Duke GLOBAL HEALTH INSTITUTE | Research Design & Analysis Core

Exploratory Data Analysis: Application of EDA to the study of causal relationships

Part of a series addressing common issues in statistical and epidemiological design and analysis

Table of Contents

Why This Core Guide Exists 2

Objectives 2

Introduction 2

Types of Exploratory Data Analysis **3**

Scenarios Relevant for Exploratory Data Analyses **4**

Significance Testing and Exploratory Data Analysis **6**

Example Dataset and Statistical Analyses *6*

Methods to Explore **7**

Method 1. Automated backward stepwise regression using a p-value threshold of 0.1 for variable selection. *8*

Statistical and scientific implications of Method 1. ..... **9**

Method 2. Automated backward stepwise regression using a p-value threshold of 0.05 for variable selection *11*

Method 3. Manual backward stepwise regression using mean squared error (MSE) for variable selection *12*

Method 4. Multiple regression models fit separately based on use of a Directed Acyclic Graph (DAG). *16*

Method 5. Random Forest procedure for variable selection. *25*

Comparison and summary of the 5 EDA Methods *29*

Recommendations *30*

Appendices *31*

Appendix 1. DAGitty code for Figure 1. ..... **31**

Appendix 2. LOESS smooth line for the bivariate association between health status and participant age. **33**

Why This Core Guide Exists

Exploratory Data Analysis (EDA) can be hard. Exploring anything, by nature, can be a meandering process and hard to formalize. While EDA is integral to the scientific process, it can be challenging to plan and execute. Given that our goals of EDA are hypothesis *generating* and not hypothesis testing, it can also often be unclear how statistical analyses (like p-values, confidence intervals, regressions, etc.), which are tools designed for *confirmatory* inference, not exploration, can be used to guide inference without inadvertently implying formal conclusions.

Objectives

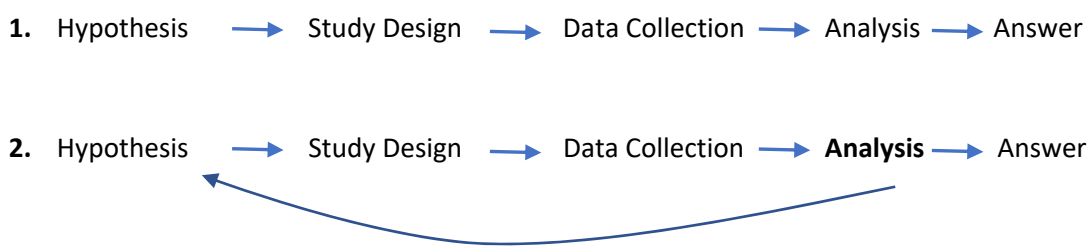
To introduce the concept of exploratory data analysis, to differentiate it from confirmatory data analysis, and to apply methods to EDA, especially related to the statistical estimation of bivariate relationships i.e. looking at how two variables relate to each other (e.g., age vs. income, smoking vs. lung function). A comprehensive overview of EDA is beyond the scope of this Core Guide. **The primary objective of this Core Guide is to focus on the numerical estimation of multivariate relationships.**

Introduction

"Finding the question is often more important than finding the answer."(1)

- John Tukey, 1980

In his 1977 seminal textbook, the statistician John Tukey describes the process of **Exploratory Data Analysis (EDA)** as "looking at data to see what it seems to say."(2) Exploratory Data Analysis can be thought of as a philosophy, or way of thinking, and not simply as a set of methods or techniques. It requires the investigator to acknowledge a feedback loop, or iterative process, to intellectual inquiry. It is through this iterative process that data are explored and hypotheses are generated. Consider two versions of scientific inquiry:



One might consider version #1 above as a **confirmatory** paradigm(1). We state a formal hypothesis, design a study to test this hypothesis empirically, and then assess whether data collected from this study support that single *a priori* hypothesis. But this paradigm presupposes that a clear formal hypothesis already exists, and our interest lies in evaluating whether the data collected for this purpose

are consistent with this single hypothesis. Occasionally, this may be the case, but more often we have to generate hypotheses, which in the quantitative social or biomedical sciences usually occur either through theory or the **exploration of existing (secondary) data**. Version #2 above is likely a more accurate representation of much of the quantitative scientific inquiry we perform in Global Health. We could also certainly imagine that hypothesis generation could come from any of the other nodes along this pathway, not only 'Analysis'. The design of a study, the act of collecting and cleaning the data, and the results, or inference itself, can inform future research questions.

Types of Exploratory Data Analysis

1. Univariate analysis (examines distributions)
    a. Numerical
    b. Visual (graphic)
2. Multivariate analysis (examines relationships)
    a. Numerical
    b. Visual (graphic)

Exploratory Data Analysis can be categorized into either the **examination of distributions** (univariate analysis) or the **examination of relationships** (multivariate analysis).

In **univariate analysis** we examine a single variable. This could be something as simple as understanding the distribution of systolic blood pressure in a population. We could summarize this by estimating the mean and standard deviation of SBP or the median and its interquartile range.

In **multivariate[1] analysis**, we are interested in understanding *relationships* between two or more variables. Bivariate analysis is a version of multivariate analysis where we limit the relationship to just two variables, and we could summarize this relationship through the calculation of a correlation coefficient, coefficients from a regression model, or any number of other statistical techniques.

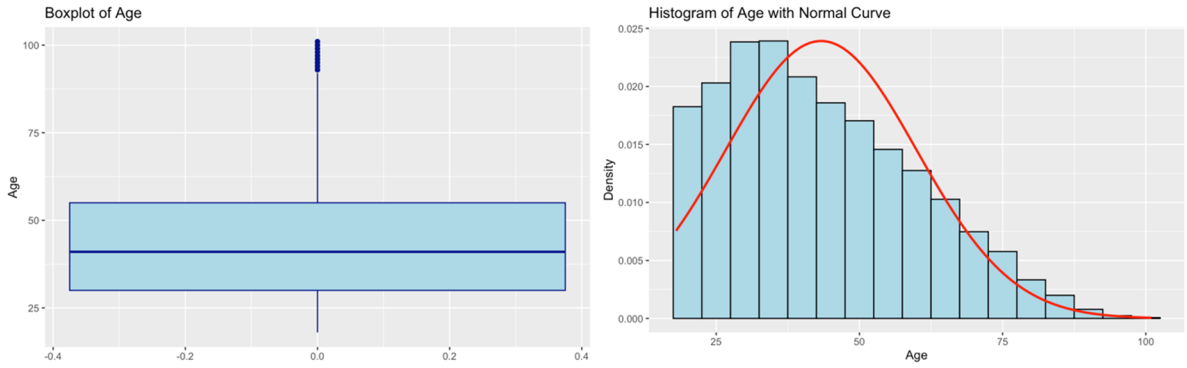
For both univariate and multivariate analyses, we can further classify EDA techniques as being either **visual** (graphical) or **numerical**. As an example, Table 1 (numerical) and Figure 1 (visual) depict results from performing univariate EDA on the variable age, from the 2013 Brazil National Health Survey of Adults. The results provide an overview of the distribution of the age variable.

Table 1.

Variable	N	Mean	StdDev	P25	P50	P75	Skew/Kurt
Age	60,202	43.3	16.7	30	41	55	0.6/2.6

[1] In EDA literature, it is common to refer to the analysis of two or more variables as "multivariate." This is distinct from the regression literature, which often refers to the regression of one dependent variable on multiple independent variables as "multivariable." When one regresses two or more variables (e.g., repeated outcome measures on the same person over time) on one or more independent variables, we call this "multivariate." Confusing, I know…

Figure 1.



The field of EDA is large and a comprehensive overview of EDA is beyond the scope of this Core Guide. Instead, **the remainder of this Core Guide will focus on the numerical estimation of multivariate relationships**. We suggest any of the references cited in this Core Guide for further reading.

Scenarios Relevant for Exploratory Data Analyses

Consider the following scenarios in which you might find yourself:

1. You have access to cross-sectional data collected at baseline of a randomized controlled trial (or cohort study – the thinking is the same). You want to analyze the data to look for correlates[2] (independent variables) of outcome Y (dependent variable). Intervention variable, I, was randomized, but the independent variables you are interested in for the present study were not.
2. You have access to (or maybe you collected the data yourself) from a cross-sectional study designed to estimate the prevalence of disease Y in a population. Maybe the sampling scheme was random, or maybe it was convenient. Despite the primary intended purpose of estimating the prevalence of Y, the dataset also contains several other variables. You want to use these cross-sectional data to estimate the correlation between these variables with disease Y.
3. You have access to data from an incident case-control study originally designed to estimate the causal[3] association of independent variable A_1 and outcome Y. Now you want to use the dataset to estimate the causal relationship between different independent variables A_2 and A_3 and outcome Y.

In each of these 3 scenarios, you are interested in **quantifying the association between two variables**. But you realize two things: A) that in each scenario the independent variable was not randomized to the study participant; and B) the data are being used for purposes that were not their original intent (i.e.,

[2] The term 'correlates' is more appropriate than 'risk factor' in the case of cross-sectional data since prevalence is usually a poor measure of risk.

[3] The word *causal* is used throughout this document to acknowledge the objective of the work, which is to determine whether an independent variable causes a dependent variable. Use of the word makes no opinion of our ability to achieve this in practice.

secondary data analysis). These 3 scenarios could all represent examples of EDA. You may also imagine that **the degree to which you might have theorized these associations beforehand is unclear or could vary substantially**. Consider further, 4 scenarios within the context of estimating the causal association of independent variable A on outcome Y for any of the above scenarios.

1. You have a large body of previous literature and/or strong social or biological plausibility to support your hypothesis that variable A is causally associated with outcome Y. Furthermore, you have a large body of literature and/or strong social or biological plausibility for the relationship between variable A and any other variables under consideration, such as potential confounders or moderators of the association.
2. You have a large body of previous literature and/or strong social or biological plausibility to support your hypothesis that variable A is causally associated with outcome Y. However, you **do not** have a large body of literature and/or strong social or biological plausibility for the relationship between variable A and any other variables under consideration, such as potential confounders or moderators of the association.
3. You have no (or a weak) body of previous literature and/or strong social or biological plausibility to support your hypothesis that variable A is causally associated with outcome Y.
4. You are throwing darts at a board to see what sticks – that is to say, you are just blindly estimating correlations or associations between multiple independent variables and a single or multiple outcomes in the dataset.

We could argue that **only #1 above is a confirmatory data analysis**. If you have a strong theory to guide your statistical analysis plan, then you might only perform a single analysis, such as a single regression model. This model would be parameterized in such a way to most validly and precisely estimate the causal effect of A on Y, and you would report results of this single model without using the results to iterate and change this model (i.e., tweak your hypothesis). It is scenario #1 that many scientists believe is a proper justification for planning all analyses **prior to execution,** and in this scenario, it is appropriate to report a p-value and interpret this p-value at some level of alpha, such as 0.05(3). But in the case of scenarios #2-4, we would almost certainly plan to do more than a single analysis and allow the results of each analysis to influence the others (i.e., allow our hypothesis to change based on our data/results). For example, we might include multiple independent variables in the same regression model and iteratively re-focus our attention on each one as the 'exposure'. Or maybe we change the variables in the model that we believe are confounders of the primary exposure variable. As a result, we would not be able to plan all analyses beforehand, and we would need to be careful in interpreting our results. We will examine why this is the case and how best to do this in the worked example below.

One might argue that even within a randomized controlled trial design to estimate the causal effect of a treatment on an outcome, we are operating within the EDA paradigm for all except the primary hypothesis (the one the study was specifically designed and powered to test.) Secondary outcomes of this trial or subgroup analyses (e.g., to test for the presence of moderation or mediation of treatment effects, or post-randomization exposures) would often be considered 'exploratory' in this framework.

Significance Testing and Exploratory Data Analysis

If Exploratory Data Analysis is hypothesis-generating and not hypothesis confirming, then one common question is, *should I be performing hypothesis testing (e.g., statistical significance testing) as part of my EDA*? There is some debate in the literature on this(4), but most people agree that the answer is usually 'No'. Many believe that it is often inappropriate, or simply not useful, to report p-values along with your EDA results. And it is certainly not useful to evaluate any formal statistical test (T-test, F-test, Chi-squared test, etc.) at a strict significance level of 0.05. The reason behind this thinking is that in EDA, we are *generating* hypotheses, not testing them. Statistically, this means that **in EDA we are purposively putting more weight in uncovering associations that may turn out to be false** (i.e., higher Type I error) than we are in failing to uncover an association because we have set the burden of proof too high (i.e., a low alpha level, such as 0.05 and high Type II error). We believe that during the EDA process, we are okay to generate hypotheses that turn out to be invalid. That is the scientific method of paradigm #2.

Because EDA often involves asking many hypotheses of the data, if we run statistical tests on these hypotheses, the overall Type I error (false positive) rate becomes inflated. As a result, evaluating a statistical test at an alpha level of 0.05 in an EDA is inappropriate and will almost certainly lead to greater than 5% Type 1 error. The epidemiologist Ken Rothman argues that the idea of hypothesis testing in this framework is flawed since null hypothesis significance testing assumes as a premise that randomness is the most likely explanation for phenomena(5) – something that scientists certainly do not believe. A detailed discussion of Type I error and multiple testing is beyond the scope of this Core Guide. Please see the Core Guide, "Multiple Testing, Part 1", or Dr. Eric Green's textbook chapter on Statistical Inference for further discussion.

Exploratory data analysis is not less rigorous or useful than confirmatory data analysis. They are both valid and complementary forms of analysis that are critical to scientific inquiry. In fact, EDA is often much harder to perform and more strongly relies on the expertise of the investigator to execute well. To again quote John Tukey,

"The picture examining eye is the best finder we have of the wholly unanticipated." (1)

The remainder of this Core Guide will introduce a motivating example and illustrate different analytic techniques and methods for reporting results within the EDA paradigm. Statistical code for both Stata and R software can be found in-line with the text.

Example Dataset and Statistical Analyses

Our objective is **to estimate the causal association of multivariate relationships.** Colloquially, we may state that we are interested in exploring or estimating "risk factors" or "correlates" of an outcome, such as a disease or health state. But there are many ways of phrasing the same or similar intent, and language tends to differ by author, journal, and discipline. We could attempt to do this from a well-controlled longitudinal study, a cross-sectional survey with convenience sampling, electronic health record data, or any number of other data sets with varying data-generating processes. Each design

presents its own analytic and inferential challenges; however, the methods outlined here are broadly applicable across these study types.

Data for the examples presented here come from the **2013 National Health Survey of Brazil - DF**. These data most closely resemble scenario #2 mentioned above – a secondary 'risk factor analysis' of a cross-sectional survey. The DF surveyed over 200,000 Brazilian households to collect information on access and utilization of health services in the country. It also sought information on population health conditions and non-communicable diseases. The sampling scheme was a multi-stage cluster random sample and produced cross-sectional data on a wide variety of individual and household-level characteristics of the sample population. The data are publicly available and can be found here. Revised and partially cleaned analytic datasets used for the purposes of this Core Guide can be found on our GitHub page. The full dataset includes approximately 60,000 records.

Our EDA will focus on the following variables in the dataset:

Variable name	Description	Distribution [levels]
federation_unit	Federal geographic unit of Brazil	Categorical [27]
female	Participant gender	Categorical [1=female; 0=male]
age	Participant age	Continuous [18yrs - 101yrs]
race	Participant race	Categorical [1=White; 2=Black; 3=Yellow; 4=Brown; 5=Indigenous; 9=Unknown]
insurance	Participant insurance status	Categorical [1=yes; 2=no]
smoke_current	Current smoking status	Categorical [1=Yes, daily; 2=Yes, less than daily; 3=Not currently smoking]
marital_status	Participant marital status	Categorical [1=Married; 2=Separated; 3=Marital status (a); 4=Widow; 5=Single]
healthy	Health status	Categorical [1=very good, good, regular; 0 = bad, very bad]]
veryhealthy	Health status, recoded	Categorical [1=very good; 0 = good, regular, bad, very bad]

Methods to Explore

In the examples below, we will examine a few common analytical methods used in EDA. Specifically, we will perform:

1. Automated backward stepwise regression using a p-value threshold of **0.1** for variable selection.
2. Automated backward stepwise regression using a p-value threshold of **0.05** for variable selection.
3. Manual backward stepwise regression using mean squared error (MSE) for variable selection.
4. Theory-driven variable selection through use of a directed acyclic graph (DAG).
5. Conditional Random Forest procedure for variable selection.

We will briefly describe each method. We will then show how to implement each method and then discuss its strengths and weaknesses. We will present software code and results of methods in-line with the text.

Disclaimer: All analyses and results of data presented here are done solely for illustrative and teaching purposes. Methods and results have been simplified to focus on key issues in EDA and should not be interpreted as valid estimates of causal effects.

Method 1. Automated backward stepwise regression using a p-value threshold of 0.1 for variable selection.

In an automated backward [4]stepwise regression, we identify all candidate variables and include them all in the model ('full model'). We then give the software a p-value criterion threshold and then run the procedure. The software runs the full model and removes any parameters where the p-value is greater than this threshold and then re-runs the model without those parameters and uses the same criterion again. The procedure stops when all parameters in the last model have a p-value at or below the p-value threshold. In this example we use 0.1 as our p-value threshold. Generally, we would not want to set this alpha level too low because we are in 'exploratory' mode and we privilege lower Type II error compared to Type 1 error. However, this decision is also often based on the size of the dataset since p-values are a function of both the size of the parameter, as well as the size of the dataset. In this dataset, we have over 60,000 observations, so we are likely to have relatively small p-values. With a smaller dataset, we might choose a p-value threshold of 0.2, but the choice is arbitrary and up to the investigator.

Stata code:

```
stepwise, pr(.1): glm healthy i.smoke_current i.insurance i.race i.marital_status age_center
female, fam(Poisson) link(log) vce(robust) eform
```

R code:

R does not currently have a function that exactly replicates Stata's results for backward stepwise selection. Instead, it uses model AIC for selecting and excluding variables, so the R code below will not replicate Table 1, which was generated using Stata. This shows the importance of fully understanding the statistical methodology and theory behind any statistical software function you use.

```
factor_vars <- c("smoke_current", "insurance", "race", "marital_status", "female")
df[factor_vars] <- lapply(df[factor_vars], factor)

full_model <- glm(healthy ~ smoke_current + insurance + race + marital_status + age_center +
    female, family = poisson(link = "log"), data = df)
step_model <- stats:::step(full_model, direction = "backward", k = qchisq(0.10, df = 1,
    lower.tail = FALSE))
```

[4] "Forward" stepwise regression is a similar method – start with an intercept-only model and iteratively add covariates instead of taking covariates away from a 'full' model. The criticisms of both methods are similar, so we focus on backwards here. Bottom line: we don't recommend either for causal inference.

```
summary(step_model)
```

Results: The stepwise procedure kept all 6 variables, but it removed several parameters associated with two variables: Race (yellow), Marital status (a, widow)[5]. All measures of association (i.e., prevalence ratios) are in the expected direction; for example, the prevalence of high health status is lower for smokers than non-smokers (Table 1).

Table 1. Model estimates from a backward stepwise regression procedure, with a p-value threshold of 0.1.

Variable	Prevalence ratio	p-value	95% CI	
Currently smoke				
Yes, daily (Ref)	1.000	--	--	--
Yes, less than daily	1.013	0.097	0.997	1.029
Not currently	1.025	<0.001	1.018	1.033
Insurance				
Yes (Ref)	1.000	--	--	--
No	0.954	<0.001	0.950	0.957
Race				
White (Ref)	1.000	--	--	--
Black	0.976	<0.001	0.968	0.984
Yellow	*dropped*	--	--	--
Brown	0.985	<0.001	0.981	0.989
Indigenous	0.936	<0.001	0.905	0.969
Marital status				
Married (Ref)	1.000	--	--	--
Separated	1.014	<0.001	1.000	1.028
Marital status (a)	*dropped*	--	--	--
Widow	*dropped*	--	--	--
Single	0.991	<0.001	0.987	0.996
Age				
Age, linear	0.997	<0.001	0.997	0.997
Gender				
Male (Ref)	1.000	--	--	--
Female	0.978	<0.001	0.974	0.982

Statistical and scientific implications of Method 1.

The backwards stepwise procedure (like most automated algorithms) is purely statistical and does not consider other scientific factors in the decision to keep or drop a variable. For example, by using a p-value threshold, the software did not consider the size of the prevalence ratio itself, and how much the adjusted (conditional) prevalence ratio (PR) may differ from the crude PR. With a smaller dataset, we might estimate a relatively large PR but have a p-value that does not meet the threshold, and it would

[5] Remember, in a regression, each level of a categorical variable (when factored) is a separate parameter and has its own test statistic and p-value.

be dropped, which we likely would not want, given our exploratory framework. Or what if the crude PR for *currently smoking* was almost identical numerically to the one presented here, adjusting for 5 other variables (8 additional parameters)? Does it make sense to include these 8 extra parameters in the model, which will decrease the precision of our PR for *currently smoke*, if the validity of the PR is not diminished? Probably not.

Second, the procedure drops parameters that are associated with a variable, but not others within that same variable. For example, it did not include a parameter for Marital status (Widow). We are to assume that the PR it would have estimated for Widow would have been close to 1.00 (the referent), but not having a PR for all categories within a variable makes the presentation of results awkward. Generally speaking, with causal inference, we would want to include all (or no) categories of a variable in a regression model.

Third, the procedure does not consider the relationship between all variables in the model. The procedure (like all regressions) allows for correlation among these 6 variables, but it assumes the variables are linearly additive (on the natural log scale in this specific model). But what if some variables interact multiplicatively? The procedure does not account for these more complex interactions.

And finally, unless we use some *a priori* knowledge in interpreting these results, we will likely interpret each of the estimated PRs from this model as the **total effect** of the exposure on the outcome, health status. But these PRs should only be interpreted as the total effect if each of the other variables is a confounder of this PR. But we know, by definition, that they are not because a variable that is a common cause (confounder) of exposure A on outcome Y will not, itself, be confounded by exposure A if the confounder-exposure association is itself causal. Instead, Exposure A will be a mediator of this variable on its association with outcome Y. Therefore, some of the PRs presented in **Table 1** should be interpreted as direct (non-mediated) effects and not total effects. See **Figure 1** for an example of this phenomenon.

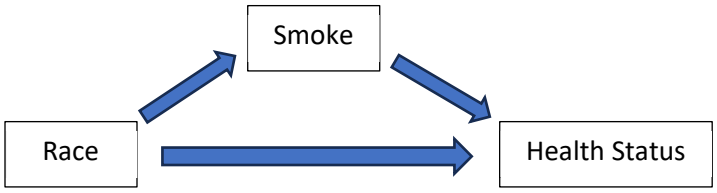


*Figure 1. Causal diagram assuming that 'smoke' is a mediator on the pathway between 'race' and 'health status'. If we add both 'race' and 'smoke' as covariates in the same regression model, the interpretation of the 'smoke' parameter is the **total** effect of smoke on health status, adjusting for the confounding effect of race. The interpretation of the race parameter is different – it is the **direct** effect of race on health status that does not pass through smoke – 'smoke' is not a confounder.*

Method 2. Automated backward stepwise regression using a p-value threshold of 0.05 for variable selection

Method 2 is identical to Method 1, except we have changed the threshold for parameter inclusion at each iteration from 0.1 to 0.05. This will reduce the Type 1 error rate but will increase the Type II error, and it will be harder to 'find' associations (prevalence ratios) that are either smaller or have larger p-values. As expected, compared to Method 1, this method has removed additional parameters (Table 2). For included parameters, the results are largely consistent with Method 1, with a small increase in precision of parameter estimates. The same statistical and scientific issues identified with Method 1 also apply to Method 2.

Table 2. Model estimates from a backward stepwise regression procedure, with a p-value threshold of 0.05.

Variable	Prevalence ratio	p-value	95% CI	
Currently smoke				
Yes, daily (Ref)	1.000	--	--	--
Yes, less than daily	*dropped*	--	--	--
Not currently	1.023	<0.001	1.016	1.030
Insurance				
Yes (Ref)	1.000	--	--	--
No	0.954	<0.001	0.950	0.957
Race				
White (Ref)	1.000	--	--	--
Black	0.976	<0.001	0.968	0.984
Yellow	*dropped*	--	--	--
Brown	0.985	<0.001	0.981	0.989
Indigenous	0.937	<0.001	0.906	0.969
Marital status				
Married (Ref)	1.000	--	--	--
Separated	*dropped*	--	--	--
Marital status (a)	*dropped*	--	--	--
Widow	*dropped*	--	--	--
Single	0.991	<0.001	0.986	0.995
Age				
Age, linear	0.997	<0.001	0.997	0.997
Gender				
Male (Ref)	1.000	--	--	--

Female	0.978	<0.001	0.974	0.982

Method 3. Manual backward stepwise regression using mean squared error (MSE) for variable selection

In Method 3, we abandon the automated procedure (good riddance!) but we continue to use a stepwise method. We also change our evaluation criterion at each step. Instead of relying solely on statistical precision (p-value) of the parameter, we also consider the validity of the estimate (Methods 1 & 2 do not explicitly consider validity). We equally weight the precision and the validity using a single statistic, *mean squared error* (MSE). In this context, the MSE is calculated as,

MSE = bias(β) + variance(β),

Where bias(β) is the square of the difference in the parameter, β, estimated in a reduced model, compared to the parameter, β, estimated in the fully adjusted (presumably the least biased) model, adjusting for factors thought to be confounders of the association between β and the outcome. And variance is the square of the standard error of β.

Simply put:

> Bias = squared difference between the reduced model parameter estimate and the full model parameter estimate.
>
> Variance = squared standard error of the parameter estimate.

The bias and the variance are estimated on the scale on which the model was fitted; in this example, it is the natural log, and not the exponentiation, of β (the prevalence ratio).

Method 3 works well when there is a primary (single) exposure of interest. This can still be within the EDA framework if we do not have good prior knowledge of the association of the exposure on the outcome (e.g., within an innovative research program). Let's consider the exposure *Currently Smoke* as an example.

In the first step, we fit a full model, including the exposure of interest, *Currently Smoke*, as well as all other variables thought to be confounders of this association. For simplicity, and for comparison to Methods 1 and 2, we will assume all 5 other variables are confounders (but we could have included only a subset if we wanted). You will notice that this code is the same model as in Methods 1 and 2 but we do not exponentiate the parameters since we want to estimate the MSE on the original (ln) scale.

Stata code:

```
glm healthy i.smoke_current i.insurance i.race i.marital_status age_center i.female,
fam(Poisson) link(log) vce(robust)
```

R code:

```
m <- glm(healthy ~ smoke_current + insurance + race + marital_status + age_center + female,
    family = poisson(link = "log"), data = df)
robust_vcov <- sandwich::vcovHC(m, type = "HC1")
coeftest(m, vcov = robust_vcov)
```

In the subsequent steps (*code not shown*), we iteratively remove each variable, and we re-fit the model and estimate the bias and the variance, and the MSE at each iteration (Table 3). Note that we do not calculate a Bias or MSE value for the full model since we are assuming the full model to be the least biased model and, therefore, serves as the reference model for the calculation of these statistics. In the associated Stata .do file and Rmd file we provide code for comparing the full model to the reduced models iteratively. The results in Table 3 are reported from Stata. The results from R will be slightly different because of differences in how Stata and R model the robust standard errors.

Table 3. MSE table for method 3

Model (Exposure = Smoke Current)*	exp(Beta), PR	Beta	Bias^2	SE (beta)	Variance	MSE
Full model (5 covariates)						
Yes, less than daily	1.013	0.013		0.008	0.000064	
Not currently smoking	1.025	0.025		0.004	0.000016	
Model removing Insurance status (4 covariates)						
Yes, less than daily	1.014	0.014	0.000002	0.008	0.000064	0.000065
Not currently smoking	1.030	0.030	0.000020	0.004	0.000016	0.000035
Model removing Insurance status and race (3 covariates)						
Yes, less than daily	1.014	0.014	0.0000003	0.008	0.000064	0.000063
Not currently smoking	1.031	0.031	0.000032	0.004	0.000016	0.000047
Model removing Insurance status, race and marital status (2 covariates)						
Yes, less than daily	1.013	0.013	0.0000001	0.008	0.000064	0.000063
Not currently smoking	1.034	0.033	0.000055	0.004	0.000016	0.000070
Model removing Insurance status, race, marital status and age (1 covariate)						
Yes, less than daily	1.029	0.029	0.000242	0.008	0.000064	0.000307
Not currently smoking	1.041	0.040	0.000209	0.004	0.000016	0.000225
Crude model (only smoke current)						
Yes, less than daily	1.030	0.030	0.000273	0.008	0.000064	0.000337
Not currently smoking	1.037	0.036	0.000116	0.004	0.000016	0.000131

*reference level for smoking is "currently smoking daily"; PR – Prevalence ratio; SE – Standard deviation; MSE – Mean Square Error

From the table of results, we see that the model with the smallest MSE is the model that had 4 covariates (removing *insurance status*). Removing insurance status does change our parameter estimates slightly, but not very much. Thus, we might be willing to report a PR for *Current Smoke* that does not adjust for insurance status. It should be noted that, given our large sample size, our variance estimates are very consistent across models. We have many degrees of freedom with which to estimate our parameters. With a smaller dataset, the variance would be expected to influence the MSE value more prominently. But in this case, changes in the MSE between models are driven primarily by changes in bias.

A few important things are worth noting about Method 3. Here, we have fit models with a focus on a single exposure, *current smoke*, but we could re-do this same method for other exposures of interest, such as *age*. We could use the same set of covariates, or we could change them, depending on how we believe these other variables may be related to the new exposure (i.e., as confounders, mediators, etc.).

But even considering the exposure, *current smoke*, we could have fit several other models, including different combinations of these 6 variables. We chose to fit these 6 models, reducing the number of

variables by 1 at each iteration[6]; but you could choose to compare other permutations of models with these 6 covariates. For example, from the Table, it appears that *age* has a strong influence on the validity of the model, as the bias goes way up when we remove *age*. If *age* is indeed a confounder of the association between current smoking and health status, it appears to be an important one. With this knowledge, we could have chosen to fit a model with only *age* and *current smoke* and compared this to our full model.

Method 3, like Methods 1 and 2, assumes that all variables added to the model are done so because they are confounders; but the distinction is that because Method 3 chooses one variable to be the 'exposure' (in this case, *current smoke*), we are only making the assumption that the other included variables are confounders of the exposure and they are not confounding each other. Furthermore, like Methods 1 and 2, the models fit in Method 3 are also assuming that parameters are linear combinations and do not assume multiplicative interactions between variables. But we could certainly include these if we wanted. For example, let's assume that we believe that current smoke and race interact multiplicatively to cause health status. The full model, including all 6 variables and the interaction between smoke current and race would be,

Stata code:

```
glm healthy i.smoke_current##i.race i.insurance  i.marital_status age_center female,
fam(Poisson) link(log) vce(robust) eform
```

R code:

```
# due to issues with estimation of interaction, remove 3 with "unknown" race
df2<-df %>% filter(race!=9)

# Fit Poisson GLM with log link; interaction
m <- glm(
  healthy ~ smoke_current * race + insurance + marital_status + age_center + female,
  data = df2,
  family = poisson(link = "log")
)

# Robust (HC1) covariance matrix ~ Stata vce(robust)
vcov_rob <- sandwich::vcovHC(m, type = "HC1")

# Exponentiated coefficients (IRRs) with robust SEs and 95% CIs (eform-like output)
est   <- coef(m)
se    <- sqrt(diag(vcov_rob))
z     <- est / se
pval  <- 2 * pnorm(abs(z), lower.tail = FALSE)

IRR    <- exp(est)
IRR_se <- se * IRR  # delta method
```

[6] Since we are comparing one model to another model, the order of which variables are included does not matter, like it does in backward/forward step-wise regression. In Method 3, we are adding variables using theory and not p-values.

```
IRR_lo <- exp(est - 1.96 * se)
IRR_hi <- exp(est + 1.96 * se)

results <- data.frame(
  term      = names(est),
  beta      = est,
  se_robust = se,
  z         = z,
  p_value   = pval,
  IRR       = IRR,
  IRR_se    = IRR_se,
  IRR_lo    = IRR_lo,
  IRR_hi    = IRR_hi,
  row.names = NULL
)

print(results)
```

Because Method 3 forces us to conceptualize a particular 'exposure', as well as a set of covariates confounding this exposure, we are moving away from a purely data-driven, empirical approach like we used in Methods 1 and 2. Instead, Method 3 may be considered a hybrid approach, where we are using some *a priori* theory of causal relationships, as well as empirically-derived statistics, like MSE, to drive our modelling choices.

Method 4. Multiple regression models fit separately based on use of a Directed Acyclic Graph (DAG).

In Method 4 we move even further away from a data-driven approach to variable selection compared to Method 3. Instead, we will rely solely on our *a priori* theory and will use a technique common in the field of Epidemiology called a directed acyclic graph (DAG)(7). This method most closely resembles a confirmatory data analysis; however, we believe we are still firmly within an EDA framework for at least two reasons: 1) we remain interested in multiple causes of health status but there is not strong evidence for these associations; 2) we remain uncertain of causal pathways between variables that may act as confounders, colliders, or mediators of the associations of interest. In addition to helping us determine which variables we should include in a model, the use of a DAG also allows us to see how our inference can change within the same statistical model, depending on the parameter of focus.

Our general scientific question remains, 'what is causally associated with health status'? We draw a causal diagram to depict how each of the variables of interest might be related to each other. **This is our best guess**. It can be purely speculative or based on previous literature or social or biological plausibility.

One DAG based on these variables might look like this. All DAGs presented here use Daggity software (www.daggity.net), but you can use other software (or paper/pencil), as you like. Code for creating this DAG can be found in the **Appendix**.

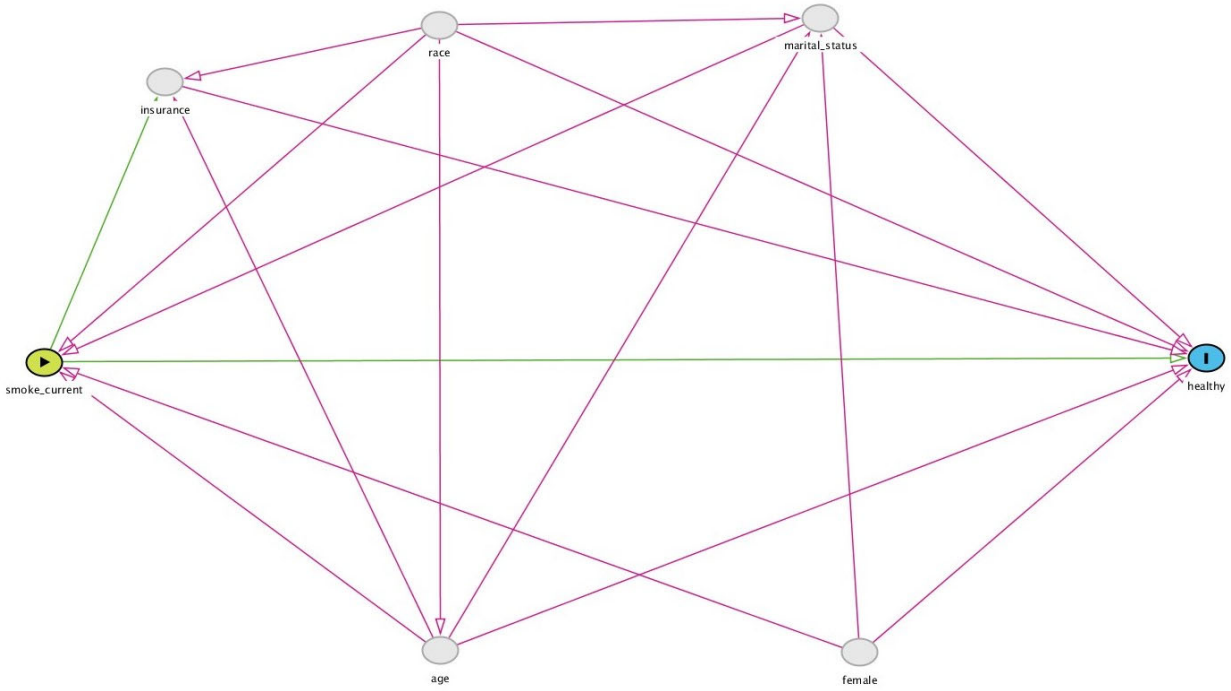


Figure 2. DAG 1, Method 4

Let's take a moment to explore this DAG. As with all methods presented, we have specified 'healthy' as our outcome of interest (blue circle). We have further specified our exposure of interest as 'smoke_current'. The statistical model we fit will not care that we specify our smoke_current as our exposure – a generalized linear model treats the main effects of all independent variables the same. However, it is often important for us to specify an exposure when drawing a DAG, **precisely because the model treats all independent variables the same**. To illustrate why this is the case, let's look at all of the other variables in the DAG and explain **why we included them and how we believe they are related to other variables in the diagram**.

Smoke_current: we know that smoking is bad for your health. We also know that rates of smoking can vary by different characteristics, such as age, gender, etc. Some of these characteristics may cause you to smoke, and other characteristics may simply covary with smoking. Regardless, if a variable is a common cause of both current smoking and of health status, then we would want to consider it a confounder and (usually) would want to condition our analysis (e.g., adjust) on this variable. Here, we believe that age, gender, race, and marital status may all be confounders of the smoke current-health status association.

insurance: if you have insurance, you may have better health status because you have better access to healthcare services. We have an arrow into insurance from smoke_current, which implies that insurance may be a mediator on the causal pathway between smoke_current and healthy. Insurance also has an arrow into it from race and age. Age may be causally associated with age as younger people may have

insurance through a caregiver (e.g., parent), but young adults may not be able to afford insurance as easily.

Race: your race may be causally associated with your health status. This could be due to biology; but more often, race is causally related to health status due to downstream (mediated) effects of race due to racism, such as access to insurance, marital status or smoke_current.

Marital status: marital status may be causally associated with health status. Perhaps being married provides social, psychological or financial support that may improve health status. Race, age and female may all be causally associated with marital status. Marital status may be causally associated with smoke_current – perhaps single persons are more likely to smoke since they don't have to live with a partner who may not like smoke.

Age [age_center]: age is a well-known causal predictor of health status. Age may also be causally associated with smoke_current, as well as insurance and marital status. There is also an arrow into age from race. This doesn't imply that race causes age but that perhaps they covary (i.e., are correlated) in the population due to an unmeasured common cause, such as racism. Maybe racism causes people of color to have a higher mortality rate, which would lead to a lower mean age of people of color in the population, compared to other racial groups. Age was mean centered [age – (mean(age))] for all models that adjusted for age. This is not necessary, but makes intercept terms in multivariable models more interpretable.

Female: gender may be a causal predictor of health status, which could vary by population and culture. Females may also have a lower (or higher) risk of smoking in this population compared with males. Females may also be more likely to be married in this population but they may also be more likely to be widowed due to higher mortality rate among males (potential partners).

As you can see, drawing a DAG takes quite a bit of thought. It is also subjective. We may be wrong!

We decided to draw the DAG in this way, but this is purely based on prior knowledge and educated guesses. But at some point we have to decide that this is the best DAG we can draw, given the data that we have and our prior knowledge of these causal processes. How would this inform our statistical analysis? The first thing we might realize is that **we will not be able to fit a single model to estimate the causal association of each of these 6 independent variables as we did in Methods 1-3**! Let's take each of these 6 variables at a time to see why.

1. **smoke_current:** [Model 1] to estimate the total causal effect of smoke_current on healthy, we would need to adjust (or condition) for race, marital status[7], age and female, since all of these variables act as confounders of this causal association. We would **not adjust for insurance** since it is a mediator, not a confounder, on this pathway. If we did adjust for insurance, we would be

[7] We might choose not to condition on marital status because in addition to acting a confounder it also acts as a 'collider' and can introduce a selection bias of our causal estimate if we condition on this variable; but that topic is beyond the scope of this core guide.

estimating the direct effect of current smoking status on healthy that does not occur through the mediated (indirect) effect of insurance status.

Stata code:

```
glm healthy i.smoke_current i.race i.marital_status age_center i.female, fam(Poisson)
link(log) vce(robust) eform
```

R code:

```
# Fit Poisson GLM with log link
m <- glm(
  healthy ~ smoke_current + race + marital_status + age_center + female,
  data = df,
  family = poisson(link = "log")
)

# Robust (HC1) covariance matrix analogous to Stata's vce(robust)
vcov_rob <- sandwich::vcovHC(m, type = "HC1")

# eform-style output: exponentiated coefficients (IRRs) with robust SEs and CIs
est   <- coef(m)
se    <- sqrt(diag(vcov_rob))
z     <- est / se
pval  <- 2 * pnorm(abs(z), lower.tail = FALSE)

IRR    <- exp(est)
IRR_se <- se * IRR                  # delta method
IRR_lo <- exp(est - 1.96 * se)
IRR_hi <- exp(est + 1.96 * se)

results <- data.frame(
  term      = names(est),
  beta      = est,
  se_robust = se,
  z         = z,
  p_value   = pval,
  IRR       = IRR,
  IRR_se    = IRR_se,
  IRR_lo    = IRR_lo,
  IRR_hi    = IRR_hi,
  row.names = NULL
)

print(results)
```

Model estimates.

Currently smoke	Prevalence ratio	p-value	95% CI	
Yes, daily (Ref)	1.000	--		
Yes, less than daily	1.014	0.07	0.999	1.030
Not currently	1.030	0.00	1.022	1.038

**Current smoking leads to lower prevalence of high health status. People who do not currently smoke have a 3% higher relative prevalence of high health status than those who smoke daily, holding race, marital status, age and gender constant.*

You may notice that we are only reporting parameter estimates for our designated exposure and not for our variables treated as confounders. This is for two reasons: 1) we want to focus our readers attention on the exposure since we have parameterized the model specifically for inference on that variable; and 2) the parameter estimates from the other variables do not have the same interpretation as 'total effects'. This fact is explained in greater detail earlier in the document

2. **insurance**: [Model 2] to estimate the total causal effect of insurance on healthy, we would need to adjust for smoke_current, age and race, since they are all potential confounders. But we would not need to adjust for female or marital status since neither is associated with insurance.

Stata code:

```
glm healthy i.insurance i.smoke_current age_center i.race, fam(Poisson) link(log)
vce(robust) eform
```

R code:

(For the rest of the R code in this section, we only replicate the glm command. To obtain robust standard errors and output the IRR is a straightforward extension of the R code in the previous code box.)

```
glm(
  healthy ~ insurance + smoke_current + race + age_center,
  data = df,
  family = poisson(link = "log")
)
```

Model estimates.

Insurance	Prevalence ratio	p-value	95% CI	
Yes (Ref)	1.000	--		
No	0.953	0.000	0.949	0.957

**Lack of insurance leads to a lower prevalence of high health status, holding smoking status, age and race constant.*

3. **race**: [Model 3] to estimate the total causal effect of race on healthy, we would probably want to adjust for age, but not for anything else. As we said before, age and race likely covary in the population but are not causally associated. That means age may be confounding the race-healthy association. **We would not adjust for smoke_current, marital status or insurance since these variables are mediators of the race-healthy association**. If we did adjust for smoke_current, insurance and marital status, we would be estimating the **direct** effect (not total effect) of race on health status that does not pass through these 3 variables – you would need to think about why you would want to estimate this direct effect before doing so. In fact, adjusting for variables, such as these, would underestimate the effect of race on healthy status due to racism. Since we want to estimate the total effect of race on health status, we will not consider those three variables here and their exclusion would not induce confounding bias of our estimates.

Stata code:
```
glm healthy i.race age_center, fam(Poisson) link(log) vce(robust) eform
```

Stata code:

```
glm(
  healthy ~ race + age_center,
  data = df,
  family = poisson(link = "log")
)
```

Model estimates.

Race	Prevalence ratio	p-value	95% CI	
White (Ref)	1.000	--	--	--
Black	0.966	0.068	0.958	0.974
Yellow	0.979	0.000	0.956	1.001
Brown	0.975	0.000	0.970	0.979
Indigenous	0.923	0.000	0.892	0.955

**Compared to white, other races have a lower prevalence of high health status, holding age at its mean value. The lowest prevalence of high health status is observed in Indigenous persons.*

4. **marital status** [model 4]: marital status may be causally associated with health status. Perhaps being married provides social, psychological or financial support that may improve health status. Race, age, and female may all be causally associated with marital status. Marital status may be causally associated with smoke_current – perhaps single persons are more likely to smoke since they don't have to live with a partner who may not like smoke.

Stata code:

```
glm healthy i.marital_status i.race age_center female, fam(Poisson) link(log)
vce(robust) eform
```

R code:

```
glm(
  healthy ~ marital_status + race + age_center + female,
  data = df,
  family = poisson(link = "log")
)
```

Model estimates.

Marital Status	Prevalence ratio	p-value	95% CI	
Married (Ref)	1.000	--	--	--
Separated	1.007	0.299	0.993	1.021
Marital status (a)	0.995	0.339	0.984	1.006
Widow	0.999	0.899	0.987	1.011
Single	0.984	0.000	0.979	0.989

**There appears to be only little difference in the prevalence of high health status by marital status, holding race and gender at their baseline levels and age at its mean level (linear). The lowest prevalence of high health status is observed in single participants.*

5. **Age (mean centered, linear)**: [Model 5] according to our DAG, the only common cause of age and healthy status is race (again, we don't think race causes age, but we believe they might co-vary in the population). We decide to adjust for race but not for any other variables. For simplicity, we treat age as a linear term in the model. We should explore this in more detail to see if this assumption holds. One method of doing this is to fit a LOESS smooth line to visually inspect this. In our case, age does not appear linearly associated (see appendix for code and graphic). But we continue with the linear term here for simplicity.

Stata code:

```
glm healthy age_center i.race, fam(Poisson) link(log) vce(robust) eform
```

R code:

```
glm(
  healthy ~ race + age_center,
  data = df,
  family = poisson(link = "log")
)
```

Model estimates.

Age (linear)	Prevalence ratio	p-value	95% CI	
Intercept (_cons)	0.950	0.000	0.948	0.953
Age (mean centered)	0.997	0.000	0.997	0.997

**if we assume that the relationship between the prevalence of health status and age is linear, then we have evidence that the prevalence of high health status goes down with increasing age, holding race constant (i.e., for any race). The intercept term from the model is 0.950, which estimates the prevalence of high health status at the mean age of the sample (43.3 yrs) is 95%.*

6. **female**: There are several factors in this population that may covary with gender, such as smoking status or marital status. But according to our DAG, we have determined the only common causes of gender and health status is age; therefore, we choose to adjust for only age.

Stata code:

```
glm healthy female age_center, fam(Poisson) link(log) vce(robust) eform
```

R code:

```
glm(
  healthy ~ female + age_center,
  data = df,
  family = poisson(link = "log")
)
```

Model estimates.

Female	Prevalence ratio	p-value	95% CI	
Male (Ref)	1.000	--	--	--
Female	0.981	0.000	0.977	0.985

**Females have a lower prevalence of healthy status in this population compared to males, holding age at its mean.*

Method 5. Random Forest procedure for variable selection.

Our final method is a significant departure from the previous 4 for several reasons. A random forest (RF) is a supervised machine learning method that is commonly used to make predictions. These predictions can either be classifications, such as “is the participant ‘healthy’, yes or no?” or it can be regression-based, which predicts a real number (e.g., 3.5, -2.7). Random forest is an ‘ensemble’ method which means that it pools results from multiple ‘trees’ to create a ‘forest’. Combining results in this manner

can reduce bias and improve generalizability of the results. Each 'tree' is created by taking a random sample (with replacement) of the data and 'bagging' this result. Many trees are created and then pooled together. Random forest is an attractive method for prediction for several reasons. Unlike a regression-based approach (Methods 1-4) we do not need to specify the functional relationship between covariates, such as non-linearity between the dependent and independent variables, or interactions between any variables, making it highly flexible.

But let's pause for a moment to acknowledge an important distinction between a random forest and the regressions in Methods 1-4. A random forest is designed to be used for **prediction**, not explanation (i.e., causal inference). Our goal with random forests is to predict the outcome, not to say whether one independent variable causes the outcome, and we assess the quality of the random forest by its ability to accurately predict this outcome. In the methods below, we will show how you can determine whether a single independent variable is 'important' in predicting the outcome, but we are not able to generate a **strength of the association** between the independent and dependent variable. For this reason, we do not recommend random forests for causal inference, but we include the method here to contrast it with other methods and to demonstrate the value in its own right.

Using our example dataset, we will predict whether the participant is 'healthy' just as we have done in the previous methods. We will also explore another outcome – whether the participant is 'very healthy'. The reason we will look at this additional outcome is that the random forest method is notoriously bad at predicting rare outcomes and very good at predicting common outcomes. Since our 'healthy' and very healthy outcomes are 93% and 87% prevalent, respectively, we will demonstrate how a random forest procedure performs on each.

Our RF procedure is quite similar to Methods 1 and 2 in that they are all highly data driven. We are using the data to tell us what is 'important' and we are not imparting any *a priori* theory as we were in Method 4. We will use the same independent variables as in Methods 1 and 2 but we will add one additional that we call 'random_var'. This variable is a randomly generated normally distributed variable with mean 0. We include it as a 'control' variable to demonstrate that this variable, since it has no relationship with the outcome, should not be 'important' in our RF results.

The RF is predicting 'healthy' and 'very healthy' with the following general form:

$$healthy \sim smoke_current\ insurance\ race\ marital_status\ age_center\ female\ random_var$$

$$veryhealthy \sim smoke_current\ insurance\ race\ marital_status\ age_center\ female\ random_var$$

The algorithm will iterate 500 times (user-specified) and each time it will draw a random sample from the 60,020 participants and will predict the outcome variable. In this version, we allow 3 variables (mtry=3 in R syntax) to be randomly selected as candidates at each split of the tree. Over all iterations, it determines which variables are the best at predicting the outcome. One common metric to report model accuracy is the 'out of bag' (OOB) error rate. We can also plot the variable importance.

Note: In a random forest, each tree is trained on a random sample of the dataset, and because this sampling is done *with replacement*, some rows get used to train that tree while others are left out. The

rows that are not used to train a particular tree are called **out-of-bag (OOB)** samples. After the tree is built, we test it by asking it to predict the outcomes of these OOB rows—the data it never saw during training. If the tree predicts one of these left-out rows incorrectly, that counts as an **OOB error**. By checking how often all the trees in the forest get their OOB samples wrong, we get the **OOB error rate**, which tells us how well the model is likely to perform on new, unseen data.

Stata code: We do not feel comfortable sharing Stata code for random forest methods due to some observed issues with estimation. We recommend using R software instead.

Stata code:

We do not feel comfortable sharing Stata code for random forest methods due to some observed issues with estimation. We recommend using R instead.

Rmd code:

```
```{r}
# Set seed for reproducibility
set.seed(1234)

# factor healthy
brazil$healthy <- as.factor(brazil$healthy)

model <- cforest(
  healthy ~ smoke_current + insurance + race + marital_status + age_center + female +
random_var,
  data = brazil,
  controls = cforest_unbiased(ntree = 500, mtry = 3)
)
## summary of the model
summary(model)

## estimating out of bag error
pred_oob <- predict(model, OOB = TRUE, type = "response")
oob_error <- mean(pred_oob != brazil$healthy, na.rm = TRUE)
cat("Out-of-bag error rate:", round(oob_error * 100, 2), "%\n")

## Code to run on 'veryhealthy' is exactly the same but changing dependent variable in the
cforest function.
```
```

The OOB error rate for the RF predicting 'healthy' is 8.9%. This is quite good, but remember, that the outcome is very common at 93% so we expect the model to perform well.

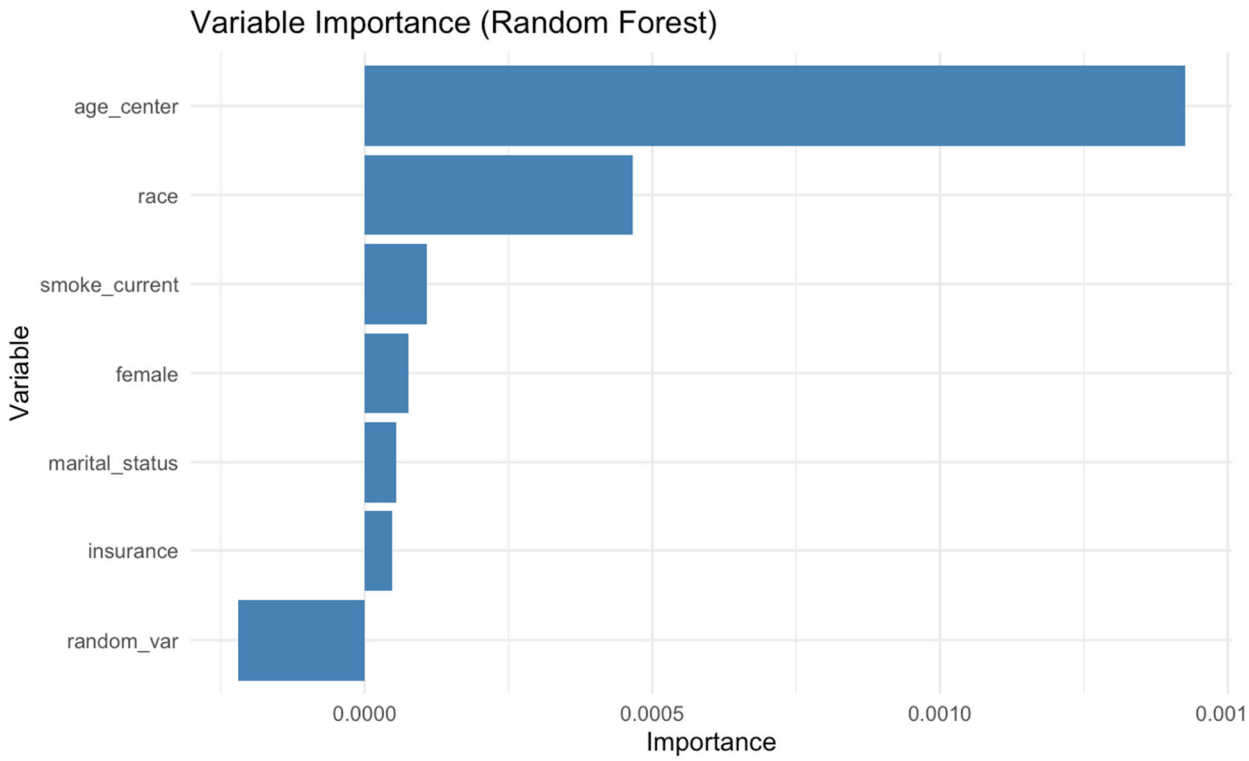


Figure 3. Variable Importance; outcome = 'healthy'

Figure 2 displays the variable importance plot for the RF predicting 'healthy'. Age is clearly the most important variable for predicting if a person is 'healthy', followed by 'race'. As expected, the random variable has no predictive ability at all. The rest look to have very marginal ability to predict so are likely not very important to include.

But when we look at the outcome 'very healthy', which is not quite as prevalent (87%), we see a different result (Figure 3). The OOB error rate has gone up to 12.4%, which was expected, but now 'insurance' is the most important predictor, following by age. The other variables all look to be unimportant in predicting very healthy status in this population.

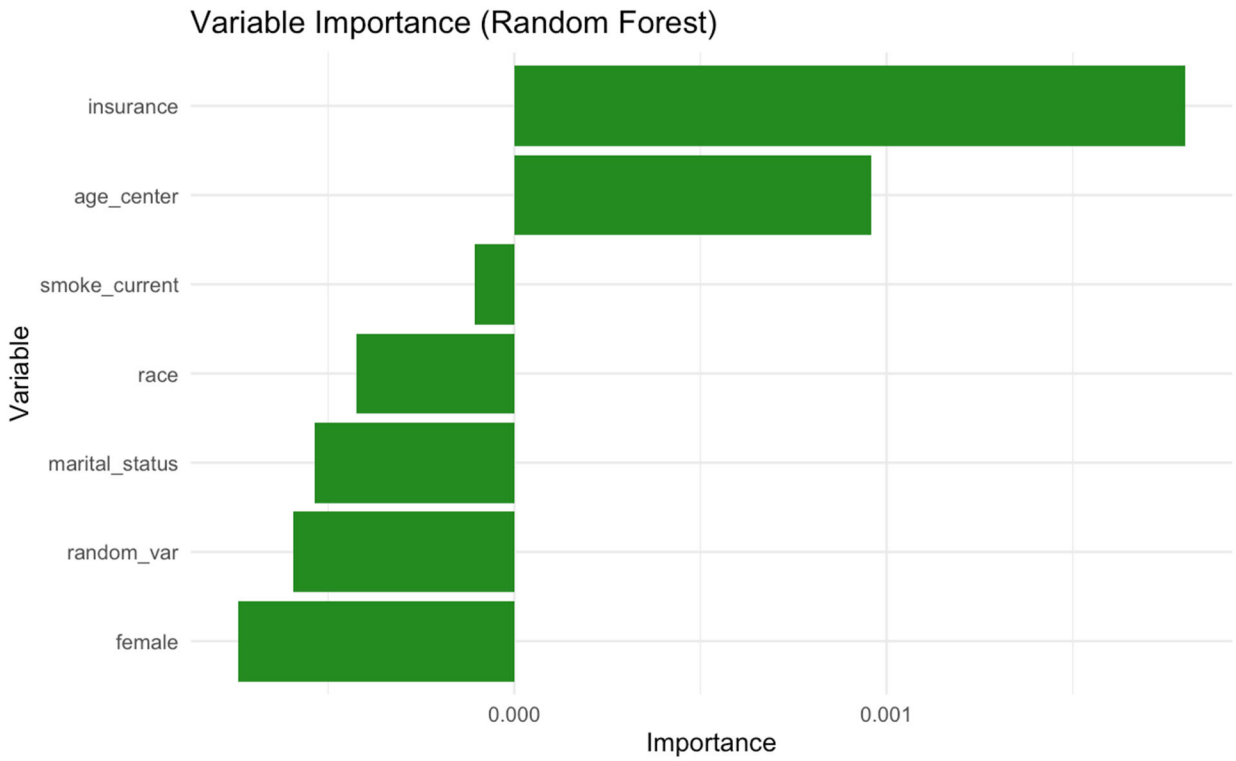


Figure 4. Variable importance; outcome = 'very healthy'

Comparison and summary of the 5 EDA Methods

Methods 1 and 2 (Automated Backward Stepwise Regression with p-value thresholds 0.1 and 0.05) are straightforward and automate variable selection based on statistical significance. Using a higher threshold (0.1) allows more variables to remain, reducing false negatives, whereas the stricter 0.05 threshold reduces false positives but risks omitting meaningful predictors. Both methods rely solely on p-values, ignore effect size and scientific context, may inconsistently include variable categories, and assume linear additive effects without considering interactions. They also conflate total and direct effects without accounting for underlying causal structures. **For these reasons, we do not recommend Methods 1 or 2 in exploratory data analysis.**

Method 3 (Manual Backward Stepwise Regression Using Mean Squared Error) improves on automated procedures by balancing bias and variance, integrating substantive knowledge through focus on a primary exposure and its confounders. It promotes more thoughtful model building, can incorporate interaction terms, and enhances interpretability. Its drawbacks include increased labor, reliance on the assumption that the full model is least biased, and limited simultaneous exploration of multiple exposures. It still assumes linearity unless interactions are explicitly modeled. Method 3 can be a good option, especially if you want to take a 'hybrid' approach of data driven and theory and when the data sample size is small enough that maximizing precision is important (when the data set is very large, precision is not an issue).

Method 4 (Theory-Driven Variable Selection Using Directed Acyclic Graphs) relies on *a priori* causal knowledge to carefully select variables, helping distinguish confounders, mediators, and colliders. This method improves causal inference transparency and avoids biases introduced by automated selection. However, it depends heavily on the researcher's expertise and assumptions, may be subjective, less exploratory, and can be challenging to implement with complex datasets. Its success hinges on the completeness and accuracy of the underlying causal theory. If you are in a purely 'exploratory' mindset, you may not feel comfortable with Method 4.

Method 5 (Random Forests) offers a flexible, non-parametric approach that can capture complex nonlinear relationships and interactions without requiring strong modeling assumptions. It handles large datasets well and provides automated variable importance measures. However, it tends to be less interpretable, does not produce direct effect estimates like prevalence ratios, and may exhibit bias in variable importance scores. Additionally, it is purely data-driven, lacking integration with causal theory. For these reasons, we do not recommend random forests, or similar machine learning approaches for causal inference (except for sophisticated causal inference methods that often use ML in estimating inverse probability weighting).

Recommendations

It is difficult to be prescriptive with EDA. It is a skill that you attain a 'feel' for over time. With that said, there are some clear good, and bad, practices.

1. If you are truly in an exploratory mindset, stay there. Enjoy exploring the data but resist the urge to let one dataset drive your inference too far. With EDA, it's more about the journey – and developing hypotheses – and not the destination.
2. If you want to engage in pursuing causal inference, use at least some theory to guide your analyses.
3. Don't over-rely on statistical tests to guide inference or next steps. Rely more on direction and size of effects since these are much less determined by sample size.

Appendices

Appendix 1. DAGitty code for Figure 1.

Paste into 'model code' window' at: https://www.dagitty.net/dags.html#

```
dag {
age [latent,pos="-0.882,1.501"]
female [latent,pos="0.188,1.510"]
healthy [outcome,pos="1.073,0.151"]
insurance [latent,pos="-1.584,-1.125"]
marital_status [latent,pos="0.087,-1.421"]
race [latent,pos="-0.885,-1.388"]
smoke_current [exposure,pos="-1.892,0.169"]
age -> healthy
age -> insurance
age -> marital_status
age -> smoke_current
female -> healthy
female -> marital_status
female -> smoke_current
insurance -> healthy
marital_status -> healthy
marital_status -> smoke_current
race -> age
race -> healthy
race -> insurance
race -> marital_status
race -> smoke_current
smoke_current -> healthy
smoke_current -> insurance
}
```

Or you can created DAGs using the dagitty library in RMarkdown:

```
``{r}

# Load the dagitty package

library(dagitty)

```

```
```{r}

# Drawing the DAG

EDA_DAG <- dagitty('

dag {

age [latent,pos="-0.882,1.501"]
```
```

```
female [latent,pos="0.188,1.510"]

healthy [outcome,pos="1.073,0.151"]

insurance [latent,pos="-1.584,-1.125"]

marital_status [latent,pos="0.087,-1.421"]

race [latent,pos="-0.885,-1.388"]

smoke_current [exposure,pos="-1.892,0.169"]

age -> healthy

age -> insurance

age -> marital_status

age -> smoke_current

female -> healthy

female -> marital_status

female -> smoke_current

insurance -> healthy

marital_status -> healthy

marital_status -> smoke_current

race -> age

race -> healthy

race -> insurance

race -> marital_status

race -> smoke_current

smoke_current -> healthy

smoke_current -> insurance

}

')


# Plot the DAG

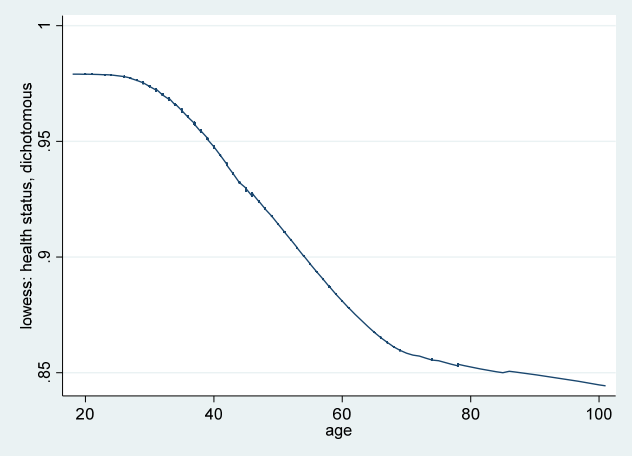
plot(EDA_DAG)
```

Appendix 2. LOESS smooth line for the bivariate association between health status and participant age.

Stata Code:

```
*Lowess exploration of healthy and age.
*large dataset so sub setting for this purpose to test.
set seed 2038947
sample 10 /*keeping just 10% of the full sample*/
count

lowess healthy age, generate(healthy_lowess_age)
line healthy_lowess_age age, sort
```



*Age does not look to be linearly associated so treating it as a linear term in the model is a simplification. Note, however, there is only little data at the right tail of the age distribution (ages > 86 yrs).

R Code:

```
## Loess curve
```{r}
# Lowess (LOESS) exploration of healthy and age

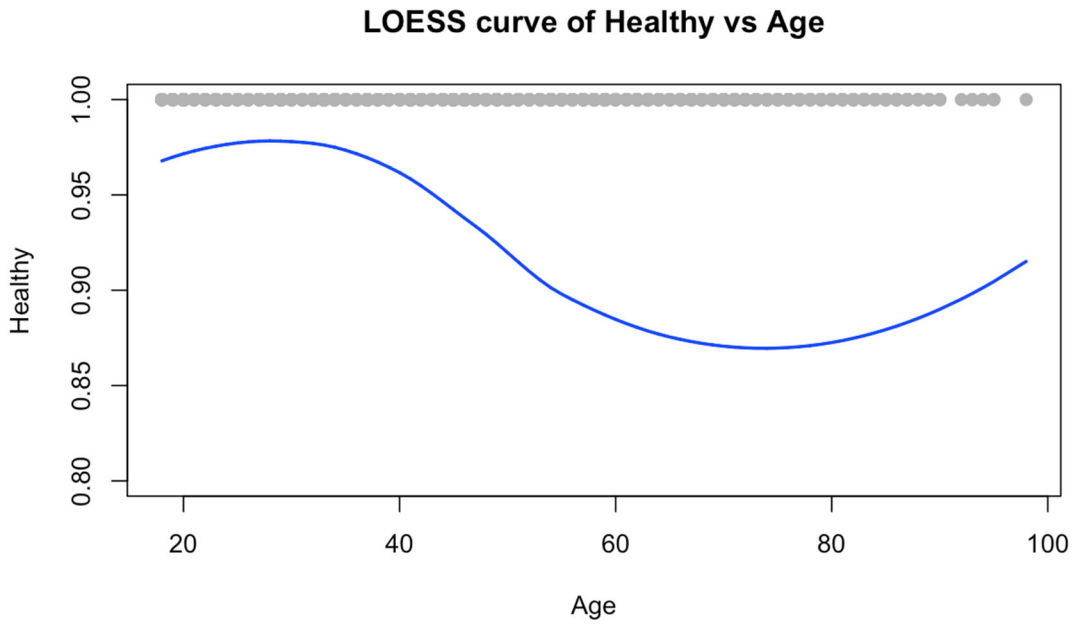
# Set seed and subset the data to 10% (similar to Stata's "sample 10")
set.seed(2038947)
sub_data <- df[sample(1:nrow(df), size = 0.10 * nrow(df)), ]

# Fit a loess (locally weighted regression) model
loess_fit <- loess(healthy ~ age, data = sub_data)

# Predict smoothed values
sub_data$healthy_lowess_age <- predict(loess_fit)

# Plot the results, setting y-axis limits from 0.8 to 1.0
plot(sub_data$age, sub_data$healthy,
    main = "LOESS curve of Healthy vs Age",
    xlab = "Age", ylab = "Healthy",
    pch = 19, col = "grey70",
```
```

```
    ylim = c(0.8, 1.0))

# Add the smoothed LOESS curve
lines(sub_data$age[order(sub_data$age)],
      sub_data$healthy_lowess_age[order(sub_data$age)],
      col = "blue", lwd = 2)
```

Curve looks a bit different from Stata but likely due to different random sample.