SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Modeling Machiavellianism

- Predicting Scores with Fewer Factors

USERS PROGRAM

Modeling Machiavellianism: Predicting Scores with Fewer Factors

Patrick Schambach and Michael Frankel

Kennesaw State University Department of Statistics and Analytical Sciences

ABSTRACT

Prince Niccolo Machiavelli said things on the order of, “The promise given was a necessity of the past: the word broken is a necessity of the present.” His utilitarian philosophy can be summed up by the phrase, “The ends justify the means.” As a personality trait, Machiavellianism is characterized by the drive to pursue one’s own goals at the cost of others. In 1970, Richard Christie and Florence L. Geis created the MACH-IV test to assign a MACH score to an individual, using 20 Likert Scaled questions. The purpose of this study was to build a regression model that can be used to predict the MACH score of an individual using fewer factors. Such a model could be useful in screening processes where personality was considered, such as in job-screening, offender profiling, or online dating sites. The research was conducted on a dataset from an online personality test similar to the MACH-IV test. It was hypothesized that a statistically significant model exists that can predict a average MACH scores for individuals with similar factors. It was also hypothesized that an individual could be classified as a MACH personality based on a logistic regression model. These hypotheses were accepted.

METHODS

The dataset used in this study was found on http://personality-testing.info, and is an online version of the MACH-IV test created by Christie and Geis. The original dataset consisted of 24 variables and 13,156 observations. All of the data was self-reported, except for the duration if the test, in seconds. The dataset was analyzed using SAS Studio.

We first cleaned and trimmed the dataset to remove all observations with nonresponses and invalid entries. Only completed surveys were analyzed. We also removed extreme outliers in test duration, so only test durations longer than 20 seconds were considered. All responses with ages under 18 were also removed. Our trimmed dataset consisted of 24 variables and 10,937 observations.

We then needed to identify potential predictors of score. First, we used variable clustering, where variables that are similar to each other are removed. Using proc varclus, we identified four clusters. The variables with the lowest 1-R^2 values were chosen. They were Q6, Q13, Q14, and test duration. When Score was modeled as a function of these variables, the R^2 value for the model was 71.19%. We then attempted forward and backward selection to choose predictors. However, both methods simply resulted in all variables being selected, so we utilized stepwise regression and found that when five variables were used, an R^2 value of 83.68% was obtained, which we deemed sufficient. Thus, we chose these five variables as our predictors. They were Q1, Q6, Q9, Q12, and Q13. We then checked for interactions between our predictors and found that there were two that were significant, Q1*Q9 and Q6*Q12. Including these interactions would have only raised our R^2 to 83.76%, but the coefficients were considered too small to make much practical contribution and were omitted from the model.

Note that since SAS suppresses all plots with over 5,000 observations, we selected a random sample of 1,000 observations to generate the visuals.

RESULTS

Figure 1: Linear Model with Clustered Predictors

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1336401	445467	9003.32	<.0001
Error	10933	540944	49.47806		
Corrected Total	10936	1877345			

Root MSE	7.03406	R-Square	0.7119
Dependent Mean	65.48185	Adj R-Sq	0.7118
Coeff Var	10.74201		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	72.38688	0.29410	246.13	<.0001	0
Q13	1	4.25605	0.05397	78.86	<.0001	1.13289
Q6	1	-4.46602	0.05636	-79.23	<.0001	1.19705
Q14	1	-3.51026	0.07406	-47.40	<.0001	1.19202

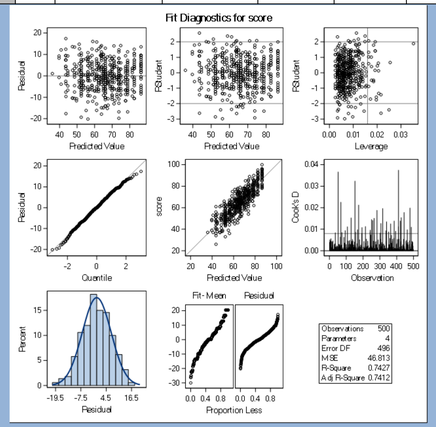
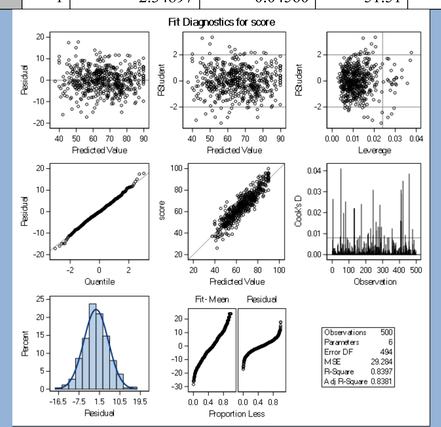


Figure 2: Linear Model with Stepwise Selection

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1571004	314201	11211.5	<.0001
Error	10931	306341	28.02496		
Corrected Total	10936	1877345			

Root MSE	5.29386	R-Square	0.8368
Dependent Mean	65.48185	Adj R-Sq	0.8367
Coeff Var	8.08447		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60.37110	0.31947	188.97	<.0001
Q1	1	2.36684	0.04841	48.89	<.0001
Q6	1	-2.92041	0.04656	-62.73	<.0001
Q9	1	-2.98785	0.04957	-60.28	<.0001
Q12	1	2.35958	0.04662	50.61	<.0001
Q13	1	2.34897	0.04560	51.51	<.0001



Modeling Machiavellianism: Predicting Scores with Fewer Factors

Patrick Schambach and Michael Frankel

Kennesaw State University Department of Statistics and Analytical Sciences

METHODS CONTINUED

We then further attempted to model Machiavellianism using a logistic regression and a binary variable that would identify an individual as being Machiavellian or not. A new variable called isMach was created where an individual scoring above the 75th percentile was classified as a 1, or an event, and those below were classified as a 0, or non event. We then attempted to identify predictors of isMAch. Before fitting our model, we randomly split the dataset into two smaller sets. We selected 65% of the data for fitting the model and the remaining 35% for validating it. In the two subsets, the distribution of isMAch was retained. To find our predictors, we first attempted backward and forward selections, but all variables except age, gender, and test duration were selected. We then used the clustered variables from earlier and, once again, retained Q6, Q12, and Q16. Test duration was not a significant predictor of isMach. Our model with clustered variables yielded a c-statistic of 0.917. We then compared this to a model with only three variables chosen by stepwise selection. The procedure yielded the predictors Q9, Q13, and Q10. Our new model also had a higher c-statistic of 0.935. Thus, the model using stepwise selection for variable selection had a higher number of concordant pairs than the one using variable clustering, so we opted to use the former.

A table of potential probabilities of a Machiavellian personality was generated and analyzed. We chose a probability of 0.200 for a Machiavellian because if we lower the probability, the false positives increase 10%. Conversely, if we raise the probability, the sensitivity and percent correct decrease.

A Machiavellian being misclassified as a non-Machiavellian, would result in a type-I error. In this case, someone who should be screened would not be screened. A non-Machiavellian being classified as a Machiavellian would result in a type-II error. In this case, someone would be unnecessarily screened who should not have been. The more detrimental case would depend on the whether or not a Machiavellian personality is considered desirable. If a Machiavellian personality would be detrimental, a type-I error should be minimized.

RESULTS

Figure 3: Logistic Model with Stepwise Selection

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9846	0.1774	30.8206	<.0001
Q9	1	-0.9698	0.0372	681.1474	<.0001
Q13	1	0.9932	0.0383	671.3771	<.0001
Q10	1	-0.8876	0.0405	479.2625	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.2	Somers' D	0.870
Percent Discordant	6.2	Gamma	0.875
Percent Tied	0.6	Tau-a	0.341
Pairs	9902309	c	0.935

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Q9	1.0000	0.379	0.353	0.408
Q13	1.0000	2.700	2.504	2.911
Q10	1.0000	0.412	0.380	0.446

Figure 4: Table of Probability levels

Classification Table										
	Correct		Incorrect		Percentages					
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	1901	0	5209	0	26.7	100.0	0.0	73.3	.
0.100	1801	3768	1441	100	78.3	94.7	72.3	44.4	2.6
0.200	1680	4360	849	221	85.0	88.4	83.7	33.6	4.8
0.300	1673	4360	849	228	84.9	88.0	83.7	33.7	5.0
0.400	1466	4805	404	435	88.2	77.1	92.2	21.6	8.3
0.500	1466	4805	404	435	88.2	77.1	92.2	21.6	8.3
0.600	1135	5033	176	766	86.8	59.7	96.6	13.4	13.2
0.700	1096	5043	166	805	86.3	57.7	96.8	13.2	13.8
0.800	744	5156	53	1157	83.0	39.1	99.0	6.6	18.3
0.900	606	5179	30	1295	81.4	31.9	99.4	4.7	20.0
1.000	0	5209	0	1901	73.3	0.0	100.0	.	26.7

Figure 5: Table of Actual vs Predicted Mach Type

Table of isMach by preds			
isMach	preds		
Frequency Percent	0	1	Total
0	2352 61.46	452 11.81	2804 73.27
1	122 3.19	901 23.54	1023 26.73
Total	2474 64.65	1353 35.35	3827 100.00

CONCLUSIONS

MACH Score can be predicted somewhat reliably using only a quarter of the questions required by the MACH-IV test. In fact, 83.68% of the variation in score can be explained by only five questions. A MACH personality can also be reliably predicted using only three questions. These questions are as follows:

Q1: “Never tell anyone the real reason you did something unless it is useful to do so.”

Q6: “Honesty is the best policy in all cases.”

Q9: “All in all, it is better to be humble and honest than to be important and dishonest.”

Q10: “When you ask someone to do something for you, it is best to give the real reasons for wanting it rather than giving reasons which carry more weight.”

Q12: “Anyone who completely trusts anyone else is asking for trouble.”

Q13:The biggest difference between most criminals and other people is that the criminals are stupid enough to get caught.

Our fitted linear model equation became:

$$\widehat{Score} = 60.37 + 2.37(Q1) - 2.92(Q6) - 2.99(Q9) + 2.36(Q12) + 2.35(Q13)$$

Our fitted logistic equation became:

$$Logit\big(P(Y = 1)\big) = 0.9846 - 0.9698(Q9) + 0.9932(Q13) - 0.8876(Q10)$$

Modeling Machiavellianism: Predicting Scores with Fewer Factors

Patrick Schambach and Michael Frankel

Kennesaw State University Department of Statistics and Analytical Sciences

APPENDIX: SAS CODE

```
*set dataset and predictors;
data machtrim3;
set _temp0.machtrim3;
%let inputs = age seconds_elapsed gender Q1 Q2 Q3 Q4 Q5
              Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16
              Q17 Q18 Q19 Q20;

*variable clustering;
proc varclus data = &d1 outtree = tree;
var &inputs;
run;

*LINEAR MODELING;
*linear model from clusters;
proc reg data = &d1;
model score = q13 q6 q14 / vif;
run;

*linear model from stepwise selection;
proc reg data = &d1;
model score = &inputs /
selection=stepwise vif rsquare sle=.015 sls=0.05;
run;
```

```
*LOGISTIC MODELING;
*create binary variable;
data machtrim4;
set machtrim3;
if score >=74 then isMach = 1;
else isMach = 0;
run;
proc sort data = machtrim4;
by isMach;
run;

*seperate dataset for logistic validation;
proc surveyselect data = machtrim4 samprate = 0.35
out = moddevfile seed = 1234567 outall;
strata ismach;
run;

*now split into two datasets;
data train valid;
set moddevfile;
if selected then output train;
else output valid;
run;
```

```
*logistic model with stepwise selection;
proc logistic data = train desc outest= betas outmodel=scoringdata;
model isMach = q9 q13 q10  / selection = stepwise
ctable pprob=(0.1 to 0.3 by 0.02)
LACKFIT RISKLIMITS;
output out = output p = predicted;
score data = valid out=score;
run;

* set cutoff point at 0.200;
data test;
set score;
if p_1 ge 0.20 then preds = 1;
else preds = 0;
run;

*grnerate table;
proc freq data = test;
tables ismach*preds / norow nocol;
run;
```

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL