STA302 Final Project Report

Ebrahim Mohamud, Raihan Amin, Rebecca Li

December 6th, 2024

Methods of Data Analysis I

University Of Toronto

STA302H1

Introduction

As online dating has become widespread, understanding the dynamics shaping engagement grows increasingly important. Factors that enhance profile attractiveness can significantly impact users' outcomes in the digital dating landscape. In this paper, we explore the following research question: "Using a linear regression model, can we determine if age, gender of interest, profile picture count, profile visit count, verification status, the number of languages spoken, and country increase the number of likes on a dating app profile?"

In a study of 1705 university students, Castro et al. (2020) found that women and younger users received significantly more likes than their male and older counterparts (p-value ¡ 0.001). In addition, Hitsch et al. (2010) analyzed 793,530 online dating observations, demonstrating that 77.6% of profile views occurred for users with multiple photos. There was no holistic country preference, however 17% of men and 41.6% of women preferred shared backgrounds. Finally, Ellison et al.'s (2017) literature review of online dating in California found that users who displayed more pictures received more interactions. Together, these articles reinforce the importance of age, gender, picture counts, profile views, and country in online dating, providing a strong foundation for our current research.

Linear regression is a suitable model for this relationship because it allows for a predictive analysis between our response variable and the predictors. This model gives suggestions on the strength and direction of these relationships, while controlling for confounding factors. By leveraging linear regression, we aim to make predictions that could inform users about effective strategies for enhancing profile appeal in the landscape of online dating.

Methods

To answer the research question, a general methodology can be followed. To begin, a preliminary multiple linear regression model should be constructed and predictors are selected based on the results from an exploratory data analysis and findings from the aforementioned articles. Next, the model should be evaluated to ensure it meets the four assumptions of a linear regression model: linearity, constant variance, uncorrelated errors, and normality. These assumptions are formally assessed using residual plots, where non-random scatter, fanning patterns, clustered patterns, and nonlinear Q-Q plots indicate violations, respectively. If normality or linearity is violated, a box cox transformation should be used. If non-constant variance is violated then a variance stabilizing transformation should be used. These corrective measures are then reassessed by checking the four model assumptions again, until the given transformation satisfies the assumptions.

Next, an $ANOVA$ overall significance test is conducted to evaluate whether the selected variables explain a significant portion of the variation in user likes. If the test result is negative, the analysis must restart with a new set of

predictors or a different transformation. If instead, the test comes out positive, we proceed by using a Partial F-Test to determine if a subset of the initial set of predictors explains as much of the variability in the response as the full model. Once again, if the reduced model performs equivalently, we continue our analysis with the reduced model; otherwise, we continue with the full model. Next, to ensure the estimates of our model coefficients are reliable, we check for the presence of multicollinearity among our predictors by calculating the variance inflation factor (VIF) of each estimated coefficient. If any VIF exceeds 5, it is essential that highly correlated predictors are removed until the VIF is reduced below this threshold, thereby improving model reliability.

The next stage of answering our research question involves conducting a final set of model diagnostics, starting with identifying problematic observations such as leverage, outliers, and influential points. This allows us to determine if these observations unfavourably influence our model's overall predictive performance. Using metrics such as cook's distance, difference in fitted values ($DFFITS$), and difference in betas ($DFBETAS$), we identify any data points that may influence model estimates, and remove them only when there is valid contextual reason to do so. Following this, we perform either automated or manual selection methods to ascertain the best regression model according four metrics: the coefficient of determination (R^2), akaike's information criteria (AIC), adjusted akaike's information criteria (AIC_c), and bayesian information criteria (BIC). Once our final model has been selected, we end our analysis by validating it using methods like resampling, leave one out cross estimations, or splitting the data into training and testing sets to assess model generalizability. This comprehensive methodology ensures our final model is statistically robust, allowing us to confidently answer our research question according to the estimates in our final model. Observe the following flowchart as a reference for our methodology.

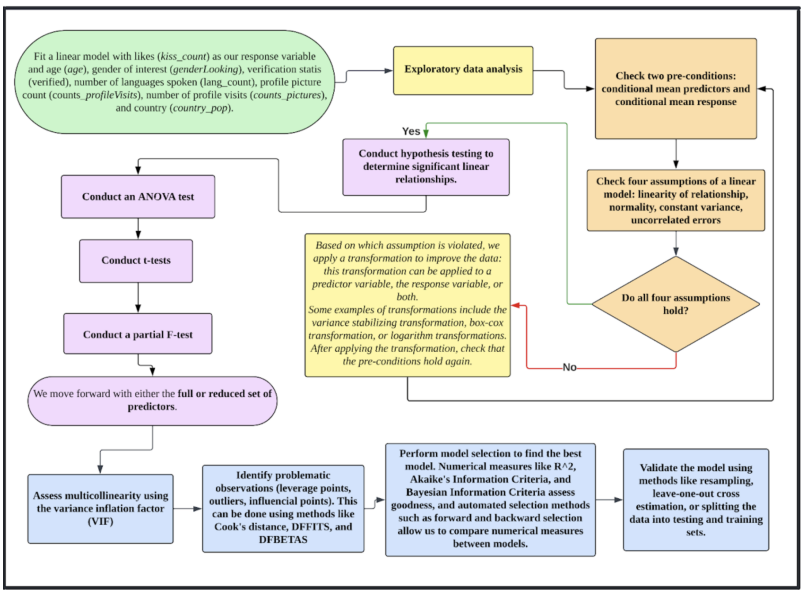

Figure 1: Analysis Flowchart

Results

We began our analysis by importing the dataset into R and performing an initial data-cleaning process. This involved renaming variables for consistency and clarity, creating a new variable to represent the full names of users' countries of origin based on their ISO country codes, along with a variable to represent each country's population. Data cleaning was essential for ensuring our analysis focused on predictors relevant to answering our research question. Next, we conducted exploratory data analysis (EDA) to summarize and visualize the data. This involved examining the distributions of our response variable as well as the various predictor variables, such as age, picture count, profile visits, country population, and language count by plotting their respective histograms. The initial EDA provided insight into the dataset and highlighted potential data issues such as skews in predictor or response variables that signify the need for a transformation. In our dataset, we observed extreme right skews across all variables, suggesting potential violations of normality and linearity assumptions which were noted for further investigation.

After analysing our dataset through the EDA and reviewing the relevant literature on our research topic, we selected seven predictor variables for our preliminary linear regression model. These variables included age, gender of interest, profile picture count, number of profile visits, verification status, the number of languages spoken, and the user's country of origin. These predictors were chosen to help explain variation in the number of likes on a dating app profile. To determine whether linear regression was a suitable model for addressing our research question, we formally checked the assumptions of linear regression using residual plot analysis. This analysis revealed a non-linear pattern in the QQ-plot, along with fanning patterns in the residuals versus fitted values and residuals versus profile visits plots. Before proceeding with any transformations, we checked both the conditional mean response and conditional mean predictor conditions to ensure the residual plots can be interpreted, and the results confirmed that these conditions were met. Since normality and constant variance were violated, we decided to apply a Box-Cox transformation selectively to variables that violated the assumptions. Using the power summary of the Box-Cox transformation, we applied a power transformation of 0.0882 to "counts kisses", 0.1184 to "country pop", and 0.2539 to "counts profileVisits". The resulting residual plots from this transformed model demonstrated significant improvement, with normality and constant variance violations successfully removed.

To determine if our set of predictors is significantly related to our response, we performed an ANOVA test, which yielded an F-statistic of 2608 with 3712 degrees of freedom. The test produced a p-value of $2.2e^{16}$, significantly below our threshold of 0.05, allowing us to conclude that at least one predictor is significantly related to the response. To determine if any predictors can be removed while still maintaining an equivalent level of variation explained, we conducted a partial F test. To identify potential candidates for removal, we performed a t-test which showed that all predictors were significant (conditional on the presence of the others). As a result, we tested the removal of each variable individually using the partial F-test. Our analysis concluded that no predictor variables should

be removed, as the maximum p-value from these tests was 0.035, which is below the significance threshold of 0.05. Consequently, we proceeded with the full model which includes all seven predictors.

Next, we evaluated our model for multicollinearity in order to ensure reliable model estimates. By analysing the variance inflation factor (VIF) we determined that some multicollinearity was present, but not enough to indicate significantly inflated variances and therefore severe multicollinearity. The highest VIF among our estimated coefficients was 1.378, which is well below the commonly accepted threshold of 5. We then analysed our model for problematic points, identifying 189 leverage points and 1 outlier. None of these points were classified as influential, as they did not significantly alter all estimated mean values or the overall regression trend. We used a manual selection method to determine our final model. We evaluated all combinations of predictor variables and assessed their quality using the following four metrics: adjusted R^2, AIC, AIC_c, and BIC. The full model and a four-predictor model (including age, county population, profile visits, and verification status) performed equally well across all metrics except BIC, where the four-predictor model prevailed. Finally, model validation was performed by splitting the data into a training or testing group, which resulted in similar adjusted R^2 values (training: 0.8625, testing: 0.8627) and relatively small MSE values (training: 0.00596, testing: 0.0839), reaffirming that our model was a good fit.

Conclusion and Limitations

Our analysis concluded that the four predictor model (with predictor variables: age, country population, profile visits, and verification status) is significantly related to the response and the best in its predictive performance according to the four metrics: adjusted R^2, AIC, AIC_c, and BIC. Our analysis showed that profile visits most impacted user likes since in its absence, the model's coefficient of determination fell to a low of 0.2418, confirming the findings of Ellison et al. (2017). Additionally, the model estimated that for a one-unit increase in the transformed profile visits, the mean number of likes a person receives increases by 0.4493 holding all other variables constant. This provides support for our hypothesis, and therefore the answer to our research question is that the number of user likes is expected to increase with profile visits and verification status, and decrease with age and country population, when using a linear regression model. These insights contribute to understanding what makes a successful profile on dating apps and highlight key factors that people may want to consider when engaging with online dating platforms, to improve overall user experience.

There are several complications with this model which should be addressed. Firstly, our analysis demonstrated a large number of problematic points present in the data. We observed 189 leverage points and while there was no indication of any points that were influential on all coefficients, 180 points were flagged as influential on their own coefficient. Therefore, it is likely that the coefficients we obtained for this model were affected by the aforementioned points and should not be taken as absolute measures. Additionally, the original dataset was collected using

a controversial method of creating two male profiles within a dating app and collecting data through the profiles that the app recommended. Thus, a majority of the data came from women who indicated a male preference. As a result, a large proportion of the variable "gender Looking" within the dataset consisted of male preference which may have skewed the relative importance of this variable. Overall, while we were able to find results that supported our hypothesis, these limitations should be taken into careful consideration before making any judgments from the model.

Ethics Discussion

In our analysis, we decided to adopt a manual selection method purely for practical reasons given that both selection methods raise equivalent ethical concerns. Fundamentally, our model aims to predict rather than describe the relationship between the various predictors and the response. Consequently, we avoid the ethical concerns of introducing bias by selecting variables based solely on statistical significance while ignoring the context of our data (in automated selection) since we are not interested in the context but rather the predictive performance. Similarly, we mitigate the ethical concerns of a subjective definition of a "best" model (in manual selection) since our model prioritizes prediction accuracy and thus defines the best model as one that minimizes the Residual Sum of Squares (RSS). Metrics such as adjusted R^2, AIC, AIC_c, and BIC align with our definition and eliminate the ethical concerns. Since both methods are ethically equivalent – neither being more negligent than the other – we opted for a manual selection method for practical reasons. For instance, since our full model only had seven predictors, it was not impractical to use an all subsets selection method. Also, automated selection methods only give an idea of the best model while the all subsets selection method gives the actual best model according to a given metric. For these reasons we opted to employ a manual selection method over an automated one. To ensure our smaller model remained reliable, we reevaluated various model diagnostics such as checking for multicollinearity and checking the linear regression assumptions before proceeding with the model validation.

References

1. Castro, Á., Barrada, J. R., Ramos-Villagrasa, P. J., & Fernández-del-Río, E. (2020). Profiling dating apps users: Sociodemographic and personality characteristics. *International Journal of Environmental Research and Public Health*, *19*(3), 1575. https://doi.org/10.3390/ijerph17103653

2. Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, *11*(2), 415-441. https://doi.org/10.1111/j.1083-6101.2006.00020.x

3. Hitsch, G. J., Hortaçsu, A., & Ariely, D. (2010). What makes you click? Mate preferences in online dating. *Quantitative Marketing and Economics*, *8*(4), 393-427. https://doi.org/10.1007/s11129-010-9088-6

4. Kaggle dataset: Mabilama Jeffrey Mvutu. (2015). Dating App Lovoo User Profiles. *Kaggle*. https://www.kaggle.com/datasets/jmmvutu/dating-app-lovoo-user-profiles

5. Original source of the Kaggle dataset: Jfreex. (2015). Dating App User Profiles' Stats - Lovoo v3. *Data World*. https://data.world/jfreex/dating-app-user-profiles-stats-lovoo-v3

6. Population data: World Bank. (2024). Population, total (SP.POP.TOTL). https://data.worldbank.org/indicator/SP.POP.TOTL?locations=1W