Introduction to Multivariable Statistical Modeling

Al M Best, PhD
Virginia Commonwealth University

Task Force on Design and Analysis in Oral Health Research
Satellite Symposium, AADR

Boston, MA: March 10, 2015

Multivariable statistical modeling from 10,000 feet

- Intro to Multivariate Statistical Modeling
  - Regression
    - Continuous outcomes: Linear ...
    - Dichotomous outcomes: Logistic ...
    - Counts: Poisson ...
    - Time to event: Proportional hazards survival
  - Unified management of confounding and effect modification in statistical models
  - Propensity methods

What is the effect of Treatment on the Outcome?

Superiority, Equivalence, & Non-Inferiority

- Treatment effect: The difference in outcomes between intervention groups
  - Continuous outcomes: difference in means
  - Binary outcomes: odds ratio, relative risk, risk diff.
- Effective: non-zero effect, as compared to control
  - Valid scientific evidence from well controlled studies for the intended use
  - Clinically meaningful endpoint
- Superiority test: a positive effect, as compared to a reference
What is the effect of Treatment on the Outcome?

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Mean</th>
<th>Std Error</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>27</td>
<td>62.3</td>
<td>3.4</td>
<td>55.5</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>25</td>
<td>79.6</td>
<td>3.5</td>
<td>72.5</td>
</tr>
<tr>
<td>Difference</td>
<td>17.3</td>
<td>4.9</td>
<td>7.5</td>
<td>27.1</td>
</tr>
</tbody>
</table>

But …

Correlation, $r = 0.49$

So, … the effect of treatment?

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Mean</th>
<th>Std Error</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>27</td>
<td>53.9</td>
<td>2.5</td>
<td>48.8</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>25</td>
<td>60.4</td>
<td>2.6</td>
<td>55.1</td>
</tr>
<tr>
<td>Difference</td>
<td>6.5</td>
<td>3.7</td>
<td>-0.9</td>
<td>13.8</td>
</tr>
</tbody>
</table>

Adjusted difference = 12.7
Difference due to confounding = 4.6
Crude difference = 17.3
So, … the effect of treatment:

Adjusted difference = 12.7

Modeling the effects of predictors on outcomes

- What we just did was build an ANCOVA model
- The model said:
  - Outcome = Expected_Value + Noise
- Where: the form of the expected value was
  - Expected = Intercept + Treatment_Effect + Slope*Prognostic
  - The effect of the Prognostic Index is the Slope
- And the assumption for noise was
  - Noise = Raw data - Prediction
  - Noise is Normally distributed with mean=0 and constant variance
- An alternate form is: Data = Signal + Noise

Modeling: Core concepts

- Modeling the statistical relationships of predictors to the outcome
- Represents the magnitude of one phenomenon as a mathematical function of “predictors” and random variation
  - The phenomenon modeled may be continuous, e.g. HDL cholesterol, or categorical, e.g. survival or death
  - An overall model consists of two submodels, for signal and noise
    - Signal model: a mathematical form of an equation to predict some aspect of the predicted phenomenon, e.g. mean cholesterol or probability of death
    - Noise model: a probability law that describes, on a group basis, how individuals vary from what the equation predicts
- The prediction equation typically includes
  - the exposure variable of interest, e.g., treatment regimen
  - other predictors of potential importance
  - potential confounders
- Predictors may be
  - continuous
  - categorical, with two or more ordered or unordered categories
  - combinations of these.
- Data are used to estimate sizes of predictor effects and random variation
- The predictor effects are estimated adjusted for other predictors by holding them constant. This mimics matching.
  - The equation relating the outcome to any single predictor, holding others constant, may be linear or of many other forms
### Classes of statistical models

<table>
<thead>
<tr>
<th>Model Class</th>
<th>Response Variable Type</th>
<th>Response Transformation</th>
<th>Noise Model</th>
<th>(Transformed) Parameter Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>Continuous</td>
<td>None</td>
<td>Gaussian (&quot;normal&quot;)</td>
<td>Difference in means.</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Dichotomous</td>
<td>Logit</td>
<td>Binomial</td>
<td>Odds-ratio (OR)</td>
</tr>
<tr>
<td>Poisson regression</td>
<td>Count</td>
<td>Logarithm</td>
<td>Poisson</td>
<td>Incidence Density (Rate, Hazard) Ratio</td>
</tr>
</tbody>
</table>

"Linear predictor" = $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 + \delta x_1 z_2$
Multivariable modeling: multiple linear regression

- Models the mean of a quantitative outcome as a function of the values of predictor variables
- Assumes independent observations with approximately Gaussian (normal) distributions
- Contrary to the name, the models need not be linear in the predictors. But they must be linear in the coefficients that multiply them
- Example: \( \mu(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 \)
  - where \( x_1 \) is the value of the exposure of interest, \( x_2 \) and \( x_3 \) are values of other variables that may biologically affect \( y \), and \( z_1 \) and \( z_2 \) are possible confounders
  - The \( x \)'s and \( z \)'s may be continuous, or values of 0 or 1 "dummy variables" representing categories of qualitative variables
- The gold Greek coefficients are estimated from the observed data, based on some measure of how well the model predicts

Multivariable modeling: multiple linear regression

- Thus, \( \beta_1 \) represents the predicted change in the mean value of \( y \) associated with an increase of one unit in the variable represented by \( x_1 \)
  - with the variables represented by the other \( x \)'s and \( z \)'s held fixed
- The parameters of different multiple linear regression models all represent differences in mean values of the outcome predicted by variations in characteristics underlying the prediction
- This type of model accommodates effect modification through product “interaction” terms, e.g., \( \delta x_1 z_2 \), which allow the effect of a change of one unit in \( x_1 \) on \( \mu(y) \) to vary with the value of \( z_2 \)
  \( \mu(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 + \delta x_1 z_2 \)
  - \( \delta \) is thus a difference of differences: how the effect on \( y \) of a one unit increase in \( x_1 \) is itself modified by a one unit increase in \( z_2 \)

Multivariable modeling: multiple linear regression

- Representing categorical predictors, e.g., 3 categories
  - "Dummy variable" = "Indicator variable" approach
    - Two variables, \( z_1 \) and \( z_2 \)
    - \( \gamma_1 z_1 + \gamma_2 z_2 \) is the contribution of the level of the categorical predictor
      - Category 1: \( z_1 = 1 \), \( z_2 = 0 \), contribution = \( \gamma_1 \)
      - Category 2: \( z_1 = 0 \), \( z_2 = 1 \), contribution = \( \gamma_2 \)
      - Category 3: \( z_1 = 0 \), \( z_2 = 0 \), contribution = 0
    - \( \gamma_1 \) represents the category 1 vs. category 3 difference
    - \( \gamma_2 \) represents the category 2 vs. category 3 difference
  - Deviation approach
    - Two variables, \( z_1 \) and \( z_2 \)
    - \( \gamma_1 z_1 + \gamma_2 z_2 \) is the contribution of the level of the categorical predictor
      - Category 1: \( z_1 = +1 \), \( z_2 = 0 \), contribution = \( \gamma_1 \)
      - Category 2: \( z_1 = 0 \), \( z_2 = +1 \), contribution = \( \gamma_2 \)
      - Category 3: \( z_1 = -1 \), \( z_2 = -1 \), contribution = \( - (\gamma_1 + \gamma_2) \)
    - Average contribution of the three levels = 0.
    - \( \gamma_1 \) represents the category 1 vs. average of the categories’ contributions
    - \( \gamma_2 \) represents the category 2 vs. average of the categories’ contributions
Multivariable modeling: multiple logistic regression

- Models the probability of a dichotomous outcome as a function of the values of predictor variables
- Assumes independent binomially distributed observations
- Same predictor functions as multiple linear regression
  - but instead of predicting means, logistic regression predicts logits, or log(odds), of probabilities of an outcome
  - Note: Odds = \( \Omega = \pi / (1 - \pi) \)
- Example: logit (\( \pi \)) = \( \log [\text{odds}(\pi)] \) = \( \log \left[ \frac{\pi}{1 - \pi} \right] \) = \( \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 \)
  - where \( \pi \) is the probability, e.g., of successful tooth restoration, and the other symbols are all as defined above for multiple linear regression
  - example: \( \beta_1 \) is the effect of \( x_1 \) ...

Multivariable modeling: multiple logistic regression

- By the same reasoning as for multiple linear regression, here
  - \( \beta_1 \) represents the predicted change in the logit of \( \pi \) associated with an increase of one unit in the variable represented by \( x_1 \), with the variables represented by the other \( x \)'s and \( z \)'s held fixed
- Thus, parameters of different multiple logistic regression models all represent
  - differences in logits of predicted probabilities, associated with the various levels of the predictor
  - differences in logits of 0 imply “no effect”
- So, since a difference in logits is the logarithm of an odds ratio:
  - “no effect” = Odds ratio = 1
  - “no effect” = Log(odds-ratio = 1) = 0
Multivariable modeling: multiple logistic regression

- Since a one unit increase may be very small indeed, e.g. 1 mmHg of SBP or 1 mg/deciliter of serum cholesterol
  - one can’t interpret such coefficients without considering the units in which the predictors are expressed
- Often it’s useful to translate the coefficients into clinically meaningful predictors.
  - Thus, the odds ratio relating SBP to stroke is much better expressed per 10 mmHg increase in SBP,
  - that is, as exp(10*β1) when x1 is in single mmHg, or
  - by defining x1 in units of 10 mmHg, than in units of 1 mmHg

Example

Multivariable modeling: multiple logistic regression

- Note in the example: the lines were NOT parallel.
  That is, the effect of RVP depended upon tooth type
- Effect modification may be incorporated into multiple logistic regression equations in much the same way as into linear regression equations, using an interaction term

\[ \log \left[ \frac{\pi}{1-\pi} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 + \delta x_1 z_2, \]
  - where exp(δ) may be interpreted as the amount by which the odds ratio relating the outcome to x1 is multiplied with a one unit increase in z_2.
- Thus, exp(δ) is a ratio of odds ratios

Multivariable modeling: multiple logistic regression

- Note that calculating an odds ratio between an exposure and disease as either

\[ \frac{\text{odds(Disease)|exposure}}{\text{odds(Disease)|no exposure}} \]

or

\[ \frac{\text{odds(exposure|Disease)}}{\text{odds(exposure|no Disease)}} \]
Multivariable modeling: multiple logistic regression

- Note that calculating an odds ratio between an exposure and disease as either
  \[ \frac{P(D|E)}{P(D|\neg E)} \]
  or
  \[ \frac{P(\neg D|E)}{P(\neg D|\neg E)} \]
  will give the same answer

- Note that calculating a risk ratio (relative risk) as either
  \[ \frac{P(D|E)}{P(D|\neg E)} \]
  or
  \[ \frac{P(\neg D|E)}{P(\neg D|\neg E)} \]
  will NOT give the same answer

Since multiple logistic regression is based on odds ratios, we can calculate these odds ratios linking exposure to disease using these models with
- cohort data, in which exposure is used to predict disease, or
- case-control data, in which disease and other variables are used to “predict” past exposure

In principle, these should give the same answers

OR’s from models for case-control data may be used to estimate risk ratios or rate ratios, when the disease is sufficiently uncommon

Multiple logistic regression model fitting comes in several forms for different types of data:
- unconditional for independent samples
- conditional for matched samples,
- conditional exact for small samples,
- hierarchical/mixed for clustered or longitudinal data

Using the wrong method of fitting can give badly biased answers

Repeated-measures logistic regression can/should be used to analyze multiple sites/teeth within the mouth

Although it can be complex, multiple logistic regression allows investigation of and simultaneous adjustment for many more variables than possible with stratification methods

As with any model, always check whether your assumptions approximate to reality
Multivariable modeling: Poisson regression

- Models the incidence density (hazard) rate of occurrence of new events as a function of the values of predictor variables
- Assumes independent observations with Poisson distributions
  - This is a "memoryless" process. Events occur at a given rate for each person, independently of events in other people, and of the number of prior events in the same person!
- Uses the same class of predictor functions as multiple linear and logistic regression above, but instead of predicting mean values or logits, these serve to predict incidence rates of new events
- Example: \( \log \lambda = \alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 z_1 + \gamma_2 z_2 + \log t \)
  where \( \lambda \) is the incidence density e.g., of implant failure, \( t \) is the time over which events are counted, and other symbols are as previously, and ...

Multivariable modeling: Poisson regression

- Parameters in Poisson regression models represent differences in log incidence densities
  - That is, \( \exp(\beta_1) \) represents an IDR = incidence density ratio (hazard, rate) ratio associated with change in the level of a predictor
- This is very useful in cohort studies in which the rate of damaging outcome events is roughly constant over time for subjects with a given combination of predictor variables
- What about when this rate is not constant over time, as for instance in a surgical clinical trial?

Example


Multivariable Modeling

- Linear regression
  - \( \mu = E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 \)
  - Parameters are differences in mean responses
- Logistic regression
  - \( \logit(\pi) = \log \left( \text{odds}(\pi) \right) = \log \left( \frac{\pi}{1-\pi} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 \)
  - Parameters are logarithms of odds-ratios (log OR)
- Poisson regression
  - \( \log \lambda = \log \text{ID} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 z_1 + \gamma_2 z_2 \)
  - Parameters are logarithms of incidence density ratios (log RR)
- Multivariable model for survival probabilities???
Survival modeling

- Modeling time-to-event data (AKA survival analysis)
  - In survival analysis, “the event” is often death
  - In dental studies, failure is usually the event of interest
- It’s of interest to model how long until the event occurs, usually comparing groups to determine characteristics related to failure (or success)
  - Record the time of each failure
  - Record the time of the last follow-up: a censored observation
- Example: How long do two types of restorations last in the posterior teeth of children? Compare amalgam and resin-based composite...

The New England Children’s Amalgam Trial


Results: adjusted for age and number of restorations in the mouth

Results: Repairs
Proportional hazards models

- Proportional hazards model:
  \[ h(t) = h_0(t) \left( \alpha + \sum \beta_i x_i + \sum \gamma_i z_i + \gamma_1 z_1 + \gamma_2 z_2 \right) \]

- With parameters representing the log of incidence density ratios
- The model assumes these incidence rates change across time but that the ratios do not.

This isn’t always appropriate!

Highly non-proportional (and changing) hazards

Unified management of confounding and effect modification in statistical models

- Linear predictor with confounders
  \[ \alpha + \sum \beta_i x_i + \sum \gamma_i z_i \]
  Treatment \((x_i)\) & other predictors \((x_2, x_3)\)

- Linear predictor with effect modifier
  \[ \alpha + \sum \beta_i x_i + \sum \gamma_i x_i z_i \]
  Treatment \((x_i)\) & other covariates \((x_2, z_1)\)
  \(z_1\) and \(x_1\) modify each others’ effects

Effect Modification

**Confounder control**

- Stratified analyses, multivariable outcome modeling, and propensity modeling, do this analytically, ex post facto

```
Confounder  Exposure  Outcome
```

- BUT: fancy analysis is no substitute for a well designed study

**Propensity analysis**

- Instead of modeling the outcome, as do all the models above, model the group (e.g., treating) assignment – that is, the “propensity” of assignment to one group or the other
  - Develop a logistic regression model for the probabilities of individuals being in each group,
  - and then analyze the outcome by either
    - stratify analysis by ranges of these propensity scores, or
    - match members of the groups being compared on their propensity scores

**Coda**

- Everyone does a little bit of lawyering, doctoring, nursing, social work, psychology, engineering, and etc. Most people recognize their limits in these areas and when their limits are reached they find someone who specializes in these areas.

- Statistics may be somewhat different in that we write “easy to use” “user friendly” software and encourage scientists to do much of their own analysis.
  - Fine; as long as we recognize our limits and seek help when needed.

- Moral of the story:
  - Find collaborative colleagues. Specifically, find a statistician who will listen to you and help you communicate your story.
  - We enjoy analysis.

**Thanks!**

... glad to help
ALBest@VCU.edu

Acknowledgement: I’m grateful to Peter B Imrey, PhD for sharing his previous version of this talk.