Bentley University

We Know Your Age

Age Predictor Based on 150 Survey Questions

Mingrui (Ray) Jin

Sujith Kurian

Joshua Zhang

Quantitative Analysis for Business (ST 625)

Moinak Bhaduri

May 1, 2019

Introduction

The dataset ‘Young People Survey’ from Kaggle is a list of survey questions that were asked responded by the friends of students at FSEV UK. The dataset has 1010 rows, 150 variables (139 quantitative, 11 qualitative). Two types of quantitative variables are presented. Height, age, and weight are numeric variables and considered as quantitative variables. The other type of quantitative variables are survey questions with the result from 1 to 5, strongly agree to strongly disagree. They are considered quantitative variables because each number is unique and is not mere representation of categorical variables. The survey questions belonged to one of the following eight categories:

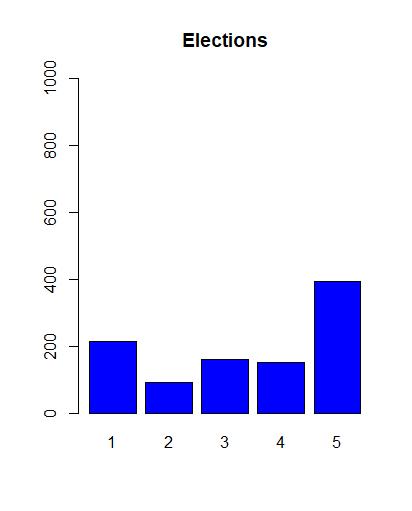
1. Music preferences (19 items)
2. Movie preferences (12 items)
3. Hobbies & interests (32 items)
4. Phobias (10 items)
5. Health habits (3 items)
6. Personality traits, views on life, & opinions (57 items)
7. Spending habits (7 items)
8. Demographics (10 items)

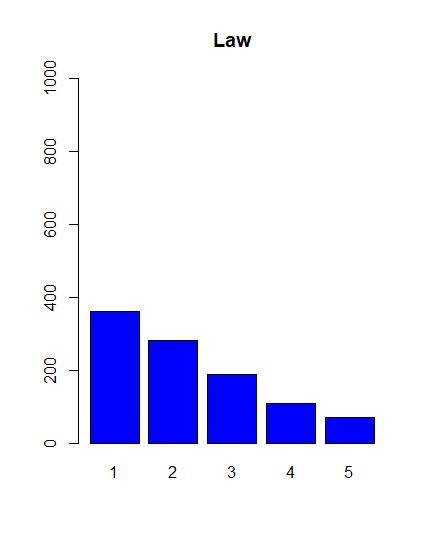
As a project team we decided to research and apply concepts learned in the Quantitative analysis class to test if age of the user can be predicted based on answers to the survey. Age is one of the many types of data that is considered customer sensitive and at the same time it is something that companies in various domains would like to know for various reasons like targeted marketing, identify an auto or health or life insurance premium, protecting the minors from an ever unsafe internet etc.

The Industry Knowledge Emulation Model (IKEM)

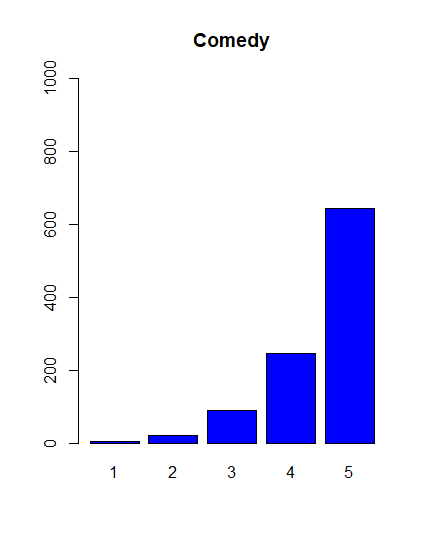
The Industry Knowledge Emulation Model (IKEM) is a model that we tried because we thought it would be interesting to compare its performance with a theoretically sound method. The model is generated in three steps: filtering, variable reduction, and model selection. The goal of the model is to predict a survey taker’s age.

The process of filtering for variables according to industry knowledge is simulated. We gave the computer one criteria, which is to only pick survey questions where less than 20% of the survey takers responded with ‘3’. The answer ‘3’ corresponds to answers of opinion such as “not interested”, “equally likely”, or “no preference”. If this criterion is satisfied, then we know that more than 80% of the responses are distributed either positively or negatively with respect to the question being asked. Using only a simple filter must involve assumptions. One such assumption is that people of similar age express similar opinions. There are a few distributions that can follow in this situation.

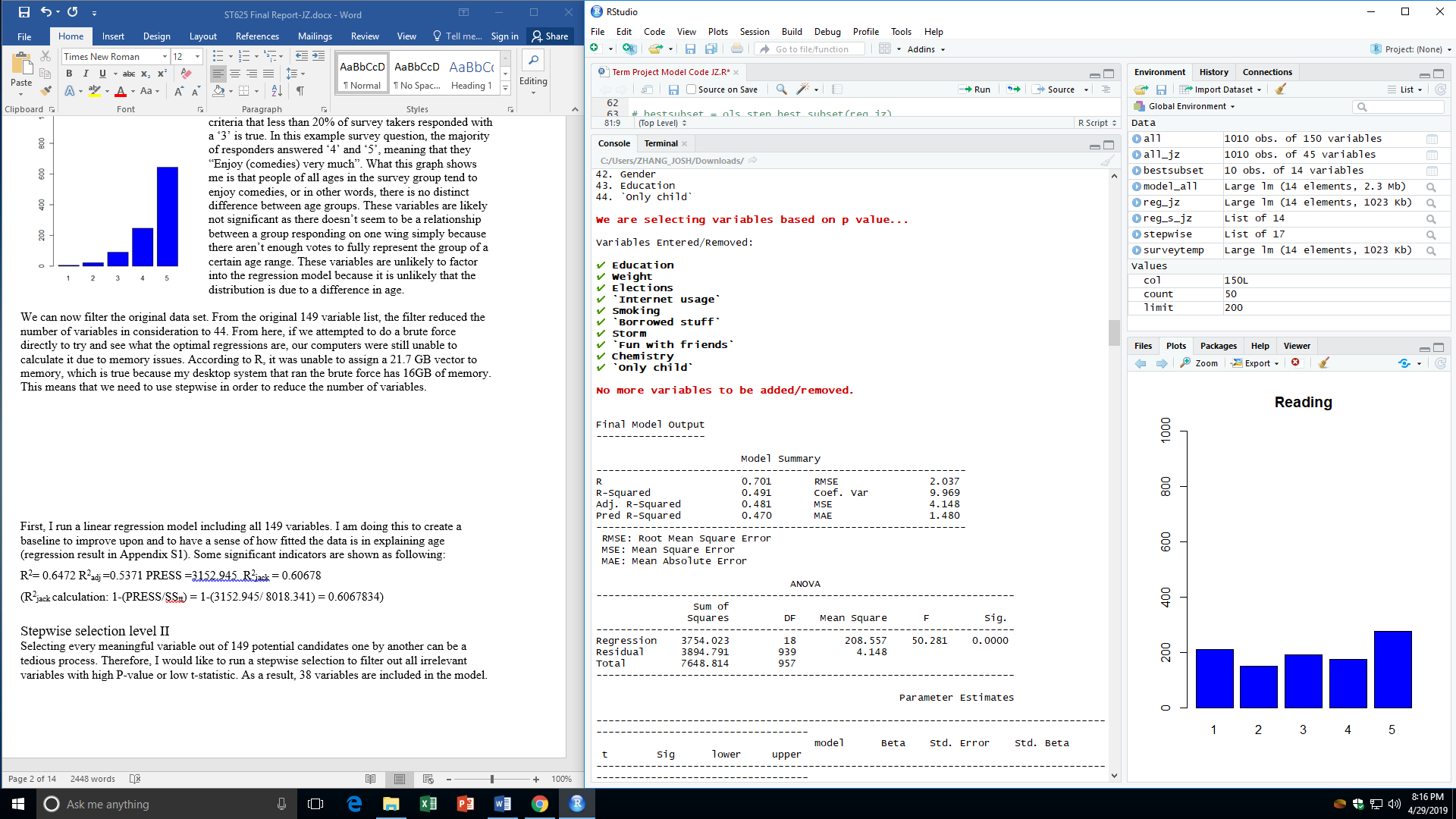
The first distribution we will look at satisfies the filter criteria, but also exhibit a somewhat balanced split between the two wings of opinion. The example survey question for this distribution is about Elections and was phrased, “I always try to vote in elections”. There were less than 200 out of 1010 responders, so the criteria that less than 20% of survey takers responded with a ‘3’ is true. The distribution is weighted towards a higher score, with interpretation of “Strongly Agree” rather than “Strongly Disagree”. The interesting observation for this survey question is that the survey takers were more likely to have a stronger response (‘1’ and ‘5’) than a weaker response (‘2’ and ‘4’). If the variable is significant, and there happens to be a connection with Age, then this kind of variable will be the strongest candidates to include in a regression model.

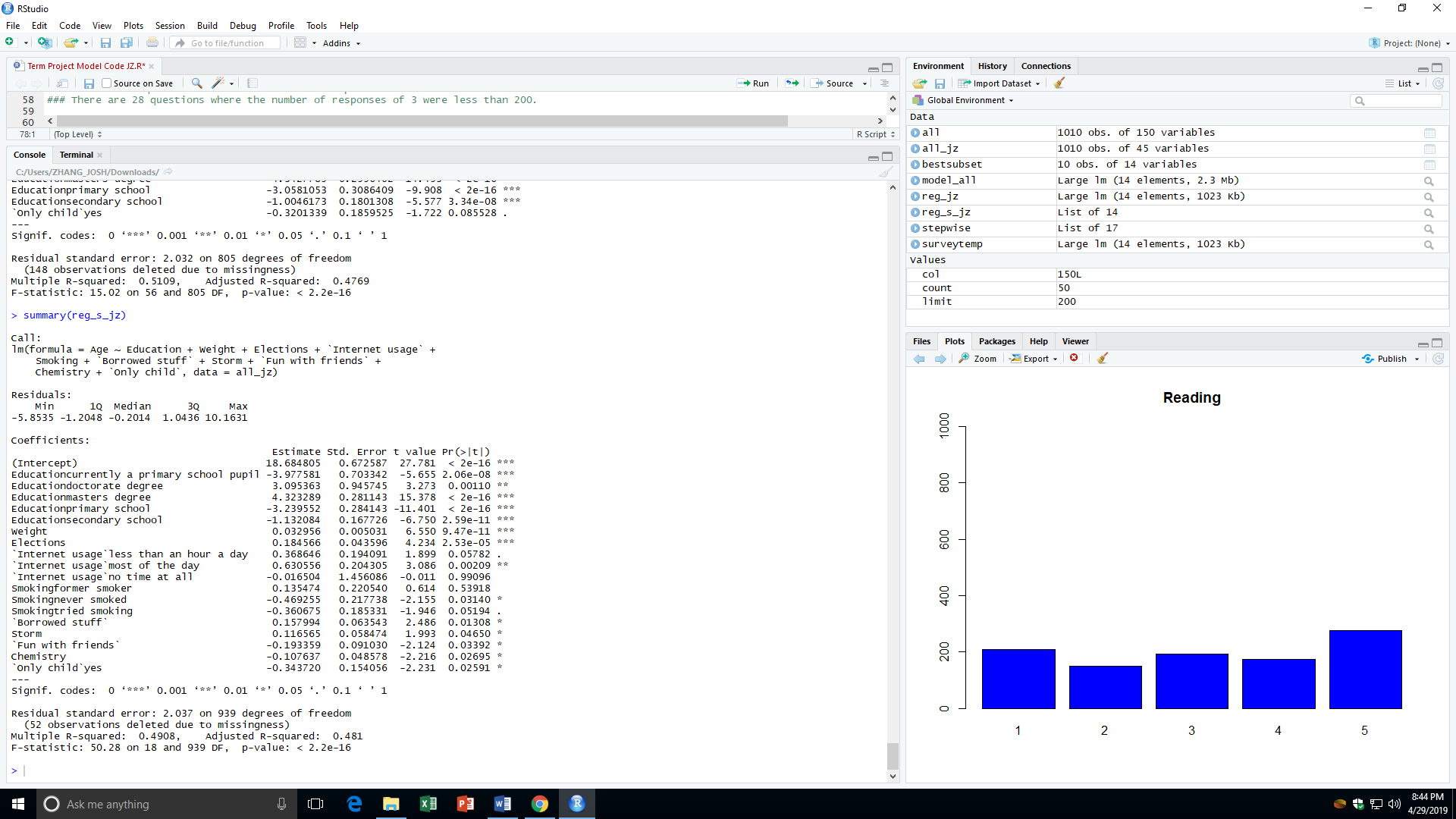


A different question with a different distribution occurs in Law. The survey question about Law was phrased, “Hobbies & Interests - Law” and the answers ranged from not interested (‘1’) to very interested (‘5’). There were less than 200 out of 1010 responders, so the criteria that less than 20% of survey takers responded with a ‘3’ is true. The distribution is weighted towards a higher score, with interpretation of overall disinterest (‘1’ and ‘2’) over overall interest (‘4’ and ‘5’). There is a better potential of statistical significance if age groups respond favorably or unfavorably. In this example, the interesting question in the distribution is who responded ‘4’ or ‘5’ to this question. Knowing that the distribution of Age in the survey dataset is slightly skewed to the right, the potential that lower aged survey responders may be the ones responding with ‘4’ or ‘5’ would make this variable tremendously interesting. Based on what we observe in reality, we tend to think that younger people are less likely to be very interested in Law when compared to older people.



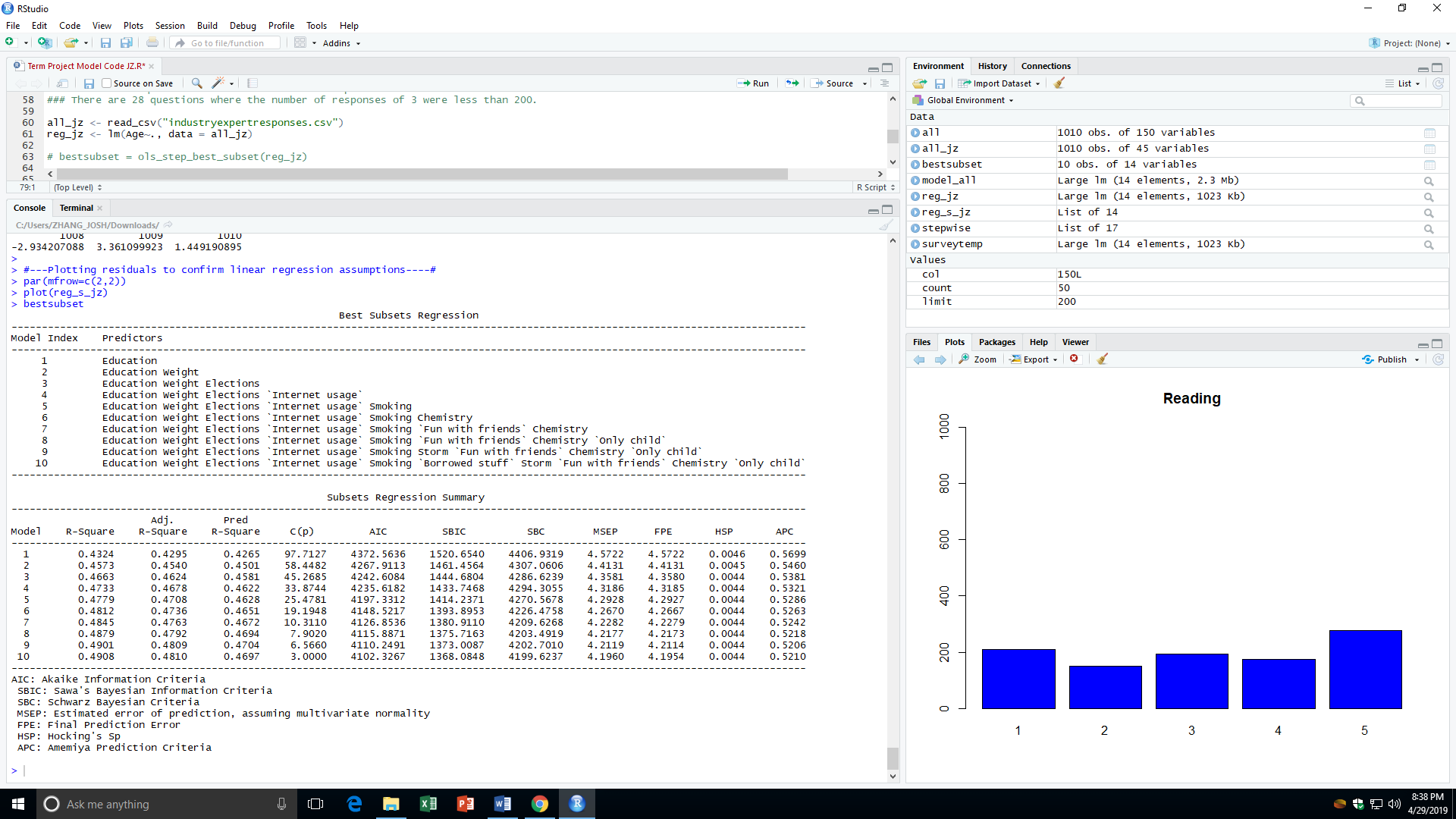
The third and final type of distribution looks at a variable where very few responses were on one side of the wing. There were less than 200 out of 1010 responders, so the criteria that less than 20% of survey takers responded with a ‘3’ is true. In this example survey question, the majority of responders answered ‘4’ and ‘5’, meaning that they “Enjoy (comedies) very much”. What this graph shows me is that people of all ages in the survey group tend to enjoy comedies, or in other words, there is no distinct difference between age groups. These variables are likely not significant as there doesn’t seem to be a relationship between a group responding on one wing simply because there aren’t enough votes to fully represent the group of a certain age range. These variables are unlikely to factor into the regression model because it is unlikely that the distribution is due to a difference in age.

We can now filter the original data set. From the original 149 variable list, the filter reduced the number of variables in consideration to 44. From here, if we attempted to do a brute force directly to try and see what the optimal regressions are, our computers were still unable to calculate it due to memory issues. According to R, it was unable to assign a 21.7 GB vector to memory, which is true because my desktop system that ran the brute force has 16GB of memory. This means that we need to use stepwise in order to reduce the number of variables. We can see in the variable selection output from the stepwise function, there were 10 variables that were identified as needed to be added. The performance of the model is somewhat reasonable, where the adjusted R2 value is at 0.481. However, since our target is to be able to predict the Age based on their answers, we are most interested in Prediction R2. The best Pred R2 so far is 0.470. Since stepwise function does not exhaustively search all the regression models, we remain open to the idea that there could be a better model. This is the motivation to perform a brute force search with the ten remaining variables derived from the stepwise function. Let’s take a look first at the regression model with the ten variables from the stepwise function.



We can clearly see that the over model is significant, where p-value < 2.2e-16. Next, we see that the adjusted R2 is indeed 0.481, consistent with the stepwise function output. Finally, we can look at the individual variable’s p-values for significance. The only variables that are not significant are for certain options of the qualitative variables. Internet usage has an option of response called “no time at all” meaning the responder does not use the internet at all. This variable can be combined with the base case of “few hours a day”. Similarly, the variable Smoking has an option of response called “former smoker”, meaning the responder has at one point in their life tried smoking. This variable can be combined with the base case of “current smoker”. A detailed break-down of the process for removing the variable is explained on pg. XXX.

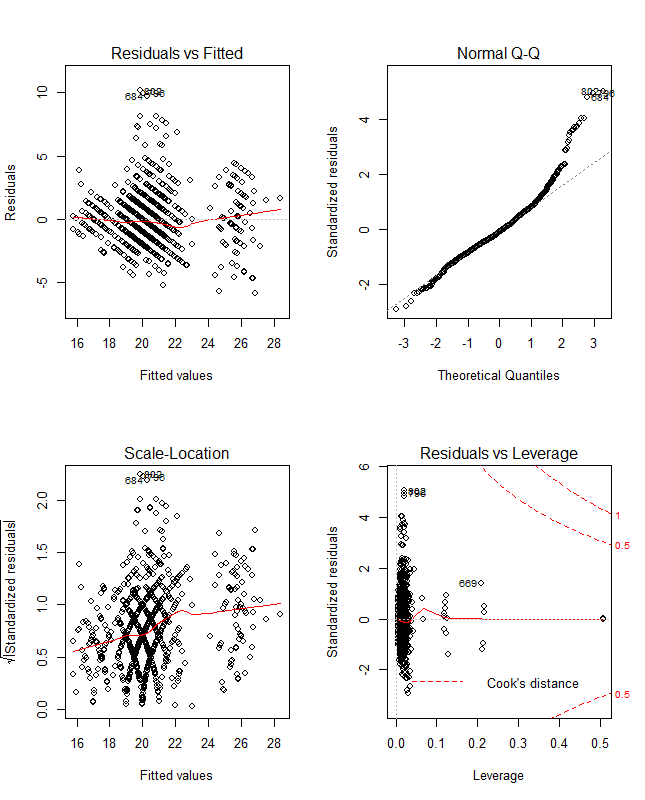
To see if the stepwise function actually found the optimal solution, or whether or not there was a variation of the variables where performance could have been improved is now trivial, as a brute force function with 10 cause variables is within the capacity of my computer. The output is below:



Here we see a slight deviation from what is expected. There are two distinct models that perform better either as a descriptive regression model or a predictive regression model. The two models of interest are the 9 variable model and the 10 variable model. By looking at the adjusted R2 results, we see that the 10 variable model (0.4810) is ever so slightly better than the 9 variable model (0.4809). To present further evidence that the 10 variable model is better at descriptive regression, we look at AIC. The 10 variable model (4102.3267) is clearly an improvement over the 9 variable model (4110.2491). A similar logic can be used to compare the C(p) for the 10 variable model (3.0000) to the 9 variable model (6.5660).

To reiterate, the purpose of our analysis is to try to predict the Age of the next survey taker, which means the principal parameter we look at to determine which model we will pick is R2 jackknife. From the output of the brute force, we can see that the 9 variable model (0.4704) actually outperforms the 10 variable model (0.4697) in R2 jackknife. Although the evidence for the better model seems to be in the 10 variable model, since three parameters we analyzed point to the 10 variable model as the optimal solution, we will adopt the 9 variable model because of its better performance in R2 jackknife.

Taking a look at the residual analysis of this output is discouraging. The plots below show an absence of normality and a questionable pattern in the variance of error.



Looking at the Normal Q-Q plot, we see that there is skewness towards the right side of the response variable. This is evident because we could see the skewed distribution directly from the collected data as well. Before justifying why we chose to continue with our model, we will point out the other interesting residual plot analysis. In the Scale-Location plot, we see that there is no pattern that suggests a need for transformation. However, perhaps due to the quasi-quantitative nature of our survey data, it forms a very beautiful weave pattern, specifically between the 18 to 22 values on the x-axis.

To justify why we chose to continue using this model even though it violates the normally assumptions that is necessary for regression analysis is that we are using pseudo quantitative variables. If we were to be completely strict about how we defined our variables, we would treat each survey question as a qualitative variable, with four dummy variables to denote the five options. This is unreasonable given the amount of manual work we would have to do just to accurately implement this into our analysis. Second, the constant variance of the residuals does not seem to have any recognizable pattern, other than the aforementioned weave pattern which is likely due to the nature of the survey data. I would also mention that sometimes there are people who deliberately ruin a survey’s validity by giving arbitrary answers, sometimes answers that are contradictory in nature. A better process for the future is to do a stricter and more diligent filtering process to try and systematically remove obvious tampering outliers from the data, which will greatly reduce any concerns stemming from the residual plots. We should also mention that

In conclusion, for our purposes, we choose to adopt the 9 variable because it is mathematically chosen as the best predictor regression model out of the ten possible models selected by best subset. We choose to overlook any residual normality assumption checks due to the nature of the data as well as factors that are inherent to surveys. The best performing model gives a prediction R2 of 0.4704, and all other quality checks pass. Next, we will introduce the other model that we built, without the influence of industry knowledge emulation.

Stepwise Selection instead of Brute Force Method

Though the brute force method will try all possible combinations among all cause variables, it is impossible to use brute force method in our dataset. If we use brute force method to run 149 causing variables, 2^149 is 7.136238463529798e44. For a regular personal laptop, it will take too long to find the result. Therefore, the only option left is stepwise selection. Although the maximum regression built for 149 variables could be (149\*150)/2= 11175 Seconds or 186.25 Minutes, not all possible regression are built. Using my personal laptop, I can finish the process under 3 minutes and find that 38 out of 149 variables are significant at explaining the variation of age.

Stepwise selection level I

First, I run a linear regression model including all 149 variables. I am doing this to create a baseline to improve upon and to have a sense of how fitted the data is in explaining age (regression result in Appendix S1). Some significant indicators are shown as following:

R2= 0.6472 R2adj =0.5371 PRESS =3152.945  R2jack = 0.60678

(R2jack calculation: 1-(PRESS/SStt) = 1-(3152.945/ 8018.341) = 0.6067834)

Stepwise selection level II

Selecting every meaningful variable out of 149 potential candidates one by another can be a tedious process. Therefore, I would like to run a stepwise selection to filter out all irrelevant variables with high P-value or low t-statistic. As a result, 38 variables are included in the model.

(regression result in Appendix S2).

Some significant indicators are shown as following:

R2= 0.585 R2adj =0.561 PRESS =3139.057 R2jack = 0.6085154

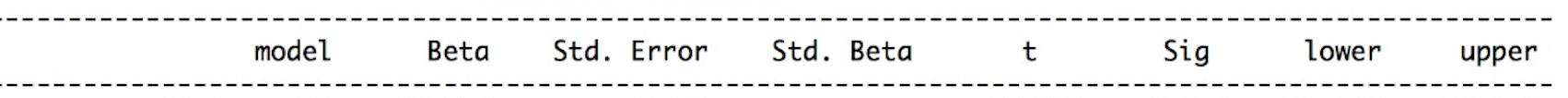
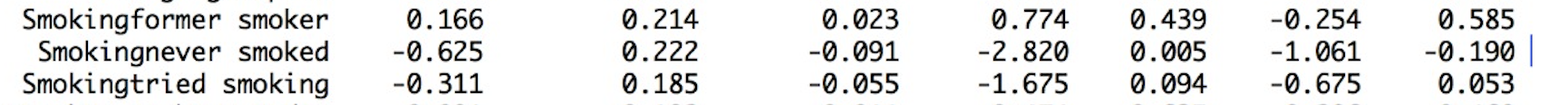
In this model, we can see a decrease in R2 because 149-38=111 variables are removed from the model. AS R2 always increase as more variables are added, decrease in R2 is as expected. However, R2adj increased from 0.561 to 0.5371. It indicates that the model is explaining more variation in age after the punishment of including irrelevant variables. Also, R2jack also changed slightly, indicating that the predicting power of level 2 is also better.

Stepwise selection level III

As one of many drawbacks of using stepwise selection, nonsense variables could be included in the model that totally make no sense. First, one variable included in the model is ‘*House-blockofflats`house/bungalow’.* The response is either grow up in block of flats or in a house.What kind of house people grow up with has no correlation with their age. Therefore, variables such as ‘*House-blockofflats`house/bungalow’* should be removed. For the same reason, the question for whether you are the only child at your house is also removed. Second, some variables with high p values and low correlation with age are also removed(how much you love the subject: Geography and how much do you love branded clothing). Third, some categorical variables are having high P-values as well. High P-value for some(not all) of the categorical variables do not means that variable is insignificant. Instead, it means that one category is not significant different from another category. For example, a high P value for internet use – no time at all indicates that it is insignificantly different from the underlying category(situation when all the dummy variables = 0). To put this in statistic term:

B1 = Beta 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | =1 if former smoker |  | =1 if never smoked |  | =1 if tried smoking |
| B1 |  | B2 |  | B3 |  |
|  | =0 if else |  | =0 if else |  | =0 if else |

If B1 = B2 = B3 =0 🡪 current smoker

From the graph above, it is clear that both former smoker and tried smoking are not significantly different from current smoker. Therefore, it is not needed to differentiate those three terms. Under the new model, three dummy variables are combined into one:

|  |  |
| --- | --- |
|  | =1 if smoker |
| B1 | =0 if never smoked |

Internet usage is also merged from four categories into two categories: sometime of the day and most of the day.

After these modifications, I ran another stepwise selection using the 149-4=145 cause variables. This time, another 38 variables are selected(Appendix S3). After those modifications, statistic indicators also improved:

R2= 0.597 R2adj =0.576 PRESS =3063.743  R2jack = 0.6179081

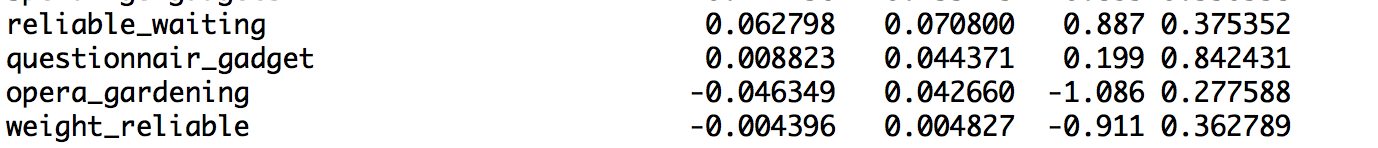
This model is doing a better job at both explain variation in current dataset as well as predicting new data set.

Stepwise selection level IV

Next, a few interaction terms are introduced into the model. Since both reliability(how reliable you think you are) and waiting(how patient you are while waiting) have positive relationship with age(target variable), I would like to know if a interaction term between them can yield a better result. Same logic applied to a few other interaction variables: questionnaire and gadget, opera and gardening, and weight and reliable. All four interaction terms are added into level III model. A new regression is executed and the result is as following(regression result: Appendix 4). Both low coefficient in relate to other variables(same unit, rating of 1 – 5) and high P-value indicate that all four interactive terms are not significant in explain the variation in age. Statistic indicators also proof the same result.

R2= 0.5983 R2adj =0.5754 PRESS =3086.577 R2jack = 0.6150604

Both R2adj and R2jack decreased compared with model III. It means that model IV is worse off at both explain variation in current dataset as well as predicting new data set.



Conclusion:

With a dataset of 1010 participants ,150 variables (survey question) and with a goal of predicting the age of the participants the number of significant variables were identified through the stepwise method. The base model was further improved by combining some categorical variables. This provided us with an improved R2a and R2jack values. Below are the results of the step wise selection at various complexity levels. The focus was on improvements observed to Adj. R2 and R2jack values since the goal is age prediction.

Step wise selection level 1 result

**R2**= 0.6472 **R2adj** =0.5371 **PRESS** =3152.945  **R2jack** = 0.60678

Step wise selection level 2 result

**R2**= 0.585 **R2adj** =0.561 **PRESS** =3139.057 **R2jack** = 0.6085154

Step wise selection level 3 result

**R2**= 0.597 **R2adj** =0.576 **PRESS** =3063.743  **R2jack** = 0.6179081

Step wise selection level 4 result

**R2**= 0.5983 **R2adj**=0.5754 **PRESS** =3086.577 **R2jack** = 0.6150604

To do a sanity check an unconventional and innovative approach of Industry knowledge emulation, Brute-Force and Step wise method was applied. It proved to us that majority of the variables that were selected through this Brute force approach were also part of the Step wise method for this type of data.

Taking this to the next level , the model was tested in class ST625 at Bentley University with a small number of users and the prediction results were encouraging. Four users responded to the survey and age prediction was within +/- 3 years of the actual age.

Below are the variables that were used to collect the data:

* Timestamp
* How old are you?
* Highest Education Achieved:
* What is your weight?
* I always try to vote in elections
* How much time do you spend online?
* Smoking Habits
* How phobic are you of Thunder and Lightning?
* Socializing as a Hobby/Interest?
* Chemistry as a Hobby/Interest?
* Are you an only child?

Result table:

|  |  |  |
| --- | --- | --- |
| **How old are you?** | **Estimated Age** | **Guess Within 2 Years of Actual?** |
| 23 | 22.969299 | 0.030701 |
| 23 | 25.457925 | -2.457925 |
| 22 | 24.661569 | -2.661569 |
| 30 | 26.435375 | 3.564625 |
| 23 | 25.035395 | -2.035395 |

It is important to note that the data collected was for the users in Bratislava which has a different socio-economic behavior when compared to users in the United States or India or China or Australia or any other country. However we think that the significant variables that were identified through this process can be used to gather data in other countries as well, thereby reducing the amount of data that needs to be collected to predict age. The variables are significant in the presence of other variables and so it will be important to gather all the significant variable data rather than just cherry picking which data may be needed.

Appendix S1

> stepwisereg = ols\_step\_both\_p(reg1)

> summary(reg1)

Call:

lm(formula = Age ~ ., data = project)

Residuals:

Min 1Q Median 3Q Max

-4.3952 -1.0250 -0.0702 0.8100 7.7191

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24.5442372 2.8097774 8.735 < 2e-16 \*\*\*

Music -0.0727630 0.1499733 -0.485 0.627761

Slowsongsorfastsongs 0.0054296 0.1084772 0.050 0.960099

Dance 0.1490838 0.0922971 1.615 0.106869

Folk 0.0367909 0.0913674 0.403 0.687359

Country 0.1241276 0.0916815 1.354 0.176364

Classicalmusic -0.1651987 0.0973071 -1.698 0.090170 .

Musical -0.0298137 0.0854626 -0.349 0.727344

Pop 0.1085872 0.0900861 1.205 0.228615

Rock 0.0435033 0.1020136 0.426 0.669961

MetalorHardrock -0.0017765 0.0836076 -0.021 0.983056

Punk -0.0754465 0.0863788 -0.873 0.382834

`Hiphop,Rap` -0.1657418 0.0780181 -2.124 0.034114 \*

`Reggae,Ska` 0.0914505 0.0822311 1.112 0.266610

`Swing,Jazz` -0.1071123 0.0865707 -1.237 0.216549

Rocknroll 0.0034150 0.0932463 0.037 0.970799

Alternative -0.0491828 0.0782575 -0.628 0.529973

Latino 0.0749343 0.0822719 0.911 0.362822

`Techno,Trance` 0.0410404 0.0729185 0.563 0.573799

Opera 0.1975437 0.0958727 2.060 0.039856 \*

Movies -0.2348038 0.1484359 -1.582 0.114299

Horror -0.0259687 0.0748948 -0.347 0.728933

Thriller 0.2033739 0.0897419 2.266 0.023854 \*

Comedy -0.1388447 0.1310842 -1.059 0.290006

Romantic 0.1240409 0.0910523 1.362 0.173700

`Sci-fi` -0.1814801 0.0782715 -2.319 0.020809 \*

War -0.0522265 0.0752950 -0.694 0.488231

`Fantasy/Fairytales` 0.1445168 0.1092042 1.323 0.186304

Animated -0.1594294 0.1010608 -1.578 0.115283

Documentary -0.0157707 0.0912231 -0.173 0.862814

Western -0.0384572 0.0923736 -0.416 0.677348

Action 0.0157612 0.0849752 0.185 0.852926

History 0.0932013 0.0846044 1.102 0.271147

Psychology -0.0715557 0.0774391 -0.924 0.355908

Politics -0.0406107 0.0885344 -0.459 0.646644

Mathematics -0.1232850 0.0835595 -1.475 0.140714

Physics 0.1060657 0.1004783 1.056 0.291644

Internet 0.0285905 0.1133676 0.252 0.800993

PC 0.1774159 0.0901846 1.967 0.049693 \*

EconomyManagement 0.0509311 0.0725076 0.702 0.482733

Biology -0.1397529 0.1057010 -1.322 0.186707

Chemistry -0.2126898 0.0884441 -2.405 0.016535 \*

Reading 0.0176328 0.0747913 0.236 0.813713

Geography 0.0558187 0.0732094 0.762 0.446140

Foreignlanguages -0.0608644 0.0874928 -0.696 0.486963

Medicine 0.2702813 0.0937396 2.883 0.004100 \*\*

Law 0.0957955 0.0831299 1.152 0.249710

Cars -0.0973715 0.0776619 -1.254 0.210490

Artexhibitions 0.1857000 0.0927406 2.002 0.045772 \*

Religion -0.0018986 0.0812367 -0.023 0.981363

`Countryside,outdoors` 0.0933942 0.0787430 1.186 0.236146

Dancing 0.0436554 0.0762310 0.573 0.567117

Musicalinstruments 0.0271320 0.0662443 0.410 0.682289

Writing -0.0219558 0.0789130 -0.278 0.780950

Passivesport 0.0402046 0.0624736 0.644 0.520158

Activesport -0.0305115 0.0677042 -0.451 0.652426

Gardening 0.0827404 0.0829298 0.998 0.318887

Celebrities 0.0493842 0.0830303 0.595 0.552257

Shopping -0.1607993 0.1050071 -1.531 0.126307

Scienceandtechnology 0.0990531 0.0894378 1.108 0.268593

Theatre -0.1286752 0.0892588 -1.442 0.150026

Funwithfriends -0.0948968 0.1378033 -0.689 0.491361

Adrenalinesports 0.0033570 0.0751672 0.045 0.964395

Pets -0.0819654 0.0594771 -1.378 0.168773

Flying -0.0076253 0.0786195 -0.097 0.922772

Storm 0.0998468 0.0923749 1.081 0.280256

Darkness 0.0041390 0.0856513 0.048 0.961477

Heights -0.0073747 0.0708766 -0.104 0.917170

Spiders 0.0112057 0.0667677 0.168 0.866782

Snakes 0.0311671 0.0748027 0.417 0.677104

Rats -0.0756917 0.0804045 -0.941 0.346950

Ageing 0.0657632 0.0683400 0.962 0.336355

Dangerousdogs -0.0444272 0.0751987 -0.591 0.554916

Fearofpublicspeaking -0.0495057 0.0878760 -0.563 0.573437

Smokingformer smoker 0.2720989 0.2793840 0.974 0.330553

Smokingnever smoked -0.2413789 0.3075029 -0.785 0.432837

Smokingtried smoking -0.0870884 0.2504431 -0.348 0.728180

Alcoholnever -0.2252885 0.3505817 -0.643 0.520762

Alcoholsocial drinker -0.0953514 0.2261377 -0.422 0.673456

Healthyeating -0.0177712 0.1019298 -0.174 0.861661

Dailyevents 0.1360503 0.0881806 1.543 0.123482

Prioritisingworkload 0.0179021 0.0869087 0.206 0.836883

Writingnotes 0.0837464 0.0690754 1.212 0.225921

Workaholism -0.0465128 0.0822035 -0.566 0.571760

Thinkingahead 0.0810356 0.0843294 0.961 0.337034

Finaljudgement -0.0146484 0.0738055 -0.198 0.842753

Reliability 0.2228370 0.1029545 2.164 0.030893 \*

Keepingpromises -0.0491173 0.1079518 -0.455 0.649307

Lossofinterest 0.0774907 0.0677593 1.144 0.253316

Friendsversusmoney 0.0043572 0.0835224 0.052 0.958415

Funniness 0.0660703 0.0826632 0.799 0.424504

Fake -0.0855544 0.0934723 -0.915 0.360469

Criminaldamage 0.0419595 0.0621159 0.676 0.499660

Decisionmaking 0.0457698 0.0806306 0.568 0.570522

Elections 0.1905657 0.0576010 3.308 0.001004 \*\*

`Self-criticism` -0.1225976 0.0787600 -1.557 0.120183

Judgmentcalls -0.1545751 0.0971150 -1.592 0.112075

Hypochondria -0.0745037 0.0833336 -0.894 0.371719

Empathy -0.0730363 0.0881803 -0.828 0.407908

Eatingtosurvive -0.0722044 0.0728058 -0.992 0.321792

Giving 0.0518676 0.0729001 0.711 0.477105

Compassiontoanimals 0.2398800 0.0832590 2.881 0.004128 \*\*

Borrowedstuff 0.0073226 0.0899175 0.081 0.935126

Loneliness 0.0655325 0.0922744 0.710 0.477908

Cheatinginschool -0.0372811 0.0764928 -0.487 0.626197

Health 0.1950366 0.0950649 2.052 0.040714 \*

Changingthepast -0.1633416 0.0755592 -2.162 0.031098 \*

God -0.0549350 0.0767208 -0.716 0.474294

Dreams -0.2309851 0.1309772 -1.764 0.078403 .

Charity -0.0644014 0.0960567 -0.670 0.502871

Numberoffriends -0.0439945 0.1015814 -0.433 0.665127

Punctualityi am often early 0.1354992 0.2038073 0.665 0.506452

Punctualityi am often running late 0.2414774 0.2093718 1.153 0.249307

Lyingnever 0.4166023 0.4553939 0.915 0.360717

Lyingonly to avoid hurting someone -0.1247651 0.3029886 -0.412 0.680672

Lyingsometimes -0.0786011 0.2818504 -0.279 0.780453

Waiting 0.0819354 0.0902907 0.907 0.364589

Newenvironment -0.0243789 0.0846492 -0.288 0.773464

Moodswings -0.1518883 0.0961878 -1.579 0.114934

Appearenceandgestures -0.3528406 0.1045951 -3.373 0.000799 \*\*\*

Socializing 0.0003658 0.0903185 0.004 0.996770

Achievements -0.0646353 0.0946196 -0.683 0.494848

Respondingtoaseriousletter -0.0474775 0.0753378 -0.630 0.528848

Children -0.0169320 0.0844479 -0.201 0.841167

Assertiveness 0.1341057 0.0801638 1.673 0.094957 .

Gettingangry 0.0542568 0.0866475 0.626 0.531476

Knowingtherightpeople -0.1159957 0.0889685 -1.304 0.192892

Publicspeaking 0.0063810 0.0816086 0.078 0.937707

Unpopularity 0.0336518 0.0778183 0.432 0.665603

Lifestruggles 0.1097286 0.0816756 1.343 0.179715

Happinessinlife -0.0462644 0.1303129 -0.355 0.722717

Energylevels -0.0004895 0.1113581 -0.004 0.996495

`Small-bigdogs` 0.0048236 0.0824399 0.059 0.953365

Personality 0.2207333 0.1501566 1.470 0.142169

Findinglostvaluables -0.0274031 0.0768741 -0.356 0.721637

Gettingup -0.0266485 0.0711815 -0.374 0.708281

Interestsorhobbies -0.0390384 0.0876904 -0.445 0.656374

`Parents'advice` -0.1423556 0.1088651 -1.308 0.191584

Questionnairesorpolls -0.2476258 0.0845091 -2.930 0.003539 \*\*

Internetusageless than an hour a day -0.2010140 0.2548763 -0.789 0.430668

Internetusagemost of the day 0.7130390 0.2685082 2.656 0.008164 \*\*

Finances 0.0568739 0.0818124 0.695 0.487261

Shoppingcentres 0.0578793 0.0903767 0.640 0.522184

Brandedclothing -0.0120319 0.0766788 -0.157 0.875375

Entertainmentspending -0.0135138 0.0897877 -0.151 0.880423

Spendingonlooks 0.0425836 0.1001431 0.425 0.670849

Spendingongadgets -0.1069235 0.0800657 -1.335 0.182322

Spendingonhealthyeating -0.0498970 0.0882490 -0.565 0.572041

Height -0.0227106 0.0145539 -1.560 0.119270

Weight 0.0341920 0.0095603 3.576 0.000381 \*\*\*

Numberofsiblings 0.0135404 0.0933414 0.145 0.884718

Gendermale 0.4601418 0.3451558 1.333 0.183077

`Left-righthanded`right handed -0.0867457 0.2841478 -0.305 0.760274

Educationcurrently a primary school pupil -4.7651693 1.5942640 -2.989 0.002934 \*\*

Educationdoctorate degree 2.8111299 1.3428501 2.093 0.036803 \*

Educationmasters degree 3.6893230 0.3478903 10.605 < 2e-16 \*\*\*

Educationprimary school -2.6330909 0.3649203 -7.216 1.94e-12 \*\*\*

Educationsecondary school -1.0123455 0.2098659 -4.824 1.86e-06 \*\*\*

Onlychildyes -0.2474513 0.2145420 -1.153 0.249286

`Village-town`village 0.1678415 0.2354136 0.713 0.476192

`House-blockofflats`house/bungalow -0.2768813 0.2085654 -1.328 0.184917

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

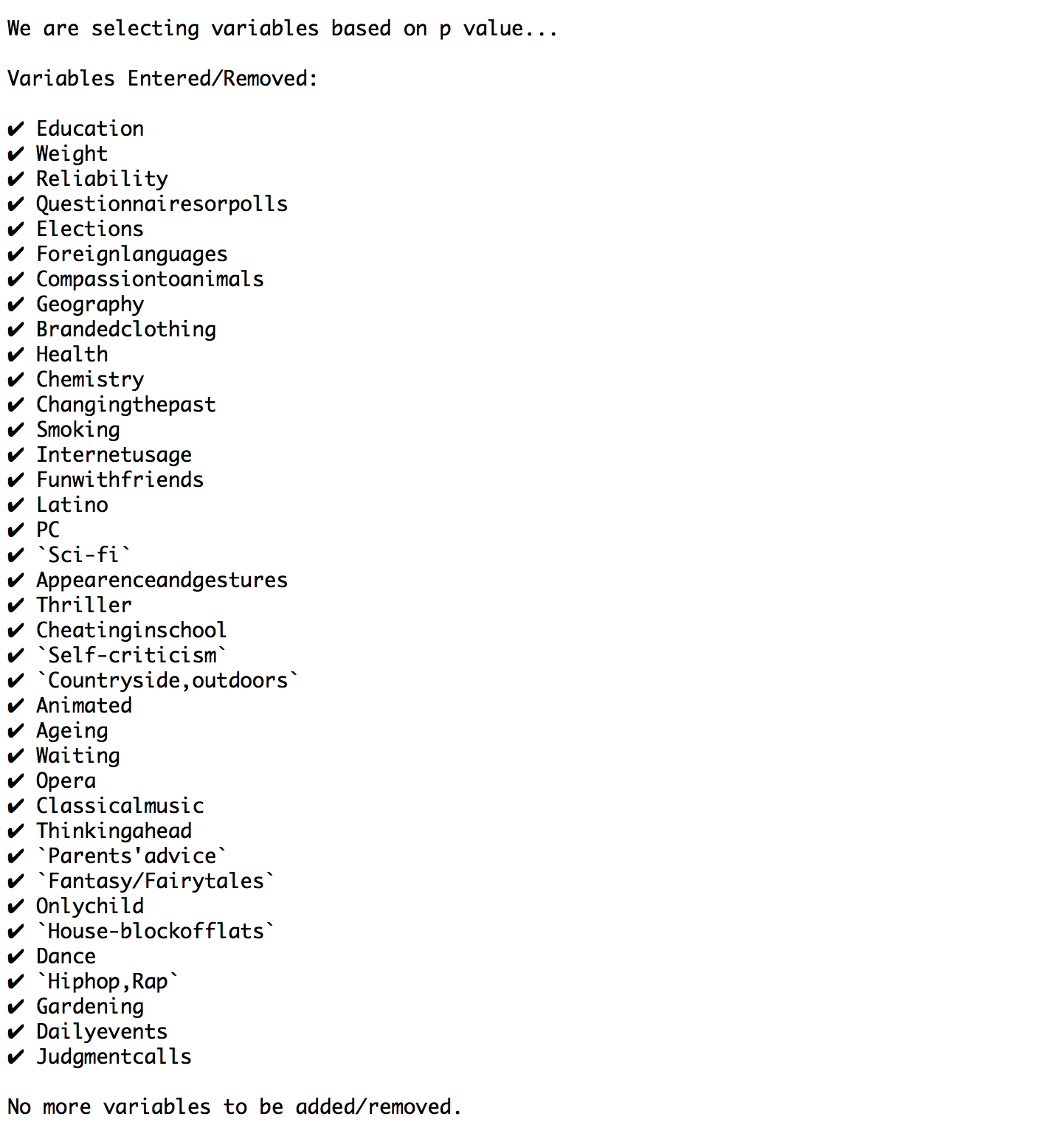
Residual standard error: 1.859 on 513 degrees of freedom

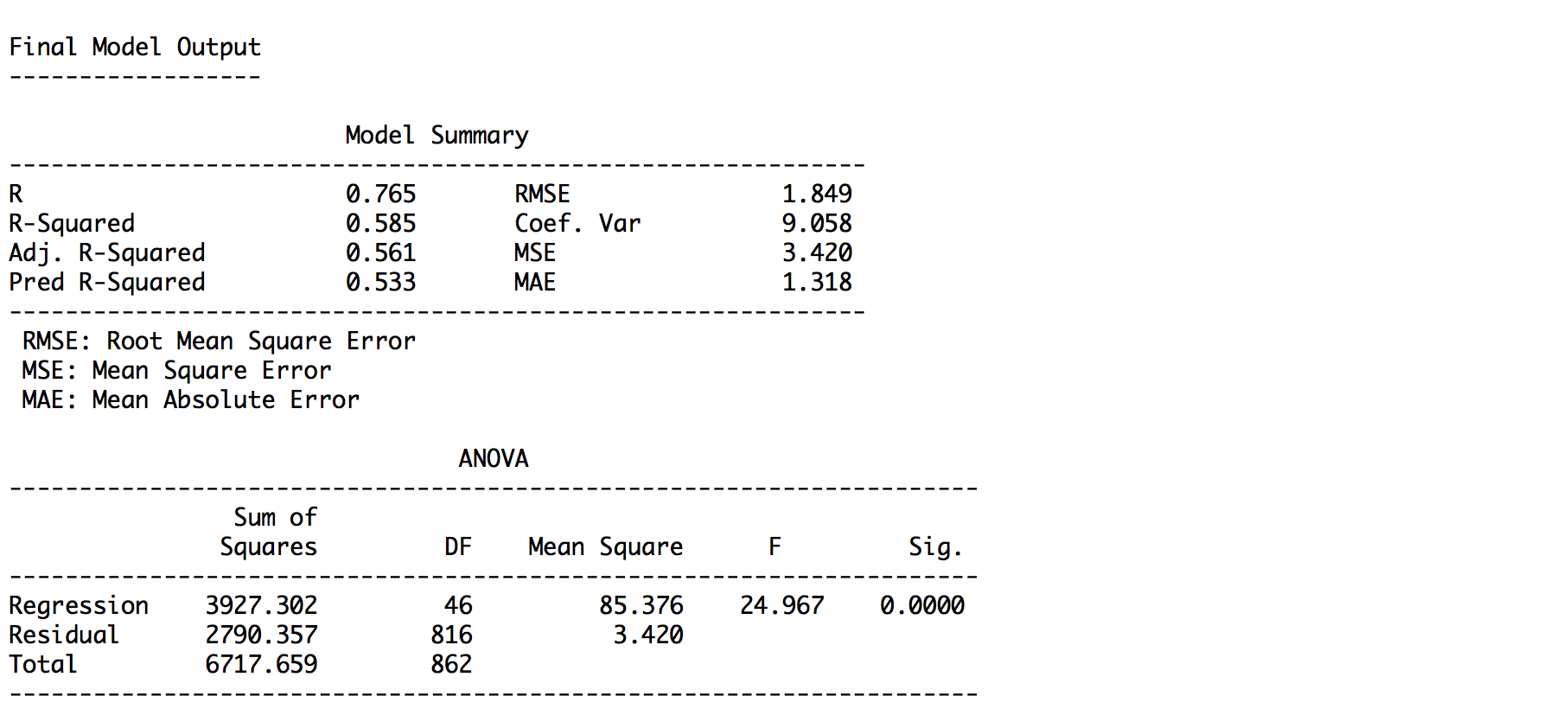
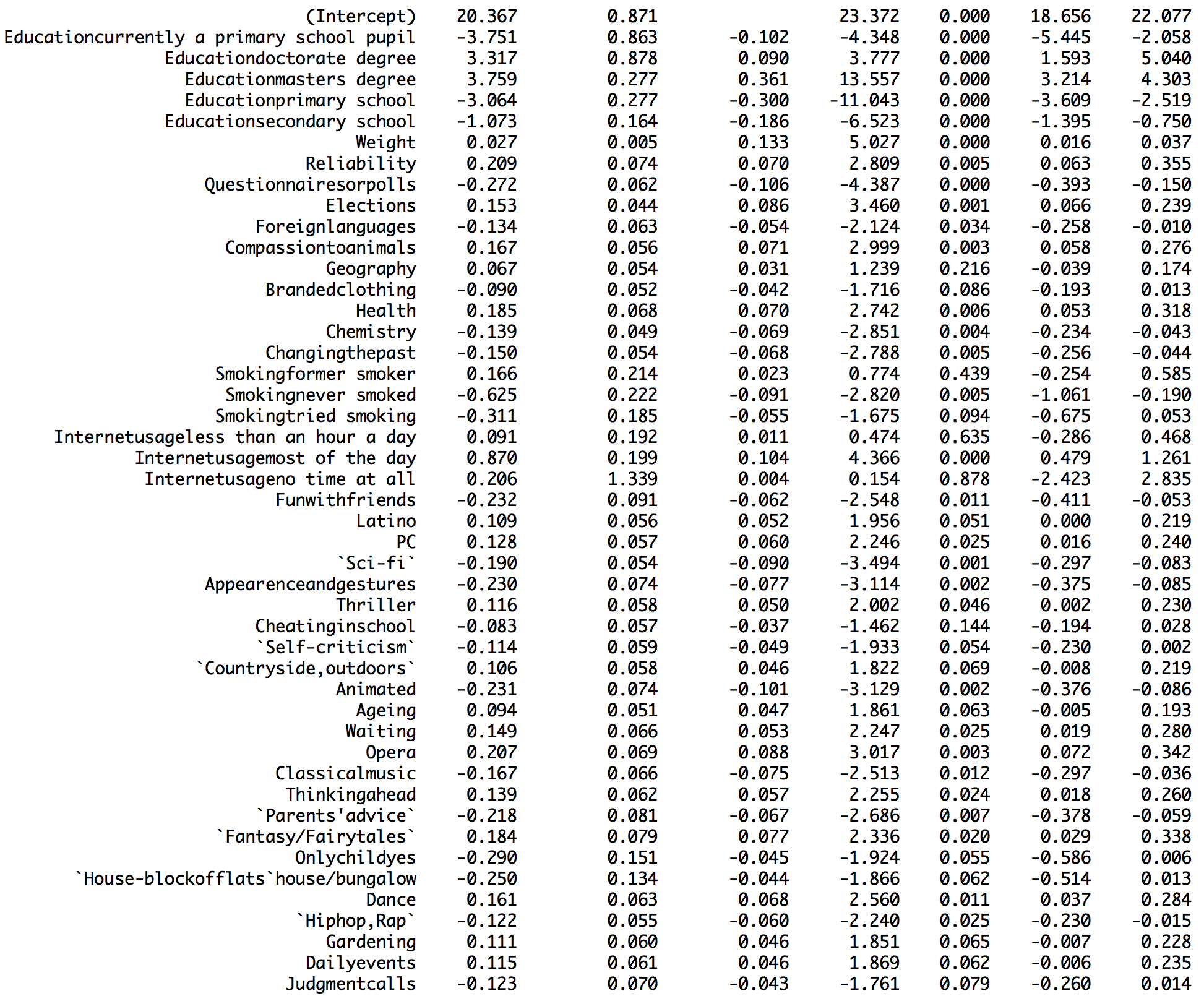
(336 observations deleted due to missingness)

Multiple R-squared: 0.6472, Adjusted R-squared: 0.5371

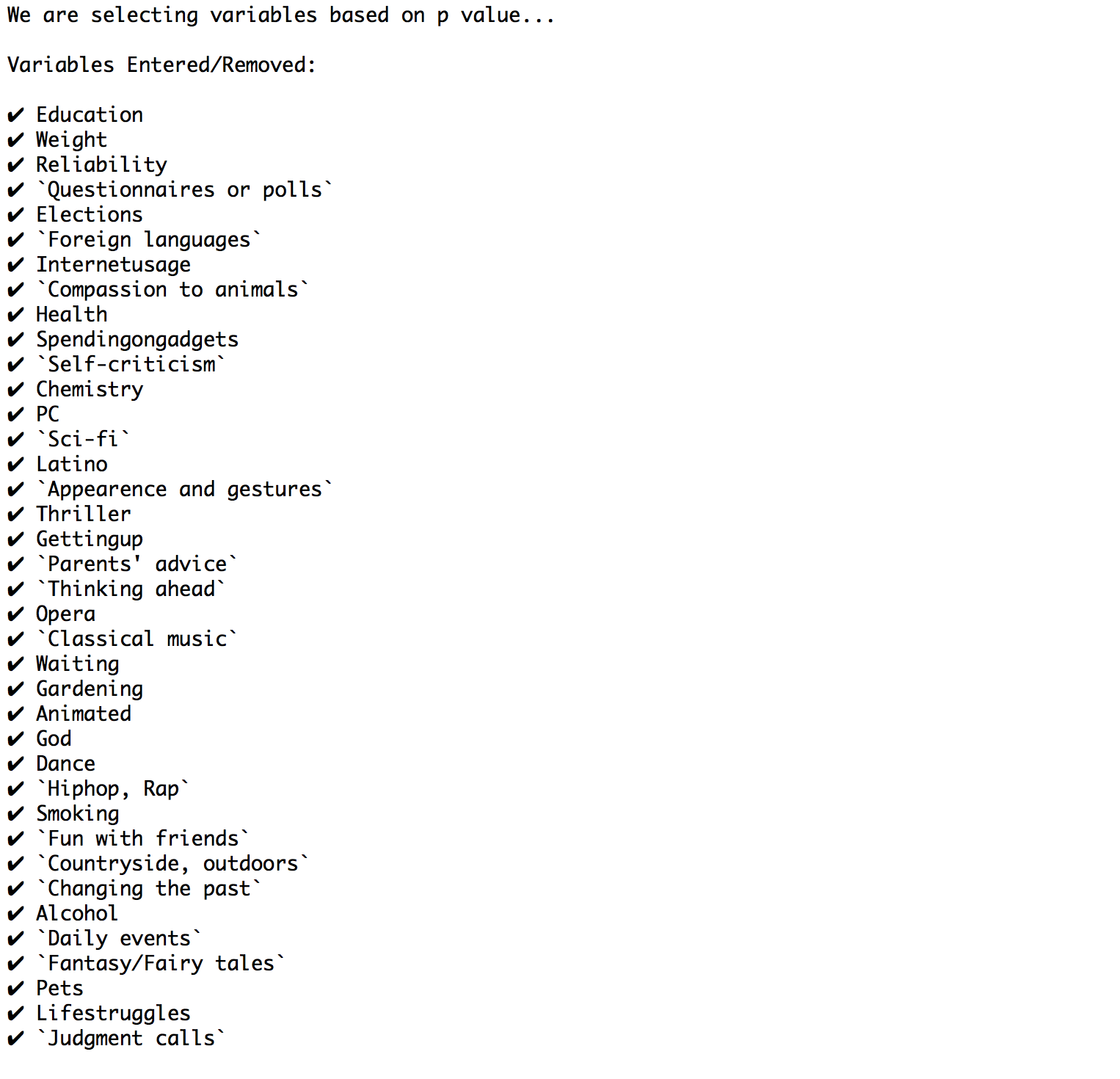
F-statistic: 5.881 on 160 and 513 DF, p-value: < 2.2e-16

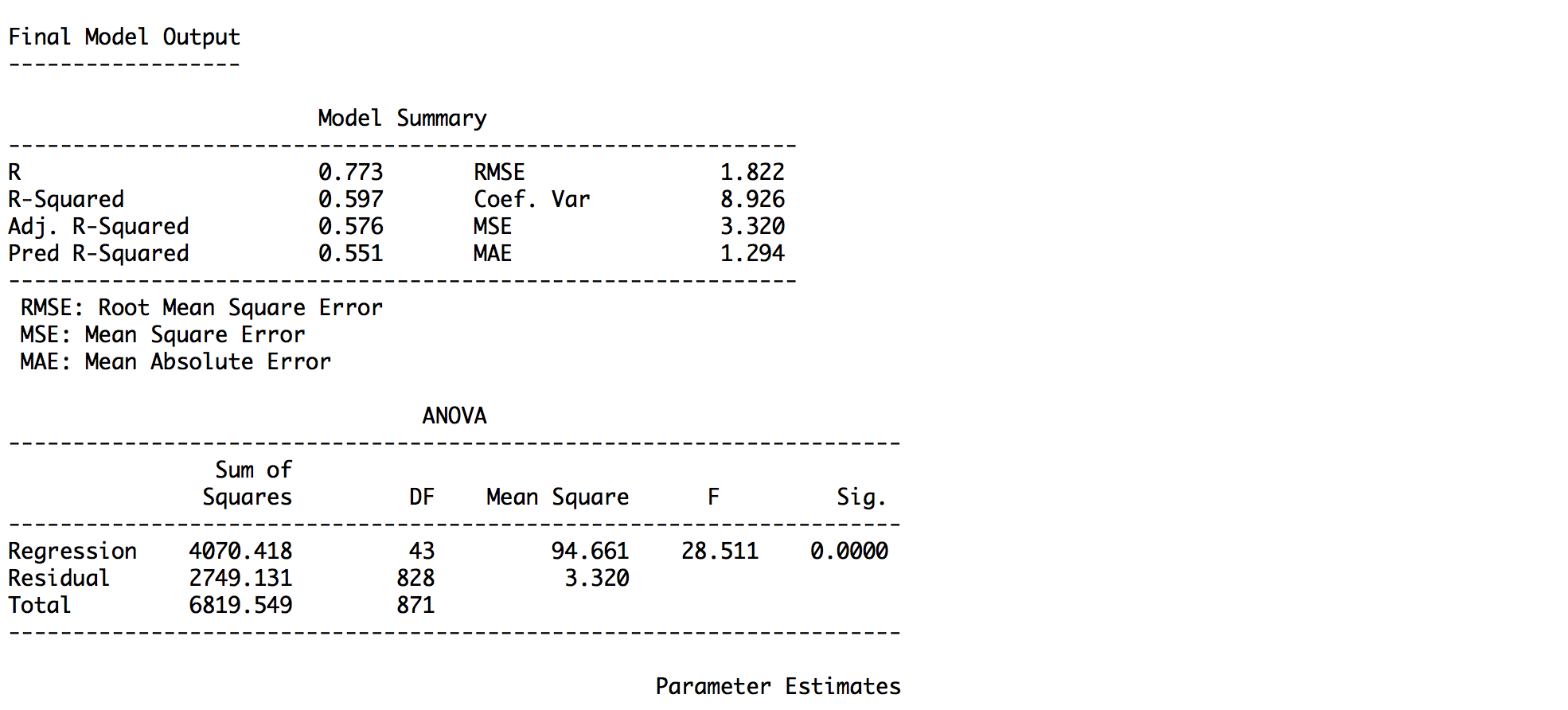
Appendix S2

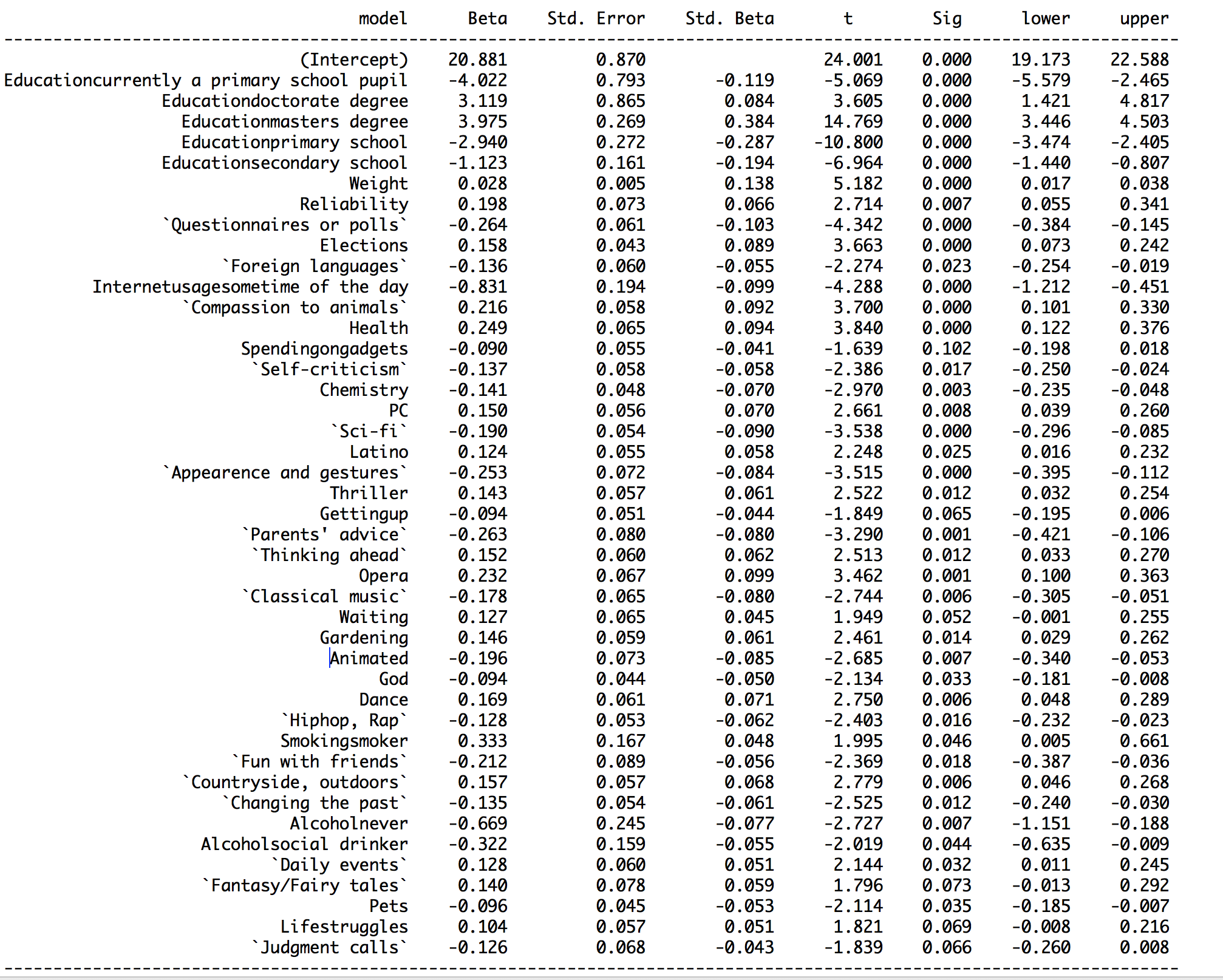




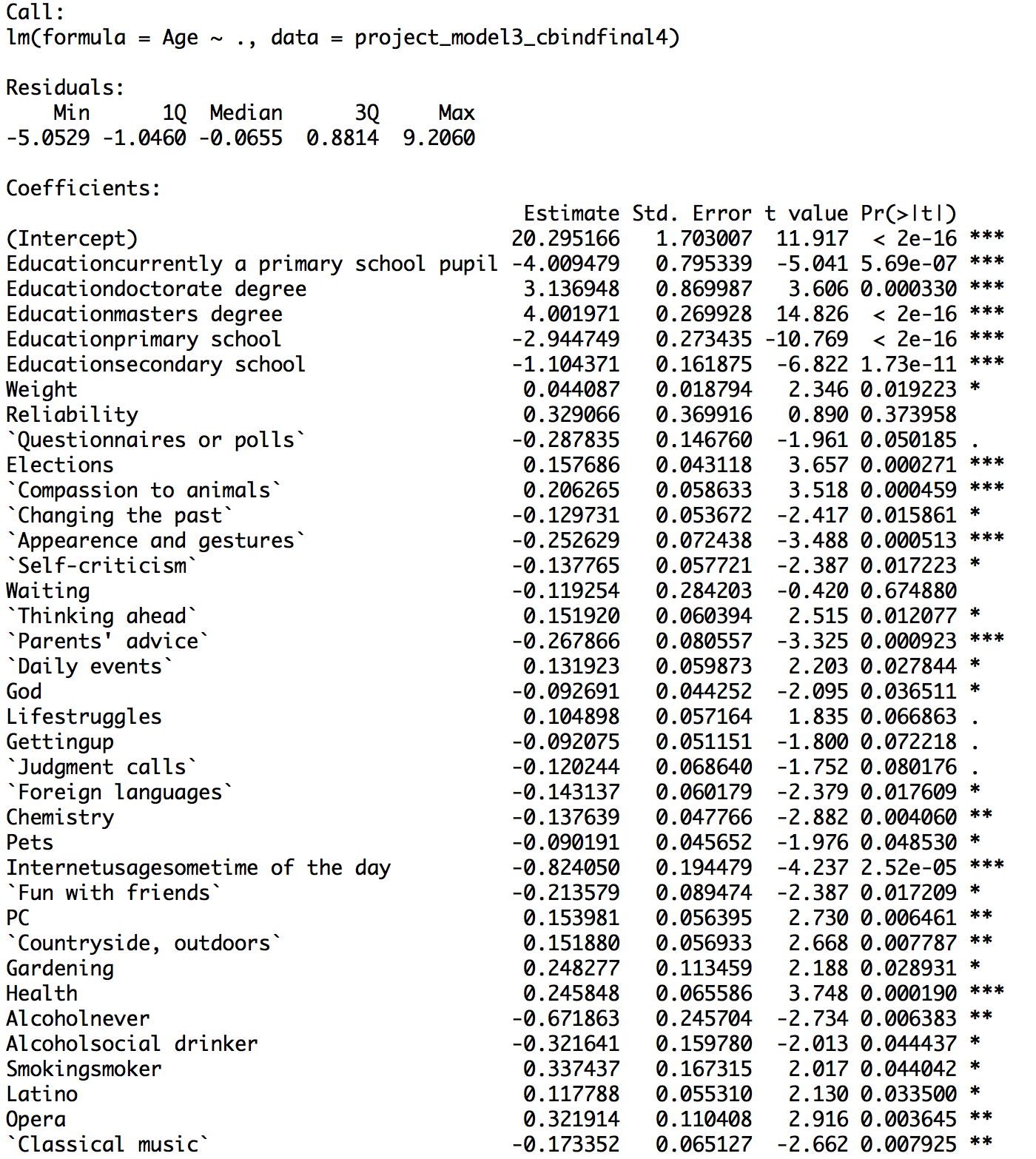
Appendix S3

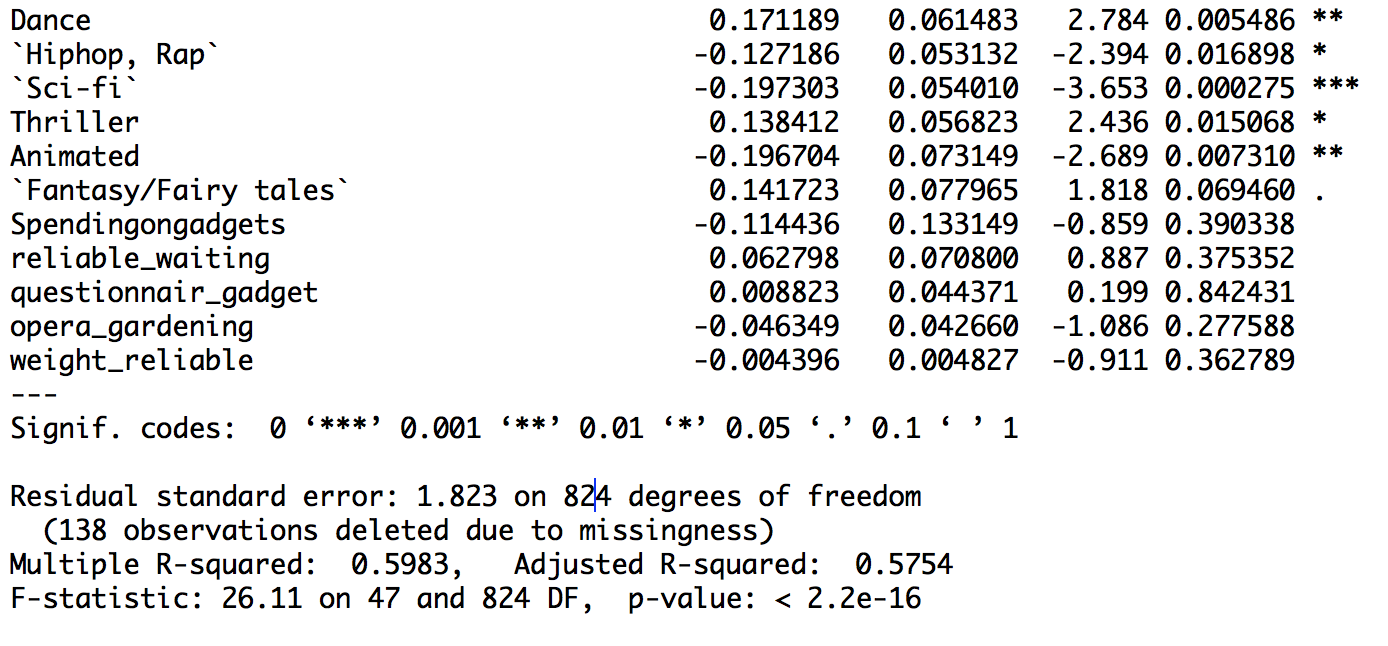






Appendix S4





Work Cited

Sabo, M. (2017). Young People Survey. Retrieved March 15, 2019, from

https://www.kaggle.com/miroslavsabo/young-people-survey