Application of Meta-Analysis in Predictive Modeling of Atopic Diseases

Abbreviations: PSY (psychological), ENV (environmental), GLM (general linear model), MSE (mean squared error)

I. Introduction

The importance of research often lies in its application and its communication. Meta-analysis is considered to contain particularly high-quality evidence, and thus it is rational to apply its results. As the original meta-analysis investigates the effects of certain factors on the health of children, the resulting application is directed towards parents.

The creation of an online program eliminates many of the discussed socioeconomic disparities in healthcare, such as driving distance. While it cannot and should not be treated as an alternative to professional care, the program's accessibility and simplicity mean it is available to any parents regardless of location or financial resources.

The additional resources included in the program, such as maternal help hotlines, can provide another benefit for any users without access to quality healthcare. Again, existing disparities in healthcare are linked to later disparities in prevalence of allergic diseases. While these resources are still not an adequate substitute for healthcare, they can help alleviate some of the effects and allow parents to better understand impacts of these exposures.

II. Methodology

The technique of bootstrapping, used to obtain sufficient training data was touched on in the original meta-analysis. As such, this section will focus more on the methods used in building the model rather than obtaining the data.

The general linear model (GLM) was selected for its its flexibility and relatively simple interpretation. Since the bootstrapped data, visualized with a kernel density plot, showed fairly normal distribution, the choice of the Gaussian family in fitting the model was supported. The robustness of the GLM to deviation from normality ensured solid model performance, even if the data did not fully fit the Gaussian distribution.

The inputs for the model were discrete values corresponding to the bootstrapped outputs. Treating each output's label (1-2500) as the exposure level simplified the user-input system, as any entered value could be converted to the 1-2500 scale.

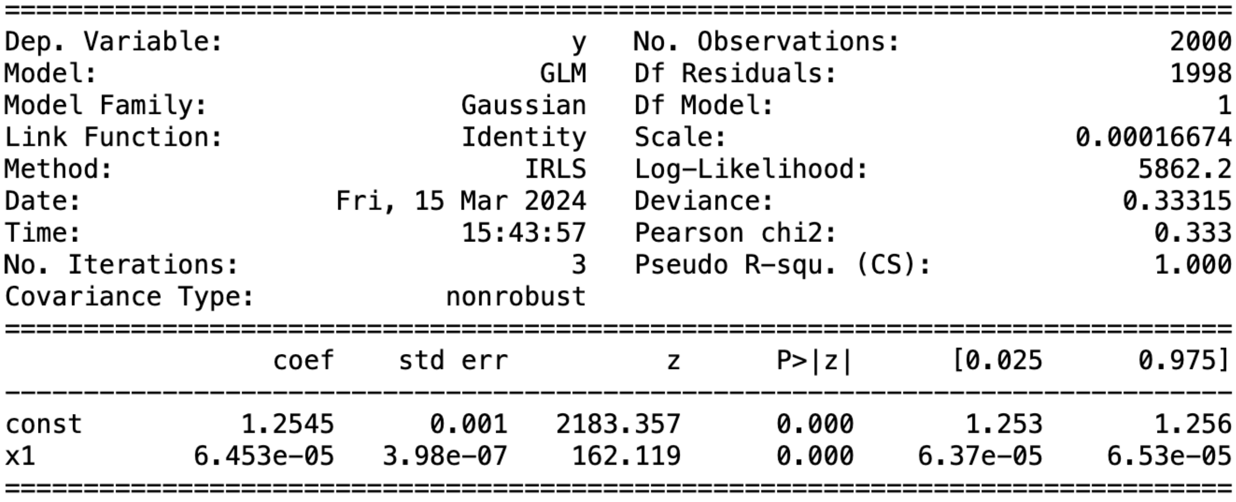
Per standard machine-learning techniques, the data was split into training and testing sets, using an 80:20 ratio. The model's performance was assessed quantitatively through mean squared error (MSE) and K-fold cross-validation and visually though the distribution of standardized deviance residuals. This same process was performed on both the PSY and ENV GLMs.

The nearest-neighbor search for ENV was conducted with the BallTree model. The pollutant concentrations for each neighboring location were obtained and compared to the hazard limits. The number of values exceeding the limit were compared to the number below, and the average value determined the input for the ENV model.

For web accessibility, the program was assembled using the package Streamlit, designed for data-based applications.

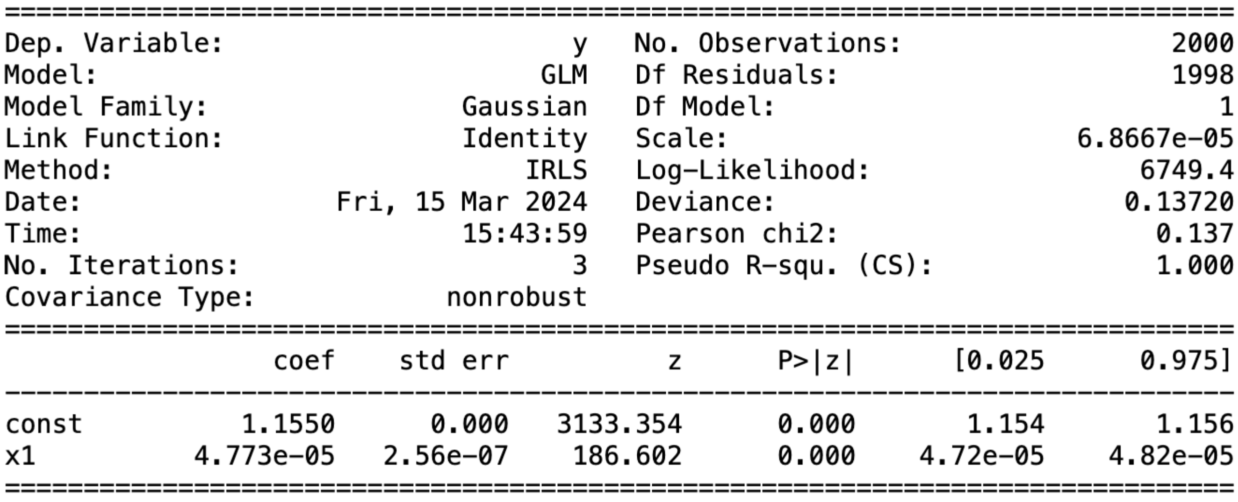
III. Results

The results of the GLMs are displayed in the following tables. The values of MSE for each model were equal to 1.507e-04 (PSY) and 6.274e-05 (ENV).



```
==============================================================================
Dep. Variable:                      y   No. Observations:                 2000
Model:                            GLM   Df Residuals:                     1998
Model Family:                Gaussian   Df Model:                            1
Link Function:               Identity   Scale:                      0.00016674
Method:                          IRLS   Log-Likelihood:                 5862.2
Date:                Fri, 15 Mar 2024   Deviance:                      0.33315
Time:                        15:43:57   Pearson chi2:                    0.333
No. Iterations:                     3   Pseudo R-squ. (CS):              1.000
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.2545      0.001   2183.357      0.000       1.253       1.256
x1          6.453e-05   3.98e-07    162.119      0.000    6.37e-05    6.53e-05
==============================================================================
```

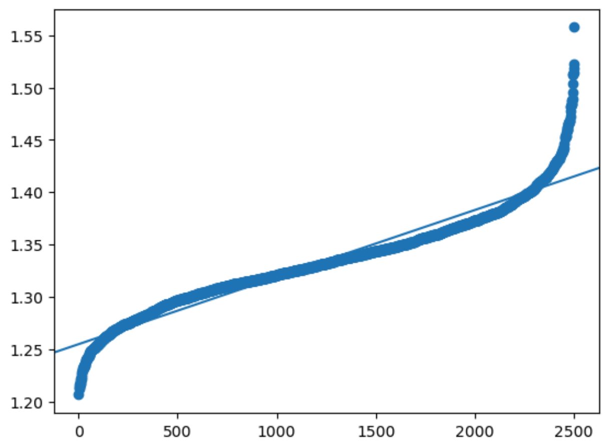
Table 1. Results of the PSY GLM. The goodness-of-fit measures indicate a very accurate fit of the model to the training data.



```
==============================================================================
Dep. Variable:                      y   No. Observations:                 2000
Model:                            GLM   Df Residuals:                     1998
Model Family:                Gaussian   Df Model:                            1
Link Function:               Identity   Scale:                      6.8667e-05
Method:                          IRLS   Log-Likelihood:                 6749.4
Date:                Fri, 15 Mar 2024   Deviance:                      0.13720
Time:                        15:43:59   Pearson chi2:                    0.137
No. Iterations:                     3   Pseudo R-squ. (CS):              1.000
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.1550      0.000   3133.354      0.000       1.154       1.156
x1          4.773e-05   2.56e-07    186.602      0.000    4.72e-05    4.82e-05
==============================================================================
```

Table 2. Results of the ENV GLM. The goodness-of-fit measures indicate a very accurate fit of the model to the training data.

These results can be also visualized using scatter plots. The following figures plot exposure level against their corresponding odds ratios. For both categories, while the majority of the data points fit the line smoothly, the tails deviate from the linear model, indicating that the model may not perform as well given more extreme exposures.



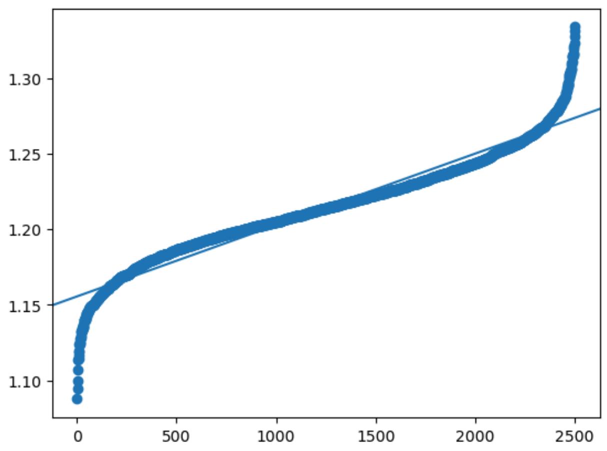
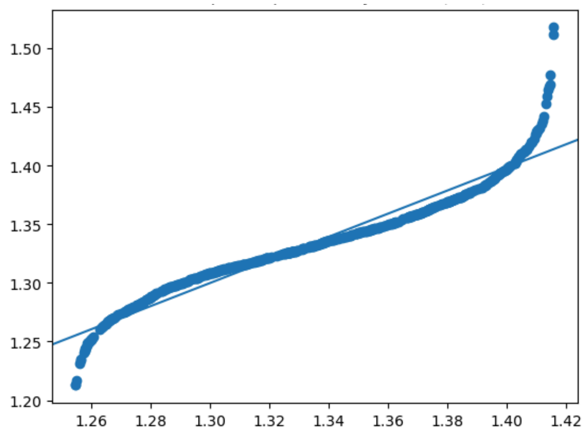


Figure 1.1. Results of the GLM fitted on PSY

Figure 1.2. Results of the GLM fitted on ENV

Accuracy of the model after testing is visualized with these next two plots, which plot the predicted values against the true values. Again, the tails deviate, indicating poorer performance at more extreme values. The middle remains smooth and very close to the plotted line.



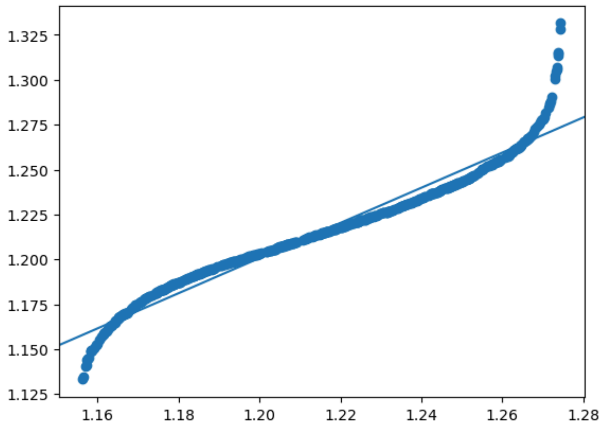


Figure 1.1. Results of the GLM tested on PSY

Figure 1.2. Results of the GLM tested on ENV

K-fold cross-validation, using 5 folds, was performed to ensure the robustness of the model with regards to new data. Train and test MSEs for PSY were equal to 1.633e-04 and 1.638e-04, respectively. These same values for ENV were equal to 6.738e-05 and 6.762e-05.

The histograms below show the distribution of standardized deviance residuals. Both show values concentrated around 0, indicating a very good fit.

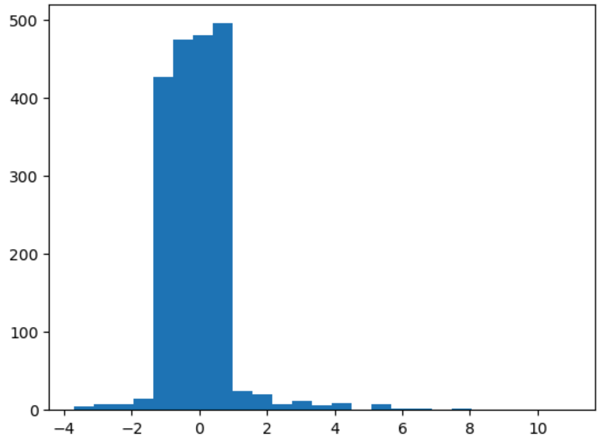


Figure 3.1. Distribution of standardized deviance residuals (PSY)

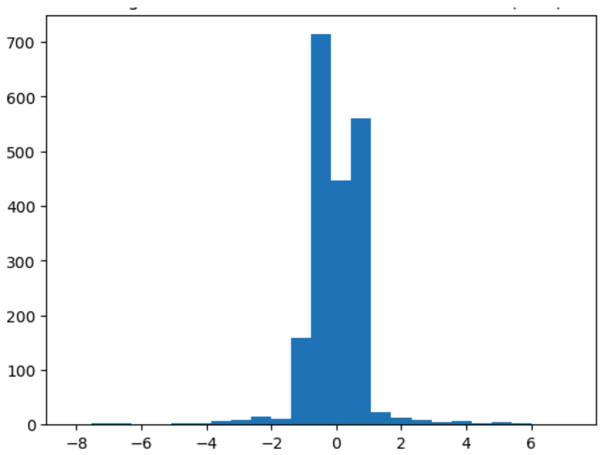
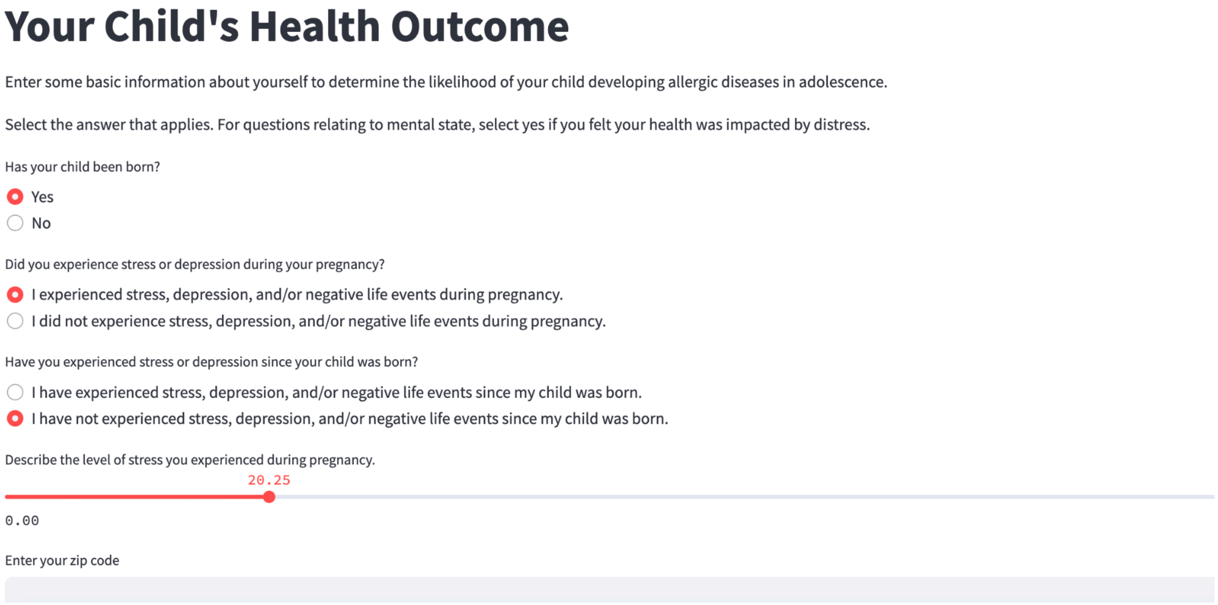


Figure 3.2. Distribution of standardized deviance residuals (ENV)

The interface of the final program is shown below. Its questions are simple and few, meaning the process is almost immediate.



Your Child's Health Outcome

Enter some basic information about yourself to determine the likelihood of your child developing allergic diseases in adolescence.

Select the answer that applies. For questions relating to mental state, select yes if you felt your health was impacted by distress.

Has your child been born?

● Yes
○ No

Did you experience stress or depression during your pregnancy?

● I experienced stress, depression, and/or negative life events during pregnancy.
○ I did not experience stress, depression, and/or negative life events during pregnancy.

Have you experienced stress or depression since your child was born?

○ I have experienced stress, depression, and/or negative life events since my child was born.
● I have not experienced stress, depression, and/or negative life events since my child was born.

Describe the level of stress you experienced during pregnancy.

20.25

0.00

Enter your zip code

Figure 4. Program interface when accessed on the web

IV. Conclusion & Future Work

The program created with the results of meta-analysis is designed to increase knowledge of the impacts of maternal distress and ambient air pollution on atopic diseases in children. Its interface is simple, but the model itself proves to be accurate and robust, showing its potential for future application on a larger scale.

Many changes could be made to this model to improve its accuracy. For example, antibiotic use, pet ownership, race, and socioeconomic status are all factors thought to be correlated with atopy, directly or indirectly. Further meta-analysis could thus benefit the program by providing data to incorporate factors. Additionally, the combined effects of these factors are not addressed in the existing framework as it treats each as a separate variable. While this may be true for the current factors, it may not hold as the program is expanded. Level of perceived distress and socioeconomic status, for instance, are likely strongly correlated. Finally, although meta-regression was conducted for individual pollutants, the program does not account for varying effects for different substances. This may not be plausible with data from the original meta-analysis, but further investigation in the future could provide the program with sufficient data. This specificity would likely increase the program's accuracy as it is unlikely that all pollutants have the same effect, which is the current assumption.

As advised in the program, users should interpret results with cautions. As mentioned above, in its current form, the program accounts for only two of a myriad of influential factors. Consulting with a health professional about potential risk factors for adolescent atopy would likely be beneficial and allow for better planning for possible atopy in children.