STATS 202 – Data Mining and Analysis
Summer 2020

Final Project Write-Up

Author: Davyd Tamrazov
Instructor: Linh Tran

November 3, 2020

1 Data

Data from five randomized controlled trials (A,B,C,D and E) for patients with schizophrenia is provided. In each study, patients are split into two groups: control and treatment. The former group is given a standard medication to treat schizophrenia, while the latter is given a new anonymized drug. The goal of the studies is to determine the effectiveness of the new treatment by measuring schizophrenia pertinent symptoms in both groups over a period of time.

Schizophrenia symptoms are recorded using the Positive and Negative Syndrom Scale (PANSS). This scale allows to quantify 30 different symptoms that are grouped into three classes: Positive (7 symptoms), Negative (7 symptoms) and General Psychopathology (16 symptoms). Each assessment record is assigned a unique ID, while the assessor, patient and site are also identified with a number. Other attributes include the country at which the assessment took place, assigned treatment group as well as the day of assessment with regards to the initial baseline visit.

Additionally, for the the first four studies (A through D), an outcome of the audit outcome is included, identifying if the assessment is potentially erroneous. As such, each assessment is assigned either Passed, Flagged or Assigned to CS label. The latter two labels are referred to as failed from this point onward in the report. Finally, while studies A through D include all of the assessments, study E omits those conducted on week 18, which will have to be predicted as a part of this project.

2 Treatment effect

In order to determine if the drug works better than the generic medication, the decreasing trends of the PANSS score have to be compared between control and treatment groups. The following hypothesis test is set up:

H_0 : PANSS score trend over time is the same for control and treatment groups;

H_A : PANSS score trend over time differs between control and treatment groups

Significance level: 0.05 (two-tailed)

Statistical model: `PANSS_Total` $\sim$ `(Study+VisitDay+TxGroup)^2`

Statistical model outlined above includes the following variables and interaction terms:

- `Study` – indicator variable for the study, included to account for the fact that the intercept of the PANSS score differs between studies;
- `VisitDay` – assessment day that describes the overall trend of the PANSS score with time for all studies;
- `TxGroup` – indicator variable for the control / treatment group, identifies if the intercept differs between groups. However, given correct randomization performed for each study, this term should be statistically insignificant;
- `Study:VisitDay` – identifies difference in the slope (trend) of PANSS score decrease for each study;
- `Study:TxGroup` – identifies difference in intercept between control/treatment groups. Similarly, this should be statistically insignificant, given that patients are randomized well;

- `VisitDay:TxGroup` – identifies difference in slope between control / treatment groups. This is the most important term of the set up model, as it will identify if the null hypothesis can be rejected. It has to be mentioned that inference from this term will be based on the assumption that the difference in slope between two groups is equivalent for all studies.

Performing this statistical test on a linear model fit to studies A–E results in the following:

```
Call:
lm(formula = (PANSS_Total) ~ (Study + VisitDay + TxGroup)^2,
    data = study.complete)

Residuals:
    Min      1Q  Median      3Q     Max
-50.245  -8.854  -0.667   8.541  69.352

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               102.083104   0.523782 194.896  < 2e-16 ***
StudyB                    -14.656090   0.763097 -19.206  < 2e-16 ***
StudyC                    -21.119407   0.566987 -37.248  < 2e-16 ***
StudyD                    -24.201689   0.712863 -33.950  < 2e-16 ***
StudyE                    -35.449877   0.758238 -46.753  < 2e-16 ***
VisitDay                   -0.301974   0.009968 -30.295  < 2e-16 ***
TxGroupTreatment           -0.919756   0.579621  -1.587  0.11257
StudyB:VisitDay             0.142948   0.011197  12.766  < 2e-16 ***
StudyC:VisitDay             0.193004   0.009985  19.330  < 2e-16 ***
StudyD:VisitDay             0.161543   0.010471  15.428  < 2e-16 ***
StudyE:VisitDay             0.262970   0.011197  23.486  < 2e-16 ***
StudyB:TxGroupTreatment    -1.383725   0.868920  -1.592  0.11129
StudyC:TxGroupTreatment     0.064417   0.641959   0.100  0.92007
StudyD:TxGroupTreatment    -0.216969   0.784632  -0.277  0.78215
StudyE:TxGroupTreatment    -0.497110   0.859508  -0.578  0.56302
VisitDay:TxGroupTreatment   0.006034   0.002082   2.898  0.00375 **
---
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘’  1

Residual standard error: 14.07 on 22893 degrees of freedom
Multiple R-squared:  0.4466,    Adjusted R-squared:  0.4462
F-statistic:  1231 on 15 and 22893 DF,  p-value: < 2.2e-16
```

From the above results the it can be first validated that all of the studies have been properly randomized, since the intercept does not vary between control and treatment groups for any of the studies. The effect of the control / treatment group on the PANSS-Visit Day slope is assigned a p-value of 0.00375 << 0.05, indicating that this term is statistically significant and is not due to random variation. Consequently, null hypothesis can be rejected, suggesting that the trend over time between two groups is different. However, it is also important to consider meaningfulness of the result. Coefficient of this term is circa 0.006 and, thus, the influence of the group on the final PANSS estimation is very limited, on the order of one decimal place. As a result, due to the effect of the group not being meaningful, it can be concluded that the drug does not provide an improvement over the generic medication.

3 Segmentation

This section aims to discover different types of patients with schizophrenia by clustering these patients into k groups using K-means clustering algorithm. Positive (P), Negative (N) and General Psychopathology (G) symptoms are aggregated into the respective groups by summation of the corresponding scores. This is particularly important for the clustering, as dimensionality of the problem has a significant effect on the accuracy of the clustering. Thus, patients are clustered in three dimensions (P,N,G).

Since only the above mentioned variables are used as features, all of the studies, including study E, are used. The complete dataset is then filtered to include only baseline measurements (day 0) in order to capture the pre-study status of the patients and identify distinct classes of schizophrenia patients prior to any treatment effects. Finally, all of the features are scaled and centered in order to eliminate the effect of the scale on the distances.

Numerous methods exist to determine the optimal number of clusters for a given dataset. Here, `factoextra` library in `R` is used to perform Elbow method. This heuristic method calculates total within sum of square multiple times and plots this value against the number of clusters. The optimal number of clusters can be determined by identifying an "elbow" where increasing the number of clusters does not significantly improve the model, thereby suggesting that beyond that point, more clusters would overfit the data.

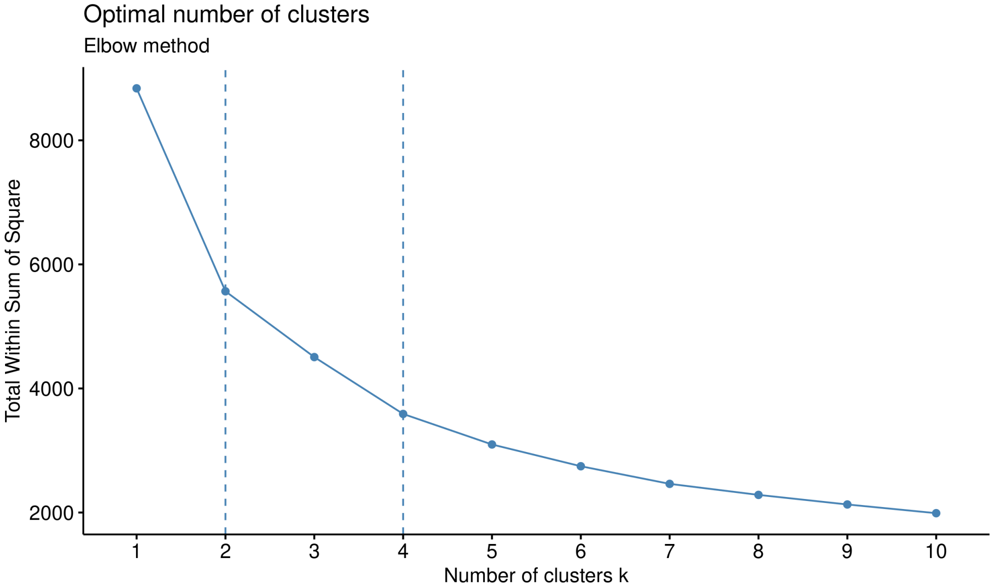

Figure 1: *Elbow method results, dashed lines identify potential elbows.*

As can be seen from the above plot, Elbow method produces a curve where the elbow can be interpreted in two ways: either at 2 or 4 clusters. In order to further investigate this, `NbClust` library is used to perform 30 statistical tests. The summary of the results is outlined below:

```
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 7 proposed 4 as the best number of clusters
```

```
    * 1 proposed 9 as the best number of clusters
    * 1 proposed 11 as the best number of clusters
    * 3 proposed 15 as the best number of clusters
```

In agreement with the Elbow test, this shows that most statistical tests also divide in the resulting optimal number of clusters. As a result, each of the options will be considered below. Segmenting patients into groups of 2 and 4 clusters results in the following cluster averages of P, N and G.

```
    K-means clustering with 2 clusters of sizes 1574, 1373

    Cluster means:
               P          N          G
    1 -0.6139333 -0.4013534 -0.6561101
    2  0.7038099  0.4601094  0.7521612
```

These resulting clusters can be broadly described as follows:

Cluster 1: Patients that have low indicators of symptoms in all categories (mild cases).

Cluster 2: Patients that have high indicators of symptoms in all categories (severe cases).

```
    K-means clustering with 4 clusters of sizes 887, 948, 778, 334

    Cluster means:
               P          N           G
    1 -0.7077116  0.5062496 -0.23009660
    2  0.3621999 -0.7474763 -0.08442211
    3  0.9240486  0.7806705  1.07460670
    4 -1.3010035 -1.0413099 -1.65244361
```

On the other hand, 4 clusters require a more specific description of each group: These resulting clusters can be broadly described as follows:

Cluster 1: Patients with larger negative but lower positive symptoms (moderate, negative dominating cases);

Cluster 2: Patients with larger positive but lower negative symptoms (moderate, positive dominating cases);

Cluster 3: Patients that have high indicators of symptoms in all categories (severe cases).

Cluster 4: Patients that have low indicators of symptoms in all categories (mild cases).

Both clusters provide meaningful results from the clinical standpoint. However, the latter case with 4 clusters provides a more flexible description of the different types of patients, capturing more unique identifiers. This can potentially help to determine a more specific treatment type for patients in each group and generally is more useful for the clinical study than just splitting into high and low symptom groups. Thus, 4 clusters should be specified.

4 Forecasting

Kaggle User Name: `davydtamrazov`
Public Leaderboard Score: `6.12258`

Prior to setting up a forecasting model, the following preprocessing is performed on the data:

- *Studies A–D*: Assessment results (e.g. PANSS individual scores) of a patient obtained on the same visit day are combined into a weighted average value. Weights are determined by the audit outcome, assigning a weight of 1 to the assessments that have passed the audit and weight of 0 to those that have failed. The aim here is to get a unique value for each given observation day for each patient. Meanwhile, previously described weights prioritize passed observation on that particular day, based on the assumption that additional assessments on that day were performed in order to redo the potentially erroneous results from the previous assessments. The average becomes weighted only when a combination of passed and failed exists, while a simple average of all same-day assessments is taken if all have the same audit results. Thus, there should not be no significant bias introduced with this process.

- *Studies A–E*: Similarly to the previous section, Positive, Negative and General Psychopathology observations are aggregated into the respective variables by summing up individual symptom indicators of each group. Since the symptoms were initially grouped into those classes based on their feature similarity, aggregation allows to achieve several improvements. First, the dimensionality of the problem is considerably reduced, from 30 scores to 3 aggregated scores, what allows to potentially perform more accurate forecasting. Secondly, certain trends in the specific scores may be unidentifiable while aggregate scores will behave in a more stable way with time, meaning that increasing or decreasing trends can be more clearly identified. This will allow to reduce the potential error caused by misidentified trends in the individual components of each class.

In order to make use of the additive nature of P, N, and G scores that form PANSS total score when combined, a separate forecasting model is set up for each variable. Splitting up into three additive models allows to identify separate trends for each of the features, as the drug can have a distinct effect on each. As a result, a more accurate prediction of the PANSS score can be obtained. In essence, forecasting model is set up in a way that the feature value at time t_{i+1} is predicted given the feature value at t_i as well as times t_i and t_{i+1}. Additionally, a group identifier (TxGroup) is added to account for the varying trends between control and treatment groups. This can be formulated as follows:

$$\begin{aligned}\mathbb{E}\left(\mathrm{PANSS}_{i+1} \mid \mathrm{P}_i, \mathrm{N}_i, \mathrm{G}_i, \mathrm{TxGroup}, t_i, t_{i+1}\right) =& \mathbb{E}\left(\mathrm{P}_{i+1} \mid \mathrm{P}_i, \mathrm{TxGroup}, t_i, t_{i+1}\right) \\ &+ \mathbb{E}\left(\mathrm{N}_{i+1} \mid \mathrm{N}_i, \mathrm{TxGroup}, t_i, t_{i+1}\right) \\ &+ \mathbb{E}\left(\mathrm{G}_{i+1} \mid \mathrm{G}_i, \mathrm{TxGroup}, t_i, t_{i+1}\right)\end{aligned}$$

Training dataset is formed from studies A-D. For each patient, a list of features includes patient group identifier, current P, N, and G values at time t_i as well as these values at time t_{i+1} represented by the next visit. This, however, means that the last observation for each patient is discarded since there is no information available for time t_{i+1}.

Time-dependent nature of the change in each of the features is captured by the R factor, which is defined with the following formula:

$$R = \frac{t_{i+1} - t_i}{t_{i+1}}$$

This factor essentially calculates the gap between the given observation time (t_i) and the forecast time (t_{i+1}) scaled by the latter. The rationale behind creating R is to capture asymptotic nature of the decrease in the effect of the time gap between observations as time progresses. This factor can be thought of as the adjustment term from the previous observation given the forecast time. The dynamics of this factor is summarized with three cases below:

- As the time gap between two observations increases, while the forecast time (t_{i+1}) remains constant, value of R factor increases linearly. This represents the increasing difference between P, N, or G scores of two observations as the time gap gets larger;
- As t_i and t_{i+1} increase by the same amount, keeping the time gap between two observations constant, value of R factor decreases asymptotically. This captures the fact that with observations at later times, the difference becomes smaller as the effect of a certain drug reaches its limit and scores level out.
- As the observation time t_i remains constant, but the forecast time increases, value of R factor grows asymptotically, again representing the leveling out of the scores with time.

Finally, each score at time t_i is expanded to include polynomial terms up to 4th degree, aiming to represent non-linear relationship between observations. The resulting linear model for each of the scores takes the following form, where `data.complete` represents the training dataset described previously:

```
lm(formula = P.Next ~ (poly(P, 4) + R):TxGroup, data = data.complete)
lm(formula = N.Next ~ (poly(N, 4) + R):TxGroup, data = data.complete)
lm(formula = G.Next ~ (poly(G, 4) + R):TxGroup, data = data.complete)
```

In order to use this model to predict 18th week observations, first the 18th week has to be defined. Since visit times are inconsistent between patients, some skipping some of the visits and others having their visits delayed, 18th week for each patient is defined as follows:

$$t_{18} = \max\left(t_{\text{last}} + 7, 7 \cdot 17 + 1\right)$$

What the above definition says is that day of the 18th week visit has to be on the latter of:

- The first day of the 18th calendar week = 120. This will generally be selected if all of the previous visits happened at least a week prior to the first day of the 18th calendar week.
- A week from the last visit. This will be selected for patients who had inconsistent schedule. Allowing at least a week between visits attempts to mimic realistic thinking of the assessors aiming to obtain reliable measurements. This is a reasonable assumption, given that for the test study, the only observations included are those prior to week 18.

Since most of the patients have multiple previous observations, applying forecasting model will result in multiple predictions – more accurate as the gap between 18th week and the observation reduces. Similarly to the dynamics of the R factor, accuracy of the predictions differs depending on the gap between the observation time and the forecast time. As a result, to prioritize observations closer to the forecast time, predictions are assigned with the weight corresponding to the inverse of R factor ($1/R$). A weighted average is then taken to obtain the final result.

5 Classification

Kaggle User Name: `davydtamrazov`
Public Leaderboard Score: `0.58910`

For the classification task, original dataset prior to preprocessing is used, accounting for same day observations and all of the features. Each observation in studies A–D has its `LeadStatus` feature corresponding to the audit assessment replaced by a binary variable: 1 if the status is ”Passed” and 0 for the rest. Using all of the observations avoids losing valid points and introducing bias towards one of the audit results.

For the purpose of this analysis, it is assumed that the assessor will determine if the observation is erroneous based primarily on two factors:

- Relative change in each of the individual scores from the previous observation, if such exists;
- Variation of the scores within each of the Positive, Negative, and General Phsychopathology groups.

Thus, training dataset that includes studies A through D is used to defined the above-mentioned features. First, the inter-group covariance is calculated, resulting in three new features: `CovP`, `CovN`, and `CovG`. Next, for every observation i corresponding to a specific patient, each of the 30 individual PANSS scores (denoted as S in the equation below) are updated to represent the relative change in value:

$$S_i = \frac{S_i - S_{i-1}}{S_{i-1}}$$

Only the interaction terms between the S_i values and the covariances are included in the model. This is done in order to magnify the S_i values in case when covariance is large and, thus, probability of being erroneous is larger. It is based on the assumption that by performing this magnification a more coherent boundary can be established. Similarly to the forecasting model, `TxGroup` feature is included to differentiate between models for different study groups. For the observation on day 0, where no prior information is available, S_i values are assumed as 0. This, however, means that all patients have the same probability of having an erroneous assessment on the first day.

A separate linear discriminant model (LDA) is then fit to each of the score groups, as shown with the formulas below, where `data.diff` represents the training dataset with the update features.

```
lda(LeadStatus ~ TxGroup:(CovP:(P1+P2+P3+P4+P5+P6+P7)), data = data.diff)
lda(LeadStatus ~ TxGroup:(CovN:(N1+N2+N3+N4+N5+N6+N7)), data = data.diff)
lda(LeadStatus ~ TxGroup:(CovG:(G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+
                            G14+G15+G16)), data = data.diff)
```

A final probability is calculated as the average of the probabilities predicted by three models. Splitting the problem into groups as shown above allows to achieve both dimensionality reduction as well as to identify different trends pertinent to each group. As a result, the final probability can be expected to be a more accurate one.