BIFX 553 - Discussion 2

Randy Johnson

January 26, 2017

Linear Regression

Recommended reading: Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (Springer Series in Statistics) 2nd ed. 2015 Edition by Frank Harrell

Why use Linear Models?

- Hypothesis Testing: Building a model to test if an exposure is statistically associated with an outcome. This can extend to multiple hypotheses, but care must be taken with interpretation of the significance measures.
- Estimation: Building a model to estimate the effect of an exposure on an outcome.
- Prediction: Building a model using training data to predict a future outcome.

Planning

If you fail to plan, you should plan to fail.

Convenience samples are often what we are left with when we fail to properly plan a study. Also, scientists are always learning and are often tempted to introduce new hypotheses into a study after the data has been collected. This isn't neccessarily a bad thing, but it can lead to problems in the analysis. Exercise caution. The resulting analyses are often plagued by difficulties including:

- The sample size is too small to answer the questions the study leaders(s) would like to ask – i.e. they are unable to test the hypotheses set forth at the beginning of the study with any reasonable level of statistical power. This often arises out of a lack of focus, or a desire to test every possible hypothesis we can think of.
- When collaborators are brought into the study, it is often discovered that important exposures have not been collected (sometimes the most important exposures are missing or are hopelessly biased).
- The study subjects making up the sample suffer from a lack of definition, they don't represent the population we planned to study (e.g. because the primary outcome has been biased or because our sample doesn't include enough variation from the population under study). Poor representation of the population we would like to study is often a function of poor selection of study sites from which to collect the data.
- Many / important variables have large amounts of missing data.
- Missing data in the sample are not missing at random.
- Many / important variables are poorly defined, leading to heterogeneity in what is actually gathered at different study sites.
- A lack of a data validation plan can lead to varying quality of data collection at different sites. This can result in some sites presenting with an excess of issues described above.

Bias: The difference between an estimator's expected value and it's true value is non-zero. Alternately, bias can be thought of as resulting in a systemic difference between a statistic and the true population parameter.

- Confounding: 1) Confounding occurs when there exists a back-door pathway between the exposure and the outcome. 2) Confounding occurs when the frequency of disease in the ovserved comparison (unexposed) population is different from the frequency of disease in the exposed population, had they been unexposed.
  - Confounding variables must cause disease
  - Confounding variables must be associated with the exposure (e.g. correlated)
  - Condounding variables may not be directly affected by the exposure

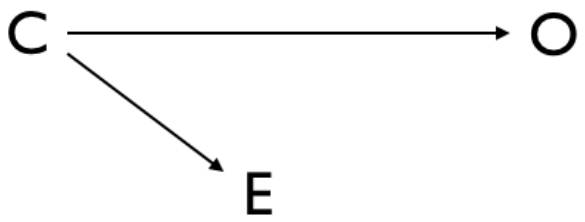


Figure 1: Back-door pathway inducing non-existant causitive association from the effect, E, to the outcome O, through a confounder, C.

- Misclassification / Information Bias: Misclassification or information bias occurs when the disease status influences how the exposure is measured. Examples include:
  - Recall bias
  - Interviewer bias
  - Data collection bias
- Selection Bias: Selection bias occurs when the measure of association in the study sample is different than in the source population. This can occur when selection for inclusion into the study is influenced by an exposure and by a cause of the outcome (in cohort studies) or by the outcome itself (in case/control studies). In the graphical representation shown below, the exposure, E, causes a factor, F, which causes selection, C. Disease, D, also is a basis for selection. An example of this is when exposure causes hospitalization.

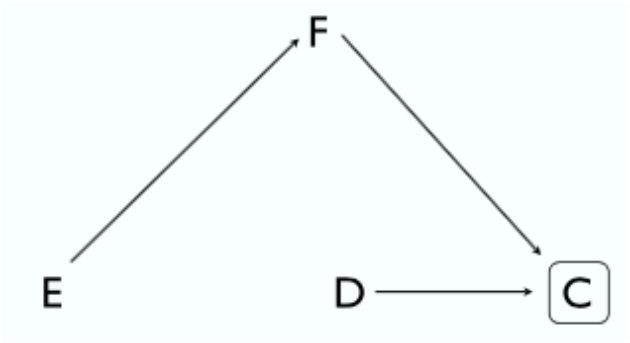


Figure 2: Selection bias is introduced when E causes F, which casues selection, C, while D also causes selection.

- Collider-Stratification Bias: This occurs when conditioning on a collider variable, C, opens a back-door pathway through some unmeasured factor(s), U, inducing a false association between and effect, E, and the outcome, O. There may or may not be a causitive relationship between C and O. For an example of

this where the effect of a mothers smoking is paradoxically associated with a protective effect on infant mortality, see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2743120/pdf/nihms-135847.pdf.

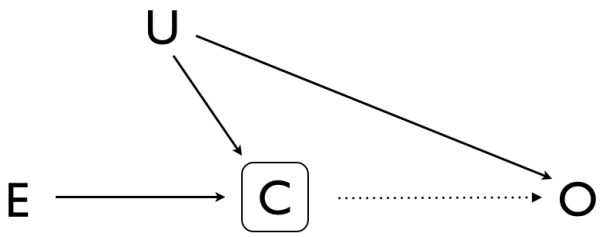


Figure 3:

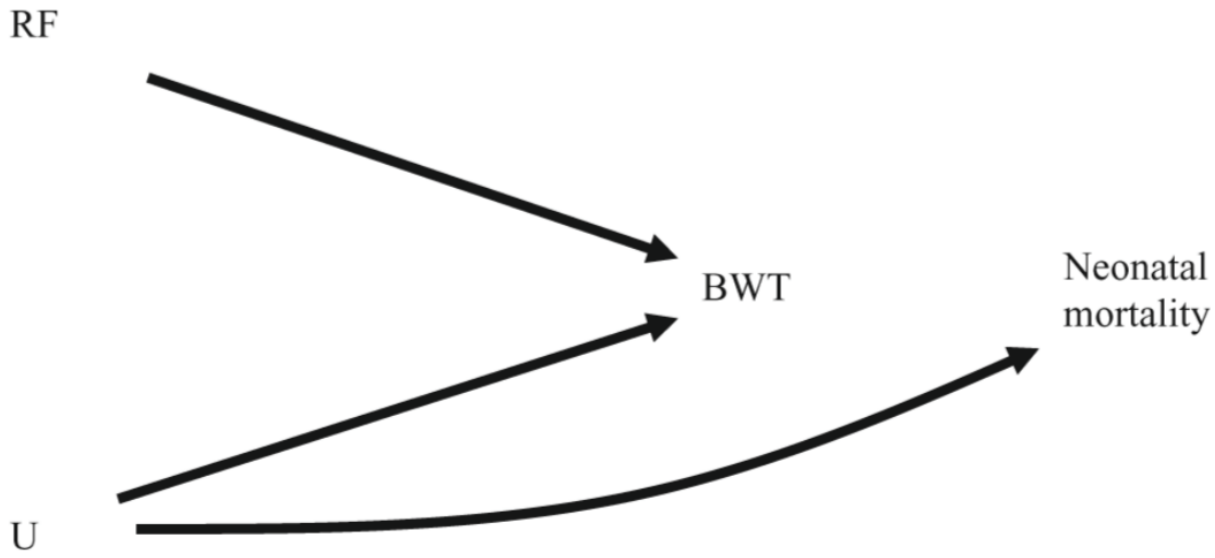


Figure 4: Whitcomb et al. Figure 1a: RF = risk factor, BWT = birth weight, U = unmeasured factors.

In general, there are a number of predictor variables that we should always consider at potential confounders. The predictors on Harrell's list will rarely all be needed for any one study, but he provides a good list to consider: - age - sex - acute clinical stability - principal diagnosis - severity of principal diagnosis - extent and severity of comorbidities - physical functional status - psychological, cognitive and psychosocial functioning - cultural, ethnic and socioeconomic attributes and behaviors - health status and quality of life - patient attitudes and preferences for outcomes

Group Activity

Building our own epidemiological model: Breast Cancer - predict the number of lymphnodes testing positive for cancer during primary tumor removal.

Notation and Formulation

```
# Read in German Breast Cancer dataset
# url('http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/gbsg.sav') %>%
#   load()
```

```
load('gbsg.sav') # in case the above doesn't work?

# some decidedly odd formatting here
str(gbsg)
```

```
## 'data.frame':    686 obs. of  12 variables:
##  $ id    :Class 'labelled'  atomic [1:686] 132 1575 1140 769 130 1642 475 973 569 1180 ...
##   .. ..- attr(*, "label")= chr ""
##  $ age   :Class 'labelled'  atomic [1:686] 49 55 56 45 65 48 48 37 67 45 ...
##   .. ..- attr(*, "label")= chr ""
##  $ meno  : Factor w/ 2 levels "postmenopausal",..: 2 1 1 2 1 2 2 2 1 2 ...
##   ..- attr(*, "label")= chr ""
##  $ size  :Class 'labelled'  atomic [1:686] 18 20 40 25 30 52 21 20 20 30 ...
##   .. ..- attr(*, "label")= chr ""
##  $ grade :Class 'labelled'  atomic [1:686] 2 3 3 3 2 2 3 2 2 2 ...
##   .. ..- attr(*, "label")= chr "1:gradd1=0 gradd2=0,2:1 0,3:1,1"
##  $ nodes :Class 'labelled'  atomic [1:686] 2 16 3 1 5 11 8 9 1 1 ...
##   .. ..- attr(*, "label")= chr ""
##  $ enodes:Class 'labelled'  atomic [1:686] 0.787 0.147 0.698 0.887 0.549 ...
##   .. ..- attr(*, "label")= chr ""
##  $ pgr   :Class 'labelled'  atomic [1:686] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..- attr(*, "label")= chr ""
##  $ er    :Class 'labelled'  atomic [1:686] 0 0 0 4 36 0 0 0 0 0 ...
##   .. ..- attr(*, "label")= chr ""
##  $ hormon: Factor w/ 2 levels "had tamoxifen",..: 2 2 2 2 1 2 2 1 1 2 ...
##   ..- attr(*, "label")= chr ""
##  $ d     :Class 'labelled'  atomic [1:686] 0 1 0 0 0 1 1 0 1 1 ...
##   .. ..- attr(*, "label")= chr "censrec"
##  $ t     :Class 'labelled'  atomic [1:686] 5.032 1.103 4.389 0.485 5.079 ...
##   .. ..- attr(*, "label")= chr "rectime/365.25"
```

```
# Convert funky formatting of data.frame to a tibble
gbsg <- as_data_frame(gbsg) %>%
  mutate(id = as.integer(id),
         age = as.integer(age),
         meno = as.character(meno),
         size = as.integer(size),
         grade = as.integer(grade),
         nodes = as.integer(nodes),
         enodes = as.double(enodes),
         pgr = as.integer(pgr),
         er = as.integer(er),
         hormon = as.character(hormon),
         d = as.integer(d),
         t = as.double(t))
save(gbsg, file = 'gbsg.RData')
```

Let's write our model using the standard notation:

$$nodes_i = \beta_0 + age_i * \beta_1 + size_i * \beta_2 + grade_i * \beta_3 + \varepsilon_i$$

Interpretation of Model Parameters

Continuous

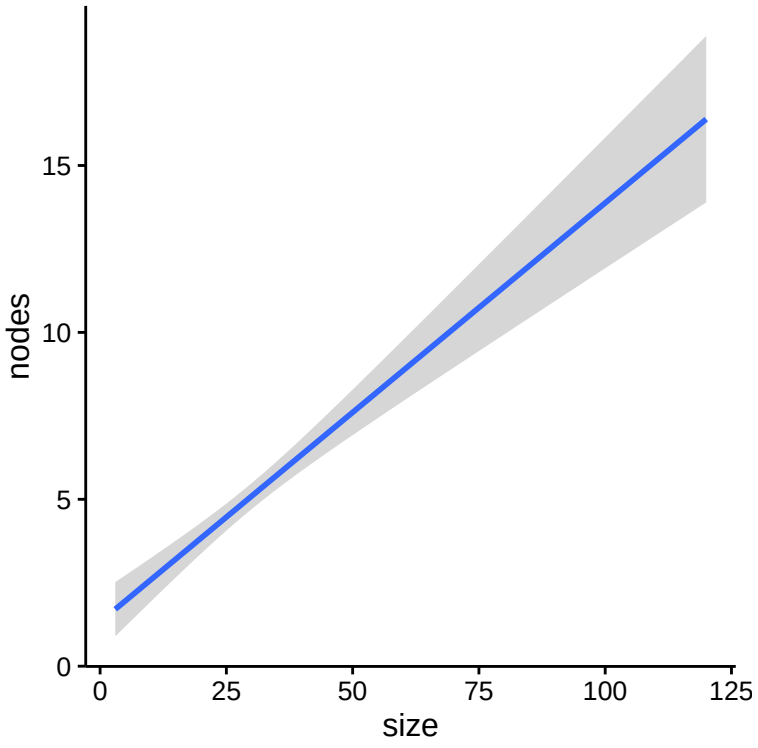
Continuous predictors describe the slope of the regression line.

$$\mathrm{E}(nodes_i) = \beta_0 + \beta_1 * size_i$$

```
lm(nodes ~ size, data = gbsg) %>%
  tidy()
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept) 1.3294634 0.4513802  2.945329 3.335664e-03
## 2        size 0.1254964 0.0138360  9.070286 1.242502e-18
```

```
ggplot(gbsg) +
  geom_smooth(aes(size, nodes), method = 'lm')
```



Categorical

Categorical predictors modify only the intercept of the regression line. In this example, the expected value of the response variable, *nodes*, is β_0 for menopausal women and $\beta_0 + \beta_1$ for pre-menopausal women. Thus, β_2 is the *difference* in *nodes* between menopausal and pre-menopausal women.

$$\mathrm{E}(nodes_i) = \beta_0 + \beta_2 * meno_i$$

```
lm(nodes ~ meno, data = gbsg) %>%
  tidy()
```

```
##                term   estimate std.error  statistic      p.value
## 1       (Intercept)  5.1186869 0.2752805 18.5944379 9.004383e-63
## 2 menopremenopausal -0.2566179 0.4233880 -0.6061057 5.446456e-01
```

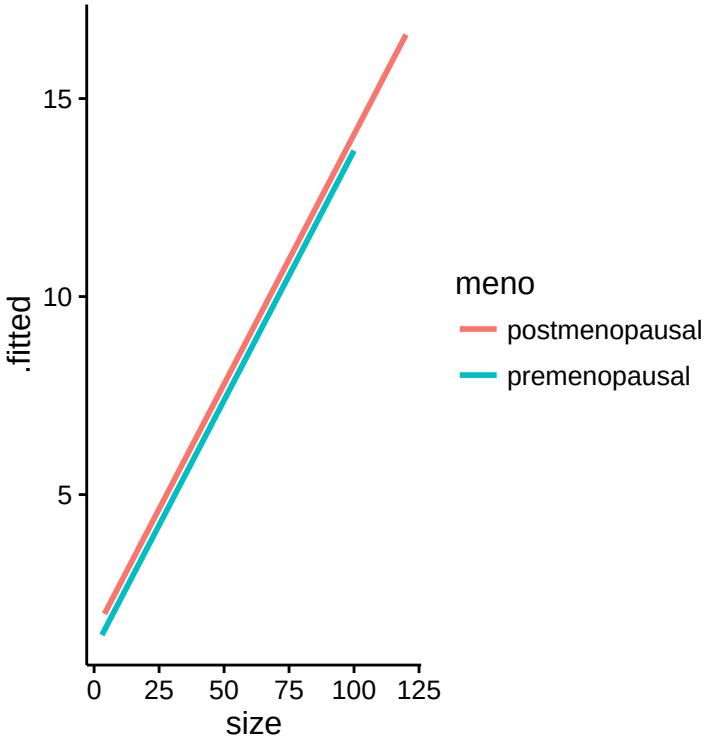
In the case of a multiple linear regression, the slopes of the regression lines are identical for each group.

$$\mathrm{E}(nodes_i) = \beta_0 + \beta_1 * size_i + \beta_2 * meno_i$$

```
tmp_lm <- lm(nodes ~ size + meno, data = gbsg)
tidy(tmp_lm)
```

```
##                term   estimate  std.error statistic      p.value
## 1       (Intercept)  1.4852609 0.47634182  3.118057 1.896946e-03
## 2              size  0.1260921 0.01384775  9.105601 9.329704e-19
## 3 menopremenopausal -0.4098674 0.40046178 -1.023487 3.064400e-01
```

```
ggplot(data = augment(tmp_lm), aes(size, .fitted, color = meno)) +
  geom_smooth(se = FALSE)
```



Interactions

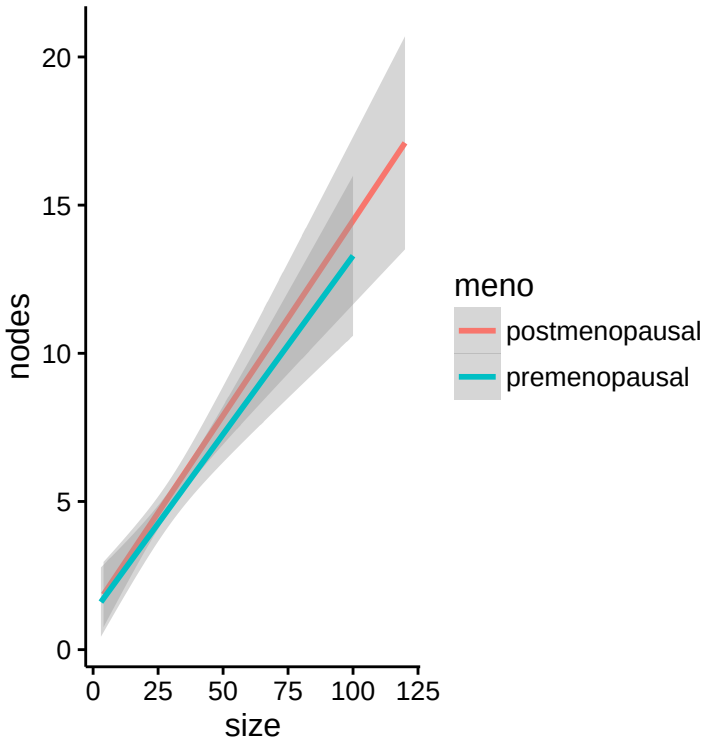
Including an interaction term between *size* and *meno* tells the model to allow the slopes of the regression lines to differ for the two different *meno* groups. In this model, the *difference* in slope for the two menopause groups is β_3.

$$\mathrm{E}(nodes_i) = \beta_0 + \beta_1 * size_i + \beta_2 * meno_i + \beta_3 * (size_i * meno_i)$$

```
lm(nodes ~ size * meno, data = gbsg) %>%
  tidy()
```

```
##                     term    estimate  std.error   statistic      p.value
## 1            (Intercept)  1.33181558 0.61543657  2.16401760 3.080944e-02
## 2                   size  0.13141714 0.01935326  6.79044119 2.434641e-11
## 3      menopremenopausal -0.08822946 0.90915321 -0.09704576 9.227186e-01
## 4 size:menopremenopausal -0.01092570 0.02772146 -0.39412412 6.936126e-01
```

```
ggplot(gbsg, aes(size, nodes, color = meno)) +
  geom_smooth(method = 'lm')
```



An asside on Methods sections

The following subsections may be included in your methods section as apprpriate. Sometimes you will see different names for these. For example the Materials subsection might be called "Genotyping" if it deals only with the genotyping methods used, or it may be split into a couple of subsections if it makes better sense and aids readability.

- Participants: A description of the study participants, what population they come from, how they were sampled and what data were collected. The experimental design may fit well here, or it might be broken out into its own subsection.
- Materials: Description of any materials and/or tools you used to collect your data. Specifics on what assays and equipment were used go in this section.
- Statistical Analysis: Description of statistical tests used, statistical correction for known biases/errors, validation of assuptions, etc. . .

When writing a methods section, it is good to keep the following things in mind:

- This section should be written in the past tense
- Check for consistency with the rest of the paper. Make sure that if you mention all statistical test and data sources. Conversely, anything that isn't relevant elsewhere in the paper doesn't need to be addressed in the methods section.

- You want to include enough information that anyone not familiar with your study could replicate what you did, but you also want to keep this section short. This is why methods sections tend to be dense.
- It is good practice to put your work down for a day or two and come back to it with fresh eyes. Some people find it helpful to read out loud to themselves. Doing this will help you find errors and increase readability.

An asside on Results sections

SanFrancisco Edit has a really nice tip sheet on how to write a results section (see http://www.sfedit.net/results.pdf). Overall, you should be sure that the results section tells a neutral, cohesive story about your study results. You should *not* make any interpretations or jumpt to any conclusions in this section! The purpose is to lay out the facts. Care should be taken to arrange the presentation of tables and figures in a way that lets the data tell their own story.

The methods used to derive every result in this setion should be contained in the methods section. You also want to arrange the methods and results sections in such a way that the methods used for a particular result will be as easy to identify as possible.

Homework

- Write a methods section describing the gbsg data set and the analysis we performed in class. Include enough information that the reader would be able to know what samples to collect and how to analyze them to reconstruct/replicate your results. **Limit: 400 words**.
- Write a results section with table(s) and figure(s) describing the results of the analysis.