Stats 5100 Project 2 Paper – Final Draft

Morgan Hales A02298007

Introduction

Ever since I graduated high school, I have had a tinder account, however, I have not met up with anyone I met on the app because I'm afraid they aren't who they say they are. The article titled "Where Dating Meets Data: Investigating social and Institutional Privacy Concerns on Tinder," suggests twenty predictive variables that might indicate a tinder users' level of genuineness. If I can create a predictive, simplified, efficient linear model to how genuine a user is, I will feel safer meeting up with people I've matched with on the app. I used the data provided to create a model to achieve this purpose.

Data

To best analyze a tinder user's genuineness, we will be performing linear regression on data collected from surveys of tinder users. The linear model will predict genuineness based on twenty predictor variables. The fifteen quantitative variables and five qualitative predictor variables are described in Figure 1 below.

Variable Name	Definition
ID	user identifier (arbitrary and meaningless other than to identify specific users)
Genuine	in terms of how they present themselves on Tinder, how genuine is the user (i.e., how honest and realistic, as opposed to fake or falsified or made-up)
SocPrivConc	how concerned is the user that other users will mis-use their private data
InstPrivConc	how concerned is the user that Tinder will mis-use their private data
Narcissism	how narcissistic is the user
SelfEsteem	how much self-esteem does the user have
Loneliness	how lonely is the user
Hookup	how interested is the user in using Tinder to hook up (especially for sex)
Friends	how interested is the user in using Tinder to build friendships
Partner	how interested is the user in using Tinder to develop a partnership
Travel	how interested is the user in using Tinder while traveling
SelfValidation	how interested is the user in using Tinder for self-validation
Entertainment	how interested is the user in using Tinder for entertainment
Orientation	user's sexual orientation (1=heterosexual, 2=homosexual, 3=bisexual, 4=other)
Gender	user's identified gender (1=male, 2=female, 3=other)
Education	user's education level (1=no schooling, 2=high school graduate, 3=some college, 4=undergrad degree, 5=masters degree, 6=doctoral degree, 7=other)
Income	user's estimated level of income (1=low, 2=medium, 3=high, 4=unknown)
Employment	user's current employment status (1=employed, 2=sef-employed, 3=out of work and looking, 4=out of work but not looking, 5=homemaker, 6=student, 7=military, 8=retired, 9=unable to work)
Age	user's age in years
ImpFitness	how important does the user think it is for members of their same gender to have physical fitness
ImpEnergy	how important does the user think it is for members of their same gender to have energy (or stamina)
ImpAttractive	how important does the user think it is for members of their same gender to have physical attractiveness

Figure 1: Variables and Descriptions

A better understanding of the data can be achieved via visualization. A histogram of the distribution of genuine levels is shown in Figure 2.

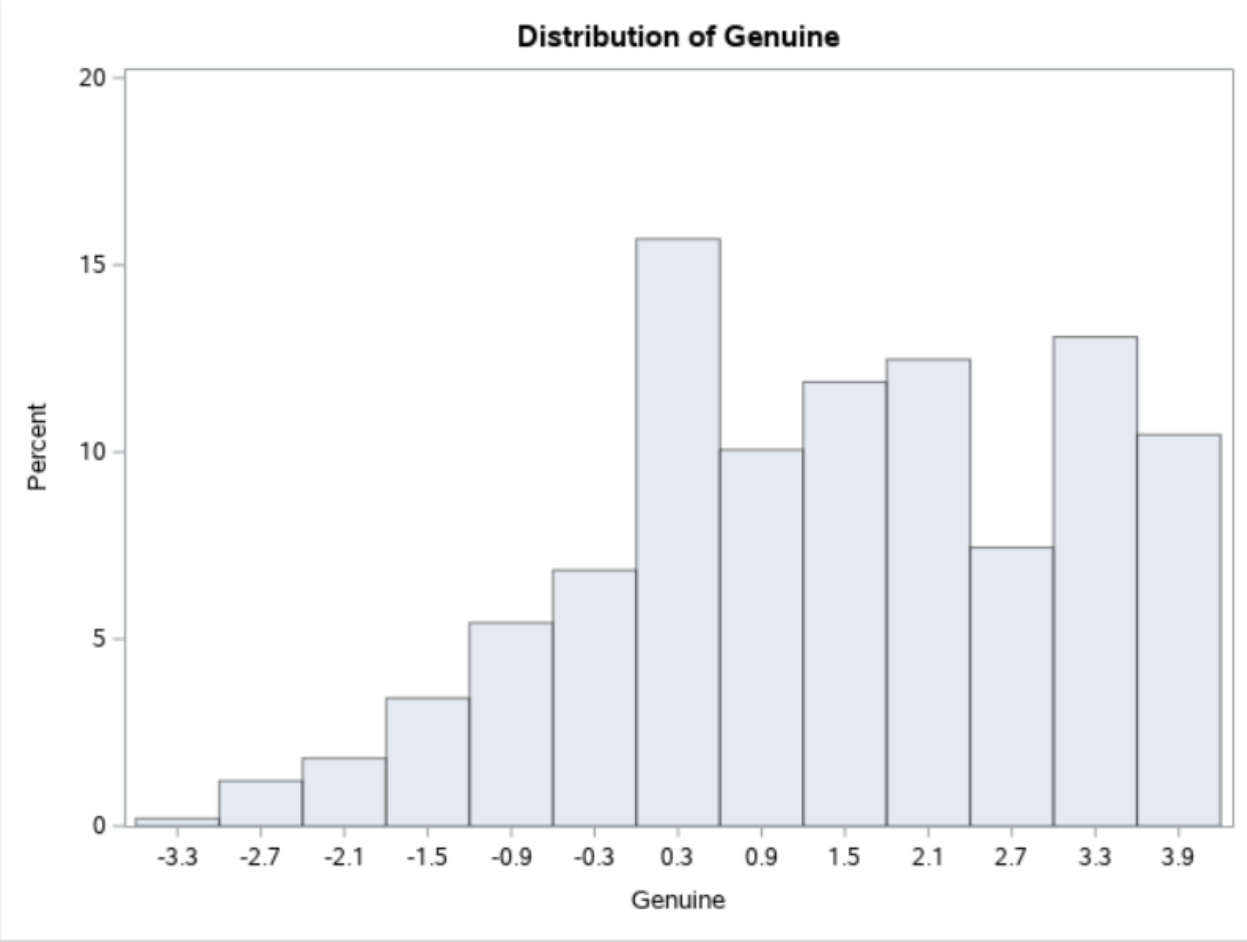

Figure 2: Histogram of Data

There may appear to be skew in the data, however, this is only due to a hard upper limit at 4. The histogram suggests that the percentage of observations increases with genuine until it reaches 0.3, then stays consistent until it reaches the hard upper limit. A scatterplot matrix of the

predictor variables can indicate whether there is a strong linear relationship between the response variable and the predictor variables.

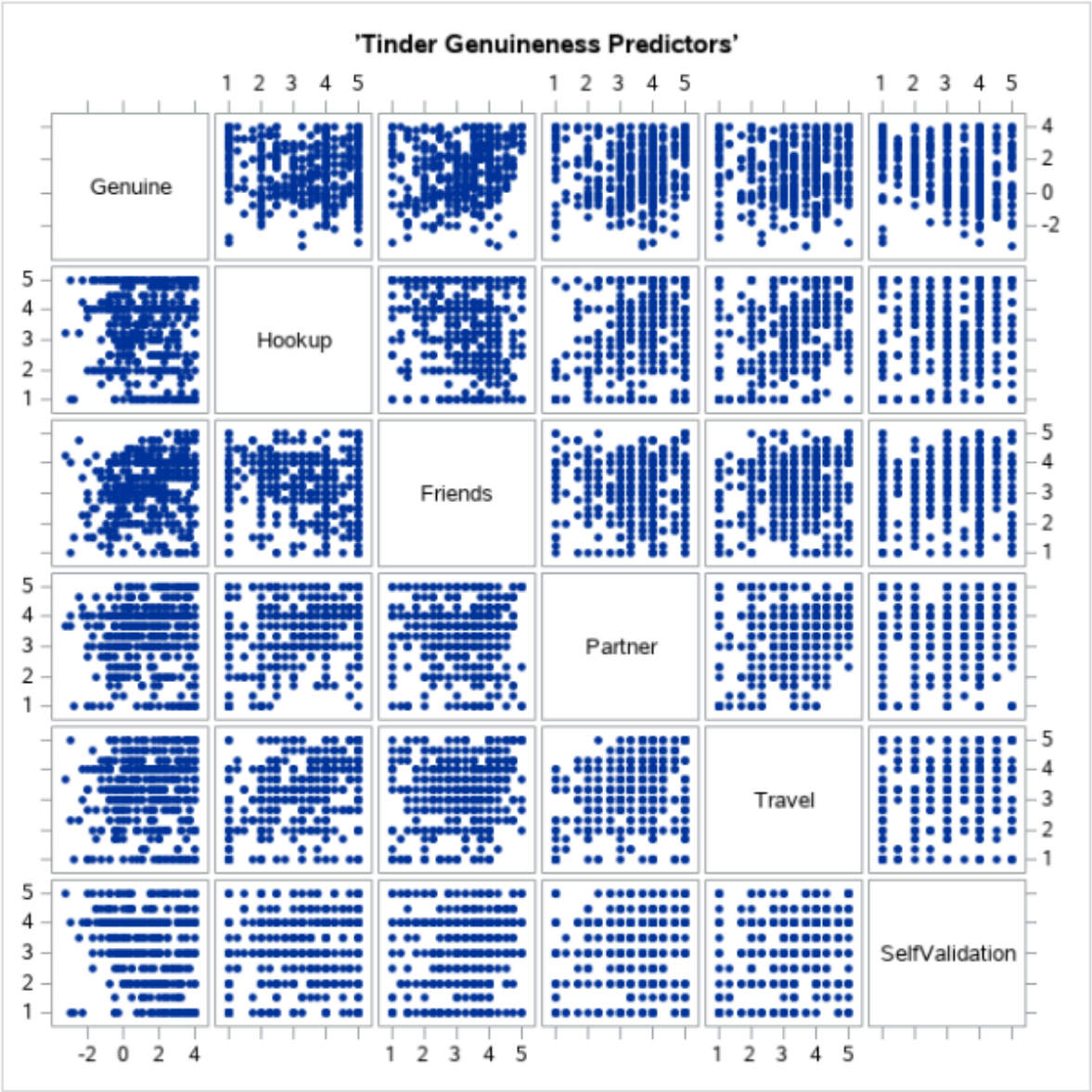

Figure 3: Scatterplot Matrices of Variables

Unfortunately, the scatterplot matrix does not tell us much about the linear relationship between the predictor variables and the response variable. This is because there are many data points, but only four to five possible value for the variables. Due to the lack of linearity, there isn't graphical evidence of multicollinearity. Even if there were multicollinearity in this model,

or our final model, it would not concern us as we are more focused on variable prediction rather than variable influence.

The final way we can visualize the data is via a box plot. This will show us if there are obvious outliers in the levels of genuineness.

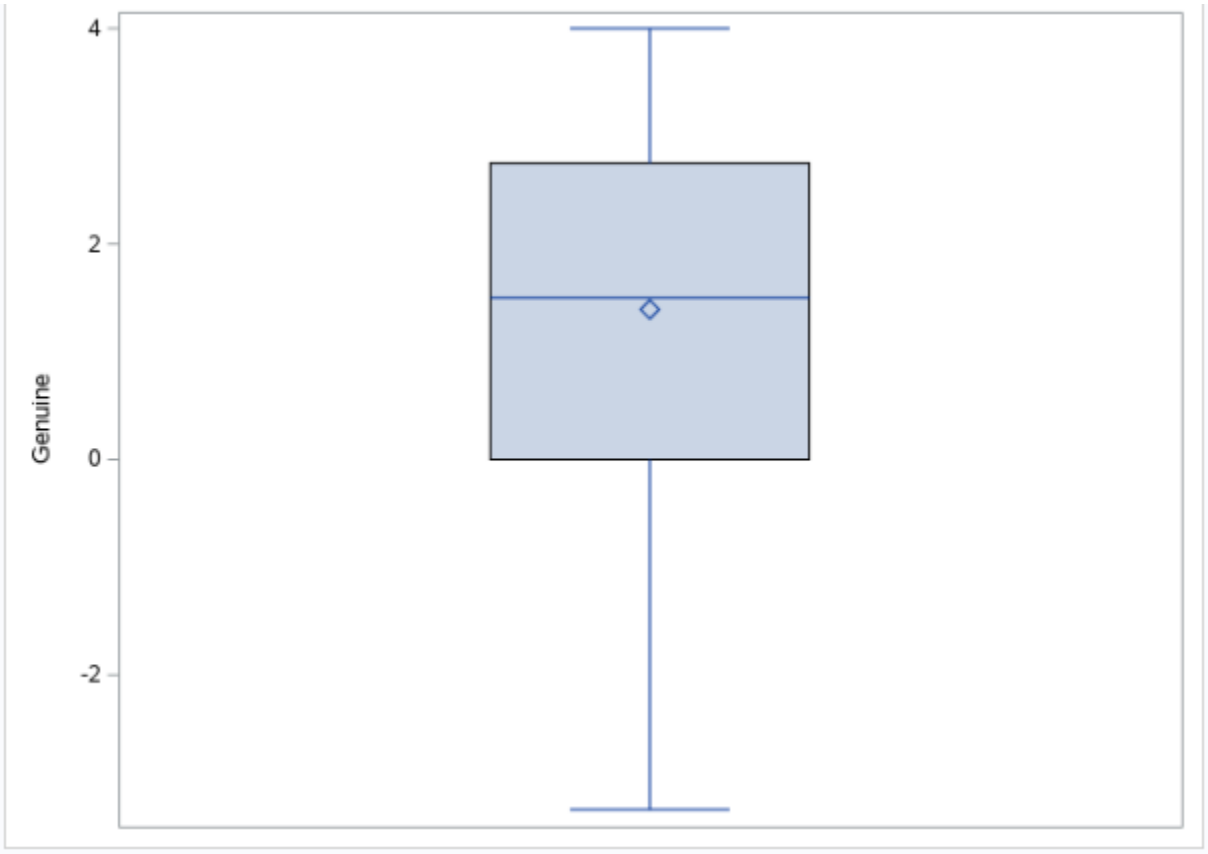

Figure 4: Boxplot of Data

The boxplot in Figure 4 suggests the data are left skewed and that there are no outliers in the response variable genuineness. Now that we understand the distribution of the data, we can take a deeper look at the data to create a predictive model.

Modeling Assumptions

Before we can develop a linear model for prediction on the data, we must check if the data meet the modeling assumptions. If the data does not fit the assumptions, our model is not valid. A linear model must have constant variance, independent residuals, and be normally distributed. One way to check this is via the Brown-Forsythe test of constant variance. The p-

value for the test is 0.2494, much greater than the threshold of 0.05. This suggests that there is not sufficient evidence of non-constant variance in the residuals, satisfying the model assumptions. The other way to check modeling assumptions is via graphical checks. A plot of the residuals can show whether there is constant or non-constant variance.

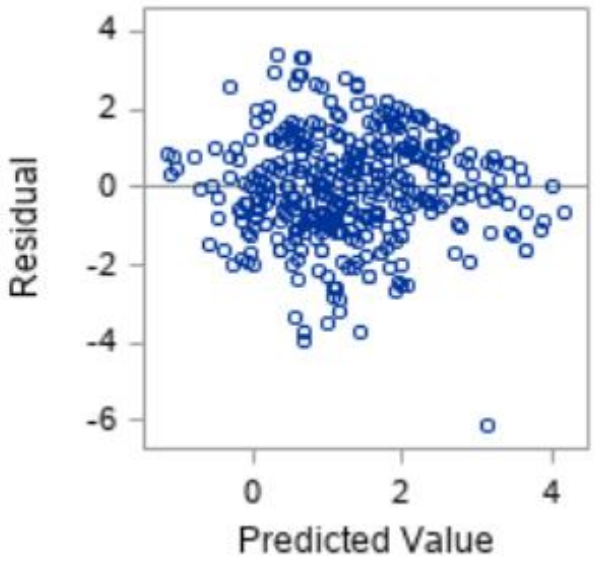

Figure 5: Scatterplot of Residuals

The plot of the residuals of the Tinder data in Figure 5 shows that variance is constant, in other words, there is no clear pattern seen as the predicted value increases. The Brown-Forsythe p-value combined with this visualization suggests that the data meets the necessary condition of having constant variance.

The sequence plot of the residuals can tell us if the residuals are independently distributed. However, since our data is ordered, the sequence plot is not justified. Instead, we must look at the histogram of the residuals in Figure 6.

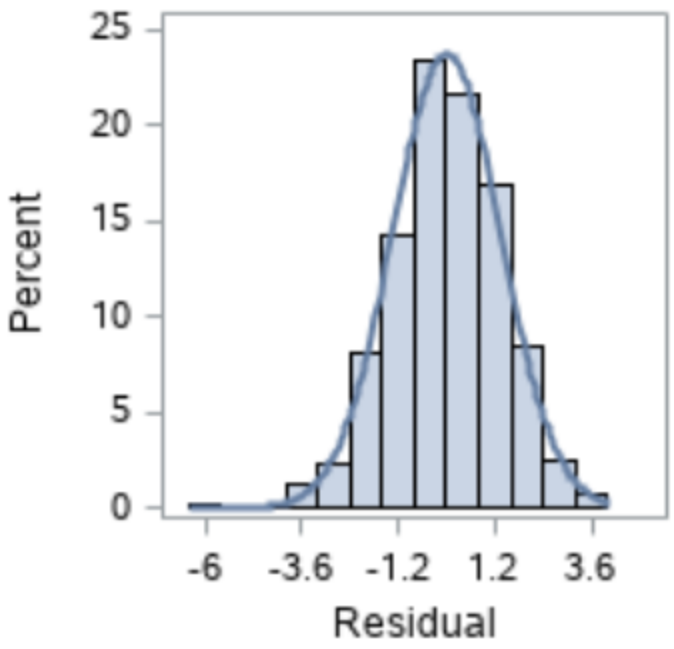

Figure 6: Histogram of Residuals

The histogram of the residuals in Figure 6 nearly follows the normal distribution. There is some skew present but not enough to invalidate the normality of the residuals. Due to this, the assumption of normality is met.

Finally, we must check that the residuals are normally distributed. This can be checked by evaluating the normal probability plot. If the data falls along the line in the plot, then the data is normally distributed.

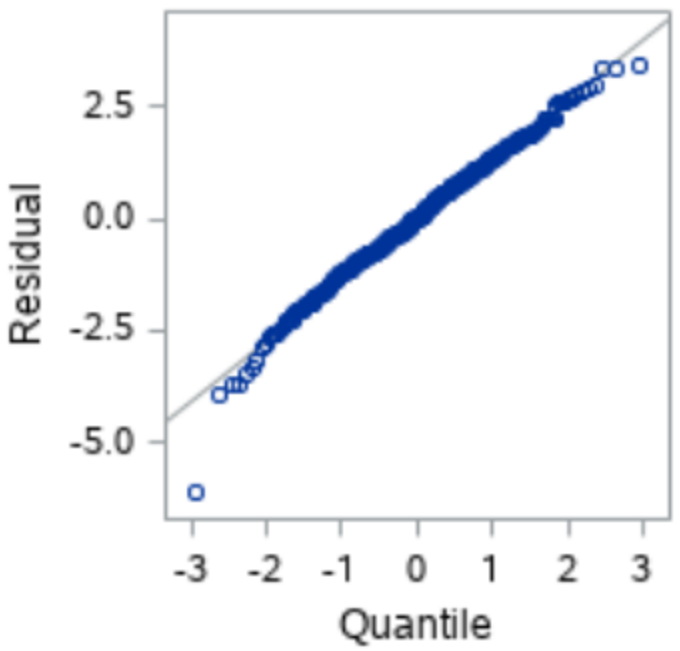

Figure 7: Normal Probability Plot of Residuals

Figure 7 shows the normal probability plot of the data. The data falls along the line, except for a possible outlier, showing no exponentiation. From this final graphical check, all the modeling assumptions have been met.

Now that assumptions have been met, we must check for outliers and influential points. If these are included in the model, the accuracy of the model is decreased. We can check for outliers graphically via a Cook's D plot.

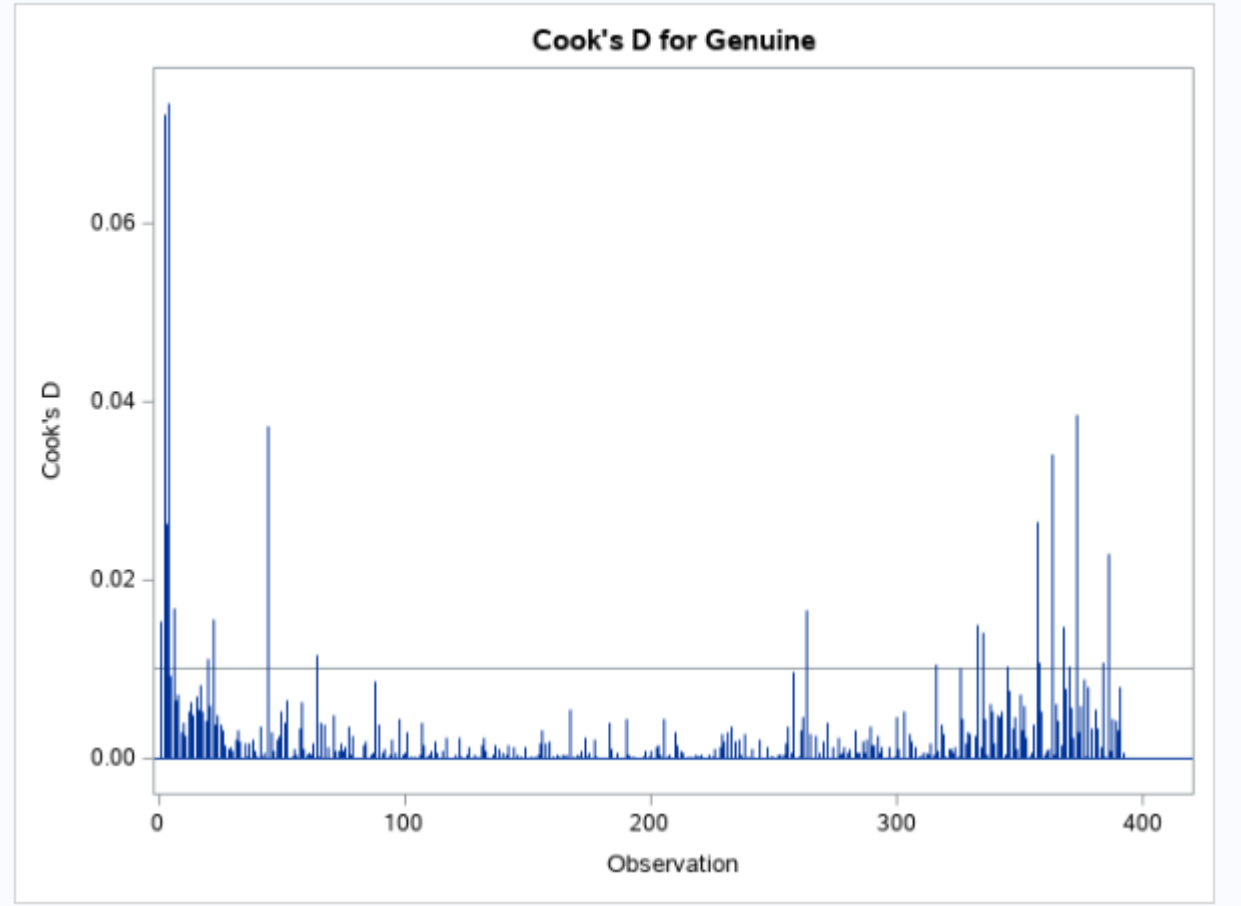

Figure 8: Cook's D Plot

The Cook's D plot for our data, shown in Figure 8, suggests that observations 2 and 4 are influential points. We can check if these points are consistently affecting our model by checking the leverage plot. The leverage plot for our data seen in Figure 9, shows that the observations 2 and 4 are outliers and leverage points.

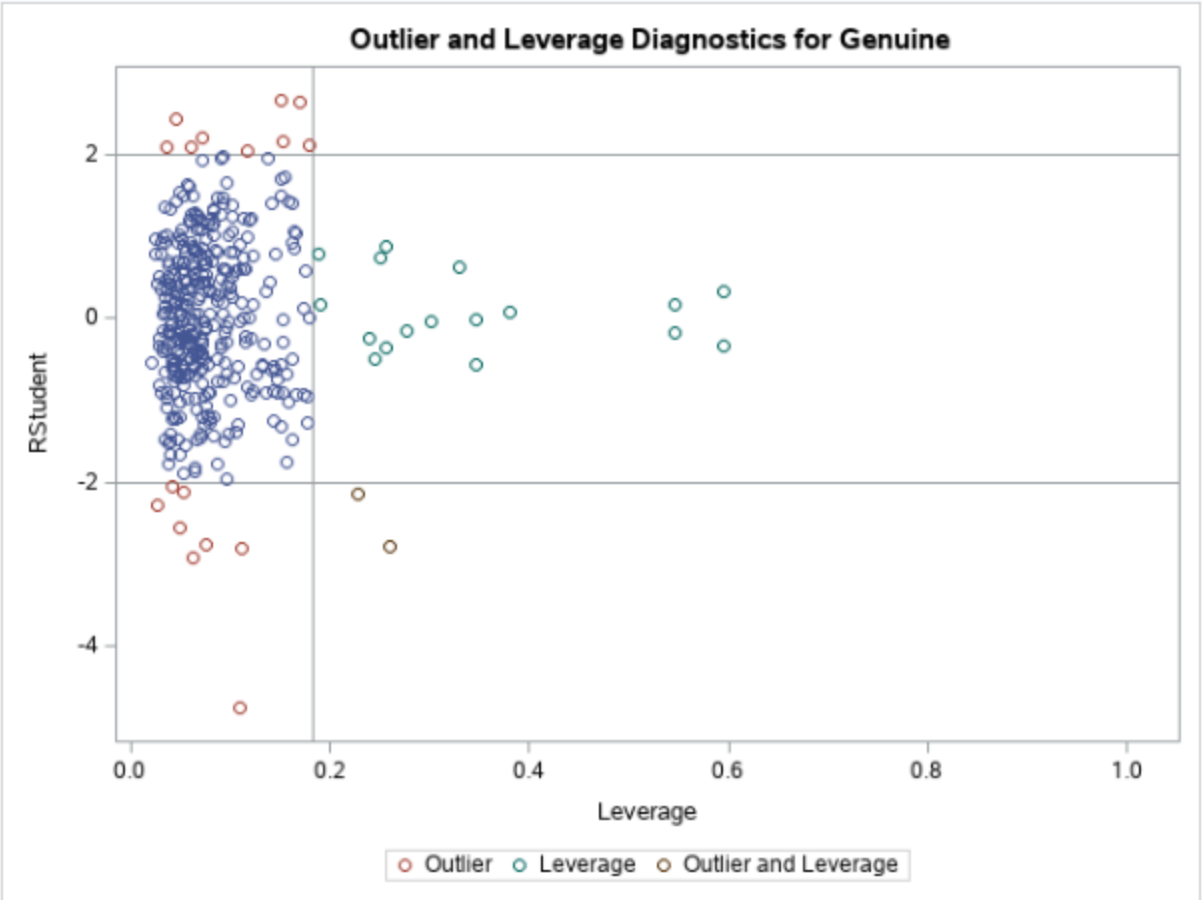

Figure 9: Leverage Plot

Since we are trying to create a model to predict on average Tinder users, outliers would affect the results and make our model less accurate. Due to this, we will remove the observations 2 and 4.

Variable Selection

The purpose of this paper is to improve on the original model provided to use with the data. We can make the model simpler and more efficient through variable selection. We will pick variables to include through backwards selection and then compare the results with the results of stepwise selection. Both stepwise collection and backward selection resulted in models that included the variables SocPrivConc, SelfEsteem, Hookup, Friends, Partner, SelfValidation, ImpFitness, Orientation, Education and Employment.

To triple check these variables, we will use another variable selection technique: LASSO. This resulted in a model that included SelfEsteem, Hookup, Friends, Partner, SelfValidation, ImpFitness, Orientation, Education, Income and Employment.

We will take the overlap of the three of these models' variables in our final model. Without interaction, our model will include SocPrivConc, SelfEsteem, Hookup, Friends, Partner, SelfValidation, ImpFitness, and Income.

After checking for interactions in the quantitative variables there are significant interactions between SocPrivConc and Partner, Friends and SelfValidation, and Partner and SelfValidation. This makes sense as these variables are somewhat related to each other. We will include these in our model. This results in the following model:

$$\hat{Y} = 1.916 + 0.306(SocPrivConc) + 0.617(SelfEsteem) - 0.223(Hookup) - 0.129(Friends) + 0.259(Partner) - 1.010(SelfValidation) - 0.089(ImpFitness) - 0.463(Income1) - 0.825(Income2) - 1.244(Income3) - 0.139(SocPrivConc * Partner) + 0.114(Friends * SelfValidation) + 0.082(Partner * SelfValidation).$$

To understand this model, the coefficient of -0.223 means that, while everything else is constant, for each unit increase in Hookup, there is a decrease of 0.223 in genuineness. In the case of interactions, the coefficients 0.129 and 0.114 mean that for every increase in unit increase in friends, genuineness increases by 0.129 +0.114(SelfValidation).

Model Validation

Since we have created a new model, we must recheck the assumptions.

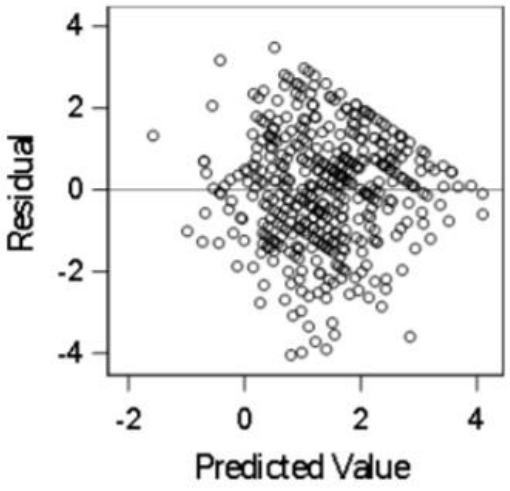

Figure 10: Simplified Scatterplot of Residuals

The plot of the residuals in Figure 10 shows non-constant variance, in other words, there is no pattern in the residuals. This satisfies one of the three necessary model assumptions.

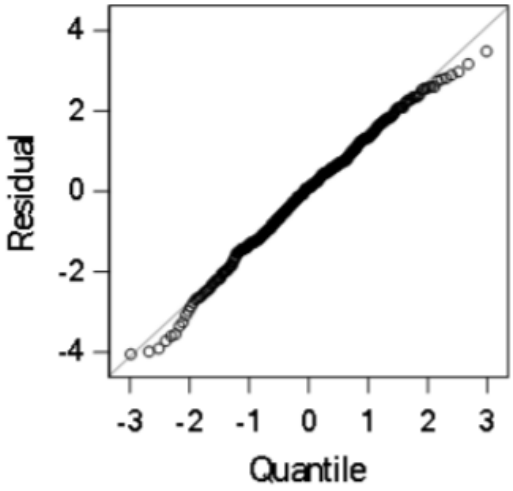

Figure 11: Simplified Normal Probability Plot

The normal probability plot of the final model in Figure 11 shows linearity, with points falling along the line. This suggests normality of the residuals, satisfying two of our three necessary model assumptions.

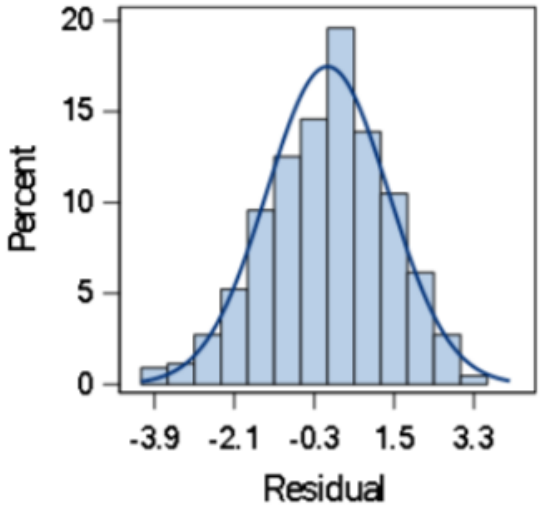

Figure 12: Simplified Histogram of Residuals

The histogram of the residual plots follows the normal curve, suggesting normality in the model. Therefore, all of the three model assumptions have been met.

Finally, we check our model error to see if it is truly a better model than the original. We compare it with the original and a model with no variables. The model with no variables reports an MSPR of 2.728, the original model reports an MSPR of 2.09, and our model's MSPR is 2.054. These values indicate that our reduced model is much more accurate than the null model and slightly more accurate than the original model. Our simplified model is more accurate and more interpretable due to its simplicity. From this, we should use the simplified/reduced model over the original model.

Conclusion

Using this simplified model, the genuineness of a Tinder user can be predicted efficiently with accuracy. The characteristics that best determine a user's genuineness include: SocPrivConc, SelfEsteem, Hookup, Friends, Partner, SelfValidation, ImpFitness, and Income. Unfortunately, when swiping through the app, I cannot tell a user's intentions with the app or their income level, so realistically, I can't use this for my own purposes. In future research it would be interesting to consider how many photos and what types of photos a user has on their profile as genuineness predictors. That way another user can more easily tell if a Tinder user is genuine.

Appendix

```
FILENAME REFFILE '/home/u49772396/5100 HW/tinder.csv';
PROC IMPORT DATAFILE=REFFILE replace
DBMS=CSV
OUT=WORK.tinder;
GETNAMES=YES;
RUN;
/* Check hard upper bound on genuine */
proc univariate data=tinder;
var genuine;
histogram genuine;
run;
/* Add ID variable where we sort by the genuine score */
proc sort data=tinder; by genuine;
data tinder; set tinder; ID = _n_;
/* Plot the genunine scores. */
proc sgplot data=tinder;
scatter x=ID y=genuine;
run;

proc glmmod data=tinder outdesign=GLMDesign outparm=GLMParm NOPRINT;
class orientation gender education income employment;
model genuine = socprivconc instprivconc narcissism selfesteem loneliness
hookup friends partner travel selfvalidation entertainment
age impfitness impenergy impattractive orientation gender
education income employment;
run;
/* Separate Into Training and Test Sets.
```

```
Only Fit Models to the Training Set. The variable
"Selected" separates training (0) from test (1) */
proc surveyselect data=GLMDesign seed=11357 out=tinder2
rate=0.1 outall; /* Withold 10% for validation */
run;
data train; set tinder2;
if Selected = 0;
run;
data test; set tinder2;
if Selected = 1;
run;
/* Look at scatterplot matrix */
proc sgscatter data=train;
matrix genuine COL2 COL3 COL4 COL5 COL6/
markerattrs=(symbol=CIRCLEFILLED size=6pt);
title1 ’Tinder Genuineness Predictors’;
run;
proc sgscatter data=train;
matrix genuine COL7 COL8 COL9 COL10 COL11/
markerattrs=(symbol=CIRCLEFILLED size=6pt);
title1 ’Tinder Genuineness Predictors’;
run;
proc sgscatter data=train;
matrix genuine COL12 COL13 COL14 COL15 COL16/
markerattrs=(symbol=CIRCLEFILLED size=6pt);
title1 ’Tinder Genuineness Predictors’;
run;
/* Look at crude initial model. */
proc reg data=train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS Residual);
```

```
id _n_;
model genuine = COL1-COL41 / vif collin;
store regModel;
output out=trainout r=resid p=pred;
run;
%resid_num_diag(dataset=trainout, datavar=resid, label='residual', predvar=pred,
predlabel='predicted');
data train1;
set train;
if _n_ ne 2;
if _n_ ne 4;
run;
proc reg data=train1 plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS Residual);
id _n_;
model genuine = COL1-COL41 / vif collin;
store regModel1;
output out=train1out r=resid p=pred;
run;
/* backward elimination */
proc reg data=train1;
model genuine = COL1-COL41
/ selection=backward slstay=.10;
title1 'Backward Elimination';
run;
/*stepwise selection*/
proc reg data=train1;
model genuine = COL1-COL41
/ selection=stepwise slentry=.10 slstay=.10;
title1 'Stepwise Selection';
```

```
run;
/* lasso */
proc glmselect data=train1 plots=(criterion ase);
model genuine = COL1-COL41
/ selection=lasso(adaptive choose=sbc stop=none);
output out=outlasso p=predlasso;
run;
/* Fit one model with variables selected without interactions. */
proc reg data=train1;
model genuine = COL2 COL5 COL7 COL8 COL9 COL11 COL14 COL30 COL31 COL32 COL33;
store regModel2;
run;
/* Define interactions. */
data train1; set train1;
COL2_5=COL2*COL5;
COL2_7=COL2*COL7;
COL2_8=COL2*COL8;
COL2_9=COL2*COL9;
COL2_11=COL2*COL11;
COL2_14=COL2*COL14;
COL5_7=COL5*COL7;
COL5_8=COL5*COL8;
COL5_9=COL5*COL9;
COL5_11=COL5*COL11;
COL5_14=COL5*COL14;
COL7_8=COL7*COL8;
COL7_9=COL7*COL9;
COL7_11=COL7*COL11;
COL7_14=COL7*COL14;
```

```
COL8_9=COL8*COL9;
COL8_11=COL8*COL11;
COL8_14=COL8*COL14;
COL9_11=COL9*COL11;
COL9_14=COL9*COL14;
COL11_14=COL11*COL14;
run;
/*Test individual interactions */
proc reg data=train1;
model genuine = COL2 COL5 COL7 COL8 COL9 COL11 COL14 COL30 COL31
COL32 COL33 COL2_9;
run;
/*Fit a model using the significant interaction terms*/
proc reg data=train1;
model genuine = COL2 COL5 COL7 COL8 COL9 COL11 COL14 COL30 COL31
COL32 COL33 COL2_9 COL8_11 COL9_11;
output out=train1out r=resid p=pred;
run;
proc reg data=train1;
model genuine = COL2 COL5 COL7 COL8 COL9 COL11 COL14 COL30 COL31
COL32 COL33 COL2_9 COL8_11 COL9_11;
store regModel3;
run;
%macro resid_num_diag(dataset,datavar,label (…)
%resid_num_diag(dataset=train1out, datavar=resid, label='residual', predvar=pred,
predlabel='predicted');
/* Fit a model with NO variables */
proc reg data=train1 noprint;
model genuine = ;
```

```
store regModel4;
run;
/* Calculate MSPR */
/* Define interactions. */
data test; set test;
COL2_9=COL2*COL9;
COL8_11=COL8*COL11;
COL9_11=COL9*COL11;
;
run;
proc plm restore=regModel3;
score data=test out=newTest3 predicted;
run;
data newTest3; set newTest3;
ASE = (Genuine - Predicted)**2;
run;
proc means data = newTest3;
var ASE;
run;
proc plm restore=regModel4;
score data=test out=newTest4 predicted;
run;
data newTest4; set newTest4;
ASE = (Genuine - Predicted)**2;
run;
proc means data = newTest4;
var ASE;
run;
```