Models

Ashley I. Naimi, PhD

Outline

Models

- Overview of Modeling in Epidemiology
- Causal versus Statistical Models
- Parametric versus Nonparametric Models
- Marginal versus Conditional Models
- Models versus Estimators
- Counterfactual Models
  - Marginal Structural Models
  - Structural Nested Models

Models

Overview of Models in Epidemiology

Models are an integral part of science (Rosenblueth and Wiener 1945). Epidemiologists rely exclusively on models to understand the relation between a particular exposure and outcome of interest. These models are often of a very particular type. Indeed, the most common approach to modeling in epidemiology is statistical regression (Freedman 2008). Logistic regression in particular has become an analytic workhorse for epidemiologists when they seek to understand the relation between an exposure and a (dichotomous) health outcome.

Typically, the use of a logistic regression model proceeds as follows:[1] 1) a researcher poses a question about the relation between an exposure and outcome of interest; 2) a host of potential threats to the validity of an assessment of the exposure-outcome relation are identified, most notably confounding variables; 3) data are collected in which the exposure-outcome relation can be quantified after mitigating the impact of the potential confounding variables; 3) the data are analyzed using logistic regression, with the measured confounders included in the model.

[1] This is a gross oversimplification. But the complexity that is being ignored here does not address the modeling issues that will be raised in subsequent sections.

The logistic model is often formulated as follows

$$\text{logit}[P(Y = 1 \mid X, C)] = \beta_0 + \beta_1 X + \beta_2 C$$

where $\text{logit}[a] = \log[a/(1-a)]$.

More practically, suppose we were intersted in the relation between s

```
aa <- read_csv("./nhefs.csv")
# original sample size
nrow(aa)
```

```
## [1] 1746
```

We'll restrict our attention to a small subset of covariates:

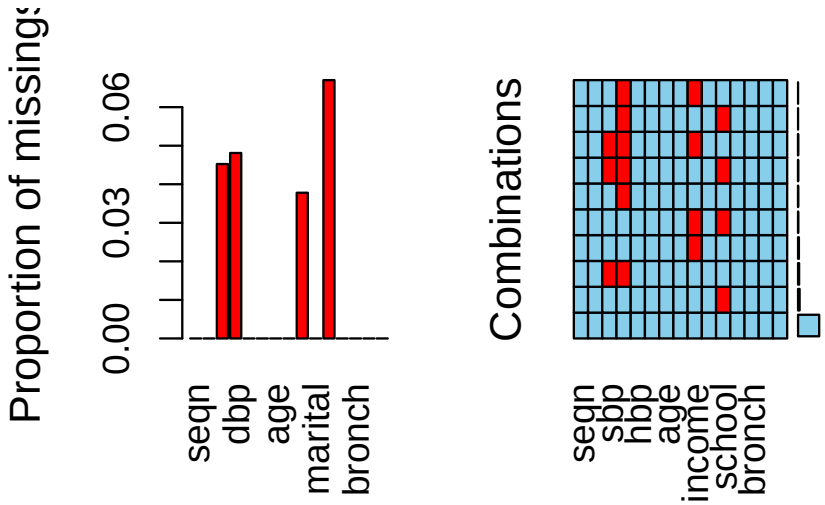
```
a <- aa %>% select(seqn, qsmk, sbp, dbp, hbp,
    sex, age, race, income, marital, school, asthma,
    bronch, diabetes, birthcontrol)
```

Missing data is always important to address. We use the `aggr` function from the VIM package to create this great plot, showing how much missing data there is, and how it's distributed in teh dataset.

To simplify, we'll restrict to complete cases. Note this is not something that should be done without careful consideration of missing data assumpitons.[2]

[2] For complete case analyses to be valid, data must be MCAR, or missing completely at random. For details, see Little and Rubin (2014).

```
aggr(a)
```

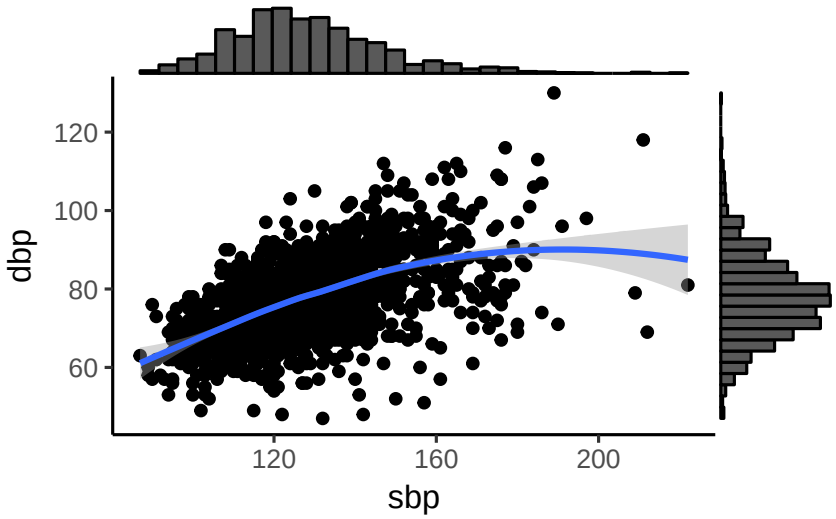



```
a <- a %>% na.omit()
# sample size remaining after restricting to
# complete case
nrow(a)
```

```
## [1] 1489
```

Let's examine the distrbution of systolic and diastolic blood pressure.

```
plot <- ggplot(a, aes(sbp, dbp)) + geom_point() +
    geom_smooth(method = "loess")
ggMarginal(plot, type = "histogram")
```



And finally, a 2×2 table for the relation between smoking and high-blood pressure.

```
a$hbp <- as.numeric(a$sbp > 130 | a$dbp > 80)
```

```
tab1 <- table(a$qsmk, a$hbp)
addmargins(tab1)
```

```
##
##          0    1  Sum
##   0    568  553 1121
##   1    148  220  368
##   Sum  716  773 1489
```

```
chisq.test(tab1)
```

```
##
##  Pearson's Chi-squared test with Yates'
##  continuity correction
##
## data:  tab1
## X-squared = 11.708, df = 1, p-value =
## 0.0006222
```

Causal Models versus Statistical Models

Parametric versus Nonparametric Models

Marginal versus Conditional Models
Models versus Estimators
Counterfactual Models

References

Freedman, D. 2008. *Statistical Models: Theory and Practice*. Revised Edition. New York, NY: Cambridge University Press.

Little, Roderick J. A., and DB Rubin. 2014. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Hoboken, N.J.: Wiley.

Rosenblueth, Arturo, and Norbert Wiener. 1945. "The Role of Models in Science." *Philosophy of Science* 12 (4): 316–21.