ACH2053 - Introdução à Estatística
EP - Parte II

Kaic Bastidas 8516048 | Matheus Cañón 8516517 | Thiago Nobayashi 8516708

July 4, 2016

1 Teoria de Resposta ao Item

A Teoria de Resposta ao Item (TRI) considera o seguinte modelo teórico para descrever a probabilidade de um aluno $\boldsymbol{j}$ acertar a questão $\boldsymbol{i}$:

$$Pr\left(A_{i,j}=1\right)=\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}$$

Foram utilizados os dados do grupo 04 para a resolução do Exercício Programa.

2 Estimador Pontual - Alunos

Para a estimação do parâmetro habilidade (θ_j), utilizou-se a técnica de Regressão Logística, considerando o próprio modelo teórico do TRI como função logito; e o Estimador de Máxima Verossimilhança (*Maximum-Likelihood Estimation* - MLE), para estimar o coeficiente da regressão. Dada a variável aleatória Y_{ij}, que segue uma distribuição de *Bernoulli*, a função de vizinhança é:

$$\begin{aligned}L\left(\hat{\theta};a,b,y\right)&=\prod_{i=0}^{n}f(y_i|a_ib_i\theta)\\&=\prod_{i=0}^{n}\left(\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)^{y_i}\left(1-\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)^{1-y_i}\end{aligned}$$

E a função que a maximiza:

$$\begin{aligned}l\left(\hat{\theta};a,b,y\right)&=\log\prod_{i=0}^{n}\left(\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)^{y_i}\left(1-\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)^{1-y_i}\\&=\sum_{i=1}^{n}y_i\log\left(\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)+(1-y_i)\log\left(1-\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}\right)\end{aligned}$$

A solução passa a ser, portanto, a resolução de um Problema de Otimização Escalar:

$$\frac{\partial l\left(\hat{\theta}; a, b, y\right)}{\partial \hat{\theta}} = 0$$

Para a busca da raíz, foi utilizado o Método da Bissecção, com os parâmetros $\boldsymbol{A}$ = -5, $\boldsymbol{B}$ = 5, $\boldsymbol{MAX}$ = 1000 e $\boldsymbol{\varepsilon}$ = 0.00001:

1. Enquanto n $<\boldsymbol{MAX}$;
2. $\boldsymbol{C} \leftarrow \frac{(\boldsymbol{A}+\boldsymbol{B})}{2}$;
3. $\boldsymbol{tmp} \leftarrow \frac{\partial l\left(\hat{\theta}; a, b, \boldsymbol{C}\right)}{\partial \hat{\theta}}$;
4. Se $\boldsymbol{tmp} = 0$ ou $\frac{(\boldsymbol{B}-\boldsymbol{A})}{2}$ ¡ $\boldsymbol{\varepsilon}$:

   Retornar $\boldsymbol{C}$;
5. Se o sinal de $\boldsymbol{tmp}$ for igual ao sinal de $\frac{\partial l\left(\hat{\theta}; a, b, \boldsymbol{A}\right)}{\partial \hat{\theta}}$:

   $\boldsymbol{A} = \boldsymbol{C}$;
6. Senão:

   $\boldsymbol{B} = \boldsymbol{C}$;

Por meio da execução deste algoritmo, o resultado abaixo foi obtido:

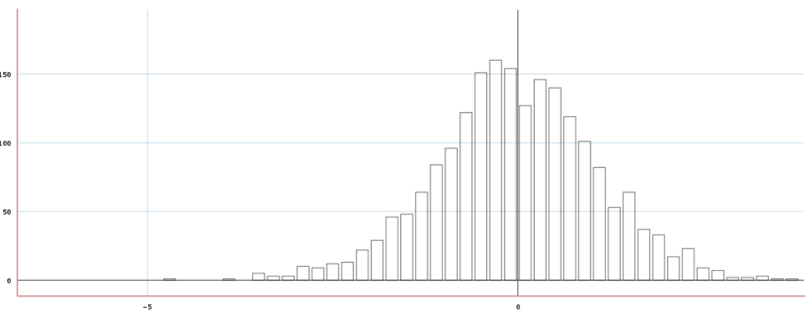

Figure 1: Histograma das habilidades dos 2000 alunos.

Os resultados tiveram os seguintes dados coletados:

- Melhor $\hat{\theta}$: 3.6740875244140625
- Pior $\hat{\theta}$: -4.6877288818359375
- μ : -0.026756973266601564
- σ^2 : 1.2157456898905663

Fica subentendido que os valores estimados tendem a uma normal padrão (média = 0 e desvio padrão = 1).

3 Selecionando o melhor aluno - Habilidade do Aluno

Para a resolução deste problema, o seguinte algoritmo foi executado:

1. primeiramente foram selecionadas as melhores provas de tamanho n (provas de 10, 20, 50 e 100 questões que maximizam as probabilidades de o aluno 5 ser melhor que o aluno 4);
2. Para os alunos 1, 2, 3, 4 e 5:
3. Para cada questão de cada exame, foi sorteado um número de 0 a 1;
4. Caso o número aleatório fosse menor ou igual a probabilidade do aluno acertar (dada pelo modelo matemátido do TRI):

   A questão é acertada;
5. Senão:

   A questão é errada;
6. A partir destes dados, o método de estimação descrito na seção anterior foi executado 1000 vezes para cada aluno em cada prova;
7. A probabilidade é obtida por meio do cálculo de quantas vezes em que a estimação da habilidade do aluno 5 for maior que a estimação da habilidade do outro aluno;

Diferente do EP1, as notas dadas agora são baseadas nas habilidades dos alunos e não na quantidade de acertos. Contudo, as probabilidades ficam parecidas, mas com leves mudanças. A superioridade da habilidade do aluno 5 em relação aos outros alunos fica mais evidente.

4 Intervalo de confiança - Habilidade do Aluno

O algoritmo da seção anterior foi utilizado, porém sem realizar a comparação de habilidades. Neste caso, os resultados obtidos por meio da estimação foram armazenados e ordenados. O intervalo de confiança foi criado a partir da extração de 5% de ambas a extremidades dos dados ordenados.

Ao remover 5% de cada extremidade, ficamos com um conjunto de 90% dos dados (como pedido no enunciado), e se considerarmos que a variável aleatória das habilidades provem vem de uma normal padrão, faz mais sentido ainda a escolha destes limites (pois estes seriam os que conteriam a menor quantidade de elementos).

Ficou claro a dificuldade do algoritmo estimar um valor para o $\hat{\theta}$ para o aluno um (θ real = -1) na prova de 10 questões, pois esse aluno conseguia errar todas as 10 questões. Fazendo assim o $\hat{\theta}$ se aproximar ao limite esquerdo, no nosso caso -5.

Além disso, nota-se que o $\hat{\theta}$ do aluno 5 é maior que o do 4, que é maior que o do 3, e assim respectivamente.

5 Intervalo de confiança - Distribuição Normal

Podemos considerar que cada resposta a uma questão é uma Bernoulli com média p e variância $p(1-p)$, e esse p é descrito pelo modelo como:

$$p = Pr\left(A_{i,j} = 1\right) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

Uma prova é um conjunto de varias variáveis Bernoulli, com parâmetros p diferentes. Podemos considerar então que a média é:

$$\mu(Prova_n) = \sum_{i=1}^{n} p_i$$

e a variância é:

$$\sigma^2(Prova_n) = \sum_{i=1}^{n} p_i(1 - p_i)$$

Dados estes valores podemos considerar que estas somas de Bernoullis tende a uma normal. Para podermos realizar o cálculo, temos de transformar essa normal obtida em uma normal padrão. Realizamos isto desta maneira:

$$S(Prova_n) = N(\mu, \sigma)$$
$$\frac{S(Prova_n) - \mu}{\sigma} = N(0, 1)$$

Tendo o mesmo raciocínio que anteriormente, o objetivo é encontrar um X tal que:

$$Pr(S(Prova_n) < X) = 0,95$$

e

$$Pr(S(Prova_n) < X) = 0,05$$

Considerando a dedução anterior, temos que:

$$Pr(\frac{S(Prova_n) - \mu}{\sigma} < \frac{X - \mu}{\sigma}) = 0,95$$
$$Pr(\frac{S(Prova_n) - \mu}{\sigma} < \frac{X - \mu}{\sigma}) = 0,05$$

Fazendo a consulta na tabela da normal padrão:

$$\frac{X - \mu}{\sigma} = \phi$$
$$X = \phi\sigma - \mu$$

O resultado desse cálculo gera os limites superiores e inferiores do intervalo.

Vale ressaltar que diferente do último intervalo calculado, este tem como resultado notas dos alunos nas provas (0 a n, onde n é a quantidade de questões da prova).

Os resultados obtidos são incrivelmente precisos. No EP1 calculamos este intervalo por meio de simulação, e ficava claro que conforme aumentávamos a quantidade de provas realizadas o intervalo ficava cada vez mais igual.

Os resultados obtidos com esse método se comparam com o método do EP1 para provas com poucas questões. Por exemplo, o aluno 1 na prova de 10 questões no EP1 teve um intervalo de 0 até 4, agora encontramos -0.11618871088737115 a 3.9891749300889714. Contudo, podemos repara que alguns intervalos ficam maiores, como por exemplo o aluno 5 na prova de 100 questões, onde no método do EP1 obteve 54 a 59, e agora 48.139376046615695 a 64.40727053097356.