TU Dortmund

Introductory Case Studies

Project 3: Logistic Regression and Odds Ratios using the Example of COVID19 Vaccinations

Lecturers:

Prof. Dr. Philipp Doebler

Dr. Crystal Wiedner

Dayasri Ravi

Author: Md Mahmudul Haque

Group members: Mehedi Rahman, Sofiul Azam Sony, and Nazmul Hasan Tanmoy

July 13, 2025

Contents

1 Introduction

The COVID-19 pandemic has heightened global awareness of vaccine hesitancy and the socio-demographic factors influencing public health decisions. Despite the widespread availability of approved vaccines, individual uptake has varied significantly across countries and population groups. This project examines the determinants of COVID-19 vaccination in Germany using microdata from the 2023 European Social Survey (ESS11). The primary goal is to identify which individual characteristics, such as: age, income, education, internet usage, and institutional trust predict the likelihood of having received at least one dose of a COVID-19 vaccine. Italy is included as a comparative case to assess cross-national differences in these predictors. The analysis is based on a binary outcome variable derived from the ESS11 item `vacc19`. Logistic regression is employed to estimate the effects of relevant socio-demographic and attitudinal variables. The modeling approach includes data cleaning, dummy coding of categorical variables, and model reduction based on statistical significance. Parameters are estimated via maximum likelihood using the Newton-Raphson algorithm. Odds ratios are computed to facilitate interpretation. Final results for Germany show that older age, higher household income, and greater trust in the legal system and politicians are positively associated with the likelihood of COVID-19 vaccination. Conversely, being female and living in rural areas significantly decrease the odds of being vaccinated. In Italy, vaccination uptake is most strongly influenced by prior COVID-19 infection and larger household size. Trust in the police increases the odds of vaccination, whereas higher trust in political parties and higher levels of education are negatively associated. Living in less urban areas also corresponds to reduced vaccination likelihood. These results highlight differing national patterns: while institutional trust and socioeconomic status are key in Germany, personal experience and social context play a stronger role in Italy.

The report proceeds as follows: Section 2 outlines the dataset, variables, and data quality considerations. Section 3 details the logistic regression methodology, including assumptions and interpretive tools. Section 4 presents the statistical analysis for Germany and Italy. Section 5 concludes with a summary of key findings and implications.

2 Problem Statement

2.1 Dataset Description

The data used in this study originate from the eleventh wave of the European Social Survey (ESS11), collected in 2023 by ESS ERIC European Social Survey European Research Infrastructure (ESS ERIC) (2025). The ESS is a cross-national survey conducted biennially across numerous European countries using a standardized questionnaire administered through face-to-face interviews. The sampling strategy applies stratification to ensure population-level representativeness within each country. This project draws on a subset of the ESS11 dataset, focusing specifically on respondents from Germany and, for comparative purposes, Italy. The sample includes adult individuals who provided valid responses to the COVID-19 vaccination and demographic questions of the survey. Observations with missing, invalid, or ambiguous values (e.g., refusal to answer, “don’t know”) have been excluded through listwise deletion. The analysis uses the following key variables:

- `vacc19`: Indicates whether the respondent received at least one approved COVID-19 vaccine dose. Values are recoded into a binary outcome variable (`vacc19_binary`) with 1 = vaccinated and 0 = not vaccinated. Responses coded as 7 (refusal), 8 (don’t know), and 9 (no answer) are excluded.
- `respc19a`: Captures self-reported COVID-19 infection status. It is dichotomized into two categories: infected/suspected (values 1 or 2) and not infected (value 3).
- `eisced`: Reflects the respondent’s highest educational attainment, mapped to the ES-ISCED classification scale (1–7). Values 55, 77, 88, and 99 are treated as missing. The value 0 is marked as not possible to harmonise into ES-ISCED
- `hinctnta`: Represents household income decile (1 = lowest, 10 = highest). This is an ordinal variable with some missing values imputed using the median.
- `hhmmb`: Number of individuals regularly residing in the household.
- `netusoft`: Frequency of internet use (1 = never to 5 = every day).
- `gndr`: Gender, recoded as binary (0 = male, 1 = female). Value 9 is considered missing.
- `maritalb`, `domicil`: Categorical variables indicating legal marital status and type of residence. These are dummy coded for regression analysis.

- `agea`: Age of the respondent. Values coded as 999 are treated as missing.
- `trstprl`, `trstlgl`, `trstplc`, `trstplt`, `trstprt`: Institutional trust variables, each measured on a scale from 0 (no trust) to 10 (complete trust).

Variables not relevant to the modeling objective have been omitted for clarity. All included variables are either numeric, ordinal, or have been appropriately transformed into dummy variables for estimation.

2.2 Data Quality and Preprocessing

The ESS data include various codes for missing or inapplicable responses (e.g., 7, 8, 9, 77, 88, 99). These have been removed or recoded depending on the context. Internet usage, income, education, and trust variables were particularly prone to item non-response. After cleaning, the dataset contains only complete cases to ensure model integrity. Categorical variables such as marital status and residential setting are transformed into dummy variables. The resulting data enable estimation using logistic regression without violating key assumptions.

2.3 Research Objectives

The content-related objective of this study is to identify and compare the socio-demographic and attitudinal factors that influence COVID-19 vaccination uptake in Germany and Italy. The central research question is: *Which individual-level characteristics significantly influence the likelihood of COVID-19 vaccination in Germany, and how do these effects compare with those in Italy?* From a statistical perspective, the goal is to model the probability of vaccination using logistic regression, where `vacc19_binary` serves as the dependent variable. The analysis ensures that all model assumptions are met, including the absence of perfect separation among predictors. Parameter estimation is conducted using Maximum Likelihood Estimation via the Newton-Raphson algorithm, and results are interpreted through odds ratios to assess substantive effects. Model selection is carried out through backward elimination to obtain a parsimonious specification. Finally, results from Germany are compared with those from Italy to examine the generalizability and country-specific variation in the identified predictors. This comprehensive approach ensures both methodological rigor and empirical relevance in addressing the central research question.

3 Statistical Methods

This section describes the statistical framework employed to identify predictors of COVID-19 vaccination using logistic regression. The analysis proceeds in accordance with the structure and assumptions of binary response modeling.

3.1 Logistic Regression Model

Given a binary outcome variable $Y_i \in \{0, 1\}$, where $Y_i = 1$ if individual i has received at least one COVID-19 vaccine dose and $Y_i = 0$ otherwise, logistic regression models the probability of vaccination as:

$$\pi_i := P(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})} \tag{1}$$

This can be rewritten using the logit transformation as:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{2}$$

where π_i denotes the probability of vaccination for individual i, β_0 is the intercept term, β_j represents the coefficient associated with predictor variable x_{ij}, and k is the total number of predictors included in the model.

Each coefficient β_j quantifies the change in the log-odds of vaccination for a one-unit increase in predictor x_j, holding other variables constant.

3.2 Parameter Estimation and Software

The model parameters are estimated using Maximum Likelihood Estimation (MLE). Given that the likelihood function for logistic regression has no closed-form solution, the Newton-Raphson algorithm is employed to iteratively approximate the estimates. This algorithm updates parameters via:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \mathbf{H}^{-1}(\boldsymbol{\beta}^{(t)}) \cdot \nabla\ell(\boldsymbol{\beta}^{(t)}) \tag{3}$$

In Equation 3, $\nabla\ell(\boldsymbol{\beta})$ denotes the gradient (i.e., the first derivative) of the log-likelihood function, $\mathbf{H}(\boldsymbol{\beta})$ is the Hessian matrix representing the second derivative, and $\boldsymbol{\beta}^{(t)}$ is the estimate of the parameter vector at iteration t.

All statistical analyses in this study are implemented using Python 3.12. Data manip ulation is conducted using the pandas package McKinney (2010), while numerical operations are supported by NumPy Harris et al. (2020). Inferential modeling relies on statsmodels Seabold and Perktold (2010). Data visualizations are produced with matplotlib Hunter (2007). The complete codebase is publicly accessible on GitHub[1]. This codebase also contains analyses for France and Spain. Only the analyses for Germany and Italy are included in this report.

3.3 Odds Ratio Interpretation

The regression coefficients β_j are typically interpreted using their exponentiated form:

$$\mathrm{OR}_j = \exp(\beta_j) \tag{4}$$

An odds ratio $\mathrm{OR}_j > 1$ indicates that higher values of x_j are associated with increased odds of vaccination, while $\mathrm{OR}_j < 1$ implies a negative association. Odds ratios provide a more intuitive interpretation of the strength and direction of predictor effects than raw log-odds.

3.4 Model Assumptions and Estimability

Logistic regression relies on several key assumptions for valid estimation and interpretation. First, the dependent variable must be binary, taking on only two possible outcomes. Second, all observations are assumed to be independent of each other. Third, the model assumes that the log-odds of the outcome are linearly related to the predictor variables. Additionally, the predictors should exhibit no perfect multicollinearity, meaning that no independent variable can be expressed as a linear combination of others. Lastly, the model must not suffer from perfect separation, where a combination of predictors deterministically classifies the outcome. These assumptions are essential to ensure the convergence and reliability of the maximum likelihood estimates obtained during model fitting.

To address potential perfect separation (as discussed by Silvapulle, 1981), categorical

[1] https://github.com/alcatraz47/logreg$_odds_ratio$

predictors such as marital status, domicile, and highest level of education are dummy coded, and cross-tabulations are inspected prior to model fitting.

3.5 Model Selection

Model selection is performed via backward elimination. Starting with all relevant predictors, the predictor with the highest non-significant p-value (above the 0.05 threshold) is iteratively removed. This continues until only statistically significant predictors remain, yielding a parsimonious and interpretable model.

3.6 Country-Level Implementation

The methodology stated in earlier sections is first applied to data from Germany. Data from Italy is analyzed using the same pipeline to enable cross-national comparisons. All data preprocessing, model fitting, and diagnostic evaluations are performed identically to ensure methodological consistency across both countries.

4 Statistical Analysis

This section presents the empirical findings based on logistic regression analyses. The discussion begins with descriptive insights and assumption checks, followed by multivariable modeling and interpretation. The structure ensures that each stage builds on the previous, facilitating transparent interpretation of the final models.

4.1 Descriptive Analysis and Assumption Checks

Before model estimation, descriptive analyses are conducted to explore the distribution of key variables in the German and Italian subsamples. Missing values coded as 7, 8, 9, 77, 88, and 99 have been removed or imputed, and dummy variables have been created for categorical predictors. Figure 1 presents bivariate patterns in Germany. A clear positive gradient is visible between vaccination rates and both income decile and frequency of internet usage. These patterns suggest the plausibility of linear or monotonic effects in the logistic model of these features. Graphs of these two features ensure that the features do not show non-linearity. Cross-tabulations have been reviewed to assess estimability.

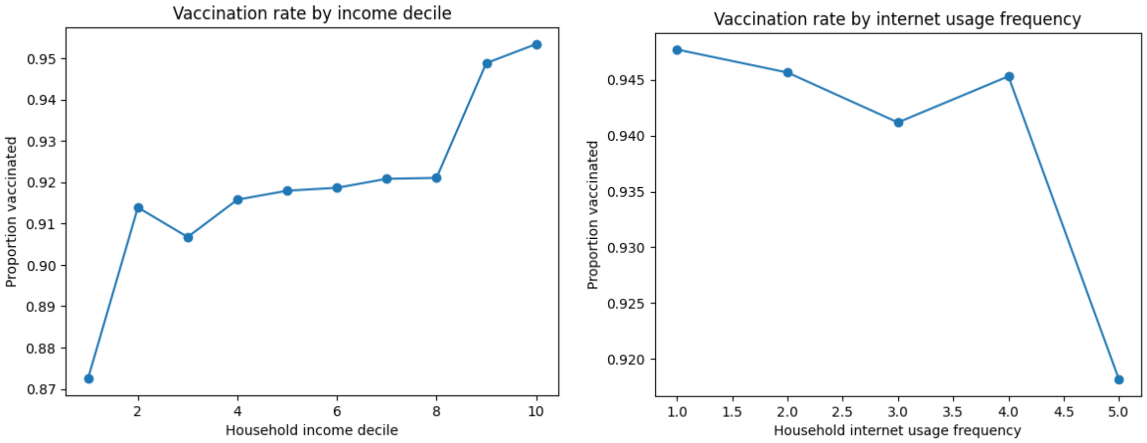
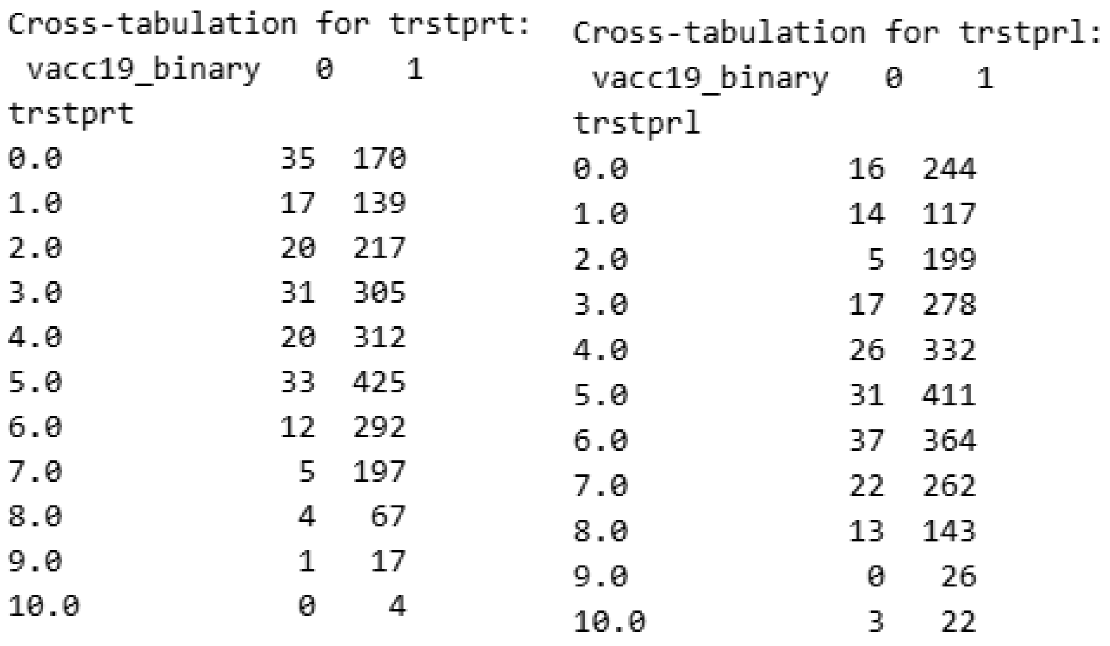


Figure 1: Vaccination rate by income decile and internet usage frequency (Germany)

Figure 2 shows that one piece of evidence of perfect separation has been found for the data from Germany. "Trust in Political Party" perfectly separated the data for one value. For Italy, the feature "Trust in Country's Parliament" separates perfectly. These features are excluded before modelling, as they show perfect separation and it might hamper the model from learning core features. All categorical variables are recoded into dummy variables using the first category as the reference. The cleaned dataset allows for valid logistic regression.

4.2 Model Results: Germany

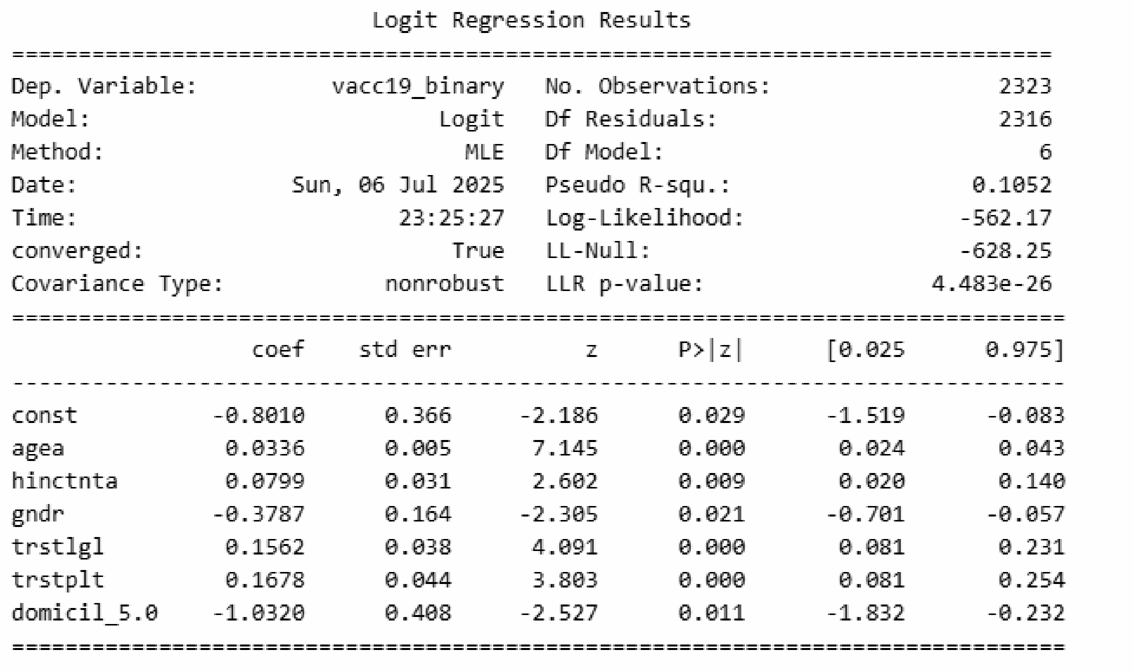
The logistic regression model is fit with the cleaned German data using Newton-Raphson-based Maximum Likelihood Estimation. The dependent variable is binary: vaccinated (1) vs. not vaccinated (0). After backward elimination, only statistically significant predictors ($p < 0.05$) are retained. The final model output is presented in Figure 3, showing parameter estimates, standard errors, and p-values. Odds ratios for the same model are visualized in Table 1. Notable predictors include age, trust in the legal system, trust in the police, household income, and internet use. It shows that institutional trust and socioeconomic status are key predictors of vaccination. Positive odds ratios for age (1.03), income (1.08), and trust in the legal system and politicians (1.17 and 1.18) suggest that higher age, income, and institutional confidence significantly increase the likelihood of being vaccinated. In contrast, being female (OR = 0.68) and living in rural areas (OR = 0.36) are associated with lower vaccination odds.



```
Cross-tabulation for trstprt:
 vacc19_binary   0    1
trstprt
0.0             35  170
1.0             17  139
2.0             20  217
3.0             31  305
4.0             20  312
5.0             33  425
6.0             12  292
7.0              5  197
8.0              4   67
9.0              1   17
10.0             0    4
```

```
Cross-tabulation for trstprl:
 vacc19_binary   0    1
trstprl
0.0             16  244
1.0             14  117
2.0              5  199
3.0             17  278
4.0             26  332
5.0             31  411
6.0             37  364
7.0             22  262
8.0             13  143
9.0              0   26
10.0             3   22
```

Figure 2: Perfect separation on one value in the feature "Trust in Political Party" for Germany (on the left) and "Trust in Country's Parliament" for Italy (on the right)



```
                           Logit Regression Results
==============================================================================
Dep. Variable:          vacc19_binary   No. Observations:                 2323
Model:                          Logit   Df Residuals:                     2316
Method:                           MLE   Df Model:                            6
Date:                Sun, 06 Jul 2025   Pseudo R-squ.:                  0.1052
Time:                        23:25:27   Log-Likelihood:                -562.17
converged:                       True   LL-Null:                       -628.25
Covariance Type:            nonrobust   LLR p-value:                 4.483e-26
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -0.8010      0.366     -2.186      0.029      -1.519      -0.083
agea            0.0336      0.005      7.145      0.000       0.024       0.043
hinctnta        0.0799      0.031      2.602      0.009       0.020       0.140
gndr           -0.3787      0.164     -2.305      0.021      -0.701      -0.057
trstlgl         0.1562      0.038      4.091      0.000       0.081       0.231
trstplt         0.1678      0.044      3.803      0.000       0.081       0.254
domicil_5.0    -1.0320      0.408     -2.527      0.011      -1.832      -0.232
===============================================================================
```

Figure 3: Final logistic regression model results (Germany)

Variable	Odds Ratio
const	0.448888
agea	1.034179
hinctnta	1.083229
gndr	0.684750
trstlgl	1.169046
trstplt	1.182688
domicil_5.0	0.356289

Table 1: Odds Ratio of the Final Model for Germany

4.3 Model Results: Italy

The same modeling pipeline is applied to the Italian sample to ensure comparability. Descriptive plots analogous to Germany's are shown in Figure 4. Again, it is ensured that these two variables do not show any arbitrary non-linearity. Hence, they are allowed to dummy encode.

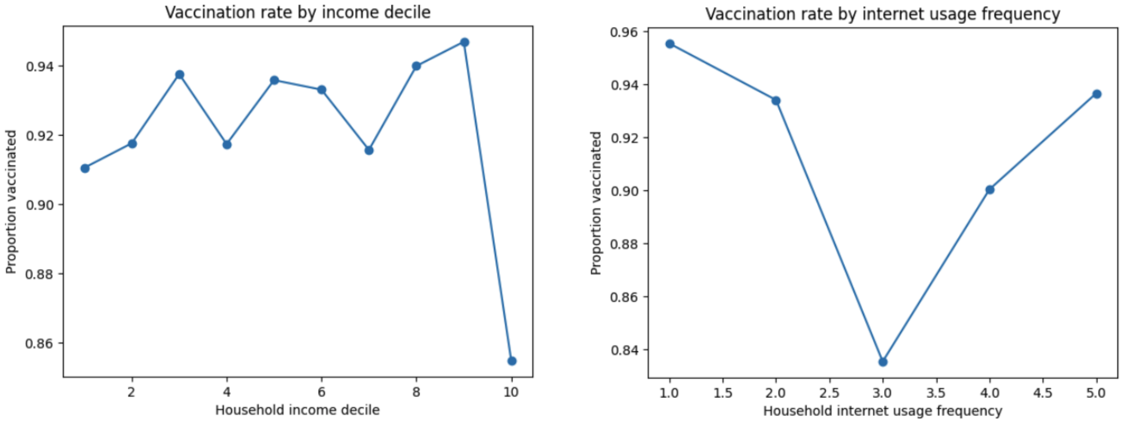
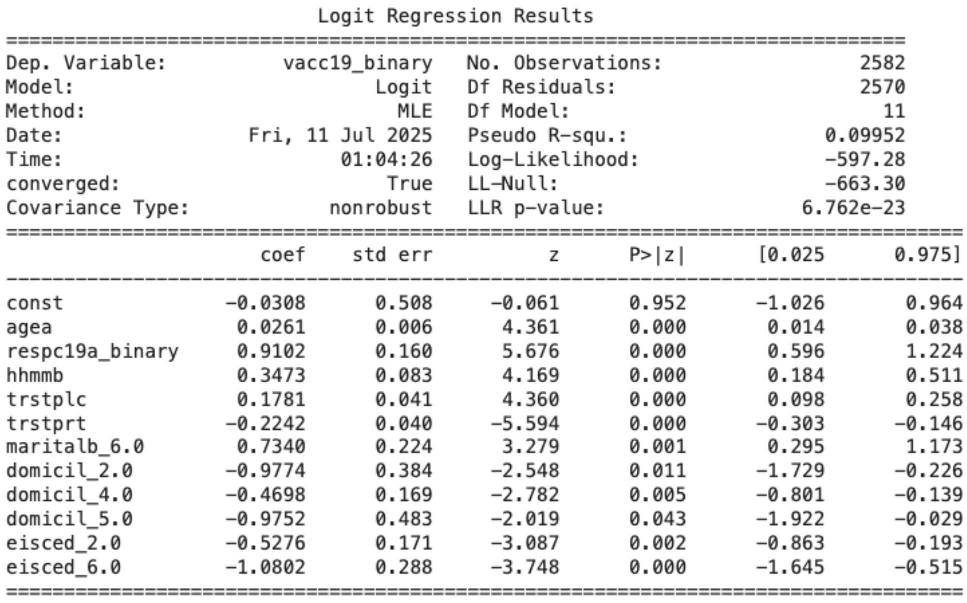


Figure 4: Vaccination rate by income decile and internet usage frequency (Italy)

Final regression results are presented in Figure 5, and the corresponding odds ratios in Table 2. The same method of dummy coding and variable filtering applies here.

In Italy, personal and contextual factors dominate. The odds ratio for prior COVID-19 infection is 2.48, indicating a strong behavioral response to past illness. Larger household size and trust in the police increase vaccination odds, while higher education levels (OR = 0.34) and trust in political parties (OR = 0.80) show negative associations. These contrasts suggest that vaccination behavior in Germany is more aligned with institutional trust, while in Italy it reflects personal risk experience and local context.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:          vacc19_binary   No. Observations:                 2582
Model:                          Logit   Df Residuals:                     2570
Method:                           MLE   Df Model:                           11
Date:                Fri, 11 Jul 2025   Pseudo R-squ.:                 0.09952
Time:                        01:04:26   Log-Likelihood:                -597.28
converged:                       True   LL-Null:                       -663.30
Covariance Type:            nonrobust   LLR p-value:                 6.762e-23
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const             -0.0308      0.508     -0.061      0.952      -1.026       0.964
agea               0.0261      0.006      4.361      0.000       0.014       0.038
respc19a_binary    0.9102      0.160      5.676      0.000       0.596       1.224
hhmmb              0.3473      0.083      4.169      0.000       0.184       0.511
trstplc            0.1781      0.041      4.360      0.000       0.098       0.258
trstprt           -0.2242      0.040     -5.594      0.000      -0.303      -0.146
maritalb_6.0       0.7340      0.224      3.279      0.001       0.295       1.173
domicil_2.0       -0.9774      0.384     -2.548      0.011      -1.729      -0.226
domicil_4.0       -0.4698      0.169     -2.782      0.005      -0.801      -0.139
domicil_5.0       -0.9752      0.483     -2.019      0.043      -1.922      -0.029
eisced_2.0        -0.5276      0.171     -3.087      0.002      -0.863      -0.193
eisced_6.0        -1.0802      0.288     -3.748      0.000      -1.645      -0.515
==================================================================================
```

Figure 5: Final logistic regression model results (Italy)

Variable	**Odds Ratio**
`const`	0.969676
`agea`	1.026411
`respc19a_binary`	2.484756
`hhmmb`	1.415171
`trstplc`	1.194957
`trstprt`	0.799142
`maritalb_6.0`	2.083343
`domicil_2.0`	0.376278
`domicil_4.0`	0.625154
`domicil_5.0`	0.377098
`eisced_2.0`	0.589998
`eisced_6.0`	0.339540

Table 2: Odds Ratio of the Final Model for Italy

4.4 Interpretation and Conclusion

The regression results reveal that the determinants of COVID-19 vaccination differ in emphasis between Germany and Italy. In Germany, vaccination uptake is primarily shaped by structural and attitudinal factors. Higher age and income, as well as greater trust in legal institutions and politicians significantly increase the likelihood of vaccination, indicating a reliance on institutional legitimacy and socioeconomic positioning. The negative association for females and rural residents suggests possible barriers in access or differential attitudes in peripheral groups.

In contrast, the Italian model highlights the role of personal experience and contextual influences. The strong positive effect of prior (suspected) COVID-19 infection and household size implies that perceived vulnerability and social exposure are major motivators. Interestingly, higher education and political trust are negatively associated with vaccine uptake, suggesting a more complex or critical view among highly educated or politically engaged individuals. These findings imply that public health strategies should be tailored: in Germany, strengthening trust in institutions may reinforce vaccination, while in Italy, personalized risk communication and community-level outreach may prove more effective.

5 Summary

This study investigates the factors associated with COVID-19 vaccination uptake using individual-level data from the 2023 European Social Survey (ESS11). The primary research question addresses which socio-demographic and attitudinal characteristics influence the likelihood of receiving at least one approved vaccine dose. Germany is used as the focal point for model construction, with Italy serving as a comparative case to examine cross-national consistency.

The dependent variable `vacc19` is transformed into a binary outcome indicating vaccination status. Responses reflecting refusal, uncertainty, or non-response are excluded. Predictor variables which do not perfectly separate include age, gender, household income, education, household size, internet usage, trust in different political and legal institutions, and prior COVID-19 infection status, the latter dichotomized for analytical clarity.

Key results for Germany show that higher age, income, trust in legal institutions, and trust in politicians are associated with significantly increased odds of vaccination. In contrast, being female and living in rural areas (as indicated by `domicil_5.0`) are associated with lower odds of vaccination. For Italy, the most influential predictor is a history of COVID-19 infection, with an odds ratio exceeding 2.4. Household size and trust in the police also increase vaccination likelihood, while higher education levels and trust in political parties show significant negative effects.

These results suggest that in Germany, structural and institutional confidence play a primary role in vaccine uptake, while in Italy, individual experience and household context appear more influential. The contrast highlights how country-specific trust patterns and behavioral dynamics shape health-related decisions.

Interpretation of these findings should consider potential limitations. The exclusion of cases with missing data may lead to selection bias. The use of dummy variables increases model complexity and may obscure interpretability when categories are unevenly distributed. Moreover, the cross-sectional nature of the ESS data limits causal inference.

Future work could extend this analysis by including additional countries from ESS11 or earlier ESS waves to assess temporal and regional trends. Multilevel models could account for contextual effects, and interaction terms may uncover moderating relationships. Furthermore, exploring how specific types of trust interact with digital engagement and health communication channels could offer actionable insights for policy and outreach strategies.

Bibliography

European Social Survey European Research Infrastructure (ESS ERIC). ESS11 - Integrated File, Edition 3.0, 2025. URL https://doi.org/10.21338/ess11e03$_0$.*Dataset*.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56. SciPy, 2010.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, pages 92–96. SciPy, 2010.

Mervyn J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(3):310–313, 1981.