How to Make a Proceedings Short Summary Submission

Hassan Nishat (University of Toronto)
1005835350

Fengbai Han (fengbai.han@mail.utoronto.ca)
1004574962

Keywords: Cognitive Science; Age of Acquisition; PCA; Linear Regression; Bayesian Ridge Regression; Automatic Relevancy Determination

Abstract

Word acquisition is a phenomenon that has been studied for many years. There are many theories on how we acquire words and much has been learnt about word acquisition through modelling acquisition with neural networks meant to imitate how the human brain works. This study investigates the age of acquisition, which is the age at which certain words are learnt. Braginsky et. al model the age of acquisition through finding a number of predicting values and applying linear regression. To replicate this study, we find our own matching data and perform our own analysis. First a principal component analysis is conducted to investigate the association between our chosen predictors and age of acquisition. Then three linear modelling techniques are applied to the data.

Introduction

A phenomenon seen in the study of how we learn language, is word acquisition, or how we learn words. There are many theories on how we learn words, especially as we learn them for the first time as infants. Although we cannot know for sure how exactly words are learnt, through implementing neural networks to model various theories, we can get a better understanding of which theories may be more plausible than others. The paper "Computational investigation of early child language acquisition using multimodal neural networks: a review of three models" by Abel Nyamapfene goes over three models that were developed to model early acquisition of words using different techniques. The first model was developed by Plunket et al. in 1992. This network aims to replicate autoassociative language acquisition, which is the idea that children take in perceptual entities, and are able to associate the entity with a single word label. The model takes in images consisting of patterns made up of dots and uses backpropagation to connect the images to their corresponding labels. Although this model does show that words can be associated with visual stimulus through a neural net, Nyamapfene argues that backpropagation is not the best way to replicate word acquisition as children do not receive constant feedback when learning new words Another model proposed by Abidi and Ahmad in 1997 combines two models to create a multi-net model. This model combines the model of words acquired from naming and pointing tasks with a model that maps non-named communicative intentions and their corresponding words. Nyamapfene points out that this network could also be problematic as it implies that the two sets of word acquisition are done through separate neural functions, whereas the current opinion in child language would suggest otherwise. The final network Nyamapfene discusses is one he proposes himself in 2007. This model is a modification of Abidi and Ahmad's model and tries to make it more biologically and psychologically plausible by using a single unsupervised neural network to simulate word acquisition rather than combining two. This model also used counterpopagation to be more biologically plausible (Nyamapfene, 2009).

More specifically, we will investigate modelling the age of acquisition (AoA). The age of acquisition refers to the age at which a child learns a certain word. There are many theories as to why children tend to learn certain words earlier than others, ranging from the formation of the letters within the word to the underlying meaning of the word. One way to better our understanding of why certain words are learnt at different times in development than others, is through the use of modelling. The paper "From uh-oh to tomorrow Predicting age of acquisition for early words across languages" by Braginsky et al. uses linear modelling to predict the age of acquisition of words based on various different predictors, such as frequency, length, concreteness, babiness (a measure of association with infancy), and the mean length of utterance (MLU) of words. Data of these predictors were collected on a group of words and a linear regression created to predict the age of acquisition of words based on the values of the given predictors of each word. Once analysis was done on the resulting regressions, it was found that frequency, babiness, concreteness, and MLU were relatively strong predictors (Braginsky et al. 2016, p 1694). For our study, we plan to replicate these findings using a different set of data and a different set of techniques. Firstly, we conduct a principal component analysis (PCA) in order to ensure a correlation between our selected data and AoA. Then we conduct our own linear regression as well as two other probabilistic methods of modelling AoA and compare our findings to those of Braginsky et. al.

Data

We collect four datasets from four different sources. (Bird, Franklin, & Howard, 2001; Brysbaert, Warriner, & Kuperman, 2013; Perlman, Little, Thompson, & Thompson, 2018; Warriner, Kuperman, & Brysbaert, 2013) Since we have four different datasets, we need to combine different predictors of the word into one dataset to find out what predictors are the most important for predicting AoA(age of acquisition).

Data Processing

In order to merge the datasets by words, we switch all the words in different datasets to lower cases. The labels of all the columns are also standardized. Four different datasets with different predictors are merged together based on words. Then we replaced all the empty values with not a number identifiers. Words with predictors marked as not a number are removed from the merged dataset. After the steps mentioned, there are 720 words left in the dataset.

AoA Ratings

The AoA ratings converted the actual age of acquiring the word into a 1-7 scale. In the scale, 1 stands for 0-2 years and 7 stands for 13 years or greater than 13 years, with interim bands of 2 years each. At last, to create ratings on a scale of 100 to 700, the mean rating for each word was multiplied by 100. (Bird et al., 2001)

Predictors

Studied from the paper by Braginsky, Yurovsky, Marchman, and Frank (2016), we selected five predictors from the dataset we created. Frequency, valence, arousal, iconicity and concreteness. The words with max and min values are given as in Table 1 shows.

Frequency

It is the logarithm of combined lemma written and spoken count divided by total words in corpus of Celex Database. (*M* = 1.56, *SD* = 0.87) (Bird et al., 2001)

Valence

It is a scale for each word by asking adult participants to rate the word from 1 to 9. The lower the score is, the participant feels unhappier about the word. On the contrary, the participant feels happy about the word when the score is high. (*M* = 5.48, *SD* = 1.40) (Warriner et al., 2013)

Arousal

It is a scale for each word by asking adult participants to rate the word from 1 to 9. The lower the score is, the participant feels calmer about the word. On the contrary, the participant feels excited about the word when the score is high. (*M* = 4.22, *SD* = 0.95) (Warriner et al., 2013)

Iconicity

It is a scale for each word by asking adult participants to rate the word from -5 to 5. It identifies the relationship between the pronunciation of the word and the meaning of the word. -5 stands for the word sounds like the opposite of what it means, 5 stands for the word sounds like what it means and 0 stands for arbitrary. (*M* = 0.82, *SD* = 1.04) (Perlman et al., 2018)

Concreteness

It is a scale for each word by asking adult participants to rate the word from 1 to 5. The lower the score is, the more abstract the word is according to the participant. On the contrary, the word is concrete when the score is high. (*M* = 3.72, *SD* = 0.98) (Brysbaert et al., 2013)

Table 1: Examples of words with the lowest and highest values for age of acquisition and each predictor.

Measure	Value	Words
AoA	max	microwave, socialist, democracy
	min	cry, foot, leg
Frequency	max	do, have, be
	min	microwave, party, quiet
Valence	max	delight, hug, happy
	min	kill, death, harm
Arousal	max	alarm, die, spider
	min	empty, comb, solid
Iconicity	max	clang, howl, hiss
	min	microwave, penguin, sea
Concreteness	max	tree, water(N), water(V)
	min	concept, belong, ought

Principal Component Analysis

Before doing the PCA, we can find that the scales of data are different, some variables may dominate other variables of small ranges. In order to analyze the contribution of each variable equally, we first need to center and scale the data based on mean and standard deviation. Here are the standardization method we apply:

$$z = \frac{x-\mu}{\sigma}$$

, where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$.

Correlation Matrix

Figure 1 shows a correlation matrix of AoA and all the variables that we have selected in the data. Through the Figure 1, we can find that AoA have a correlation with all other variables except for Arousal.

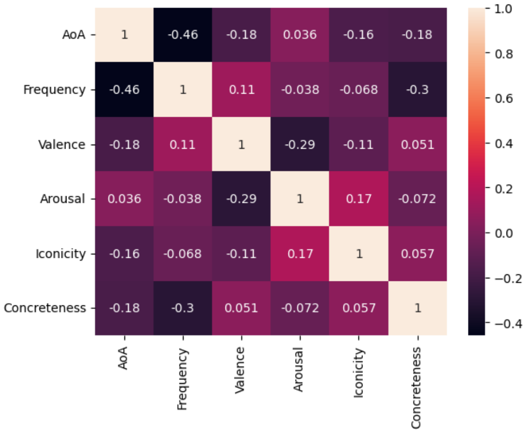


Figure 1: Correlation matrix of six variables.

Scree Plot and Variance Explained Plot

Before doing PCA, we need to create a scree plot and a Variance Explained Plot, as Figure 2 shows. Scree plots are used to visualise eigenvalues, which specify the magnitude of principle components. Variance Explained Plot shows the cumulative sum of the percentage of variances explained by each of the principle components. From the left plot of the Figure 2, we can find that inflection points exists while component is equal to either 2 or 4. According to Cattell's scree test, 1 or 3 PCs can a proper value to choose. Kaiser's rule also indicates that we should select PCs that have eigenvalues of at least 1. From the right plot of the Figure 2, together with the data from the program, we can know that 3 components can explain 69.55% of the variances. So, in this case, we are doing a 3D PCA for visualization .

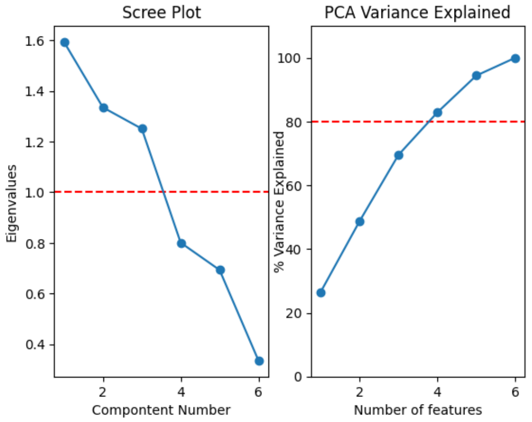


Figure 2: Scree plot showing the point of inflection and PCA Variance Explained plot showing the variances explained by each of the components.

3D PCA

Figure 3, Figure 4 shows us how 3D PCA plot looks like. We can observe that there are more nouns in the dataset. From Figure 4, we can find that the words with lower AoA tend to concentrate on the bottom right of the graph, while words with higher AoA tend to concentrate on the top left of the graph. Figure 5 is also created for observation. We can observe that some of the words with similar measurements also gathered together. The result of three figures clearly shows the relationship between AoA and other variables.

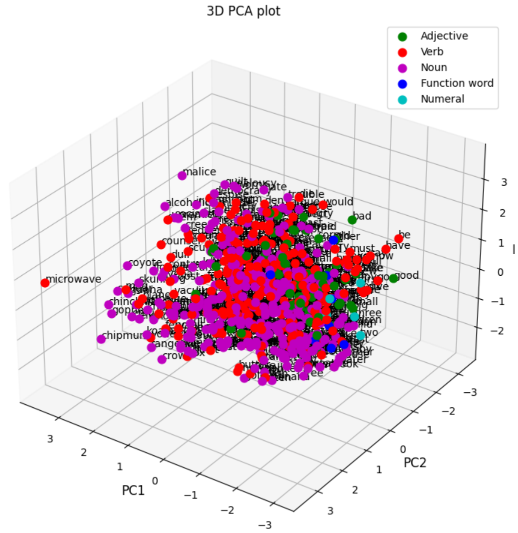


Figure 3: 3D PCA with all the words kept, labeled by word types.

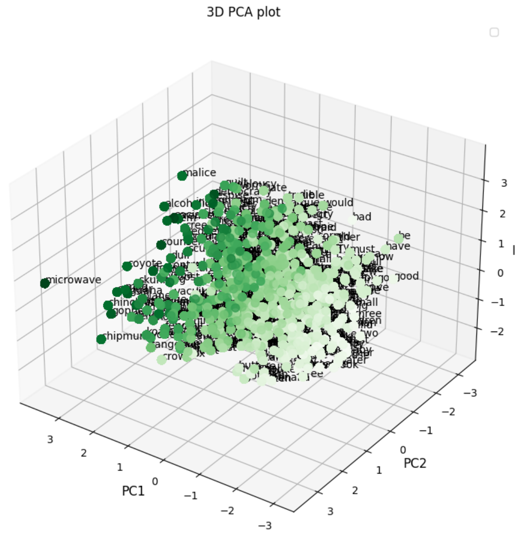


Figure 4: 3D PCA with all the words kept, colorized with AoA values.

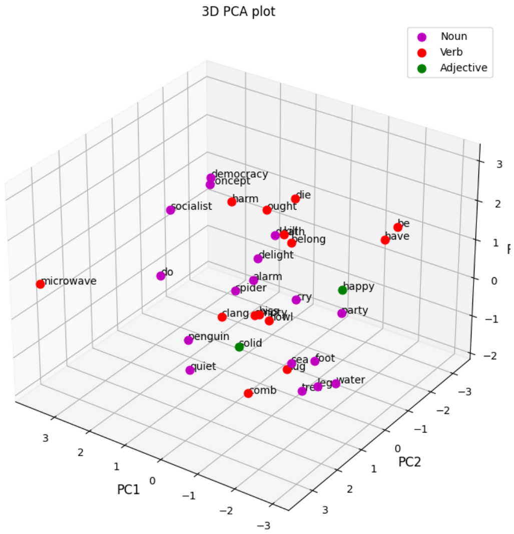


Figure 5: 3D PCA with the value words from Table 1 kept, labeled by word types.

Modelling

For modelling the data we chose to implement three different methods; linear regression, Bayesian ridge regression, and automatic relevancy determination.

Linear regression is the modelling method used by Braginsky et. al in their investigation. The regression is used to predict a dependent variable using one or more independent variables. This is done by determining a linear relationship between the dependent variable and the independent variables. In our case, the dependent variable is AoA, and the independent variables are our chosen predictors. A regression line is determined for the relationship between each independent variable and the dependent variable through a method called least squared estimation. The line that predicts the dependent variable with the lowest sum of the squared distances between the predicted values and the actual values is chosen as the regression line.

We also implemented Bayesian ridge regression as a part of our analysis to determine whether there are any key missing predictors of AoA in our analysis. Bayesian ridge regression is a variation of linear regression that instead of determining the regression line through least squared estimation, applies a gaussian distribution to determine the coefficient of each predictor prior to observing the data. As data is observed, the distribution is changed for each predictor using Bayes theorem to maximize the accuracy of the predictions. Given that each predictor's coefficient is assumed to have a normal distribution, a predictor that is not actually associated with AoA it will not impact the predictive model's accuracy as it's coefficients will simply be taken from a normal distribution with mean zero.

Automatic Relevance Determination (ARD) is another form of linear regression. ARD is a variation of Bayesian ridge regression that aims to determine which of the predictors is most effective in predicting AoA and gives that predictor the most weight in the model. ARD works in a similar way to Bayesian ridge regression as it also determines the coefficients of the predictors from a gaussian distribution which is updated as data on the predictors is observed. ARD then determines which of the predictors are most relevant and minimizes the effect of those that are irrelevant by shifting their coefficients closer to zero. By minimizing the effect irrelevant predictors, ARD ensures that its predictions are not skewed by irrelevant data. ARD was included alongside Bayesian ridge regression in our study to strengthen our analysis of whether or not any of our predictors are irrelevant in predicting AoA as if we see a significant improvement in the accuracy of both Bayesian ridge regression and ARD from linear regression, it could be said that at least one of the predictors chosen actually is not correlated with AoA.

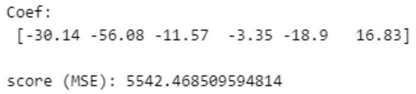
These three models were implemented using the python sklearn packages. First the data is split into training and test sets, then each model is trained and tested on the data splits.

Results

Models

Once all three models had been fitted, the found coefficients for each predictor were reported alongside the mean squared error for each model. The coefficients are reported in the order of concreteness, frequency, valence, arousal, iconicity, and length.

Linear Regression results as Figure 6 shows.

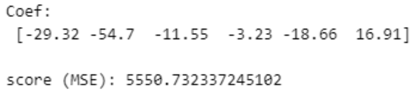


```
Coef:
 [-30.14 -56.08 -11.57  -3.35 -18.9   16.83]

score (MSE): 5542.468509594814
```

Figure 6: Linear regression results.

Bayesian Ridge Regression results as Figure 7 shows.

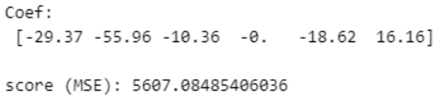


```
Coef:
 [-29.32 -54.7  -11.55  -3.23 -18.66  16.91]

score (MSE): 5550.732337245102
```

Figure 7: Bayesian ridge regression results.

ARD results as Figure 8 shows.



```
Coef:
 [-29.37 -55.96 -10.36  -0.   -18.62  16.16]

score (MSE): 5607.08485406036
```

Figure 8: ARD results.

Finally, a comparison between the coefficients of each predictor was visualized with the Figure 9 below.

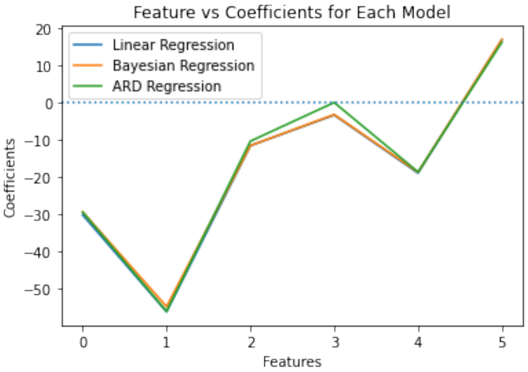


Figure 9: Feature vs Coefficients for each model.

Discussion

Given our results, we can see that there is actually little difference in prediction accuracy between the three models as they all reported a similar mean squared error. This tells us that our data and chosen predictors are all associated with AoA. If there were certain predictors that were not associated with AoA, we would see less error within the Bayesian ridge and ARD regressions. Instead, we see that there is actually a overfitting done by ARD. It seems as though ARD fits the coefficient of arousal to 0 as it was determined to not be relevant in predicting AoA, this could be seen as a misrepresentation of the association between arousal and AoA and could be an explanation as to why ARD has the highest error among the three regressions. We can also determine which of the predictors are given the highest weight in each model by looking at the coefficients found in each model. It can be seen that in all models, concreteness had the highest impact, followed by frequency. Valence, iconicity, and length were seen to have smaller, but still significant weight in the models while arousal was relatively low in each model. This matches the findings by Braginsky et. al as they also reported a high correlation in concreteness and frequency.

We believe that some interesting ways to continue this work could be through continuing to look for predictors of AoA. One variable that we hypothesized to be a potential predictor is that of overextension. Overextension occurs when a child applies one word to multiple meanings, for example, using the word dog to refer to all furry animals that walk on four legs. We believe that overextension could be a useful predictor of AoA and could lead to an interesting future study. Another way to further this investigation is to apply different prediction techniques. Neural networks could be a different approach to modelling AoA and it would be interesting to see how it compares to linear modelling.

Conclusion

For PCA, the result not only shows the possible relationship between AoA and other selected variables, it also shows us a way using PCA to find the possible relationship between one variable and other variables quickly and easily. However, the length of the paper restricted us from analyzing the PCs matrix. It is possible for us to have valuable findings if we can dissect the principle components in pairs with AoA values marked. In conclusion, the age of acquisition is the age at which we learn new words. There have been many studies into how we learn words and what factors affect when we learn certain words. One study by Braginsky et. al took a look at a number of predicting variables and modelled AoA with linear regression. For our report we conducted a similar study by modelling AoA with similar predictors through linear regression as well as probabilistic regression methods. We found that there was little difference between the three chosen regression methods, showing that the predictors chosen were all correlated with AoA, this matched our findings from our PCA as well. We also found that the concreteness and frequency of words are the strongest predictors of AoA and discussed how to further this work by exploring new predictors and more advanced methods of modelling..

Data Availability

All data and code for these analyses are available at `https://github.com/hassan3301/COG403Project`

References

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, amp; Computers*, *33*(1), 73–79. doi: 10.3758/bf03195349

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. *Cognitive Science*.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911. doi: 10.3758/s13428-013-0403-5

Nyamapfene, A. (2009, Oct). *Computational investigation of early child language acquisition using multimodal neural networks: A review of three models - artificial intelligence review.* Springer Netherlands. Retrieved from `https://link.springer.com/article/10.1007/s10462-009-9125`

Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in signed and spoken vocabulary: A comparison between american sign language, british sign language, english, and spanish. *Frontiers in Psychology*, *9*. doi: 10.3389/fpsyg.2018.01433

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. doi: 10.3758/s13428-012-0314-x