Statistics and Methodology Assignment

Group 39

Dennis Hokke 2045370/u589426: Data preparation and write-up
Arati Sharma: 2047448/ u289062: Inferential modeling and write-up
Yaohua Liu : 2041746/ u844976: Predictive modeling and write-up

Data Preparation

Missing data processing

At first observation of the data it appears as though there is no missing data in the dataset. However, there are an immense amount of negative values. These negatives numbers could mess with our further analyses and interpretations. Therefore the first step was to transform those negative values into NA labels (missing data). After transforming all negative values in NA's we performed several tests to see how many NA values there are, how many observed values,we ran a missing data pattern function, covariance coverage, the range of the covariance coverage and problematic variables. We then extracted the variables we would use to answer the questions and used imputation on the missing data within these variables. We used several imputation methods, such as norm, norm.boot, logreg and pmm. Ultimately we used the predictive mean matching (pmm) imputation which is a method used for numerical values. We created a pdf for traceplots and density plots to determine whether the imputation converges well, see Appendix A and B.

Outlier detection

To detect any outlier in the imputed dataset we put all the lists -a total of 10- into a single list and used mdOutliers to detect any outliers. Then we created a table which counted each time an observation was flagged as an outlier, see appendix C

Choosing the dataset

Ultimately since we need one dataframe to use for analysis in our questions, we picked list number 5 of the 10 available lists from our imputed data. Mainly, because list 5 seems to be the most successful in fulfilling NA labels with an appropriate value.

Inferential Modeling

The question we chose to answer for the inferential modeling task is: Are conservative attitudes good or bad for your psychological well-being? Therefore, we want to understand if having conservative attitudes and beliefs have a positive or a negative impact on the overall happiness of an individual.

In order to answer this question, we decided to use a multiple regression model with one dependent variable and multiple independent variables.

We selected the variable 'V10: Feeling of happiness' as our dependent variable as we expect it to be a good proxy for the psychological well-being of an individual.

When selecting the independent variables to include in our model, we considered various factors that could assess conservative or liberal attitudes. Specifically, we considered attitudes towards gender roles, views about religion and science, feelings about people who are different i.e a different nationality, race, religion. Negative feelings towards gender roles, science and people who were different were taken to be conservative. For example, for the question: When

jobs are scarce, men should have more right to a job than women, participants who responded with 1: Agree, were considered to be more conservative than those who answered 2: Neither or 3: Disagree.

We used two multiple regression models to do the inferential modeling task. In the first model we only included variables that represent conservative attitudes as independent variables. In the second model we added variables that do not represent conservative attitudes but have an impact on the overall feeling of happiness of an individual as controls. These variables include the respondent's marital status, their financial situation in the household and their health state.

The final variables selected for the first model (uncontrolled) are:

Table 1. List of variables for the first model (uncontrolled)

V45 : When jobs are scarce, men should have more right to a job than women
V47 : If a woman earns more money than her husband, it's almost certain to cause problems
V52 : A university education is more important for a boy than for a girl
V139 : Democracy: Women have the same rights as men.
V46 : When jobs are scarce, employers should give priority to people of this country over immigrants.
V107 : How much you trust: People of another nationality

The additional variables used in the second model as controls are:

Table 2. List of control variables for the second model (controlled)

V57 : Marital Status
V59 : Satisfaction with financial situation of household
V11 : State of health (subjective)

The inferential modeling was conducted using the two following multiple regression equations:

Model 1 (uncontrolled)
Psychological well being = $\beta_0 + \beta_1$ V45+ β_2V47 + β_3 V52 + β_4 V139 +β_5 V46 + β_6 V107

Model 2 (controlled)
Psychological well being = $\beta_0 + \beta_1$ V45 + β_2V47 + β_3 V52 + β_4 V139 + β_5 V46 + β_6 V107 + β_7 V57 + β_8 V59 + β_9 V11

The results of the two models are summarized in the following tables:
Model 1:

Table 3. Results model 1 (uncontrolled)

MODEL INFO:
Observations: 13156
Dependent Variable: V10
Type: OLS linear regression

MODEL FIT:
$F(6,13149) = 18.10$, $p = 0.00$
$R^2 = 0.01$
Adj. $R^2 = 0.01$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	1.90	0.04	48.77	0.00
V45	0.02	0.01	2.01	0.04
V47	-0.04	0.01	-5.43	0.00
V52	0.00	0.01	0.50	0.62
V139	-0.00	0.00	-0.36	0.72
V46	-0.04	0.01	-5.05	0.00
V107	0.04	0.01	5.78	0.00

As shown in the table above, model one does not capture the variance in the dependent variable well. The R^2 of 0.01 shows that only 1% of the variation in the dependent variable is explained by the independent variables. This suggests that conservative attitudes do not have a strong correlation with the overall happiness an individual feels.

This model shows that the variable V45 (When jobs are scarce, men should have more right to a job than women) has a statistically significant impact on overall happiness (est: 0.02, t: 2.01, p: 0.04). Since the estimated coefficient for this variable is positive, it suggests that respondents who agreed that when the jobs are scarce, men should have priority over women when searching for jobs (more conservative) are likely to be happier than those who do not agree. Similarly, variable V107 (How much you trust: People of another nationality), is also significant with a positive coefficient (est: 0.04, t :5.78 and p <0.05). In this case, participants who said they trusted people of another nationality (less conservative), were likely to be happier. The coefficient on two other variables, V47 (If a woman earns more money than her husband, it's

almost certain to cause problems) and V46 (When jobs are scarce, employers should give priority to people of this country over immigrants) are negative and statistically significant (V47: est: -0.04, t: -5.43, p: <0.05 and V46: est: -0.04, t:-5.05, p<0.05). However, once the model is adjusted so that the variance in the dependent variable is better explained by the model, we do not expect these variables to be significant.

Model 2

Table 4. Results of Model 2 (controlled)

MODEL INFO:
Observations: 13156
Dependent Variable: V10
Type: OLS linear regression

MODEL FIT:
$F(9,13146) = 461.30$, $p = 0.00$
$R^2 = 0.24$
Adj. $R^2 = 0.24$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	1.63	0.04	40.73	0.00
V45	0.01	0.01	1.79	0.07
V47	-0.01	0.01	-0.99	0.32
V52	-0.00	0.01	-0.47	0.64
V139	0.00	0.00	0.22	0.83
V46	-0.01	0.01	-1.66	0.10
V107	0.01	0.01	1.51	0.13
V57	0.03	0.00	9.88	0.00
V59	-0.07	0.00	-30.38	0.00
V11	0.27	0.01	44.26	0.00

As shown on the tables above, the second model with additional control variables is significantly better at explaining the variation in the dependent variable than the first model ($\Delta R^2$0.23,). Therefore, this suggests that the additional variables included in this model that capture a person's health state (V11: est: 027, se: 0.01, t: -30, p: <0.01), marital status(V57: est: 0.03, se:0.00, t:9.88, p <0.01) and satisfaction with financial status (est: -0.7, se: 0.00, t: -30.38, p <0.01) have a significant impact on the overall happiness an individual feels with 99% statistical

significance. There is a higher probability that respondents who were in a good state of health, were married, and were satisfied with their financial status were likely to be happier than those who weren't. Additionally, the variables that were significant in the first models with 99% significance, are no longer statistically significant with 99% significance.

Predictive Modeling

For our predictive modeling task, we wanted to design a model that would predict a person's satisfaction with life. The dependent variable we picked for this task is 'V23: Satisfaction with life'.

When selecting the predictor variables, we considered several factors that would impact a person's satisfaction in life including financial status,the type of job (in particular we focused on distinguishing between whether person was engaged in an intellectual, rewarding job versus a repetitive job), feeling of safety in their community, and how much they trusted their current ruling government, courts and schools in their community.

To pick out the best model, we used the data to run 10 fold cross validation comparing four different models. Each of the models are described below:

Base model: In the base model we included two dependent variables: V242: age and V240: sex

Model 1: In the first model we used variables that captured a person's financial situation and job type to predict their satisfaction in life. The variables we considered are:

Table 5. List of variables for prediction model 1

Variable	**Description**
V239	Scale of incomes
V237	Family savings during past year
V233	Nature of tasks: independence
V231	Nature of tasks: manual vs. intellectual
V232	Nature of tasks: routine vs. creative
V229	Employment status
V190	In the last 12 month, how often have you or your family: Gone without needed medicine or treatment you needed
V191	In the last 12 month, how often have you or your family: Gone without a cash income

V188	In the last 12 month, how often have you or your family: Gone without enough food to eat
V182	Worries: Not being able to give one's children a good education
V181	Worries: Losing my job or not finding a job

Model 2: In the second model we used variables that captured whether an individual felt connected to their community, and how they viewed the overall governance of their country. The variables we considered are:

Table 6. List of variables for prediction model 2

Variable	**Description**
V226	Vote in elections: local level
V227	Vote in elections: National level
V141	How democratically is this country being governed today
V142	How much respect is there for individual human rights nowadays in this country
V213	I see myself as part of my local community

Model 3: In the third model, we used variables that captured how safe individuals felt in their community as we expect this to have an impact on how satisfied a person is with their life

Table 7. List of variables for prediction model 3

Variable	**Description**
V189	In the last 12 month, how often have you or your family: Felt unsafe from crime in your own home
V184	Worries: A terrorist attack
V183	Worries: A war involving my country
V180	Respondent's family was victim of a crime during last year
V179	Respondent was victim of a crime during the past year
V178	Things done for reasons of security: Carried a knife, gun or other weapon

V177	Things done for reasons of security: Preferred not to go out at night
V173	How frequently do the following things occur in your neighborhood: Police or military interfere with people's private life
V174	How frequently do the following things occur in your neighborhood: Racist behavior
V175	How frequently do the following things occur in your neighborhood: Drug sale in streets
V172	How frequently do the following things occur in your neighborhood: Alcohol consumed in the streets
V171	How frequently do the following things occur in your neighborhood: Robberies
V170	Secure in neighborhood

Model 4: In this model, we incorporated variables from all of the models, specifically picking out those that we expect have a strong correlation with satisfaction in life.

Table 8. List of variables for prediction model 4

Variable	**Description**
V239	Scale of incomes
V237	Family savings during past year
V232	Nature of tasks: routine vs. creative
V233	Nature of tasks: independence
V231	Nature of tasks: manual vs. intellectual
V229	Employment status
V189	In the last 12 month, how often have you or your family: Felt unsafe from crime in your own home
V190	In the last 12 month, how often have you or your family: Gone without needed medicine or treatment that you needed
V191	In the last 12 month, how often have you or your family: Gone without a cash income
V188	In the last 12 month, how often have you or your family: Gone without enough food to eat

V171	How frequently do the following things occur in your neighborhood: Robberies
V170	Secure in neighborhood
V174	How frequently do the following things occur in your neighborhood: Racist behavior
V179	Respondent was victim of a crime during the past year
V180	Respondent's family was victim of a crime during last year
V173	How frequently do the following things occur in your neighborhood: Police or military interfere with people's private life

The cross validation errors we received for the following models when run using the test data are summarized in the table below:

Table 9. Summary of cross validation errors of the four prediction models

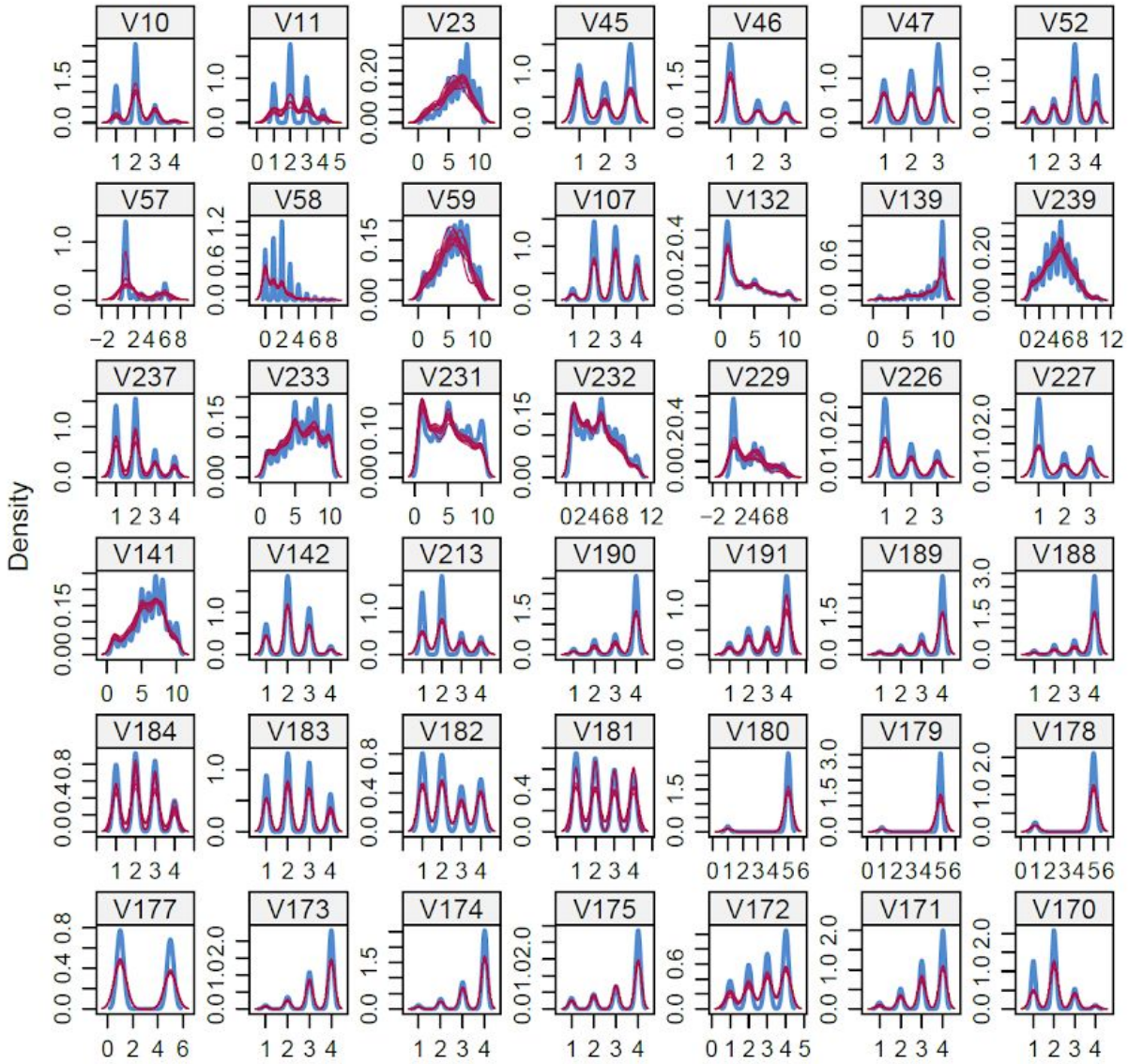
Model	Cross-validation error
Model 1	3.80
Model 2	4.30
Model 3	4.28
Model 4	3.74

Based on the cross validation errors on the train set, we picked Model 4 as our final model to do the prediction. The final mean standard error (MSE) for the test set using Model 4 was 3.80.

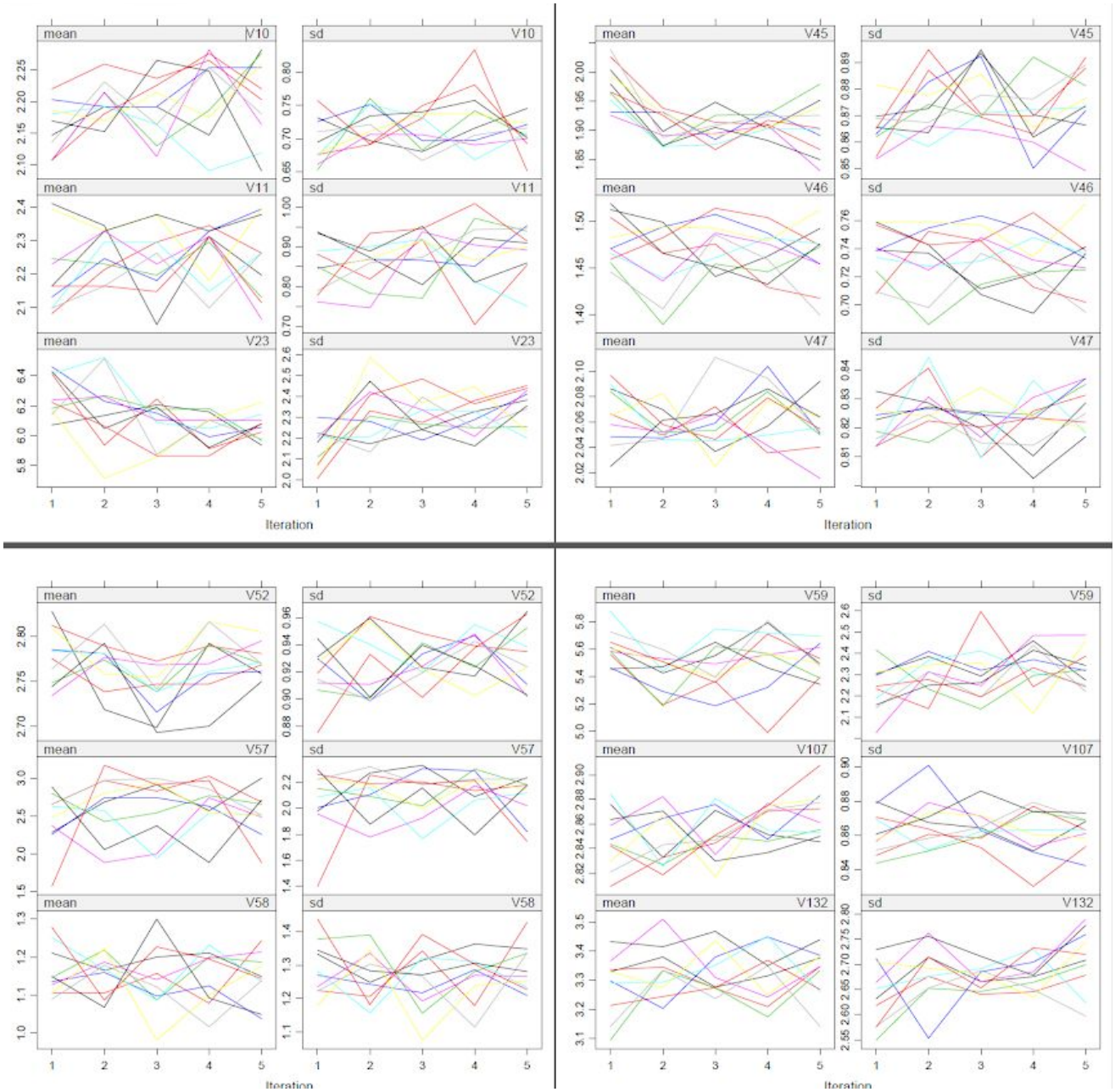
The comparison of cross validation errors of our models suggest that models are able to better predict the dependent variable when the number of predictor variables (independent variables) are increased. However, this could be specific to the models we used, as the model with the greatest number of predictor variables has relevant variables. Randomly adding variables to the model that do not help explain the dependent variable is not likely to improve the performance of the model, although it will have many variables.

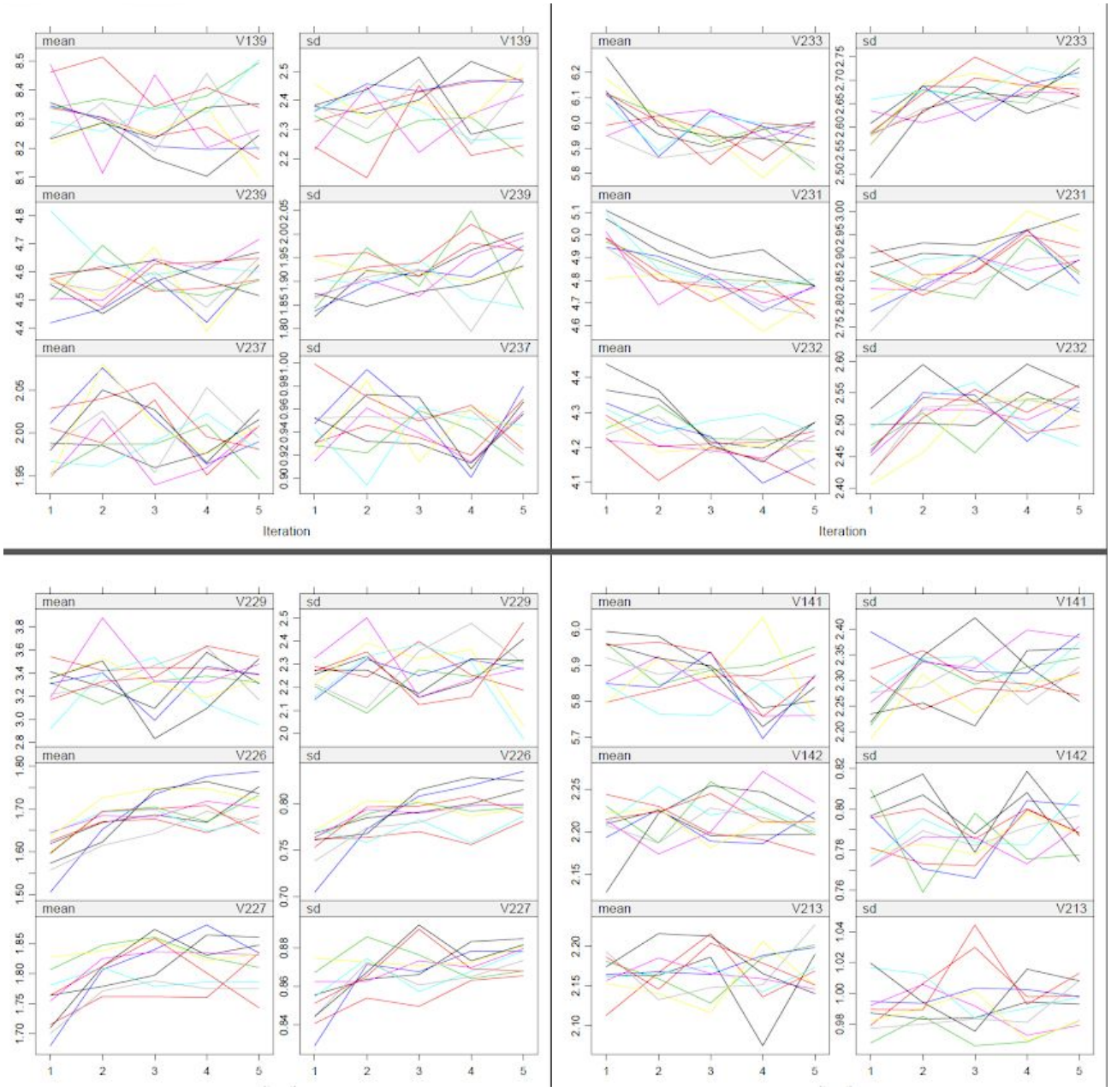
Appendix A.

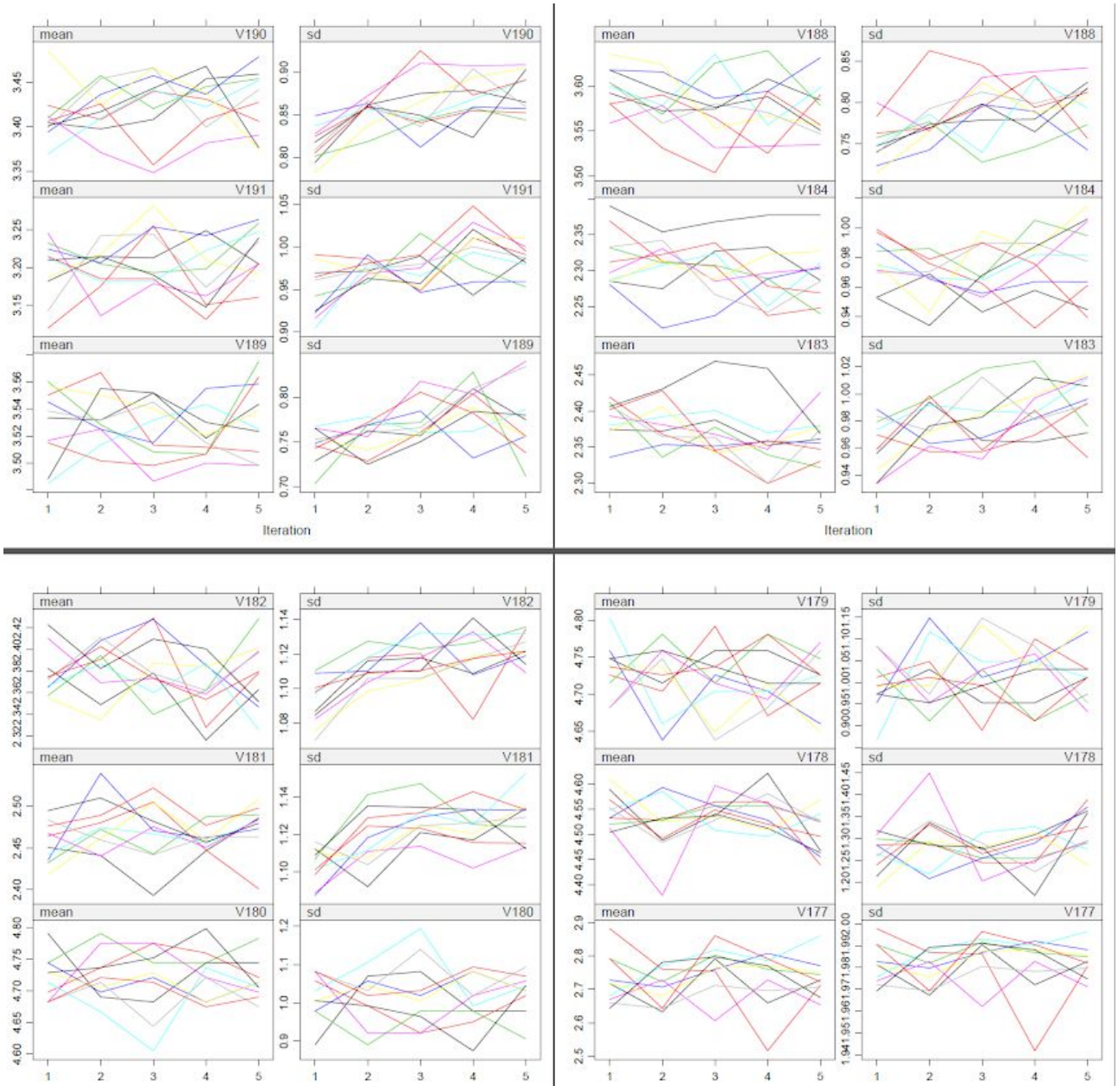
Density plots

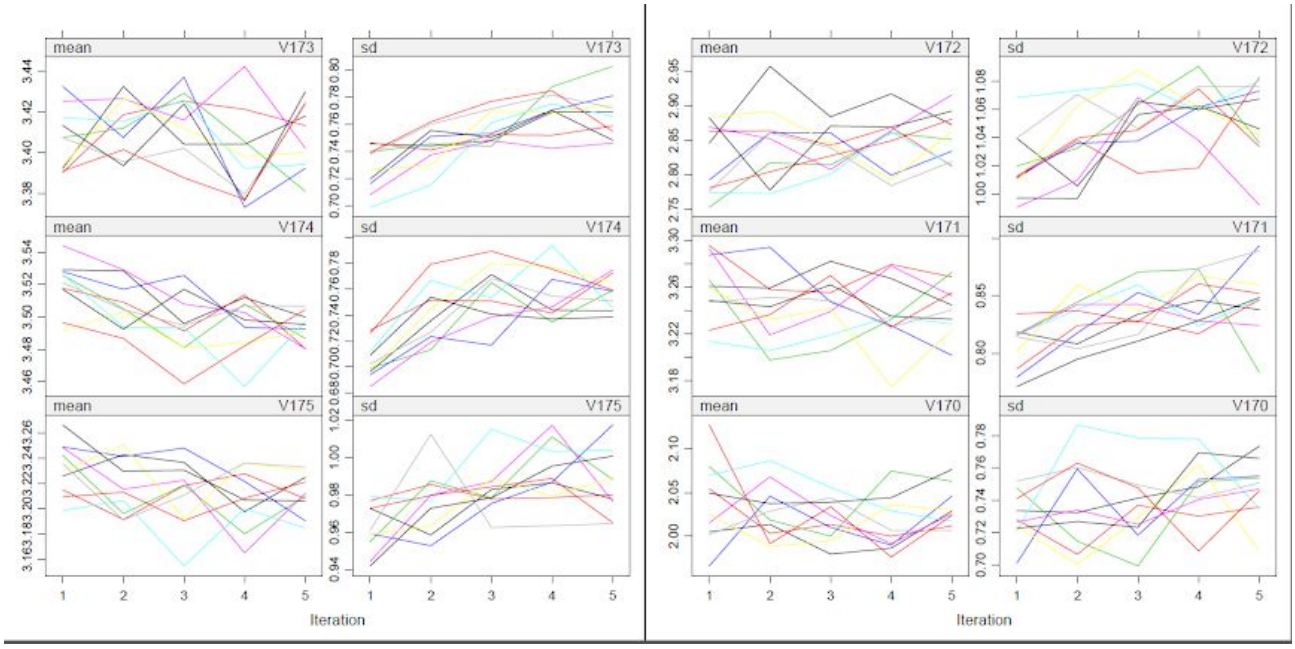


Appendix B Traceplots




mean V139
sd V139
mean V239
sd V239
mean V237
sd V237
mean V233
sd V233
mean V231
sd V231
mean V232
sd V232
Iteration
Iteration
mean V229
sd V229
mean V226
sd V226
mean V227
sd V227
mean V141
sd V141
mean V142
sd V142
mean V213
sd V213


mean V190
sd V190
mean V191
sd V191
mean V189
sd V189
mean V188
sd V188
mean V184
sd V184
mean V183
sd V183
Iteration
Iteration
mean V182
sd V182
mean V181
sd V181
mean V180
sd V180
mean V179
sd V179
mean V178
sd V178
mean V177
sd V177



Appendix C

Outlier counts

```
 37  49  52  53  54  57  59  60  77  82  96  99 100 103 116 120 125 126 127 130 164 181 198
  1  20  20  18  20   3  20   1  20   4  20   6   6  20   2  20  19  19   1   1  20  20   3
224 226 231 235 245 253 254 267 288 301 331 336 342 343 344 352 356 368 372 373 375 387 390
  4  20  15   4  20  20  20  16  20  20   2  20   2  11  20  20   1  20  20  20  20   1  17
391 397 414 417 419 434 443 452 456 463 470 497 499 501 510 512 516 524 528 529 540 542 548
 20   6  10   7  20   2   2  13  19  11   1  16  20  20   6  19  20  20  20   8  16   1  20
549 554 555 556
 20   8  20  20
```