Lareisha Lewis

Statistical Model SPR 2022

March 31, 2022

Objective:

In project 1 I will explore the correlation between being bilingual/monolingual and the impact it has on ones IQ. After determining the correlation between the two variables, I will look at other variables such as age, sex, and education of participates to explore their impact.

Hypothesis: Individuals who are bilingual generally have a higher IQ than monolingual speakers.

**Data source:**

<https://researchdata.reading.ac.uk/207/>

**EDA:**

In this analysis I will explore a dataset I retrieved from researchdata.reading.ac.uk. The dataset has 114 variables, however for the purpose of the objective I will only explore 8 of them. I chose to explore this dataset because it contains two key variables that will help test my hypothesis, IQ, and language status. This dataset contains 25 Bengali-English speaking participants, and 25 English-speaking participants. Each participant has an IQ value assigned to them based on their accuracy of the Raven’s standard IQ test. The Raven’s IQ test has 60 test items. As I explore the data, I will be testing different models to see which one is the best fit. I will examine other variables to see the effect they have on IQ, such as sex, age, education, response time of Stroop trial, accuracy of color-shape switch task, and backwards digit span task.

I took the approach of eliminating insignificant variables by testing the correlation between IQ and each of the independent variables one by one. In doing this I was able to determine if any of the other variables have a greater impact on IQ than language status. The four plots below represent the worse possible models. I started with the variables I believed to have the least correlation with the response variable. Starting with IQ and education I produced a simple linear regression model and correlation test. The education variable represents the number of years a participant has been in school. It was my assumption that the amount of education one has received would have some impact on IQ. Surprisingly the two variables have almost no correlation at 0.08. The model produced an overall p-value of 0.58 and r-squared value of 0.00633. Model A below visualizes the weak correlation produced as an outcome of this model. As you may notice in the plot none of the points fit around the red regression line, and the line is almost completely horizontal. This indicates the points do not follow a specific trend, and a poor correlation between the two variables.

Next, I created a model for IQ and Age. The age variable states the age of each participant. Participants’ ages range from 22 to 45. Many may assume age is directly related to IQ, and in some cases that may be true. For example, if the age range was larger (ages 14 to 60) we may be able to see a greater correlation between the variables. However, in this dataset age has a 0.07 correlation with IQ. The linear regression model produced an overall p-value of 0.63. This p-value is greater than the significance level of 0.05(0.63 > 0.05), therefore the variable is not considered significant. Model B displays the weak relationship between IQ and age. The points are widely spread out, and do not follow any trend. Predictions would be inaccurate with this model.

After exploring education and age, I moved on to variables that test participants ability to complete assigned task. The stroop variable represents the response time to the stroop trial. Participants were instructed to look at a list with colors spelled out in color ink. Ignoring the fact that some words do not match the color they are presented in, the individual will state the color of the words as quickly as possible (ex. Red printed in red ink, or red printed in blue ink). To get the incongruent response time researchers look at the response time to words that do not match the color. Higher values indicate a longer response time in seconds. As I explored the output of the model summary, I concluded that Stroop vs. IQ is also a poor model. It produced a p-value of 0.85 and r-squared value of 0.0007723. Model C is a clear depiction of the poor correlation of -0.03 between the two variables. Based on the p-value and other factors this model is even worse than model B. Its coefficients also do not meet the standards for this model to be considered significant.

Attempting to eliminate all weakly correlated variables, I move on to explore IQ vs. Color-shape task accuracy. This is another variable that requires participants to complete a task, and scores them based on their accuracy. The color-shape task requires participants to look at colored shapes one at a time. The computer will display the word “shape”, or “color” to indicate which one they want you to focus on. If “shape” is indicated one must ignore the color and state, the shape. If “color” is indicated one must ignore the shape and state, the color. I chose this variable because I was curious to see if individuals with higher IQ’s held a greater accuracy rate than those with lower IQ’s. The model gave a p-value output of 0.63 and r-squared value of 0.005016. This model displays the poor correlation between IQ and Color-shape task accuracy of -0.07. Although the correlation is poor between the two variables the plot shows that of the individuals that scored above the average of 93-accuracy rate, 84% are bilingual, and 48% are monolingual. This finding leads me to believe that bilingual speakers are more likely to be able to effectively shift their attention between tasks than monolingual speakers.

|  |  |
| --- | --- |
| Chart  Description automatically generated  **Model A** : IQ vs. Education | Chart, scatter chart  Description automatically generated  **Model B**: IQ vs. Age |
| Chart, scatter chart  Description automatically generated  **Model C**: IQ vs. Stroop (Incongruent) | Chart, scatter chart  Description automatically generated  **Model D**: IQ vs. Color-shape Task Accuracy |

**Comparing Models:**

After closely exploring the data’s correlation between some of the variables, I was able to label some variables as insignificant. I tested the correlation between my dependent variable (IQ) and each of the four independent variables (age, education, Stroop Task, and Switch task). Through running each test, I learned that all four variables have no significant correlation to one’s IQ. In this section I will continue to focus on finding the best model. Now that I have a better understanding of the insignificant variables, I will look at simple and multiple linear regression models. You will see models built between the dependent variable and language status, sex, and backwards digit span. I will also explore how variables interact with each other, and if they are able to improve the model.

I ran a multiple linear regression model using the following code in R: mdl\_all <- lm(IQ ~ Age + Sex + Group.name + Education + Stroop.IC + Switch.trial.acc + Digitspanbackward , data = BiMonlingIQ2). This model produced an adjusted r-squared value of -0.09, all the variables generate p-values greater than 0.05 except for sex, and the overall p-value of the model is 0.87. As expected, this is not a significant model due to the high p-value and low r-squared. I already have a good idea of the variables with poor correlation from the correlation test I previously ran. Moving forward I removed the following variables: Age, Education, Stroop.IC, and Switch.trial.acc. The next model I produced contained IQ as the response variable and sex, language status, and backwards digit span as the explanatory variables. Although “Group.name” and “Digitspanbackward” both showed not to be significant in the previous model, I decided to keep them for the next model. “Group.name” represents the language status of each participant. It has two values, bilingual and monolingual. The simple linear regression model for IQ and “Group.name” produced p-values of nearly zero for each value and a r-squared value of 0.99. I concluded the “Group.name” variable would be a good addition to my model due to the coefficients, and it relates directly to the objective. The “Digitspanbackward” variable consist of values produced by another task the participants had to complete. This task required participants to observe a certain number of numbers being presented to them one by one. After the numbers are finished being displayed the participant must recall the numbers in a backwards order and list them. The value assigned to the variable represents the number of numbers they listed correctly in the correct order.

Following running the new model I reviewed its summary. The summary stats indicate an adjusted r-squared of 0.99 and overall p-value of nearly zero. “Group.name” and “Digitspanbackward” still maintain a p-value greater than 0.05 when ran inside the current model. “Group.name” shows a high p-value when the values monolingual and bilingual are combined, however separately they each hold an acceptable p-value of nearly zero. Due to this fact, I believe “Group.name” is still an effective variable. The plot below shows a visual of the model. The yellow regression line shows a negative correlation for bilingual speakers and positive correlation for monolingual speakers. The actual data does not follow the trend of the regression lines. The minimum digit span was 4 and the maximum was 7. The orange stars represent the IQ predictions for digit spans up to 10. The predictions also fail at following the yellow regression line, adding to the fact that this is a poor model.

Chart, scatter chart

Description automatically generated

I removed the digit span variable and ran the model again with IQ as my dependent variable and sex and group name as my independent variables. This model gave me an overall p-value less than 0.05, and a r-square of 0.99. In addition to the low overall p-value and high r-square, each variable maintained a p-value close to zero. Therefore, I was able to conclude that the model with IQ, “Sex”, and “Group.name” would be the best model for the purpose of this project. The graph below shows a visual of this model. In the graph red points indicate bilingual and black points indicate monolingual. The red line shows the average IQ for bilingual speakers, and the black line shows the average IQ for monolingual speakers. Bilingual speakers have a slightly higher average IQ at 43.5 than monolingual speakers at 43. This visual allows you to see the number of participants that score above average according to their language style group.

Chart, scatter chart

Description automatically generated

**Diagnosis/ Conclusion:**

Lastly, I ran the model containing all the variables through the “step” function to make sure I picked the best model. The function identifies IQ vs Sex as the best possible model. The results are similar to my findings; however, I kept the “Group.name” variable for the purpose of the topic. I was surprised to see “Group.name” was one of the first variables to be dropped in the step function. The AIC dropped from 165 to 163 when the “Group.name” was removed. IQ vs. Sex generated an AIC of 154. My model appears to be a great model based on the coefficients I observed; however, it is not normally distributed. In the Q-Q plot below the points fall perfectly on the line close to the average. The average overall IQ is 43, and most of the values are within the forties, which is why we see most of the points in the middle of the plot close together and on the line. As we move to the right, we notice the points start to deviate away from the normal distribution line. In the upper right corner two points are labeled with their index numbers, and one point in the lower left corner. These points are outliers. The max IQ in this dataset is 54. Only two individuals from the dataset scored 54, which is why they are positioned furthest from 0 on the x-axis. If some of the values with high leverage and high residuals were removed the model would possibly be a better fit.

Chart, line chart, scatter chart

Description automatically generated

Based on the results of the model my hypothesis is accurate. One’s IQ and whether they are monolingual, or bilingual has a significant relationship. In other words, the results we observed did not occur by chance. I was also able to find some interesting facts about the variables. I immediately notice a difference in the deviation of the two language statuses. Monolingual participants have a wider spread. This group holds the title for both lowest and highest IQ scores with a standard deviation of 5.45. Bilingual speakers have a smaller standard deviation of 3.8. While monolingual speakers have reached a higher IQ score, bilingual speakers have a higher percentage of participants above the overall average IQ of 43. 48% of bilingual participants scored above 43, and 36% of monolingual participants scored above 43. I wonder if the language status and IQ variables would have a closer correlation if we explored beyond bilingual into trilingual, or multilingual individuals.

**Problems:**

* Group.name and Group.ID both represent the same thing in the dataset. Bilingual is the Group.name and it has the Group.ID 1. Monolingual is the Group.name and it has the Group.ID 0. Although they both represent language status, I kept get different coefficients when I used each of them in the model. The issue was with Group.ID, as it holds the values o and 1, it is considered a numeric variable, when it should be a factor. I used the as.factor function to convert the values.
* The final model appears to be significant, however the r-squared value of 0.99 is a bit alarming. While the goal is to reach a r-squared value of 1, getting a high r-squared can be the result of a problem somewhere else in the model or dataset.
* The final model is not normally distributed. The points two standard deviations to the right of the average seem to have higher residuals.

Code:

**IQ for bilingual and monolingual plot**

ggplot(BiMonlingIQ2, aes( IQ, Group.ID)) +

geom\_point(aes(col = Sex)) +

geom\_smooth(method = "lm", se=F, col="red")+

geom\_point(data = prediction\_data1, color = "black", shape = 5, size = 2) +

labs(title = "IQ for Bilingual vs. Monolingual ")+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5, size=15, face="bold"))

**IQ vs. Age plot**

ggplot(BiMonlingIQ2, aes(Digitspanbackward, IQ)) +

geom\_point(aes(color = Sex)) +

geom\_smooth(method = "lm", se=F, col="yellow")+

geom\_point(data = prediction\_data, shape = 8, color = "orange")+

facet\_wrap(vars(Group.name)) +

labs(title = "IQ vs Digit Span Backwards ")+

theme\_bw()+

theme(plot.title = element\_text(hjust = 0.5, size=15, face="bold"))

**Models**

* mdl\_sex\_group\_dig <- lm(IQ ~ Digitspanbackward + Sex + Group.name + 0, data = BiMonlingIQ2)

Table

Description automatically generated

* mdl\_sex\_GN <- lm(IQ ~ Sex + Group.name + 0 , data = BiMonlingIQ2)

Table

Description automatically generated

**Step**

step(mdl\_all, direction = "both", data = BiMonlingIQ2, trace = TRUE)

Table

Description automatically generated