Project STAT 5620

Konstantin Zuev, Mohammed Usama Jasnak, Siqi Wang

2024-10-08

Abstract

Nowadays, Mental exhaustion and depression among medical students stand as a formidable challenge. This project explores the intricate relationships between empathy, mental health, and burnout among medical students, utilizing a cross-sectional dataset comprising of 887 medical students from the University of Lausanne from a study published in a medical journal in 2022 by Carrard et al. The project focuses on how variables such as age, gender, academic year, mental health, empathy, and burnout relate to depression (CES-D) and emotional exhaustion (MBI-Ex). The analysis involves using generalized linear models (GLMs) and generalized additive models (GAMs) to capture the relation among the explanatory and response variables. Key findings suggest a notable decrease in emotional exhaustion and depression symptoms with few variables. The study provides valuable insights into the mental health challenges faced by medical students.

Keywords: Medical Students, Mental Health, Burnout, MBI Exhaustion, CES-D (Depression), Generalized Linear Models (GLMs), Generalized Additive Models (GAMs)

1. Introduction

Medical students often are under high levels of stress and pressure due their demanding curriculum and competitive environment. This leads to depression, anxiety, and burnout which not only affects students' academic performance and professional development but also their overall well-being. Due to this, our project is focused on understanding the factors influencing depression and burnout among medical students at the University of Lausanne. We have used a cross-sectional design using a data set of 887 students from the University of Lausanne.

Since we have concerns about mental well-being of the students, we will focus on and will address the following research questions:

- RQ1: What factors contributed to depression symptoms in students (`cesd`)?
- RQ2: Which students were more likely to feel emotionally overextended and exhausted (`mbi_ex`)?

Together, with these research questions we will explore and find out how empathy, depression, social skills, academic factors and demographics intersect and extract meaningful insights which can be utilized to better support students.

2. Project description

2.1 Dataset

- **Name**: The relationship between medical students' empathy, mental health, and burnout

- **Source**: Carrard, V., Bourquin, C., Berney, S., Schlegel, K., Gaume, J., Bart, P.-A., Preisig, M., Schmid Mast, M., & Berney, A. (2022). Dataset for the paper "The relationship between medical students' empathy, mental health, and burnout: A cross-sectional study" published in Medical Teacher (2022) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5702895
- **Sampling**:
  - **Population**: medical students from curriculum years 1 – 6 (Faculty of Biology and Medicine at University of Lausanne & Lausanne University Hospital)
  - **Sampling method**: random sampling
  - **Sample size**: 887
- **Study design characteristics**:
  - **Study type**: cross-sectional
  - **Data collection method**: self-report questionnaires and an emotion recognition test
- **Data description**: this study included medical students from curriculum years 1 – 6. Dataset dimensions {rows, columns}: {886, 20}

Variable Name	Variable Label	Variable Scale
id	Participants ID number	nominal
age	Age	quantitative ratio
year	CURRICULUM YEAR	ordinal: 1=Bmed1; 2=Bmed2; 3=Bmed3; 4=Mmed1; 5=Mmed2; 6=Mmed3
sex	GENDER	nominal: 1=Man; 2=Woman; 3=Non-binary
glang	MOTHER TONGUE	nominal: 1=French; 15=German; 20=English; 37=Arab; ... ; 121=Other
part	PARTNERSHIP STATUS	nominal: 0=No; 1=Yes
job	HAVING A JOB	nominal: 0=No; 1=Yes
stud_h	HOURS OF STUDY PER WEEK	quantitative ratio
health	SATISFACTION WITH HEALTH	ordinal: 1=Verydissatisfied; 2=Dissatisfied; 3=Neithersatisfiednordissatisfied; 4=Satisfied; 5=Verysatisfied
psyt	PSYCHOTHERAPY LAST YEAR	nominal: 0=No; 1=Yes
jspe	JSPE total empathy score	quantitative ratio
qcae_cog	QCAE Cognitive empathy score	quantitative ratio
qcae_aff	QCAE Affective empathy score	quantitative ratio
amsp	AMSP total score	quantitative ratio
erec_mean	GERT mean value of correct responses	quantitative ratio
cesd	CES-D total score	quantitative ratio
stai_t	STAI score	quantitative ratio
mbi_ex	MBI Emotional Exhaustion	quantitative ratio
mbi_cy	MBI Cynicism	quantitative ratio
mbi_ea	MBI Academic Efficacy	quantitative ratio

2.2 Foundational principles behind some key variables

Given that half of these variables consist of scores or indices derived from various tests, so we will provide a brief explanation about them.

JSPE total empathy score: the Jefferson Scale of Physician Empathy (JSPE) is a widely used instrument designed to measure empathy in healthcare professionals and students in the medical field. It was developed by researchers at the Center for Research in Medical Education and Health Care at the Sidney Kimmel Medical College of Thomas Jefferson University in Philadelphia. The scores range from 20 to 140 where high score indicates higher level of empathy [2].

QCAE Cognitive and Affective empathy score: the QCAE (Questionnaire of Cognitive and Affective Empathy) is a tool used to measure both cognitive and affective empathy.

The Cognitive Empathy score on the QCAE measures an individual's ability to understand and perceive others' emotions and thoughts wheras the Affective Empathy score on the QCAE measures an individual's ability to share and respond to others' emotions. Both Cognitive and Affective empathy score typically range between 0-80. There is no fixed threshold but higher the score mean higher empathy [3].

AMSP total score: Ability to Modify Self-Presentation. This metric measures the ability to alter your behaviour in social situation [4].

GERT (Geneva Emotion Recognition Test): GERT is a test which is used to assess an individual's ability to recognize emotions based on facial expressions. The score is calculated based on the number of correctly identified emotions from a set of facial expressions. The range of values can vary depending on the specific version of the test and how it's scored [5].

CES-D (Center for Epidemiological Studies Depression): the Center for Epidemiological Studies-Depression (CES-D), originally published by Radloff in 1977, is a 20-item measure that asks caregivers to rate how often over the past week they experienced symptoms associated with depression, such as restless sleep, poor appetite, and feeling lonely. Response options range from 0 to 3 for each item (0 = Rarely or None of the Time, 1 = Some or Little of the Time, 2 = Moderately or Much of the time, 3 = Most or Almost All the Time). Scores range from 0 to 60, with high scores indicating greater depressive symptoms. The CES-D also provides cutoff scores (e.g., 16 or greater) that aid in identifying individuals at risk for clinical depression, with good sensitivity and specificity and high internal consistency [6].

STAI (State-Trait Anxiety Inventory): the State-Trait Anxiety Inventory (STAI) is a widely used psychological test designed to measure both state anxiety and trait anxiety, which are two different components of anxiety. It is to diagnose anxiety and to distinguish it from depressive syndromes. (I could not find any definite cutoff or value scale online or using GPT so I didn't add it) [7].

Burnout (MBI Cynicism, MBI Emotional Exhaustion, MBI Academic Efficacy):

The Maslach Burnout Inventory (MBI) is a psychological assessment instrument pertaining to occupational burnout [8]. The metrics included in the paper are:

- **MBI Cynism:**

  This scale measures the development of negative, cynical attitudes and feelings towards studies. It assesses the tendency to distance oneself emotionally from academic-related activities and interactions. The typical range of values of MBI-Cynicism are:

  - Low: scores typically range from 0 to 6.
  - Moderate: scores typically range from 7 to 12.
  - High: scores typically range from 13 and above.

- **MBI Exhaustion:**

  This scale assesses feelings of being emotionally drained, depleted, and overwhelmed by one's work. The typical range of values of MBI Exhaustion are:

  - Low: scores typically range from 0 to 16.
  - Moderate: scores typically range from 17 to 20
  - High: scores typically range from 20 and above.

- **MBI Academic Efficacy:**

  This scale measures feelings of competence and successful achievement in one's work. It is akin to the Personal Accomplishment scale. Lower scores correspond to greater experienced burnout.

  - Low: scores typically range from 15 and above.
  - Moderate: scores typically range from 15 to 25
  - High: scores typically range from 25 and above.

3. EDA

3.1 Univariate Analysis

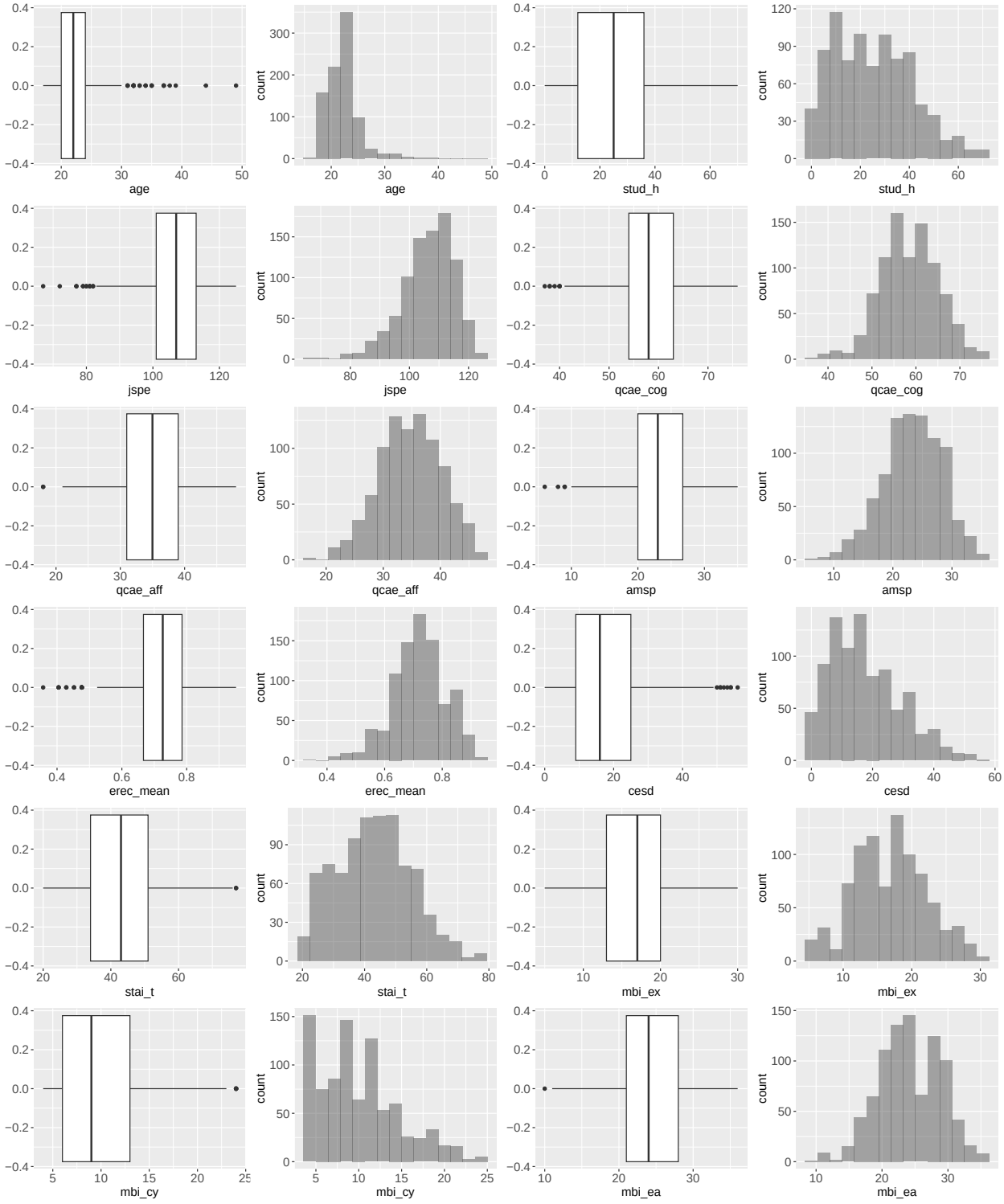
We start our EDA by producing a concise data quality report for both categorical and numeric variables. It is necessary for identifying the most obvious data issues that may need to be fixed right away, i.e. missing values, datatype issues etc.

```
##                V1            V2      V3      V4      V5      V6      V7
## Statistic Missing Unique Values    Mean  Median     Std Minimum Maximum
## id           0.00        886.00  889.71  876.00  515.56    2.00 1790.00
## age          0.00         24.00   22.38   22.00    3.30   17.00   49.00
## year         0.00          6.00    3.10    3.00    1.76    1.00    6.00
## sex          0.00          3.00    1.70    2.00    0.47    1.00    3.00
## glang        0.00         19.00   14.33    1.00   32.37    1.00  121.00
## part         0.00          2.00    0.56    1.00    0.50    0.00    1.00
## job          0.00          2.00    0.35    0.00    0.48    0.00    1.00
## stud_h       0.00         61.00   25.29   25.00   15.93    0.00   70.00
## health       0.00          5.00    3.78    4.00    1.06    1.00    5.00
## psyt         0.00          2.00    0.22    0.00    0.42    0.00    1.00
## jspe         0.00         49.00  106.37  107.00    8.78   67.00  125.00
## qcae_cog     0.00         40.00   58.53   58.00    6.57   37.00   76.00
## qcae_aff     0.00         29.00   34.78   35.00    5.38   18.00   48.00
## amsp         0.00         29.00   23.15   23.00    4.99    6.00   35.00
## erec_mean    0.00         24.00    0.72    0.73    0.09    0.36    0.95
## cesd         0.00         55.00   18.05   16.00   11.48    0.00   56.00
## stai_t       0.00         57.00   42.90   43.00   11.98   20.00   77.00
## mbi_ex       0.00         26.00   16.88   17.00    5.26    5.00   30.00
## mbi_cy       0.00         21.00   10.08    9.00    4.59    4.00   24.00
## mbi_ea       0.00         27.00   24.21   24.00    4.63   10.00   36.00
```

Based on this report, we can conclude the following:

- There are no missing values;
- In general, numeric variables are well-behaved, i.e. no variables that have extreme values, no negative values for strictly positive features.
- Variable `glnag` has a lot of unique categories that may require additional attention during the modelling stage.

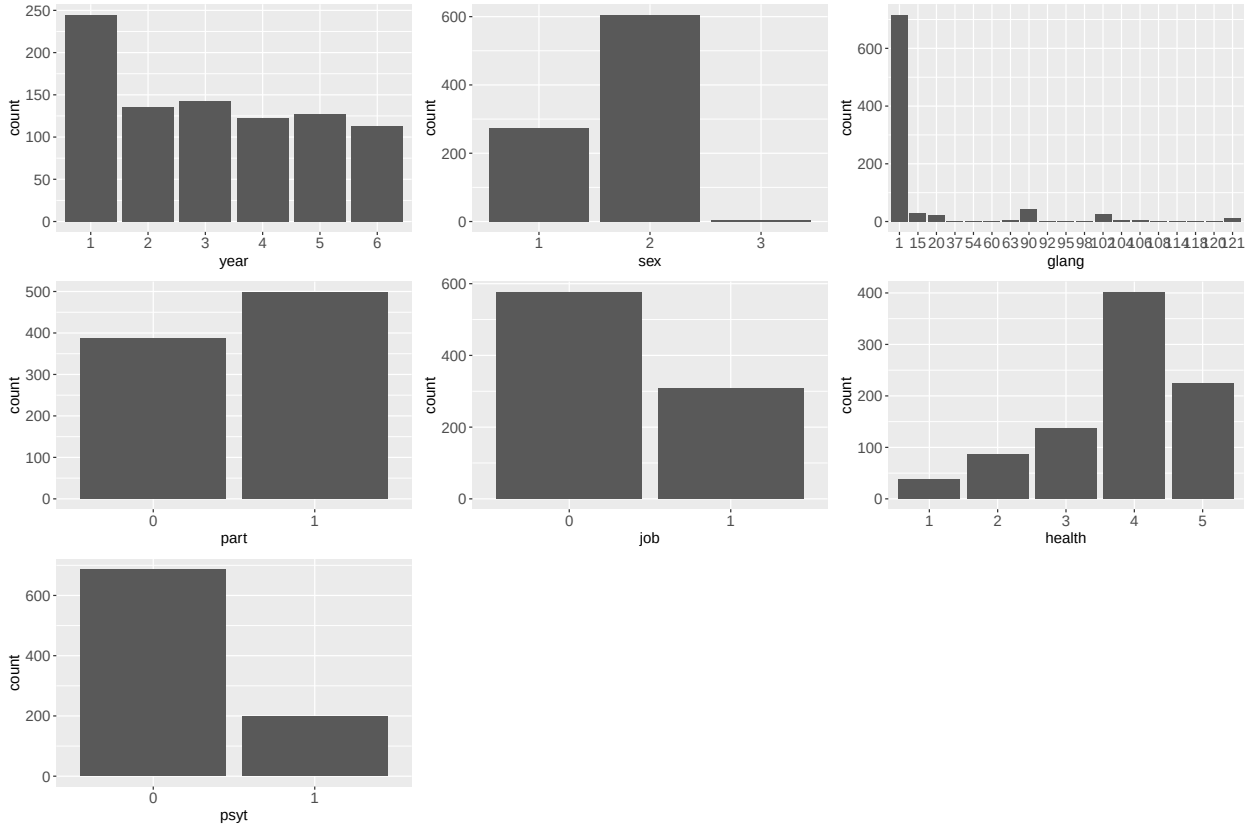
Before analysing how the target variables of interest (`mbi_ex` and `cesd`) interact with independent variables, it is important to visualise each variable on its own:



Unconditional distributions of our variables might not be particularly important. However, useful information can still be extracted: we have strictly positive, integer-valued (semi-continuous) variables, some are relatively normal, others are skewed. In this setting, when building statistical models, we may consider Poisson distribution due to the fact that variables are not truly continuous, Gamma distribution because of the

skewness as well as non-negative nature of our data, and even Gaussian. The fact that we do not really have significant outliers is also evident.

Of course, both the distribution of $Y|X$ and the impact of potential outliers on our model will have to be determined more rigorously during the model fitting stage.



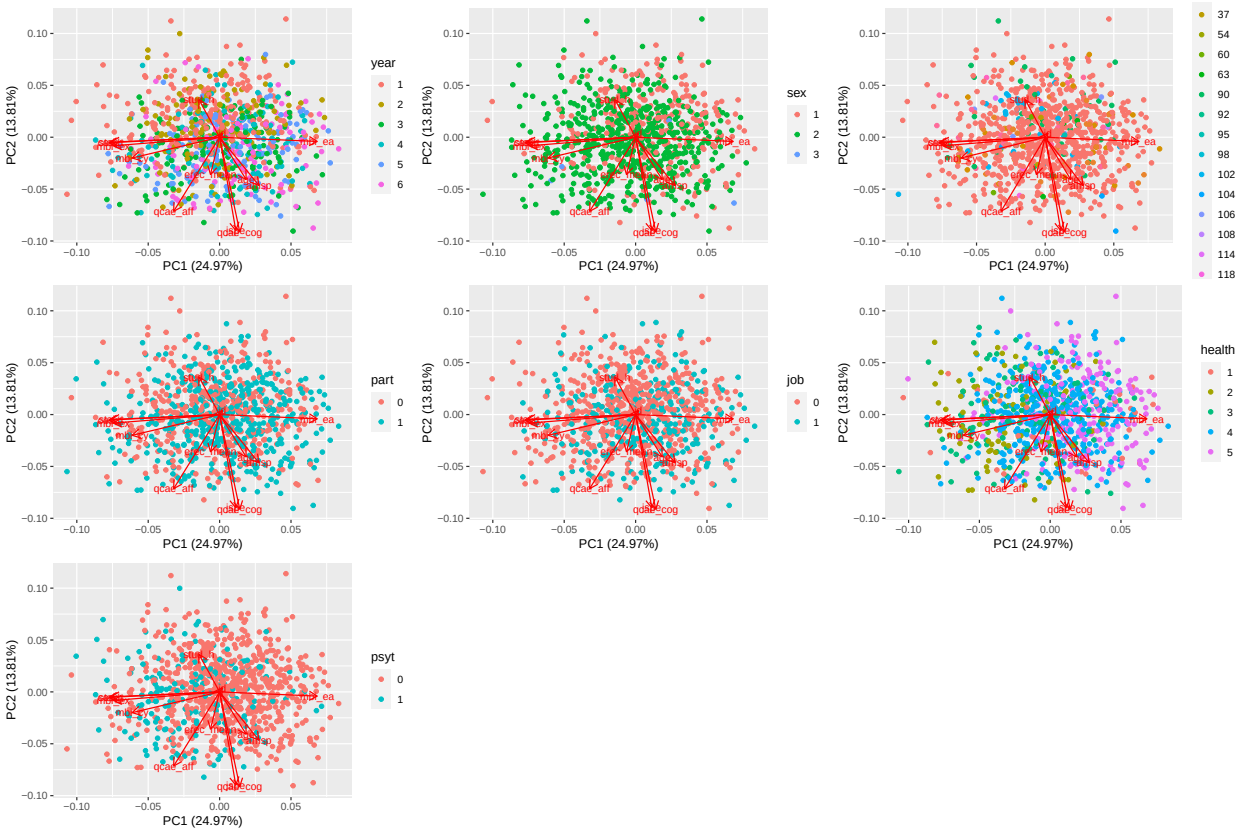
Based on the plots above, we can also confirm that some categorical variables are unbalanced, which may impact the analysis.

3.2 Principal Component Analysis

By using PCA we were able to map a high-dimensional space into fewer components (linear combinations).

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8018 1.3397 1.1902 1.05026 1.00064 0.98879 0.85771
## Proportion of Variance 0.2497 0.1381 0.1090 0.08485 0.07702 0.07521 0.05659
## Cumulative Proportion  0.2497 0.3878 0.4968 0.58160 0.65862 0.73382 0.79041
##                           PC8     PC9   PC10    PC11    PC12    PC13
## Standard deviation     0.82508 0.72253 0.6954 0.64391 0.60555 0.50683
## Proportion of Variance 0.05237 0.04016 0.0372 0.03189 0.02821 0.01976
## Cumulative Proportion  0.84278 0.88294 0.9201 0.95203 0.98024 1.00000
```

We can see that each principal component explains a relatively low percentage of overall variance, suggesting that variables are not strongly correlated and that their ability to explain variance in data is limited.



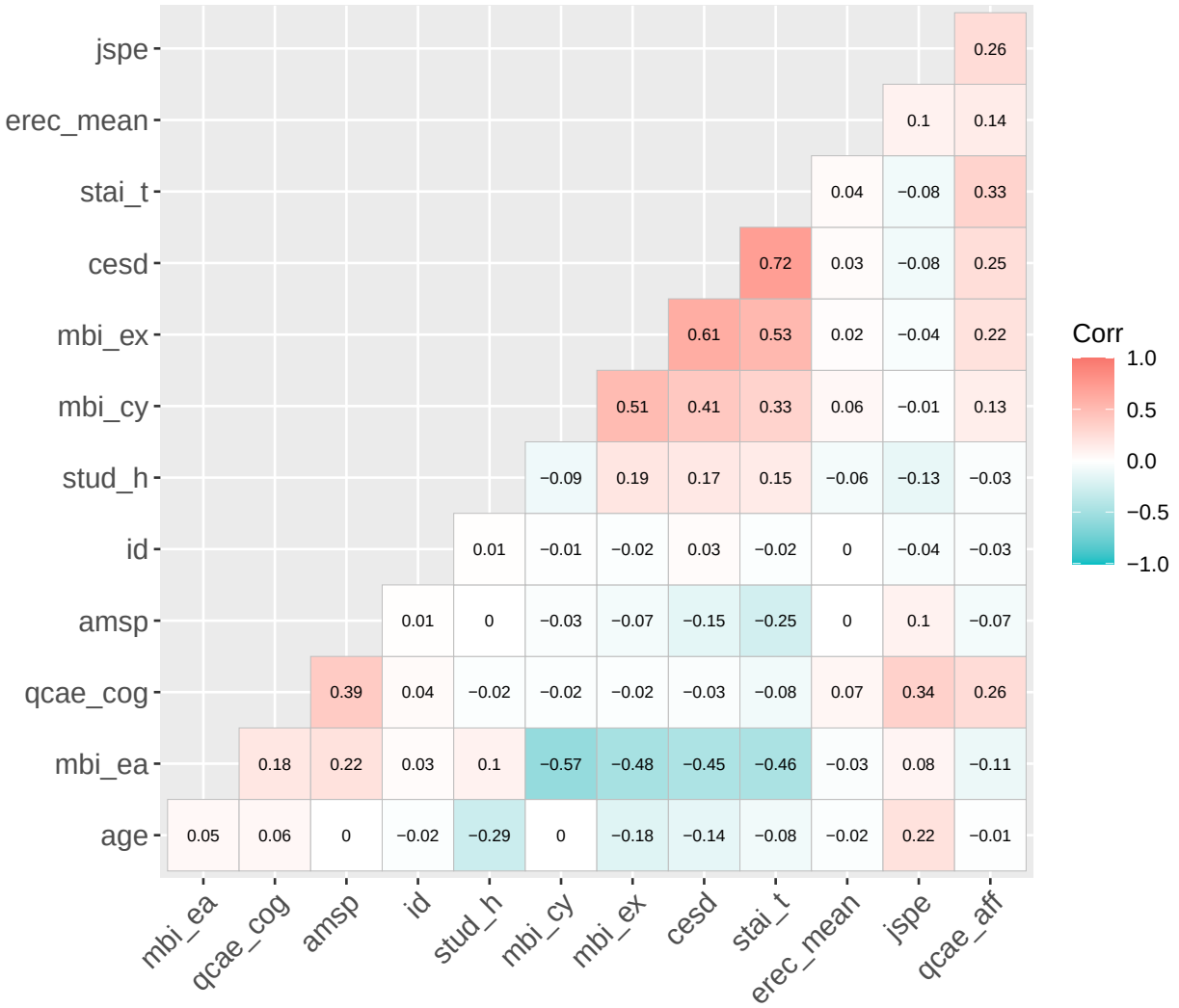

```
##                 PC1         PC2
## id          0.01158860  0.02344081
## age         0.11040488 -0.24184682
## stud_h     -0.08573801  0.21440695
## jspe        0.06725305 -0.53249218
## qcae_cog    0.08363952 -0.54067176
## qcae_aff   -0.19075125 -0.42325988
## amsp        0.16344900 -0.27549775
## erec_mean -0.03793587 -0.21264079
## cesd       -0.45980064 -0.03165191
## stai_t     -0.44969651 -0.02811249
## mbi_ex     -0.44079458 -0.04902991
## mbi_cy     -0.36322399 -0.11861551
## mbi_ea      0.40545169 -0.02422116
```

Producing biplots for the first two components and colouring them based on categorical features also doesn't show any clear patters. One exception is the plot 6 where the categorical variable of interest is `health`.

Since the data was standardised, we can interpret loadings as correlations. PC1 was primarily defined by `cesd`, `stai_t`, `mbi_ex`, `mbi_cy`. All variables represented burnout, anxiety and depression and were negatively correlated, i.e. larger values of PC1, smaller values of the aforementioned variables.

Plot 6 shows a more or less expected pattern: the higher the satisfaction with one's health is, the larger the value is for PC1 (less depression and anxiety).

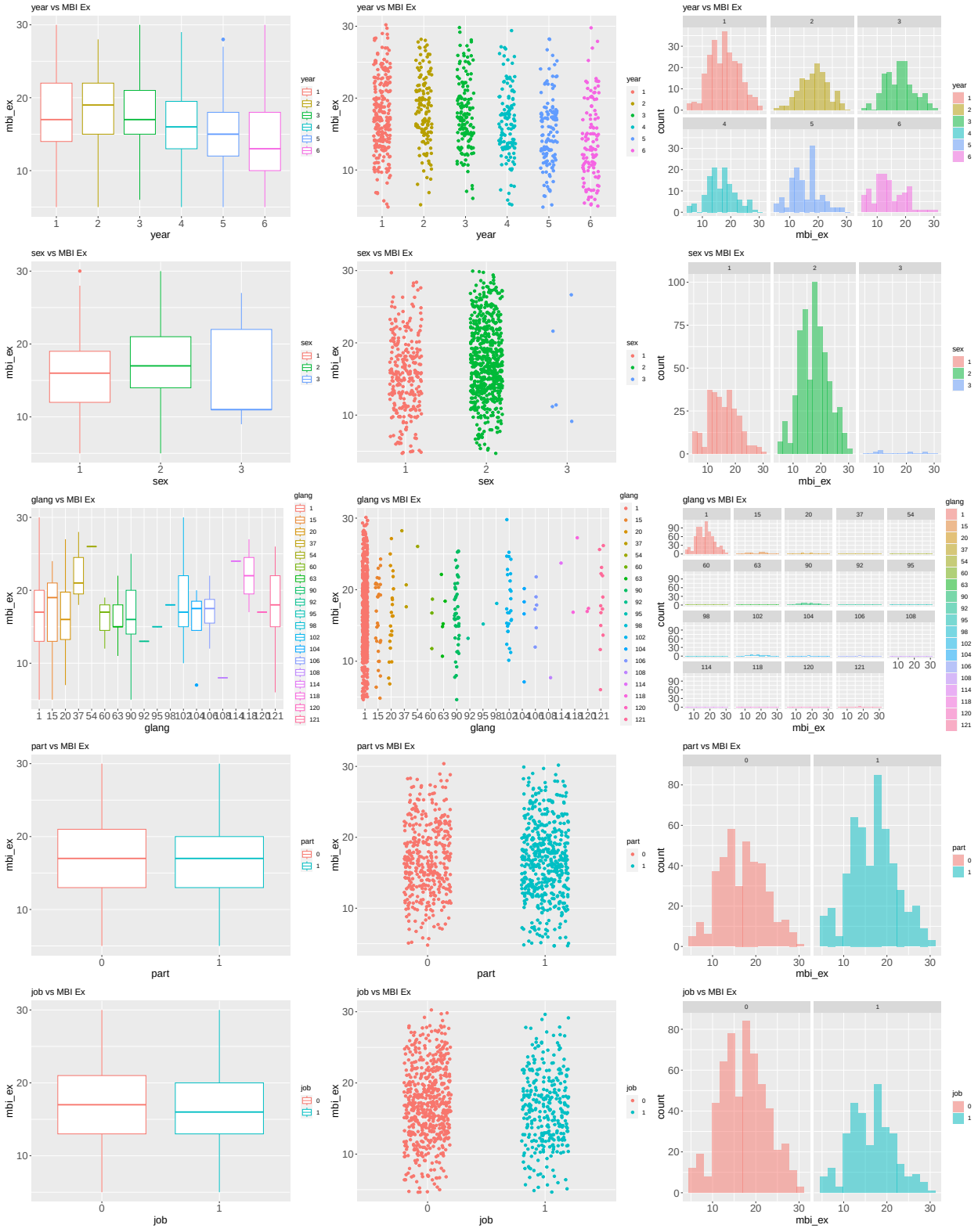
Before focusing on the target variables and their interactions with other features, a correlation matrix was created to explore linear relationships between variables.

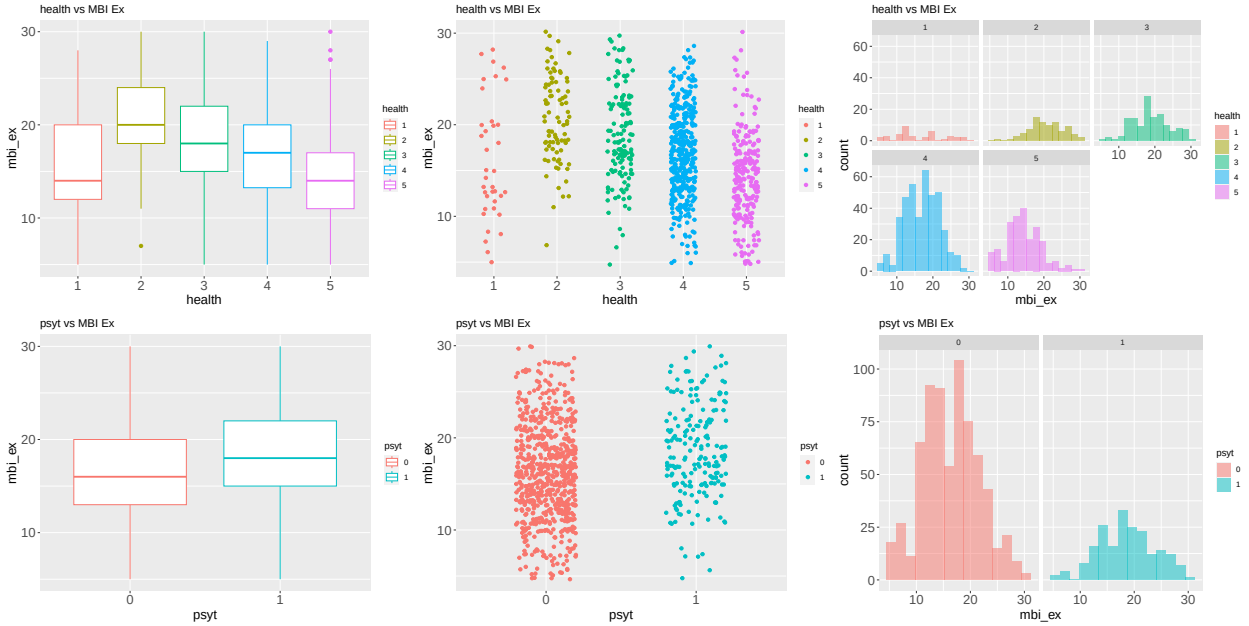


Overall, it is evident that most variables do not exhibit strong linear relationships between each other. The maximum absolute value of the correlation coefficient is 0.72, between the target variable `cesd` and `stai_t`.

3.3 Multivariate Analysis. MBI Exhaustion

To capture the relationship between categorical variables and the target variable `mbi_ex`, a combination of boxplot, jitterplot and histogram was used. A histogram allows for assessing the distribution of a variable for a given category, a jitterplot helps in estimating the number of observations for each group and a boxplot provides more information with respect to potential outliers.

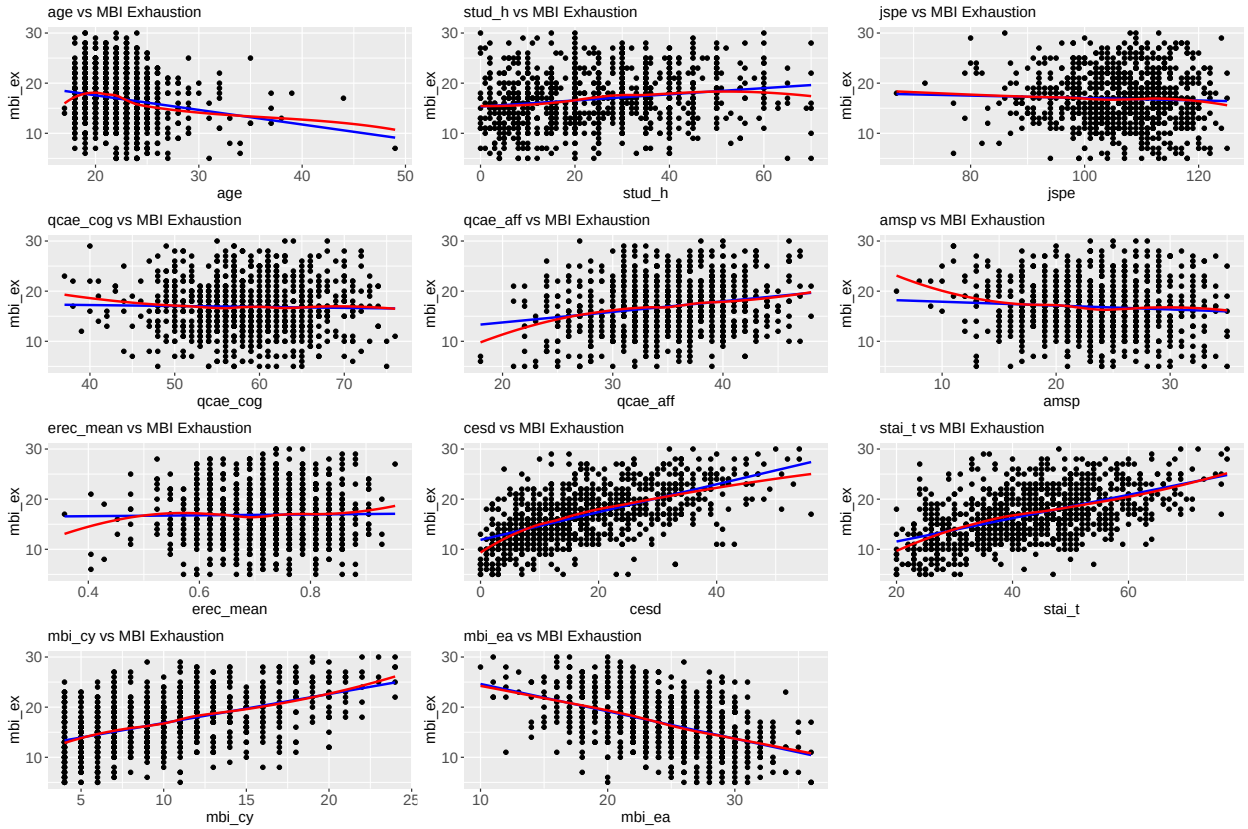

year vs MBI Ex
mbi_ex
year
year vs MBI Ex
mbi_ex
year
year vs MBI Ex
count
mbi_ex
sex vs MBI Ex
mbi_ex
sex
sex vs MBI Ex
mbi_ex
sex
sex vs MBI Ex
count
mbi_ex
glang vs MBI Ex
mbi_ex
glang
glang vs MBI Ex
mbi_ex
glang
glang vs MBI Ex
count
mbi_ex
part vs MBI Ex
mbi_ex
part
part vs MBI Ex
mbi_ex
part
part vs MBI Ex
count
mbi_ex
job vs MBI Ex
mbi_ex
job
job vs MBI Ex
mbi_ex
job
job vs MBI Ex
count
mbi_ex



Based on produced figures, the following can be concluded:

- Variable `year`(curriculum year) shows a decreasing trend in emotional exhaustion as students progress through their studies. Although the effect doesn't seem to be perfectly linear.
- Emotional exhaustion levels in women seem to be higher than in men. However, whether the difference is statistically significant or not will have to be investigated.
- The partnership status (`part`) and having a part-time job (`job`) don't seem to have any noticeable effect on emotional exhaustion.
- Variable `health` (satisfaction with health) has a similar patter as `year`. The more students are satisfied with their health, the lower the emotional exhaustion is.
- Lastly, students who had psychotherapy sessions `psyt` reported higher levels of emotional exhaustion.

Scatterplots were drawn to analyse the relationship between continuous variables and `mbi_ex`.



From the plots presented, we can conclude that:

- There are no strong non-linear patterns.
- In most cases, slopes are relatively gradual. Sometimes even visually flat: `erec_mean` or `qcae_cog`.
- Variable `age` also seems to be significantly affected by a single observation (a student who was 49 years old).

```
##    Column_name Pearson_correlation Spearman_correlation
## 11      mbi_ea              -0.481               -0.471
## 1          age              -0.183               -0.177
## 6         amsp              -0.073               -0.074
## 3         jspe              -0.041               -0.032
## 4     qcae_cog              -0.024               -0.022
## 7    erec_mean               0.015                0.000
## 2       stud_h               0.186                0.199
## 5     qcae_aff               0.216                0.187
## 10      mbi_cy               0.505                0.478
## 9       stai_t               0.530                0.514
## 8         cesd               0.606                0.613
```

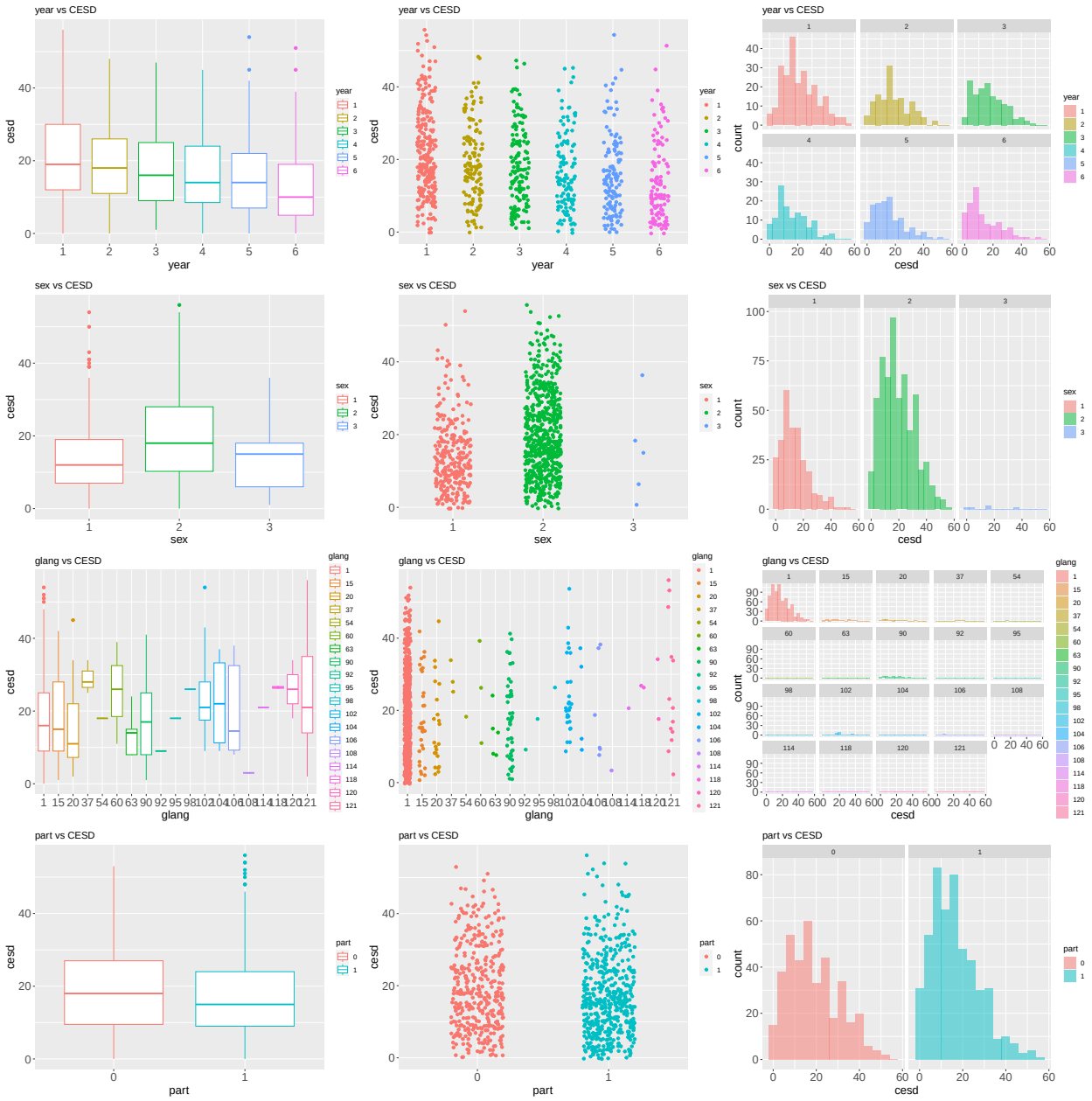
Analysing correlation coefficients, we can find support for the conclusions made based on scatterplots. Spearman correlation coefficients were calculated to account for potential non-linear relationships, but they barely diverged from Pearson coefficients.

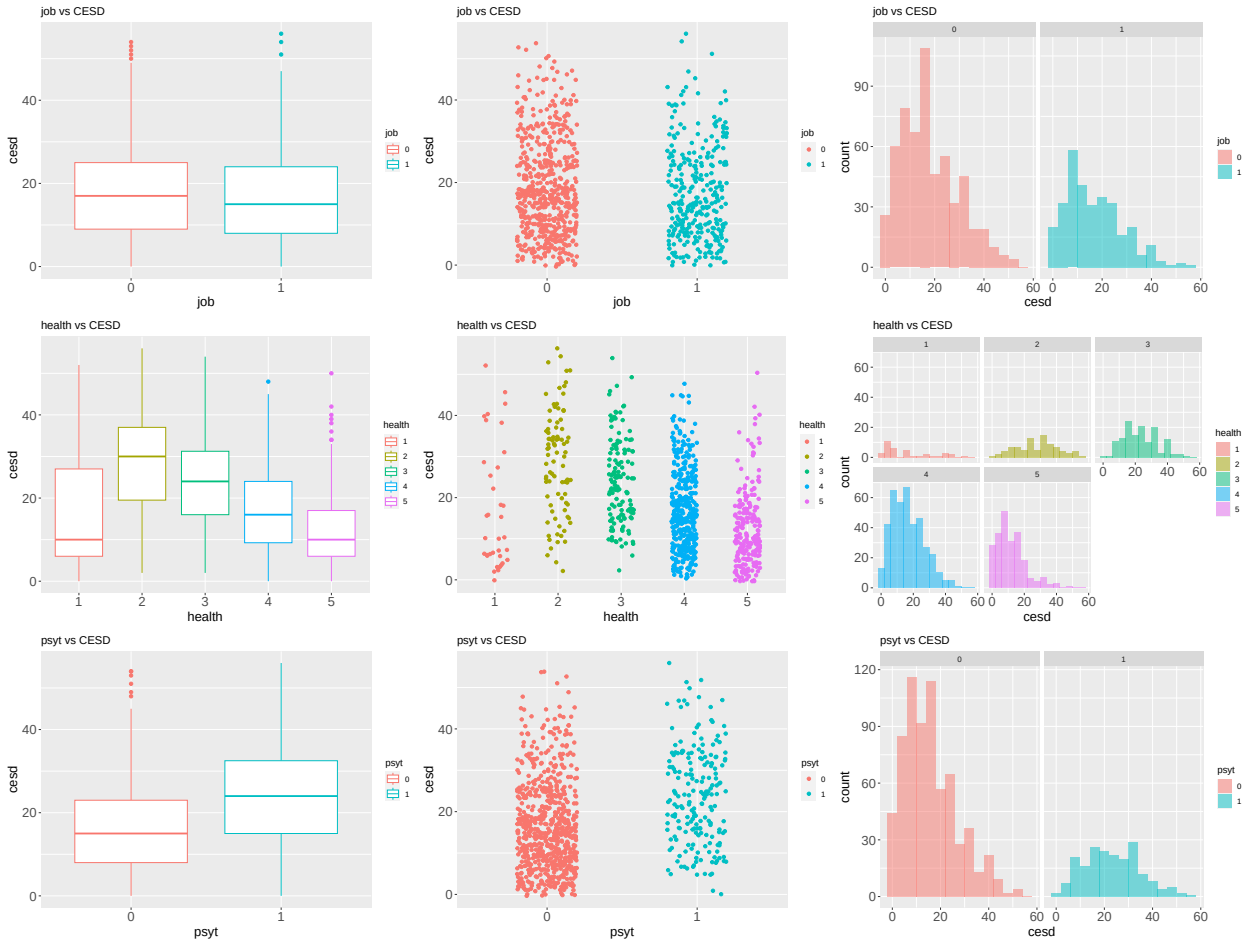
Lastly, we fitted analysis of variance models and conducted Kruskal-Wallis rank sum tests to complement our visual analysis of categorical variables.

```
##    Variable ANOVA_p_value Kruskal_p_value
## 1      year        0.0000          0.0000
## 2       sex        0.0000          0.0000
## 3     glang        0.4968          0.6187
## 4      part        0.7168          0.8677
## 5       job        0.0512          0.0552
## 6    health        0.0000          0.0000
## 7      psyt        0.0000          0.0000
```

As expected, `glan` and `part` were not significant, while `job` was significant at the level of $\alpha = 0.1$.

3.4 Multivariate Analysis. CESD

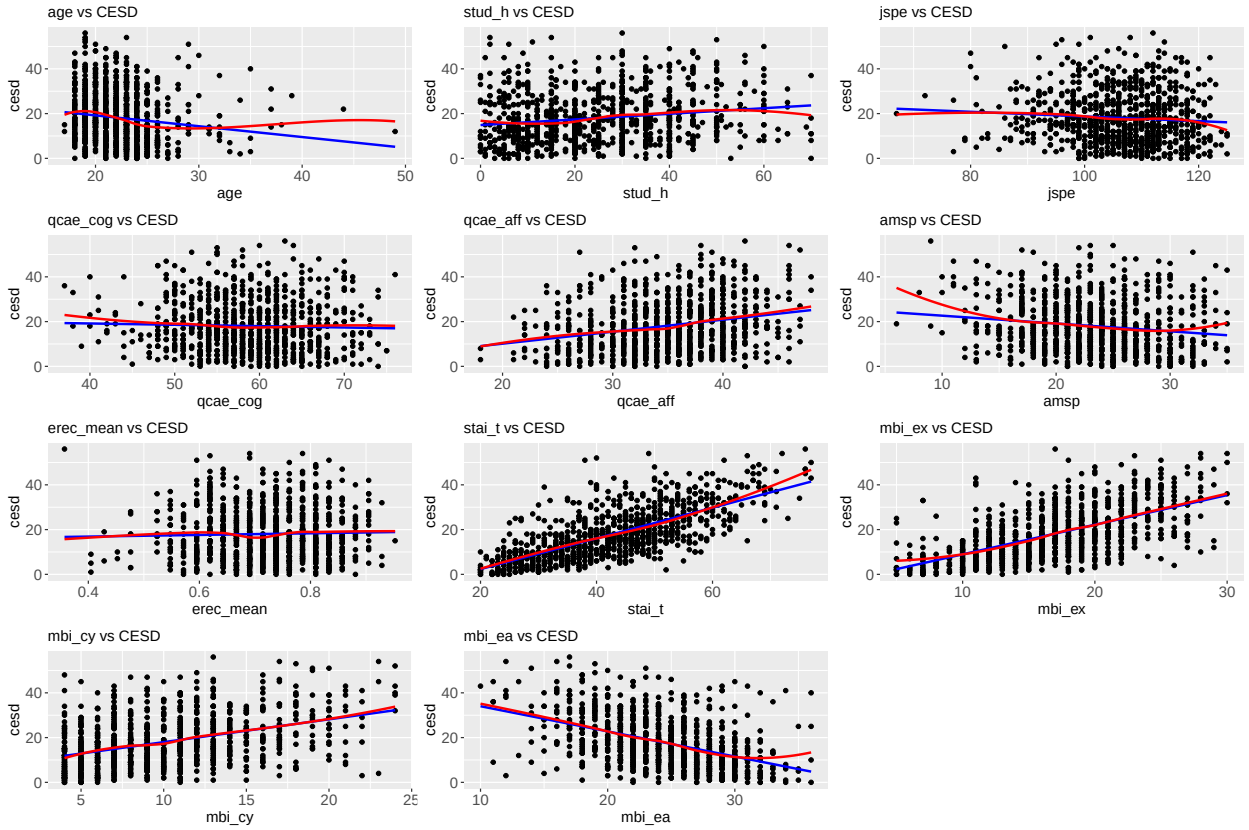




Based on produced figures, the following can be concluded:

- Variable `year`(curriculum year) shows a decreasing trend in CESD (Center for Epidemiologic Studies Depression Scale) as students progress through their studies. The effect seems to be linear.
- CESD in women seem to be higher than in men. However, whether the difference is statistically significant or not will have to be investigated.
- The partnership status `part` and having a part-time job `job` don't seem to have any noticeable effect on CESD (although it could be statistically significant).
- Except for Level 1 (students who were very dissatisfied with their health), CESD decreases linearly as health satisfaction increases.
- Lastly, students who had psychotherapy sessions `psyt` reported higher levels of CESD.

Scatterplots were drawn to analyse the relationship between continuous variables and `cesd`.



From the plots presented, we can conclude that:

- Although most trends appear to be linear, some variables exhibit signs of non-linear patterns.
- In many cases, slopes are relatively gradual. Sometimes even visually flat: `erec_mean` or `qcae_cog`.
- `stai-t`, `mbi_ea`, `mbi_ex` have relatively strong relationships with `cesd`.

```
##    Column_name Pearson_correlation Spearman_correlation
## 1          age              -0.139               -0.201
## 2       stud_h               0.174                0.194
## 3         jspe              -0.080               -0.089
## 4     qcae_cog              -0.034               -0.043
## 5     qcae_aff               0.251                0.248
## 6         amsp              -0.152               -0.159
## 7    erec_mean               0.030                0.037
## 8       stai_t               0.716                0.718
## 9       mbi_ex               0.606                0.613
## 10      mbi_cy               0.408                0.394
## 11      mbi_ea              -0.454               -0.462
```

Analysing correlation coefficients, we can find support for the conclusions made based on scatterplots. Spearman correlation coefficients were calculated to account for potential non-linear monotonous relationships, but they barely diverged from Pearson coefficients.

Lastly, we fitted analysis of variance models and conducted Kruskal-Wallis rank sum tests to complement our visual analysis of categorical variables.

```
##   Variable ANOVA_p_value Kruskal_p_value
```

```
## 1     year        0.0000          0.0000
## 2      sex        0.0000          0.0000
## 3    glang        0.1082          0.0926
## 4     part        0.0015          0.0015
## 5      job        0.0752          0.0676
## 6   health        0.0000          0.0000
## 7     psyt        0.0000          0.0000
```

Compared to `mbi_ex`, p-values for non-significant variables were considerably lower. At the level of $\alpha = 0.05$, both `glang` and `job` were not statistically significant.

4. Methods

We broke down the model fitting process in the following steps:

- Discussing theoretical assumptions for choosing an appropriate distribution (Gaussian / Poisson / etc.).
- Testing in practice how those theoretical assumptions work by fitting different models.
- Removing uninformative variables.
- Dealing with `glan` variable as it has a lot of unbalanced levels: using mixed models, merging categories and treating this new variable as fixed.
- Interpreting final models.

One thing to keep in mind is that every step is interconnected. For example, choosing a distribution can effect what variables are important. But at the same time, dropping some variables may influence how well a model will work under a given distribution. Thus, there is no perfect solution unless you grid search every possible combination which is probably not very feasible. Thus, even though we defined a sequence of steps for building models, we still tried to be flexible in our approach. For example, when selecting a distribution we used AIC-based variable selection as a quick test of model stability, which was done by comparing residuals before and after using `step()` function for a given distribution.

Note, model assumptions and limitations will be discussed separately in the proceeding chapters for each target variable of interest.

5. Analysis and results

5.1 Analysis. MBI Exhaustion

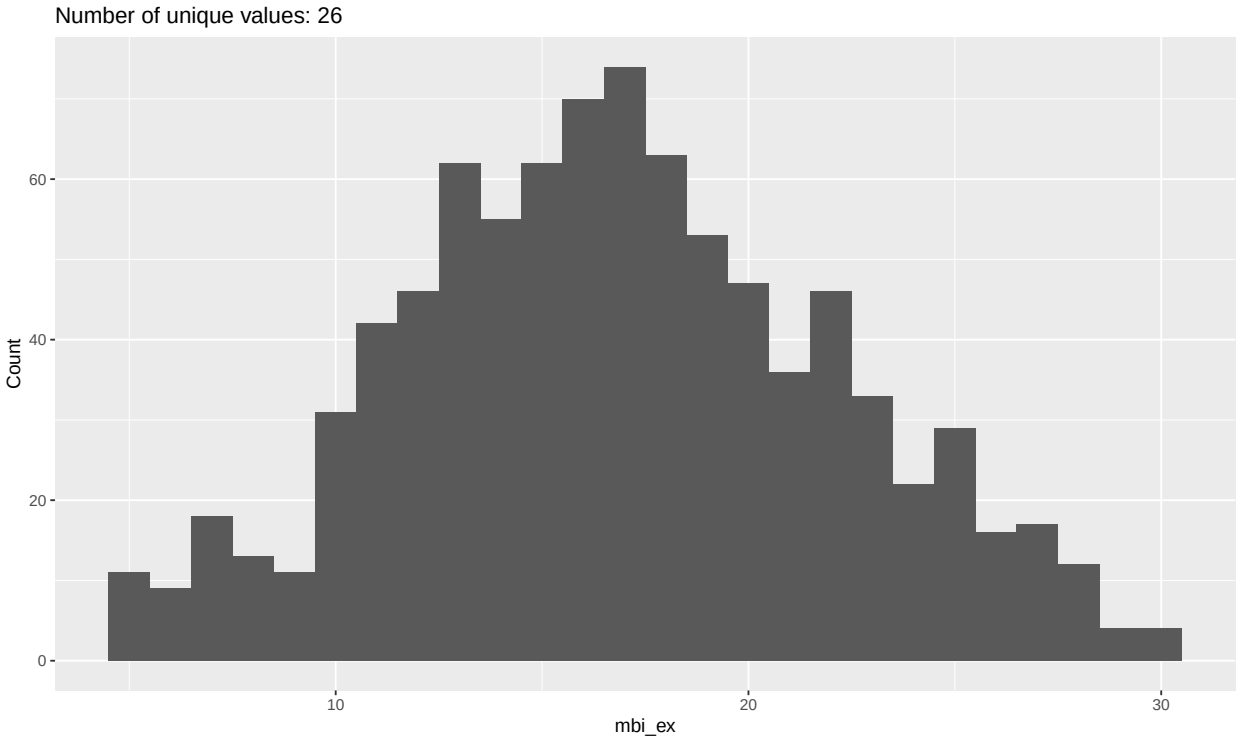
After conducting EDA, it became evident that for `mbi_ex` there weren't any complex non-linear patterns that we could model with GAMs. We also did not have any spatio-temporal data making spatial and spatio-temporal GLMMs unnecessary. In light of that, we decided to consider GLMs and GLMMs to account for potentially having non-normal distribution of the response variable and also efficiently control for `glang` variable with a large number of unbalanced categories.

(a) Distribution considerations. Theoretical assumptions

Variable `mbi_ex`:

- Integer-valued and bounded between 5 and 30.
- 26 unique values.

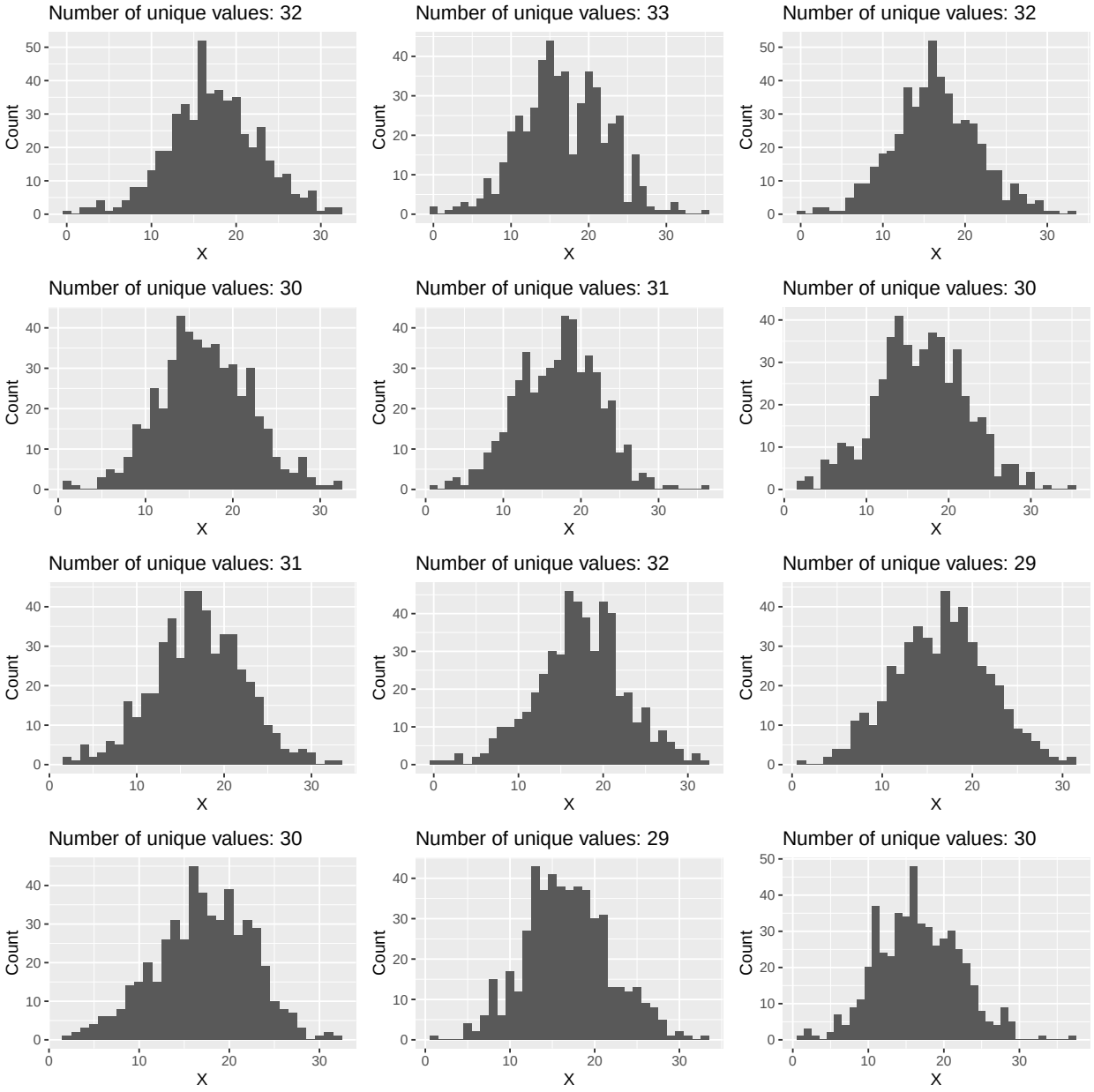
- The mean value is equal to 16.9 and median to 17.
- The standard deviation is 5.3.
- Unconditional distribution of `mbi_ex`:



As a result, we have to consider what distributions are going to be appropriate for modelling this variable.

- Negative binomial distribution is the first that come to mind. Although it is frequently used specifically for count data, it can be a good fit in case of integer-valued variables. Also, since we have no reasons to believe that $\mu = \sigma^2$, Poisson distribution seems less reasonable. However, since this variable does not allow zeros, using either negative binomial or Poisson distribution could be problematic.
- Gaussian distribution may also be appropriate. Normal data is continuous and unbounded. However, if we have a relatively large number of distinct values (26 integers for `mbi_ex`), observations are not concentrated around extreme values, then it is possible to have a distribution that will be approximated by a normal one to a sufficient degree. Remember, all models are wrong but some are useful.

Let's create 12 normally distributed samples with the mean and standard deviation of `mbi_ex`, convert values to integers and plot histograms:



It can be seen that in this experiment using a normal distribution for approximating those samples would work well enough. Of course, under the hood samples are in fact normal, but the key point is that even after converting them to integers (if some reasonable assumptions hold), using a normal distribution would work.

Of course, it doesn't definitely prove anything, and in fact for model building we are interested in $Y|X{\sim}N(\mu, \sigma^2)$ and not Y, but the core idea is that we should not ignore Gaussian distribution right away even if data characteristics do not seem to be suitable.

- Lastly, the most important argument for or against a certain distribution is model results. Thus, trying out different options and analysing residuals to assess how well a given model fits the data will be the ultimate reason why a specific distribution is chosen.

(b) Distribution considerations. Testing assumptions

For assessing how well models fits the data, we primarily relied on residual analysis:

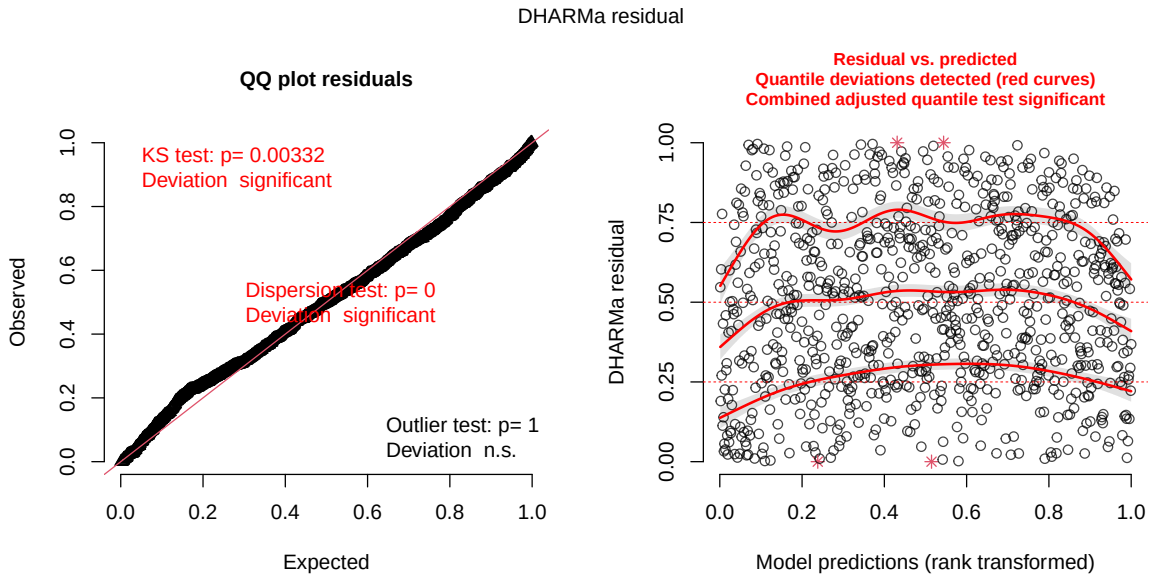
- For a linear regression with a constant variance and the identity link we have a very natural way of defining raw residuals $\varepsilon_i = y_i - X\beta$ via additive decomposition. For GLMs, in general, we lose the property of constant variance, and the relationship between $X\beta$ and μ becomes non-linear – no additive decomposing is possible. Thus, we cannot simply search for a lack of patterns in raw residuals anymore. That is why we need other methods like deviance residuals to analyse the results.
- A simulated-based approach (parametric bootstrapping) can be another great tool for producing interpretable residuals by taking the value of empirical CDF for each observation. Here CDF is produced from simulated data from the fitted model assuming that the model is correct. To do so, we used DHARMa residuals.

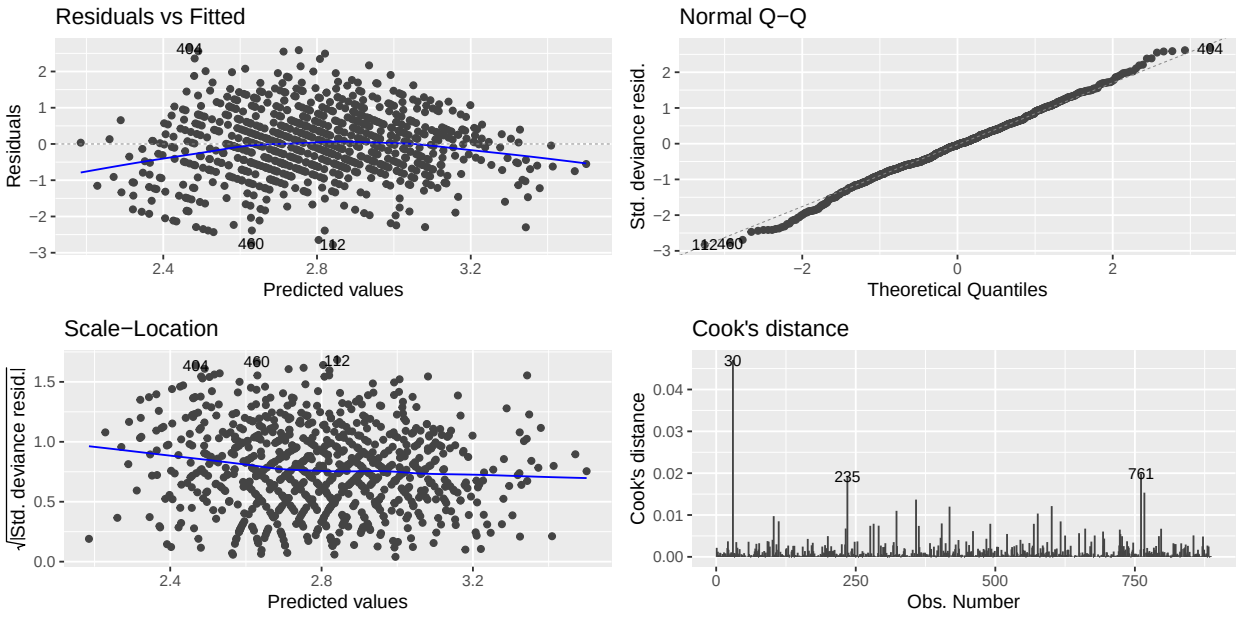
So, first, we fitted two GLMs with negative binomial distribution:

- Model with all variables included except `glang`:

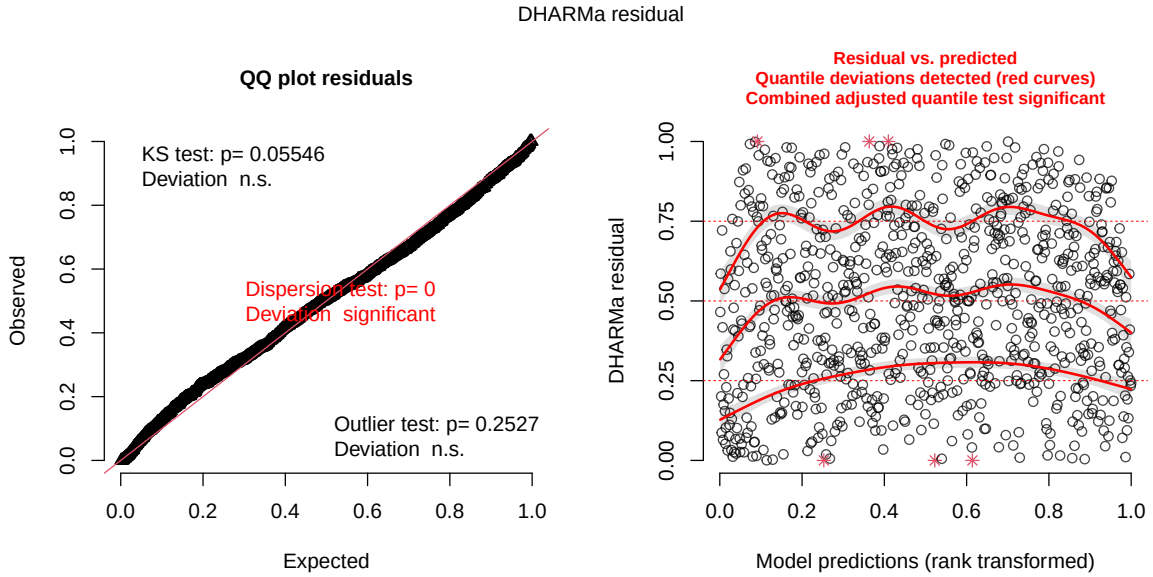
```
##
## Call:
## MASS::glm.nb(formula = mbi_ex ~ age + stud_h + year + sex + part +
##     job + health + psyt + jspe + qcae_cog + qcae_aff + amsp +
##     erec_mean + cesd + stai_t + mbi_cy + mbi_ea, data = data,
##     init.theta = 292864.3228, link = log)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.543e+00  1.656e-01  15.357  < 2e-16 ***
## age         -6.213e-03  3.344e-03  -1.858  0.06318 .
## stud_h       1.902e-03  6.229e-04   3.053  0.00227 **
## year2        1.038e-01  2.657e-02   3.908 9.32e-05 ***
## year3        8.646e-02  2.940e-02   2.941  0.00327 **
## year4        2.469e-02  3.336e-02   0.740  0.45922
## year5       -5.108e-02  3.409e-02  -1.499  0.13398
## year6       -7.646e-02  3.933e-02  -1.944  0.05187 .
## sex2         3.550e-02  2.077e-02   1.709  0.08738 .
## sex3         3.466e-02  1.155e-01   0.300  0.76402
## part1        4.010e-02  1.710e-02   2.344  0.01906 *
## job1        -3.886e-03  1.852e-02  -0.210  0.83380
## health2      6.656e-02  4.851e-02   1.372  0.17006
## health3      4.110e-02  4.624e-02   0.889  0.37411
## health4      4.857e-02  4.352e-02   1.116  0.26435
## health5     -2.326e-02  4.562e-02  -0.510  0.61010
## psyt1       -1.392e-02  2.022e-02  -0.688  0.49115
## jspe        -8.994e-05  1.094e-03  -0.082  0.93448
## qcae_cog    -2.722e-04  1.500e-03  -0.181  0.85600
## qcae_aff     2.393e-03  1.824e-03   1.312  0.18955
## amsp         2.983e-03  1.863e-03   1.601  0.10928
## erec_mean   -8.125e-02  9.140e-02  -0.889  0.37402
## cesd         6.529e-03  1.087e-03   6.007 1.89e-09 ***
## stai_t       3.010e-03  1.053e-03   2.857  0.00428 **
## mbi_cy       1.859e-02  2.203e-03   8.437  < 2e-16 ***
## mbi_ea      -9.992e-03  2.408e-03  -4.149 3.34e-05 ***
```

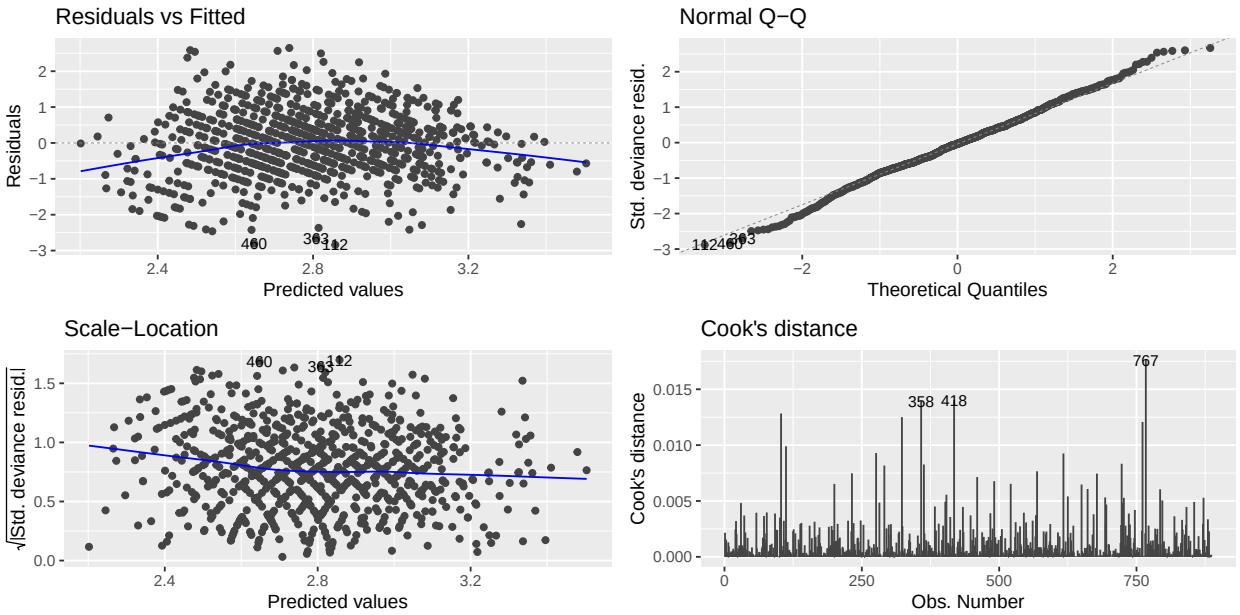
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(292864.3) family taken to be 1)
##
##     Null deviance: 1502.42  on 885  degrees of freedom
## Residual deviance:  720.54  on 860  degrees of freedom
## AIC: 4867.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  292864
##           Std. Err.:  1467400
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -4813.794
```

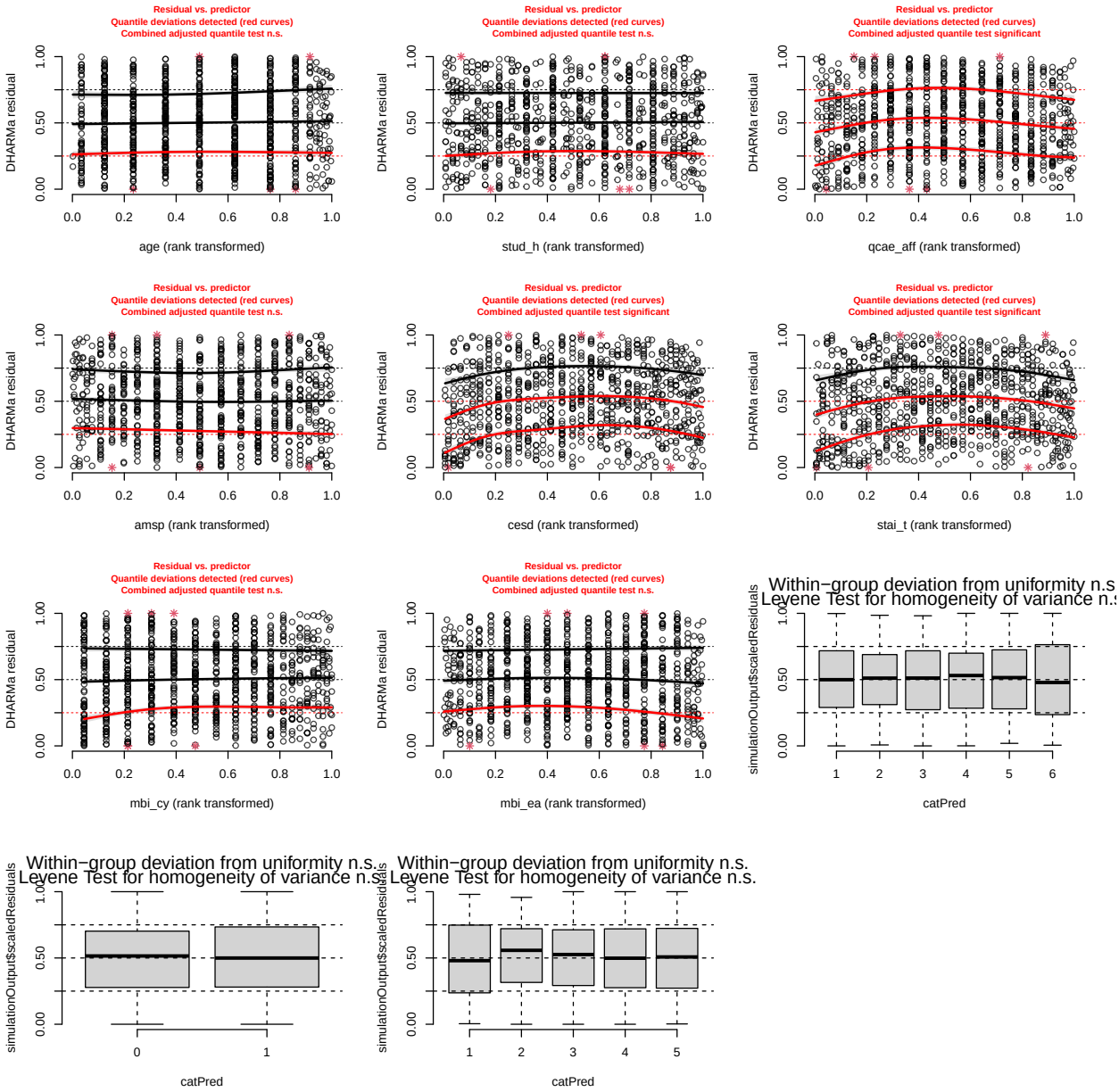




- Model after using `step()` function:




Residuals vs Fitted
Residuals
Predicted values
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
Scale−Location
√|Std. deviance resid.|
Predicted values
Cook's distance
Cook's distance
Obs. Number
767
358
418



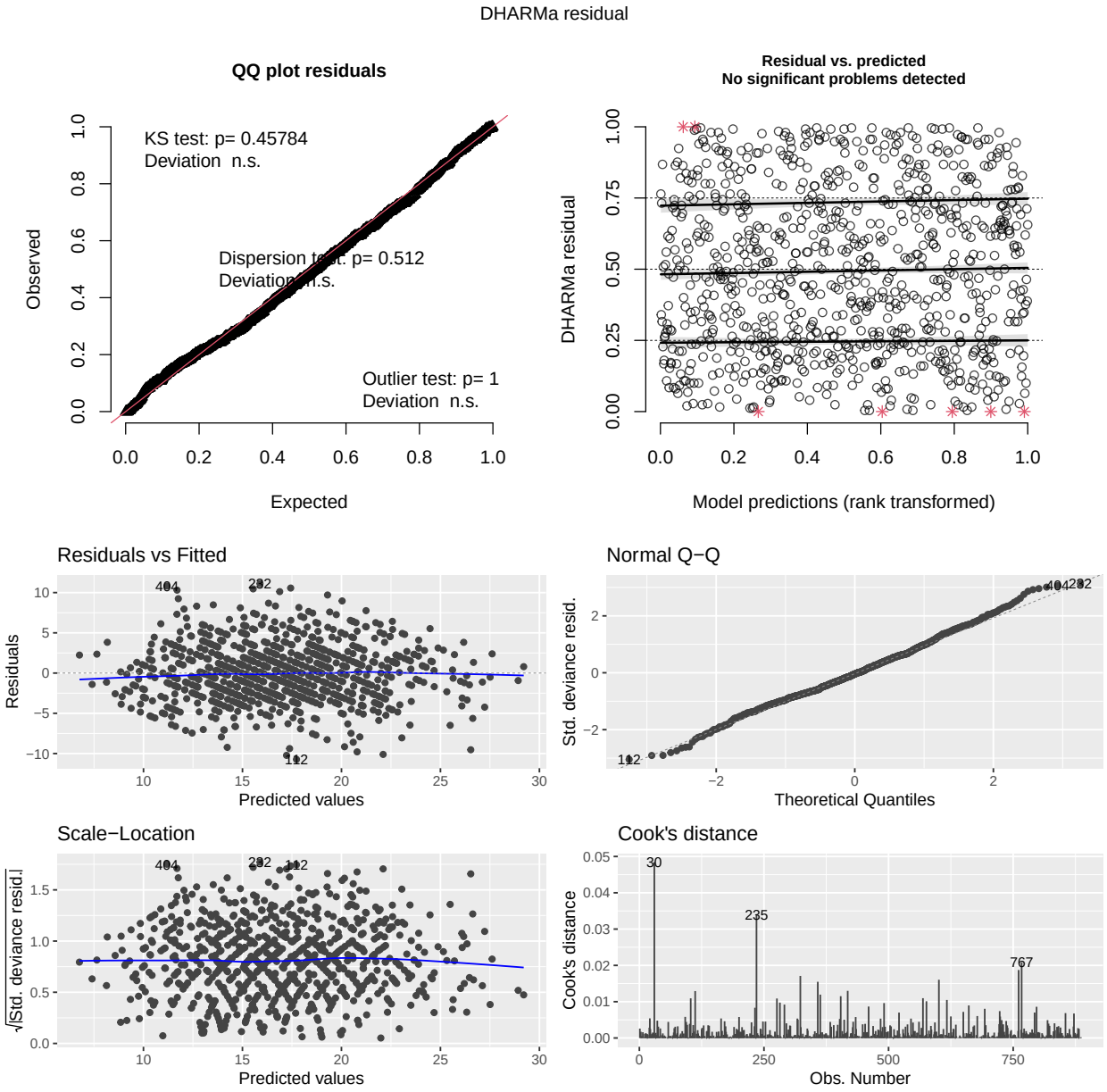
At this stage we were more interested in preliminary results of how well a given distribution worked for our data (no in-depth analysis of selected variables or attempts to interpret them).

We then fitted two GLMs with Gaussian distribution, i.e. linear regression models:

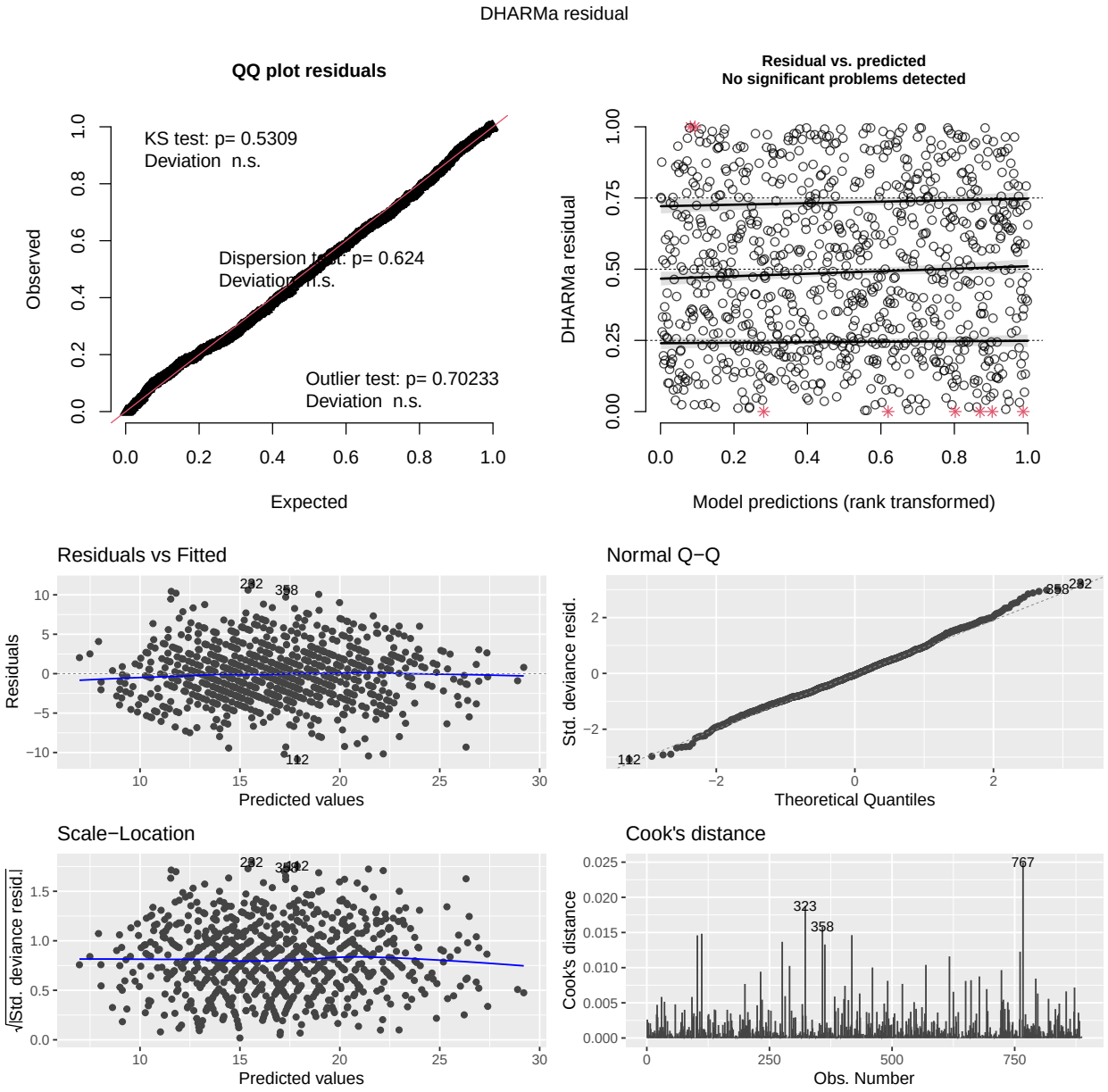
- Model with all variables included except `glang`:

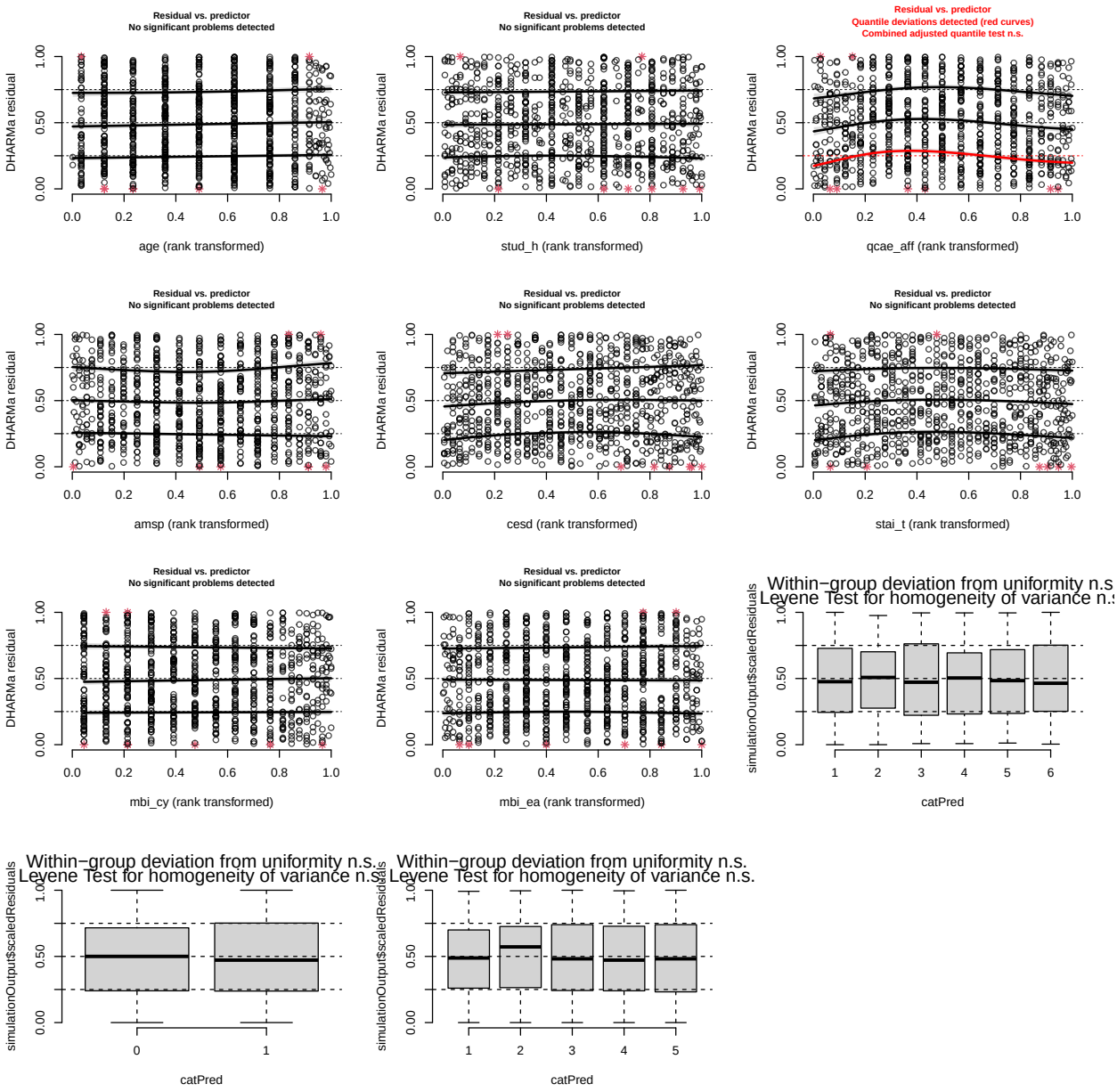
```
##
## Call:
## glm(formula = mbi_ex ~ age + stud_h + year + sex + part + job +
##     health + psyt + jspe + qcae_cog + qcae_aff + amsp + erec_mean +
##     cesd + stai_t + mbi_cy + mbi_ea, family = gaussian, data = data)
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.854612   2.424732   4.889 1.21e-06 ***
## age         -0.107847   0.047292  -2.280 0.022824 *
## stud_h       0.030596   0.009290   3.293 0.001030 **
## year2        1.744172   0.403898   4.318 1.76e-05 ***
## year3        1.474602   0.439994   3.351 0.000839 ***
## year4        0.404497   0.490023   0.825 0.409336
## year5       -0.786306   0.489339  -1.607 0.108449
## year6       -1.038267   0.557623  -1.862 0.062951 .
## sex2         0.552836   0.299867   1.844 0.065585 .
## sex3         0.417019   1.655344   0.252 0.801161
## part1        0.683641   0.252640   2.706 0.006945 **
## job1        -0.114200   0.269834  -0.423 0.672239
## health2      1.028293   0.725615   1.417 0.156806
## health3      0.376769   0.677560   0.556 0.578309
## health4      0.422971   0.624478   0.677 0.498386
## health5     -0.517943   0.645821  -0.802 0.422779
## psyt1       -0.265565   0.308742  -0.860 0.389946
## jspe         0.005264   0.016122   0.327 0.744117
## qcae_cog    -0.004268   0.022202  -0.192 0.847611
## qcae_aff     0.037378   0.026874   1.391 0.164621
## amsp         0.055180   0.027954   1.974 0.048710 *
## erec_mean   -1.501658   1.348420  -1.114 0.265743
## cesd         0.124938   0.016696   7.483 1.79e-13 ***
## stai_t       0.046411   0.015945   2.911 0.003699 **
## mbi_cy       0.331451   0.033603   9.864  < 2e-16 ***
## mbi_ea      -0.162348   0.035683  -4.550 6.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12.85592)
##
##     Null deviance: 24449  on 885  degrees of freedom
## Residual deviance: 11056  on 860  degrees of freedom
## AIC: 4804.6
##
## Number of Fisher Scoring iterations: 2
```



- Model after using `step()` function:


DHARMa residual
QQ plot residuals
KS test: p= 0.5309
Deviation n.s.
Dispersion test: p= 0.624
Deviation n.s.
Outlier test: p= 0.70233
Deviation n.s.
Observed
Expected
Residual vs. predicted
No significant problems detected
DHARMa residual
Model predictions (rank transformed)
Residuals vs Fitted
Residuals
Predicted values
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
Scale−Location
√|Std. deviance resid.|
Predicted values
Cook's distance
Cook's distance
Obs. Number
767
323
358



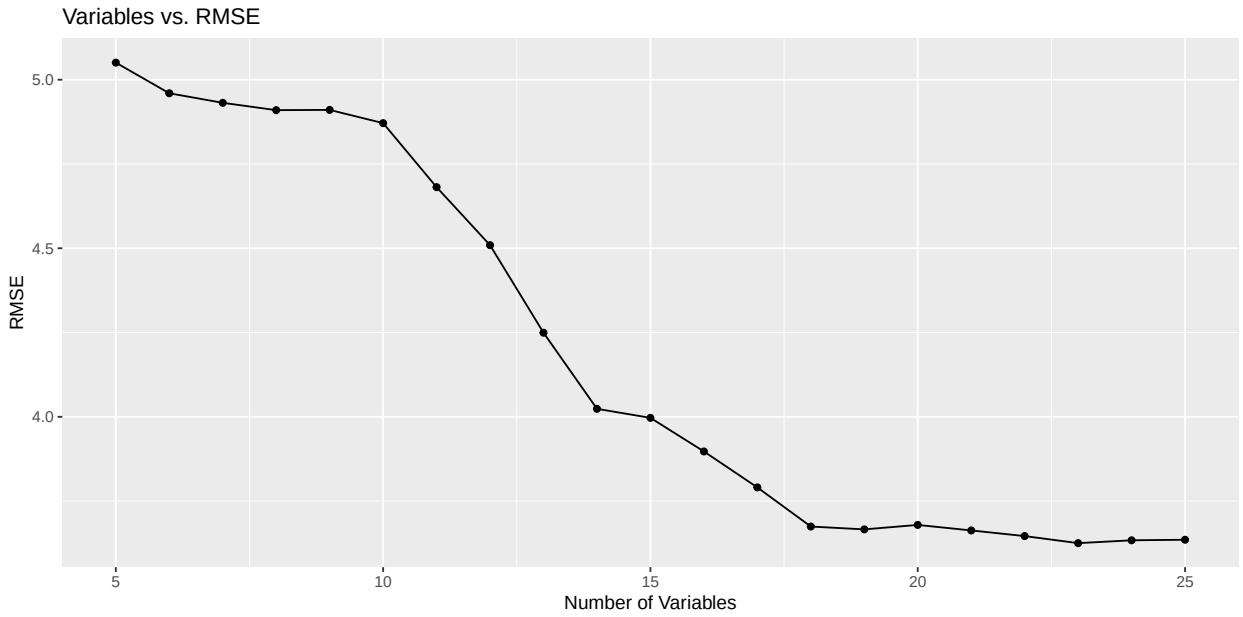
Interpretation:

- Both DHARMa residuals and standardised deviance residuals should exhibit no patterns (as much as it possible). However, it was not the case for both negative binomial models.
- The θ parameter for negative binomial model was very large, making basically equivalent to Poisson model, suggesting that $\mu \approx \sigma^2$ (Poisson distribution). In general, it seems more reasonable to fit a negative binomial regression as it will either converge to Poisson distribution or account for overdispersion to some extent.
- Using normal distribution yielded considerably better results: no noticeable trends / patterns in residuals.
- AIC was also lower for Gaussian GLM (lm) than for NB GLM.

As a result, we believe that choosing a normal distribution was justified.

(c) Variable selection

In this section, we attempted to summarise our findings in terms of variable importance based on AIC, cross-validation RMSE due to the fact that $Y|X$ was found to be well-approximated by a normal distribution (which makes it equivalent to maximising log likelihood) and EDA.



If variable was important, we used "+" to signify it.

```
##    variables cv_rmse aic eda
## 1       year       +   +   +
## 2        sex       +       +
## 3       part       +   +
## 4     health       +   +   +
## 5        job       +
## 6       psyt       +       +
## 7        age           +   *
## 8   qcae_cog
## 9       amsp           +
## 10    mbi_cy       +   +   +
## 11    mbi_ea       +   +   +
## 12      cesd       +   +   +
## 13    stud_h           + +/-
## 14      jspe
## 15    stai_t           +   +
## 16 erec_mean       +
## 17  qcae_aff           + +/-
```

Interpretation:

- In the final model we included variables that were important according to all 3 approaches (CV, AIC, EDA): `year`, `health`, `mbi_cy`, `mbi_ea`, `cesd`.
- Variables that were significant only according to AIC or CV but not EDA were not included in the final model. They gave only a slight boost in predictive power, but were not statistically significant: `job`, `amsp`, `erec_mean`.

- Variables that were important according to 2 out of 3 metrics were included as well, but more attention was paid to them when constructing the final model: `sex`, `part`, `psyt`, `study_h`, `stai_t`, `qcae_aff`.
- Variable `age` had an outlier (a very old student) that made the relationship seem significant. Since AIC works with the entire dataset, it can fail to circumvent such issues. Using cross-validation in this case can be more helpful since out-of-sample error will be high due to the outlier.

(d) Addressing potential clustering issues

From EDA we conducted, we observed that `glnag` was not an important predictor, primarily because the majority of students were French speaking. Nevertheless, we decided to make sure that we were not missing some crucial information by not including this variable.

We decided to test two options: GLMM and GLM with a modified version of `glang`. In this case, using GLMM was done not because we were interested in treating `glang` as a random draw from the general population, but rather because we wanted to use other benefits of partial pooling and account for possible among group variation. Other approach was to modify `glnag` by grouping all languages (except French) into another group and then fitting a GLM.

- GLMM (`glang` as random effect, i.e. random intercept only):

```
## boundary (singular) fit: see help('isSingular')
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: mbi_ex ~ year + sex + part + health + psyt + cesd + stud_h +
##     stai_t + qcae_aff + mbi_cy + mbi_ea + (1 | glang)
##    Data: data
##
## REML criterion at convergence: 4801.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.98164 -0.69583 -0.03924  0.63334  3.16222
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  glang    (Intercept)  0.00    0.000
##  Residual             12.93    3.596
## Number of obs: 886, groups:  glang, 19
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 10.134014   1.525910   6.641
## year2        1.528987   0.392335   3.897
## year3        1.118874   0.402062   2.783
## year4       -0.147594   0.431630  -0.342
## year5       -1.363154   0.429292  -3.175
## year6       -1.700826   0.475563  -3.576
## sex2         0.464613   0.291505   1.594
## sex3         0.254864   1.653460   0.154
## part1        0.640766   0.250206   2.561
## health2      1.002554   0.725503   1.382
## health3      0.293020   0.677409   0.433
## health4      0.422426   0.622750   0.678
```

```
## health5     -0.504397   0.643408  -0.784
## psyt1       -0.259952   0.308387  -0.843
## cesd         0.125418   0.016730   7.497
## stud_h       0.030727   0.009266   3.316
## stai_t       0.040001   0.015694   2.549
## qcae_aff     0.042363   0.025500   1.661
## mbi_cy       0.341867   0.033367  10.246
## mbi_ea      -0.154585   0.034789  -4.444

##
## Correlation matrix not shown by default, as p = 20 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

- GLM (**glang** merged into 2 categories):

```
##
## Call:
## glm(formula = mbi_ex ~ year + sex + part + health + psyt + glang_merged +
##     cesd + stud_h + stai_t + qcae_aff + mbi_cy + mbi_ea, family = gaussian,
##     data = data)
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.133105   1.526600   6.638 5.62e-11 ***
## year2          1.520445   0.392937   3.869 0.000117 ***
## year3          1.122704   0.402327   2.791 0.005378 **
## year4         -0.151468   0.431904  -0.351 0.725900
## year5         -1.368874   0.429660  -3.186 0.001495 **
## year6         -1.694251   0.475986  -3.559 0.000392 ***
## sex2           0.467554   0.291704   1.603 0.109336
## sex3           0.299114   1.656909   0.181 0.856783
## part1          0.634000   0.250736   2.529 0.011630 *
## health2        1.019163   0.726698   1.402 0.161138
## health3        0.314258   0.679234   0.463 0.643721
## health4        0.444818   0.624868   0.712 0.476744
## health5       -0.488079   0.644643  -0.757 0.449178
## psyt1         -0.268155   0.309024  -0.868 0.385773
## glang_merged2 -0.147877   0.316125  -0.468 0.640059
## cesd           0.125603   0.016742   7.502 1.56e-13 ***
## stud_h         0.030891   0.009277   3.330 0.000906 ***
## stai_t         0.040581   0.015750   2.577 0.010142 *
## qcae_aff       0.040722   0.025752   1.581 0.114162
## mbi_cy         0.342609   0.033419  10.252  < 2e-16 ***
## mbi_ea        -0.153272   0.034918  -4.390 1.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12.93951)
##
```

```
##     Null deviance: 24449  on 885  degrees of freedom
## Residual deviance: 11193  on 865  degrees of freedom
## AIC: 4805.5
##
## Number of Fisher Scoring iterations: 2
```

```
##   anova_pvalue kruskal_pvalue
## 1       0.4988         0.3272
```

Interpretation:

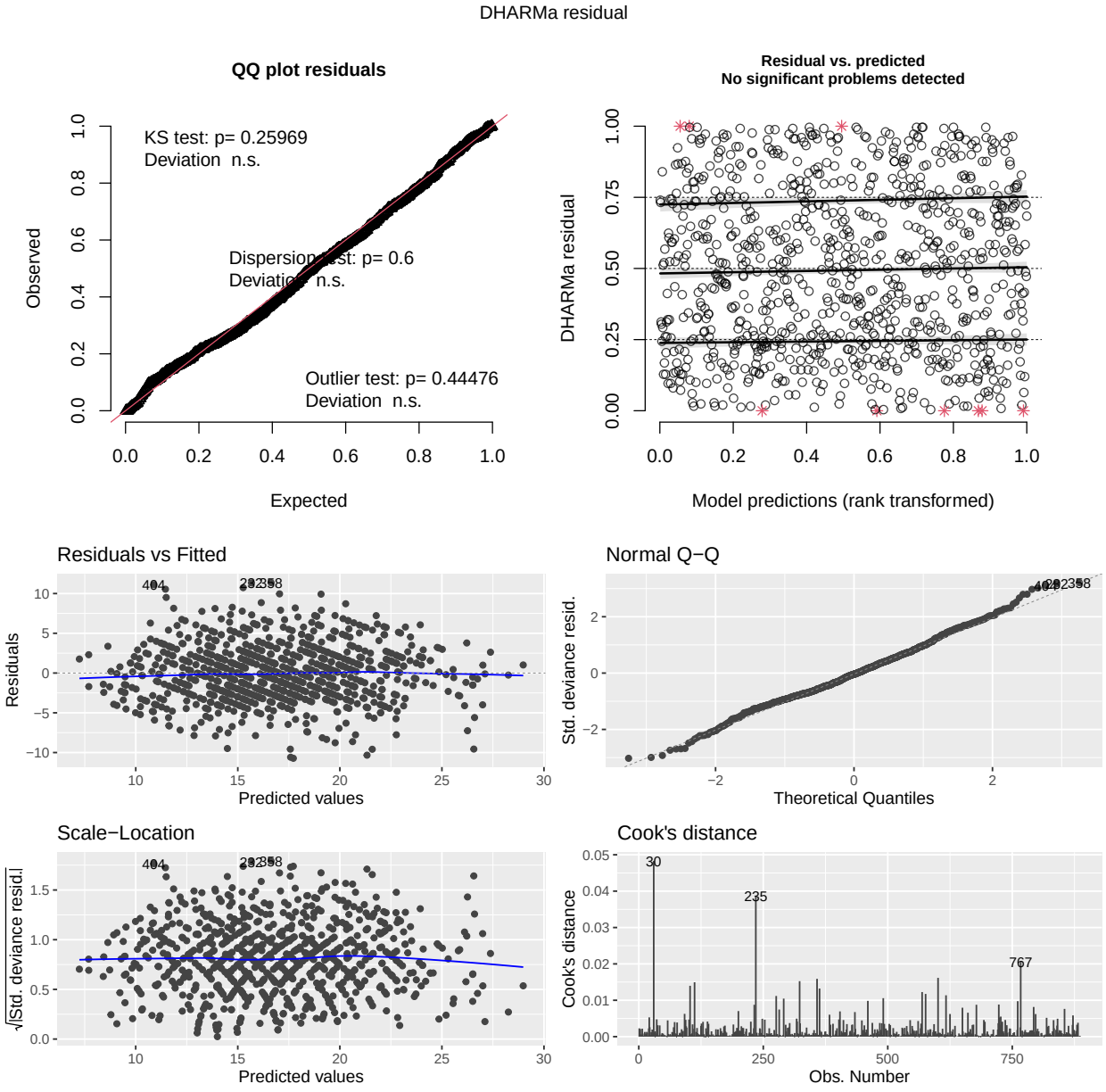
- The variance associated with the `glang` in the GLMM was 0, which meant that this factor wasn't important.
- Merging `glang` in two more general categories also did not contribute to improving the GLM.
- Thus, we confirmed our findings from the EDA stage.

5.2 Results. MBI Exhaustion

Fitting the final model:

```
##
## Call:
## glm(formula = mbi_ex ~ relevel(year, ref = 2) + sex + part +
##     relevel(health, ref = 2) + psyt + cesd + stud_h + stai_t +
##     qcae_aff + mbi_cy + mbi_ea, family = gaussian, data = data)
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 12.665554   1.503871   8.422  < 2e-16 ***
## relevel(year, ref = 2)1     -1.528987   0.392335  -3.897 0.000105 ***
## relevel(year, ref = 2)3     -0.410112   0.444767  -0.922 0.356742
## relevel(year, ref = 2)4     -1.676581   0.473706  -3.539 0.000423 ***
## relevel(year, ref = 2)5     -2.892140   0.471675  -6.132 1.32e-09 ***
## relevel(year, ref = 2)6     -3.229812   0.513378  -6.291 4.99e-10 ***
## sex2                         0.464613   0.291505   1.594 0.111337
## sex3                         0.254864   1.653460   0.154 0.877536
## part1                        0.640766   0.250206   2.561 0.010607 *
## relevel(health, ref = 2)1 -1.002554   0.725503  -1.382 0.167367
## relevel(health, ref = 2)3 -0.709533   0.498812  -1.422 0.155257
## relevel(health, ref = 2)4 -0.580128   0.449568  -1.290 0.197252
## relevel(health, ref = 2)5 -1.506951   0.500301  -3.012 0.002670 **
## psyt1                       -0.259952   0.308387  -0.843 0.399494
## cesd                         0.125418   0.016730   7.497 1.62e-13 ***
## stud_h                       0.030727   0.009266   3.316 0.000951 ***
## stai_t                       0.040001   0.015694   2.549 0.010979 *
## qcae_aff                     0.042363   0.025500   1.661 0.097013 .
## mbi_cy                       0.341867   0.033367  10.246  < 2e-16 ***
## mbi_ea                      -0.154585   0.034789  -4.444 1.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12.92784)
```

```
##
##     Null deviance: 24449  on 885  degrees of freedom
## Residual deviance: 11196  on 866  degrees of freedom
## AIC: 4803.7
##
## Number of Fisher Scoring iterations: 2
```



Interpretation:

- Alpha level was taken as 0.05.
- Variable `year`: our exploratory data analysis revealed a noticeable trend: starting from the second year, levels of `mbi_ex` tended to decrease more or less linearly. For instance, being a sixth-year student was associated with an average decrease of 3.23 points in emotional exhaustion compared to being a

second-year student. In the third year, we observed only a slight drop in 'mbi_ex,' which was not statistically significant. However, the second year, acting as the reference category, deviated from this linear trend. Freshmen exhibited lower levels of emotional exhaustion, comparable to those of third or even fourth-year students, albeit significantly higher than those of senior students.

- Variable `sex`: based on EDA, we concluded that `sex` was an important variable. However, when controlling for other relevant factors, our model did not find statistically significant differences between sexes, although women, on average, exhibited higher levels of emotional exhaustion than men. We chose to retain this variable as a control since it contributed to a "better: model, as indicated by improvements in AIC and CV RMSE scores.
- Variable `health`: similar to the trend observed for 'year,' there was a linearly decreasing relationship between `mbi_ex` and increasing levels of health satisfaction. The only exception was observed in group 1, consisting of students who were extremely unsatisfied with their health. However, due to the underrepresentation of this group, no definitive conclusions could be drawn regarding their emotional exhaustion levels. It's worth noting that the only statistically significant difference was observed in group 5, comprising students who were extremely satisfied with their health. Compared to group 2, their average levels of 'mbi_ex' decreased by 1.5 points.
- A cluster of variables, including depression `cesd`, anxiety `stai_t`, and cynicism `mbi_cy`, showed a positive relationship with `mbi_ex`. All variables were statistically significant. Notably, `mbi_cy` had the largest impact, with a slope of 0.34, despite having the narrowest scale, ranging from 5 to 25. Therefore, even after standardization, it would exhibit the highest impact. Cynicism measured by this scale serves as a coping mechanism for distancing oneself from exhausting job demands.
- Variable `stud_h` was also statistically significant and had a positive correlation with the target. However, the slope was very gradual.
- Last but not the least, higher academic efficiency `mbi_ea` corresponded to lower `mbi_ex`, suggesting that feelings of competence and successful achievement in one's work made a significant difference in lowering exhaustion.

It was also interesting to observe that `mbi_ex` was distributed approximately normally. We would usually expect such variables to be skewed as less people normally report high level of emotional exhaustion.

```
##           mean_high_mbi_ex std_high_mbi_ex mean_low_mbi_ex std_low_mbi_ex
## erec_mean              0.7             0.1             0.7            0.1
## mbi_cy                15.6             5.1             6.2            3.1
## mbi_ea                20.1             4.5            28.4            3.6
## age                   21.6             2.5            24.4            4.6
## amsp                  23.8             5.9            24.8            5.4
## mbi_ex                26.5             1.5             7.1            1.3
## stud_h                28.8            16.9            20.1           17.1
## cesd                  30.9            11.6             6.6            7.7
## qcae_aff              36.9             5.2            30.6            5.8
## stai_t                52.7            12.3            28.0            7.4
## qcae_cog              59.3             6.5            59.4            6.6
## jspe                 107.5             8.3           109.1            9.3
```

If we compare which variables significantly changed in the bottom 10% as opposed to the top 10%, the following becomes apparent:

- Students with high levels of `mbi_ex` reported being more indifferent, `mbi_cy` scores more than doubled for those students.
- Those students who had low levels of exhaustion spent considerably less time studying `stud_h` and yet had higher levels of academic efficiency `mbi_ea`.
- Moreover, students in bottom 10% reported a 5-fold decrease in depression scores `cesd`. That was the most significant difference.

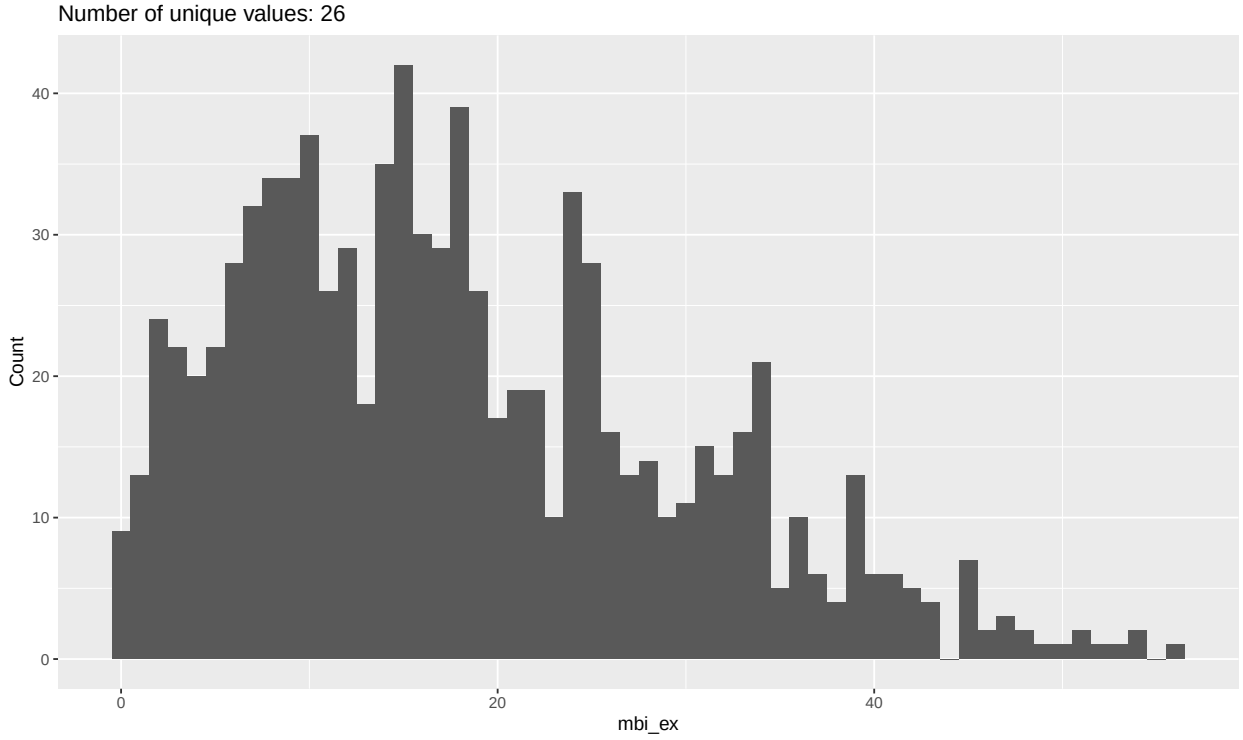
5.3 Analysis. CESD

After conducting EDA, we detected signs of non-linearity, particularly in the variable `stai_t.` Consequently, we opted for a Generalized Additive Model (GAM) to test whether the effect was significantly non-linear. Additionally, since we determined that `glang` was not significant, we chose not to fit a Generalized Linear Mixed Model (GLMM).

(a) Distribution considerations. Theoretical assumptions

Variable `cesd`:

- Integer-valued and bounded between 0 and 56.
- 26 unique values.
- The mean value is equal to 18 and median to 16.
- The standard deviation is 11.5.
- Unconditional distribution of `cesd`:

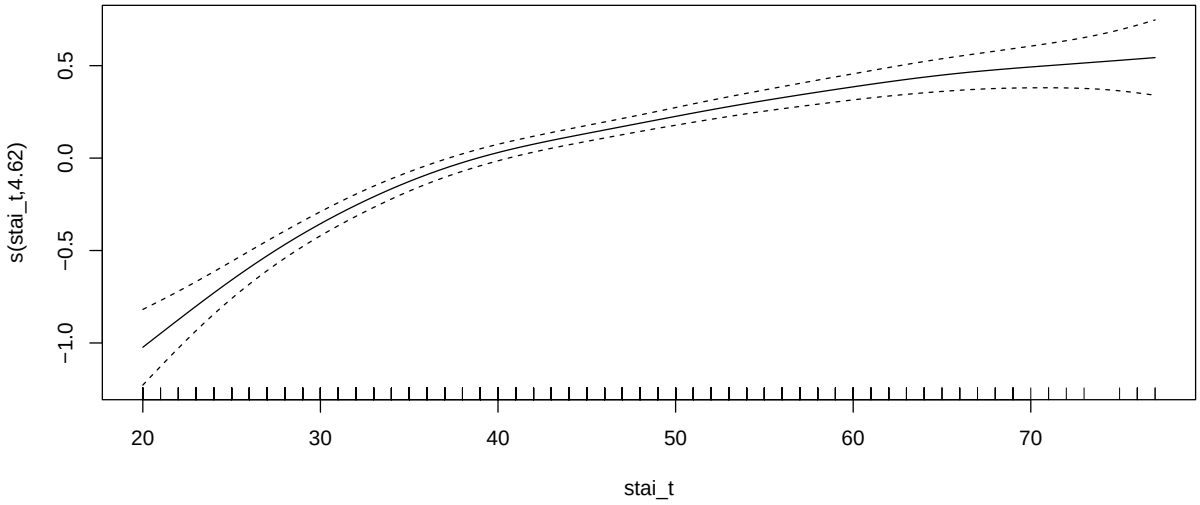


What we considered was utilising the Tweedie distribution due to its high flexibility. Depending on the power parameter p, it can span from Poisson to Gamma distributions, allowing us to accommodate integer-valued data, the presence of zeros, and other factors.

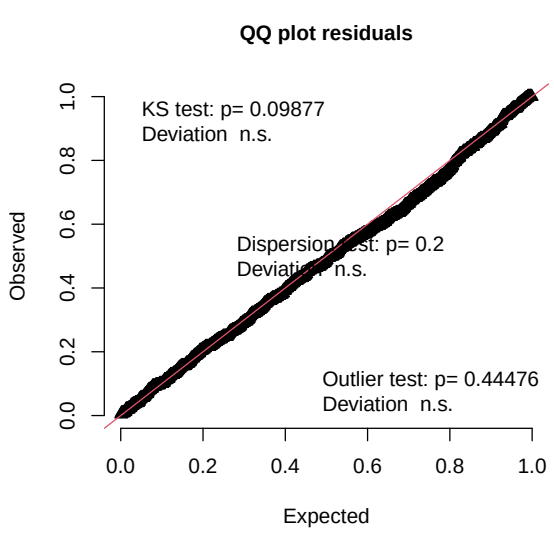
(b) Distribution considerations. Testing assumptions

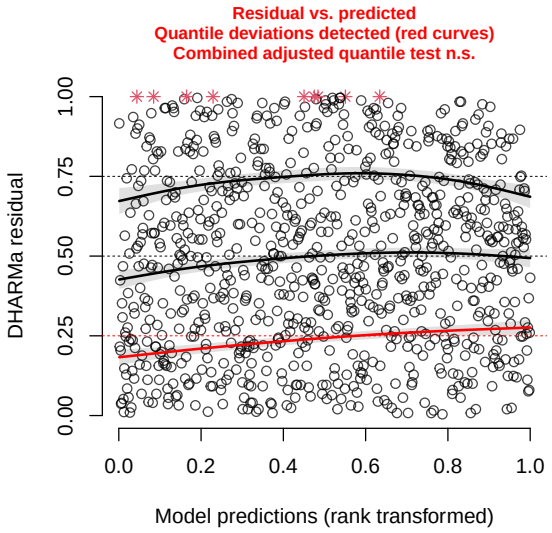
```
##
## Family: Tweedie(p=1.173)
## Link function: log
##
## Formula:
```

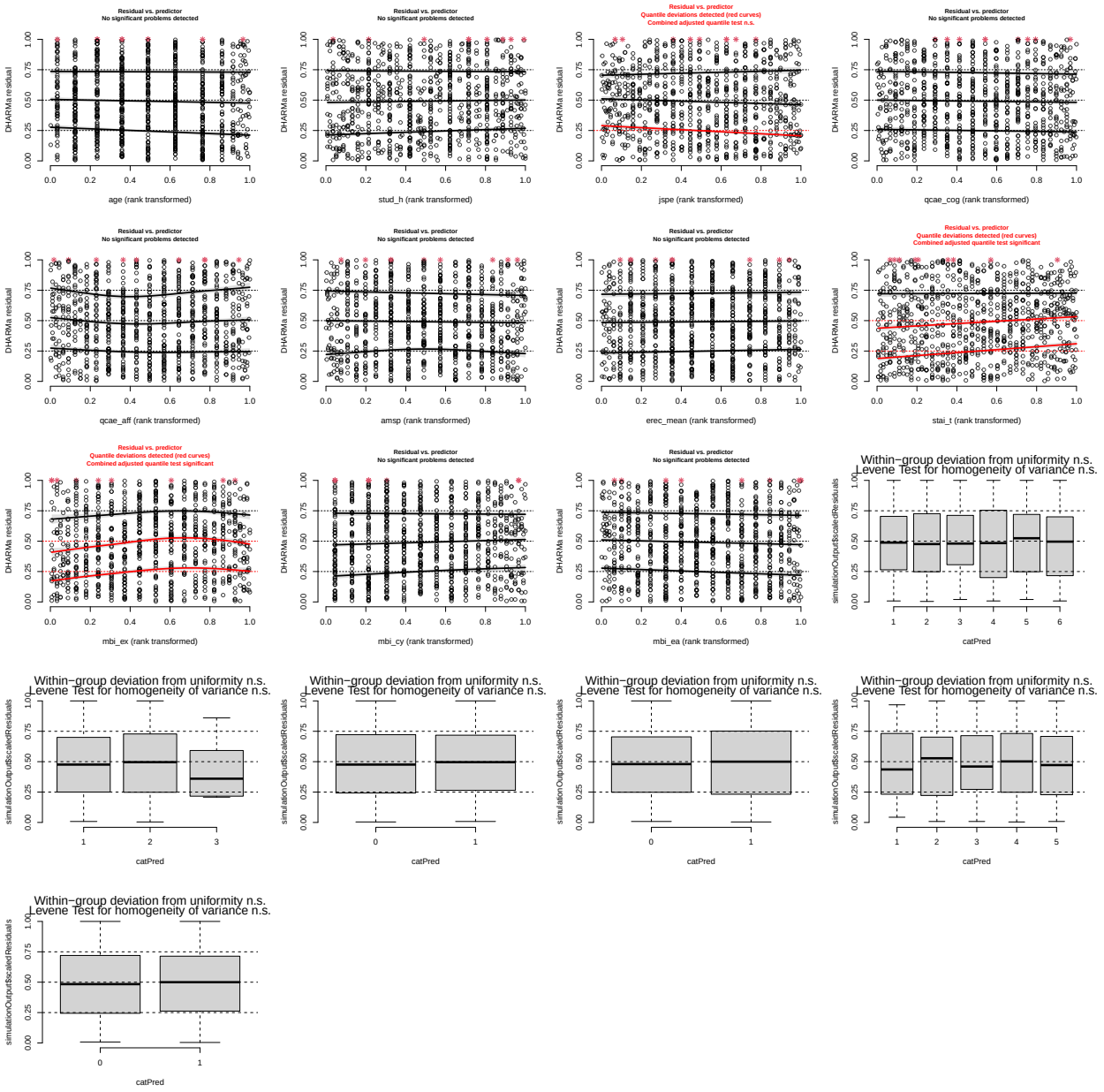
```
## cesd ~ age + stud_h + year + sex + part + job + health + psyt +
##     jspe + qcae_cog + qcae_aff + amsp + erec_mean + mbi_ex +
##     mbi_cy + mbi_ea + s(stai_t, k = 10)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1210369  0.2746690   7.722 3.19e-14 ***
## age          0.0067427  0.0054231   1.243 0.214083
## stud_h       0.0009144  0.0010353   0.883 0.377328
## year2       -0.1294200  0.0449062  -2.882 0.004050 **
## year3       -0.1862665  0.0490221  -3.800 0.000155 ***
## year4       -0.1675191  0.0555309  -3.017 0.002631 **
## year5       -0.2106119  0.0563776  -3.736 0.000200 ***
## year6       -0.2106823  0.0650140  -3.241 0.001239 **
## sex2         0.0802446  0.0360496   2.226 0.026277 *
## sex3        -0.1460719  0.2036970  -0.717 0.473506
## part1       -0.0685659  0.0286841  -2.390 0.017046 *
## job1         0.0061880  0.0311919   0.198 0.842791
## health2      0.1463247  0.0802621   1.823 0.068639 .
## health3      0.1208631  0.0778484   1.553 0.120901
## health4      0.0189845  0.0749565   0.253 0.800118
## health5     -0.0563263  0.0793318  -0.710 0.477892
## psyt1        0.0784869  0.0324065   2.422 0.015644 *
## jspe        -0.0005271  0.0018124  -0.291 0.771240
## qcae_cog     0.0020350  0.0024793   0.821 0.411989
## qcae_aff     0.0023427  0.0030963   0.757 0.449488
## amsp        -0.0002768  0.0030548  -0.091 0.927811
## erec_mean    0.0093276  0.1516138   0.062 0.950958
## mbi_ex       0.0290360  0.0037461   7.751 2.58e-14 ***
## mbi_cy       0.0097750  0.0037927   2.577 0.010124 *
## mbi_ea      -0.0087832  0.0041162  -2.134 0.033143 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(stai_t) 4.621  5.681 42.59  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.61   Deviance explained = 61.3%
## -REML = 2986.7  Scale est. = 1.7687    n = 886
```



DHARMa residual







Interpretation:

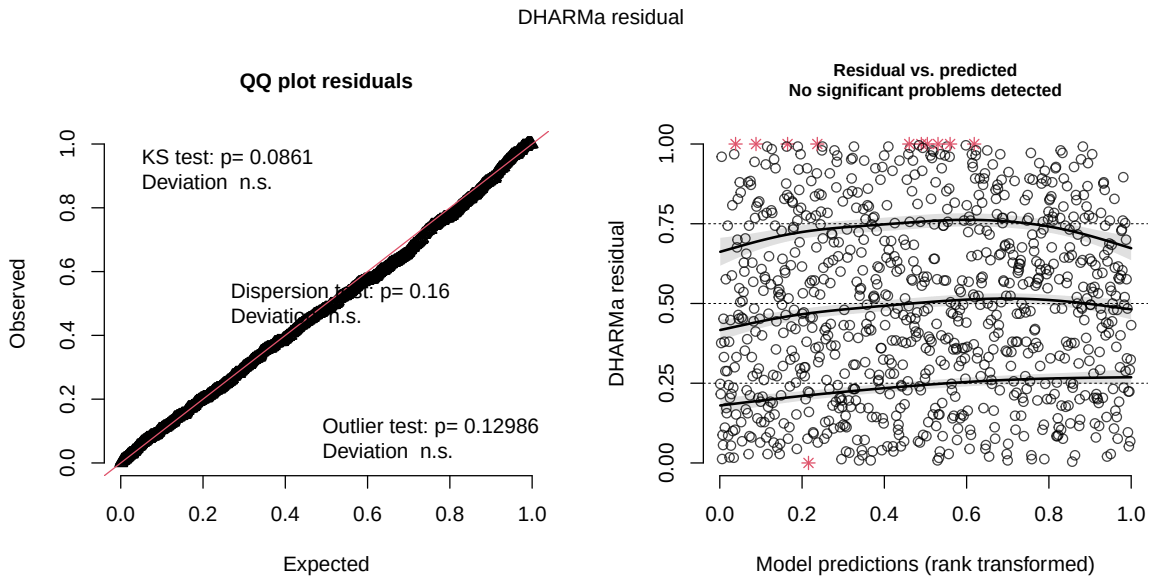
- The effective degrees of freedom for `stai_t` were equal to 4.621, suggesting the presence of a statistically significant non-linear effect.
- Plotting DHARMa residuals against `stai_t` showed deviations from uniformity. After including `stai_t` as a non-linear effect this issue should be resolved.

(c) Variable selection

Since `step()` function was not available for GAMs and using RMSE also didn't make sense for the Tweedie distributed target variable, we opted to rely on EDA as well as other metrics such as adjusted R-sq, explained deviance and AIC. For the initial model, adjusted R-sq was 0.61, deviance explained was 61.3% and AIC was 5856.

Based on EDA, `age`, `erec_mean`, `qcae_cog`, `amsp`, `jspe` did not contribute to explaining `cesd`. Moreover, they also had extremely high p-values, but we decided not to rely on p-values in selecting important variables as type I errors will add up (multiple comparison problem).

```
##
## Family: Tweedie(p=1.173)
## Link function: log
##
## Formula:
## cesd ~ stud_h + year + sex + part + job + relevel(health, ref = 2) +
##     psyt + qcae_aff + mbi_ex + mbi_cy + mbi_ea + s(stai_t, k = 10)
##
## Parametric coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.4368792  0.1645149  14.813  < 2e-16 ***
## stud_h                      0.0009318  0.0010318   0.903 0.366737
## year2                      -0.1175191  0.0431971  -2.721 0.006649 **
## year3                      -0.1676965  0.0447324  -3.749 0.000189 ***
## year4                      -0.1460663  0.0493446  -2.960 0.003159 **
## year5                      -0.1805448  0.0495410  -3.644 0.000284 ***
## year6                      -0.1743952  0.0563064  -3.097 0.002017 **
## sex2                        0.0782556  0.0349539   2.239 0.025422 *
## sex3                       -0.1389449  0.2024897  -0.686 0.492783
## part1                      -0.0637799  0.0282349  -2.259 0.024139 *
## job1                        0.0112156  0.0308243   0.364 0.716055
## relevel(health, ref = 2)1 -0.1458381  0.0796793  -1.830 0.067549 .
## relevel(health, ref = 2)3 -0.0273759  0.0479083  -0.571 0.567861
## relevel(health, ref = 2)4 -0.1338820  0.0441840  -3.030 0.002518 **
## relevel(health, ref = 2)5 -0.2114706  0.0528941  -3.998 6.94e-05 ***
## psyt1                       0.0788392  0.0322987   2.441 0.014849 *
## qcae_aff                    0.0025966  0.0029308   0.886 0.375877
## mbi_ex                      0.0286844  0.0037147   7.722 3.18e-14 ***
## mbi_cy                      0.0099981  0.0037586   2.660 0.007957 **
## mbi_ea                     -0.0077493  0.0039821  -1.946 0.051974 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(stai_t) 4.642  5.704 44.86  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.61   Deviance explained = 61.2%
## -REML = 2967.1  Scale est. = 1.7669    n = 886
```



Interpretation:

- After removing the aforementioned variables, adjusted R-sq became 0.61, deviance explained dropped slightly from 61.3% to 61.2% and AIC decreased from 5856 to 5847 (a very minor improvement).
- Although `qcae_aff` was insignificant, it helped improve the model (residual-wise).

5.4 Results. CESD

```
##
## Family: Tweedie(p=1.173)
## Link function: log
##
## Formula:
## cesd ~ stud_h + year + sex + part + job + relevel(health, ref = 2) +
##     psyt + qcae_aff + mbi_ex + mbi_cy + mbi_ea + s(stai_t, k = 10)
##
## Parametric coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.4368792  0.1645149  14.813  < 2e-16 ***
## stud_h                       0.0009318  0.0010318   0.903 0.366737
## year2                       -0.1175191  0.0431971  -2.721 0.006649 **
## year3                       -0.1676965  0.0447324  -3.749 0.000189 ***
## year4                       -0.1460663  0.0493446  -2.960 0.003159 **
## year5                       -0.1805448  0.0495410  -3.644 0.000284 ***
## year6                       -0.1743952  0.0563064  -3.097 0.002017 **
## sex2                         0.0782556  0.0349539   2.239 0.025422 *
## sex3                        -0.1389449  0.2024897  -0.686 0.492783
## part1                       -0.0637799  0.0282349  -2.259 0.024139 *
## job1                         0.0112156  0.0308243   0.364 0.716055
## relevel(health, ref = 2)1 -0.1458381  0.0796793  -1.830 0.067549 .
## relevel(health, ref = 2)3 -0.0273759  0.0479083  -0.571 0.567861
## relevel(health, ref = 2)4 -0.1338820  0.0441840  -3.030 0.002518 **
```

```
## relevel(health, ref = 2)5 -0.2114706  0.0528941  -3.998 6.94e-05 ***
## psyt1                      0.0788392  0.0322987   2.441 0.014849 *
## qcae_aff                   0.0025966  0.0029308   0.886 0.375877
## mbi_ex                     0.0286844  0.0037147   7.722 3.18e-14 ***
## mbi_cy                     0.0099981  0.0037586   2.660 0.007957 **
## mbi_ea                    -0.0077493  0.0039821  -1.946 0.051974 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(stai_t) 4.642  5.704 44.86  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.61   Deviance explained = 61.2%
## -REML = 2967.1  Scale est. = 1.7669    n = 886
```

Interpretation:

- Alpha level was taken as 0.05.
- `year`: the coefficients for `year2` to `year6` are all negative, indicating lower `cesd` scores compared to first-year students, with all differences being statistically significant. This suggests a general decrease in `cesd` scores over time. For example, for a given sample, being a fifth-year student decreased depression symptoms by a factor of $e^{-0.18} = 0.84$ compared to the reference category (`year1`).
- `sex`: `sex2` i.e. female has a positive coefficient (0.0783), suggesting higher `cesd` scores for this group compared to the reference sex group (male), with statistical significance. There were not enough observations in the group `sex3` to reliably detect the difference between categories.
- `part1`: shows a negative coefficient, suggesting lower `cesd` scores for this group compared to its reference, with statistical significance. That means that having a partner helped students battle depression.
- `job1`: having part time job or not did not have any statistical significance.
- `health`: significant effects are observed for health levels 1, 4, and 5 compared to the reference level 2, with levels 4 and 5 showing negative effects on `cesd`. This indicates students who were satisfied with their health had lower CESD scores. But just like for `mbi_ex`, there were very few students who were very dissatisfied with their health (`health1`), making it impossible to properly assess the effect for this category.
- `pyst` (psychological therapy): indicates a positive association with `cesd` scores, suggesting that undergoing psychological therapy is associated with higher `cesd` scores, with statistical significance. It may seem contradictory, but we pose that students who had mental health issues would more likely to go to a specialist.
- `qcae_aff` (QCAE Affection), `mbi_ex` (MBI Exhaustion), `mbi_cy` (MBI Cynicism), and `mbi_ea` (MBI Efficacy): of these, `mbi_ex` and `mbi_cy` show significant positive relationships with `cesd` scores, indicating that higher levels of exhaustion and cynicism are associated with higher `cesd` scores.
- `mbi_ea` shows a negative but marginally significant ($p = 0.051974$) relationship, suggesting that higher professional efficacy might be associated with slightly lower `cesd` scores. But based on the chosen alpha level, it was not significant.

Interpretation of Smooth Terms: The smooth term for `stai_t` is statistically significant, with an effective degree of freedom (edf) of 4.642. So it means the smooth term captures a nonlinear relationship between `stai_t` and `cesd`. The non-linear effect persisted from 20 to 40, and after that it essentially became linear.

6. Conclusion

We conducted an in-depth exploratory data analysis and ensured that the models we built were well-specified, using appropriate distribution, analysing residuals, selecting important variables, and ensuring that no clustering effects would render our models invalid.

It is important to keep in mind that our findings should not be simply generalised to students from other universities and/or countries, as the sample we studied was very homogeneous, namely medical students from the University of Lausanne and Lausanne University Hospital. Nevertheless, it is possible that the underlying data generation process is similar for students globally, resulting in some conclusions being relevant for other samples as well, which can be inferred based on meta-analysis of works in the same field.

We found that senior students were considerably less exhausted and depressed. Additionally, academic efficiency was barely dependent on the number of hours students studied. Furthermore, those who reported spending considerably more time studying, i.e., younger students, were significantly more emotionally overextended.

These findings are important as more attention should be paid to guiding younger students in their academic journey. For example, they can benefit from learning about stress management and seeking support from on-campus specialists. More emphasis should be put on long-term academic success and sustainability. Moreover, polling senior students to learn how they were able to overcome these challenges could provide valuable insights.

There were also some gender differences, suggesting that females were feeling more depressed. Further analysis is required to understand why this discrepancy held true in the given sample.

References:

[1] Carrard, V., Bourquin, C., Berney, S., Schlegel, K., Gaume, J., Bart, P.-A., Preisig, M., Schmid Mast, M., & Berney, A. (2022). Dataset for the paper "The relationship between medical students' empathy, mental health, and burnout: A cross-sectional study" published in Medical Teacher (2022) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5702895

[2] Hojat, M., DeSantis, J., Shannon, S. C., Mortensen, L. H., Speicher, M. R., Bragan, L., LaNoue, M., & Calabrese, L. H. (2018, December). *The Jefferson Scale of empathy: A nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students.* Advances in health sciences education???: theory and practice. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6245107/

[3] Reniers, R. L., Corcoran, R., Drake, R., Shryane, N. M., & V?llm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment*, *93*(1), 84-95. https://doi.org/10.1080/00223891.2010.528484

[4] Cramer, K. M., & Gruman, J. A. (2002). The Lennox and Wolfe Revised Self-Monitoring Scale: latent structure and gender invariance. *Personality and Individual Differences*, 32(4), 627-637. https://doi.org/10.1016/S0191-8869(01)00065-4

[5] Schlegel, K., Grandjean, D., & Scherer, K. R. (2012). Geneva emotion recognition test. *PsycTESTS Dataset*. https://doi.org/10.1037/t36935-000

[6] American Psychological Association. (n.d.). *Center for Epidemiological Studies Depression (CESD)*. American Psychological Association. https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/depression-scale

[7] American Psychological Association. (n.d.-b). *The state-trait anxiety inventory (STAI)*. American Psychological Association. https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/trait-state

[8] Wikimedia Foundation. (2023, June 22). *Maslach Burnout Inventory*. Wikipedia. https://en.wikipedia.org/wiki/Maslach_Burnout_Inventory#Maslach_Burnout_Inventory_Scales