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

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.

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 quitting smoking and weight gain. We can examine this relation using data from the NHANES 1 Epidemiologic Follow-Up Study.[2]

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

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

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

[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.

[2] These data are available on the website for the forthcoming book, Causal Inference, by Hernán and Robins. See: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

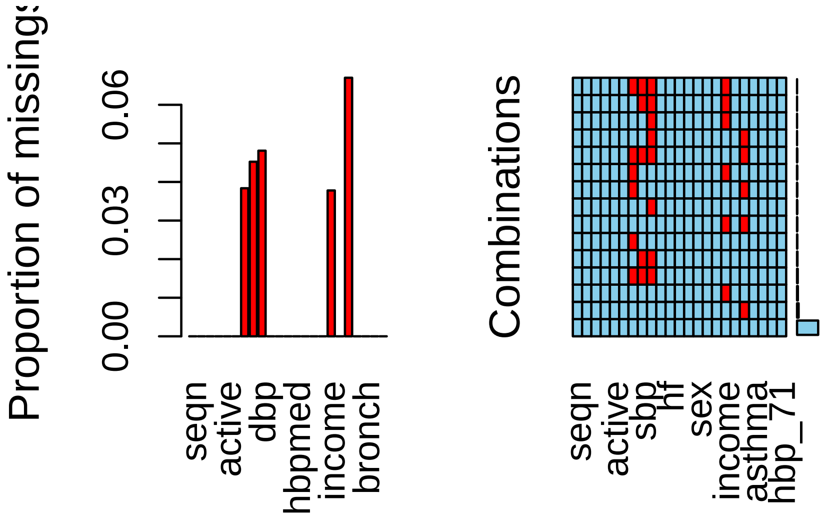
```
a <- aa %>% select(seqn, qsmk, smkintensity82_71, smokeintensity, active, exercise, wt82_71,
    sbp, dbp, hbp, hf, ht, hbpmed, sex, age, hf, race, income, marital, school, asthma, bronch,
    diabetes)
a$hbp_71 <- a$hbp
```

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 the dataset.

To simplify, we'll restrict to complete cases. Note this is not something that should be done without careful consideration of missing data assumptions.[3]

[3] 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)
```

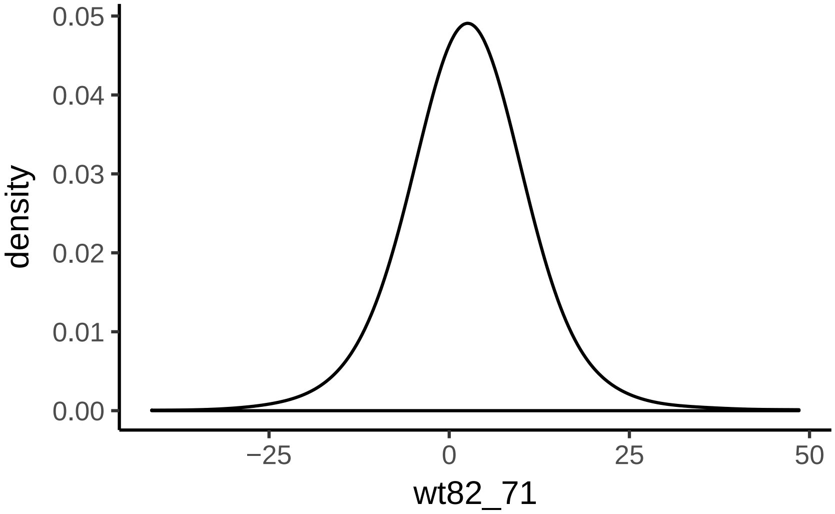


```
a <- a %>% na.omit()
# sample size remaining after restricting to complete case
nrow(a)
```

```
## [1] 1476
```

Let's examine the change in weight between 1971 and 1982.

```
ggplot(a, aes(wt82_71)) + geom_density(bw = 5)
```

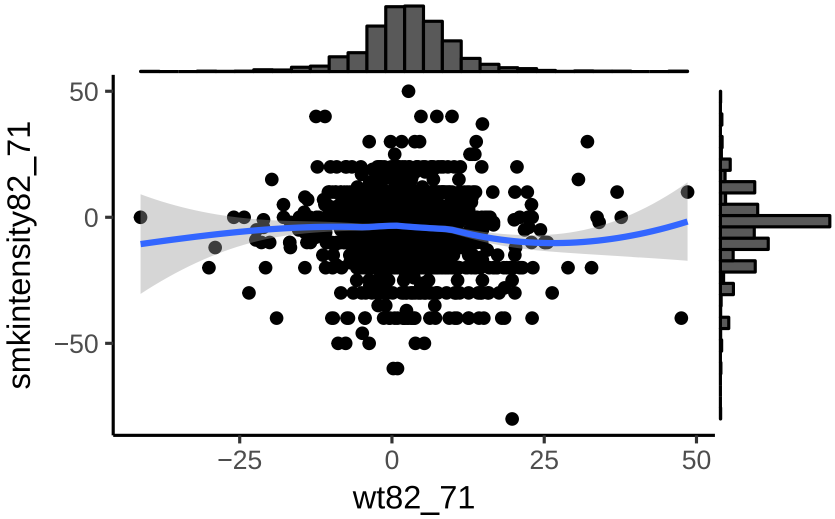


```
quantile(a$wt82_71, probs = seq(0, 1, 0.2))
```

```
##         0%        20%        40%        60%        80%       100%
## -41.280470  -2.496482   1.130996   3.971964   7.708958  48.538386
```

And the distrbution of weight change and smoking intensity change.

```
plot <- ggplot(a, aes(wt82_71, smkintensity82_71)) + geom_point() + geom_smooth(method = "loess")
ggMarginal(plot, type = "histogram")
```



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

```
a$delta <- as.numeric(a$wt82_71 > 0)
```

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

```
##
##          0    1  Sum
##   0    390  724 1114
##   1     94  268  362
##   Sum  484  992 1476
```

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

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab1
## X-squared = 9.7298, df = 1, p-value = 0.001813
```

Traditionally, the approach to quantifying the relation between quitting smoking and gaining weight is to identify a set of confounders. Let's assume for our purpose that the relevant confounders are as listed in the model below. The most common approach to estimating the effect is to fit a logistic model, adjusting for these confounders:

```
model1 <- glm(delta ~ qsmk + sex + age + race + income + marital + school + active + hf + smokeintensity +
    exercise + diabetes + hbp_71, data = a, family = binomial(link = "logit"))
summary(model1)
```

```
##
## Call:
## glm(formula = delta ~ qsmk + sex + age + race + income + marital +
##     school + active + hf + smokeintensity + exercise + diabetes +
##     hbp_71, family = binomial(link = "logit"), data = a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1206  -1.2013   0.7074   0.8954   1.6002
##
```

```
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.870156   0.620795   3.013  0.00259 **
## qsmk            0.624658   0.143764   4.345 1.39e-05 ***
## sex             0.158408   0.122464   1.294  0.19584
## age            -0.042106   0.005301  -7.943 1.98e-15 ***
## race           -0.020490   0.181773  -0.113  0.91025
## income          0.035621   0.026425   1.348  0.17765
## marital         0.024831   0.058920   0.421  0.67343
## school         -0.009861   0.021672  -0.455  0.64909
## active         -0.101246   0.094256  -1.074  0.28275
## hf             -0.255381   0.760797  -0.336  0.73712
## smokeintensity  0.004230   0.005267   0.803  0.42188
## exercise       -0.054735   0.084633  -0.647  0.51781
## diabetes        0.276079   0.210229   1.313  0.18910
## hbp_71         -0.302561   0.217931  -1.388  0.16503
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1867.7  on 1475  degrees of freedom
## Residual deviance: 1758.5  on 1462  degrees of freedom
## AIC: 1786.5
##
## Number of Fisher Scoring iterations: 4
```

The output from this model tells us that the odds of weight gain among those who quit smoking is 1.87 times the odds of weight gain among those who did not quit (with 95% CIs of 1.41, 2.48).

It might be tempting to conclude that this is the effect of quitting smoking on weight gain. One might be further tempted to interpret this effect as a comparison of odds if everyone quit versus if no one quit. Unfortunately, there are a number of considerations related to the types of models we are using that jeapordize the validity of such an interpretation.

Causal versus Statistical

Any statistical association between an exposure and outcome of interest can be caused by a number of relations:

- Direct causation
- Reverse causation
- Confounding
- Selection (collider)
- Chance

These relations are codified in the causal model, but not the staitstical model. For example, depending on how/when it was measured, we might draw a directed acyclic graph in which diabetes is a common cause of both quitting smoking (becuase diabetics will refrain from smoking) and weight gain, or in which diabetes mediates (quitting smoking changes diabetes risk) the relation between smoking and high blood pressure.

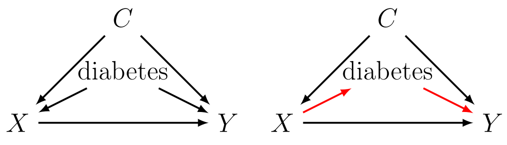


Figure 1: Two scenarios depicting the relation between smoking (X), diabetes, and high blood pressure (Y).

The critical point is that if the interest if the total effect of quitting smoking on the outcome is of interest, and if diabetes is in fact a mediator of this relation, then it should **not** be adjusted for.[4] Otherwise, a key part of the effect will be blocked. Alternatively, if diabetes is confounding the relation between quitting smoking and weight gain, then it must be adjusted for to reduce confounding bias.

Technically, the two DAGs are Markov (or observationally) equivalent, or are in the same **equivalence class** (J. Pearl 2009). Two or more DAGs are said to be Markov equivalent if and only if they have the same skeletons (i.e., the same nodes) and the same set of colliders (Verma and Pearl 1990). In non-technical terms, a set of DAGs that

[4] Note that at times, clinicians/researchers are primarily interested in the effect of smoking independent of it's effect on diabetes. In this case, methods for mediation analysis are required.

are Markov equivalent means that no statistical analysis can be used to distinguish between them.

Parametric versus Nonparametric

A second key consideration is the nature of the assumptions invoked when specifying the logistic regression model. When data are collected for an observational study, assuming the causal model (DAG) is correct, the extent of what we know about how the outcome relates to an exposure and confounders can be written as follows:

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}), \tag{1}$$

where $g(\bullet)$ represents a function of X and C. In an observational cohort study, the exact form of the exposure and outcome models is usually completely unknown (Robins 2001). However, despite this lack of knowledge, by using the logistic model, we are willing to assume the outcome follows a very particular model:

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}) = \text{expit}\{\beta_0 + \beta_1 X + \beta_2 \mathbf{C}\} \tag{2}$$

even though the true model may be any number of models, for example:

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}) = \text{expit}\left\{\frac{\beta_0}{\beta_1 X} + \beta_2 \mathbf{C}\right\} \tag{3}$$

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}) = \beta_0 \times X^{\beta_1} \times C^{\beta_2} \tag{4}$$

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}) = \frac{\beta_0}{\beta_1 X + \beta_2 \mathbf{C}} \tag{5}$$

There are ways to avoid assuming the outcome follows a logistic model conditional on the exposure and confounders. Briefly, this would entail using a nonparametric estimator (such as, e.g., machine learning algorithms) to quantify $g(X, \mathbf{C})$.

Marginal versus Conditional

A third complication with interpreting the estimate for quitting smoking as “what would be observed if everyone quit versus if no one quit" is the distinction between marginal and conditional models. This distinction becomes more complicated when interest lies in estimating non-collapsible parameters (i.e., non-linear model) such as the odds ratio or the hazard ratio (Greenland 2005).

Briefly, a conditional effect estimate can be obtained by fitting a conditional model such as:

$$g[E(Y \mid X, C)] = \beta_0 + \beta_1 X + \beta_2 C$$

A marginal model can be obtained by fitting a marginal model using inverse probability weighting (which we will see in the next section), or by marginalizing a conditional model (which we will see in the g computation section).

For a linear model, the marginal and conditional effect estimate will be equivalent, unless there is an interaction between the exposure and a covariate (in which case they may, but need not be, equivalent). For example, if we consider the relation between quitting smoking and weight gain after adjusting for sex, we can obtain a conditional and marginal risk difference as follows:

```
linear_model <- glm(delta ~ qsmk + sex, data = a, family = gaussian("identity"))
round(coef(linear_model)[2] * 100, 2)
```

```
## qsmk
## 9.24
```

Thus the conditional risk difference is 9.24 excess cases of weight gain among quitters versus nonquitters per 100 participants. We can marginalize over the distribution of sex in the sample as:

```
aa0 <- a
aa0$qsmk <- 0
risk0 <- predict(linear_model, newdata = aa0)

aa1 <- a
aa1$qsmk <- 1
risk1 <- predict(linear_model, newdata = aa1)

round(mean(risk1 - risk0) * 100, 2)
```

```
## [1] 9.24
```

Thus the marginal risk difference is 9.24 excess cases of weight gain among quitters versus nonquitters per 100 participants, the same as the conditional risk difference.

If we interact quitting smoking and sex, the conditional and marginal are no longer equivalent because there are two effects for the conditional model (one for each quit smoking level), but only one for the marginal.

```
linear_model2 <- glm(delta ~ qsmk + sex + qsmk:sex, data = a, family = gaussian("identity"))
# effect among sex = 0
round(coef(linear_model2)[2] * 100, 2)
```

```
## qsmk
##  7.2
```

```
# effect among sex = 1
round((coef(linear_model2)[2] + coef(linear_model2)[4]) * 100, 2)
```

```
##  qsmk
## 11.45
```

The conditional effect among `sex = 0` is 7.2 but among `sex=1` is 11.45. On the other hand, if we marginalize over sex in this conditional model with an interaction, we get a marginal effect estimate of:

```
aa0 <- a
aa0$qsmk <- 0
risk0 <- predict(linear_model2, newdata = aa0)

aa1 <- a
aa1$qsmk <- 1
risk1 <- predict(linear_model2, newdata = aa1)

round(mean(risk1 - risk0) * 100, 2)
```

```
## [1] 9.38
```

This becomes more complicated if the estimand of interest is non-collapsible. For example, if we are interested in the odds ratio, the conditionally adjusted odds ratio is:

```
logit_model <- glm(delta ~ qsmk + sex, data = a, family = binomial("logit"))
round(exp(coef(logit_model)[2]), 2)
```

```
## qsmk
## 1.55
```

but if we marginalize over sex, we get:

```
aa0 <- a
aa0$qsmk <- 0
risk0 <- predict(logit_model, newdata = aa0, type = "response")

aa1 <- a
aa1$qsmk <- 1
risk1 <- predict(logit_model, newdata = aa1, type = "response")

num <- mean(risk1)/(1 - mean(risk1))
den <- mean(risk0)/(1 - mean(risk0))

round(num/den, 2)
```

```
## [1] 1.55
```

The complication here is that, while these numbers are numerically equivalent, the estimands are not mathematically equivalent. In situations where number of events is higher, or when there are many covariates in the model, the conditionally adjusted OR will be further from the null than the marginally adjusted OR (Muller and MacLehose 2014). This can create problems for interpretation, becuase a conditionally estimated OR does not always correspond to the marginal contrast, interpreted as what would be observed if everyone versus no one were exposed.

References

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

Greenland, Sander. 2005. "Collapsibility." In *Encyclopedia of Epidemiologic Methods*, edited by Mitchell H. Gail and Jacques Bénichou. John Wiley & Sons, Ltd.

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

Muller, Clemma J, and Richard F MacLehose. 2014. "Estimating Predicted Probabilities from Logistic Regression: Different Methods Correspond to Different Target Populations." *Int J Epidemiol* 43 (3): 962–70.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Robins, JM. 2001. "Data, Design, and Background Knowledge in Etiologic Inference." *Epidemiol* 12 (3): 313–20.

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

Verma, T, and J Pearl. 1990. "Equivalence and Synthesis of Causal Models." In *Uncertainty in Articial Intelligence 6*, 220–27. Cambridge: Elsevier Science Publishers.