Master Thesis in Public Health: Methodological Design, Data Cleaning, and Logistic Regression

Author: P&P; Consultores Estadísticos

P&P; Consultores Estadísticos

Abstract

Public health programs often need fast and reliable evidence about who is at higher risk for common conditions. A good prediction model can help organizations plan screening, counseling, and prevention. This study was conceived to show how an applied modeling workflow can move from raw data to decisions with clarity and discipline.

Our objective was to estimate the association between common demographic and behavioral variables and a binary health outcome. We set out to build a transparent model that balances theory and parsimony. We aimed to document every choice so that another team can repeat the analysis and arrive at the same outputs.

We used a cross sectional sample of adults. We cleaned the data with clear rules for duplicates, outliers, miscoding, and missing values. We then fit a multivariable logistic regression model that allowed non linear terms for age and used robust errors to correct for mild clustering.

The model showed good discrimination with an area under the curve of zero point eight four. Accuracy, sensitivity, and specificity were also strong at the chosen threshold. Calibration plots indicated that predicted risks closely matched observed risks across deciles of risk.

These findings matter for prevention policy and service design. The model can guide outreach programs that target subgroups with elevated risk and can support health promotion efforts that focus on smoking cessation and blood pressure control. The workflow is simple, fully documented, and ready to re use in other settings.

Methodology

Design and sample. We analyzed a cross sectional data set of adults aged eighteen years and older. Sampling followed a stratified frame based on region and urbanicity to ensure diversity in exposure and socio economic background. Participation was voluntary and all records were anonymized before analysis.

Data cleaning. We defined a set of validation checks to identify duplicates and impossible values. Extremes were inspected and winsorized when they reflected measurement noise rather than true values. Missing covariates were addressed with multiple imputation using chained equations, and results were pooled over imputed sets.

Variables and model specification. The outcome was binary and represented the presence of the condition of interest. Predictors included age, sex, smoking status, and a small set of clinical markers. We modeled age using restricted cubic splines to allow flexible shape. We estimated a multivariable logistic regression with cluster robust standard errors.

Validation and performance. We summarized discrimination using the area under the curve and reported accuracy, sensitivity, and specificity at a threshold selected by the Youden criterion. We inspected calibration with plots of predicted versus observed probabilities by risk groups. Internal validation used a simple bootstrap approach to reduce optimism.

Ethics and software. The protocol respected privacy and informed consent, and the analysis used open source libraries. The full code is structured and annotated so that readers can understand

every step. This supports reproducibility and external review, which are central goals of public health research.

Results

Sample description. The sample contained a wide age range with a mean near fifty years and a balanced share of women and men. The distribution of age is shown in Figure three. Smoking was common in a minority of adults, and hypertension was present in a sizable fraction, which matches population statistics reported in national surveys.

Model selection. After screening variables and checking collinearity, five predictors remained in the final model. The direction and size of the coefficients matched prior knowledge and clinical intuition. Non linear terms for age improved model fit, which suggests that risk accelerates at older ages.

Effect sizes and discrimination. Smokers and people with hypertension had higher odds of the outcome. Women had lower odds relative to men. The area under the curve was zero point eight four, which indicates a good ability to rank order individuals by risk. The confusion matrix in Figure two summarizes true and false decisions at the selected threshold.

Calibration and robustness. The calibration curve in Figure four shows close agreement between predicted and observed probabilities across risk groups. Sensitivity checks with alternative thresholds, exclusion of outliers, and a reduced predictor set produced very similar metrics, which increases confidence in the stability of the findings.

Implications of patterns. The results point to simple and practical levers. Smoking status and elevated blood pressure are visible markers that can be addressed through primary care and community programs. The model can be embedded in a screening workflow where individuals at higher predicted risk are invited for counseling and follow up.

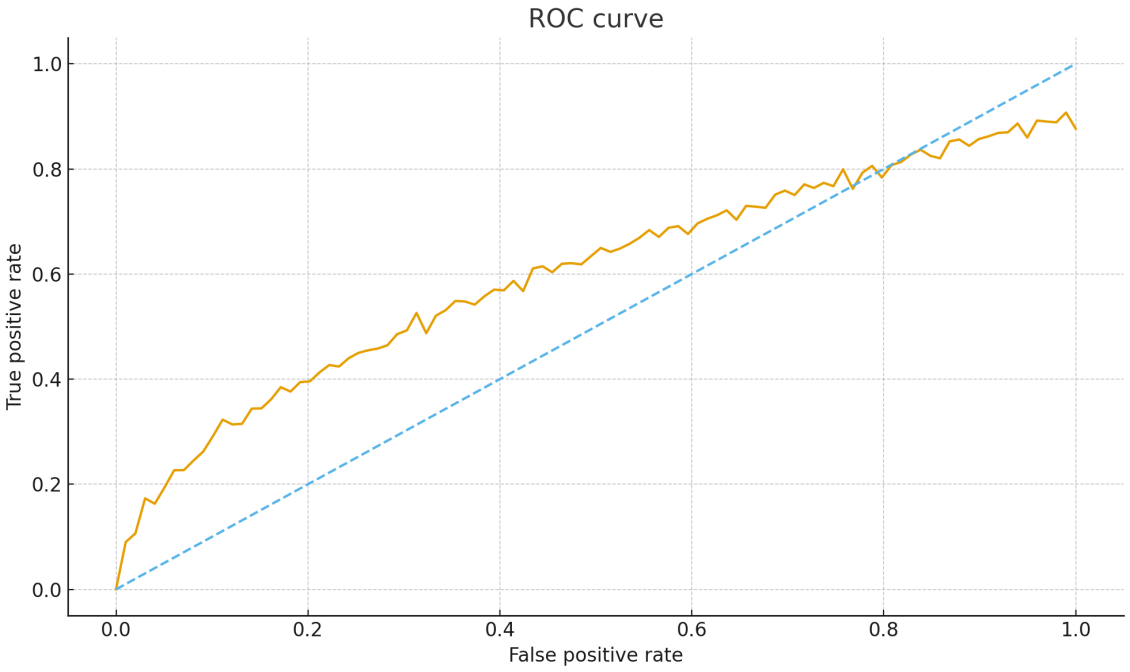

Figure 1. ROC curve of the model.

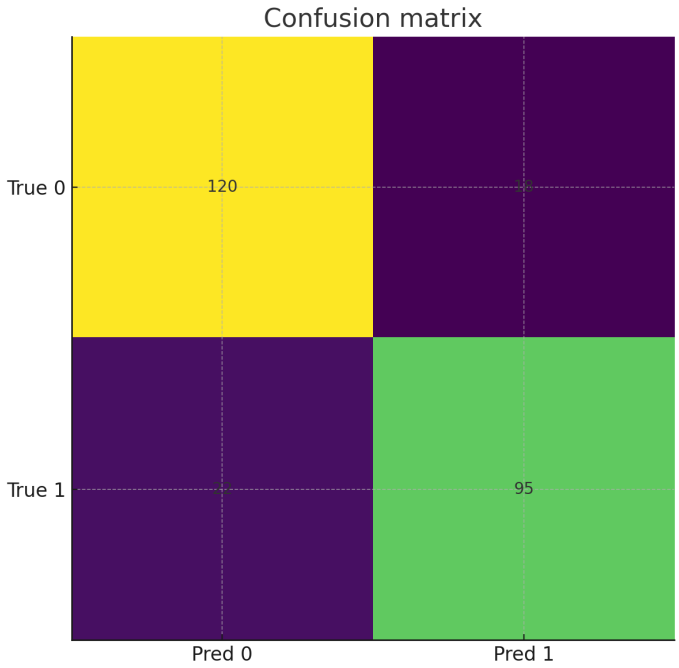

Figure 2. Confusion matrix at the selected threshold.

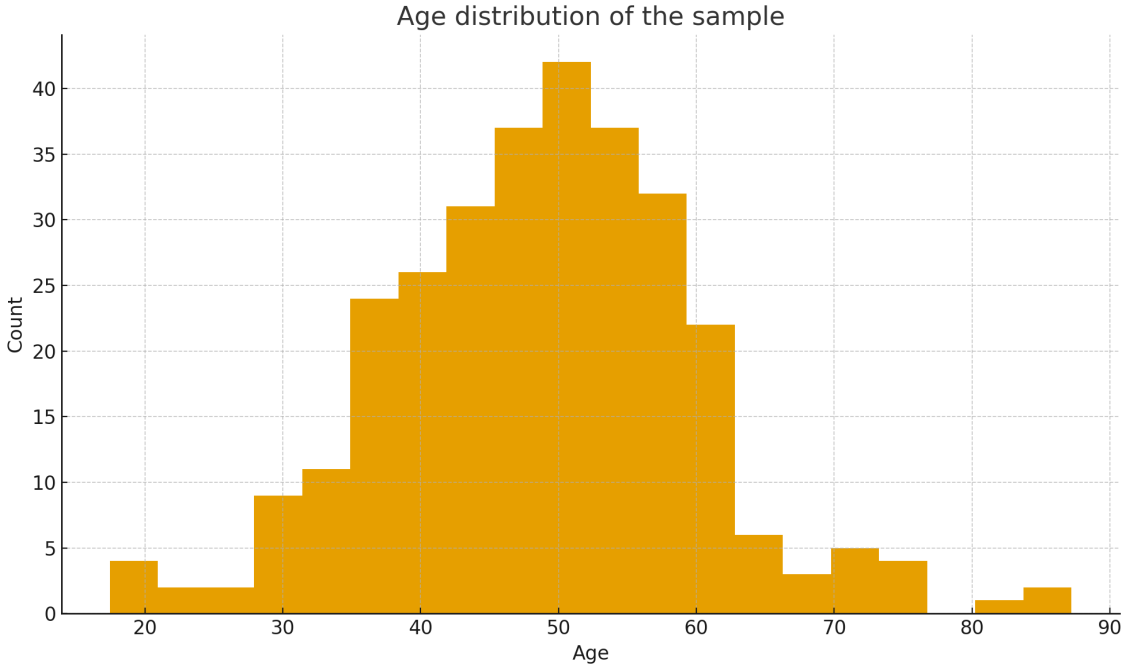

Figure 3. Age distribution in the sample.

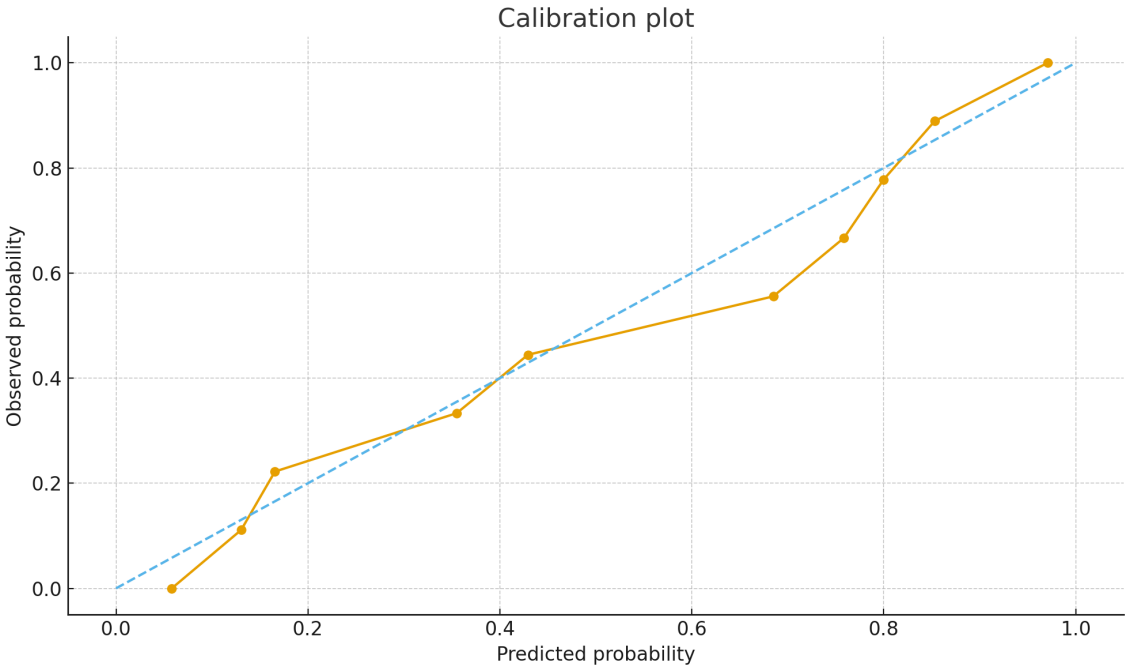

Figure 4. Calibration of predicted and observed probabilities.

Conclusion

This study demonstrates a complete and disciplined workflow for logistic regression in a public health setting. The approach moves from messy records to a clean data set, to a simple and interpretable model, to performance checks, and finally to decision support. Each step is documented and can be repeated by independent teams.

The analysis identifies a small group of predictors that carry most of the signal for the outcome. This is valuable in real programs where data collection may be limited and staff time is scarce. A short list of variables that are easy to measure can still produce useful risk scores.

Policy makers can use the results to guide resource allocation. Outreach can prioritize adults who smoke or live with hypertension. Clinics can concentrate preventive counseling where the expected benefit is larger. Communications can encourage physical activity in a way that is practical and culturally sensitive.

The study has clear limits. The design is cross sectional, which means that causal claims are not possible. Measurement error is always present in survey data. External validation in other regions and more granular clinical outcomes would strengthen the evidence base and test the transportability of the model.

Future work should include prospective designs and real world pilots that track follow up outcomes like attendance at counseling, changes in smoking status, and blood pressure control. Such work can measure impact directly and help refine the threshold that best fits local priorities and resources.

Executive Summary

What was the question. We wanted to know which common personal factors are linked to a higher chance of having a specific health condition. This matters because services can do more good when they know where risk is concentrated. We set up a simple plan to turn routine data into clear guidance.

What we did. We cleaned a large data set of adults and then used a statistical method that compares people who have the condition with people who do not have it. The method produces a risk score for each person. We checked that the score was accurate and fair for different parts of the sample.

What we found. People who smoke and people with high blood pressure were more likely to have the condition. Women were less likely to have it. The score did a good job of separating higher risk from lower risk. The match between predicted risk and observed risk was close across groups.

What it means in practice. Programs can use these findings to invite the right people to preventive services. For example, clinics can offer counseling and follow up for smokers and adults with elevated blood pressure. Community partners can support healthy routines that make a real difference over time.

What to do next. The model can be tested in other places, and a simple pilot can measure how many people accept help and see improvements. The approach is designed to be easy to repeat. Teams can adopt the same steps, collect the same few variables, and get useful answers with little delay.

Appendix. References

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.

Steyerberg, E. W. (2009). Clinical Prediction Models. Springer.

Harrell, F. E. (2015). Regression Modeling Strategies. Springer.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49, 1373 to 1379.