MSDS 6372 Project 3

**Team members:**

Kottegoda Manjula

Messa Araya

Kumar Rajeev

Date:2016-12-01

**Statistical Analysis and Prediction of Wikipedia’s Perceived Usefulness**

**Introduction**

Since its inception in 2001, Wikipedia, the largest free online encyclopedia has steadily become a one of the most visited websites on the internet. With over 5 million articles in 250 languages it currently ranks as the fifth most visited website across the globe. Searching on a subject on your favorite search engine or simply asking Siri a question on your iPhone often pulls up references to Wikipedia articles at the very top of the search results.

Wikipedia’s technical infrastructure is managed by Wikimedia Foundation, a nonprofit organization, while its content is created and maintained by a vast community of volunteer subject matter experts with the aim of providing a ‘neutral compilation of verifiable, established facts’ [1]. Content on Wikipedia is often kept accurate and current by a mechanism of citations and are updated with the verification from reliable sources [2].

Wikipedia is widely used as a basic reference when trying to familiarize yourself with a new subject. But even with controls to maintain content accuracy it is not always look at as a reliable source in academia based on the quality and completeness of the material.

In this study, we will examine the perception of faculty members from two Spanish universities on using Wikipedia as a credible reference source. The wiki4HE dataset from the UCI Machine Learning Repository [3] will be used in this analysis and Logistic Regression will be used as the methodology in performing the analysis. This paper will examine which variables are correlated and which one’s influence in the perception of Wikipedia’s reliability.

**Logistic Regression**

This section provides a brief background on the statistical technique employed to predict the probabilities of Wikipedia reliability. Since the response, namely perception on Wikipedia usefulness indicator, is categorical (binary) and has values Yes (useful) or No (not useful), ordinary least squares regression cannot be used as **assumptions of normality** of the responses and **homoscedasticity** of the residuals will be violated. The response is binomial and the mean of the distribution, which is the probability of usefulness (π), is to be modeled as a function of the Independent Variables Gender, Age, WIKIUSERS, DOMAIN, etc. This function cannot be linear since, theoretically, the predictions can range from - ∞ to +∞ but probabilities lie between 0 and 1. Hence a nonlinear transformation, log odds (Logit), is applied to the DV which is then expressed as a linear function of the IVs in the following manner [Agrestic, 1996]:

The above functional form of modeling the probabilities has the following advantages:

1. The estimated Logits are free to lie anywhere between -∞ to +∞.
2. The model performs even when the responses (enrollment probabilities) are non-normal.
3. The model has a linear form and the parameter estimates can be directly related to the Logit of wiki reliability.
4. The corresponding probabilities of wiki reliability can be obtained by transforming back the estimated Logit equation.

The estimates of the β parameters of the logistic response function equation are obtained by the method of maximum likelihood estimation.

**Descriptive Statistics**

Data on users of Wikipedia in our paper generally consists of information on whether using Wikipedia is reliable or not (PU\_1), years of experience, age etc. Table 1 gives a list of variables in the data while identifying the Independent (IV) and Dependent (DV) variables and their valid ranges. These variables are considered as potential predictors and are hence included in the model development. The outcome variable is the reliability of using Wikipedia which is binary with values Yes=1 (reliable) or No=0 (not reliable).

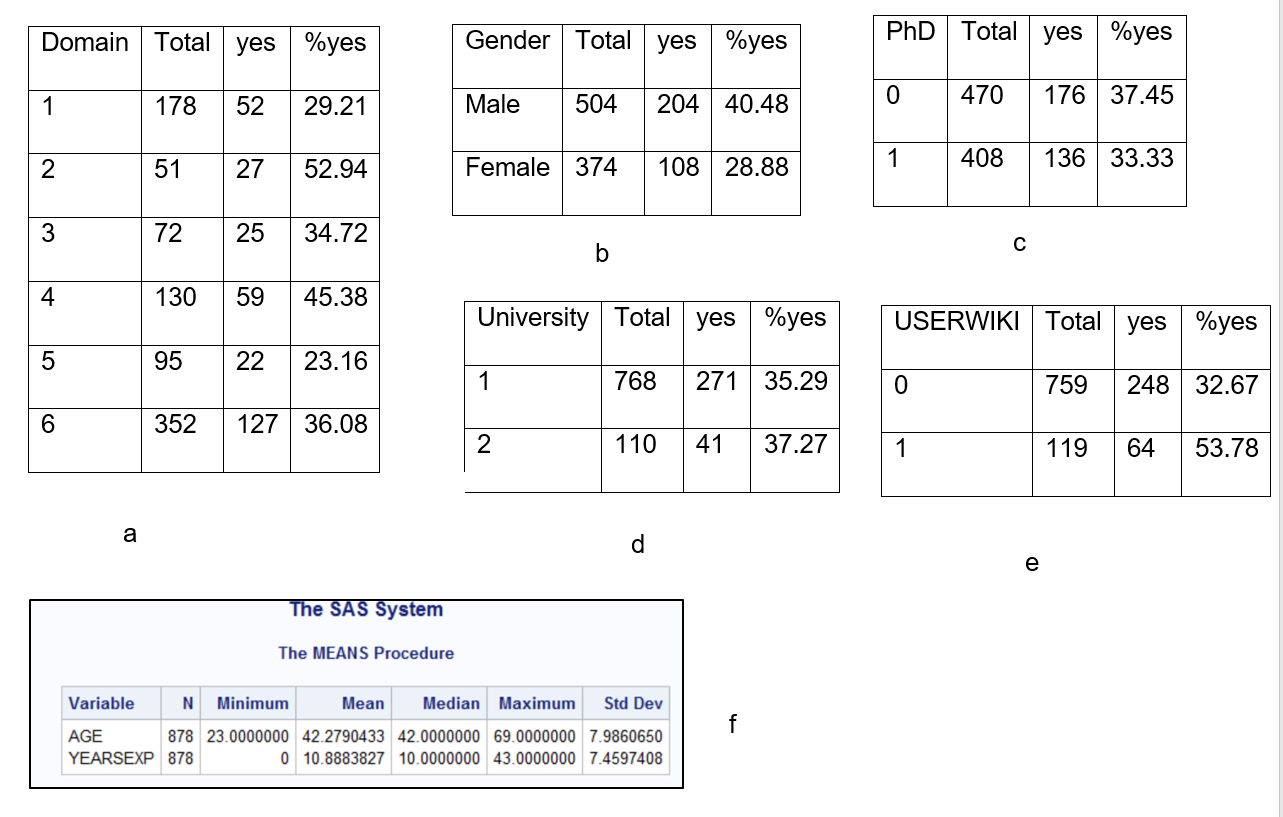
Table 1. Dependent and Independent Variables to be modeled

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Complete Name | IV/DV | Valid Range | Variable Type |
| Age | Age of university instructors | IV | >0 | **Continuous** |
| Gender | Gender of Ins. 0-Male, 1-Female | IV | 0-1 | Categorical |
| Domain | 1=Arts & Humanities; 2=Sciences; 3=Health Sciences; 4=Engineering & Architecture; 5=Law & Politics) | IV | 0-5 | Categorical |
| PhD | Highest degree | IV | 0-1 | Categorical |
| Years | Years of experience | IV | >=0 | **Continuous** |
| University | University | IV | 1-2 | Categorical |
| USERWIKI | User registered on wiki | IV | 0-1 | Categorical |
| PU\_1 | Reliability 1- Reliable, 0-Non-Reliable | DV | 0-1 | Categorical |

Table 2 (a) – (f) on page 4 gives data on total and the # of responders that say yes for the reliability of Wikipedia. The % gives the percentage of responders who said yes. In addition, Table 2 (f) shows the means, standard deviation, maximum, and minimum data for the continuous IVs (YearsExprience, Age).

The normality plot for the continuous variable of Age is fairly normal but the normality plot for YearsExprience had gross departures from normality (Figure 1(a)).

Table 2. Break down of variables.



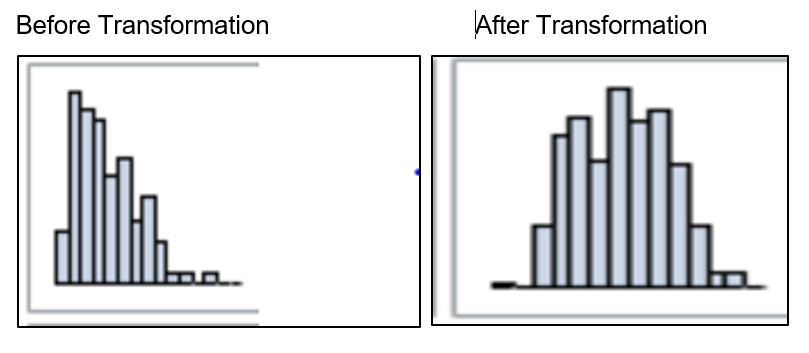
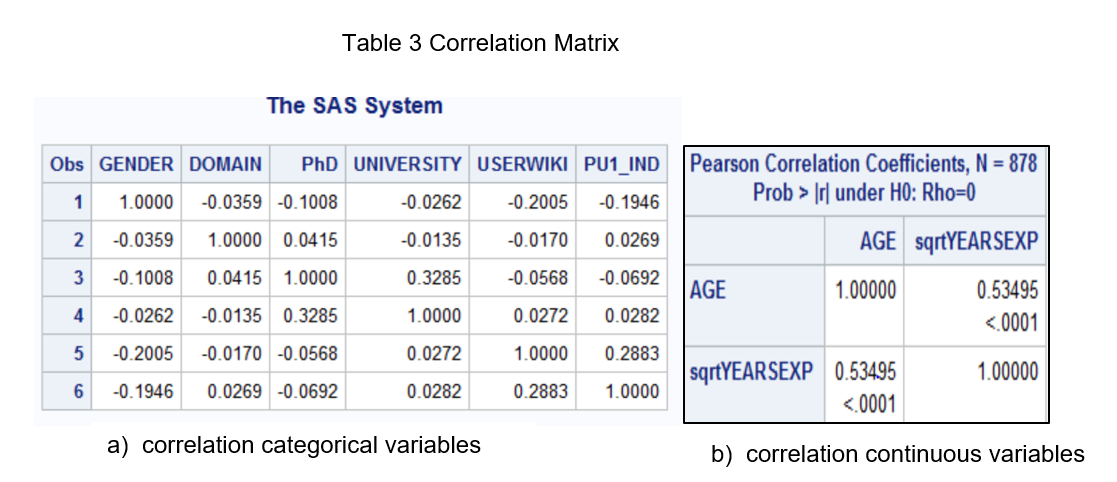


Fig 1. Histogram for YearsExprience before and after square root transformation

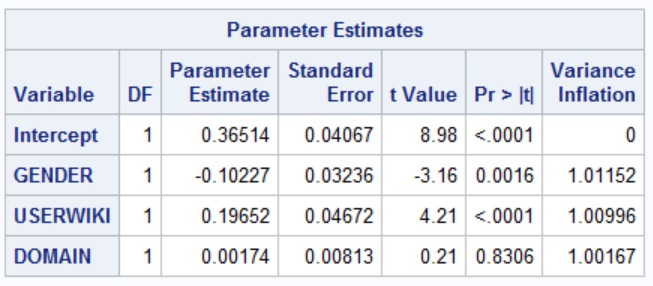
**Correlation between Variables**

We have checked the correlation between variables to check if there is multicollinearity between variables. The correlation analysis was done for continuous and categorical variables separately. There is no significant correlation between the categorical variables (Table 3(a)) whereas there is correlation (R2=0.54) between the continuous variables (Table 3(b)). We will later check if the continuous variables will be included.



We also checked the variance inflation factor (VIF) for the selected variables using logistic regression. These variables are GENDER, WIKIUSER, and DOMAIN. The value of VIF for the chosen variables is 1.00 for all of them (Table 4). Since it is a VIF less than 10, the multicollinearity is not a problem.

Table 4. Variance Inflation Factor (VIF)



**Initial Modeling**

In our initial model which included all variables, the first logistic regression model eliminated age, Ph.D., sqrtYEAREXP and University. And, the result was same with the stepwise automatic variables selection as well. Table 5 shows the result of stepwise selection. The p-values of the initial model after stepwise selection are shown in Table 6. The Wald chi-square score for USERWIKI , Gender and Domain are significantly higher than other elements in the initial model and all other variables reject the null hypothesis. USERWIKI , Gender, and Domain were the biggest contributing factors to the overall dataset, with male 148% more likely to find the usefulness of Wikipedia as shown below in Tables 6 and 7.

Table 5. Stepwise Selection

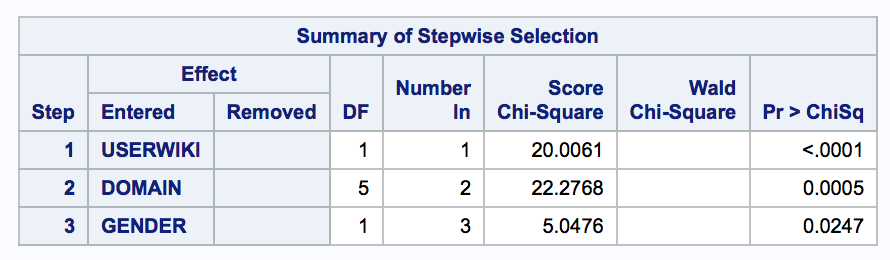
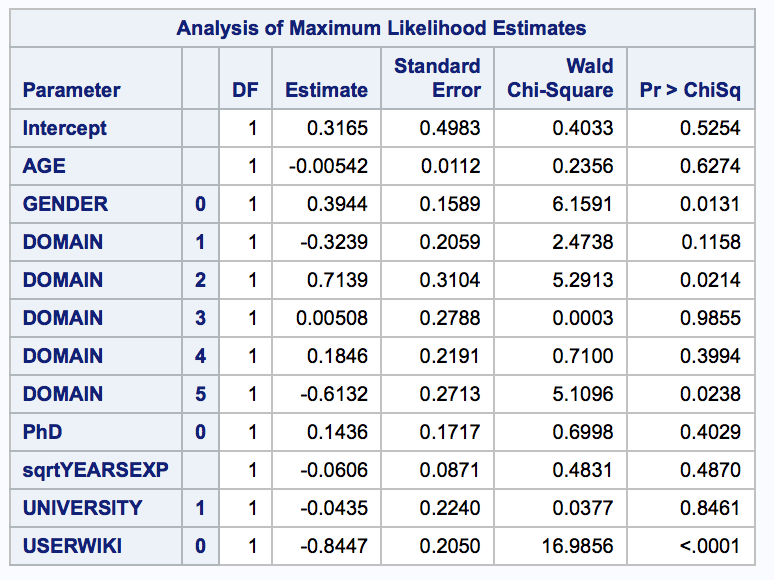
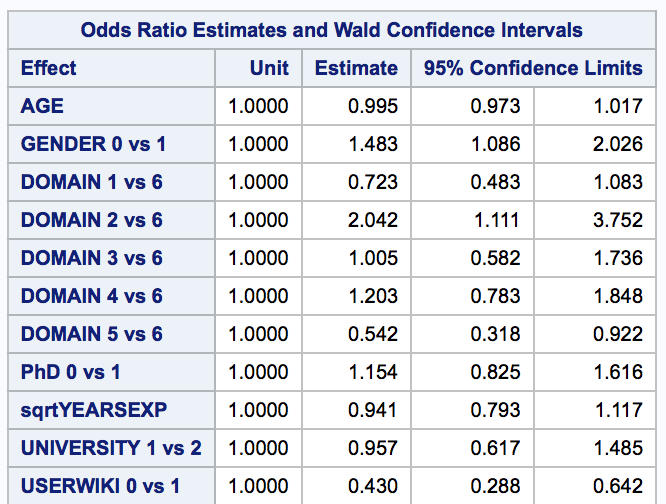


Table 6. Likelihood Estimates for Initial Model. Table 7. Odds Ratio for Initial Model

**Final Model Generation**

From our initial model we then narrowed it down to three major factors that influence Wikipedia useful ratings. Then, we attempted to determine if the model with interactions are better than the model without interaction. From Table 8, it is evident that none of the interactions are statistically significant.

Table 9. Final Model Fit Statistics

Table 8. Interaction Model

|  |  |
| --- | --- |
|  |  |

Therefore, we decided to keep the model without interactions of variables. We regenerated the final model while retaining only significant variables in the model. From table 9, Null hypothesis testing confirms that model is statistically significant than an empty model.

|  |  |
| --- | --- |
| Table 10. Final *Model* | Table 11. *Odd* Ratio |

From Table 10, it is evident that the odds for male to find Wikipedia favorable is 1.418 as high as female. The odds for non-registered users on the Wikipedia website is 42.6% as compared to registered users. Also, those who are in science domain (domain-2) gave a more favorable rating to Wikipedia (198.3% higher), and Engineering students (domain-4) gave a more favorable rating (126.3% higher). However, Arts & Humanities (domain-1), Health Sciences (domain-3) and Law & Politics (domain-5) students gave 73.7%, 98.7% and 53.7% lower odd ratio to Wikipedia.

From these analyses, we come to our final model shown below for the logistic regression of predictors for Wikipedia Perception:

Where x1 = gender, x2 is UserWiki, x3 is domain(1), x4 is domain(2),x5 is domain(3), x6 is domain(4), x7 is domain(5).

**Conclusion**

With the changing landscape of technology new forms of content delivery is becoming mainstream in academia. While some of the content is mere digital copies of its physical form there is also content that is produced through collaboration by groups of knowledgeable experts and scholars. Wikipedia falls under the latter category and as such is subject to scrutiny as to the quality and reliability of its content. In this study we examined several variables that could help explain the favorability, or lack thereof, that faculty members from Open University of Catalonia and Pompeu Fabra University might show towards Wikipedia.

Based on the model from this study it was determined that the age of the faculty member, university affiliation, years of teaching experience and having a Doctorate had little effect in influencing favorability towards using Wikipedia. It was also apparent that the gender of the faculty member played a significant role in determining favorability where male faculty were 142% more likely than their female counterparts in embracing Wikipedia. Being registered as a user of Wikipedia also played a role where non-registered users were 42% likely to be unfavorable towards Wikipedia. Furthermore faculty in STEM fields had a much higher favorability towards Wikipedia where educators from the Sciences were 198% more likely to accept Wikipedia. Faculty from Law and Politics had the lowest level of comfort having 54% lower odds than the rest.

When looking at the overall picture one can see that most of the faculty do not recommend relying on Wikipedia as an acceptable source for knowledge. Male registered faculty in science and engineering, the very demography that makes up a large portion of the volunteer content producers[4] seem to be the biggest advocates of the site.

One can argue that with time to come, knowledge portals such as Wikipedia will organically adopt more mainstream content governance procedures there by gaining wider acceptance and participation from university educators. Perhaps a broader study could help mold the governance structure needed to enable the inevitable future of open collaborative knowledge sharing.

# Bibliography

[1]: <https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_brief>

[2]: <https://en.wikipedia.org/wiki/Wikipedia:Verifiability#What_counts_as_a_reliable_source>

[3]: <https://archive.ics.uci.edu/ml/datasets/wiki4HE>

[4]: <https://www.technologyreview.com/s/520446/the-decline-of-wikipedia/>

**Appendix**

*PROC IMPORT DATAFILE= "/folders/myfolders/Data/Wikipedia4HE.csv"*

*OUT= wikipedia*

*DBMS=csv*

*REPLACE;*

*GETNAMES=YES;*

*RUN;*

*/\* descriptive statistics for the continuous variables \*/*

*proc means data=wikipedia n min mean median max std;*

*var AGE YEARSEXP;*

*Run;*

*/\* descriptive statistics for the catagorical variables and crosstabulation with the response \*/*

*proc freq data=wikipedia;*

*tables GENDER DOMAIN PhD UNIVERSITY USERWIKI PU1\_IND\*GENDER*

*PU1\_IND\*DOMAIN PU1\_IND\*PhD PU1\_IND\*UNIVERSITY PU1\_IND\*USERWIKI;*

*run;*

*/\* Scatter Plot \*/*

*proc sgscatter data=wikipedia;*

*matrix AGE GENDER DOMAIN PhD YEARSEXP UNIVERSITY USERWIKI PU1\_IND / diagonal=(histogram);*

*run;*

*/\* Square Root Transformation of Years of Experience \*/*

*data wikipedia1;*

*set wikipedia;*

*sqrtYEARSEXP = sqrt(YEARSEXP);*

*Run;*

*/\* Scatter Plot \*/*

*proc sgscatter data=wikipedia1;*

*matrix AGE GENDER DOMAIN PhD sqrtYEARSEXP UNIVERSITY USERWIKI PU1\_IND / diagonal=(histogram);*

*run;*

*/\* Checking correlation for continous\*/*

*proc corr data = wikipedia1;*

*var AGE sqrtYEARSEXP;*

*run;*

*/\* Checking correlation for catagory\*/*

*proc freq data = wikipedia1;*

*tables (GENDER DOMAIN PhD UNIVERSITY USERWIKI PU1\_IND )\*(GENDER DOMAIN PhD UNIVERSITY USERWIKI PU1\_IND) /plcorr;*

*ods output measures=mycatcorr (where=(statistic="Tetrachoric Correlation"*

*or statistic="Polychoric Correlation")*

*keep = statistic table value);*

*run;*

*proc print data=mycatcorr;*

*run;*

*data mycatcorrt;*

*set mycatcorr ;*

*group = floor((\_n\_ - 1)/6);*

*x = scan(table, 2, " \*");*

*y = scan(table, 3, " \*");*

*keep group value table x y;*

*run;*

*proc print data = mycatcorrt;*

*run;*

*proc transpose data = mycatcorrt out=mymatrix (drop = \_name\_ group) ;*

*id x;*

*by group;*

*var value ;*

*run;*

*proc print data = mymatrix;*

*run;*

*/\*Logistic Regression \*/*

*proc logistic data=wikipedia1 descending;*

*class GENDER DOMAIN PhD UNIVERSITY USERWIKI /param=ref;*

*model PU1\_IND = AGE GENDER DOMAIN PhD sqrtYEARSEXP UNIVERSITY USERWIKI/risklimits;*

*Run;*

*/\*Automatic Selection Process \*/*

*proc logistic data=wikipedia1 plots=ALL descending ;*

*class GENDER DOMAIN PhD UNIVERSITY USERWIKI/param=ref;*

*model PU1\_IND = AGE GENDER DOMAIN PhD sqrtYEARSEXP UNIVERSITY USERWIKI/CLODDS=pl selection=stepwise sle=0.05 sls=0.05;*

*run;*

*/\*Logistic Regression of selected variables only \*/*

*proc logistic data=wikipedia1 descending;*

*class GENDER USERWIKI DOMAIN/param=ref;*

*model PU1\_IND = GENDER USERWIKI DOMAIN /risklimits;*

*Run;*

*/\*Logistic Regression for Interaction of selected variables \*/*

*proc logistic data=wikipedia1 descending;*

*class GENDER USERWIKI DOMAIN /param=ref;*

*model PU1\_IND = GENDER|USERWIKI|DOMAIN/risklimits;*

*Run;*

*/\*Checking collinearity(VIF) for selected variables from logistic reg \*/*

*proc reg data=wikipedia1;*

*model PU1\_IND= GENDER USERWIKI DOMAIN / vif;*

*run;*

|  |  |
| --- | --- |
| Grade: 93 | |
| Mechanics | |
| Spelling & Grammar | Good |
| Flow of thought | Good. |
| Sections | Well defined |
| Figures & tables | Tables are all numbered and placed in right context. |
| Code | Included. |
| Title | Good |
| Analysis | |
| Data Description | The dataset has been described well prior to analysis. |
| EDA | EDA is included. |
| Method & interpretation | The analysis procedure is explained in details. The interpretation of odds ratio estimates needs to be reconsidered though. |
| Diagnostic | Some diagnostic of residual is recommended for outlier identification and assumption validation. |
| Conclusion | Reasonable conclusion is included. |