Lesson 5: Regression models for item description

Patrícia Martinková

Department of Statistical Modelling
Institute of Computer Science, Czech Academy of Sciences

Institute for Research and Development of Education
Faculty of Education, Charles University, Prague

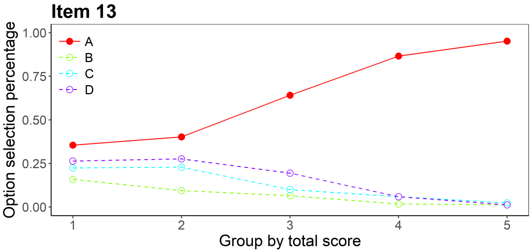
NMST570, October 30, 2018

Table of contents

Traditional item analysis

Traditional item analysis describes item properties by

- Percentages of correct response
- Proportions of those who selected given distractor
- Differences of percentages for groups by total score
- Correlations of item score with total score

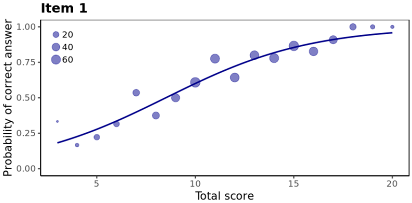


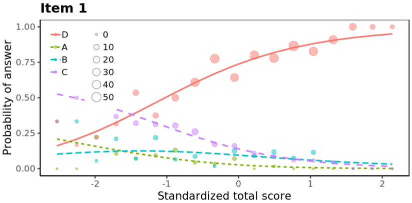
Traditional item analysis

- Difficulty
    - Ratio of correct answers
    - Average item score
    - Scaled average item score
- Discrimination
    - Upper-lower index (ULI)
    - Generalized ULI
    - Correlation Item – Test (RIT)
    - Correlation Item – Rest (RIR)
    - Cronbach's alpha without item
- Distractor analysis
- Analysis of non-reached items

Towards regression models for item description

Models describing mean item score or probability of correct answer with respect to total (or standardized total) score

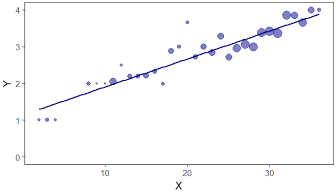




- Better description of item functioning
- Using few parameters per item only
- Possibility to test differences in item score
  - For different groups of respondents (gender, ethnicity)
  - For different types of items
- Possibility to account for specific data features
  - Hierarchical structure, etc.

Regression analysis

Statistical procedures for estimating the relationships among variables.



Simple linear regression:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

- Dependent variable / response Y
- Independent/explanatory variables X (predictors/covariates/regressors)
- Unknown parameters β (intercept β_0)
- Random error ϵ

Regression analysis

Interpretation:

- β_0 is value of Y when $X = 0$ (intercept)
- β_1 describes change of Y with one-unit increase of X (slope)

Estimation procedure:

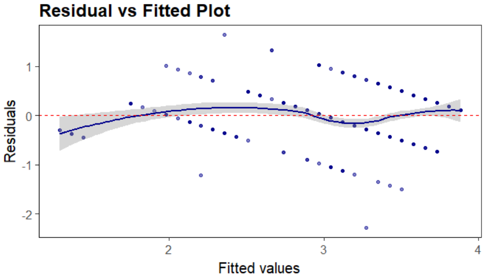
- Predicted value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Residuals $e_i = Y_i - \hat{Y}_i$
- Ordinary least squares estimation: Minimizes $RSS = \sum_{i=1}^{n} e_i^2$
- Model fit: R^2, F test of the overall fit, t tests
- Model selection:
    - Likelihood ratio test (LRT)
    - AIC, BIC (Akaike/Bayesian information criterion)

Assumptions of linear regression

- Representativeness of the sample
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The errors are uncorrelated
- The independent variables are measured with no error
- The predictors are linearly independent
- The error variance is constant across observations (homoscedasticity)

Task 1: How would you check these asumptions are fulfilled?
Task 2: Provide examples of cases when these assumptions don't hold.
Task 3: Which extensions may be applied in such cases?

Checking assumptions of linear regression



To learn more, see courses:

- NMSA407: Linear regression
- NMST432: Advanced regression models

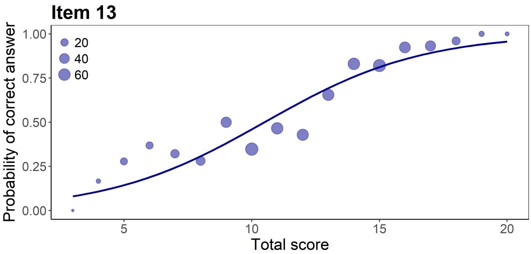
Logistic regression

For each item j, item properties are described by parameters b_{0j} and b_{1j} of logistic function

$$\pi_{ij} = P(Y_{ij} = 1 | X_i, b_{0j}, b_{1j}) = \frac{\exp(b_{0j} + b_{1j}X_i)}{1 + \exp(b_{0j} + b_{1j}X_i)}$$

Also can be written as:

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = b_{0j} + b_{1j}X_i$$



Logistic regression

Terminology/Interpretation:

- logit = log-odds = logarithm of the odds $\pi/(1-\pi)$ of answering the item correctly vs. uncorrectly
- b_0 is value of log-odds when $X = 0$ (intercept)
- b_1 is change of log-odds associated with one-unit increase of X (slope)

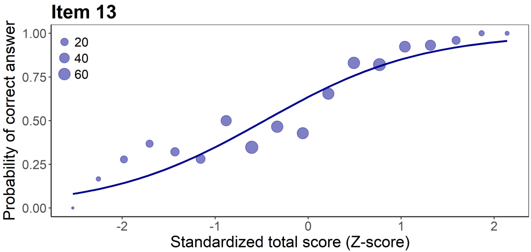
Notes:

- Logistic model belongs to larger class of Generalized linear models (GLM)
- In GLM, linear model is related to response variable via a link function
- Link functions: logit, probit (inverse of the cumulative distribution function), etc.

Logistic regression on Z-scores

Logistic regression on Z-score

$$\pi_{ij} = P(Y_{ij} = 1 | Z_i, b_{0j}, b_{1j}) = \frac{\exp(b_{0j} + b_{1j} Z_i)}{1 + \exp(b_{0j} + b_{1j} Z_i)}$$



Interpretation:

- b_0 is value of log-odds for average respondent $Z = 0$
- b_1 is change of log-odds associated with one-unit increase of Z, i.e., with 1 SD increase of X

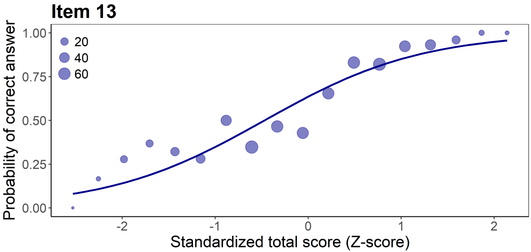
Logistic regression on Z-scores, IRT parametrization

Logistic regression on Z-score, with IRT parametrization

$$\pi_{ij} = P(Y_{ij} = 1 | Z_i, a_j, b_j) = \frac{\exp[a_j(Z_i - b_j)]}{1 + \exp[a_j(Z_i - b_j)]}$$

Also can be written as:

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = a_j(Z_i - b_j)$$



Logistic regression - IRT parametrization

Interpretation:

- Z_i standardized total score of person i (Z-score)
- b_j difficulty of item j, location of inflexion point, Z_i such that $P(Y_{ij} = 1|Z_i) = 0.5$
- a_j discrimination of item j, slope at $Z_i = b_j$, change of log-odds associated with one-unit increase of Z_i

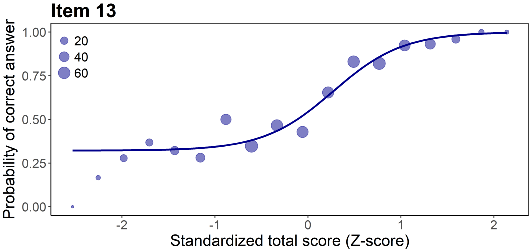
Nonlinear regression

$$\pi_{ij} = P(Y_{ij} = 1 | Z_i, a_j, b_j, c_i) = c_i + (1 - c_i)\frac{\exp[a_j(Z_i - b_j)]}{1 + \exp[a_j(Z_i - b_j)]}$$

b_j difficulty of item j

a_j discrimination of item j

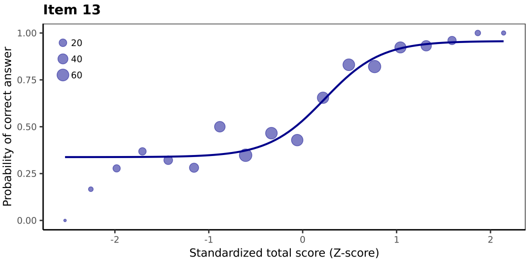
c_j probability of guessing of item j (lower asymptote)



Nonlinear regression

$$\pi_{ij} = P(Y_{ij} = 1|Z_i, a_j, b_j, c_i, d_i) = c_i + (d_i - c_i)\frac{\exp[a_j(Z_i - b_j)]}{1 + \exp[a_j(Z_i - b_j)]}$$

b_j difficulty of item j
a_j discrimination of item j
c_j probability of guessing of item j
d_j probability of innatention on item j (upper asymptote)

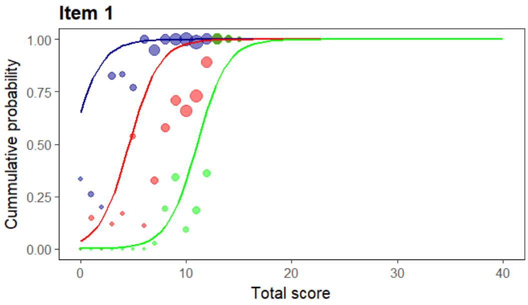


Ordinal regression - cumulative models

Cumulative probabilities are modelled using logistic regression:

$$\pi^*_{ijk} = P(Y_{ij} \geq k | Z_i, a_j, b_{jk}) = \frac{exp[a_j(Z_i - b_{jk})]}{1 + exp[a_j(Z_i - b_{jk})]}$$

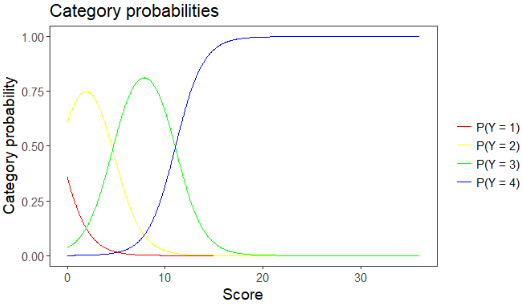
b_{jk} locations of inflection points of cumulative functions



Ordinal regression - cumulative models

Response category probabilities are given by difference of cumulative probabilities:

$$\pi_{ijk} = P(Y_{ij} = k | X_i, a_j, b_{jk}) = \pi^*_{ijk} - \pi^*_{ij(k+1)}$$



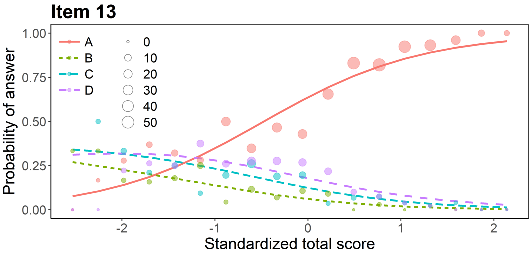
Multinomial Regression - divide-by-total models

Modells log odds of choosing distractor vs. correct answer (baseline)

$$\log \frac{\pi_{ijk}}{\pi_{ij0}} = a_{jk}(Z_i - b_{jk})$$

Response category probabilities are defined as the ratio between category-related functions and their sum:

$$\pi_{ijk} = P(Y_{ij} = k | Z_i, a_{j0}, \ldots, a_{jK_j}, b_{j0}, \ldots, b_{jK_k}) = \frac{\exp(a_{jk}(Z_i - b_{jk}))}{\sum_{r=0}^{K_j} \exp(a_{jr}(Z_i - b_{jr}))}$$



Conclusion

Regression models provide more flexible approach to item description than traditional item analysis

- Description of item functioning across whole ability (total scores) distribution using few parameters per item only
- Possibility to test differences in mean item score for given total score
  - In different groups of respondents (gender, ethnicity)
  - In different types of items
- Possibility to account for specific data features
  - Hierarchical structure
  - Correlations between some items, etc.

Thank you for your attention!

www.cs.cas.cz/martinkova

Vocabulary

- Regression analysis
  - Dependent variable / response Y
  - Independent variables / predictors X
  - Unknown parameters β
  - Residuals $e_i = Y_i - \hat{Y}_i$
  - Ordinary least squares: minimizes RSS
  - Model fit R^2
  - Model selection: LRT, AIC, BIC
- Regression models
  - Linear regression
  - Generalized linear regression
    - Link function: logit, probit,...
  - Nonlinear regression
  - Ordinal regression
  - Multinomial regression