1.

Thank you for coming to my presentation. I currently work as a data scientist in Laureate Education which is a company doing for profit education. The clients/customers of the company are the students.

But we don’t deal with students directly. I’m in the business intelligence department. Our clients are various business units of the company.

We do simple things such as creating reports and visualization, and also more complicated things as data analysis, building forecasting models and optimization models.

Today I’ll show some applications of statistics in the education business.

2.

Our clients often want to try new approaches to improve business. For example, new software, new ways of contacting students etc. And they want to know whether the new approaches are better. In this regard, it’s a little like clinical trials. But there are differences. First it’s not as crucial as testing new medicine. Although money and resource are involved, but it’s not about life and death. Many times it’s hard to do randomized controlled trials because the business partners don’t want to do so. So a lot of times we are dealing with observational data.

3.

Today I’m going to focus on two types problems. The first one is the thing that I just mentioned, that how to evaluate new approaches.

The second type of problem usually occurs in this scenario. The clients know that a new method is good.

For example, their agents calling students get good feedback. But the problem is that they don’t have the man power to call every relevant student. So they want us to score or rank the students so that they can prioritize using the scores. This typically involves building a statistical learning model to forecast certain probabilities.

4.

Our typical work flow is as this:

First we meet with business contact who has domain knowledge. We try to understand the request, formulate the problem and learn data source. Then we collect data and do some cleaning. Usually we use SAS and SQL for this task. The next step is data analysis, typically we use R. Finally if a statistical learning model is needed, we usually use R or Python to build such a model.

5.

Now I’m going go over the first case study. The problem came to us last year. A business unit wanted to improve the second term retention of new students. Here the second term retention indicates whether a new student is enrolled in his second term after completing the first term. Simply put, they wanted to keep more students enrolled. They came up with a new approach and applied it to the new students started in June of 2017. And the result is quite positive compared with 2016. Hence their request to us is whether it’s really effective since it costs resource to put the approach into production level.

6.

This and the next few slides show the data. Firstly even though the problem was real, the data are not, because I’m not allowed to use real student data outside of company. The treatment group consists of 455 new students who enrolled in 2017 and the control group are 594 students from a year before, also started in June. Here tc stands for treatment or control, and ret stands for whether the student was retained in the 2nd term or not.

7.

Here are more variables. For gender, 1 stands for male and 0 for female. And prog is program, and here we only consider two programs. Balance indicates whether there is a balance in their account.

8.

Here honesty actually is a button on the website when they signed in for a module. In some sense, it indicates how serious they were. And payment plan, generally there are two types, 1 installment and 3 installment, hence is a binary variable here. Assignment indicates whether they have completed the module assignment when the approach is applied. And forum is the number of forum posts that they have posted up to that point.

9.

Finally there are the regions they are from.

10.

Of course the most difficult part is to find all confounders. After discussion with business contact and considering the available data, we will only consider the previous listed variables, in other words we will ignore other potential unobserved confounders.

What business people usually do is just to compare the proportions of retention of the two groups. But because this was an observational study, we had to worry about whether the covariates are balanced across the two groups. Usually we check the balance by several ways. Naturally first we check some plots for a feel.

11.

This a simple bar chart to compare the proportions of programs across the two groups. The pink bar stand for the education program EDD while the green ones are for Management. They are clearly unbalanced between treatment and control. If there are very few covariates, we can certainly condition on program first. For our case, we proceed without literally conditioning on program.

12.

Of course a statistical test can be more helpful for more rigorously checking the balance. Here we can use the z-test (with continuity correction) to check.

The p-value is very small, so it seems reasonable to conclude the data are not balanced with respect to the covariate prog.

13 -- 16.

Similarly we can check the balance of other covariates. Here is payment plan which is also kind of unbalanced. But the proportion of students who have clicked the honesty declaration button seem to well balanced.

17.

Another popular way that we often use to check balance is to use the standardized mean difference. Which by definition is just the difference in means (or proportions for binary variables) divided by pooled standard deviation.

The rule of thumb is that if the SMD is < 0.1 then we have good balance for the corresponding covariate.

If it’s between 0.1 and 0.2, then it’s OK, not too bad. Greater than 0.2 indicates serious problem.

18.

In R, it’s quite easy to calculate SMDs using the tableone package. This table lists the result for the original dataset. The second column gives the mean of each covariate on the control group. The numbers inside the parentheses are the corresponding standard deviation. Here we can see that program is highly unbalanced, payment plan is kind of unbalance, while the honesty indicator is well balanced.

Since there exists unbalance, we need to do some matching. We could try directly matching on all covariates, but sometimes it’s hard to adjust so many variables. Typically we use the so called propensity score matching.

19.

The propensity score is the probability of receiving treatment, rather than control, given the covariates.

It is balancing score, in other words, if we match on the propensity score, we should achieve balance of all the covariates. Of course, in reality for observational studies It’ hard to get the exact propensity score, so usually we just estimate it. Typically a logistic regression model is used, like the R script here shows: glm with binomial family gives us logistic regression, and we are regressing treatment/control to the rest of the covariates. The model is trained on all the data and then applied to the same data to get the estimation of propensity scores. Although logistic regression model is most frequently used, some other more advanced models can also be used.