Data Science in Educational Testing

Yunxiao Chen, PhD

Department of Statistics, LSE

Overview

- This year, GCSE and A level exams are cancelled due to covid and predicted grades are used.

- Issues have been found with the predicted grades.

- In this talk, we introduce statistical methods for grade prediction and for ensuring fairness (but not specifically comment on the GCSE and A level prediction). We will learn from the educational testing system in the United States.

Compare Examinees based on Different Test Forms

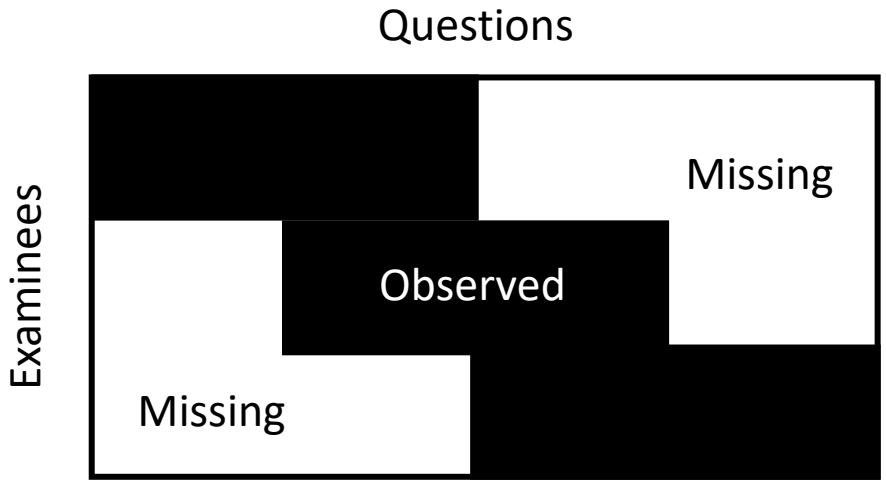
- Scholastic Assessment Test (SAT) and American College Testing (ACT) are two major tests for college admission in the US. Unlike A level test, SAT and ACT are offered 7 times a year.

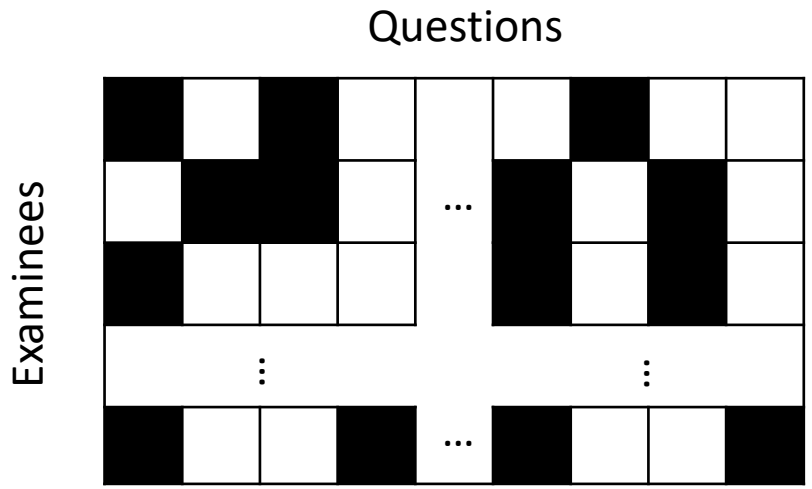
- Graduate Record Examinations (GRE) test uses an adaptive design, meaning that test questions received by different examinees are different.

- The question is: Consider students who received different test forms (which may not be equally difficult). How do we compare them?

Compare Examinees based on Different Test Forms

- The test data may look like this, or this:





Compare Examinees based on Different Test Forms

- Say student A took form 1 and student B took form 2.

- Their total scores (i.e., marks) on their own forms may not be comparable, as for example, form 1 may be more difficult (due to randomness in test assembly or adaptive design).

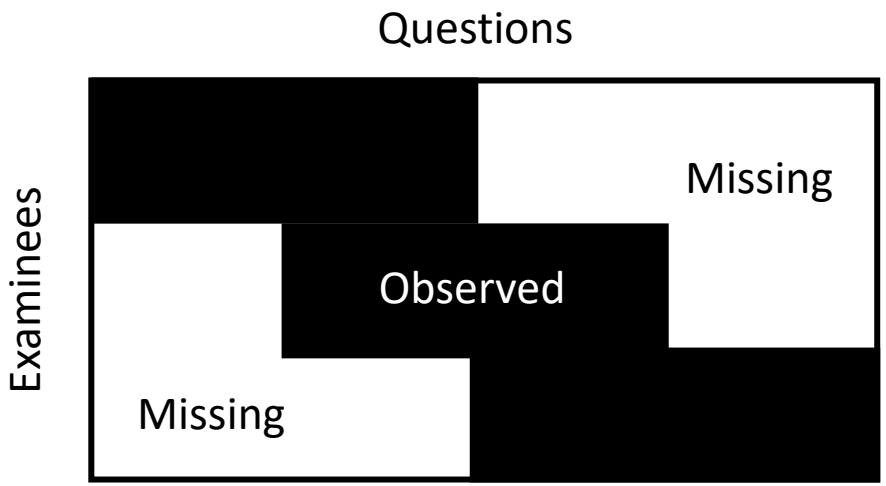
Compare Examinees based on Different Test Forms

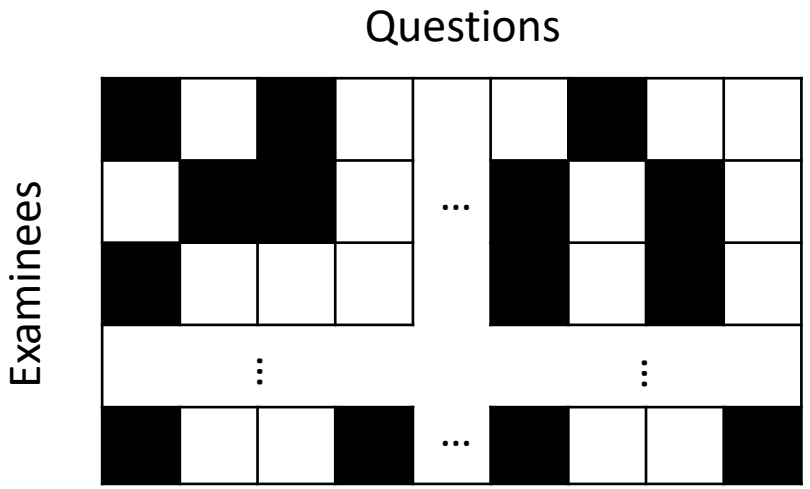
- It is possible that two students received some common items.

- The comparison is in some sense unbiased if we compare their performance based on these common items. However, we lose information by discarding their answers to the other items. Thus, this comparison has high uncertainty.

Compare Examinees based on Different Test Forms

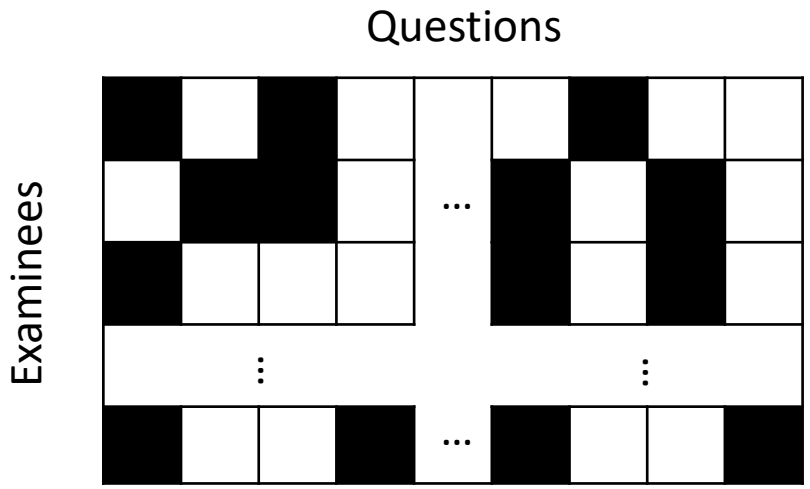
- Ideally, if all the students answer all the questions (which is impossible due to practical concerns such as test security), then we can compare students based on their total scores.
- Idea: Impute all the missing values based on what we observe.





Missing Data Imputation from Matrix Completion Point of View

- Let N be the number of students (i.e. rows) and J be the number of questions (i.e., columns).
- Let $\omega_{ij}, i = 1, \dots, N, j = 1, \dots, J$, be indicators of missingness, where $\omega_{ij} = 1$, if the (i, j) entry is observed, and $\omega_{ij} = 0$ otherwise.
- Let y_{ij} indicate the performance of person i on question j. We only observe y_{ij} when $\omega_{ij} = 1$. For example, $y_{ij} = 1$ if correct, and $y_{ij} = 0$ if incorrect.
- **We are interested in the (expected) values of y_{ij} for which $\omega_{ij} = 0$.**



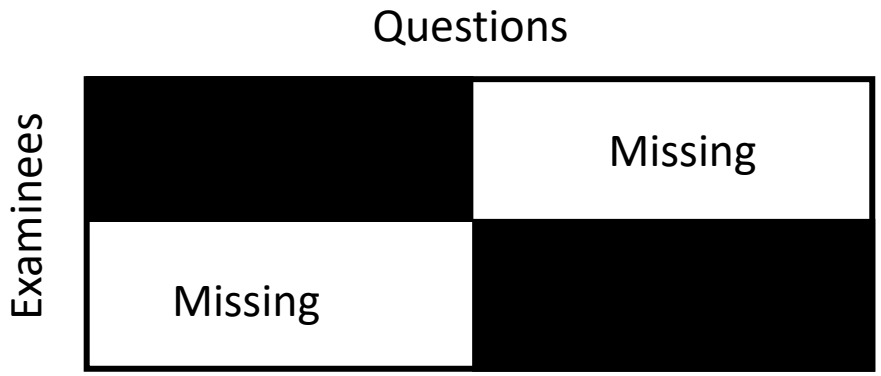
Missing Data Imputation from Matrix Completion Point of View

- This is known as a matrix completion problem, which seems unsolvable at first glance.

- However, this is possible if we know the matrix is generated with certain pattern.
  - For example, say we know $y_{ij} = \theta_i + \beta_j$.
  - Then by the observed y_{ij}s, we can write $n = \sum_{i,j} \omega_{ij}$ equations, for which we have $(N + J)$ unknown parameters.
  - Roughly speaking, as long as $n > N + J$, then we have more equations than parameters. As a result, we can find unique solutions for θ_i and β_j.

Missing Data Imputation from Matrix Completion Point of View

- Of course, the missingness pattern needs to satisfy some regularities. For example, matrix completion is not possible under the assumptions above if missing data are structured like this, as there is no common items for comparing the two groups.



- More generally, it is often assume the (y_{ij}) matrix is of low rank, i.e., $y_{ij} = \theta_{i1}\beta_{j1} + \cdots + \theta_{iK}\beta_{jK}$, for some small K.

Missing Data Imputation from Matrix Completion Point of View

- In practice, the (y_{ij}) matrix cannot be exactly a low rank matrix. Need a probabilistic model to capture this deviation from the mathematical model.

- Moreover, we should take into the fact that y_{ij} is a 0/1 variable (or a categorical variable, e.g., $y_{ij} \in \{0,1,2\}$).

Item Response Theory Model

- Item response theory (IRT) is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables.

- An IRT model is a probabilistic model for the joint distribution of $y_{ij}, i = 1, \dots, N, j = 1, \dots, J.$

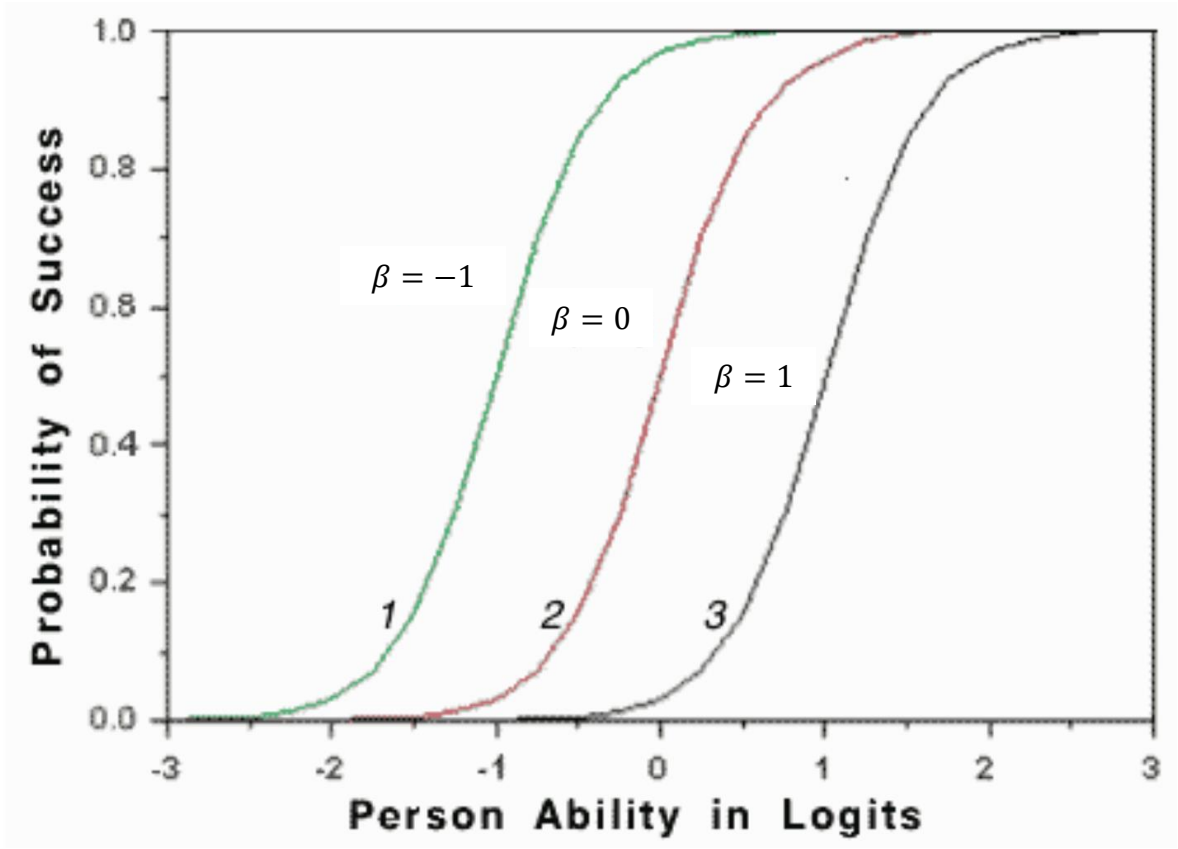
Item Response Theory Model: Assumptions

- Rasch model (Rasch, 1960)
  - Let θ_i be a person-specific parameter. It is often interpreted as the ability of examinee i.
  - Let β_j be an item-specific parameter. It is often interpreted as the difficulty of item j.
  - It is assumed that

$$y_{ij} \sim Bernoulli\left(\frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}\right).$$

  - That is, the probability of correctly answering an item increases with one's ability parameter and decreases with the difficulty parameter of the item.

Item Response Theory Model: Assumptions



Item Response Theory Model: Assumptions

- $y_{ij}, i = 1, \dots, N, j = 1, \dots, J$, are assumed to be independent, given $\theta_i, \beta_j, i = 1, \dots, N, j = 1, \dots, J$.
- This leads to the likelihood function (recall the definition of likelihood function)

$$L(\theta_1, \dots, \theta_N, \beta_1, \dots, \beta_J) = \prod_{\omega_{ij}=1} p_{ij}^{y_{ij}} (1 - p_{ij})^{1-y_{ij}}.$$

- The unknown parameters can be estimated by maximum likelihood estimator, or Bayes and empirical Bayes estimators (where further assumptions are needed).

Item Response Theory Model: Scoring

- Now, say we have estimated the parameters accurately based on the likelihood function (regularity conditions are needed for accurate estimation, for example, large numbers of people and items, as well as, missing data patterns).

- Denote the estimates by $\hat{\theta}_i$ and $\hat{\beta}_j$. Then $\hat{p}_{ij} = \exp\left(\hat{\theta}_i - \hat{\beta}_j\right)/(1 + \hat{\theta}_i - \hat{\beta}_j)$, which predict the expected score of person i on item j.

Item Response Theory Model: Scoring

- Thus, the predicted total score for person i is $\hat{T}_i = \sum_{j=1}^{J} \hat{p}_{ij}$.

- This predicted total score can be used to compare students.

- In fact, under the Rasch model, $\hat{T}_i$ is an increasing function of $\hat{\theta}_i$. That is, $\hat{T}_i > \hat{T}_k$, if $\hat{\theta}_i > \hat{\theta}_k$, and vise versa. It is equivalent to compare students based on the estimated ability parameters.

Item Response Theory Model: Inference

- Sometimes, point estimation is not enough. For example, consider comparing two students. It is possible that $\hat{\theta}_i > \hat{\theta}_k$ due to random error. It is of interest to know whether student i has significantly higher ability than k.

- Then we may test hypotheses: $H_0: \theta_i = \theta_k$ versus $\theta_i > \theta_k$.

- We can do this by using, for example, z-test.

$$z = \frac{\hat{\theta}_i - \hat{\theta}_j}{\sqrt{\widehat{var}(\hat{\theta}_i - \hat{\theta}_j)}}$$

Item Response Theory Model

- Rasch model is the simplest IRT model. There are more complex IRT models.

- In practice, we need to check the goodness of fit of the used IRT model. The evaluation of goodness of fit is similar to the Pearson's Chi-square test (but not the same; will not get into the details here).

Back to A-level Grade Prediction

- Many items from a large system (quiz, mock exam, etc.)
- An IRT model for producing predicted score, which leads to grade prediction.
- Of course, results based on a well-designed and carefully administrated test is more accurate and fair.

Fairness

- Another question in educational testing is to ensure all the test questions are fair to the students with different backgrounds.

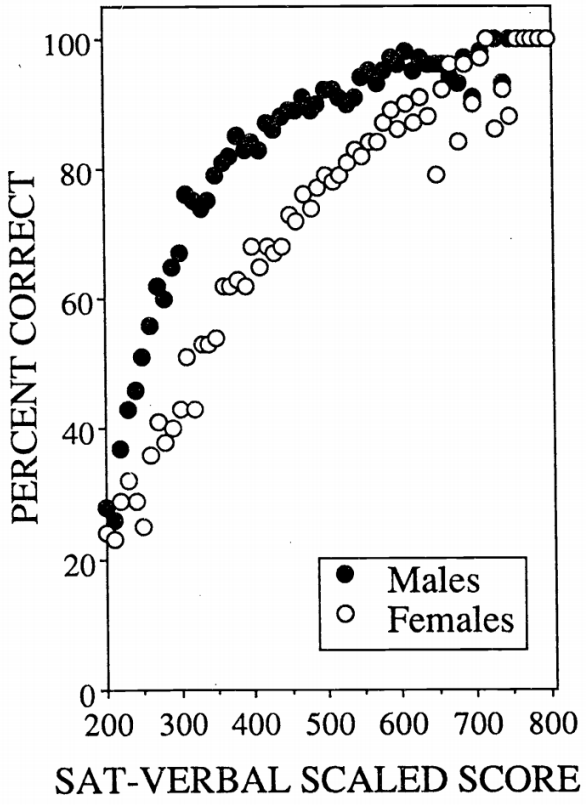
- Differential item functioning occurs when groups (such as defined by gender, ethnicity) have different probabilities of answering a given item correctly after **controlling for the ability being tested in the exam**.

Differential Item Functioning

- Example:

The content for this item, which appeared on the December 1977 form of the SAT, reveals why there is such large DIF on this item. It is a verbal analogy item,

DECOY : DUCK :: (A) net : butterfly (B) web : spider
(C) lure : fish (D) lasso : rope (E) detour : shortcut.



- To detect DIF, we need most of the items in the test to be DIF-free, so that the scaled score accurately represents students' ability being tested.

Differential Item Functioning

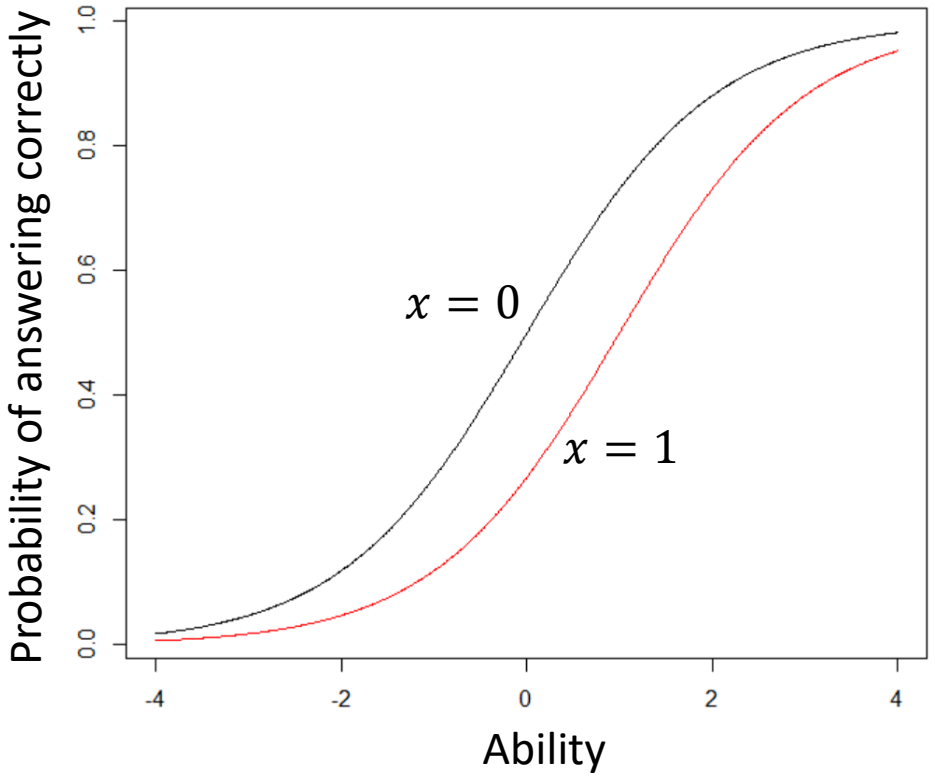
- Detecting DIF under IRT:
  - Let $x_i \in \{0,1\}$ indicate the group, where $x_i = 0$ is the reference (i.e., baseline) group and $x_i = 1$ is the focal group.
  - Assume

$$y_{ij} \sim Bernoulli\left(\frac{\exp(\theta_i - \beta_j + x_i\gamma_j)}{1 + \exp(\theta_i - \beta_j + x_i\gamma_j)}\right).$$

  - There are a small number of non-zero γ_js.

Differential Item Functioning

- E.g., Item response functions when $\beta_j = 0, \gamma_j = -1$.



Differential Item Functioning

- To test DIF for item j, we may test

$$H_0: \gamma_j = 0 \ \ VS \ \gamma_j \neq 0,$$

or test

$$H_0: \gamma_j = 0 \ \ VS \ \gamma_j < 0,$$

Special Features of Data Science in Educational Testing

- Compare with data problems in some areas (e.g., algorithms for recommending tik-tok videos), the prediction and inference problems in educational testing are of higher-stake.
- We often not only care about the prediction accuracy, but also have other considerations, such as fairness.
- The analysis of educational testing data requires methods that are not commonly used in traditional statistics (e.g., regression models).
- People with strong statistics and data science background are of strong need in this field.

Summary

- Grade prediction as a missing data imputation problem.
- Item response theory as the statistical framework for solving this problem.
- Differential item functioning for ensuring fairness.

Thank you!