HW2 - Nachi Lieder 314399114

Contents

Introduction

In this project we will evaluate a specific dataset and in specific answer the questions to 2 different models that we will assess.

The data we are dealing with questions patients several times over 3 days , and asks them about their thoughts of the future. Among the different sets of features we will be dealing with, we have :

- Day
- Time
- Age
- Gender (male/ female)
- A : agreeableness
- C : conscientiousness
- E : extraversion
- N : neuroticism
- O : openness to experience

Where the last 5 are considered the Big Five traits, and the patients are asked to rate themselves from 1-7.

In the first question , we will be asked to build a model that will answer a binary classification question of whether a patient will think of the future as a funtion of the time of day. We will assess the models results and determine whether there is some sort of statistical significance.

In the second model we will address a question of predicting how many times from the total amount of times that the observants were asked , did they think of the future. We will assess different types of models such as poisson models and quasi-poisson models and perform some sort of comparison.

Among the different analysis we will be performing are over-disperssion tests , T tests on means of distributions, anova tests to compare the different models, model selection using Stepwise-AIC . Lets start off with the data exploration prior to the predicitve sections.

File reading

Lets look at the following subset:

Lets look at the time . We are talking about time since midnight , having a maximum at 86400 (seconds in a day). As we see in the plot , there arent any non-rational values in the set. What we can see are three observations that were taken sometime near (0:00 - 0:30). Since the data is somewhat continuous and there is a cycle that needs to be taken into place (value 1 = 86401 = 0:00:01 AM) - I will convert these values to 86400+X to represent midnight since then they will be scaled the same way as the rest and the partition wont be drastic. The next observation occurs around 30000 seconds later => (8:30 AM). There are no missing data points , which is encouraging.

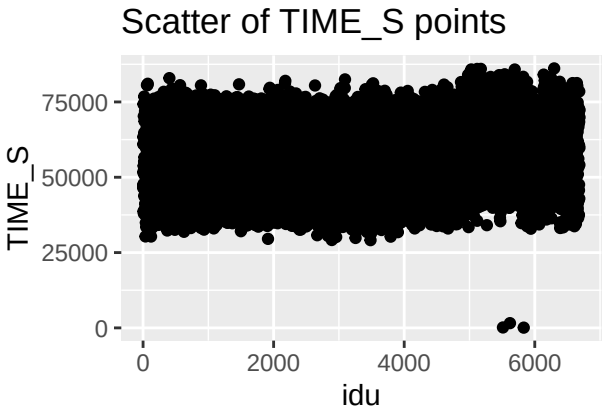
As for the Day variable , there too is no missing data , and the range looks fine [1,3] We are also talking about almost equal sets, having all above 200, though Day1 might have the most.

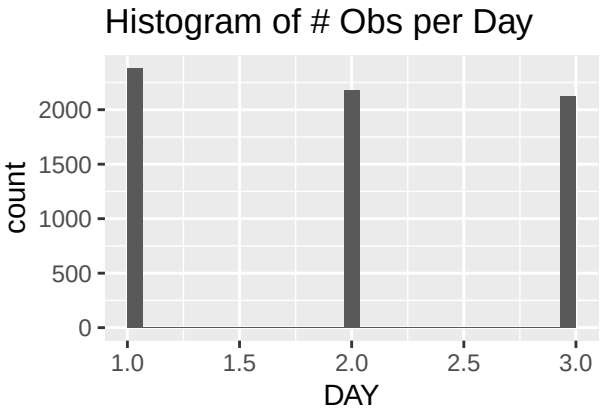
Regarding the Subject variable , we are looking at 492 subjects, and I wanted to inspect the average time that each subject has its observations measured. Below is the scatter plot with per subject the average time_S. What is interesting here is that there is a large variation of measurement times. We can see averages per Subject_i ranging from 30000 to 80000. We will take this into mind when analyzing the results of our analysis since this can be taken as a bias for the model.

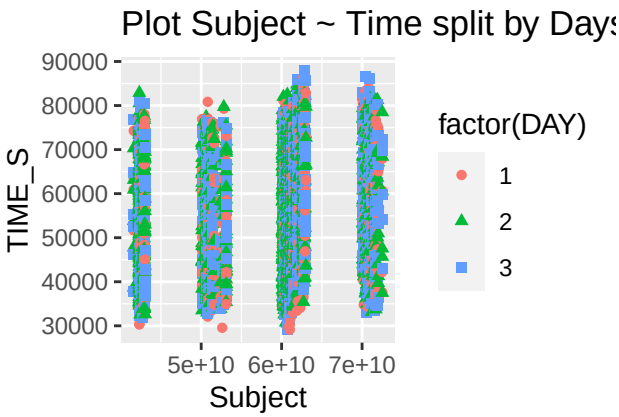
Also , I verified the diversity of the time per day per subject , in the following plot we can see that the there is no pattern of time of the day relative to the day of the observation and the subject.

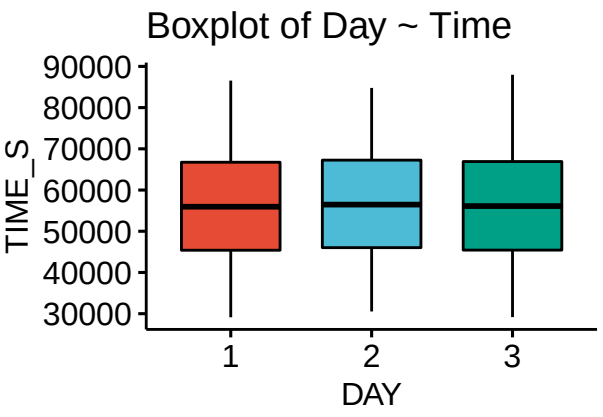
To conclude this part of the analysis I validated that per day , the average time is somewhat similar. In the boxplot below we can see that the medians are very close and we can assume similarity between the distributions of the three classes.

This concludes the exploration analysis per the 4 features









Here we can see the results of the t test that I perfromed between Time and the different DAYs , validating that there isnt a difference between the different days and time's mean. We can see that th P val is high and we can assume similarity between the groups.

On the other hand , we can see here the test of the Future ~ Day , wehere we can see that it is quite significant the difference between the behavior of the subjects in the days with respect to their Future. The P val is quite low here which confirms that there is a difference that we should consider in the future.

Table 1: T Test future ~ DAY

Effect	DFn	DFd	F	p	p<.05	ges
DAY	1	6684	0.369	0.543		5.53e-05

Table 2: T Test TIME_S ~ DAY

Effect	DFn	DFd	F	p	p<.05	ges
DAY	1	6684	26.3	3.07e-07	*	0.004

Logistic Regression Stage

Lets assess a classic Logistic Regression with the Time as the predictor to preedict the probability of a future thought..

We can see that standalone , the time has a pretty siginificant Z val with -6.3 , and a very low P val.

Below we also have the plot of the curve which is very lean , and from the points witin the scatter you can see that the diferentiation is not very significant. Also , the QQplot of the residuals presents a very non normal distribution of the residuals.

What is interesting here is that while predicting the sample test set , we get a range of probabilities from [0.21,0.37]. A suggestion would be to enforce the model with some more predictors. There seems to be a bias towards the results , where we would expect a broader spread of results rather than the small range. Below we can observe a histogram of the resulted probabilites. We can see that within the range [0.25,0.35] it is prety well spread, and seems gaussian , but not normal.

We can also see that while splitting the sets to train and test , to evaluate the accuracy , we can see that this specific model is able to acheive an accuracy of 70% . This could give us an indication of the fittness of the model and another benchmark and pivot point for model selections in the future. Although this model was not fine tuned , it is a benchmark that we can pivot off of.

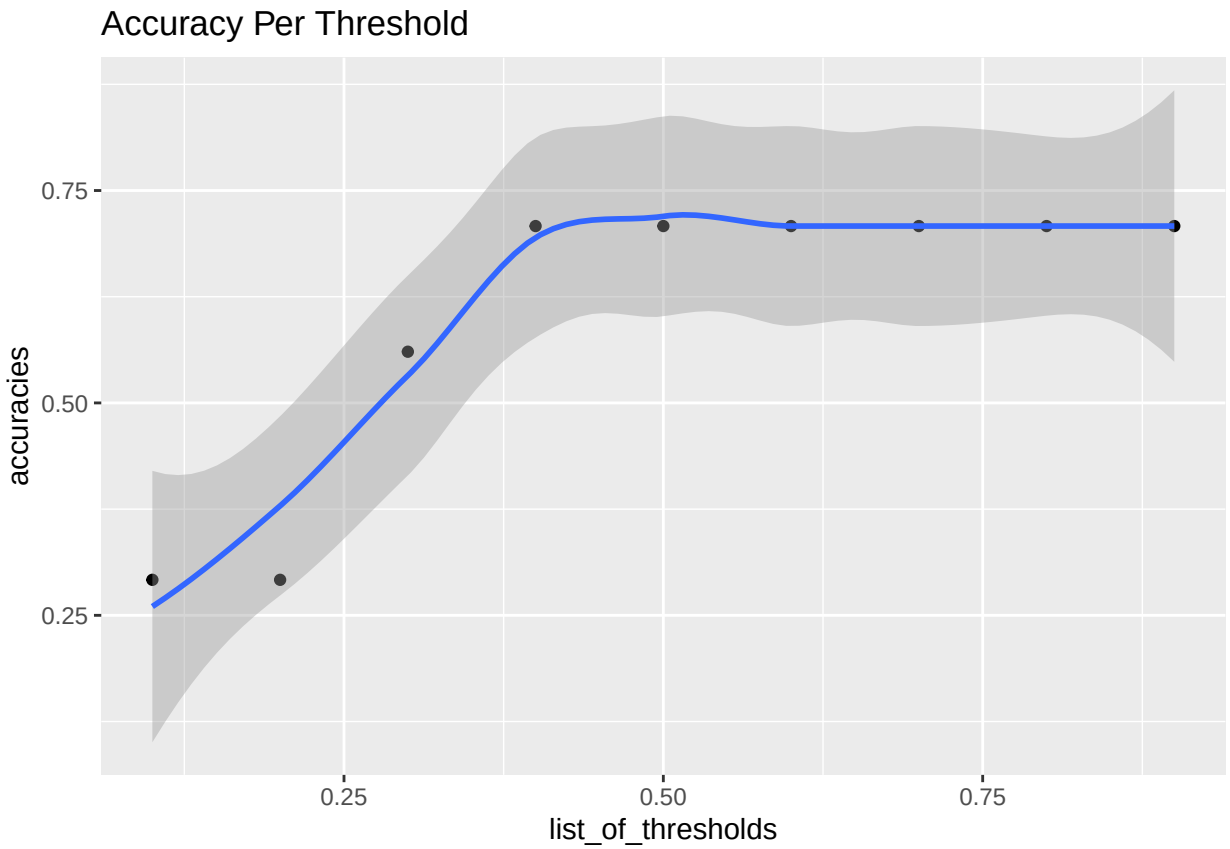
The improvement of the dispersion from the null disperssion is quite small , shifting from 8073 to 8033.This shows that the additional input of our single predictor is not quite good enough. our score is equal to 0.04 .

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.893	0.0271	-33	8.08e-239
TIME_S_standardized	-0.171	0.0272	-6.3	2.95e-10

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	8073 on 6685 degrees of freedom
Residual deviance:	8033 on 6684 degrees of freedom

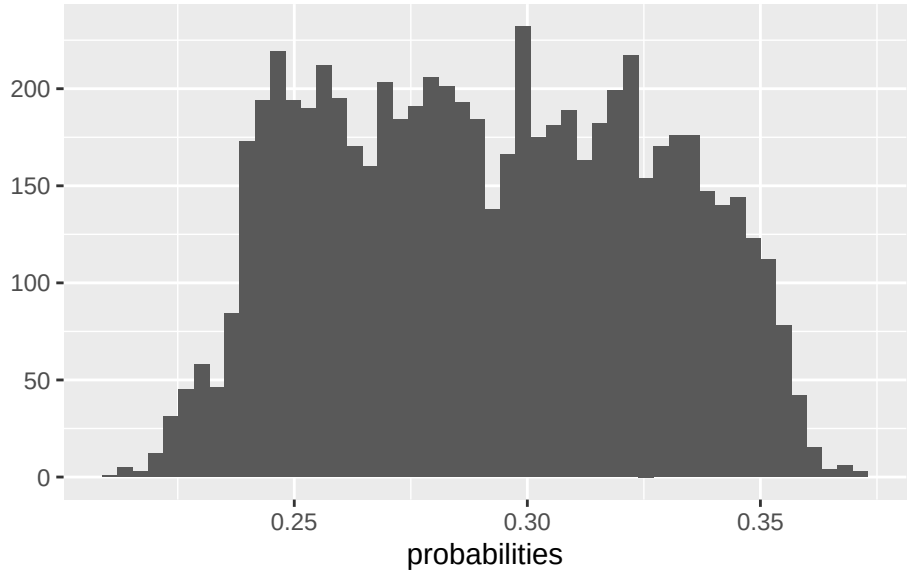
Below we can view the transition and behavior of the model in different thresholds. We can see that once we hit the 50% threshold, we reach a plateu , where this can be due to the distribution of the probabilities which we will address. We may consider a different threshodld to uphold different presicion/ recall scores , where the balanced accuarcy may be around 0.4.



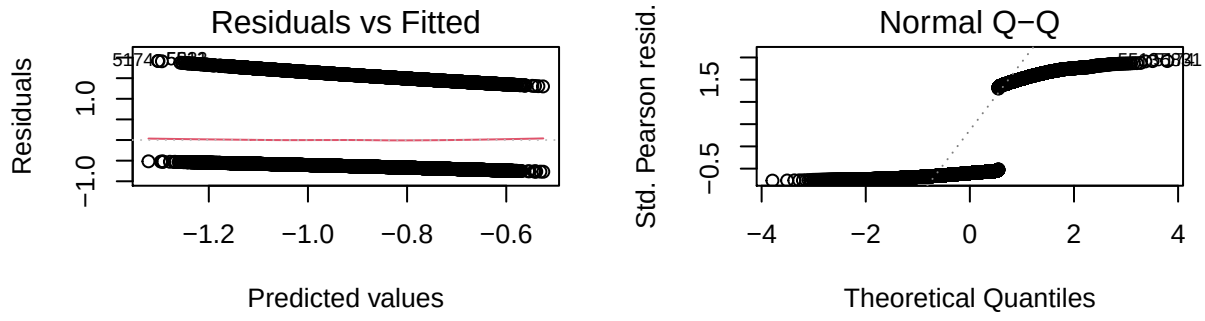
We can see that generally speaking the range is not well spread , and that probably the model is not complex enough to create a true prediction only using the one predictor and needs more complexity and depth to define a well differentiation. We would assume that given a ratio od 1:3 between positive and negative targeted values would give a better spread of fitted values , but here we see a dense fit around the distribution of the target.

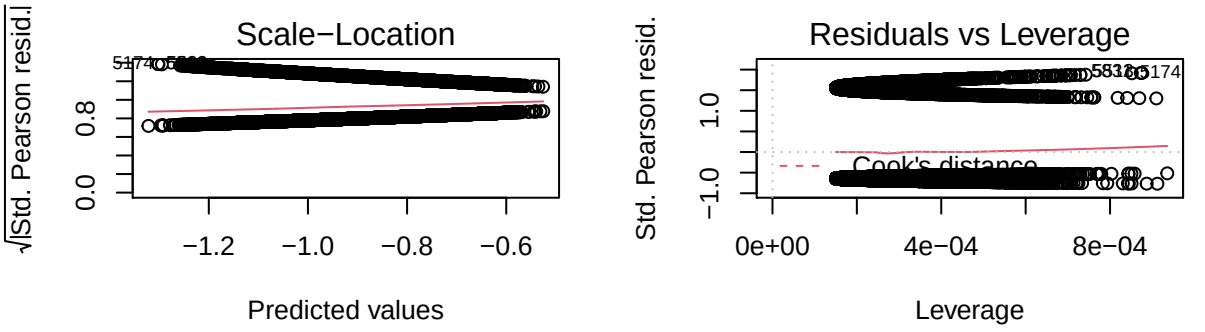
Since this is indeed surrounding the 0.3 (which is more or less the distribution mean of the models target), we can assume that model took into account the distribution “too much” and we would need to create a stronger regularization to the model of pad it in order to save this mean of the distributed probabilites , but spread the values over [0,1] with a wider gaussian projection.

Another option here would be to change the threshold from 0.5 to 0.3 per say and assess accuracy.




```
## [1] 0.004960184
```





The next step is to try to evaluate the targeted variable using a mixed model with time of day as a fixed effect

When observing the standard error we can see that given the day within the subject ,there is a much lower variance , than just the subject as it is. There is a differnce of 0.08 to 0.53. This means that the model and

its variance can be explained by fitting an intercept specifically per subject , and that the main noise indeed comes from the Day.

In addition , while looking at the fixed effects , we can see that both the regular intercept and also the Time (standardized) have significant estimates. So per subject we recieve its own intercept based on the usual intercept + the time standardized which is tailor made per subject. The coefficients are obviously equal to all in both models.

We can see that the AIC in the mixed model evaluation is slightly lower than the linear model above , shifting from 8037 to 7803 , indicating that this model is slightly more predictive and well defined.

The addition of adding the level of the DAY into the mixed model gives only a slight if not any improvement. What is interesting is that adding another model with only the Day as the mixed random variable , we get a higher AIC which indicates that the major factor here is indeed the subject (which makes sense)

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: future ~ TIME_S_standardized + (1 | Subject/DAY)
##    Data: newdata
##       AIC       BIC    logLik  deviance  df.resid
##  7803.723  7830.954 -3897.861  7795.723      6682
## Random effects:
##  Groups      Name        Std.Dev.
##  DAY:Subject (Intercept) 0.2885
##  Subject     (Intercept) 0.7291
## Number of obs: 6686, groups:  DAY:Subject, 1379; Subject, 492
## Fixed Effects:
##         (Intercept)  TIME_S_standardized
##             -0.9946              -0.1903
```

To compare in depth these two models (linear model vs the mixed affect model) I performed an anova test which summarizess the comaprison. Here we can see the difference between the AICs , as well as the LogLikelyhood . Also notice the Chi Squared high level , and very low P val which conclude the rejection of the null hypothesis of equally contributed models in terms of the deviance and addition of new features. This concludes that the addtition of the predictor Subject and Day are significantly considered contributers.

In terms of agreement , it seems like the two models somewhat behave the same more or less , with one having a slight improvement. The difference here would be that for the advanced mixed model , we have a more tailor made model , fitting the intercept to have a more individual set , where each subject and observation would have its own additional contribution to the intercept. By lowering the generalization we are able to receieve slightly better results.

Table 5: Anova Test - Mixed model vs LR

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
fit.lr	2	8037	8051	-4017	8033	NA	NA	NA
mixed_model	4	7804	7831	-3898	7796	238	2	2.43e-52

Question 2

I will retreive the 7 given features (age , sex , and the big five features) and create an aggregated dataframe representing the aggregation grouped by subjects , and their descriptive analysis.

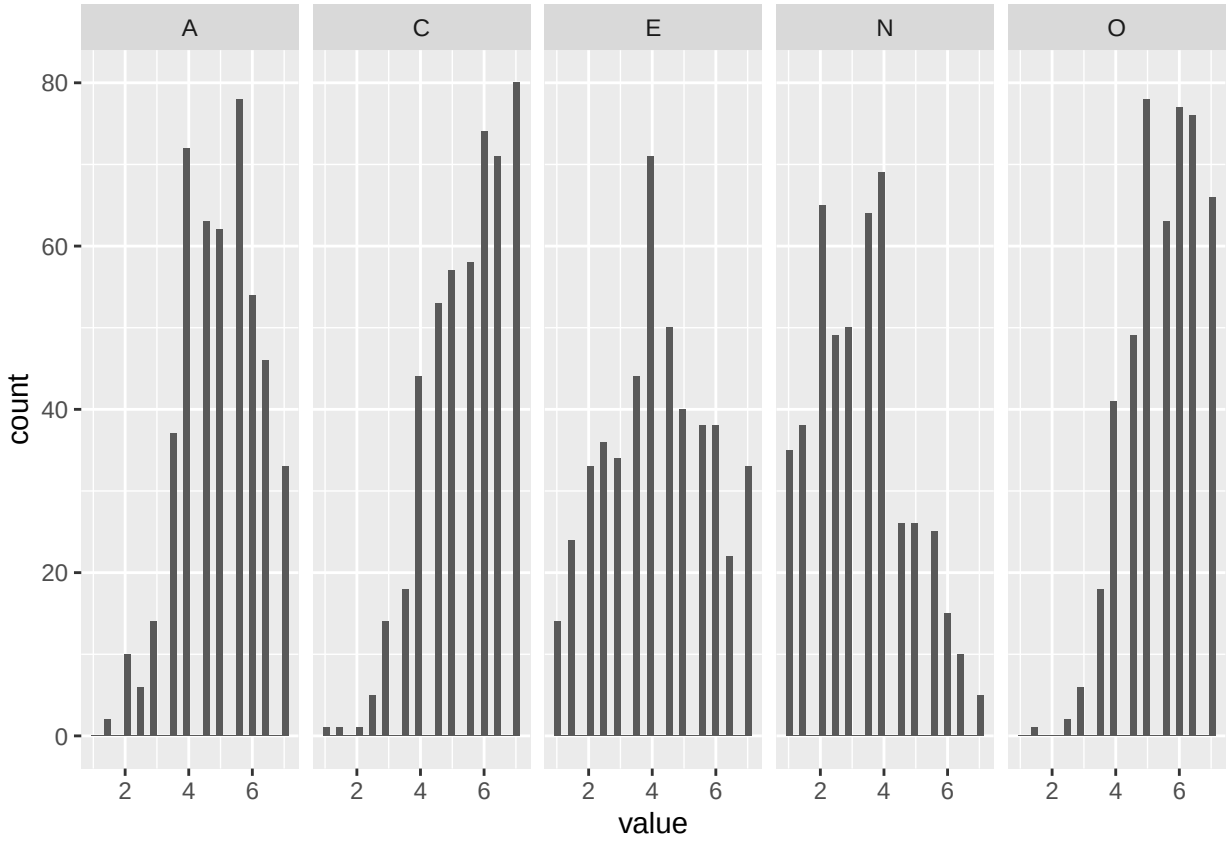
while plotting out the aggregation count , we can see that different subjects have differnt amount of observations. The smallest set was 1 observation per subject , and largest 18.

Lets try to observe each subjects big five by representing its categories in the form of means per subject. The following set of histograms per each category represents the means dsitribution . As you can see there are very different distributions per category.

For an instance , categories A, E are much more gaussian ,while the density of C and O are emphasized on the right, and category N on the left in terms of skeweness.

Since the scale of the big 5 is from 1-7 , we can say that traits O and C (openness to to experiences and conscientiousness) were found to have much higher ranks. These are self imaging traits that are expected to have higher values. Values of agreeableness and Extraversion are with slightly lower means in overall and representing a slighlty more unbiased set of questions. Neuroticism is considered a negative trait in some way, and like the first two , is expectd to have a more biased look at themselves in terms of self image.

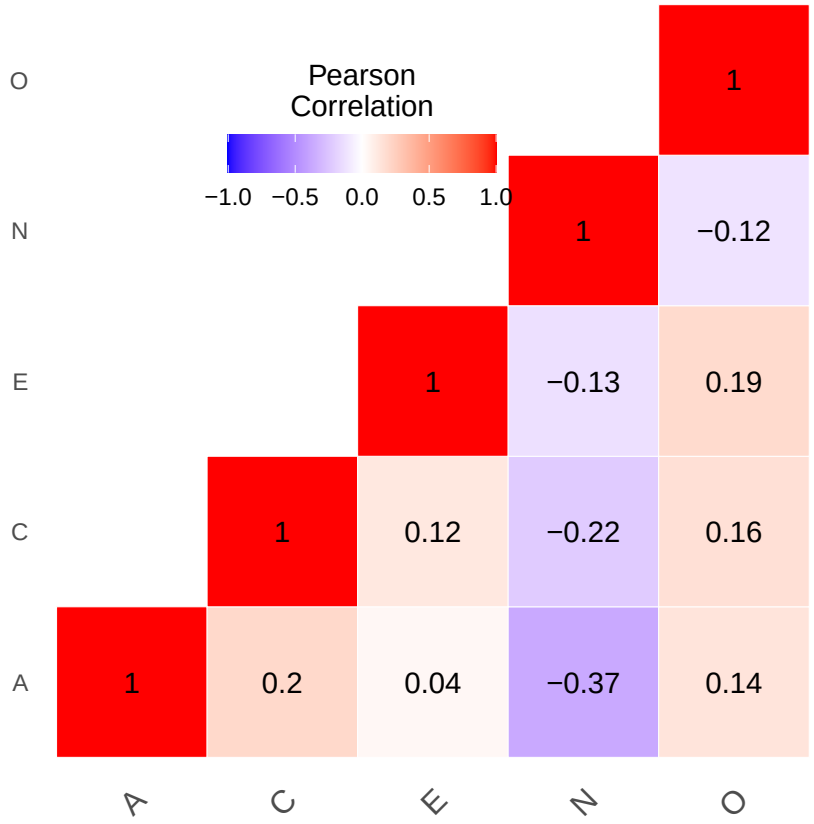
Lets take this behavior under consideration while building our model.



Lets build a model in the form that will answer the following question: Predict the number of Future thoughts per subject using the descriptive information regarding the big 5 traits + age and gender.

We will start off using the means of the Big 5 per subject, and add a numeric value of the age and a categorical predictor of thr gender. I will attempt to fit Poisson model and evaluate the following. Next I will add and define the offset of the number of observations per Subject into the model. Last, I will fit a Stepwise AIC model to perform a proper model selection and filter the remaining contributing columns.

First off we will observe the histogram of the target variable, and a heatmap of the pearson correlations between the averages of the Big 5. Bellow are the plots. We can see from the heatmap that there is a slightly strong negative correlation between N and the other 4 traits , in particular with A. We would expect the model selection to remove one of these two due this finding.



From the results Below we can see that the model selection has dropped the Gender and N predictors , and has resulted in a pretty significant prediction model.

Its interesting to see that the Z value for the age is negative , and significant reflecting on a negative relation to the predictor (lower the age , the more people think of the future).

Another interesting finiding is the relation between "O" and the target , where O represents Openness to experiences.We find that the test's results indicates that the higher the grade the observant gives itself to this predictor (O) , the more they think of the future. One of the more interesting facts that this trait includes is active imagination , which we can derive a strong connection to future thinking.

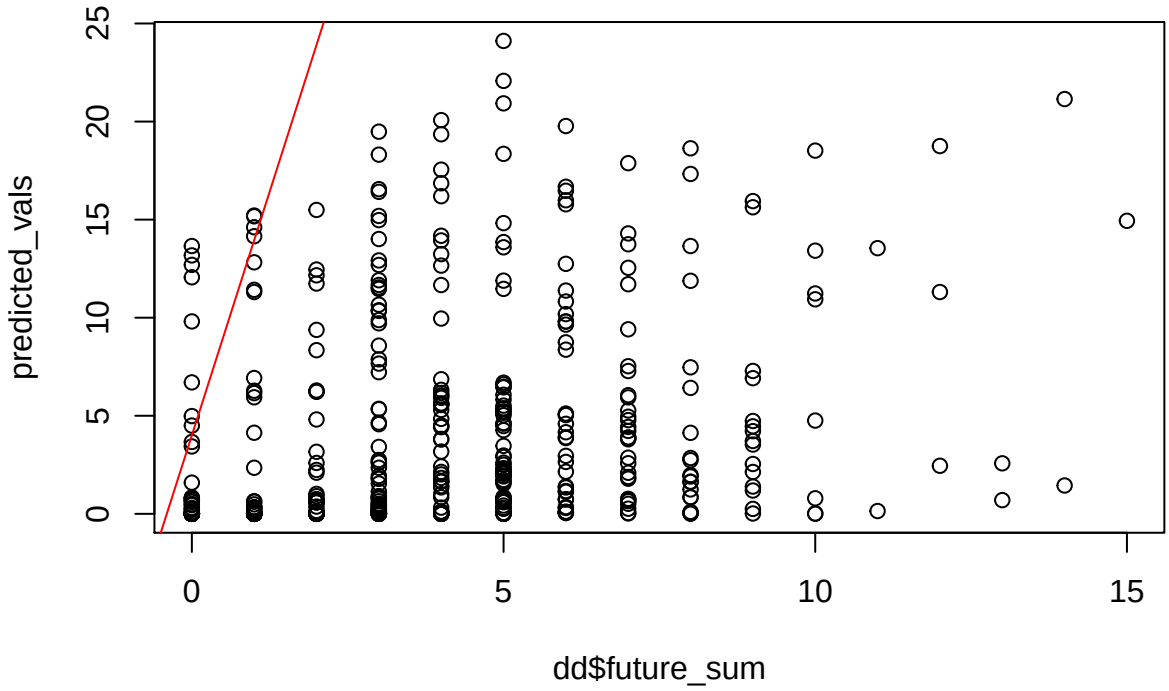
On the other hand , a negative relationship we see is with predictor "C" , being Conscientiousness. Under this definition , we see this as a trait of carefulness and deliberation. This is surprising, since we would expect that people that are careful and think carefully before they act would have a high relationship with the number of times they thought of the future.

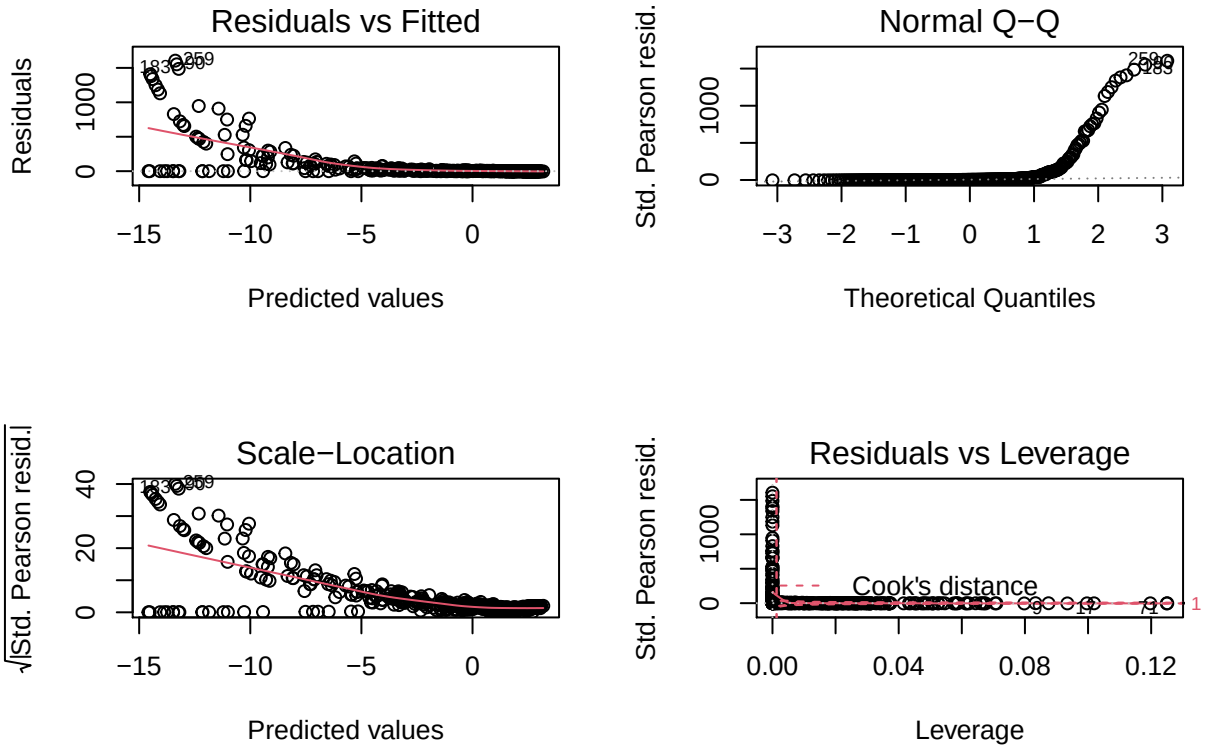
Lets also view the predicted vs actual values to esitmate how well this is fitting. We can see that the scaling here is off , in terms of the predicted being a bit higher in general than the actual. Or in other words , the fitted model performs some sort of bias that we need to take under consideration.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.2	0.171	-95	0
A	0.0512	0.0208	2.46	0.014
C	-0.0669	0.0207	-3.23	0.00123
E	0.0454	0.0152	2.99	0.00283
O	0.186	0.0228	8.16	3.37e-16
age	-0.00709	0.00235	-3.01	0.00261

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	7104 on 476 degrees of freedom
Residual deviance:	6991 on 471 degrees of freedom





I decided to look at the selected model and compare to the default one, using an Anova to test the deviance of the two models. We can see a Chi Squared that is high, which means that we will not reject the null hypothesis , and most likely treat these two models with the same deviance.

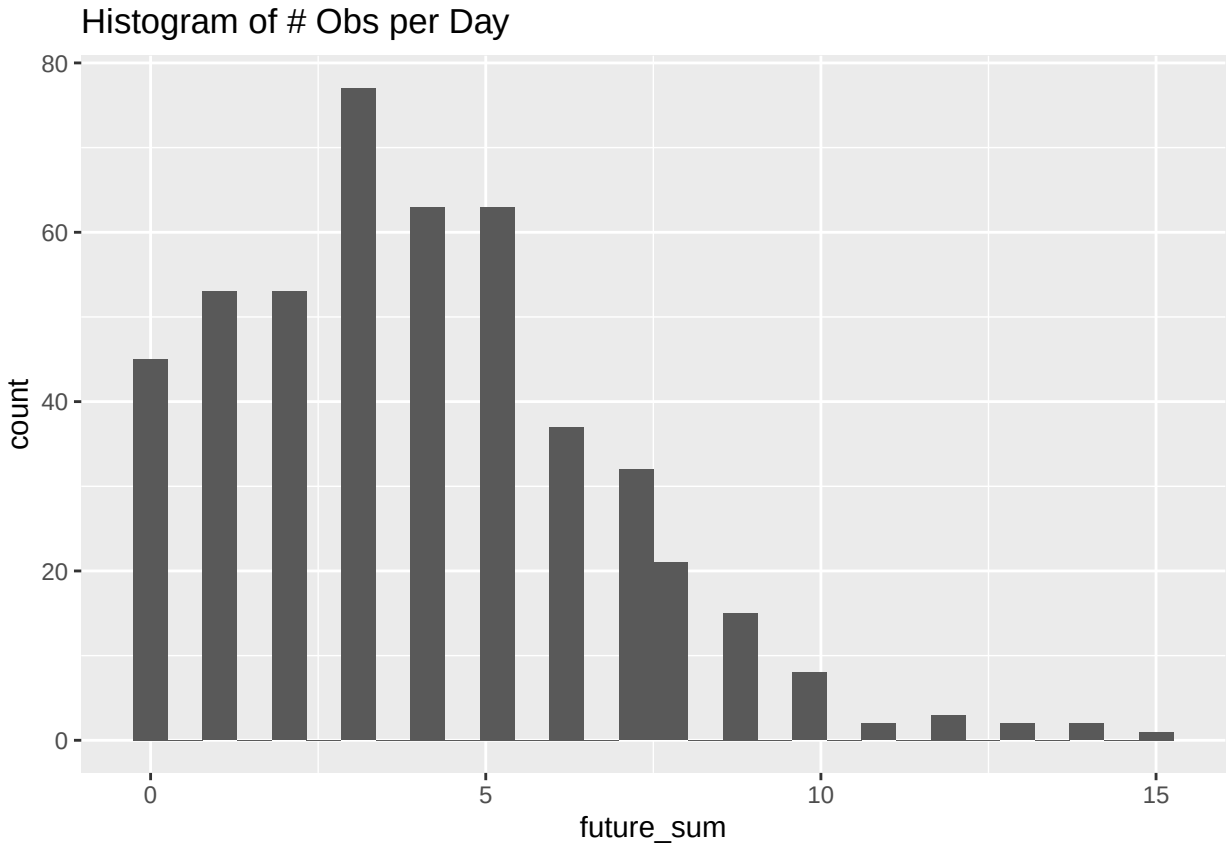
Table 8: Anova offset W/O Step AIC

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
469	6990	NA	NA	NA
471	6991	-2	-1.36	0.507

Next we will check out the model of the quasi poisson. As we can see , there can be an assumption that there is overdipersion in the data. I also performed an overdispersion test to verify whether the posson model fits well. We can see that the despersion ratio is not close to 1, and since its larger than one , it probably indicates an over dispersion. We also can identify this by the very low P value (~ 0). Also ,below is the histogram where we can visually identify the overdisperssion skewing towards the left with a slight inflation of zeros.

Therefor we will attemt a quasipoisson model and compare the results. First off ,we can see that the estimates stay the same (by definition) and due to the factorization of the quasi-poisson model , the std errors start to grow , and T values become insignificant. The may show that this model is not contributing to the fit. We can also see that the diserssion parameter is 53756.73 which is significantly high.

We are assuming that the fitted values will stay the same and that the only effect will take place in the factorization which we can see from the change in the std error.



- **chisq_statistic**: *25183301*
- **dispersion_ratio**: *53696*
- **residual_df**: *469*
- **p_value**: *0*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-16.2	39.7	-0.41	0.682
A	0.0512	4.83	0.0106	0.992
C	-0.0669	4.8	-0.0139	0.989
E	0.0454	3.53	0.0129	0.99
O	0.186	5.3	0.0352	0.972
age	-0.00709	0.546	-0.013	0.99

(Dispersion parameter for quasipoisson family taken to be 53756.73)

Null deviance:	7104 on 476 degrees of freedom
Residual deviance:	6991 on 471 degrees of freedom

Table 11: Anova test - Quasi vs Stepwise

Resid. Df	Resid. Dev	Df	Deviance
471	6991	NA	NA
471	6991	0	0

Summary - Q2

To sum up this research , we attempted to predict the number of times a patient has future thoughts, using the Big 5 traits plus some demographic information. I tried to use the poisson and Quasipoisson , and the results have room for improvement with additional touches. We can assume that there might be a different fit that may help.

Future Thoughts

We saw here an attempt to fit a poisson model on the given predictors. A couple thoughts for the future would be to possibly engineer the predictors and refine them. The current predictors were the avergaes based on a small set per subject, where the variance might be high per subject and this may have affected the predicitivnes and fittness. With more data per subject this might have been improved.

Another thought would be to add the STD per predictor in the model as well , and use both the mean and STD to predict. this would possibly enrich the dataset wihout harming the dimensionality too much.

We can attempt to fit a different type of model that may give us better results (such as a negative binomeal , exponential etc. . .) - in the code attached I attempted the NB and saw that there was no improvement (possibly due to the number of zeros)