

Lab 2 - Multilingualism and Cognitive Performance

Oviya Adhan, Helin Yilmaz, Nory Arroyo

2025-04-17

Introduction

Past research on multilingual individuals has shown mixed results on the benefits multilingualism has on cognitive ability. As globalization continues to blend cultures across borders around the world, the rate of multilingualism increases ¹. By understanding the concurring traits, we can get some insight into the characteristics that we can expect to become more prevalent along with multilingualism, particularly in regards to cognitive ability. To explore the question “What is the relationship between multilingualism and cognitive ability?” we sourced the “Multilingualism and Cognitive Performance” dataset from <https://dataverse.nl/>, a research data repository.

Data

This data was collected from adults over the age of 65 from three provinces in the northern Netherlands, a reportedly highly multilingual location. With a focused opportunity to investigate multilingualism, we can uncover relationships that can inform the approach to implementing multilingual education earlier in life. For each participant, the data includes results from two cognitive tests and a questionnaire covering aspects of their language knowledge, wellbeing, health, and personal traits. The results of the first cognitive test noted the number of errors made in a Wisconsin Card Sorting (cognitive flexibility) test, while the second tested Flanker Task (attention and inhibitory control).

To measure cognitive ability, we focused on the errors made in the Wisconsin Card Sorting (WCST) which tests for the test taker’s cognitive flexibility. They are asked to sort cards based on certain characteristics, such as color, shape, or number. Throughout the duration of the test, the rules of the task changes and scoring is based on how many errors are made following rule changes ². This variable is numeric, discrete, and ranges from 4 to 29. To measure multilingualism, we focused on the variable that represented the number of languages known by the participant. This variable is numeric, discrete, and ranges from 1 to 5. With this localized dataset, controlling for geographic location and stage of life, we set out to explore the relationship between multilingualism and cognitive ability.

Step 1 - Create Regression Models

First, we wanted to explore the relationship between multilingualism and cognitive ability without input from other variables in the dataset. We chose the number of languages variable from the questionnaire to represent multilingualism and number of errors made on the cognitive flexibility test to represent cognitive ability. Below is the summary of this univariate regression model:

X: Number of Languages

Y: Errors made in Cognitive Flexibility Test (WCST)

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	13.8815199	1.393125	9.9643050	6.547620e-20
##	number_lang	-0.2466321	0.319438	-0.7720814	4.407955e-01

From the summary we can see that the first, univariate regression model shows a minimal inverse relationship between number of languages and number of errors on the cognitive flexibility test suggesting that an increase of number of languages is linked to a decrease in errors. The coefficient for number of languages is -0.2466, indicating that for each additional language known, there is an average decrease of 0.24 errors made in the cognitive flexibility test. Being a small, statistically insignificant coefficient estimate at a p-value of 0.441, this alone is not enough to show a strong relationship between multilingualism and cognitive flexibility.

From plotting the number of languages against the errors made in the cognitive flexibility test (Appendix - Table 1), we can see a slightly negative relationship, confirming our statistical findings.

There are likely other variables that might impact this relationship. To explore this possibility, we revised our model to include other x variables covering standard socioeconomic factors including gender, age, education, and income to hopefully add more color to our analysis. Summary of the revised, multivariate model:

X: Number of Languages, Gender, Age, Education, Income

Y: Errors made in Cognitive Flexibility Test

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	9.72151991	4.71845603	2.0603180	0.0404199669
##	number_lang	-0.06182842	0.31530408	-0.1960914	0.8447005200
##	gender	-1.04944585	0.62148690	-1.6886049	0.0925623126
##	age	0.15527779	0.05618491	2.7636921	0.0061473865
##	education	-1.05115907	0.30150446	-3.4863798	0.0005796855
##	income	-0.14816778	0.24135806	-0.6138920	0.5398539658

We can see that the coefficient estimate of the number of languages in the multivariate regression model shows a less strong relationship (a 0.06 decrease in errors for every additional language) with cognitive flexibility test errors compared to the univariate regression model while maintaining a statistically non significant coefficient estimate. The two other variables that stand out are age and education which show statistically significant coefficient estimates. With all else constant, for every additional year of age, there is a 0.15 increase in errors made while with every additional level of highest completed education in Dutch, there is a 1.05 decrease in the number of errors made on the WCST. As for the remaining two, gender and income level, they do not have statistically significant coefficients. For gender, as gender increases by 1, the number of errors made decreases by 1.05. With gender being noted as 1 for male, 2 for female, this shows better cognitive flexibility scores for women on average, although practically insignificant. For income, as income level by month increases by 1 level, the WCST errors decrease by 0.14, suggesting practical insignificance.

With the highly diverse and interesting outcome, we decided to focus on our second model to further explore the relationship between these variables.

Step 2 - Evaluate Assumptions of Model 2

Evaluate the IID Assumption

Given that Model 2 has multiple predictors for errors made in the cognitive flexibility test, we can visually inspect if there are any patterns like clustering using various scatter plots. Visually, as observed in the

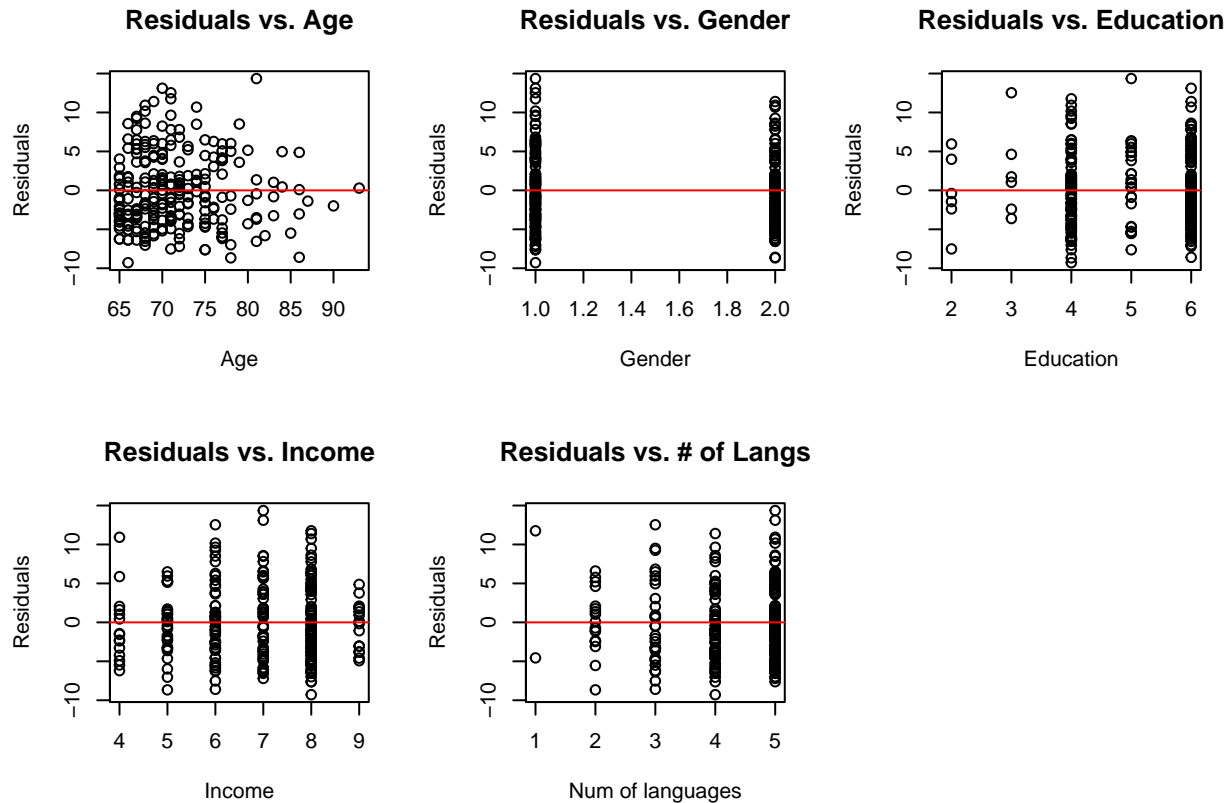
scatter plots (Appendix - Table 2) of the predictors versus the cognitive flexibility errors, there are no strong visual clusters of data suggesting that IID is not violated.

Looking into the data collection method, the data were collected from a sample of 387 65-95 year old adults, recruited through open calls in local media (e.g. radio, newspapers), local fliers, and by word-of-mouth through the personal networks of the research team. This method of convenience sampling, especially through the researchers' personal networks, may be a limitation to the generalizability of our findings due to the highly localized scope. Common regional influences like education, healthcare, and sociocultural environment could have some unknown effect on the general distribution of the observations. These limitations suggest that findings should be interpreted carefully, especially when considering broader applicability beyond the studied population. However, considering that the participants were each recruited and tested individually, therefore having no direct influence on one another, we can comfortably assume that the resulting data of each individual is independent and identically distributed. ³

Evaluate the Linear Conditional Expectation Assumption

To evaluate the linear conditional expectation assumption, we can plot the residuals vs. fitted values for a multivariate regression.

The residuals vs fitted values plot (Appendix - Table 3) shows a fairly random scatter, suggesting that the Linear Conditional Expectation (LCE) assumption is reasonably met.



Based on the plots of residuals versus fitted values and residuals versus each predictor variable, the Linear Conditional Expectation (LCE) assumption appears to not be strongly violated. There are no strong, visually apparent non-linear patterns, although it is important to note that variance somewhat decreases at higher ages. Also, it's more difficult to discern a trend for Gender due to the binary nature of the variable. Thus, given the complexity of the model and how these factors interact, it's possible that some effects are

nonlinear or interact in ways not captured by a basic linear model. While residual plots and prediction diagnostics help flag major issues, they don't guarantee that the model fully reflects the data structure. So, although linear regression is a useful starting point, it's important to be cautious when interpreting coefficients as reflecting simple, linear relationships.

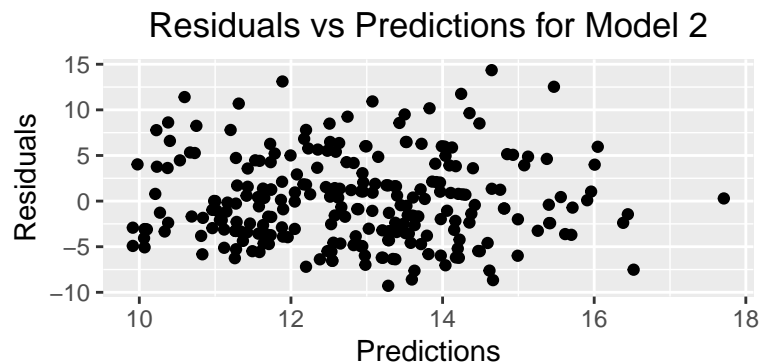
Evaluate the No Perfect Collinearity Assumption

As observed when running our regression models, all of our variables are still included. If we had perfect collinearity between our input variables, the model would automatically drop one of our x variables. For near perfect collinearity, we can compare the relationship between standard error and coefficient estimates, notably the number of languages and income have a higher standard error than the absolute value of their coefficient estimates. Additionally, their coefficients are not significant, which can lead us to believe that they may have a relationship with the other factors that have significant coefficients.

```
## number_lang      gender      age      education      income
##      1.058058      1.062037      1.004101      1.167871      1.181965
```

To further detect multicollinearity, we can take a look at the variance inflation factor (VIF) of each of our x terms. The VIF of each term in our model ranges from 1.00 to 1.18 portraying a low correlation between the factors, but practically insignificant, being on the lower end of 1 to 5, to accept as a violation of near perfect collinearity for any of these terms.

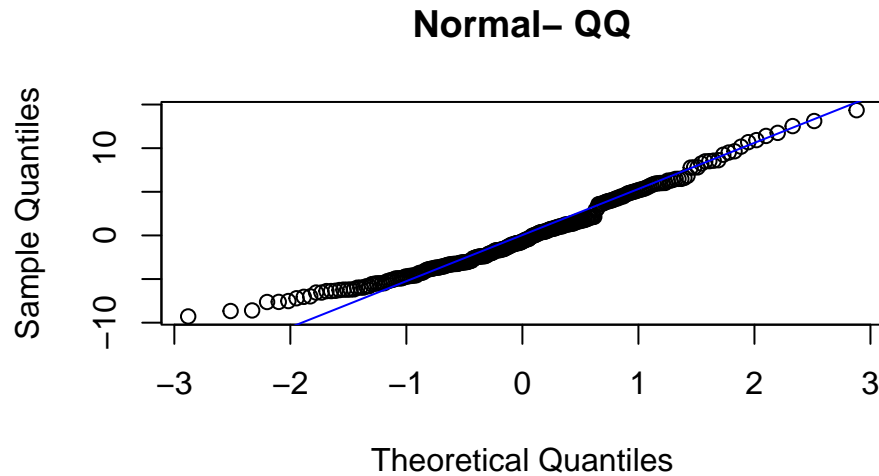
Evaluate the Homoskedastic Conditional Variance (Constant Conditional Variance) Assumption



```
## Analysis of Variance Table
##
## Model 1: model_1_residuals2 ~ 1
## Model 2: model_1_residuals2 ~ 1 + number_lang + gender + age + education +
##      income
##   Res.Df    RSS Df Sum of Sq   F Pr(>F)
## 1      251 225445
## 2      246 223222   5    223.1 0.49 0.7836
```

For our ocular test we can observe that the distribution between residuals and predictions is randomly spread out with no distinct clustering or patterns. To statistically check for heteroskedasticity we then performed a Breusch-Pagan test, by hand. We found that there is no evidence to reject the null hypothesis of homoscedasticity. Therefore, we can likely validate this assumption.

Evaluate the Normally Distributed Errors Assumption



As seen in our QQ-plot, the theoretical quantiles and sample quantiles follow a linear pattern that falls approximately across the line. Therefore, we can likely validate the normality assumption.

Conclusion

These models aim to investigate the relationship between multilingualism and cognitive performance. The first model, depicting the relationship between number of languages and errors made on WCST, showed a statistically and practically insignificant relationship. The second model, depicting the relationship between number of languages, age, income, gender, education, and the number of WCST errors, showed a statistically and practically insignificant relationship between gender, income, and number of languages with errors made on WCST. On the other hand, age and level of education, showed statistically significant relationships with WCST errors. However with respective coefficients of 0.15528 and -1.05116, it appears that only level of education has a practically significant impact on cognitive flexibility.

For future studies, looking further into these characteristics and how they impact cognitive performance across varied age ranges and locations can offer a diverse perspective into this unique relationship.

Appendix

<https://github.com/mids-w203/lab-2-girl-power/>

References

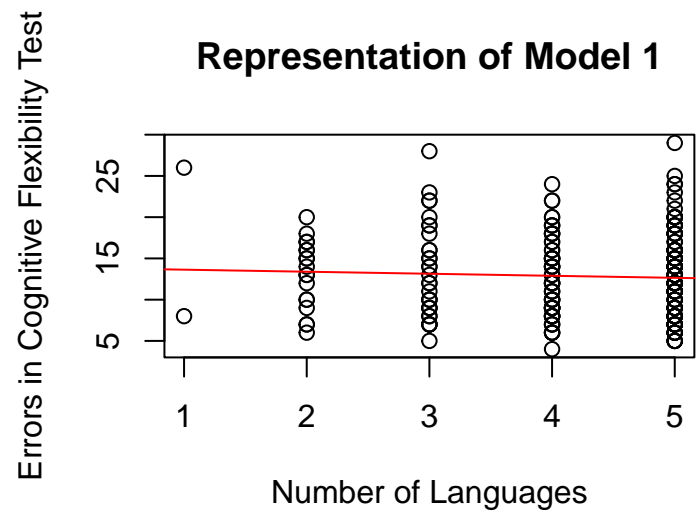
1 - Housman, P. (2023, November 15). Multilingualism on rise in US: Illusion or reality? <https://www.american.edu/cas/news/multilingualism-on-rise-in-us-illusion-or-reality.cfm>

2 - Grant, D. A., & Berg, E. A. (1948). Wisconsin Card Sorting Test. Psychological Assessment Resources. <https://www.parinc.com/products/WCST>

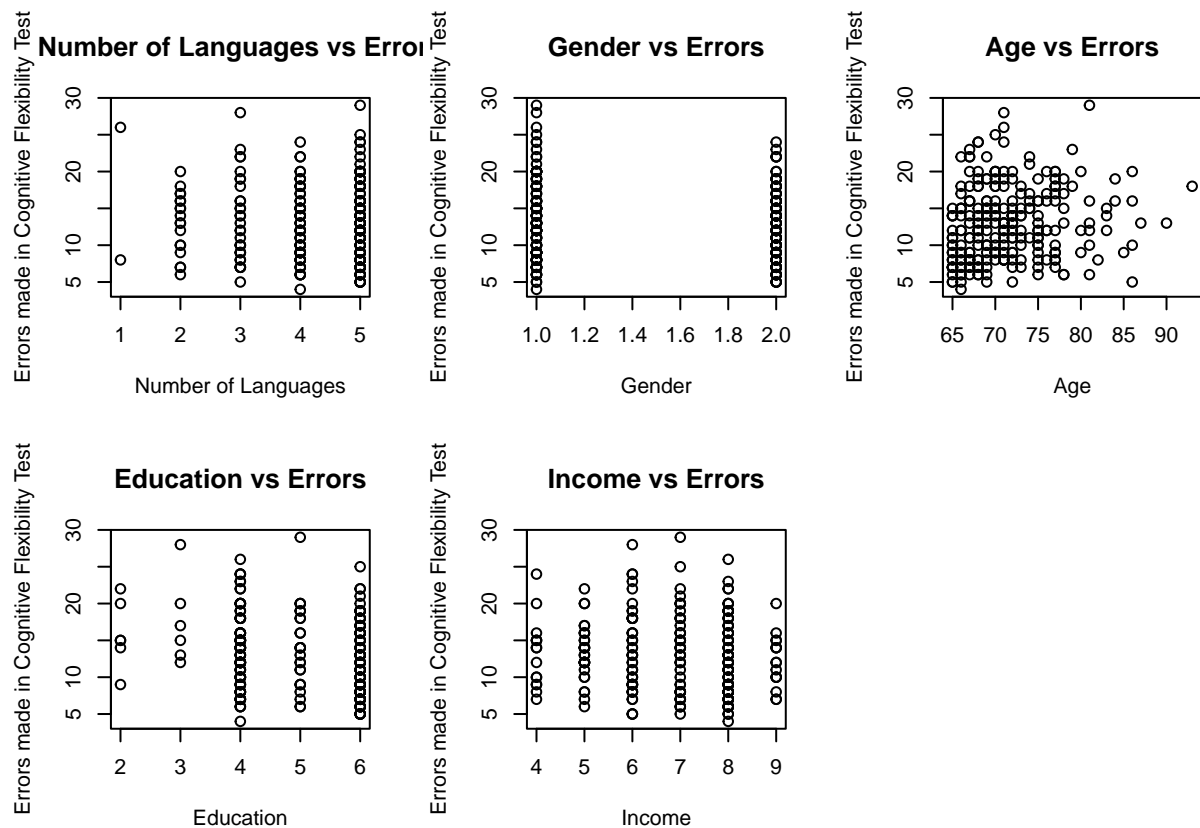
3 - Pot, A., Keijzer, M., & de Bot, K. (2018). Intensity of Multilingual Language Use Predicts Cognitive Performance in Some Multilingual Older Adults. Brain sciences, 8(5), 92. <https://doi.org/10.3390/brainsci8050092>

Tables

[Table 1] Model 1 Scatterplot



[Table 2] IID Scatterplots



[Table 3] Residual v. Fitted

resid_v_fitted

Assessing LCE of Model 2: Residuals vs Fitted V_k

