Universidade de Brasília
Intituto de Ciências Exatas
Monografia apresentada como requisito parcial para obtenção do título de bacharel em Estatística pela Universidade de Brasília.

Testes de Adequabilidade de Ajuste Para Modelos de Teoria de Resposta ao Item Aplicados a Prova do ENEM 2013

Frederico Barros Diniz

Brasília

2015

Frederico Barros Diniz

Testes de Adequabilidade de Ajuste Para Modelos de Teoria de Resposta ao Item Aplicados a Prova do ENEM 2013

Monografia apresentada como requisito parcial para obtenção do título de bacharel em Estatística pela Universidade de Brasília.

Orientador: Prof. Dr. Antônio Eduardo Gomes

Brasília

Novembro de 2015

À minha mãe Fátima e minha irmã Rafaela

Agradecimentos

Agradeço, primeiramente, ao meu orientador, Prof. Dr. Antônio Eduardo Gomes que, além de ser um grande pesquisador, é um grande professor. Sempre muito solícito e atencioso, me apoiou em todos os momentos deste trabalho com toda sua dedicação e competência.

Agradeço ao Prof. Dr. Joaquim José Soares Neto pelo apoio e prestatividade em compartilhar seu vasto conhecimento sobre o assunto.

Agradeço aos meus colegas de graduação, que se tornaram queridos amigos ao passar desses cinco anos. Em especial ao Tomás Moura da Veiga que desde o meu primeiro dia na graduação me incentivou a continuar.

Agradeço aos meus amigos, que fiz ao longo da minha vida, sendo sempre importantes para o meu desenvolvimento.

Dedico este trabalho a minha mãe, Fátima, a minha irmã, Rafaela.

Resumo

Neste trabalho, são apresentados testes que analisam se os modelos não paramétricos se adequam ao modelo logístico de três parâmetros, chamamos esse procedimento de teste de adequabiliade de ajuste para modelos de teoria de resposta ao item (TRI) iremos aplica-los a uma amostra das questões do Exame Nacional do Ensino Médio (Enem), utilizando de técnicas de regressão isotônica e de Regressão via o Estimador de Nadaraya-Watson. Os testes foram realizados no software R, além disso, foram feitas simulações para observar seu desempenho na avaliação da adequabilidade de modelos paramétricos para a curva característica do item.

Palavras Chave: Teste de Adequabilidade, Regressão Isotônica, Estimador de Nadaraya-Watson, Teoria de Resposta ao Item, Exame Nacional do Ensino Médio.

Sumário

Lista de Quadros e Tabelas

Capítulo 1

Introdução

Os modelos estatísticos de Teoria de Resposta ao Item (TRI), em suas muitas formas, são amplamente utilizados em programas de avaliação. É uma forma mais versátil de medir o conhecimento dos candidatos. Bem diferente dos modelos clássicos de avaliação, onde o escore final de acertos é contabilizado. A TRI não contabiliza apenas o número total respostas corretas no teste, o foco principal da avaliação é o item.

Podemos listar algumas provas e testes que utilizam a TRI:

- Sistema de Aliação da Educação Básica (SAEB) - Primeira prova que utilizou a Teoria de Resposta ao Item no Brasil;

- TOEFL - Exame que avalia a proeficiênca do candidato de falar e entender o inglês em nível acadêmico. Mundialmente usado como instrumento de entrada para estrangeiros em instituições de ensino de língua inglesa;

- SAT (Scholastic Aptitude Test ou Scholastic Assessment Test) - Prova americana semelhante ao que conhecemos como vestibular. Ele é um dos sistemas de avaliação para estudantes de ensino médio que desejam ingressar em universidades norte americanas.

O Exame Nacional do Ensino Médio (Enem) é uma das provas que utilizam modelos da TRI para avaliar os candidatos. É considerada a maior avaliação realizada no Brasil, pois, além de servir como porta de entrada para o ensino superior também é um excelente indicador de qualidade do ensino básico.

Avaliar o ensino é de suma importância para o desenvolvimento do país. Olhando por esse lado, esse trabalho não tem somente foco em modelos de Teoria de Resposta ao Item e seus respctivos testes de adequabilidade. Podemos ir além, e tratar o Exame como um dos instrumentos de pesquisa. Com o aprimoramento de técnicas estatísticas que melhorem a avaliação dos candidatos, estamos contribuindo também para um ensino mais justo e igualiário.

Aqui serão exposto os principais modelos de Teoria de Resposta ao Item (TRI) e suas aplicações, sempre lembrando que o interesse final é comparar modelos já consolidados detro da TRI com métodos ainda novos e pouco explorados. Como é o caso dos modelos não paramétricos que iremos abordar. Os métodos não paramétricos são computacionalmente mais simples e exigem menos processamento das máquinas.

Para a realização do trabalho, utilizaremos uma amostra dos dados extrídos da prova do Exame Nacional do Ensino Médio (Enem) no ano de 2013. Por ser um Exame com o número muito grande de candidatos, não podemos utilizar a população de respondentes da prova, que no ano de 2013 teve por volta de 7 milhões de inscritos. A análise dos dados foi feita no Softwere R.

Foram realizados testes de adequabilidade de ajustes para comparar os modelos paramétricos e os modelos não paramétricos propostos. Os testes de adequabilidade ou testes de aderência são técnicas estatisticas que permitem a comparabilidade entre formas de modelos. No nosso caso, vamos comparar os modelos não paramétricos obtidos via Regressão Isotônica e via o Estimador de Nadaraya-Watson com o modelo logístico de três parâmetros, que é utilizado na prova do ENEM.

Depois de aplicar os devidos testes, obtemos boas estimativas de que os modelos não paramétricos são boas alternativas ao modelos logísitico de três parâmetros. Analisando os resultados, a hipótese de que as curvas não paramétricas se ajustam bem as curvas logísticas de três parâmetros foi aceita na grande maioria das questões estudadas.

Podemos considerar que modelos não paramétricos são boas alternativas para estudos relacionados a Teoria de Resposta ao Item (TRI) e que, além de serem computacionalmente mais simples apresentam boas estimativas da curva característica do item (CCI).

Capítulo 2

O Exame Nacional do Ensino Médio

Nesta sessão será apresentado o Exame Nacional do Ensino Médio (Enem). Para entender a importância do tema, vamos fazer um breve resumo do contexto em que o Exame foi criado, e quais eram os objetivos iniciais da Avaliação. Além de explicar os motivos que impulsionaram a mudança nos moldes da prova e a importância da Teoria de Resposta ao Item nos critérios avaliativos do Enem.

2.1 História do Exame Nacional do Ensino Médio (Enem)

Criado em 1998, o Exame Nacional do Ensino Médio (Enem) tinha como finalidade avaliar os alunos das escolas secundaristas brasileiras. Os resultados obtidos na prova serviriam para elaboração de politicas públicas em prol do desenvolvimento educacional no país.

No primeiro momento o exame não tinha um caráter seletivo, ou seja, não era utilizado na admissão dos alunos em instituições de ensino superior. Era meramente avaliativo e não obrigatório, por isso, em sua edição inicial houve um número discreto de participantes, apenas 157 mil. Além disso, a prova era restrita unicamente à alunos concluintes do ensino médio no ano de aplicação do Exame.

A estrutura original da prova era composta por 63 questões de múltipla escolha divididas em cinco grandes áreas;

1 Ciências da Natureza e suas Tecnologias;

2 Ciências Humanas e suas Tecnologias;

3 Linguagens, Códigos e suas Tecnologias;

4 Matemática e suas Tecnologias;

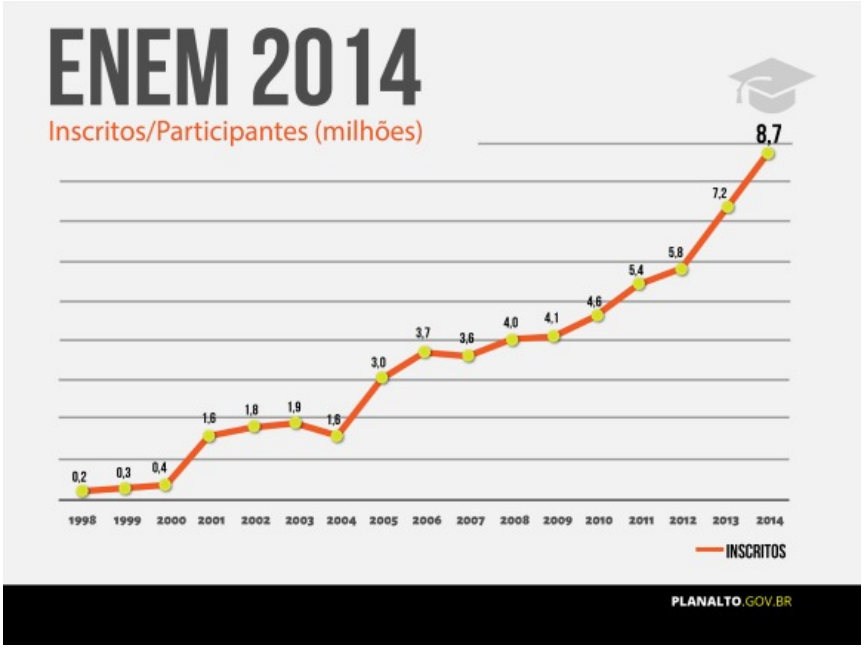
5 Redação dissertativa com tema variado.

Não demorou muito para que o Ministério da Educação percebesse a influência da prova no cenário educativo brasileiro. Seria possível utilizar o Enem como uma poderosa ferramenta de inclusão social. Logo algumas mudanças estruturais foram feitas no Exame dando a ele uma maior versatilidade.

Em 2004 foi lançado o programa governamental Universidade para todos (ProUni). Seriam distribuídas bolsas de estudo em instituições particulares de ensino superior, e o programa utilizaria o exame nacional do ensino médio (Enem) como critério de seleção. As técnicas de avaliação do estudante ainda eram falhas e mais simples, além de não uma ampla divulgação como ocorre atualmente.

A partir da criação do ProUni, o número de candidatos aumentou significativamente, impulsionados também pela isenção da taxa de inscrição para alunos de escolas públicas. Com todo o dinamismo que o Exame mostrava ter, o Ministério da Educação percebeu que poderia ir além, e fazer uma prova unificada de admissão para universidades públicas federais.

Abaixo temos um gráfico com a série histórica do número de participantes no Exame Nacional do Ensino Médio até 2014..



2.2 Reformulação do Exame Nacional do Ensino Médio (Enem)

A ideia de unificar a prova de admissão para as principais instituições de ensino superior (IES) públicas do país foi vista como uma revolução para o sistema educacional brasileiro. Houveram muitas críticas quanto a viabilidade do projeto, mas em 2009, depois de uma total modificação nos critérios avaliativos, o Exame Nacional do Ensino Médio (Enem) foi aplicado como meio de seleção integral ou parcial para dezenas de Universidades Federais.

O Exame continuou sendo o meio de seleção para o programa Universidade para Todos (ProUni). Fazendo com que, em 2009, a prova chegasse a marca histórica de 4,1 milhões de inscritos.

2.2.1 Mudanças Significativas

As mudanças estruturais no Exame foram bastante significativas. O desafio de implementar uma única prova que medisse igualitariamente o conhecimento dos

alunos de todas as regiões do Brasil era grande. Nas primeiras edições da prova, temos observados vários problemas relacionados a segurança e vazemento de provas.

Foi criado um modelo mais completo de prova, onde o candidato é avaliado em todos os quesitos necessários para o ingresso no ensino superior. O Exame que antes continha por 63 questões, agora tem 180 itens de múltipla escolha. Além disso, devido a complexidade da prova, o que antes era realizado em apenas um dia de avaliação passou a demandar 2 dias para a execução, .

O Enem era destinado apenas à alunos concluintes do ensino médio, porém, passou a ser aplicado a todas as pessoas que possuíam o segundo grau completo e desejavam realizar o Exame. Essa medida é justificável por se tratar do processo seletivo de universidades, tendo em vista que nem todos obtêm a aprovação na primeira vez que realizam a prova.

2.2.2 Vantagens da Reformulação

Se um aluno optasse por fazer o processo seletivo para mais de uma Instituição de Ensino Superior (IES), caberia a ele o deslocamento entre cidades para a realização do vestibular. Com a unificação do Enem, ele poderia fazer uma única prova e concorrer em mais de uma universidade. Esse é um importante fator de fluxos migratórios no país, gerando uma imensa interação entre as regiões brasileira, históricamente segregadas.

O Ministério da Educação (MEC) adotou outras medidas positivas em relação ao Enem. Além de servir como processo de seleção para diversas IES e para o ProUni, a prova também serve como avaliação para a conclusão do ensino médio, ou seja, caso um aluno se ache preparado para receber o diploma de concluinte do ensino médio e for maior de 18 anos, ele pode realizar Enem, e dependendo do resultado, receber junto ao MEC o diploma de concluinte do ensino médio.

2.2.3 Desvantagens da Reformulação

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão governamental ligado ao Ministério da Educação e responsável pela realização da prova, tem sérias dificuldades em relação a segurança da informação.

O Exame Nacional do Ensino Médio, devido a sua complexidade, acaba necessitando de muito pessoal para a realização. O problema com segurança é eminente a partir do momento em que muitas pessoas são envolvidas na organização da prova.

O Brasil é um país de vasta extensão territorial, onde há áreas com grandes problemas de acessibilidade. Atingir as regiões mais inóspitas do país é um grande desafio e exige uma logística gigantesca para concluir tal ação.

Mesmo com tantas dificuldades a prova vem sendo um sucesso e motivando estudos para que melhore cada vez mais. É o caso deste trabalho, que visa buscar formas alternativas que contribuam com o aprimoramento do Exame Nacional do Ensino Médio (Enem).

2.2.4 Uso da Teoria de Resposta ao Item na Prova do Enem

O Ministério da Educação tem como objetivo principal, ao adotar o Enem como forma de avaliação para admissão nas principais Instituições de Ensino Superior, tornar a acessibilidade à educação mais igualitária.

A Teoria de Resposta ao Item (TRI) tem como característica marcante o fato de ser mais descriminante quanto a proficiência dos candidatos avaliados. A utilização da TRI no Enem deu-se por dois motivos;

1 Permite a Comparabilidade dos resultados entre provas diferentes;

2 Permite a aplicação do Exame várias vezes ao ano (Há limitações financeiras que ainda não permitem essa medida.)

No próximo capítulo desse trabalho iremos explicar os fundamentos teóricos que nos permitem afirmar as condições apresentadas acima e justificar sua usabilidade no Exame Nacional do Ensino Médio.

Capítulo 3

Teoria de Resposta ao Item (TRI)

Nesse capítulo vamos descrever os fundamentos básicos da Teoria de Resposta ao Item (TRI). Onde tentaremos elucidar os seguintes tópicos:

- A História da Teoria de Resposta ao Item;
- Os Conceitos Básicos da Teoria de Resposta ao Item;
- Modelo Logístico de Três parâmetros ;
- Os Modelos Não-Paramétricos;
- Usabilidade em Avaliações.

O conceito chave para que possamos entender o que é a Teoria de Resposta ao Item é o de Teste de Avaliação.

Os testes de avaliação, são procedimentos sistemáticos de observação e registro de amostras sobre comportamentos ou respostas de indivíduos com o objetivo de mensurar características ou proficiência.

3.1 Breve Histórico da Teoria de Resposta ao Item (TRI)

Desde o início do século XX, pesquisadores trabalham com o objetivo de aprimorar as técnicas aplicadas nos testes avaliativos. A evolução no começo era lenta e limitada pela falta recursos computacionais. Com o passar dos anos vários avanços foram obtidos, entre eles a elaboração da Teoria de Resposta ao Item (TRI) de forma viável, que poderia ser executada computacionalmente. Por depender de modelos logísticos, a TRI demanda muito recurso tecnológico para ser elaborada.

Na primeira metade do século, a hoje conhecida como Teoria Classica dos Testes (TCT) era a principal metodologia aplicada em modelos de avaliação. A TCT apresentava muitos problemas estruturais, o que levava muitos testes a falharem em sua mensuração. Se fossemos analisar a proeficiencia de um candiato utilizando a Teoria Classica dos Testes, deveriamos observar o escore obtido por ele, ou seja, a soma dos itens respondidos corretamente. Nesse caso os resultados seriam analisados pelos parâmetros de discriminação e de dificuldade, fazendo com que o instrumento construído dependa diretamente da dificuldade da prova, da nota do candidato e da amostra de candidatos respondentes do teste.

Os primeiros estudos sobre a viabilidade de comparar as habilidades e os conhecimentos de examinandos submetidos a provas diferentes surgiram na década de 50. Os modelos utilizavam um único parâmetro e os itens eram corrigidos de maneira dicotómica, ou seja, apenas acertos ou erros. Estes modelos foram, em um segundo momento, desenvolvidos na forma de uma função ogiva normal com dois parâmetros e, depois, evoluíram para modelos logísticos com três parâmetros, utilizados atualmente na prova do Enem.

3.2 A Teoria de de Resposta ao Item

A TRI é usada para analisar dados provenientes de respostas a itens presentes em instrumentos avaliativos, e sugere formas de representar a probabilidade de um individuo dar uma determinada resposta a um item levando em conta os suas habilidades (traços latentes) e algumas características do item. Essa relação é modelada através de funções de ligação simétricas. Tal relação é conhecida como Curva Característica do Item (CCI).

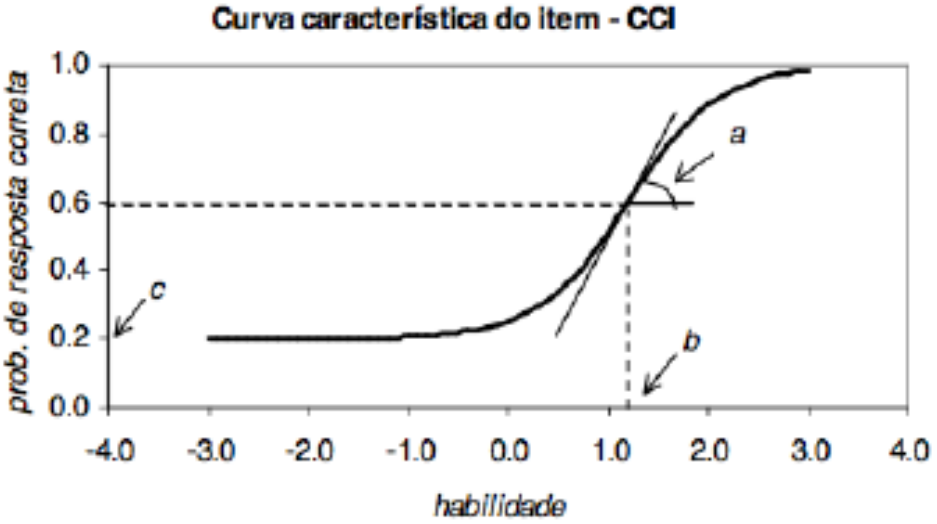
Os modelos logísticos para itens dicotômicos são os modelos de resposta ao item mais utilizado, sendo que há basicamente três tipos, que se diferenciam pelo número de parâmetros que utilizam para descrever o item. Eles são conhecidos como os modelos logísticos de 1, 2 e 3 parâmetros, que consideram, respectivamente:

i) Somente a dificuldade do item;

ii) A dificuldade e a discriminação;

iii) A dificuldade, a discriminação e a probabilidade de resposta correta dada por indivíduos de baixa habilidade.

Os modelos mais utilizados dentro da TRI atualmente são paramétricos e assumem forma funcional especifica para a curva característica do item (CCI), normalmente logística. A CCI do item evidencia o progresso da probabilidade de acerto do item de acordo com a proeficiência do respondente:



O parâmetro de discriminação *a* é proporcional à inclinação da CCI. Quanto maior esse parâmetro, maior é a capacidade do modelo em diferenciar a probabilidade de acerto de um item em relação ao nível de habilidade. Ou seja, esse parâmetro mede a qualidade de descriminação do item

Para resultados baixos de *a*, essa probabilidade não apresenta muita variação em diferentes níveis de habilidade.

O parâmetro de acerto ao acaso ou chute *c* mede a probabilidade de acerto de um candidato com pouca habilidade no tema. Ele evidencia que pessoas com pouco conhecimento sobre o tema tambem tem uma probabilidade de acertar o item. No exemplo mostrado acima, o parâmetro *c* é igual a 0,2.

O parâmetro de dificuldade *b* indica a habilidade necessária para uma probabilidade de acerto do item equivalente a $(1 + c)/2$. Isto é, quanto maior o valor de *b*, maior o nível de dificuldade do item. Esse parâmetro, portanto, mede a dificuldade do item.

O modelo utilizado pelo Exame Nacional do Ensino Médio, é o logístico de três parâmetros, ou seja, para calcular a proeficiência do candidato em determinado item, todos os parâmetros apresentados acima são utilizados.

Essa probabilidade de acerto pode ser calculada da seguinte forma:

$$P(U_{ij} = 1/\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}. \tag{3.1}$$

Segundo (Santos, 2000) os vários modelos propostos na literatura dependem fundamentalmente de três fatores:

i) Da natureza do item - dicotômicos ou não dicotômicos;

ii) Do número de populações envolvidas - apenas uma ou mais de uma;

ii) E da quantidade de traços latentes que está sendo medida - apenas um ou mais de um.

3.3 Modelos Não Paramétricos da TRI

Como já foi apresentado, o aprimoramento do modelos de avalição é constante. Devido a complexidade computacional de muitos modelos paramêtricos, surgestões alternativas começaram a ser desenvolvidas. Modelos Não paramétricos são de alta usabilidade nesse caso, pois exigem menos de recursos computacionais.

Segundo Schlemper (2010) para a construção de um teste, devemos levar em consideração que nem sempre alguns aspctos importantes como interpretação, aplicabilidade e viabilidadem convergem para uma mesma direção. A contrução de uma escala rozoável na TRI é, portanto, correlacionada com a capacidade de encontrar um equilíbrio entre os requisitos.

Podemos listar três motivos, além de facilidade computacional, para a utilização de modelos Não Paramétricos:

1 Conhecimento mais amplo de como modelos paramétricos se comportam;

2 Proporcionam um leque maior de possibilades para aplicações que modelos paramétricos não conseguiriam atingir;

3 Quando há poucos respondentes, possibilitam meios de utilização mais fáceis.

A grande vantagem dos modelos não paramétricos é que neles não existe a necessidade de forçar que o modelo possua um formato logístico ou orgiva da distribuição normal. Na estimação não paramétrica não é preciso que os candidatos sejam ordenados por suas proeficiências estimadas que são não decrescentes.

Segundo Schlemper (2010), outra vantagem é que os modelos não paramétricos podem possuir representação gráfica não ajustada, como no caso paramétrico. Além disso, a diferença (distância) entre o ponto paramétrico e não paramétrico pode ser calculado pelos valores de θ, como veremos na próxima sessão.

Capítulo 4

Testes de adequabilidade de ajuste para modelos de Teoria de Resposta ao Item Aplicados à Prova do ENEM 2013

Nessa Sessão iremos apresentar os testes de ajuste para modelos de Teoria de Resposta ao Item (TRI) que serão aplicados à prova do ENEM 2013. São eles;

- Regressão Isotônica

- Regressão de Nadaraya-Watson

4.1 Regressão Isotônica

A regressão Isotônica tem como objetivo encontrar uma função não decrescente que minimiza a soma de quadrados dos erros. Com isso podemos testar a hipótese de que o item segue um modelo de três parâmetros ou não.

Esse método é utilizado em Teoria de Resposta ao Item para a estimação não paramétrica da Curva Característica do Item, uma vez que, na construção dessas curvas, elas a não decrescem, quando se aproximam da assintota de resposta correta.

O primeiro passo para a estimação é gerar um conjuntos de respostas m para o item a partir da Curva Característica do Item (CCI) paramétrica e ajustada.

Em seguida devemos ajustar a CCI paramétrica e a não paramétrica para cada um dos m conjuntos de dados e calcular a distância, $d_i(\hat{f}_p^j, \hat{f}_n^j)$ com j=1, 2, 3,..., m, entre elas.

A estatística do teste observada, S, é definida por:

$$S = d_i(\hat{f}_p, \hat{f}_n) = \int_{-\infty}^{\infty} |\hat{f}_p(\theta) - \hat{f}_n(\theta)|\phi(\theta)d\theta, \tag{4.1}$$

Sendo assim,

$$d_2(\hat{f}_p, \hat{f}_n) = \int_{-\infty}^{\infty} |\hat{f}_p(\theta) - \hat{f}_n(\theta)|^2\phi(\theta)d\theta. \tag{4.2}$$

Onde ϕ é a função de densidade de probabilidade da distribuição N(0,1).

Essa é a estatística que nos permite testar se a regressão não paramêtrica obtida segue uma regressão logística de três parâmetros.

Com isso temos as seguintes hipóteses:

H_0: O modelo segue um modelo logístico de três parâmetros.
H_a: O modelo não segue um modelo logístico de três parâmetros.

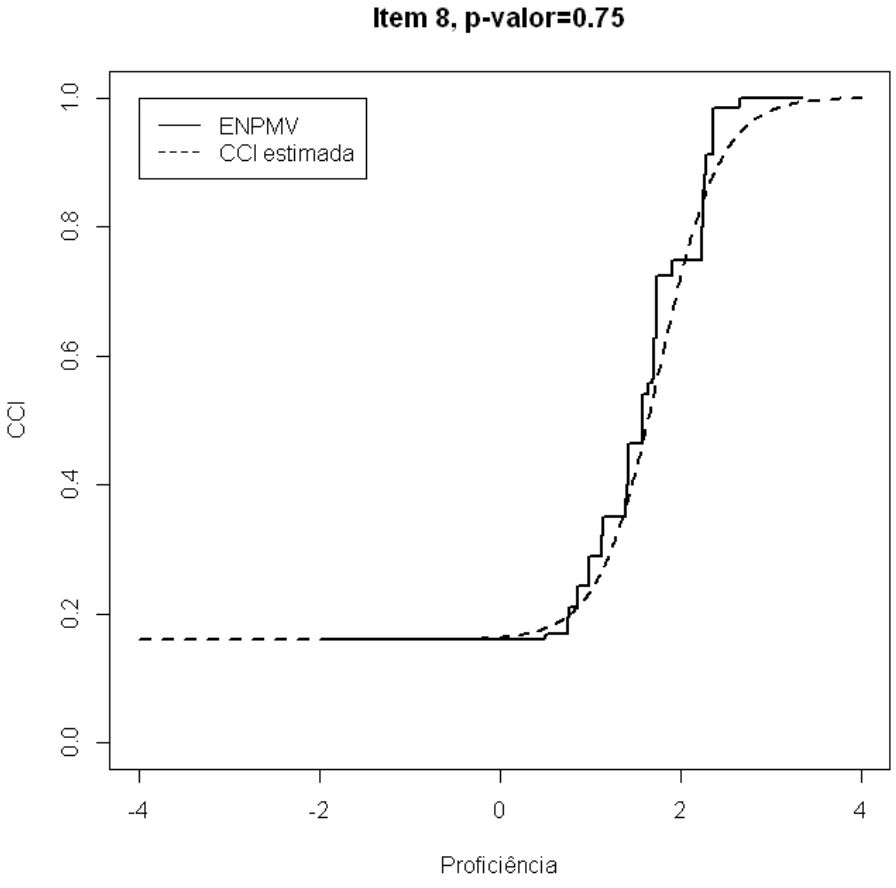
Também podemos obter o p-valor para testar a adequabilidade do modelo.

$$p - valor = \frac{N : d_i(\hat{f}_p, \hat{f}_n) > S}{m}. \tag{4.3}$$

Onde N significa o número de distâncias maiores que a estatística do teste S.

Se o p-valor for muito pequeno, devemos rejeitar H_0

Abaixo segue um gráfico de como a regressão isotônica se ajusta quando comparada com uma regressão linear.



Para realizar os testes necessitamos obter o p-valor associado ao valor correspondente a estatística do teste . Em nosso caso, quanto maior o valor da estatística de teste maior a evidência contra a suposição que o modelo ajustado se ajusta bem aos dados. O p-valor vai nos indicar se o valor observado da estatística é atipicamente grande sob a suposição de veracidade da hipótese nula (suposição de que os dados seguem um modelo paramétrico ajustado.). Para isso necessitamos conhecer a distribuição da estatística de teste sob Ho.

Uma maneira de obter o valor p consiste em gerar um grande número de valores a partir da distribuição da estatística de teste sob Ho e observar o posicionamento do valor da estatística de teste para os dados observados em relação a esses valores gerados a partir da estatística de teste. A proporção de valores gerados que estejam acima do valor observado da estatista será nossa estimativa do p-valor. Temos evidências para a rejeição de Ho se o p-valor for muito pequeno.

Para obter valores da estatística sob Ho geramos respostas (certas ou erradas) para um determinado item em função da probabilidade de acerto determinada pelo modelo paramétrico ajustado em função da proficiência de cada respondente. Com estes dados gerados ajusta-se o modelo paramétrico e calcula-se o estimador não paramétrico da CCI, obtendo-se então um valor para a estatísitcca do teste calculando-se S (estatística do teste) isso é repetido um número grande de vezes (10 mil).

4.2 Regressão Via Estimador de Nadaraya-Watson ou Regressão Kernel

Análogo a regressão isotônica, temos o estimador não paramétrico via Nadaraya-Watson, que é calculado com os valores de zeros e um dos respondente.

$$\hat{f}_n(\theta) = \frac{\sum_{j=1}^{n} y_j k\left(\frac{\theta - \theta_j}{n}\right)}{\sum_{j=1}^{n} k\left(\frac{\theta - \theta_j}{n}\right)} \tag{4.4}$$

Onde k representa o Kernel, ou seja, a função densidade de probabilidade simétrica em zero.

O método de suavização via Kernel é um dos mais usados para estimação não paramétrica de CCIs com a finalidade de verificar a adequabilidade do ajuste.

O cálculo do p-valor será igual ao modelo anteriormente apresentado.

4.3 Metodologia

Nessa sessão vamos apresentar a metodologia adotada no trabalho. Principalmente a parte de;

1 - Apresentação dos dados;

2 - Preparação dos dados.

4.3.1 Apresentação dos Dados

A prova do Enem, por razão de segurança, é dividida em quatro cadernos de respostas:

1 Azul;

2 Verde;

3 Amarelo;

4 Vermelho.

Por ser uma prova que utiliza a teoria de resposta ao item, as questões não são as mesmas em todas as provas, mas analisam as mesmas habilidades.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) garante que alunos sentados próximos uns dos outros não compartilhem informações no decorrer da avaliação.

A prova do Enem em 2013 atraiu mais de sete milhões de inscritos, número recorde até então. Por ser tratar de um volume muito grande de dados, foi feito uma amostra desses participantes para poder processar os testes de adequabilidade de ajuste no Softwere R.

O Distrito Federal teve pouco mais de 114 mil incritos no Exame Nacional do Ensino Médio. Entre esses candidatos apenas 89 mil realizaram a prova. Mesmo com o número elevado do abstenção, o caderno azul, que era o objeto de estudo, teve pouco mais de 20 mil candidatos. Número suficiente para a realização dos testes de adequabilidade de ajuste.

Os criérios adotados para seleção dos candidatos foram os seguintes:

1 Selecionar apenas candidatos presentes na prova;

2 Selecionar apenas candidatos do Distrito Federal;

3 Aplicar os testes de adequabilidade apenas nos 45 itens de Matemática e Suas Tecnologias;

4 Selecionar apenas candidatos respondentes do caderno azul.

4.3.2 Preparação dos Dados

Após definir os parâmetros do modelo, é necessário estimá-los. Em Teoria de Resposta ao Item, esse processo é conhecido como calibração dos itens. Também é necessário estimar a habilidade dos indivíduos. Ou seja, a calibração consiste em calcular cada um dos paramétros para os itens observados, para depois gerar as curvas caracteristicas dos itens. Há vários métodos com tal finalidade. O pacote MIRT (Multidimensional Item Response Theory) no softwere R é bastante conhecido e recomendado para a calibração dos itens, e foi nossa primeira escolha para realizar a estimação dos parâmetros. Após realizar uma rotina para calibrar os dados com o MIRT, os resultados obtidos não foram bons, pois, as CCI estimadas estavam muito diferentes das estimativas não paramétricas para todos os itens. A alternativa para contornar esse problema foi rodar os dados com a função TPM do pacote "ltm". Nesse segundo momento, O R apontou problemas com a matriz hessiana, mas forneceu as estimativas desejadas.

Capítulo 5

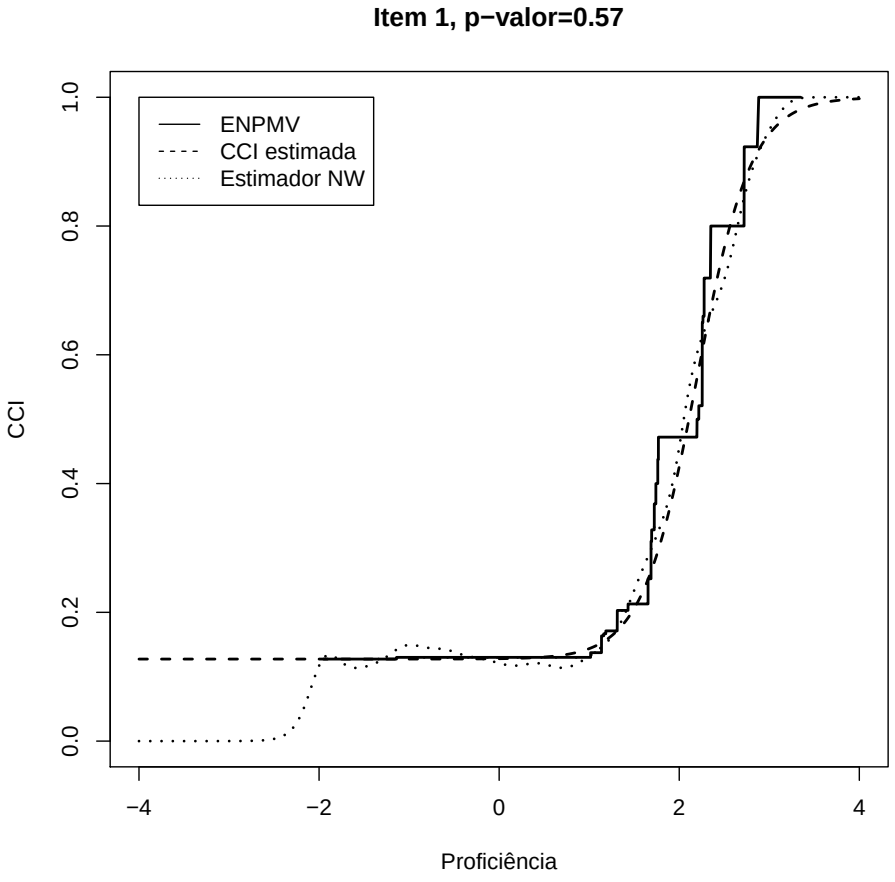
Resultados

Depois de realizados os procedimentos para tratar os dados, aplicamos os testes de adequabilidade para o modelo paramétrico de três parâmetros. O objetivo final era avaliar se os itens, após rodar as regressões não paramétricas, teriam comportamento semelhante ao modelo paramétrico.

O resultado para a grande maioria dos casos foi positivo, ou seja, os modelos não paramétricos se aproximaram dos paramétricos. Porém houve casos em que os itens não seguiram o comportamento esperado. Como será mostrado a seguir.

Os gráficos abaixo irão elucidar melhor os resultados obtidos. Com isso, serão mostrados cinco item que tiveram comportamento bem distintos seguindo as mesmas metodologias. As curvas não paramétricas pontilhadas são referentes aos modelos de regressão isotônica e regressão de Nadaraya-Watson e são denominadas ENPMV e Estimador NW, respectivamente. A CCI estimada corresponde a Curva Característica do Item do modelo logístico de três parâmetros.

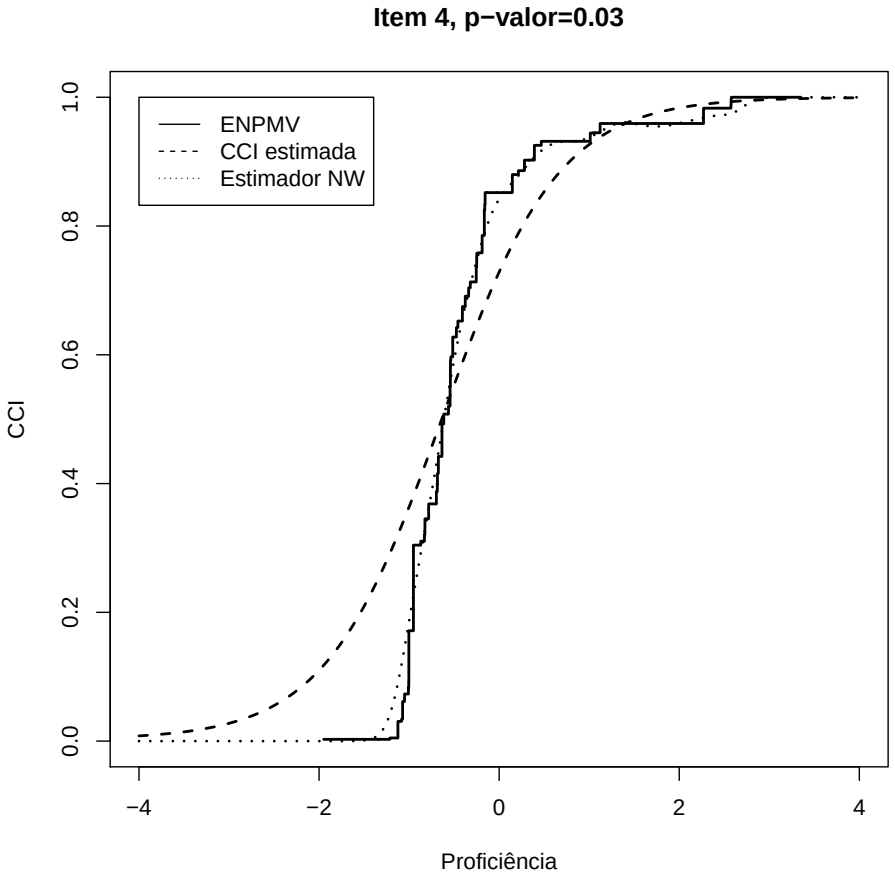
Figura1 - Item 1



O Item 1 foi bem ajustado pelos modelos não paramétricos. Podemos observar que a CCI paramétrica ajustada, o estimador não paramétrico via regressão isotônica e o Estimador NW, convergiram bem durante toda a curva do item.

Outra forma de analisar o item é olhando o p-valor obtido na Tabela 2 e sua respectiva distância. Pelo valor observado nesse item, temos evidências para não rejeitar a hipótese de que os modelo paramétrico seguem um modelo logístico de três parâmetros.

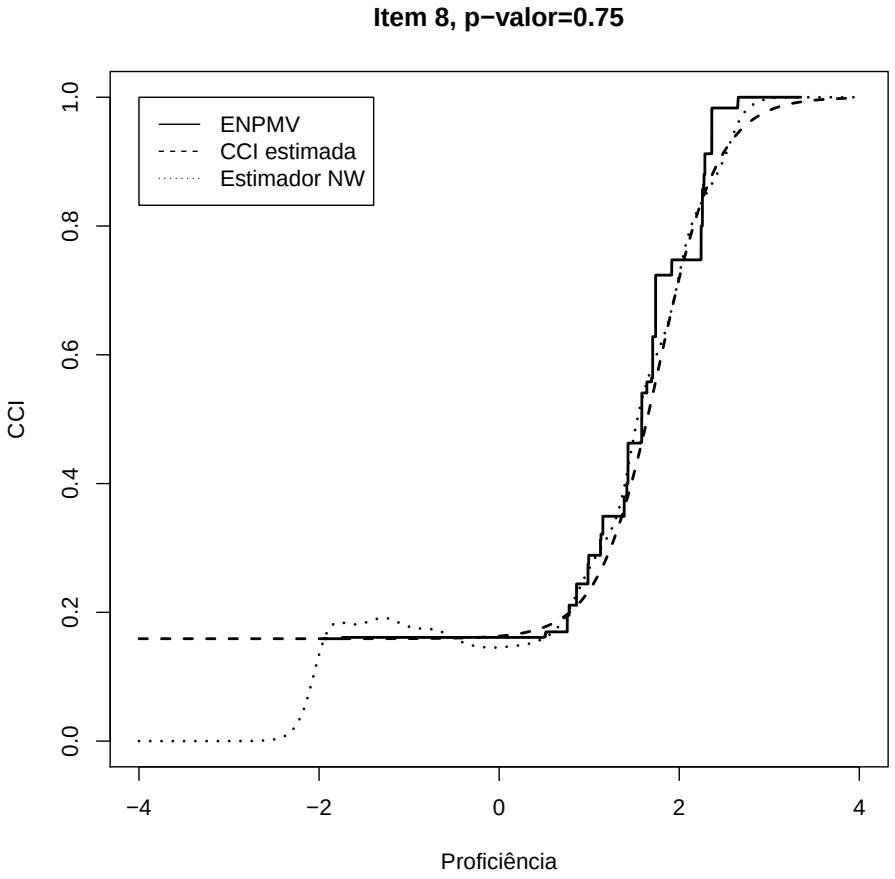
Figura 2 - Item 4



O Item 4 aparentemente foi bem ajustado pelos modelos não paramétricos. Porém, observamos que há uma divergência nas caldas em relação a CCI paramétrica ajustada, o estimador não paramétrico via regressão isotônica e o Estimador NW, não convergiram muito bem em relação a CCI paramétrica.

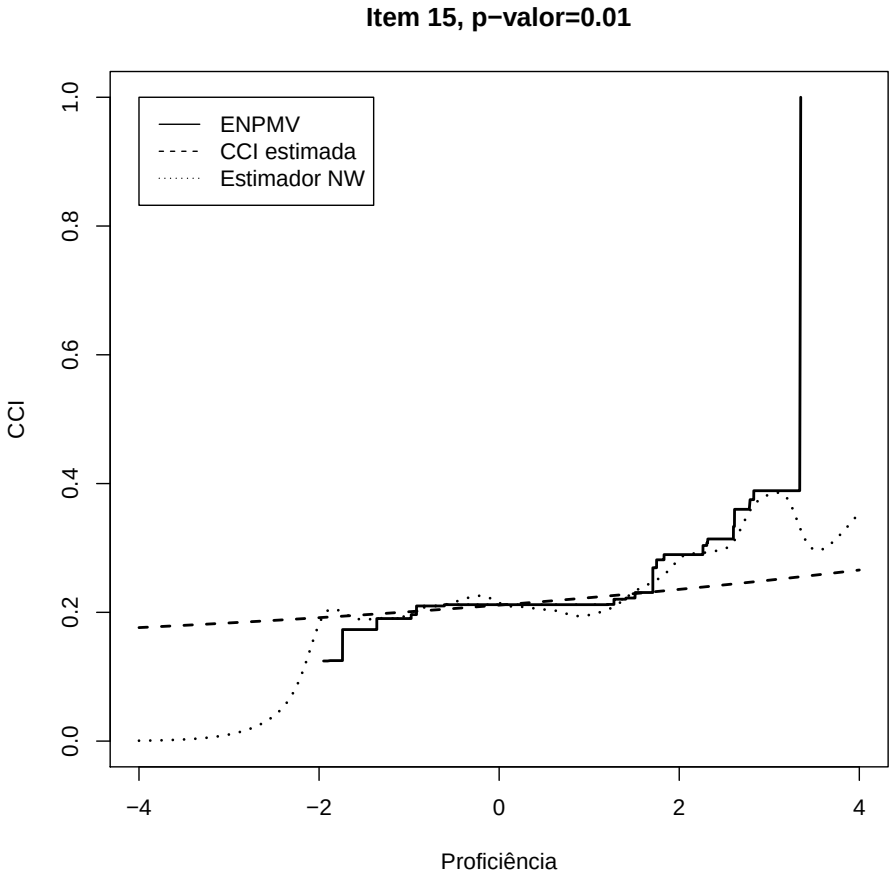
Quando analisamos o item pelo p-valor obtido, temos evidências para rejeitar Ho devido ao valor pequeno de p.

Figura 3 - Item 8



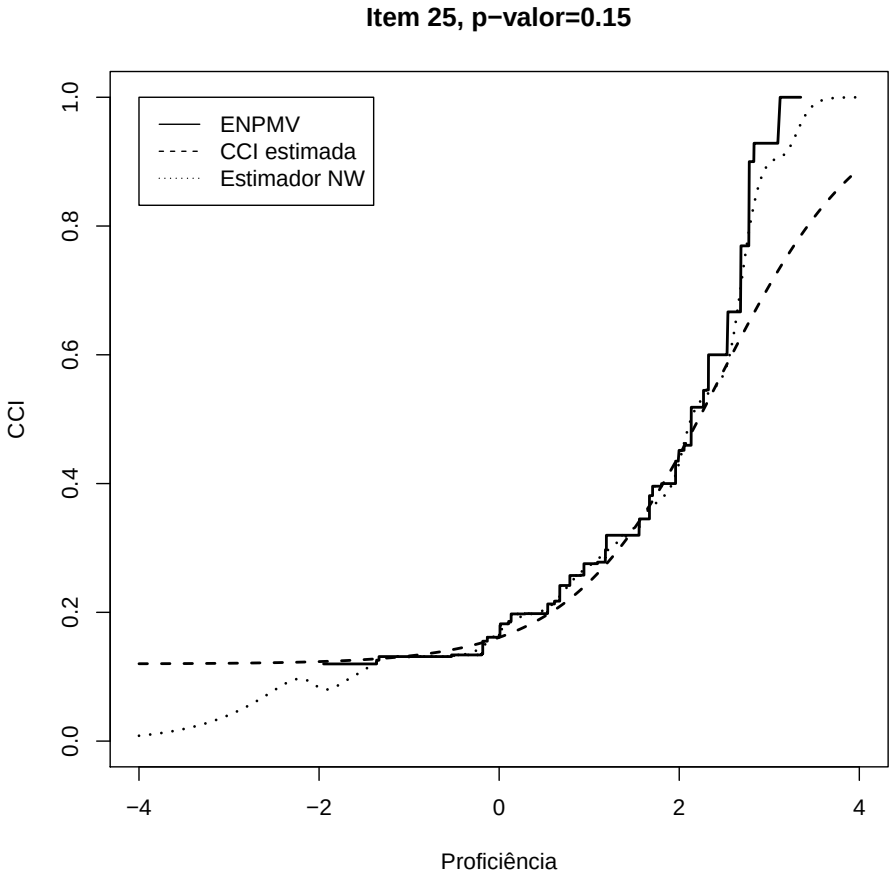
Análogo ao item 1, o item 8 foi muito bem ajustado pelos modelos não paramétricos. Podemos observar que a CCI paramétrica ajustada, o estimador não paramétrico via regressão isotônica e o Estimador NW, convergiram muito bem durante toda a curva do item.

Figura 4 - Item 15



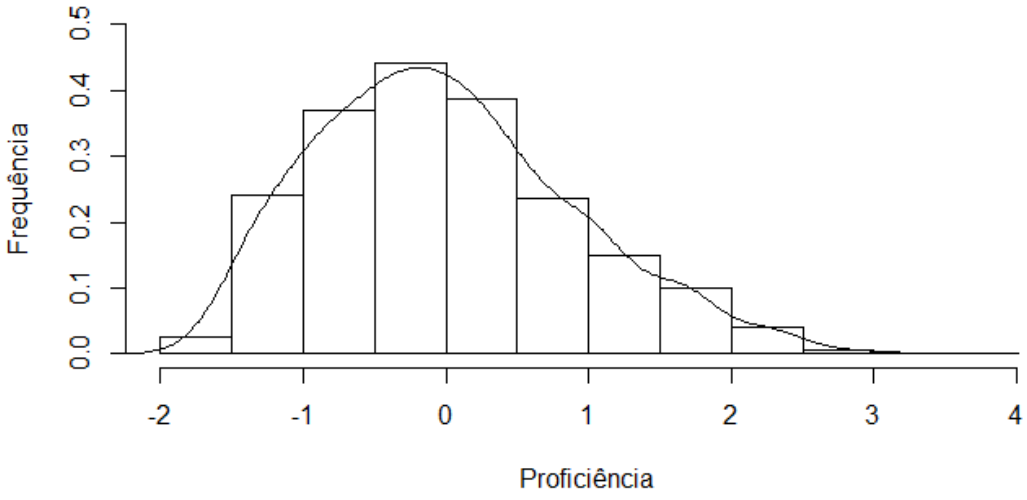
Dentre todos os itens, o 15 foi o pior ajustado. Sua Curva característica foi bem divergente das regressões não paramétricas propostas. Com isso, dizemos que ele não segue um modelo paramétrico de três parâmetros, além de apresentar evidências para afirmar que o item não discrimina bem certa proficiência.

Figura5 - Item 25



O item 25 é um item que podemos considerar como bem ajustado, mesmo com certa divergências nas caldas ele apresentou um p-valor mais razoável dizer que ele é bem ajustado.

Figura6 - Histograma



O Histograma apresentado acima representa a distribuição da proficiência dos candidatos analisados. Podemos ver que ela tende a convergir para uma distribuição normal, porém ainda está levemente assimétrica.

Tabela 5.1: Estimativas dos Parâmetros do Modelo

Item	Parâmetro a	Parâmetro b	Parâmetro c
1	0.1273942	2.197037	3.312840
2	0.1267806	1.605712	3.553575
3	0.1686461	1.387556	2.689681
4	0.0027413	-0.63446	1.549812
5	0.1046385	-0.36126	1.828314
6	0.2427418	2.006995	6.106446
7	0.1080437	1.751496	2.859930
8	0.1590301	1.770924	3.009716
9	0.1307971	1.709215	4.044169
10	0.1590186	1.925541	3.269709
11	0.2163188	0.706296	2.698458
12	0.1793916	0.927305	2.687393
13	0.1617907	0.529403	2.547716
14	0.0124744	-0.00729	1.243684
15	0.1244386	15.79861	0.139683
16	0.0626385	1.564243	2.558959
17	0.2218415	0.896457	1.491702
18	0.1000974	1.478650	2.884070
19	0.0796380	0.834645	0.869077
20	0.1804636	-19.8691	-0.12418
21	0.0924330	2.087949	3.781930
22	0.1860697	2.015931	2.251577
23	0.1814498	2.124924	2.997622
24	0.1866184	1.157073	2.241977
25	0.1198504	2.436113	1.237068
26	0.1855790	4 17.874	0.088648
27	0.1125866	2.027601	3.373979
28	0.2538362	1.965788	2.529861
29	0.2094698	2.223778	5.824056
30	0.1496779	1.939168	2.905057
31	0.1537091	2.676327	2.015642
32	0.0023441	-5.67440	-0.34719
33	0.0000021	1.255972	0.676559
34	0.1471792	0.904357	2.525654
35	0.1118618	-0.03254	1.514138
36	0.0149938	48.59241	0.034899
37	0.2516324	2.618139	2.378580
38	0.1497524	-4.11611	-1.04451
39	0.2115647	2.234723	2.099998
40	0.2334715	2.723088	1.089743
41	0.0987986	-4.02601	-0.99588
42	0.1445644	2.308858	2.243000
43	0.2115587	1.749425	1.908124
44	0.1278933	2.040018	1.650284
45	0.1139626	2.667609	1.431960

Tabela 5.2: Distâncias Entre o Estimador Não Paramétrico da CCI e a CCI Paramétrica Ajustada

Item	Distâcia a	P-Valor
1	0.01581417	0.57
2	0.02670183	0.84
3	0.04486755	0.71
4	0.04485509	0.03
5	0.02856195	0.07
6	0.01691219	0.94
7	0.03040202	0.88
8	0.02703473	0.75
9	0.01984392	0.55
10	0.02168906	0.49
11	0.05628288	0.09
12	0.05574136	0.29
13	0.05960318	0.12
14	0.05278328	0.03
15	0.06846542	0.01
16	0.04415187	0.78
17	0.07283459	0.14
18	0.04013441	0.66
19	0.08705556	0.04
20	0.05594105	0.02
21	0.01373866	0.53
22	0.03135801	0.32
23	0.01924837	0.61
24	0.05972910	0.19
25	0.05166424	0.15
26	0.09145263	0.00
27	0.01806802	0.92
28	0.03020515	0.35
29	0.01086162	0.71
30	0.02408425	0.44
31	0.01894895	0.09
32	0.09089769	0.02
33	0.10618602	0.01
34	0.05990915	0.82
35	0.04805323	0.07
36	0.09935916	0.03
37	0.01694266	0.22
38	0.02213202	0.03
39	0.02734944	0.48
40	0.04982887	0.11
41	0.02661942	0.05
42	0.02319235	0.74
43	0.04692865	0.57
44	0.04700640	0.21
45	0.03540894	0.07

Capítulo 6

Conclusão

Os testes de adequabilidade de ajuste para modelos de teorias de resposta ao item corresponderam a expectativa. Foi possível evidenciar nos resultados que, há questões onde a Curva Característica do Item é perfeitamente acompanhada pelos modelos não paramétricos. Porém em poucos casos, os testes de adequabilidade não retornam o valor esperado para o item em análise.

Dentre os 45 itens analisados, 37 apresentaram adequabilidade ao modelo logístico de três parâmetros. Ou seja, quando estimados por modelos não paramétricos seguem o modelo paramétrico. 8 Itens não ficaram bem ajustados, mas não quer dizer que necessariamente o modelo não paramétrico é ruim, pode ser que o item não discrimine bem, como ocorreu no item 15.

Com esse trabalho, podemos concluir que os modelos não paramétricos são boas alternativas aos modelos paramétricos. Por serem computacionalmente mais acessíveis os estudos nessa linha devem crescer ainda mais.

Referências Bibliográficas

[1] Andrade, J. M. d., Laros, J. A., & Gouveia, V. V. (2010). O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. *Avaliação Psicológica*, 9(3):421–435.

[2] de Andrade, D. F., Tavares, H. R., & da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*.

[3] de Araujo, E. A. C., de Andrade, D. F., & Bortolotti, S. L. V. (2009). Teoria da resposta ao item. *Revista da Escola de Enfermagem da USP*, 43(spe):1000–1008.

[4] de Castro, M. H. G. & Tiezzi, S. (2004). A reforma do ensino médio e a implantação do enem no brasil1. *Desafios*, 65(11):46–115.

[5] Douglas, J. & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25(3):234–243.

[6] Fini, M. E. (2005). Erros e acertos na elaboração de itens para a prova do enem. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Exame Nacional do Ensino Médio (ENEM): fundamentação teórico-metodológica*, pages 101–106.

[7] Glas, C. A. & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27(2):87–106.

[8] Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, 78(3):417–440.

[9] Lopes, A. C. & López, S. B. (2010). A performatividade nas políticas de currículo: o caso do enem. *Educ. rev*, 26(1):89–110.

[10] Maydeu-Olivares, A. & Montaño, R. (2013). How should we assess the fit of rasch-type models? approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, 78(1):116–133.

[11] MILDNER, T. & DA SILVA, A. (2002). O enem como forma alternativa ou complementar aos concursos vestibulares no caso das àreas de conhecimento lingua portuguesa e literatura: Relevante ou passível de refutação? *Avaliação*, 7(2):49–79.

[12] Neto, J. J. S., de Jesus, G. R., Karino, C. A., & de Andrade, D. F. (2013). Uma escala para medir a infraestrutura escolar. *Estudos em Avaliação Educacional*, 24(54):78–99.

[13] Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4):611–630.

[14] Sinharay, S. (2003). Bayesian item fit analysis for dichotomous item response theory models. *ETS Research Report Series*, 2003(2):i–47.

[15] Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, 42(4):375–394.

[16] Soares, T. M. (2005). Utilização da teoria da resposta ao item na produção de indicadores sócio-econômicos. *Pesquisa Operacional*, 25(1):83–112.

[17] Sueiro, M. J. & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educational and Psychological Measurement*, page 0013164410393238.

[18] Wells, C. S. & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education*, 21(1):22–40.