Akademia Górniczo-Hutnicza
Wydział Matematyki Stosowanej

Qualitative Data Analysis

Final Project

Mateusz Mglej

Kraków, January 28, 2025

1 Dataset and Formulation of the Research Problem

1.1 National Poll on Healthy Aging

The **NPHA Wave 1** dataset contains information from the National Poll on Healthy Aging, which was conducted in the United States in April 2017 [1]. The survey was carried out by the GfK Group on behalf of the University of Michigan. It targeted non-institutionalized adults aged 50 to 80 who reside in the United States.

GfK selected households using the *KnowledgePanel* system in a way that ensured the sample was representative of the U.S. population. Once assigned to the survey, respondents received an email notification informing them that the survey was available for participation. The message included a link to the questionnaire; no login or password was required. Participants were asked questions about health insurance, household composition, sleep issues, dental care, prescription medications, and medical care. The median time to complete the survey was 8 minutes. The main survey completion rate was 75% for respondents aged 50–64 and 80% for those aged 65–80.

1.2 Dataset for Analysis

For this analysis, we use a subset of the original dataset, which is available in [2]. The authors selected 14 features related to health and sleep to be used in the task of predicting the number of different doctors a respondent had seen over the course of a year. They then removed all responses with missing values for any of the selected features, leaving 714 records.

I processed the dataset in the following way:

- The target variable, the number of visits to different doctors, had 3 categories: *0–1, 2–3, 4 or more*. For the purposes of binary analysis, I merged the last two categories, resulting in: *0–1, 2+*.
- I noticed that the variable *Age*, which in the original dataset had values *50–64* and *65–80*, only includes the latter in the reduced dataset. Therefore, it is irrelevant for the analysis and was removed.
- In the survey, one of the possible responses was "refused." There are 18 such cases in the dataset. These can be considered missing values, so I removed them. The dataset now contains 696 records.
- The values in the dataset were originally coded numerically. For easier analysis, I replaced these numeric codes with textual labels, using the instructions provided in [2]. Additionally, I shortened the names of variables that seemed too long.

1.3 Description of Variables

All variables considered in the analysis are qualitative. Several of them (*Count.Visits*, *Phys.Health*, *Mental.Health*, *Trouble.Sleeping*, *Sleep.Medication*) are also ordinal. The dependent variable is **Count.Visits**, which indicates whether the respondent saw at most one doctor or more than one during the past year. All variables, along with their descriptions and possible values, are presented in Table 1.

Code	Variable	Description	Possible Values
CV	Count.Visits	Number of different doctors visited in the past year.	0–1, 2+
PS	Phys.Health	Self-assessment of physical health.	Poor, Fair, Good, VeryGood, Excellent
MH	Mental.Health	Self-assessment of mental health.	Poor, Fair, Good, VeryGood, Excellent
DH	Dental.Health	Self-assessment of oral or dental health. (*Dentures – prosthesis*).	Dentures, Excellent, Fair, Good, Poor, VeryGood
Emp	Employment	Employment status of the respondent.	Full-Time, Not-Working, Part-Time, Retired
SvS	Stress.vs.Sleep	Whether stress affects the ability to fall asleep.	No, Yes
MvS	Medication.vs.Sleep	Whether medication affects sleep.	No, Yes
PvS	Pain.vs.Sleep	Whether physical pain interferes with sleep.	No, Yes
BvS	Bathroom.vs.Sleep	Whether the need to use the bathroom affects sleep.	No, Yes
UvS	Unknown.vs.Sleep	Whether unknown factors affect sleep.	No, Yes
TS	Trouble.Sleeping	Degree to which sleep is a problem.	LittleOrNot, Some, GreatDeal
SM	Sleep.Medication	Information on any sleep medications prescribed to the respondent.	Dont.Use, Occasionally, Regularly
Rac	Race	Racial or ethnic background of the respondent. (*Hispanic – of Hispanic origin*).	2+Races.NoHis, Black.NoHis, Hispanic, Other.NoHis, White.NoHis
Gen	Gender	Gender identity of the respondent.	Female, Male

Table 1: Presentation of variables

1.4 Research Problem

The goal of this analysis is to predict the number of different doctors a respondent visited in the past year based on their answers to the survey. We apply logistic regression to this binary classification task: whether the respondent visited at most one doctor in the previous year, or more than one. A well-performing model may be useful in planning public health strategies.

2 Data Analysis

2.1 Dependent Variable

The dataset under consideration contains 696 observations. For the dependent variable **Count.Visits**, we observe the following distribution:

- 0–1: 126 observations
- 2+: 570 observations

A clear imbalance is visible in this variable. As many as 82% of respondents visited at least two different doctors in the past year. Only 18% fall into the first category — those who visited at most one doctor during that time.

Such an imbalance in the data may lead to a situation where the majority class dominates the classification results, potentially reducing the performance and interpretability of predictive models.

2.2 Testing Variable Independence

Before proceeding with model construction, it is important to examine the variables for independence. A natural starting point is the chi-squared test of independence. Unfortunately, in many cases this test returns warnings. The reason for this becomes clear in Table 2, which presents a contingency table for the variables *Mental.Health* and *Phys.Health*.

MH \ PH	**Poor**	**Fair**	**Good**	**VeryGood**	**Excellent**
Poor	1	1	0	0	0
Fair	4	18	10	1	0
Good	7	47	95	14	2
VeryGood	5	40	113	110	10
Excellent	4	15	68	109	22

Table 2: Contingency table for the variables *Mental.Health* and *Phys.Health*

Although the dataset contains nearly 700 observations, some cells in this (and many other) tables have very small counts, which makes the chi-squared test unsuitable. While many variables are nominal, some are ordinal, so it is worthwhile to use Fisher's exact test for independence. We use the `fisher.gamma.test()` function with 10,000 simulations, whose implementation was presented during the lecture [**?**].

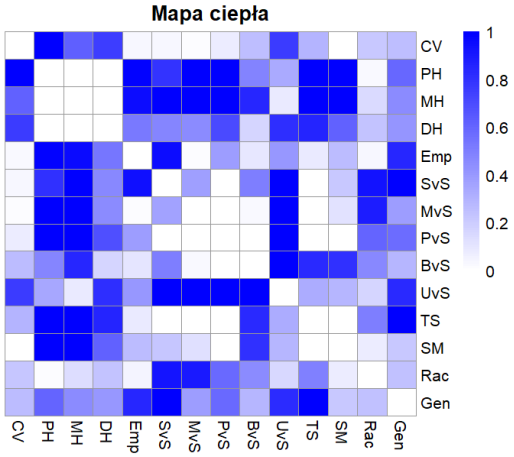

Figure 1: Heatmap for Fisher's exact test results

Figure 1 presents the results of the test. They are not conclusive. We can observe that there may be some dependence, for example, between the variables *Phys.Health*, *Mental.Health*, and *Dental.Health*, as well as among *Medication.vs.Sleep*, *Pain.vs.Sleep*, and *Bathroom.vs.Sleep*. At this stage, no variable is excluded from further analysis.

2.3 Preparation for Model Building

After building the models, we will evaluate their predictive performance. To do this, we will split the dataset into a training set and a test set in an 80/20 ratio. We will use the `createDataPartition()` function from the `caret` package, which ensures class balance in the dependent variable. The random seed is set to 123, using R version 4.4.0.

2.4 Model Evaluation Criteria

As previously noted, our dataset is imbalanced, so we must use appropriate metrics to assess the models' performance. In this case, we will use the **F1-score**, which is the harmonic mean of **precision** (the proportion of

predicted positives that are correct) and **recall** (the proportion of actual positives that were correctly identified). The F1-score is calculated as follows:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{where}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}, \quad \text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}.$$

To be thorough, we will also include another metric: the **Matthews correlation coefficient (MCC)**. Both metrics are suitable for imbalanced datasets and are intended to minimize both false positives (FP) and false negatives (FN).

3 Models with a Qualitative Dependent Variable

The dependent variable is binary, so we will build logistic regression models. We will use the `glm()` function with the parameter `family=binomial`. First, we construct two logit models: a full model and a reduced one based on variable selection using stepwise regression with the `step()` function. We then repeat this process using the *cloglog* (complementary log-log) link function.

3.1 Model 1

We build a saturated logit model without interactions. After training it on the training dataset, we obtain an AIC value of 539.63, which is higher than that of the null model (529.37). The `anova()` test shows that the full model's variables significantly improve the model's fit compared to the null model. However, the model does not fit some of the data well, which is visible in the plot of standardized Pearson residuals in Figure 2.

The predictive performance, after selecting the optimal decision threshold, is just over 50%. This is not a good result, as shown by the F1-score and MCC values (both of which should be close to 1 for a perfect model).

```
          Reference
Prediction 0-1 2+
       0-1  22 65
       2+    3 49
> f1
       F1
0.3928571
> MCC
[1] 0.2459104
```

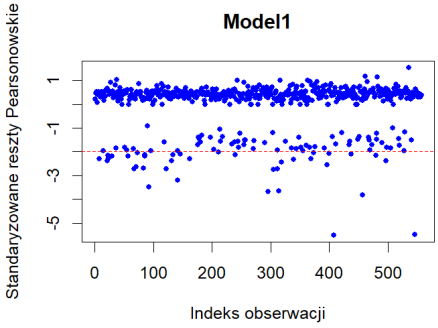

Figure 2: Prediction results and standardized Pearson residuals for Model 1

3.2 Model 2

Using the `step()` function, we build another logit model. We perform backward stepwise regression, where variables contributing the least to model fit are gradually removed from the full model. The algorithm returns a model with 5 variables: *Dental.Health*, *Medication.vs.Sleep*, *Sleep.Medication*, *Race*, and *Gender*. The AIC of this model is 519.3, which is lower than both the saturated and null models. The p-value of the test `anova(model1, model2)` is 0.69, meaning there is no evidence that the full model is better. On the other hand, the Pearson

residuals are still concerning, and prediction on the test set using the optimal decision threshold performs worse than the full model. This time the model predicts the majority class more accurately but performs significantly worse for the minority class, as seen in the calculated evaluation metrics.

```
          Reference
Prediction 0-1 2+
       0-1  12 32
       2+   13 82
> f1
       F1
0.3478261
> MCC
[1] 0.1645651
```

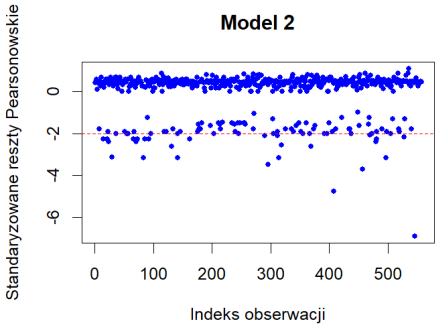

Figure 3: Prediction results and standardized Pearson residuals for Model 2

3.3 Model 3

We now build another full model, but this time using a different link function: complementary log-log (`link="cloglog"`). The resulting AIC is 537.57. Again, the null model performs better in this respect (529.37). As before, the `anova()` test rejects the null hypothesis that the additional variables in the full model do not improve model fit. The Pearson residuals look similar to previous models.

However, when calling `summary(model3)`, we finally see three variables marked with asterisks, indicating statistical significance. We obtain p-values below 0.05 for the nominal contrast *Dental.HealthExcellent* and the linear contrasts *Phys.Health.L* and *Sleep.Medication.L*. This model achieves the best classification performance so far.

```
          Reference
Prediction 0-1 2+
       0-1  18 44
       2+    7 70
> f1
       F1
0.4137931
> MCC
[1] 0.2580913
```

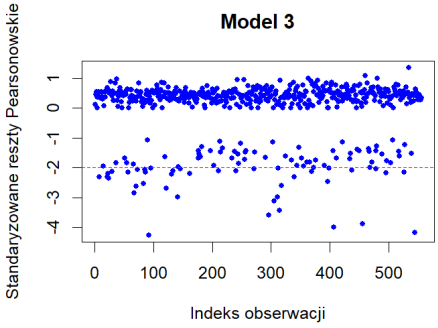

Figure 4: Prediction results and standardized Pearson residuals for Model 3

3.4 Model 4

The final model is the result of backward selection using the *cloglog* link function. The resulting model includes six variables: *Phys.Health*, *Dental.Health*, *Stress.vs.Sleep*, *Sleep.Medication*, *Race*, and *Gender*. Its AIC is 518.49, which is lower than the full model with the same link function. The test `anova(model3, model4)` yields a p-value of 0.907, providing no evidence that either model is superior.

The model summary shows significance for the same three variables observed in Model 3. The standardized Pearson residuals still reveal issues with model fit for some observations. Prediction on the test set yields results very similar to those from Model 3 (see Figure 4).

```
          Reference
Prediction 0-1 2+
       0-1  13 34
       2+   12 80
> f1
       F1
0.3611111
> MCC
[1] 0.1800329
```

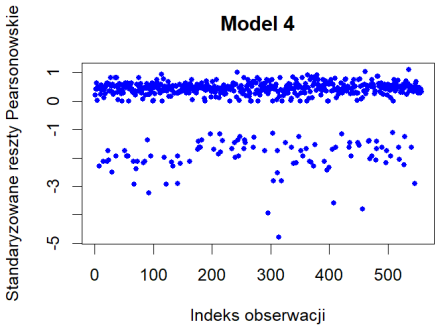

Figure 5: Prediction results and standardized Pearson residuals for Model 4

4 Summary

4.1 Interpretation of Statistical Analysis Results

We built four different regression models for our problem. None of them performed exceptionally well — neither in terms of model fit (high AIC values), nor in terms of classification performance (low F1-scores and MCC values). Nevertheless, we can conclude that Model 3 performed best in terms of classification, achieving a moderate F1-score of 0.41. On the other hand, Model 4 provided the best fit to the data, with its coefficients and related statistics presented in Table 3. Variables that were statistically significant at the 5% level are highlighted in blue.

From the analysis, we observe a linear relationship for the variables *Phys.Health* and *Sleep.Medication*. This may suggest that as physical health improves, the probability of visiting at least two different doctors in a year decreases. Conversely, individuals who use sleep medications more frequently are increasingly likely to visit more doctors. Interestingly, the same effect is observed for people who rate their dental or oral health as excellent. One possible explanation is that good dental health may result from frequent medical visits.

Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	3.27686	60.25594	0.054	0.9566
Phys.Health.L	-0.65320	0.31134	-2.098	0.0359
Phys.Health.Q	-0.05373	0.25559	-0.210	0.8335
Phys.Health.C	0.04331	0.18599	0.233	0.8159
Phys.Health^4	-0.09991	0.11812	-0.846	0.3976
Dental.HealthExcellent	0.67971	0.29529	2.302	0.0213
Dental.HealthFair	-0.19336	0.24799	-0.780	0.4356
Dental.HealthGood	0.18726	0.23769	0.788	0.4308
Dental.HealthPoor	0.07633	0.33512	0.228	0.8198
Dental.HealthVeryGood	0.21799	0.24426	0.892	0.3722
Stress.vs.SleepYes	0.22140	0.13521	1.637	0.1016
Sleep.Medication.L	0.52270	0.25725	2.032	0.0422
Sleep.Medication.Q	0.01554	0.28120	0.055	0.9559
RaceBlack.NoHis	-2.48563	60.25567	-0.041	0.9671
RaceHispanic	-3.13841	60.25586	-0.052	0.9585
RaceOther.NoHis	-3.10106	60.25621	-0.051	0.9590
RaceWhite.NoHis	-2.62248	60.25539	-0.044	0.9653
GenderMale	0.19813	0.11662	1.699	0.0893

Table 3: Coefficients of Model 4

4.2 Conclusions

The research problem posed at the beginning was not fully solved. Logistic regression was not sufficient to effectively handle the classification task. In future attempts, it would be worth exploring alternative classification methods such as decision trees, random forests, or neural networks. Difficulties may also stem from the dataset itself — limited sample size (both in observations and number of features), frequent binary (Yes/No) responses, and narrow ordinal scales for subjective ratings. The research question remains open for further exploration.

References

[1] P. N. Malani, J. Kullgren, E. Solway, *National Poll on Healthy Aging (NPHA), [United States], April 2017*, Inter-university Consortium for Political and Social Research, 2019-05-29. Available at: `https://doi.org/10.3886/ICPSR37305.v1`.

[2] *National Poll on Healthy Aging (NPHA)*, UCI Machine Learning Repository, 2017. [Online]. Available at: `https://archive.ics.uci.edu/dataset/936/national+poll+on+healthy+aging+(npha)`.