Tilburg University
MSc Data Science and Society

Statistics & Methodology
April 3, 2020

Group Project | Group 29

Name: Sathya Krishna Sharma Jagannatha
Analytic Role: Data Preparation
Student Number (SNR): 2033987
Administrative number (ANR): u580435

Name: Estée Coenraad
Analytic Role: Inferential Modeling
Student Number (SNR): 2013242
Administrative number (ANR): u225986

Name: Martijn Hooijman
Analytic Role: Predictive Modeling
Student Number (SNR): 2034200
Administrative number (ANR): u996412

Name: Chi Hung
Analytic Role: Write-Up
Student Number (SNR): 2034109
Administrative number (ANR): u677127

Data Preparation

After running the pre-processing script, the data is prepared for further analysis. Figure 1 shows the process of data preparation. The dataset consists of 13156 observations of 179 variables. A selection of 25 variables are selected in order to operate the inferential and predictive modelling. The selection of variables is shown in table 1. Negative values in the data are related to missing or removed values and are treated as missing data by assigning a NA value to them.

Univariate outliers are removed from the data using the boxplot method. This method is an easy way to compare the shapes of distributions, find central tendencies, assess variability and identify outliers. The boxplot method is used to identify possible and probable outliers in all variables of the dataset. Value that falls outside of the inner fence of the boxplot are flagged as a possible outlier and values that falls outside of the outer fence are flagged as a probable outlier. Table 2 shows the number of possible and probable outliers that were found in the selected variables. All probable outliers are removed from the dataset by assigning a NA value to them.

The next step is to analyze and account for the missing values in the data. The proportion of missing data per variable ranges from 0.02% to 17.51% with a mean of 4.64% missing data per variable. The covariance coverage ranges from 77.718% to 99.997%. Multiple imputation is used to deal with the missing values. Before running the multiple imputation algorithm, we converted binary and nominal values to factors and defined the method for multiple imputation for each variable. Using multiple imputation with 10 iterations, 20 imputed datasets are generated. Density plots of the imputed datasets show that the imputation has gone right. It is clear from these plots that ordinal variables are treated as continuous variables. Figure 1 shows the density plots for all 20 datasets.

The last step of the data preparation is to account for multivariate outliers. We used the robust Mahalanobis Squared distances method to detect multivariate outliers. This method uses robust estimates of the central tendencies and dispersion and this makes the measures themselves insensitive to outliers. Multivariate outliers in each of the imputed datasets are identified using a critical probability of 0.99. The number of times that each value is classified as a multivariate outlier in all imputed datasets are summed. Using a removal threshold of 10, all values that occurred 10 times or more often are removed from all imputed datasets using their indexes. This results in 760 values being removed in each imputed dataset.

Predictive modelling

The outcome to explain using predictive modelling is *Satisfaction with life*. In order to operationalize this problem, the corresponding variable V23 (Satisfaction with your life) is chosen as the dependent variable. Besides, a set of independent variables are manually selected from the dataset based on the fact that they would be related to satisfaction of life by human judgement. Table 3 shows an overview of the selected variables.

To find the best model to predict satisfaction with life, we aimed to find the best selection of variables that does not use interaction in the model and the best selection of variables that does use interaction in the model. A loop is created to constructively test models to obtain the best performing model based on the lowest cross validation error. At first the simplest set of models is generated by combining the dependent variable with each of the independent variables in an individual model. The cross-validation error for all models is obtained using a 10-fold cross-validation procedure and the model with the lowest error is kept as the best model. In the next step, the current best model is combined with each of the resulting independent variables, leading to a new set of models. This procedure is continued until no decrease in cross validation occurred. Using the loop procedure, we obtained the best performing model that uses interaction and the best performing model that does not use interaction. The best model with interaction is the following: "V23 ~ V59 * V10 * V55 * V11 * V181 * V24", having a cross validation error of 2.140. The best model without interaction is the following: "V23 ~ V59 + V10 + V55 + V248 + V11 + V24 + V181", having a cross validation error or 2.120. The results of these two procedures are shown in table 4 and table 5.

The best model of the two models that are obtained using the loop procedure is the model without interaction. This model is selected as our final model. The test-set prediction error for this model is MSE = 2.076.

Inferential Modeling

For inferential modelling, we chose the following question:
“How do gender politics relates to economic beliefs? “.

To conduct an inferential analysis, we started to convert the question into a hypothesis:
H0: Gender politics is related to economic beliefs
H1: Gender politics is not related to economic beliefs

To build the model, variables have been chosen from the codebook. The variables have been manually chosen based on the type of questions which are related to the hypothesis. The variables that have been used to conduct the analysis are the following:
V7, V8, V45, V48, V53, V81, V96, V121, V139, V239 and V240. (See appendices for: “List Variables”)

To answer the hypothesis, several linear models have been used to determine if the hypothesis should be accepted or rejected. To begin with, the following questions are created based on the variables which will help to answer the main question.
1)H0: Gender politics does play a role in job scarcity
2)H0: Gender politics does not relate to income inequality
3)H0: Economic belief is related to the importance of economic growth
4)H0: Gender politics does not relate if losing my job is important

To test if variables explain a significant part of the variance, several variables are appended to a multiple linear regression model to find out if that would make a significant difference.

The following procedure was taken to find out:

1. Create a regression model
2. Summarize the model
3. Summarize the pooled estimates
4. Pool R squared and adjusted R squared
5. Find the R^2
6. Compute increase in R^2
7. Find the significant increase in R^2 (ANOVA)

The first step for interpreting the multiple regression analysis is to examine the p-value. P-values are obtained by fitting linear models to the multiple imputed datasets and use pool estimates. The outcome of this is shown in the appendix. The level of statistical significance is determined when the p-value is equal or less than ≤.05. Also, we decided to follow up with a standard dummy coding for the linear models.

Model 1

H0: Gender politics does play a role in job scarcity

H1: Gender politics does not play a role in job scarcity

fit1.1	<-lm.mids(V45 ~ V7,data = miceOut3)
fit1.2	<-lm.mids(V45 ~ V7+ V48,data = miceOut3)
fit1.3	<-lm.mids(V45 ~ V7+ V53,data = miceOut3)
fit1.4	<-lm.mids(V45 ~ V7+ V48 + V53,data = miceOut3)

At the different summaries for Model 1 (See: Appendix Table 6) is shown that the p-values for V7 and V48 are $P \geq 0.05$ when these variables are combined together, resulting that no effect was observed. For the R^2 , when a variable is added the R^2 increased. However, when an ANOVA test was applied to determine if there is a significant increase in R2. The outcome shown for model fit1.2 ~ fit1.1 and fit1.3 ~ fit 1.2 showed a high p-value = 3.209492e-06. Based on this high number, the conclusion showed that variable V7 and V48 cannot be statistically proven. Further, the R^2 is “high” and does not shows enough prove to approve to accept the hypothesis. Therefore, the null hypothesis will be rejected for Model 1.

Model 2:

H0: Gender politics does not relate to income inequality

H1: Gender politics does relate to income inequality

fit2.1	<-lm.mids(V96 ~ V240, data = miceOut3)
fit2.2	<-lm.mids(V96 ~ V240+V45, data = miceOut3)
fit2.3	<-lm.mids(V96 ~ V240+V45+V7, data = miceOut3)
fit2.4	<-lm.mids(V96 ~ V240+V45+V7+V139,data = miceOut3)

For the summarized models (table 7), the variable V240 is odd due to the high p-value>.05. For the four models the p-value was, 2.232, 1.175, 8.548 and 2.0343. V240 is a dummy variable that stands for the gender of the person and the V2402 can be translated to the sex$Female. When we pool the R^2 and the adjusted R^2, the outcomes shown that the adjusted R^2 are slightly lower than the pooled R^2. For example, the comparison with the R^2 and adjusted R^2 for the model fit2.1 R2 = 0.00158 and the adjusted R^2 = 0.00149. Also, the adjusted R^2 increase only when a new variable is added and improves the model. To come to a conclusion for the model, an ANOVA - test has been used to check if the variables are significant to each other. The result shown that the p-value <.05. With this outcome and the adjusted R^2 values we will reject the null hypothesis for Model 2

Model 3

H0: Economic belief is related to the importance of economic growth
H1: Economic belief is not related to the importance of economic growth

fit3.1 <-lm.mids(V8 ~ V81, data = miceOut3)
fit3.2 <-lm.mids(V8 ~ V81 +V121, data = miceOut3)
fit3.3 <-lm.mids(V8 ~ V81 +V121 +V97, data = miceOut3)
fit3.4 <-lm.mids(V8 ~ V81 +V121 +V97 +V239,data +miceOut3)

In the summaries (table 8), the variable V81, divided in V812 and V813 responds differently on the model. To take model 3.3 the p-value for p(V812) = 5.821>.05 and p(V813) = 2.859 >.05 and V121 is the only constant variable that shows a significance difference p<.05. For the R^2 and the adjusted R^2 it is the same as in model 2. The adjusted R^2 slightly increase when a term is added. The outcome of the ANOVA test, shows the increase between fit3.3-fit3.2 is p= .118. Considering the findings of this analysis there is not enough prove to accept the null hypothesis.

Model 4

H0: Gender politics does not relate if losing my job is important
H1: Gender politics relate if losing my job is important

fit4.1 <-lm.mids(V181 ~ V45, data = miceOut3)
fit4.2 <-lm.mids(V181 ~ V45 + V240, data = miceOut3)
fit4.3 <-lm.mids(V181 ~ V45 + V240+ V8, data = miceOut3)

Firstly, the summaries of the pooled estimates (table 9) indicates that the p-value for all variables are equal to zero P≤.05 except for V240(~female). Secondly, for the R^2 and the adjusted R^2 after the second variable (V181 ~ V45 + V240) the p-value is increased by p = .0934. This decreased variable is not improving the model by sufficient amount and this is also shown in the ANOVA test [(F 851.4450,1) =1.003, p = 0.317]. To conclude, the previously results from the R^2 and the adjusted R^2 does not shows enough prove to accept the null hypothesis.

Conclusion

To come to a conclusion the data analysis shows that there is not enough evidence to accept the main hypothesis. Therefore, the null hypothesis is rejected, meaning that gender politics is not significantly related to economic beliefs.

Appendix

Table 1: " List Variables " : Variables used for inferential and predictive modelling.

Variables	Question	Scale
V7	Important in life: Politics	1.- Very important 2.- Rather important 3.- Not very important 4.- Not at all important
V8	Important in life: Work	
V10	Feeling of happiness	1.- Very happy 2.- Rather happy 3.- Not very happy 4.- Not at all happy
V11	State of health (subjective)	1.- Very good 2.- Good 3.- Fair 4.- Poor
V23	Satisfaction with your life	1.- Completely dissatisfied 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Completely satisfied
V24	Most people can be trusted	1.- Most people can be trusted 2.- Need to be very careful
V45	When jobs are scarce, men should have more right to a job than women	1.- Agree 2.- Neither 3.- Disagree
V46	When jobs are scarce, employers should give priority to people of this country over immigrants	1.- Agree 2.- Neither 3.- Disagree
V48	Having a job is the best way for a woman to be an independent person.	1.- Agree 2.- Neither 3.- Disagree
V53	On the whole, men make better business executives than women do	1.- Agree strongly 2.- Agree 3.- Disagree 4.- Strongly disagree

V55	How much freedom of choice and control over own life	1.- No choice at all 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- A great deal of choice
V59	Satisfaction with financial situation of household	1.- Completely dissatisfied 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Completely satisfied
V60	Aims of country: first choice	1.- A high level of economic growth 2.- Making sure this country has strong defense forces 3.- Seeing that people have more say about how are done at their jobs and in their communities 4.- Trying to make our cities and countryside more beautiful
V81	Protecting environment vs. Economic growth	1.- Protecting the environment should be given priority, even if it causes slower economic growth and some loss of jobs 2.- Economic growth and creating jobs should be the top priority, even if the environment suffers to some Extent 3.- Other answer
V96	Income equality	1.- Incomes should be made more equal 2.- 2 3.- 3 4.- 4

		5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- We need larger income differences as incentives for individual effort
V97	Private vs state ownership of business	1.- Private ownership of business and industry should be increased 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- Government ownership of business and industry should be increased
V121	Confidence: Banks	1.- A great deal 2.- Quite a lot 3.- Not very much 4.- None at all
V123	Confidence: Women's organizations	1.- A great deal 2.- Quite a lot 3.- Not very much 4.- None at all
V139	Democracy: Women have the same rights as men.	1.- Not an essential characteristic of democracy 2.- 2 3.- 3 4.- 4 5.- 5 6.- 6 7.- 7 8.- 8 9.- 9 10.- An essential characteristic of democracy
V143	Thinking about meaning and purpose of life	1.- Often 2.- Sometimes 3.- Rarely 4.- Never
V181	Worries: Losing my job or not finding a job	1.- Very much 2.- A great deal 3.- Not much

		4.- Not at all
V239	Scale of incomes	1.- Lower step 2.- second step 3.- Third step 4.- Fourth step 5.- Fifth step 6.- Sixth step 7.- Seventh step 8.- Eigth step 9.- Nineth step 10.- Tenth step
V240	Sex	1.- Male 2.- Female
V242	Age	10-29.- Up to 29 30-49.- 30-49 50-102.- 50 and more
V248	Highest educational level attained	1.- No formal education 2.- Incomplete primary school 3.- Complete primary school 4.- Incomplete secondary school: technical/ vocational type 5.- Complete secondary school: technical/ vocational type 6.- Incomplete secondary school: university-preparatory type 7.- Complete secondary school: university-preparatory type 8.- Some university-level education, without degree 9.- University - level education, with degree

Table 2: Variables with univariate outliers

Variable	**Questions**	**Outliers**	
V8	Important in life: Work	Possible	697
		Probable	0
V10	Feeling of happiness	Possible	181
		Probable	0
V139	Democracy: Women have the same rights as men	Possible	846
		Probable	358

Table 3: Selected variables for predictive modelling.

Variables	Question	Type of variables
V10	Feeling of happiness	independent
V11	State of health (subjective)	independent
V23	Satisfaction with your life	dependent
V24	Most people can be trusted	independent
V55	How much freedom of choice and control over own life	independent
V59	Satisfaction with financial situation of household	independent
V143	Thinking about meaning and purpose of life	independent
V181	Worries: Losing my job or not finding a job	independent
V248	Highest educational level attained	independent

Table 4: Results of loop procedure to obtain the best model without interaction.

	Testing Model	**Add predictor variables**	**Best predictor variable**	**After testing the models without interaction, we found the best model was:**	**Cross validation error**
1	V23	V10, V11, V24, V55, V59, V143, V181, V248	V59	V23 ~ V59	2.942
2	V23 ~V59	V10, V11, V24, V55, V143, V181, V248	V10	V23 ~ V59 +V10	2.441
3	V23 ~ V59+10	V11, V24, V55, V143, V181, V248	V55	V23 ~ V59 +V10 +55	2.184
4	V23 ~ V59+10+55	V11, V24, V143, V181, V248	V248	V23 ~ V59+V10+55 +V248,	2.155
5	V23 ~ V59+10+V55+V248	V11, V24, V143, V181	V11	V23 ~ V59+V10+55+V248 +V11	2.133
6	V23 ~ V59+10+V55+V248+V11	V24, V143, V181	V181	V23 ~V59+V10+55+V248+V11 +V181	2.129
7	V23 ~V59+V10+55+V248+V11+V181	V24, V143	V24	V23 ~V59+V10+55+V248+V11 +V181+V24,	**2.125**

Table 5: Results of loop procedure to obtain the best model with interaction.

	Testing Model	**Multiply the relevant variables**	**Best predictor variable**	**After testing the models with interaction, we found the best model was:**	**Cross validation error (CVE).**
1	V23	V10, V11, V24, V55, V59, V143, V181, V248	V59	V23 ~ V59	2.942
2	V23 ~V59	V10, V11, V24, V55, V143, V181, V248	V10	V23 ~ V59 * V10	2.427
3	V23 ~ V59 * V10	V11, V24, V55, V143, V181, V248	V55	V23 ~ V59 * V10 * V55	2.181
4	V23 ~ V59 * V10 * V55	V11, V24, V143, V181, V248	V11	V23 ~ V59 * V10 * V55 * V11	2.147
5	V23 ~ V59 * V10 * V55 * V11	V24, V143, V181, V248	V181	V23 ~ V59 * V10 * V55 * V11 * V181	2.141
6	V23 ~ V59 * V10 * V55 * V11 * V181	V24, V143, V181	V24	V23 ~ V59 * V10 * V55 * V11 * V181 * V24	**2.140**

Table 6: Results of model 1.

Table 6.1: Summary Pooled Estimates Model 1.1

Term	**Estimate**	**Std.Error**	**Statistic**	**df**	**P.Value**
(Intercept)	2.096899036	0.024839696	84.417258	2535.119	0.0000000
V7	0.009744872	0.008940048	1.090025	2070.072	0.02758292

Table 6.2: Summary Pooled Estimates Model 1.2

Term	**Estimate**	**Std.Error**	**Statistic**	**df**	**P.Value**
(Intercept)	2.01386538	0.030452627	66.131089	2053.355	0.000000e+00
V7	0.01047252	0.008935232	1.172048	2050.091	2.413139e-01
V48	0.05278686	0.011247757	4.693101	1324.186	2.970069e-06

Table 6.3: Summary Pooled Estimates Model 1.3

Term	**Estimate**	**Std.Error**	**Statistic**	**df**	**P.Value**
(Intercept)	0.99438868	0.030560140	32.538747	2817.024	0.00000000
V7	0.01765696	0.007980824	2.212423	2573.131	0.02702497
V53	0.40833770	0.007867070	51.904673	1739.607	0.00000000

Table 6.4: Summary Pooled Estimates Model 1.4

Term	Estimate	Std.Error	Statistic	df	P.Value
(Intercept)	0.92635605	0.034207726	27.080317	2505.266	0.000000e+00
V7	0.01825579	0.007975110	2.289096	2575.364	2.215436e-02
V48	0.04423166	0.10093204	4.382321	1356.172	1.264728e-05
V53	0.40776542	0.007873158	51.791852	1660.430	0.000000e+00

Table 7: Results of model 2.

Table 7.1: Summary Pooled Estimates Model 2.1

Term	Estimate	Std.Error	Statistic	df	P.Value
(Intercept)	3.9649897	0.03409570	116.290021	5399.684	0.000000e+00
V240	-0.2094605	0.04948499	-3.233011	2169.038	2.401947e-05

Table 7.2: Summary Pooled Estimates Model 2.2

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	3.2940026	0.06655337	49.494154	2187.690	0.000000e+00
V240	-0.3066261	0.05023857	-6.103401	1693.641	1.284373e-09
V45	0.3387898	0 .02893783	11.707505	1657.215	0.000000e+00

Table 7.3: Summary Pooled Estimates Model 2.3

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	3.9169470	0.09415691	41.600207	4155.329	0.000000e+00
V240	-0.2494062	0.05060738	-4.928258	1552.568	9.184269e-07
V45	0.3359963	0.02875744	11.683805	1810.964	0.000000e+00
V7	-0.2444143	0.02699368	-9.054503	2350.369	0.000000e+00

Table 7.4: Summary Pooled Estimates Model 2.4

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	4.9774947	0.13381109	37.197923	4071.134	0.000000e+00
V240	-0.2397411	0.05043249	-4.753703	1480.659	2.034255e-06
V45	0.3789874	0.02901878	13.060072	1557.044	0.000000e+00
V7	-0.2411472	0.02676922	-9.008377	2654.304	0.000000e+00
V139	- 0.1345802	0.01231615	-10.927131	2138.508	0.000000e+00

Table 8: Results of model 3

Table 8.1 Summary Pooled Estimates Model 3.1

Term	Estimate	Std.Error	Statistic	df	P.Value
(Intercept)	1.62966938	0.01036306	157.257554	2312.2583	0.00000000
V81 ~2	0.09497705	0.01683979	5.640039	719.0530	2.444473e-08
V81~3	0.18068745	0.04646322	3.888828	975.4801	1.075527e-04

Table 8.2: Summary Pooled Estimates Model 3.2

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	1.27925527	0.020904872	61.194121	1969.8676	0.000000e+00
V81~ 2	0.07073124	0.016744426	4.224166	624.5207	2.755603e-05
V81~3	0.10123130	0.046210735	2.190645	843.1634	2.875064e-02
V121	0.15696335	0.008390546	18.707168	912.1113	0.000000e+00

Table 8.3: Summary Pooled Estimates Model 3.3

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	1.305147393	0.026181313	49.850342	2653.2485	0.000000e+00
V81~ 2	0.068221697	0.016850719	4.048593	605.9404	5.820995e-05
V81~3	0.101294671	0.046194108	2.192805	848.5746	2.859223e-02
V121	0.156670351	0.008400085	18.651043	892.4459	0.000000e+00
V97	-0.004562689	0.002914655	-1.565431	945.1055	1.178167e-01

Table 8.4: Summary Pooled Estimates Model 3.4

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	1.350147464	0.032977651	40.941286	5677.7337	0.000000e+00
V81~ 2	0.068398921	0.016835595	4.062756	614.4503	0.0000547831
V81~3	0.102411258	0.046209098	2.216257	840.9145	0.0269408440
V121	0.156670351	0.008406690	18.516379	907.7678	0.000000e+00
V97	0.005119139	0.002923290	1.751157	970.0383	0.0802349584
V239	-0.008600591	0.003929411	-2.188774	6680.2977	0.0286477887

Table 9: Results of model 4.

Table 9.1: Summary Pooled Estimates Model 4.1

Term	Estimate	Std.Error	Statistic	df	P.Value
(Intercept)	2.0254002	0.02796358	72.42994	931.0152	0
V45	0.1667689	0.01207160	13.81498	1192.9869	0

Table 9.2: Summary Pooled Estimates Model 4.2

Term	Estimate	Std.Error	Statistic	df	P.Value
(Intercept)	2.01869969	0.02864537	70.472102	948.5338	0.00000000
V45	0.16486560	0.01221491	13.497075	1259.3789	0.00000000
V240	0.02168714	0.02123129	1.021471	1242.5042	0.3072303

Table 9.3: Summary Pooled Model 4.3

Term	Estimate	Std.Error	Statistic	DF	P.Value
(Intercept)	1.49107937	0.03299927	45.1852217	745.7735	0.0000000
V45	0.13306285	0.01180144	11.2751395	1193.9683	0.0000000
V240	-0.02028256	0.02052460	-0.9882075	1111.1917	0.3232662
V8	0.36808211	0.01236437	29.7695731	776.2551	0.0000000

Figure 1: Density plots of imputed datasets