for the missing data. In the smoking women example, the reason is guilt, shame, embarrassment. Suppose that in the original data set, $\mathbb{E}[u|x_1, \ldots, x_k] = 0$. It means that, conditional on observables, guilt, shame and embarrassing are constant. So, it is as if those characteristics are fairly randomly distributed in the population (conditional on observables), and therefore are not confounders. Mind you, they could determine birth weight, but they are not related to smoking conditional on $x_2, \ldots, x_k$. However, if you eliminate those observations that did not answer the smoking question, you may no longer be able to say that $\mathbb{E}[u|x_1, \ldots, x_k] = 0$. For example, conditional on having 13 years of education, being married, and having an income of \$80,000, what is the expected guilt? It doesn't matter. It is probably slightly lower than what it used to be, because a few smokers were eliminated from the sample due to omission. Now, take the group with 8 years of education, not married, and income of \$30,000. What is the expected guilt of this group compared to the last? This group has proportionally much more smokers, and therefore more people eliminated for guilt-caused omission. Therefore, the remaining group has proportionally less guilt than the previous. So, guilt is varying depending on the covariates, which is the definition of endogeneity!

You see that, if the data is not missing-at-random, you could cause a problem of sample selection if you simply eliminate the observations with missing information. How can you know if the information is missing-at-random? Perhaps the easiest way is to find out is to compare the information you do have for both groups. So, take the women that answered the smoking question, and get their expected income, education, etc. Now, get the women that did not answer the question, and get their expected income, education etc. If they are the same, it is a good indication that the missing information is fairly random. This is not a perfect test, but it works very well in practice, and it is very quick to implement.

If the data is not missing-at-random, then you must understand the reason it is missing. If you can proxy for it, then you can go ahead and eliminate the observations. However, usually you cannot proxy for the reasons. In such cases, if the missing information refers to the dependent variable or the variable of interest, you must either concede defeat, or depending on the case proceed and be upfront about the issue. If there is missing information on one of the controls, you can keep those observations, and add an extra control, a dummy for missing. For example, if there is missing information for x_2, then the dummy is $x_{k+1\,i}$ =1 if x_{2i} is missing, and $x_{k+1\,i}$ =0 if we have x_{2i}. This is not a perfect

approach, but it creates a crude proxy for the reason why the information is missing, whatever it is, and does help with the endogeneity. If x_2 is missing for a great proportion of the observations, it is also a good idea to add a few more controls as well, by interacting the dummy $x_{k+1\,i}$ with some of the most important covariates.

If either your dependent variable or your variable of interest is censored, you should know that there is quite a bit of research on this topic. It can be worthwhile to attack the problem head-on, instead of taking the measures above. Look up "censored regression models" and "Tobit models."

3 Near-multicollinearity

Multicollinearity: Multicollinearity is when the variables are linearly related in the sample. If you can write

for all $i = 1, \ldots, n$.

Pure multicollinearity is not a problem, it is a mistake. It means that you have variables in your model which are so redundant that they are completely predicted by other covariates. You should simply eliminate them, after all they are not confounders.

The real problem is near-multicollinearity, which we discussed before. It happens when a variable is almost a linear function of the others. Why would this happen? It turns out this is quite common when using dummies. If one of the categories is rare, then the dummy for that category will be zero almost all the time. So, suppose the rare category is x_1, then

---

This is not a true identity, because sometimes $x_{1i} = 1$, which ruins the equality, but it is almost always true.

- What is the problem caused by near-multicollinearity? _____

Near-multicollinearity is a very serious problem. Usually we are concerned with bias issues, and near-multicollinearity is a variance problem. However, in practice it is a common problem which cannot be overlooked.

Observe that near-multicollinearity is a problem related to the sample size. Even if a category is rare, if we had an awful lot of observations there would be enough information to form reliable conclusions. For example, American Indians are less than 1% of the population. This means that if my data set has 1,000 observation, only about 10 of those would be American Indian. What can I possibly say about them with any certainty? 10 observations is not enough information to say anything. However, if I had 100,000 observations, then about 1,000 of those would be American Indian. Even though they would be few in comparison to other ethnicities, I have enough information to reach conclusions about them.

What should you do if you have a rare category? The first thing to consider is whether this particular variable is very important to your question. Suppose it is. For example, suppose you want to study racism. You would like to have a good separation of all the races. In this case, the first measure you should take is to look for another data set, especially one that was created for studying racial issues. You may find a data set with stratified sampling for race:

- **Stratified sampling:** _____

_____

Data sets produced to study race anticipate the problem of studying rare races, and will thus oversample those. This is not a problem, it is a great practice, as long as it is done carefully. It is important that each of the categories be sampled randomly, in the sense that within the category there is no selection of the sampled individuals. The data set must also provide the weights, which is a way to convert the regression using the oversampled categories back to the population proportions.

If you cannot find such a data set, then you must take other measures. Can you combine categories? For example, instead of talking about all races, you may decide to combine American, Native Hawaiians and other Pacific Islanders, Alaska Natives, and other races. This should make up about 7% of the population. Perhaps you are interested in racism against blacks, in which case you would be willing to combine Asians and those with multiple-races as well, so that you would have 3 categories: whites, blacks, and other. The "other" race category has 11% of the population, so it is not so rare anymore.

You must realize that when you combine categories, your question often has to change. Your model was originally about races, but now you can only talk about blacks and whites in relation to other races. You cannot talk, for example, about blacks and whites in relation to American Indians, or to Asians. Sometimes this is not too big of an issue, sometimes it misses important relations.

You can also decide to eliminate those observations of the rare category. So, for example, you could eliminate all the American Indian from the sample, and simply not study them.

If the rare category is very special, very different from the rest, this is actually preferable to combining it to a group that shares nothing with it. Combining categories arbitrarily can ruin the linearity of the model, and eliminating some observations may be preferable. However, do explain why you did this in your paper. Researchers don't like it when you throw away data, even though this is sometimes the best solution.