Concatenating ALL notes

Biostatistics, **Lecture #1** Notes, 21 September 2015 (Brian Healy, PhD)

1.1: Introduction: expectations, goals, etc.

1.2: Overview of Syllabus & Introduction to STATA

BH says that the STATA help manual is about the best stats book there is

workspace very like MATLAB’s: command window, variable window, command summary

1.3: Data types, summarizing data, interpret scatter plots

think of stats/analysis BEFORE doing expt.

study only as good as DESIGN:

What are you trying to learn? How will you prove this? (Think like a reviewer.)

Sample selection: Who are you going to study?

Data collection: What should be collected

Data analysis:

Was there any effect? What does it all mean? To whom do results apply?

Key concept: Population vs. Sample

We want our sample to be representative and random. (Lecture #3)

Key concept: Description (e.g. std. dev.) vs. Inference (e.g. SEM; 95% CI)

1.4/1.5: Data description

longitudinal study of MS: 252 patients; quality of life survey

variable: something measured in all people in our sample

variable types:

continuous (age, expression level)

dichotomous/binary (dead/alive)

nominal/categorical (race, group membership)

ordinal (mild/moderate/severe)

count (# of lesions, # of children)

time-to-event (time to death)

All data has some kind of DISTRIBUTION that we would like to describe

Ways to summarize data: 1) summary statistics, 2) graphics

summary: mean, median, std. dev., variance, interquartile range (diff. btw. 75%ile and 25%ile)

STATA command: ‘summarize’ var\_name

Histogram: ‘histogram’ var\_name

skewness: mean pulled towards rare, high/low observation

Box plot: median, 25th/75th %iles, max/min

ALWAYS look at your data first, before you do analyses

easier to detect errors in data entry (‘clean’ data)

**Sampling variability**: differences between random samples from a population

confidence interval: ‘ci’ var\_name

1.6: summarizing data

**dichotomous varia**bles: summarize as a proportion: ‘tabulate’ var\_name

surgery joke: interp. a dichotomous variable as a continuous one (“We’ll only do 72% of it, since it has been reported that 28% of all surgery is unnecessary.”)

aside: political polls (51% of people support Hillary Clinton); margin of error (= 95% CI)

Why is there so much variability from poll to poll?

1. Chance: sampling variability
2. Bias (Has the population being sampled changed? e.g. entire population vs. likely voters)
3. How you gather your sample affects your conclusions

nominal/categorical: summarize with a table: ‘tabulate’ var\_name

**ordinal variables** (most challenging to deal with)

1. order but no magnitude
2. order and info about magnitude (fatigue score)

**count variables**

1. very common in epidemiology
2. mean is usually an interesting statistic
3. histogram useful (= bar chart, really)

**time to event variables**

1. survival time (median; Kaplan-Meier curve)

Summary: Each data type has specific summary stats and graphics.

1.8: Correlation

Descriptions vs. Comparisons: for comparison, there is an explanatory variable (e.g. male vs. female) and an outcome variable (e.g. height)

Contingency table: relationship between 2 dichotomous variables (lecture 6)

Correlation: relationship between two continuous variables; use a scatterplot (explanatory var. on x-axis; outcome var. on y-axis)

BH mantra: “The best statistician you know is . . . you.” (b/c you know your data better than anyone else)

Biostatistics, **Lecture #2** Notes, 28 September 2015 (Brian Healy, PhD)

2.1 Ways of saving your work from a STATA session:

1. copy output and paste into Word doc
   1. STATA helpful hint when copy/pasting, convert font to Courier New, 11 pt.
2. Can also open up a ‘log’ file in STATA, and it will save everything you do. (File -> Log -> Begin)
3. Command window keeps track of all commands, so you can copy and paste this into a Word doc.
4. ‘do’ file, which is like a MATLAB script (m-file)

2.2 Basic probability:

STATA commands:

* display binomial(n,x,p)
* display normal(z)

def’n of probability: relative freq. of a given outcome over an infinitely large # of trials

P(A) = prob. of event A

P(not A) = complement = 1 – P(A)

P(A and B) = intersection = P(A,B) = prob. of BOTH A and B

P(A and/or B) = union = prob. of A or B or both

Venn diagrams

2.3 Conditional probability

P(A/B) = prob. of event A given that event B has occurred

conditional probability RESTRICTS the sample space

e.g. Given that a person has blonde hair, what is the prob. the s/he has blue eyes?

P(A/B) = P(A,B) / P(B)

both sensitivity and p-values are examples of conditional probabilities (i.e. p-value is prob. of as or more extreme than data, GIVEN that H0 is true)

**mutually exclusive**: 2 events A and B that cannot occur at the same time: P(A/B) = 0

when events are mutually exclusive, we sum the probs.:

P(A and/or B) = P(A)+ P(B)

**exhaustive**: a set of events that covers all possible events

**exhaustive and mutually exclusive**: prob. sums to one

2.4 Independent events:

occurrence of B does not affect the prob. of A: P(A/B) = P(A); P(A,B) = P(A)\*P(B)

diagnostic tests:

sensitivity: prob. of a positive test given that the patient HAS the disease: P(T+/D+)

(complement = false negative); P(T+/D+) + P(T-/D+) = 1

specificity: prob. of a neg. test given that the patient does NOT have the disease: P(T-/D-)

(complement = false positive); P(T+/D-) + P(T-/D-) = 1

2.5 Probability distributions

1. binomial distribution: distribution for the number of successes, k, (with prob. of success, p) in a specific # of trials (n); 2 options for each trial, with each trial independent
   1. discrete random variable
   2. P(X=x) = (nchoosex) \* p^x \* (1-p)^(n-x)
   3. expectation = mean = n\*p
   4. variance = np(1-p)
2. normal distribution: one of most important probability distributions in biostatistics
   1. continuous random variable; range(-inf,+inf)
   2. two parameters: mean, mu and variance, sigma^2
   3. symmetric
   4. typically want to know prob. of being between two values or the prob. of being less than/great than a given value
   5. can always convert to standard normal by z-transform: X – mu / s.d
   6. i.e. standard normal has mean = 0 and s.d. = 1
   7. +/- 1 s.d. = 68% of area under Nl. (P(-1 <= Z <= 1) = 0.68
   8. P(Z <= 1) = 0.16; P(Z >= 1) = 0.16
   9. +/- ~2 s.d. = 95%; P(-1.96 <= Z <= 1.96) = 0.95
   10. lower tail: P(Z <= z)
   11. **key point**: If we know that something follows a normal distribution, we can calculate the probabilities of specific events
3. other normal distributions: any normally distributed variable can be converted to standard normal (i.e. z-transform)
   1. P(Z <= z): display normal(z)
   2. P(Z > z) = 1 – P(Z <= z)
   3. “normal” range is often defined as the middle 95% of the distribution
   4. convert to X from Z by inverse z-transform (z\*s.d. + mu)
4. Distribution of the mean (standard error of the mean)
   1. means of multiple observations: What is distribution of the mean of a group of observations?
   2. if we take a bunch of samples of size, n, and take the mean, this distribution will always be normal (central limit theorem) with mean = sample mean and s.d. (sem) of sampleSD/sqrt(n)
5. Central limit theorem: (by simulation)
   1. <http://onlinestatbook.com/stat_sim/sampling_dist/index.html>
   2. better URL: <http://onlinestatbook.com/stat_sim/sampling_dist/>
   3. looks like an entire on-line statistics book: <http://onlinestatbook.com/rvls.html>

Lecture #3: Hypothesis Testing viewed 07 October 2015

3.1 “Class data set”: think about how you are going to collect your data; run a pilot to see what goofy answers you get so that you can clarify your questions

3.2: Review: p-value; hypothesis tests (1- and 2-sample t-test); confidence intervals

define “normal range” to be middle 95% of the data: -1.96 <= X <= +1.96 (for std. nl. dist.)

3.3: Most of the time, we do not know the parameters of the distribution, so we have to do inference: estimating population parameters based on a sample

example from multiple sclerosis (MS): Is the avg. cognitive function in patients with MS different from healthy controls? prob.: How do you measure cognitive function? “symbol digit modalities test” (SDMT)

We want our sample to be both 1) representative of the population and 2) random

In nl. ctrls., mean SDMT is 55; MS sample mean is 54.5

Reasons for differences between groups:

1. actual effect; i.e. a real difference
2. chance: sampling error
3. bias: something in the design that led to the result (e.g. give the healthy people the drug and the sick people the placebo)
4. confounding

Statistical tests are designed to determine if the observed difference between groups was likely due to chance.

3.4: Chance experiment: Heads I win $1; tails you win $1 (intuition about chance: 2 Hs in a row . . . 15 Hs in a row: When do you start to suspect that the coin is not fair?)

concept of a null hypothesis (H0); How likely is the observed data under H0? (p-value)

alternative hypothesis (HA or H1): must cover all values not included in the null

e.g. H0: mean = 55; HA: mean ~= 55 (in STATA, ‘not equals’ is ‘!=’)

p-value is computed under the assumption that H0 is true (conditional probability)

**p-value**: probability of the observed result or something more extreme under H0

Type 1 error: reject H0 when H0 is true (determined by alpha)

Type 2 error: accept H0 when H0 is false (determined by beta)

Steps for hypothesis testing:

1. state H0
2. state type of data (e.g. continuous, categorical, etc.)
3. determine appropriate statistical test
4. state summary statistics if possible
5. calculate p-value (stat package)
6. decide whether to reject or not reject H0
7. write conclusion

3.5: Example of hypothesis test with MS data

calculate Z for MS data: Z = (54.5 – 55) / (12.1 / sqrt(252)) = -0.66

p(Z <= -0.66) = area of lower tail in Nl. distribution = 0.25 (display normal(z)); one-sided p-value

2-sided p-values: p(Z <= -0.66) OR p(Z >= 0.66); 2-sided = 1-sided \* 2 (for a symmetrical distribution) = 0.51 (fail to reject H0)

3.6: One-sample t-test

In previous example, we assumed we KNEW the population standard deviation, but we almost never know this. If we do not know the sample SD, then we use a t-test. Assumption of t-test is that the underlying population distribution is normal and the observations are independent.

t-distribution is very much like the normal distribution; t-dist. has slightly heavier (fatter) tails, so it is a little more difficult (i.e. conservative) to reject H0 with a t- than with a normal-

t-value for 95% CI: (‘n’ is # of samples or degrees of freedom for t-dist.)

n=10, 2.23

n=100, 1.98

n=1000, 1.96 (same as z)

t-statistic is same as z-statistic, we just replace sigma with s (sample std. dev.)

3.7: 1-sided vs. 2-sided tests

ttest sdmtfin == 55

2-sided test is more conservative, because the rejection region is split between the high and low side; therefore you need a bigger effect to reject H0

What if p > 0.05? Does NOT mean that H0 is true; it means we fail to reject H0.

3.8: Effect size and Confidence Intervals

CI: a set of values that we believe are plausible estimates of population mean based on the sample we have drawn; an interval around our sample mean that allows us to have a certain amount of confidence that our estimate is within that interval

CI = (X – 1.96 \* s.e.m., X + 1.96 \* s.e.m)

3.9: CIs cont’d

**Key point concerning confidence intervals**: population mean is NOT a random variable; rather, the INTERVAL is the random variable that is subject to variability.

-This is why we do NOT say that there is a 95% chance that the TRUE population mean is in the interval.

-Rather, we say we are 95% confident that the interval covers the true mean

Simulation: <http://students.brown.edu/seeing-theory/statistical-inference/index.html>

Simulation: <http://www.ruf.rice.edu/~lane/stat_sim/conf_interval/index.html>

width of CI is based on 3 things:

1. confidence level (How confident do we want to be that the interval covers mu?)
2. variance
3. sample size

one-sided CI: either a lower (or upper) bound because we now say that we are 95% confident that the mean is above (or below) a given value. To have 0.05 in the lower (upper) tail, the cut-off from the nl. dist. is -1.645 (+1.645)

3.10: comparison of two groups

Example: Is immune system marker different in 2 groups of MS patients? (relapsing-remitting vs. progressive); measure of immune function: CD-26

H0: means of two populations is the same; or diff. in means is 0

2-sample t-test: underlying population distribution is normal and observations independent

STATA code:

ttest expression, by(group) unequal

end of week #3 lecture

Week #4 lecture (viewed on 10/14/2015) ANOVA

4.1 starts with a “practice problem” done in class (BMI and sweetened beverages), then reviewed

4.2: left-overs from t-test last week (1-sample, 2-sample, paired)

1-sample: “one-arm trial” (everyone in trial gets the drug; compare to historical controls)

2-sample: compare mean of one sample to mean of other sample (drug vs. placebo)

\***Many instances in statistics are a trade-off between assumptions and power.** (more assumptions = more power; down side is more assumptions = more problems with generalizing or being wrong)

4.3: Assumptions of the t-test, 2 main ones:

-independent observations

-data come from a normal distribution

* The t-test is valid even without normality if the sample size is large (n>50; Lovely et al.)
* If observations are not independent, paired t-test or correlated outcomes methods are needed

ways to assess normality:

* inspection of plot of data (histogram; qq plot)
* formal statistical tests: failing to reject H0 means assumption of normality is reasonable
  + Shapiro-Wilk test (swilk in STATA)
  + Skewness/kurtosis test (sktest in STATA)

What if normality does not hold?

* transform the data (e.g. log)
* change to a nonparametric test: for 2-sample t-test = Wilcoxon rank sum or permutation test

Two types of t-tests: equal vs. unequal variance:

* easier to just use the unequal variance version (more robust)
* test for equal variance: failing to reject H0 means that assumption of = variance is reasonable
* Is lack of equal variance interesting in and of itself? (e.g. males vs. females for many cognitive measures)

4.4 ANOVA

* continuous outcome / categorical predictor
* between group vs. within group variance
* F-distribution
* Bonferroni correction

Example: MRI study of four groups: 1 healthy control and 3 categories of MS patients

H0: all means are equal; HA: at least one of the means is different

One approach: use 2-sample t-test to make all pairwise comparisons (6 comparisons for 4 groups)

Problem: multiple comparisons

Type I error rate, alpha = P(reject H0 / H0 is true)

P(fail to reject H0 / H0 is true) = 1 – alpha

If we perform 6 pairwise tests, what is the chance that we would reject the null that all means are the same?

Two types of error rate:

component-wise error rate (i.e. Type 1 error rate per comparison)

family-wise error rate: Type 1 error rate for the entire group of comparisons

= P(find any two groups are different / all groups are the same)

FWER = 1 – (1 – alpha)^n where n = # of comparisons; alpha = per comparison

= 1 – 0.95^6 = 0.265

solution is to adjust alpha-per-comparison downwards so that FWER remains at 0.05 (several ways to do this) OR use 1-way ANOVA

\*critical to think about H0 when deciding how many comparisons you need to correct for

4.5 details of 1-way ANOVA

H0: all means are equal; HA: at least one mean is different

\*Although we are analyzing variances, we are making conclusions about means.

Assumptions (similar to 2-sample t-test)

* underlying distributions in the populations are normal
* variances of each group is equal (homoskedastic)

Two sources of variability:

* within group: individual samples within the same group
* between group: difference from group to group

If between-group variability is large (compared to within-group), the means of the two groups are likely not the same.

4.6 ANOVA notation: see my tutorial on curve fitting with MATLAB and the sequential F-test

Sources of variability: the distance of each observation from the grand mean can be broken into TWO pieces:

1. distance of obs from its group mean (= within group variability; xij - xj
2. distance of its group mean from the grand mean (= between group variability; xj – x)

Total sum of square = within group sum of squares (MSW) + between group sum of squares (MSB)

Fk-1,n-k = MSB / MSW

If F is large, we tend to reject H0

F-distribution: ratio of variances, skewed distribution

ANOVA table: breaks down variance into different sources

4.7 Example ANOVA (MRI measurements from MS patients)

step 1: box plot of the four groups

step 2: compute ANOVA, F = 5.04; p = 0.0027

step 3: conclusion: mean is different in at least one group

STATA: oneway hypo group, tabulate

‘hypo’ = outcome variable (continuous)

‘group’ = group identity (categorical)

STATA also automatically gives Bartlett’s test for equal variances

alt. STATA command: anova hypo group

Two-way ANOVA: effect of two factors

If we reject H0, we often go on to complete pairwise comparisons

4.8 Pairwise comparisons & Bonferroni correction

\*There is a benefit to comparing the groups using the ANOVA model (as opposed to multiple t-tests), because we use information from all groups to estimate the variance.

alternative alpha for multiple comparisons (see above)

Bonferroni: use alpha / # of comparisons (simplest way to deal with multiple comparisons)

STATA: oneway hypo group, tabulate bonferroni

produces table of all 6 comparisons of the 4 groups with Bonferroni adjusted (praw \* 6) p-values

Confidence intervals: MSE is the residual variance estimated from all of the groups and alpha\* is the Bonferroni corrected level

4.9 Bonferroni cont’d, plus Tukey

What if we only care about each patient group vs. healthy controls? Now we are only making 3 comparisons, so we correct accordingly (i.e. multiply each p-value by 3 instead of 6).

Other corrections:

* Sidak’s correction: 1 – (1 – alpha)^(1/# comparisons)
* Scheffe’s correction
* Tukey-Kramer (more powerful if you are performing ALL possible pairwise comparisons)
* Dunnett’s correction (best if comparing each group to a control group)
* Many others
* False discovery rate: P(H0 true / we rejected H0)

Tukey-Kramer approach uses a different distribution (studentized range distribution) to compute the test statistic; gives more power for multiple comparisons

STATA: tkcomp group

4.10 ANOVA Assumptions

1. independent observations
2. equal variance in each group (homoskedacity)
3. normality

In most cases, there is an alternative statistical test that relaxes some of these assumptions.

As with t-test, ANOVA is quite robust to departures from normality and homoskedacity as the sample size increases.

If observations are not independent, then use repeated measures ANOVA

Equal variance assumption:

* ANOVA is robust to departures from this assumption, especially with similar sample size in each group
* If homoskedacity does not hold, there are alternative options (similar to t-test with unequal variance)
* STATA provides Bartlett’s test of equal variance as standard output for the ANOVA

Normality assumption:

* same tests for normality as for the t-test (e.g. Shapiro-Wilks test)
* assumes normality within EACH group
* ANOVA is robust to departures from normality for large sample size
* if not normal: can transform data or use non-parametric (Kruskal-Wallis test; permutation test)

Using graphs (box plots for each group):

* look for errors in data entry
* look to see whether assumptions appear reasonable

BH’s thoughts: When assessing ANOVA assumptions, consider four things:

1. What do I think the data are?
   1. all statistical tests are subject to errors (underpowered?)
   2. test for equal variance or normality may reject or fail to reject incorrectly
2. What is done in my field?
   1. previous work is a good guide
3. How big is my sample size?
   1. If in doubt with a small sample size, use a nonparametric test to deal with nonnormality
4. Am I going to fit more complex models?
   1. assumptions matter more here, so best to look at it with the simpler model

My wording of FWER: probability of NO false positives, which is 1 – probability of all true negatives

end of week #4 lecture

Week #5 Lecture: Nonparametric tests (lecture viewed on 10-21-2015)

5.3: paired t-test: observations paired as in pre-/post-tests on same person

* same as 1-sample t-test on the difference between the two measures (so really, simple)

5.4: Nonparametrics (NP): they do not make zero assumptions; just fewer assumptions

simplest NP test is the **sign test**:

* H0: median of differences in *paired data* is zero
* under H0, predict that prob. of a positive difference is 0.5
* so then it’s really just binomial probabilities
* in STATA: signtest var\_name=0
* most NP tests just give a hypothesis test; not a measure of effect size or a confidence interval

5.5: Wilcoxon signed rank test: NP analog of 1-sample t-test

* The Wilcoxon test is generally preferred over the sign test because it uses more of the data and therefore has more power. (i.e. not just sign, but both sign and rank)
* Wilcoxon signed rank test assumes symmetry and sign test does not.
* pair up data, calculate differences, then compute the rank of each difference
* “Exact Wilcoxon test” (uses permutation test)
* first do “approximate Wilcoxson test”
* under H0, we expect the sum of the positive ranks to be equal to sum of negative ranks, which is equal to ½ of the total sum of the ranks
* once we calculate mean and s.d. under H0, we just use the z-score to compute p-value

5.6: Wilcoxon rank sum test: NP analog of 2-sample t-test

* WRST is not JUST a test on the diff. of medians, but rather a test on diff. of the distributions
* BUT: assumes that the two distributions have the same basic shape
* first step: replace data with its rank
* compute sum of ranks for each group and compare with expected under H0
* logic is then the same as for WSRT (above)
* in STATA: ranksum var\_name, by (group)

For the ANOVA, the NP-equivalent is the Kruskal-Wallis test

5.7: Introduction to the bootstrap: instead of re-sampling from the population, we re-sample from the sample to get estimate of mean and std. error

* in STATA: bootstrap . . .

5.8: Bootstrap example

5.9: Permutation test: key is to break association being tested by randomly re-assigning group labels

* with small Ns, can compute test statistic for all possible grouping of data (i.e. permutations)
* example with 3 values each in group A and group B: 20 different possible arrangements

5.10: Exact Wilcoxon signed rank test:

* test statistic is sum of positive ranks
* can brute-force compute the probability of all possible sums of positive ranks and read p-value directly from this distribution
* in STATA: permute n\_permutations

End of Lecture #5

Additional resources: Bootstrap

[1] The Stata website provides a description of the commands used in bootstrap sampling and estimation, along with annotated output from some examples:

http://www.stata.com/features/overview/bootstrap-sampling-and-estimation/

[2] UCLA provides annotated Stata output for two bootstrap examples:

http://www.ats.ucla.edu/stat/stata/faq/ownboot.htm

[3] UW-Madison provides some examples on bootstrapping results from Stata commands, bootstrap options, and bootstrapping results you’ve calculated:

http://www.ssc.wisc.edu/sscc/pubs/4-27.htm

[4] Methods Consultants of Ann Arbor; another example of using the bootstrap command in Stata:

http://methodsconsultants.com/tutorial/8/The-Bootstrap-in-Stata

Additional resources: Permutation test

[1] Phil Ender (UCLA) provides background on permutation tests and an example of using the permute command in Stata:

http://www.philender.com/courses/intro/notes/permute.html

[2] UCSF provides background on permutation tests, several examples with annotated Stata output:

http://www.epibiostat.ucsf.edu/biostat/sen/statgen/permutation.html

Lecture #6: Analysis of proportions (dichotomous outcomes) viewed on 10-27-2015

6.1 Margin of error.

political polls: try to estimate proportion of population who support something

e.g. “47% of people plan to vote for Obama; margin of error +/- 3%”

What is the “margin of error”: measure of the accuracy of an estimated proportion (confidence interval)

for p=0.5, n=1000, z=1.96; accounts for sampling variability (NOT bias or other sources of error)

for the approximately normal distribution of binomial proportions: Var(p) = p(1-p)/n

95% CI = p – zalpha/2 \* sqrt(Var(p)), p + zalpha/2 \* sqrt(Var(p))

for n=1000, p=0.5 and alpha=0.05, z = 1.96 and 95% CI = +/- 3%

(to get z in MATLAB, use: norminv(1 – alpha/2)

What does the margin of error NOT tell us?

Sample design (affects generalizability of result):

Who was sampled?

How was sampling done?

Was there any missing data?

Were all people treated the same?

6.2 One-sample test of proportion

Is the proportion that progress in symptoms on drug = 20%?

Of 190 patients on treatment, 25 progressed: est. = 25/190 = 0.132; p = x/n

need an est. of the variability; What is the variance of a proportion? from binomial:

Var(p) = p(1-p) / n

When n\*p is sufficiently large, the distribution of the number of successes (X) is approx. normal

mean = n\*p; var = np(1-p)

Z = phat – p0 / sqrt(p0(1-p0)/n), where phat is the observed proportion and p0 is p under H0

* compare Z to the standard normal distribution to calculate p-value
* normal approx.. is generally deemed approp. if np(1-p) > 5
* for n = 15, p = 0.2, we get 2.4, so mini-quiz answer is ‘no’ (not OK to use normal assumption)
* in STATA: prtest prog == 0.2

6.3 Binomial Test (= exact test for one-sample proportion)

If normal approx. is not appropriate, can compute p-value directly from binomial:

this is just binorat, or binocdf(x,n,p)

P(X <= x) = binocdf(x,n,p) = binocdf(25,190,0.2) = 0.0092

in STATA: bitest prog == 0.2

OR

STATA: bitesti 190 25 0.2, detail

. bitesti 190 25 0.2, detail

N Observed k Expected k Assumed p Observed p

------------------------------------------------------------

190 25 38 0.20000 0.13158

Pr(k >= 25) = 0.994765 (one-sided test)

Pr(k <= 25) = 0.009184 (one-sided test)

Pr(k <= 25 or k >= 52) = 0.018034 (two-sided test)

Pr(k == 25) = 0.003949 (observed)

Pr(k == 51) = 0.004947

Pr(k == 52) = 0.003306 (opposite extreme)

Same as normal est. because we have a reasonably large sample size and large proportions

6.4 Confidence interval for a proportion (binofit in MATLAB)

If normal approx. is appropriate, we just derive CI with Z, using s.e.m. of the binomial: sqrt(p(1-p)/n)

* estimate of proportion is used to calculate the variance
* known as the Wald confidence interval (but there are problems)
* better: use binomial confidence interval (= binofit); in STATA:
* four CIs available in STATA: exact (default), Wald, Wilson, Agresti and Coull

In STATA: cii 190 25

. cii 190 25

-- Binomial Exact --

Variable | Obs Mean Std. Err. [95% Conf. Interval]

-------------+---------------------------------------------------------------

| 190 .1315789 .0245235 .086994 .188082

. cii 190 25, wald

-- Binomial Wald ---

Variable | Obs Mean Std. Err. [95% Conf. Interval]

-------------+---------------------------------------------------------------

| 190 .1315789 .0245235 .0835139 .179644

In MATLAB:

[phat,pci] = binofit(25,190)

phat = 0.1316; pci = 0.0870 0.1881

6.5 Comparison of proportions: Normal approx.

Ex.: presence/absence of susceptibility SNP for MS; outcome: presence/absence of dz. progression

1. Normal approximation: prtest outcome\_var, by(pred\_var)
2. Uses diff. in proportions (H0: diff. = 0) and compares this with pooled est. of variance:
   1. Var(p1) = p1(1-p1) / n1
   2. Pooled: (n1\*p1 + n2\*p2) / n1 + n2
   3. Compute std. Z statistic

6.6. Comparison of proportions: Chi-square test

|  |  |  |  |
| --- | --- | --- | --- |
| **OBSERVED DATA** | SNP+ | SNP- | Total |
| Progression | 12 | 13 | 25 |
| No Progression | 62 | 103 | 165 |
| Total | 74 | 116 | 190 |

1. Contingency table
   1. Assume that the margins are set (i.e. fixed)
   2. H0: under null of no effect, what would we expect the table to look like?
   3. Under H0, the probability of Progression is 25/190; prob. of No Progression is 165/190
   4. Use these to calculate expected numbers

|  |  |  |  |
| --- | --- | --- | --- |
| **EXPECTED under H0** | SNP+ | SNP- | Total |
| Progression | 74\*(25/190) = 9.73 | 116\*(25/190) = 15.3 | 25 |
| No Progression | 74\*(165/190) = 64.3 | 116\*(165/190) = 100.7 | 165 |
| Total | 74 | 116 | 190 |

Question, as always, is how different is the observed data from that expected under H0?

Compute statistic: sum of squared difference: sum((O-E)^2/E))

Turns out that this statistic follows a chi-square distribution with 1 deg. of freedom under H0

Cool nugget: If x is a normal random variable with mean=0 and var=1, then x2 has a chi-sq distribution with 1 deg. of freedom:

since z=1.96 is std. nl. cut-off for p=0.05, our cut-off for chi-sq is simply 1.962 = 3.84

in STATA: tabulate predictor\_var outcome\_var, chi2

Why is there 1 deg. of freedom, when there are 4 numbers? Well, because margins are set (row and column totals), as soon as you know one number, you can figure out the other 3. i.e. It is a measure of the complexity of the data.

Normal approx.: can use when expected # in every cell is 5 or greater

6.7 Fisher’s Exact Test

If we can’t use the normal approximation, we can exhaustively compute the number of contingency tables that are as or more extreme than the one we got: Fisher’s exact test

tabulate predictor\_var outcome\_var, exact

6.8 Estimate of effect (Intro)

quick review of conditional probability

Study design: cohort study

1. patients identified based on exposure: exposed vs. not exposed
2. patients followed over time to see if they develop the outcome (e.g. disease)
3. can be prospective (more common) or retrospective
4. proportion of subjects who develop dz. in each exposure group is compared
5. P(dz/exposure) vs. P(dz/no exposure)

Study design: case-control study (can be matched or non-matched)

1. patients identified based on OUTCOME
2. patients’ prior exposure is then determined
3. used with rare diseases
4. proportion of patients with prior exposure is compared
5. P(exposure/dz)

6.9 Measures of association: RD and RR

What are we trying to measure?

RD: Risk Difference = P(dz+/exposure+) – P(dz+/exposure-)

under H0: RD = 0

RR: risk ratio (relative risk) = P(dz+/exposure+) / P(dz+/exposure-)

under H0: RR = 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Exposure? Yes | Exposure? No | Total |
| Disease? Yes | a | b | n1 |
| Disease? No | c | d | n2 |
| Total | m1 | m2 | N |

P(dz+/exp+) = a/m1 = p1

P(dz+/exp-) = b/m2 = p2

RD = a/m1 – b/m2

calculate conf. int. under asymptotic normal assumption

for RR = (a/m1) / (b/m2)

to do this, use log transformed data to convert a ratio to a difference

recall: eln(RR) = RR

se[ln(RR)] = sqrt((c/a\*m1) + (d/b\*m2))

so we do calculation on log-transformed data, then raise to e to get back

e.g. from previous chi-sq problem, in STATA: csi 12 13 62 103

6.10 Odds ratio and hypothesis test with CI

Odds: p / (1-p)

Odds ratio: odds exposure+ / odds exposure-

= [P(Dz+/Exp+) / 1 – P(Dz+/Exp+)] / [P(Dz+/Exp-) / 1 – P(Dz+/Exp-)]

under H0, the Odds Ratio = 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Exposure? Yes | Exposure? No | Total |
| Disease? Yes | a | b | n1 |
| Disease? No | c | d | n2 |
| Total | m1 | m2 | N |

Odds(Dz+/Exp+) = a/c

Odds(Dz+/Exp-) = b/d

OR = (a/c)/(b/d) = ad/bc

This is the est. of the odds ratio from a cohort study

For case-control study:

Odds(Exp+/Dz+) = a/b

Odds(Exp+/Dz-) = c/d

OR = (a/b)/(c/d) = ad/bc

This is the SAME THING as the OR for a cohort study!!!

In STATA:

‘csi’ = cohort study conf. interval calculation

‘cci’ = case-control study conf. interval calculation

End of week #6 lecture

Week #7: Linear Regression and Correlation

7.3 Introduction

Correlation: Pearson & Spearman

Linear Regression: continuous outcome with continuous, dichotomous or categorical predictor

Equation: E(Y|X=x) = β0 + β1x; hypothesis testing: are betas = 0?

Interpretation of coefficients (betas)

correlations vs. regression

Big Picture:

Assoc. between 2 continuous variables: e.g. person’s weight and height

E.g.: brain parenchymal fraction (BPF): Does BPF decrease with age in MS patients?

Start with scatter plot

Nugget: One of the most important issues that get the least attention in statistics courses is the importance of having clean data.

**Correlation**: the degree to which 2 continuous variables are *linearly* related

rho (ρ) ranges from -1 to +1

Corr(x,y) = Cov(x,y) / (sqrt(Var(x)) \* sqrt(Var(y)))

In words: If covariance between 2 variables is the same as the variance in the two variables we have perfect correlation, because all of the variability in x and y is explained by how the two variables change together.

7.4 Pearson & Spearman correlation

How do we estimate that correlation? Peason’s correlation coefficient, rho (ρ)

* assumes both x and y are normally distributed
* uses mean, so can be sensitive to outliers

Hypothesis test: Is rho = 0; H0: rho = 0

* need an estimate of the standard error of rho
* se(r) = sqrt((1-r^2) / (n-2))
* resulting test statistic is a t-distribution with n-2 degrees of freedom
* t = r – 0 / se(r) = r\*sqrt((n-2)/(1-r^2))
* STATA: pwcorr age bpf, obs sig
* Can also test other hypotheses, rho = rho0 other than 0 (use Fisher’s z test)

Nonparametric correlation: if x,y not normal (due to, e.g., outliers or ordinal data)

* Spearman correlation coefficient: based on ranks in data
* Very similar to Pearson’s, but uses ranks
* test statistic, t, is basically same (once you compute rs)
* STATA: spearman var1 var2, stats(rho obs p)

Quick Thoughts:

1. Correlation does not equal causation.
2. “Statistically significant” correlation is not the same as an “important/meaningful” correlation.
   1. With very large samples, even very small rho can be significant

7.5 Linear Regression = “most commonly used statistical technique”

same example of Age and BPF in MS patients: fit a straight line to the data, minimizing vertical distance (squared) between the data points and the line

review of y = mx + b; when we add noise in the real world, two ways to write the equation:

1. yi = β0 + β1xi + εi
2. E(Y|X) = β0 + β1 xi

Since we do not know the line’s equation, we must estimate it—How do we do this?

One method is ‘least squares’. Minimize squared vertical distance from line to each point.

7.6 Linear Regression example (Age and BPF in MS patients)

in STATA: regress outcome\_var predictor\_var

for regression, H0: β1 = 0

need an estimate of variance for our β-hats: math worked out, used to compute the same basic idea of comparing difference in our result (β1) compared to the null (β1 = 0) in comparison to the variability

-the result is a t-statistic

7.7 Confidence intervals for regression

our test statistic for beta is a t-statistic, so we use this distribution for calculating Cis

same as it ever was

7.8 Assumptions of linear regression

* **independence** (each point in scatter plot is independent of others)
  + correlated data points (e.g. make measurements in same Ss over time) can be taken into account using multivariate and longitudinal methods
* **linearity**: relationship between x & y is linear
* **homoscedasticity of residuals**: residuals, εi, have the same variance
* **normality of the residuals**

7.9 R-squared

Comparison of linear regression and correlation: in worked example, we got the exact same p-values. This is because the two approaches are mathematically the same. Also a relationship between the Pearson correlation coefficient (ρ) and the so-called “coefficient of determination” (R2), which is the proportion of variance in the outcome explained by the model: R2 = (ρ)2

NOTE: R2 becomes more valuable for models with more than one predictor.

7.10 Prediction

Once we have solved for our coefficients (Betas), we are in a position to make a prediction. We just plug in a new predictor value to our regression equation. (And can also get CIs on predictions)

7.11 Linear regression with dichotomous predictor variables

Big plus of LR: we can incorporate predictor variables of multiple types (continuous, dichotomous, categorical); e.g. effect of gender of BPF

This involves use of an “indicator variable” (1 = male; 0 = female; a.k.a. “dummy variable”):

yi = β0 + β1\*male + εi

(“male” is the indicator variable; drops out for females)

What is the meaning of the coefficients? β0 is the mean BPF for females; β0+β1  is for males

(β1 corresponds to diff. between males and females)

7.12 Hypothesis test for LR w/ dichotomous predictor variable

same as it ever was; test H0 that β1= 0

also same method for CI

get exact same result (p-value) with a t-test (under equal variance assumption)

End of week #7 lecture

Stats web sites for local lectures:

http://bcb.dfci.harvard.edu/cal/

http://www.hsph.harvard.edu/biostats/events/seminars/

<http://www.hsph.harvard.edu/epidemiology/epi-seminar-series/>

Week #8: Study Design Henry Feldman, PhD viewed 09 November 2015

8.2 biostatistics overview; as a tree:

* roots: variables & distributions: binary . . . continuous; probability . . . odds
* trunk: description & inference: populations . . . sample; bias . . . standard error; confidence interval; hypothesis test . . . p-value
* main branches: comparisons & associations: correlation . . . regression; chi-squared . . . t-test
* top level: study design: sample size . . . power; confounding . . . interaction

8.3 Study design outline: finding the research question

Study design consists of the following:

1. The Big Picture: Finding the handle on a research question
   1. Handle 1: the study population
      1. The study population: unit of analysis or unit of intervention is who or what you want to study (e.g. adults, children, doctors, mice, schools, etc.)
      2. identify study population and key variables; form hypotheses
      3. method of selection or recruitment is part of the design
      4. criteria for inclusion/exclusion
   2. Handle 2: the independent variable (‘X’ or ‘explanatory’ variable)
      1. X is an attribute of the unit of analysis (e.g. hgt., wgt., sex, genes, exposure, drug dose)
      2. idea is the X drives the outcome (precedes Y; ‘causal path’)
      3. may not be able to manipulate X, but would if could
      4. X is known ‘exactly’ (or at least more exactly than Y)
      5. X can be fixed (controlled by experimenter) or random variable (observed)
   3. Handle 3: the dependent variable (‘Y’ or ‘outcome’ variable)
      1. Y is a random variable (there is always noise)
      2. Y is an attribute of the unit of analysis
      3. Y is the effect in the cause-and-effect mechanism
      4. Secondary outcomes are permissible
2. The Three Elements: Study population, X (independent) and Y (dependent)—see above
   1. Once we know the 3 elements, we can formulate hypotheses
   2. e.g. To evaluate the efficacy of two oral supplementation regimens (X) for prevention of Vitamin D deficiency (Y) in healthy adolescents (study pop.).
3. The Wider World: Covariates, confounders, and effect modifiers
   1. covariates:
      1. independent variables of secondary interest, operating jointly with X to influence Y
      2. can be fixed (planned strata) or random (observed, 'baggage'—e.g. like firing rate in analysis of effects of FB on correlations)
      3. something you want to control for or adjust for
      4. may account for the influence of X, partially or completely (mediator, confounder)
      5. may modify the influence of X (interaction)
   2. example of adjusting for a covariate:
      1. std. dev. of serum Vit. D levels in 3-yr. study was 15.2 ng/mL (which is alarmingly high)
      2. seasonal effect: winter levels lower than summer levels (sunlight)
      3. adjust for this effect: group winter and summer and measure s.d. within each group (i.e. relative to each group mean)
      4. adjusting for covariates can reduce residual variance (unknown sources), improve precision and power
      5. can also adjust for continuous variables: e.g. physical activity on Vit. D (linear regression)
      6. now calc. s.d. relative to regression line (i.e. residuals treated like yi – mu)
      7. e.g. Feldman always includes sex and age as covariates in every model
   3. confounders (a special kind of covariate)
      1. a covariate that accounts for the influence of X on Y, either partially or completely
      2. a “loose cannon” that may be responsible for effects attributed to X; may blur interpretation, cast doubt on inference regarding hypotheses
      3. How do you recognize one? by its relation to both X and Y.
   4. confounding: example
      1. in a steel plant, female workers had better lung function than males (FEV, % predicted for sex, age, wgt.)
      2. BUT: lung function declined with longer employment
      3. add regression line for FEV vs. length of employment: males & females all fall on same line
      4. length of employment (C) was related to both sex (X) and FEV (Y).
      5. The sex difference in lung function is explained away by length of employment.
   5. How to defeat confounders
      1. In an observational study design: identify potential confounders and plan to measure all you can think of.
         1. plan to use multiple regression or stratification to assess the influence of potential confounders
      2. In an experimental study design: Treatment (X) is assigned randomly, so X cannot be associated with any covariate (pre-randomization), including potential confounders.
   6. Interaction (a.k.a. “effect modification”)
      1. one covariate may modify (modulate or “mess with”) another covariate’s influence on Y
      2. an effect modifier (E) does not modify the other covariate (X); it modifies the other covariate’s effect on Y.
      3. interacting covariates are not necessarily correlated
      4. interaction is mutual: if E modifies the effect of X, the X also modifies the effect of E (seems counterintuitive to me)
   7. Interaction example:
      1. in mid-to-late adolescence (age 15-20) boys are heavier than girls by an average of 8.4 kg
      2. age trends over that time are not parallel
      3. on avg., boys gain 2.8 kg/yr.; girls gain 1.2 kg/yr
      4. i.e. effect of sex on weight is modified by age (and vice versa)
   8. Confounding (observational studies only) vs. Interaction (both O & E studies)

Y

X

E

Y

X

C

i. interaction appears explicitly in the model equation; confounding occurs ‘behind the scenes’:

ii. confounding: Y = a + bX + cC (coef. b ≠ 0, c ≠ 0, X is correlated w/ C)

iii. interaction: Y = a + bX + cE + dXE (note product term)

= a + (b+dE)X + cE

= a + bX + (c + dX)E

4) Statistical Methods and Power Analysis (put off for a later lecture)

1. Morals: Emphasizing human issues
   1. study design (SD) matters: for scientific strength, generalizability, feasibility, approvability, fundability, publishability, communicability
   2. SD is a process: so start early
   3. SD is harder than it looks; so start early
   4. Every SD is unique—you can’t get it out of a cookbook
   5. discuss SD with colleagues: Would this be exciting if we did it?
   6. starting at the end (i.e. sample size) is unproductive; need to know other things, like the primary outcome, etc.

Ends with an example of project flow from his own experience: this is precious! **End of lecture #8**

Lecture #9: Power and sample size Henry Feldman, PhD

MATLAB: sampsizepwr

9.2: sample size, precision, power, detectable effects: all intertwined

Ex. 1: mean wgt. of US newborns in 1950 was 3360 g. How many babies do we need to sample to test whether the current mean is different?

Ex. 2: Does George have enough patients (N = 35) in his study to detect a clinically meaningful difference in bone density of 0.075 g/cm2?

Ex. 3: Is acanthosis nigricans more common in obese than in normal adolescents? How many subjects do we need in each group?

9.3 What is precision?

**Estimator**: a sample statistic intended to approximate an unknown population parameter

**Standard error**: standard deviation of the estimator

**Precision** is defined as the inverse size of the standard error (i.e. good precision is a small s.e.)

3 important estimators and their standard errors:

1. sample mean: σ \* sqrt(1/n)
2. difference of independent means: σ \* sqrt(1/n1 + 1/n2)
3. binomial proportion: sqrt(π(1-π) / n), where π is the true proportion (or probability, p)

Ex. 1: To estimate mean birthwgt. within 50 g., how many babies must we sample?

Need more information:

* With what level of confidence? e.g. 95%
* Over what level of noise? e.g. σ = 545 g.
* Calculate: 50 g = radius of 95% confidence interval (2 std. errors)
  + - = 1.96 \* SE(mean)
    - = 1.96 \* σ / sqrt(n)
    - =1.96 \* 545g / sqrt(n)
    - so, n = 456 babies

Nearly 500 babies is a lot, so what about other designs?

Plan A (Original): choose 500 babies and measure wgt.

Plan B: choose 250 babies and measure each wgt. twice? (ad absurdum)

What’s wrong with this? ans. not independent samples

Levels of Variance:

|  |  |  |  |
| --- | --- | --- | --- |
| Level | Number | Indep. Obs. | Variance |
| Units: | U | U | V(U) |
| Subunits: | S (per unit) | UxS | V(S) |
| Individuals: | I (per subunit) | UxSxI | V(I) |

Std. Error of the mean (because variances add):

**SEM2 = V(U)/U + V(S)/UxS + V(I)/UxSxI**

for wgt. of babies:

baby-to-baby: V(B) = (539 g)2

weighing-to-weighing: V(W) = (80 g)2

[NB: For my money, he missed an important big picture issue, which is why the variances add. It would be useful to develop an intuition for this.]

9.5 Inferential errors, power, clinical screening & statistical inference

**Inference**:

* draw conclusions about distribution
* compare parameters to ref. values or to each other
* test hypothesis, yes/no, p-value
* Here, POWER is the issue

Two kinds of inferential errors:

* Type 1 (alpha error): false alarm; reject H0 when H0 is true
* Type 2 (beta error): miss; fail to reject H0 when H0 is false

Power = 1 – beta (analogy with the power of an engine)

= probability of not missing a true difference

graph of Prob. of rejecting H0 (y-axis) vs. true difference

What affects power?

* sample size (bigger sample = more power)
* noise level (lower noise = more power)
* experimental design (e.g. paired more powerful than independent)
* choice of critical p-value (alpha)

Analogy: clinical screening and statistical inference

|  |  |  |
| --- | --- | --- |
|  | Disease | No Disease |
| Screening test + | Sensitivity | False Positive |
| Screening test - | False Negative | Specificity |

|  |  |  |
| --- | --- | --- |
|  | Association | No Association (H0) |
| Statistical test + | Power | Type 1 (α) error |
| Statistical test - | Type 2 (β) error |  |

9.6: Example power calculation

1-sample z-test: n = (σ/Δ)2 \* (zα/2 + zβ)2 where:

n = sample size, σ = standard deviation, Δ = actual difference, α = type 1 error, β = type 2 error

(NB: for power of 80%, β=0.2, zβ = 0.84) in MATLAB: norminv(0.8,0,1) = 0.8416

for alpha = 0.05, zα/2 = 1.96; in MATLAB: norminv(0.975,0,1) = 1.9600

Overall: The five trade-offs in study design are

1. sample size
2. power
3. p-value
4. noise level
5. effect size

9.7: More examples of power formulas: plug and chug

Just a bunch of formulas. So what. Should show how you can do this with simulations.

**One nugget**: Sample sizes are always high for proportions (i.e. binomial data)—why is this? Because you are throwing away information when you reduce it to a yes/no proposition. In a sense, you’ve added noise. Moral: often better to look at actual measurements than to simply classify as, e.g., ‘normal’ vs. ‘abnormal.’

9.8: More examples

*generic power formula*: H0: Δ=0, then power is: zα/2 + zβ = ± (Δ-0) / SE(Δ)

for most calculations, zα/2 = 1.96 and zβ = 0.84, so left side = 2.8016

\*for a multi-level design (i.e. levels of variance), to maximize power, increase n at the level of greatest variance

9.9 The Big Picture: sample size & power

Statistical power is influenced by:

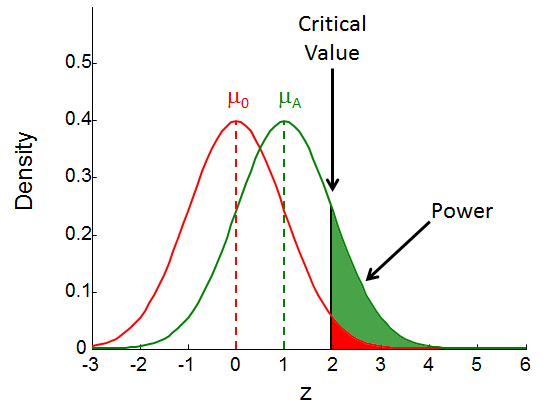
1. sample size
2. magnitude of effect
3. choice of outcome variable (binary vs. continuous)
4. prevalence of condition (binary), precision of measure (continuous)
5. study design (grouped vs. paired; presence of confounding variables)
6. method of analysis (e.g. t-test vs. nonparametrics; moral: assumptions give more power)
7. choice of Type 1 error rate

9.10: Morals about power

1. Power analysis is an exercise in speculation. It's OK to use rough numbers.
2. Power analysis is a set of trade-offs and compromises. Be flexible and ready to sacrifice one optimal design feature for another.
3. You don't have to promise p<0.05, but you do have to promise p<0.05 with reasonable likelihood. 80% is the conventional minimum. 90% is more persuasive.

**Note Well!** The best intuition about power was conveyed in about 3 minutes by the STATA YouTube channel’s “A conceptual introduction to power and sample size calculations using STATA”

Here is the key figure, which I reproduced in MATLAB (PowerDemo.m)



On-line sample size calculators:

David Schoenfeld (MGH Biostatistics Center):

<http://hedwig.mgh.harvard.edu/sample_size/size.html>

Russ Lenth (Iowa):

<http://homepage.stat.uiowa.edu/~rlenth/Power/>

UBC:

<http://www.stat.ubc.ca/~rollin/stats/ssize/>

Creative Research Systems:

<http://www.surveysystem.com/sscalc.htm>

Raosoft:

<http://www.raosoft.com/samplesize.html>

End of Week #9 lecture

Week #10 viewed on Monday, 30 November 2015

Study Design in Research Proposals and Papers (a.k.a. Statistical Communication), Henry Feldman, PhD

--Really heavily targeted to clinical trials and not at all useful for our students.

Nice little riff from Tufte during the last segment, but otherwise nothing to note.

End of Week #10 notes

Week #11 Multiple Linear Regression Brian Healy, PhD watched 12/9/2015

1. Review of Linear Regression
   1. relationship between a continuous predictor and continuous outcome
   2. example of rel. btw. age and BPF in RRMS patients
   3. BPFi = β0 + β1\*age + *e*i
   4. importance of graphing data: slope usually param. of interest
   5. can also use LR to explore rel. btw. dichot. and continuous variables (use of “indicator” variable: e.g. 0 for female, 1 for male in BPFi = β0 + β1\*male + *e*i). In this case β1-hat is the estimated difference between the groups and β0-hat is the estimated mean for females.
   6. Four assumptions of LR:
      1. independence of outcome variables
      2. linearity (i.e. rel. btw. outcome & predictor variables)
      3. homoscedasticity of residuals
      4. normality of residuals
2. Intro to Multiple Regression (MR)
   1. big reason: can include multiple predictor variables in same model
   2. same as LR, just have multiple betas and x’s
   3. same fitting procedure: minimize squared residuals
   4. same assumptions as for LR
   5. as number of predictors increases, the number of potential models and potential hypotheses increases
   6. example: BFP as function of sex and age (confound if men are younger)
      1. BPFi = β0 + β1\*male + β2\*age + *e*i
      2. for females, ‘male’=0; BPFi = β0 + β2\*age + *e*i
      3. for males, ‘male’ = 1; BPFi = (β0 + β1) + β2\*age + *e*i
      4. So, in this model, we only test the intercept, we do not allow for an effect of sex on the slope (i.e. the change in BPF with age)
      5. Interpretation of coefficients:
         1. β0 : the average BPF when age is 0 and the patient is female
         2. **β1** : the average difference in BPF between males and females, HOLDING AGE CONSTANT
         3. β2 : the average increase in BPF for a one unit increase in age, HOLDING SEX CONSTANT
3. Regression Diagnostics:
   1. We know from the design that our observations are independent
   2. plot of ‘residuals vs. fitted’: random scatter and no gross departures from linearity and homoscedasticity
   3. the normal quantile (“Q-Q plot”) plot of the residuals shows assumption of normality is also reasonable
4. Relationship between MR and analysis of covariance (ANCOVA)—just a special type of MR
   1. The example above (#2) **is** ANCOVA
   2. commonly used in comparison of 2 treatment groups
   3. Assumptions:
      1. equal slopes (to relax this A, use interaction terms (lecture #14)
      2. other A’s implicit in MR
5. Hypothesis tests of Betas (parameters or predictors):
   1. In this example, we only considered tests of single parameters
   2. could ask: Are any of the predictors significantly associated with outcome? (Global test)
   3. Do a specific subset of predictors add significantly (Partial F-test)

MR example #2: univariate analysis (starting at lect. 11.6)

* comparing volume of medulla oblongata in two groups of patients: relapsing MS vs. progressive MS
* hypothesis: MO volume is different in two groups
* potential confounders: age at onset of disease; duration of disease
* Nice STATA plot, called a “scatterplot matrix”:
  + graph matrix mov ageonset dd\_symptom rr
  + allows us to visually inspect for outliers
* First step: unadjusted analysis: fit model without confounders included
  + t-test or LR with indicator variable
  + MOVhat = 0.911 + 0.164\*RR (where RR=0 for progressive; 1 for relapsing/remitting)
* Second step: now adjust for potential confounders (based on scientific knowledge *without* regard to statistical considerations)
  + new model: MOVi = β0 + β1\*RRi + β2\*Ageonseti + β3\*DDi + *e*i
  + β1 is the diff. in mean MOV btw. dz. groups holding age-at-onset and disease-duration constant
  + β2 is the change in mean MOV for a one unit increase in age-at-onset holding dz. group and disease-duration constant
  + β3 is the change in mean MOV for a one unit increase in disease-duration holding dz. group and age-at-onset constant
  + STATA: regress mov rr ageonset dd\_symptom
  + generically: regress outcome predictor\_1 predictor\_2 predcitor\_3
  + MOVihat = 0.92 + 0.15\*RRi + 0.0004\*Ageonseti – 0.0025\*DDi
  + Hypothesis test: β1 = 0; t=2.70 (63 df), p=0.009
  + degrees of freedom changed from 65 for unadjusted analysis to 63 for adjusted
    - dof for MR is # observations minus # coefficients estimated
    - rule-of-thumb: need at least 10 observations per coefficient
    - i.e. sample size limits model complexity
  + for plot of residuals in STATA: rvfplot

MR Example #3: interaction, problem of batch effect (lect. 11.8)

* mean expression of an immunologic marker in patients vs. healthy controls
* since all of the samples could not be analyzed in a single experiment, the experimental design was to measure the outcome variable in two batches (w/ patients and controls represented in each batch)
* 40 ctrls. and 40 patients:
  + batch 1: 20 ctrls, 20 patients
  + batch 2: 20 ctrls, 20 patients
* model: Expi = β0 + β1\*Diseasei + β2\*Batchi + *e*i
  + Disease = 1 for patients, 0 for ctrls
  + Batch = 1 for batch 2, 0 for batch 1
* Expi = 3.43 + 0.32\*Diseasei + 0.38\*Batchi
  + both batch and disease seem to affect the outcome
* H0: no diff. btw. groups (pt. vs. ctrl): β1 = 0
* NOTE: In this model, we assume that the diff. btw. groups is the same in batch 1 and batch 2
* What if the difference between the two groups was different in the two batches?
  + This implies an *interaction* between group and batch

MR example #4: effects of multiple continuous predictor variables

* effect of fatigue and depression on mental quality of life in MS patients
* regression eqn: MCSi = β0 + β1\*Fatigi + β2\*Depi + *e*i
* STATA: regress mcs fatig dep
* soln: MCSi = 71.0 – 0.16\*Fatigi – 0.63\*Depi
* interp. of R2: percentage of variance in data explained by model
  + R2 = SST – SSE / SST
    - SST = total sums of squares (total amt. var. in data)
    - SSE = residual sums of squares

Week #12 Multiple Linear Regression (MLR) **diagnostics**

ASSUMPTIONS of MLR and strategies for dealing with departures from those A’s.

12.1: review of last week’s lecture on interpreting MLR coefficients

12.2: practice problem review

12.3: diagnostics for continuous data

* Two recommended books on MLR:
  + “Applied Regression Analysis and Other Multivariable Methods” by Kleinbaum, Kupper, Nizam & Muller (good introduction)
  + “Regression Modeling Strategies” by Harrell (for fancier stuff)
* model diagnostics BEFORE analysis:
  + assessment of data quality
  + 90% of problems w/ MLR due to errors in data, often due to data entry
  + continuous data:
    - make a histogram of each variable separately (both predictors and outcome)
      * missing data, outliers, shapes of distributions
    - look at 5 highest and 5 lowest values for each variable
    - in STATA: summarize age, detail
  + categorical data:
    - make a table: tabulate gender
  + scatterplot of outcome vs. explanatory variable(s)
    - “scatterplot matrix”: graph matrix outcome\_var pred\_var1 pred\_var2 etc.
* model diagnostics AFTER analysis:
  + ASSUMPTIONS of MLR:
    - independence (most important)
    - linearity
    - homoscedasticity of residuals
    - normality of residuals (least important)
  + regression diagnostics
    - BUT, best to determine if A’s apply PRIOR to choosing your model
    - MLR diagnostics designed to find gross departures from MLR A’s
    - estimated residual: observed value – estimated value (from regression eqn)
    - Two approaches to calculating residuals in STATA:
      * first calc. fitted values, then subtract from orig. data
        + predict yhat, xb
        + generate resid=bpf-yhat
      * calculate residuals directly
        + predict resid2, resid
    - Residual plots:
      * “Residual vs. fitted” plot
        + STATA: rvfplot (after a regress command)
        + y-axis = residuals; x-axis = val. estimated from regression
        + if A’s of MLR are met, we should observe a random scatter of points on this plot
      * histogram of residuals: histogram resid
      * Q-Q plot: qnorm resid
        + want points to fall on main diagonal
* Departures for the A’s of MLR
* 12.6: departure from independence:
  + most important assumption, because it is not fixed by large sample size
  + often not obvious from graphics, but are known from the *study design*
    - e.g.: change over time in BPF (3 measures from each of 3 patients)
      * observations correlated within a given patient (i.e. patients tend to have the same BPF over time, across measurements)
    - e.g.: kids from the same school
* 12.7/8: departure from linearity:
  + to check: scatterplot of outcome vs. predictor:
    - e.g. there appears to be a relationship, but it is not linear
    - e.g. BPF vs. age: linear, but with steep fall-off > 65 y.o.
    - possible remedy: transformation of the x-variable:
      * add a quadratic term (x2, polynomial regression):
        + BPFi = β0 + β1\*age + β2\*age2 + *e*i
        + STATA: regress bpf age agesq
        + Is there an overall assoc. btw. age & BPF?

H0: β1 = β2 = 0

* + - * + Does the quadratic term add significantly to the model with a linear term?

H0: β2 = 0

* + - * fit a spline (allows different slopes for diff. parts of the curve)
* 12.9/10: Homoscedasticity (equal variance across values of predictor)
  + most common: variability goes up with value of the predictor
    - common remedy: log-transform or square-root transform the y-value (“variance stabilizing transformation”)
    - other remedies:
      * weighted least squares
      * “robust variance estimation”
        + in STATA: use the vce(robust) or vce(hc3) at end of regress statement
    - paper on web site by Long & Ervin all about dealing with heteroscedasticity
  + formal tests for homoscedasticity:
    - STATA: estat hettest (test for unequal variance; want p > 0.05; similar in spirit to Bartlett’s test for ANOVA)
* 12.11: departures from normality (i.e. normality of residuals):
  + not important if you have a large data set (see paper on web site by Lumley et al.)
  + common remedies:
    - transform outcome variable
      * log transformation
      * cube-root transformation (for volume data)
    - non-parametric methods
      * median regression
      * robust regression
    - switch to another type of regression
      * logistic regression
      * Poisson regression
    - Bootstrap the standard errors
      * in STATA: add vce(bootstrap) at the end of regress
      * coefficients remain the same, but std. errors calculated without assuming normality
* 12.12: outliers/influence
  + extreme values in vertical direction (y-axis): “discrepancy”; easy to see on residual plot
  + extreme values in horizontal direction (x-axis): “leverage”; residual may not be extreme, but it can still have an undue influence on the regression
  + diagnostics:
    - “dfbeta”: calculates the change in the regression coefficients when each observation is removed
    - “cooksd”: calculates Cook’s distance, which combines the information
    - jackknife residuals (leave one out analysis)
  + robust regression that down-weights extreme values (e.g. MLE)
  + \*best to think about how you will handle outliers BEFORE you look at results so you can be confident regarding your inference (i.e. not talk yourself into a positive result)

End of lecture #12

Lecture #13: MLR, Part 2 Brain Healy watched in Milwaukee, 22 Dec. 2015

loose end from last week: multiple kinds of residuals:

1. observed – predicted (‘raw’ residual)
2. ‘standardized’ residual
3. ‘studentized’ (jack-knifed)

ANOVA and the F-test

* extension of idea of using regression with an indicator variable to do a t-test: use multiple indicator variables (NOTE: ‘indicator’ variable a.k.a. ‘dummy’ variable)
* Using indicator variables for an n-group comparison, include n-1 of the variables in the equation and leave out the one that is your reference group.
* The intercept is the mean value of the outcome for the control group.
* This is called ‘reference cell coding’ (vs., e.g. ‘effect coding’)
* hypothesis testing:
  + global test: all betas = 0
  + in STATA, after running regression: test (β1 = 0) (β2 = 0) (β3 = 0)
  + get exact same result as running a one-way ANOVA
  + advantage of regression over ANOVA is that we can also include confounder variables in our regression model (e.g. to control for age)
* degrees of freedom: regression vs. ANOVA:
  + In ANOVA, the degrees of freedom are based on the sample size and the number of groups (k-1, n-k).
  + In linear regression, the degrees of freedom are based on the number of predictors in the model (#large-#small, n-#large).
  + When we have more than two groups, we can do global comparisons and pairwise comparisons using the regression output.
* can follow up the global test (all betas = 0) with individual pair-wise comparisons
  + individual betas = 0
  + these p-values are NOT adjusted for multiple comparisons
  + NOTE: These are all tests for each group vs. healthy controls
  + If we want to compare two patient groups, we test, e.g. β1 = β2 (or β1 − β2 = 0)
  + in STATA: test β1 = β2
* ALWAYS WRITE DOWN THE MODEL!!!

13.5: F-test:

* test of a ratio of variances: (variance explained by the model) / (residual variance)
  + equivalent to the ANOVA comparison of the between-group variance (var. explained by model) to the within-group variance (residual variance)

13.7: Example of SAT scores and High school grades as predictors of college performance

* a bunch of predictors, all of which seem to correlate with outcome
* BUT, also, all of which are correlated with each other
* When we look at univariate regression of HS-english, we find a significant effect; however,
* When we now include HS-math, we find that the sig. effect of HSe goes away—it was just along for the ride, due to its correlation with HSm scores
* each coefficient represents the **independent** contribution of that predictor (i.e. controlling for the other predictors in the model)

13.8: Partial correlation:

* Beyond the regression coefficients, we might want to understand the correlation between each predictor and the outcome controlling for the other factors in the model.
* As correlation coefficients are related to regression coefficients, partial correlation coefficients are also related to multiple regression coefficients.
* These provide information regarding the residual correlation between two measures after accounting for other variables.

|  |  |  |  |
| --- | --- | --- | --- |
| (corr. matrix) | GPA in college | HS math | HS English |
| GPA in college | 1 |  |  |
| HS math | 0.44 | 1 |  |
| HS english | 0.29 | 0.45 | 1 |

What is the partial correlation of college GPA and HS math, controlling for correlation between HS math and HS English?

r GPA,HSM|HSE = [rGPA,HSM – (rGPA,HSE x rHSM,HSE)] / [sqrt(1 – r2GPA,HSE) x sqrt(1 – r2HSM,HSE)]

= [0.44 – (0.29x0.45)] / [sqrt(1 – 0.292) x sqrt(1 – 0.452)]

= 0.359

NOTE: The coefficients from a regression model are directly related to the partial correlation coefficients. They both provide information regarding the residual correlation between two measures after accounting for the other variables.

in STATA: pcorr gpa hsm hse

13.10: practice problem #2: full regression model of SAT data set:

>> regress gpa satv satm hsm hss hse

college GPA as explained by: SATverbal, SATmath, HSmath, HSscience, HSEnglish

full model: GPAi = β0 + β1\*satv + β2\*satm + β3\*hsm β4\*hss + β5\*hse + *e*i

Question: Do SAT scores add anything after we know HS performance? H0: β1 = β2 = 0

in STATA: test (satv=0) (satm=0)

Way cool: use nested regression in STATA:

>> nestreg : regress gpa (hsm hss hse) (satv satm)

directly compares the full with the reduced model (partial F-test?)

Is the big model (i.e. including SAT scores) better than the smaller model (i.e. just HS performance)?

13.11: Collinearity

Example: relationship between GPAcollege and SATverbal, SATmath in a reduced sample of 18 economics majors.

first look at correlation matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| (corr. matrix) | GPA in college | SAT math | SAT verbal |
| GPA in college | 1 |  |  |
| SAT math | 0.68 | 1 |  |
| SAT verbal | 0.58 | **0.80** | 1 |

Note the very high corr. between SATv and SATm.

Model: GPAi = β0 + β1\*satvi + β2\*satmi + *e*i

results of regression: neither coefficient is significant in the presence of the other, but the overall model is significantly better than the null model (i.e. there is a significant association of the combined information)—this result is due to the collinearity between SATv and SATm, or, in other words, the information they provide is highly redundant.

“variance inflation factor”: Having two collinear terms in a model leads to their variance being **over**-estimated.

End of lecture #13

Lecture #14: interpretation of interaction terms in MLR Brian Healy viewed 5 Jan. 2016

Example #1: BPF (“brain parenchymal fraction”) in relapsing remitting MS (RRMS) with two predictors: sex and age

* Model #1 (simplest): no predictors (i.e. a flat line), y-intcpt (β0) = mean BPF
  + BPFi = β0 + ei
  + same as 1-sample t-test
* Model #2: effect of gender only (offset, but no effect of age; i.e. 2 flat lines (BPF vs. age))
  + BPFi = β0 + β1\*malei + ei (where male = 0 for females)
  + Same as 2-sample t-test (with equal variance)
* Model #3: effect of age only
  + BPFi = β0 + β1\*agei + ei
* Model #4: effect of both gender and age:
  + BPFi = β0 + β1\*malei + β2\*agei + ei
  + 2 parallel lines: same effect of age within sex (i.e. slopes same)
* BIG POINT: In order to interpret the regression coefficients, we must consider ALL of the factors in the model.
* Model #5: interaction between two variables: What if the effect of age is different for men vs. women?
  + BPFi = β0 + β1\*malei + β2\*agei + β3\*agei\*malei + ei
  + Break this down by sex:
    - For females (male=0): BPFi = β0 + β2\*agei + ei
    - For males (male=1): BPFi = (β0 + β1) + (β2+ β3)\*agei + ei
    - i.e. allows for different slope and y-intcpt for the two sexes
  + What is the meaning of β3?
    - **Heuristic**: The effect of age on BPF is different in males and females. (or, equivalently, the effect of sex on BPF changes with age)
    - **Mathematical**: The change in the effect of age comparing males to females. (or, equiv., the change in the difference between males and females for every one-unit increase in age)
  + How do we test for an interaction?
    - H0: β3 = 0
    - In STATA: regress bpf male age inter
  + How do we “handle” (i.e. interpret) the interaction?
    - If β3 is sig. diff. from 0, we report the 2 diff. slopes
    - If not, we remove interaction term and use model #4.
  + What about the other coefficients? (more complicated)
    - In the interaction model, β2 is the effect of age on BPF for females.
  + Test for coincidence: i.e. the two lines are the same (both slope and y-intcpt are the same): β1 = β3 = 0 (NOTE: In lect. 14.5, for this last step, he switches the order of the coefficients in the model, so that what was β1 becomes β2)

Example #2: predictors of weight in an adult sample: sex, age, diabetes, others

* First step: plot of wgt. vs. age for females (red) and males (blue)
  + Males look to be, on avg., heavier than females
  + Older cohort: wgt. tends to decrease with age
* Do age and sex interact in terms of their effect on weight?
  + wgti = β0 + β1\*malei + β2\*agei + β3\*agei\*malei + ei
* Note that this is identical to our model #5 above (except we substitute wgt for BPF)
  + For females (male=0): wgti = β0 + β2\*agei + ei
  + For males (male=1): wgti = (β0 + β1) + (β2+ β3)\*agei + ei
  + In STATA: regress weight i.male c.age i.male#c.age
* Whether age and gender interact in terms of their effect on weight can be asked as the following questions:
  + 1) Does the difference in weight between males and females change as age changes?
  + 2) Is the effect of age on weight different in the males and females?
  + 3) Is there an interaction between age and gender?

Example #3: two categorical predictors (with non-interaction model):

* Another common study design, especially in basic science and social science, is to investigate the relationship of two dichotomous or categorical variables and a continuous outcome.
* One of the categorical variables may be the object of interest and the other a confounder, or each may be of interest.
  + E.g. treatment/gender (var. of interest)
  + E.g. treatment/site (confounder)
* Example: effect of two factors on the expression of a marker in mice:
  + Genotype: WT vs. KO
  + “Air” condition: air (ctrl) vs. gas
  + 2 factors, each with 2 levels → 4 potential groups
  + 6 mice in each of the 4 groups: classic 2-way ANOVA
  + exi = β0 + β1\*KOi + β2\*gasi + ei
    - KO = 0 for WT, 1 for KO
    - Gas = 0 for air, 1 for gas
  + In STATA: regress val ko air

Example #4: two categorical predictors (with interaction model):

* In factorial design (above) we often want to test for an interaction between the two factors.
* exi = β0 + β1\*KOi + β2\*gasi + β3\*KOi\*gasi + ei
* Break this down by groups:
  + G1 (WT/Air): exi = β0 + ei
  + G2 (WT/Gas): exi = β0 + β2 + ei
  + G3 (KO/Air): exi = β0 + β1 + ei
  + G4 (KO/Gas): exi = β0 + β1 + β2 + β3 + ei
* In STATA: regress val ko air inter
* \*\*NOTE: BH advice: Whenever you want to know anything about statistics, google the general keywords + “ucla”
* To see diff. btw. ANOVA and regression approaches see: www.ats.ucla.edu/stat/stata/faq/main\_effects.htm

Final practice problem (“pathological case”): effect of 2 genes on cholesterol levels

Coding is presence (1) vs. absence (0) of each gene

|  |  |
| --- | --- |
| G1: gene1=0/gene2=0  Est. mean = 171 = β0 | G3: gene1=0/gene2=1  Est. mean = 190 = β0 + β2 |
| G2: gene1=1/gene2=0  Est. mean = 188 = β0 + β1 | G4: gene1=1/gene2=1  Est. mean = 166 = β0 + β1 + β2 + β3 |

Model 1: effect of gene1, ignoring gene2: no effect

Model 2: effect of gene2, ignoring gene1: no effect

.

.

.

Model 5: effect of gene1 only in subset of people who have gene2: sig. effect

Model 6: effect of gene1 only in subset of people who DO NOT have gene2: sig. effect

BUT: effects are in opposite direction, so they cancel in the univariate analyses

Model 4: include interaction: sig. effect of interaction term

“a perfect storm” of interaction: gene-gene interaction, with the effect of one gene being negative in the presence of the other and positive in its absence

End of lecture #14

Lecture #15 John Orav, PhD viewed on 12 Jan. 2016

Regression: Approaches to Variable Selection (Model Building)

good news: principles apply to any sort of regression

bad news: no single right answer; requires judgment based on certain principles

1) The simple "univariate" (i.e., one predictor versus one outcome) effect estimates and tests that you learned earlier can produce inaccurate and/or biased results if you do not account for other measurements (covariates) that are related to the outcome.

2) Or, if you incorporate covariates that you should not, you can create a bias or a false negative result.

Three types of measurements:

1. outcome
2. predictors: the things you care about testing
3. covariates (e.g. age, sex) – not the goal of the study, but they always affect the outcome (disease)

Typical uses of regression include:

1. Determine whether a specific predictor is causally related to the outcome.
   1. confounders must be included in the model to eliminate alternative explanations
   2. collinear covariates must be excluded from the model in order to preserve power
   3. other significant predictors could be included in the model in order to enhance power
2. Identify all predictors that are causally related to the outcome.
   1. as above
   2. prioritize between retaining confounders and eliminating collinear covariates
   3. convention is that confounding usually takes precedence over collinearity
3. Build a model that can accurately predict the outcomes for individual patients. Predictors should:
   1. be few and (highly) significant
   2. be easy to measure
   3. be measured in advance of the outcome
   4. be able to be used to help the patient
   5. enhance the sensitivity and specificity of the model (or the r-squared)
   6. make clinical sense

Potential problems: bias through confounding

1. **Bias** occurs if the effect estimate, even with an infinite sample available, would not equal the true relationship in the population.
   1. bias can occur due to flaw in study design
   2. bias can occur because of naturally occurring differences between patients in the groups being compared
   3. bias can make an effect appear either weaker (i.e. towards H0) or stronger
   4. Sources of bias:
      1. confounders: covariates that are *related* to the predictor of interest and also independently to the outcome. Always include in model.
         1. confounding can go in two directions: positive (i.e. effect on outcome is in the same direction as that of the predictor) or negative (opposite direction); so, when you include a positive confounder, the effect of the predictor gets smaller
         2. Think of positive confounders as alternative explanations of the outcome.
      2. Intermediary variables are covariates that are on the "causal pathway" between the predictor of interest and the outcome. (a.k.a. “mediator” variables)
         1. if included in analysis, the effect of the primary predictor will be weakened
         2. Rule of thumb: Do **NOT** include intermediate covariates in your analysis
         3. (NOTE: This should not be properly considered as a source of bias.)
      3. Effect modifiers: covariates such that the effect of the primary predictor on the outcome is different according to different levels of that covariate
         1. consequence: You can no longer report simply the one effect of the predictor; you must report a different effect for each level of the effect modifier
         2. example: a new chemotherapy may improve survival in Stage 1 patients, but show no efficacy in Stage 3 patients
         3. requires inclusion of an *interaction term* in the regression model
2. **Imprecision** occurs when the study outcomes are not consistent from patient to patient because of the variability in patient characteristics.
   1. leads to large standard deviations and standard errors for our effect estimates
   2. leads to wide confidence intervals
   3. leads to higher p-values
   4. may lead to false negative results
   5. sources of imprecision:
      1. collinearity: covariates that are related to the predictor but NOT to the outcome (can also be thought of as a *redundant predictor variable*)
         1. Can think of this as that the model “gets confused” as to which predictor accounts for the variance in the outcome, which leads to a larger standard error
         2. Including a collinear variable tends not to affect the value of the regression coefficient for the predictor (beta), but affects the standard error (makes it larger)
         3. Rule of thumb: if including the covariate affects beta of predictor by less than 20%, call it a collinear variable and exclude it from the model
         4. This could lead to a false negative result
      2. imprecision through significant covariates: i.e. an independent covariate that partially predicts the outcome but is not related to the main predictor of interest
         1. by accounting for significant covariates and reducing the variability in the outcome, the std. error of the predictor-of-interest will also become smaller
         2. therefore, one should include sig. covariates in the model
         3. **Important point**: In a randomized clinical trial, in principle you don’t need to include covariates in your model, because they should be balanced across treatment groups (due to the randomization). However, if you know of covariates, it sometimes still makes sense to include them in your model, because they will reduce the variability in the data and give you more power (i.e. since std. error is in the denominator of your t-statistic, reducing std. error helps to decrease your p-values).

Summary of Rules of Thumb for creating regression models:

1. include all positive and negative confounders
2. exclude intermediary covariates
3. include effect modifiers (as interaction term) and report separate results for each level of the effect modifier
4. exclude collinear covariates (which can inflate the std. error and lead to false neg. results)
5. include significant covariates (which can help reduce std. error and lead to true pos. results)

Implementing the above rules:

1. confounders: If you think a covariate might be a confounder: Add it to the existing model. If the coefficient (beta) for the existing predictor changes by more than 10% (20%?), then the covariate in question confounds the relationship and should be kept in the model.
2. collinearity: If you think that a covariate might be collinear, add it to the existing model and either:
   1. run collinearity diagnostics or
   2. if the std. error for an existing predictor increases substantially (> 10%), then the covariate is collinear with the predictor
   3. if the effect estimates do not also indicate confounding (i.e. change in beta of main predictor), then take that covariate back out of the model
3. significant predictors: If you think that a covariate might be a predictor:
   1. add it to the existing model
   2. if its p-value is < 0.05 and it is neither a confounder (change in beta) nor collinear (change in std. error), then it is a predictor and it should be kept in the model
4. intermediary variables: no statistical test; You need to answer the scientific question of whether the covariate is a consequence of the predictor (i.e. it is in the causal chain between predictor and outcome)
5. effect modifiers: also no statistical test; If you think that a covariate might be an effect modifier, put the interaction term into the model and see if it is significant

15.8 Other considerations: Multiple testing, overfitting, choosing predictors:

1. Multiple Testing: If you perform many statistical tests, then the chance of one or more false positives can climb much higher than 5%.

|  |  |
| --- | --- |
| # of tests | Family-wise false positive rate (at p < 0.05 / test) |
| 1 | 5% |
| 2 | 10% |
| 3 | 14% |
| 4 | 19% |
| 5 | 23% |
| 6 | 27% |

FWER = 1 – (1 – α)n

where α is the FP rate per test and n is the number of tests

solutions:

a) run as few tests as possible

b) use Bonferroni correction (run each test at p < (0.05 / # tests)

1. Overfitting:
   1. If you have too many predictors in your regression model, the equation could become numerically unstable.
   2. The coefficients and their standard errors become large and meaningless.
   3. Rule of thumb for linear regression: # predictors < (# subjects / 10) or (# subjects / 20)
2. Methods of choosing predictors include:
   1. Based on science and prior studies (“current preference” in the biostats literature)
   2. Use all available predictors and covariates in a single multiple regression model (safe for confounding; unsafe for collinearity)
   3. Use a "univariate screen" to test each covariate versus the outcome (and/or versus the primary predictor if one exists)
      1. take each of your potential covariates and test it individually against the outcome variable (criterion of p < 0.05 or 0.10 or 0.20)
      2. produces a “short list” of candidates
      3. good for eliminating collinear covariates
   4. Use an automated selection procedure (safe for collinearity; unsafe for confounding)
      1. “backward elimination”: start with all predictors and covariates in one model. If any have p > 0.05, then remove the one with the largest p-value; repeat until only significant coefficients remain
      2. “forward selection”: from the list of all potential variables, choose the one with the smallest p-value when tested against outcome (as long as p < 0.05); then add each of the remaining variables to this existing one-predictor model and choose the one with the smallest marginal p-value; continue adding one variable at a time until no remaining variable has a p-value < 0.05
         1. good for generating a prediction rule because you end up with a small number of significant predictors
         2. multiple testing problem? potentially huge and largely “blind” to reader
      3. “purposeful selection”

15.10: Evaluating model-building performance (using simulations)

--very interesting; try to get paper

end of lecture #15

Lecture #16 Logistic regression John Orav watched 25 Jan. 2016

16.1 & 16.2 clean-up from last week—still model building

1. Include all positive and negative confounders in order to avoid biased effect estimates and test results.
2. Exclude intermediary covariates (at least in initial analyses) in order to avoid biasing effect estimates and tests toward non-significance.
3. Include effect modifiers and report separate results for each level of the effect modifier.
4. Exclude collinear covariates which can inflate the standard error and lead to false negative results.
5. Include covariates which are significantly related to the outcome and can help to reduce the standard error and lead to true positive results.

16.2 choosing predictor variables

Methods of choosing predictors include:

1. Based on science and prior studies
2. Use all available predictors and covariates in a single multiple regression model (safe for confounding; unsafe for collinearity)
3. Use a "univariate screen" to test each covariate versus the outcome (and/or versus the primary predictor if one exists)
4. Use an automated selection procedure (safe for collinearity; unsafe for confounding)
   1. Reviewers and book authors dislike automated selection intensely (for valid reasons).
   2. Data analysts like automated procedures.
   3. An automated selection procedure may be more useful as part of a larger strategy, rather than as the only strategy.
   4. Orav likes the “purposeful selection” strategy:
      1. run a univariate screen for each covariate against outcome
         1. if p < 0.25, put it on the “candidate” list
         2. if p > 0.25, put it on the “non-candidate” list
      2. put all covariates from candidate list into multivariate model
      3. remove covariates from model if they are not significant (p > 0.10) AND not a confounder (i.e. removal does not change a coefficient by more than 20%)
      4. add back covariates from the non-candidate list, but keep them only if they are significant (p < 0.10)
      5. repeat step 3 (covariate removal), but only for the newly added non-candidate covariates
   5. NOTE: The purposeful selection strategy focuses on both significance and confounding, but it still suffers from multiple testing.
   6. Also NOTE: When this method was compared against other automated selection processes using simulations (with known covariates) none of the procedures did particularly well. WTF???
5. Change-in-AIC based methods (AIC = ‘Akaike Information Criterion’)
   1. use a forward (or backward) selection method, but choose (i.e. retain) covariates that reduce the AIC for the model
   2. focus is on overall performance of the model, rather than on individual predictors
   3. AIC = -2log(likelihood) + (2\*#ofPredictors)
      1. so a predictor has to substantially improve the model
      2. smaller AIC value means a better model
   4. AIC works for many types of regression (e.g. linear and logistic)
   5. let’s you compare very different sorts of models

**Logistic Regression**

1. Linear regression is for a numerical, normally distributed outcome variable.
2. Logistic regression is for a binary outcome such as Success/Failure.

ln [P(success) / P(failure)] = β0 + β1\*predictor1 + β2\*predictor2

e.g. from Salzman et al. 1992 J. Neurosci.:

ln(P/1-P) = β0 + β1\*stim + β2\*corr + β3\*stim\*corr,

where:

P is the probability of making a preferred-direction decision

β0 represents choice bias

β1 represents the effect of microstim (indicator or “dummy” variable; 1 for stim trials, 0 for ctrl)

β2 represents the effect of the visual stimulus strength (slope of psychometric function)

β3 represents interaction between visual stimulus and microstimulation (not sig., so dropped)

P = 1 / 1 + exp[-(β0 + β1\*stim + β2\*corr + β3\*stim\*corr)]

NOTE: generally have more power using the continuous outcome (vs. dichotomous)

Nice mathematical insight for why we use ln(odds) on left-hand size: since right-hand side of eqn. is not bounded and can assume negative values, we use log-odds on the left-hand side to accommodate this. (i.e. the link function ‘ln’ just makes the math tractable)

Px = P(success, given risk factors x1, x2, x3 . . .)

Px = exp (β0 + β1\*x1 + β2\*x2 + β3\*x3 + . . .) / [1 + exp (β0 + β1\*x1 + β2\*x2 + β3\*x3 + . . .)]

Procedure:

1. Estimate the βs.
2. Test each β to see if it is zero.
3. Calculate confidence intervals for each β.
4. For any “type” of subject, estimate the probability of success.
5. Put confidence intervals on each probability estimate.

**Example** of logistic regression: probability of ovulating as a function of androstenedione level on day 3 of menstrual cycle

Example 2 was on “somatitization” with a slew of predictor variables

End of Lecture #16

Lecture #17 more on logistic regression

17.1: In logistic regression the log odds of the outcome is used on the left side of the equation. (really just finishing up lecture #16)

* main difference between linear and logistic regression is the nature of the outcome variable: continuous (linear) vs. binary (logistic)
* linear regression: any type of predictor variable is allowed; same for logistic
* same (non-)rule on predictors apply to Poisson and Cox regression as well—the only thing that distinguishes them is the nature of the outcome variable

Presenting results of logistic regression:

1. pick prototypical values for predictor variables (e.g. lower & upper values of interquartile range) and calculate probability; only works well for a low number (i.e. <= 2) of predictor variables
2. if Xi is a yes/no covariate, then you can exponentiate its beta to get the odds ratio for success in the outcome variable for a patient with Xi = 1 as opposed to Xi = 0 (adjusted for the other predictor variables)
3. if Xi is a continuous variable (e.g. age), then exp(βi) is the odds ratio for success for every additional year of age

Interpreting odds ratios:

recall that an odds ratio is:

{P(success | Xi = 1) / 1 - P(success | Xi = 1)} / {P(success | Xi = 0) / 1 - P(success | Xi = 0)}

relative risk is:

P(success | Xi = 1) / P(success | Xi = 0)

The two are related but not identical:

1. if RR = 1, then OR = 1
2. if RR > 1, then OR > 1
3. if RR < 1, then OR < 1
4. if outcome is rare (i.e. P(success) is very small), then RR = OR
5. OR is always “exaggerated” compared to RR (i.e. RR closer to 1 than OR)

17.5/6: turning a logistic regression model into a prediction rule

pi = P(event | Xi) = exp(βXi) / {1 + exp(βXi)}

How good is our prediction rule?

1. c-statistic
   1. calculate probability, pi, for a patient with the event and for another, pk, without the event
   2. if model is good, pi >pk
   3. for a given pair of patients, outcome can be concordant (pi >pk), discordant (pi <pk), or a tie (pi =pk)
   4. to compute c-statistic, do this exercise for all possible pairs of patients with/without disease (M = I \* K), then C = (# concordant + 0.5\*ties) / M
2. area under ROC curve
   1. use the regression model to calculate the predicted probability of an event, pi, for each patient
   2. then select a criterion (e.g. 0.5) to classify patients as having the event (pi > 0.5) or not having the event (pi < 0.5)
   3. compare the model’s prediction with the actual outcome for each patient (i.e. the 2x2 contingency table) so that we can calculate ‘hits’ (p(model predicted event) | event occurred) and ‘false alarms’ (p(model predicted event) | event did not occur)

|  |  |  |
| --- | --- | --- |
|  | Event actually happened | Event did not happen |
| Model predicted event | ‘hit’, a | ‘false positive’, b |
| Model predicted no event | ‘miss’, c | ‘correct reject’, d |

sensitivity = a / (a+c)

specificity = b / (b+d)

* 1. for each possible criterion (0 → 1), plot sensitivity on y-axis vs. 1-specificity on x-axis
  2. area under ROC curve gives us the equivalent of the c-statistic

17.7: logistic regression example: use of B-type natriuretic peptide (BNP) in the diagnosis of heart failure

1. Put BNP into logistic regression model as continuous variable
2. Calculate ROC curve from this model
3. The area under the ROC curve (c-statistic) is 0.91.
4. The BNP value with highest sensitivity is 50 pg/ml (sensitivity=0.97).
5. The BNP value with highest specificity is 150 pg/mL (specificity=0.83).
6. For prediction models, we don’t really care about individual covariates, we just care about the overall performance of the model (i.e. confounders don’t really bother us)

17.8: assessing fit of a logistic regression model

1. p-values and odds ratios for individual predictors (low p-values and high odds ratios are, in general, good, but it’s hard to say how good)
2. likelihood-ratio test: H0 = β1 = β2 = β3 = β4 = 0
   1. i.e. versus a model with intercept only
   2. i.e. “better than nothing”
   3. analogous to the global F-test for linear regression
3. sensitivity and specificity for a given threshold (as in the ROC calculation)
4. c-statistic (or, equivalently, the area under the ROC curve)
5. Hosmer-Lemeshow test (goodness-of-fit test for logistic regression)
   1. tests whether your model is predicting well or not
   2. a test for groups of patients—whether the actual number of successes is close to the model’s predicted number of successes
   3. gives a chi-square est. of goodness-of-fit; large p-values are good (i.e. We want predicted and actual to be close to each other → low chi-sq → high p-value)
6. Individual patient residual diagnostics compare what actually happened to a patient (Y=0/1) to the predicted probability from the model for that patient (p)
   1. known as “Pearson residuals” (Lecture #17 ends here)

End of Lecture #17

Lecture #18: Start of Unit 2b on “Study Design” Henry Feldman, PhD

**NOTE**: I skimmed these, as they are very tailored for clinical studies.

Observational study designs (overview of epidemiology approaches)

1. **Observational study designs**:
   1. case series
   2. cross-sectional survey
   3. case-control
   4. cohort
2. The causal path goes forward in time *from exposure to disease*.
   1. “risk” is not just a probability of something occurring or not; it implies an arrow in time from exposure to disease
   2. relative risk: P(dz/exposure) / P(dz/no exposure)
   3. summarized in 2x2 contingency table
3. Measure Relative Risk directly if possible; indirectly if not.

18.2 Overall Types of study design

1. In an observational study data is collected 'passively'; measurement but no intervention. The investigator controls eligibility but no exposure or related variables.
2. In an experimental study exposure is controlled by the investigator and is usually randomly assigned
3. In a prospective study new data is generated after the start of the study, e.g., interviews, questionnaires, anthropometrics, diagnoses, labs.
4. In a retrospective study data already exist before the start of the study, e.g., from medical records, registry, national database, prior study, earlier phases of ongoing study.
   1. etc. Observational study designs:
5. In a cross-sectional study each subject is measured at one point in time:
   1. A sample is taken from a defined population.
   2. Disease and exposure are assessed at a single point in time, along with covariates of interest and potential influence.
   3. Association of disease and exposure is assessed, controlling for confounders and effect modifiers.
   4. Strengths:
      1. Provides disease & exposure prevalence
      2. Measure of association
      3. Can be small or large
      4. Multiple national data sources available.
   5. Weaknesses:
      1. Single point in time
      2. limits causal inference
      3. Only survivors are measured ('prevalence-incidence bias')
      4. Large prospective surveys are expensive
   6. Keys to success:
      1. Representative sampling
      2. Valid, precise, consistent measurement.
6. In a longitudinal study each subject is measured on two or more occasions.
7. case-control study:
   1. Sample is from a defined population, stratified by disease.
   2. Covariates may be matched between D+ and D-.
   3. Association of disease and exposure is assessed, controlling for confounders and effect modifiers.
   4. Strengths:
      1. Useful for rare diseases
      2. Provides prevalence of exposure, measure of association
      3. Small or large, exploratory or definitive.
   5. Weaknesses:
      1. Prone to bias from recall, survival
      2. Comparable controls hard to find
      3. No estimate of disease prevalence
   6. Keys to success:
      1. Representative sampling of the disease
      2. Controls from the same population
      3. Valid, precise, consistent ascertainment of cases
      4. Valid, precise, consistent recall or retrieval of exposure.
8. cohort study:
   1. Sample from a defined population, free of disease.
   2. Occurrence of disease is assessed prospectively.
   3. Exposure and covariates assessed at baseline.
   4. Final disease status related to exposure.
   5. Follow-up schedule may include exposure and covariate updates.
   6. Exposure may be representative or stratified
   7. Strengths:
      1. True risk estimate, incidence or hazard rate
      2. Measure of association
      3. Exposure prevalence (if so designed)
      4. Temporal verity
      5. Exposure updates
      6. Outcome adjudicated in real time
      7. Can address known adverse exposures
      8. Potential projection to population
      9. Can plan for comprehensive data collection, banking.
   8. Weaknesses:
      1. Lengthy, expensive
      2. Not practical for rare diseases
      3. Self-selected exposure
      4. Confounding rampant
      5. Loss to follow-up, possibly selective
      6. Information bias in D if not blinded to E.
   9. Keys to success:
      1. Representative sampling of at-risk population
      2. Valid, precise, consistent ascertainment and adjudication of events
      3. Valid, precise, consistent assessment of exposure
      4. Maintain the cohort
      5. Get the confounders.
9. Morals of Observational Studies:
   1. It is easy to find fault in observational studies due to possible unclear temporal sequence, confounding, and lack of representativeness.
   2. On the plus side, observational studies involve no intervention and few ethical issues so that adverse exposures can be studied. Observational studies can also be retrospective; there are many databases available.
   3. Observational studies play an important role in the study of human health and disease. Every design is unique and has advantages and drawbacks.

End of Lecture #18

Lecture #19 **Clinical Trials** viewed on 09 February 2016 Henry Feldman, PhD

Clinical trials:

1. answer clinically important questions: can disease be cured, mitagated, delayed, prevented?
2. overcome many design flaws of observational studies
3. are culminations of much preliminary work
4. are likely to provide a definitive answer, worthy of FDA approval
5. need to be feasible in order to justify
   1. time
   2. cost
   3. risk to human subjects
6. Structure:
   1. an experimental study (scientist controls the treatment → outcome)
   2. recruit volunteers from a disease-free population
   3. assign treatment **randomly**
   4. outcome is assessed prospectively
   5. double-blind: neither patient nor researcher knows who got what
   6. same basic 2x2 table defines **analysis**:

|  |  |  |
| --- | --- | --- |
|  | Outcome + | Outcome - |
| Treatment + | a | b |
| Treatment - | c | d |

* + 1. risk of outcome:
       1. treated: a / a+b
       2. untreated: c / c+d
    2. measures of association:
       1. risk ratio (relative risk, RR): [a/(a+b)] / [c/(c+d)]
       2. risk difference (absolute risk reduction, ARR): [a/(a+b)] – [c/(c+d)]
    3. Projection:
       1. number needed to treat (to prevent 1 outcome): 1 / ARR
    4. all of these measures have std. errors and conf. intervals
    5. Tests: H0: RR=1, ARR=0
  1. Special Issues
     1. choice of T-
        1. placebo?
        2. usual care?
        3. nothing?
     2. random assignment
        1. allocation ratio (T+ : T-)
        2. time, mechanics & preservation of blinding
        3. uncertain balance in small subgroups (i.e. sampling error w/ randomization)
           1. “stratified randomization”: random assignment is done within known subgroups (e.g. ethnic groups)
        4. homogeneity in time (i.e. confounding season of year w/ treatment)
           1. alternate: ABABABAB

bad, because predictable

* + - * 1. permuted pairs: AB,AB,BA,AB,BA,BA

still sort of predictable

* + - * 1. permuted blocks (2,4): AB-ABBA-BABA-BA-AB-BBAA

gives good temporal balance

hard to predict

* + - 1. predictability: no one should be able to predict who got what
    1. blinding (opp. is “open label” = not blind)
       1. who knows what?
       2. when do the patients find out?
       3. emergency need to know
  1. Data & Safety Monitoring Board (DSMB):
     1. monitor data: Is there a clear result which would merit halting the study?
     2. monitor patient safety: Is Rx harming patients?
  2. Interim Analysis:
     1. stop early if Rx is showing a clear, dramatic effect
     2. stop early if Rx is showing a clear adverse effect
     3. stop early if insufficiently powered (given what data has been collected since the beginning of the trial; aka, “futility analysis”)
     4. \*requires budget for Type I errors (i.e. multiple comparisons problem)
        1. O’Brien-Fleming rule:
           1. equally spaced looks at data (e.g. 25% enrollment, 50%, etc.)
           2. test treated vs. ctrl. at each look
           3. use a stringent critical p-value at first look (e.g. p < 0.00005; z=4), then relax on subsequent looks; chosen so that FWER is 0.05
  3. Different “phase” clinical trials:
     1. A phase I trial is for safety and may include animal studies; it is small scale and exploratory.
     2. A phase II trial is for dose-finding (drug only); it is pilot scale and uses rough efficacy and looks at more safety.
     3. A phase III trial is randomized and controlled; it is full scale, has power to demonstrate efficacy and safety, and is definitive for the FDA.
     4. A phase IV trial is for post-marketing surveillance, long-term safety and efficacy, and is wide scale.

End of Lecture #19

Lecture #20 “How good is your measure?” Henry Feldman 15 February 2016

a.k.a. “agreement studies”

1. Correlation is **how well y agrees with other measures of the same 'truth'**.
2. Reliability, reproducibility, and precision are **how well y agrees with itself**, as measured on other occasions.
3. Kappa and components of variance are **how well y agrees with itself, as measured by different raters**.
4. Validity, bias, calibration, sensitivity, specificity, false +/-, predictive value, and ROC analysis are **how well y agrees with the Gold Standard**.
5. Calibration to a gold standard:
   1. e.g. an insulin meter compared to blood measured by a lab
      1. use linear regression: measures are well correlated, but there is an offset (bias) that can be accounted for (i.e. use the regression equation to back correct: subtract y-intcpt and multiply by slope)
   2. correlation: tendency of two random variables to fluctuate in the same or opposite direction
      1. Pearson correlation: mean of [(y1 – ŷ1) x (y2 – ŷ2)] / (s1 x s2)
      2. two measures of the same event ought to be identical
         1. e.g. two readers rate the same x-ray
         2. LDL measured directly vs. calculated from total cholesterol and triglyceride
      3. different measures with common determinants
         1. e.g. height and weight
         2. vitamin D level and dairy consumption
      4. repeated measurements on the same person
         1. annual measures of child growth
         2. three successive blood pressure measurements
         3. baseline and follow-up fat intake in a diet study
      5. case and matched control
         1. e.g. two 80-y.o. females, one with osteoporosis, one without
         2. paired t-test: takes advantage of correlation to improve power
   3. reliability (reproducibility)
      1. same measure on same person on 3 different days, but have measurements by two different technicians:
         1. within tech sd = 12 mm (measurement error)
         2. between tech sd = 25 mm (rater variability)
         3. between subjects sd = 80 mm (“true”, subject-to-subject, sd)
         4. SD2 = 802 + 252 + 122
         5. ICC (intraclass correlation) = True SD2 / Total SD2 = 802 / (802 + 252 + 122) = 0.89
         6. interp. of ICC: 89% of variance is true variance, and 11% is measurement error
      2. reliability of a binary variable (McNemar’s test)
         1. e.g. inter-rater agreement
            1. 2 x 2 contingency with binary assessments of 2 raters

|  |  |  |
| --- | --- | --- |
|  | Rater 1, A | Rater 1, B |
| Rater 2, A | 71 | 6 |
| Rater 2, B | 7 | 16 |

* + - * 1. agreement: (71 + 16) / 100 = 0.87
        2. “chance” agreement:

P(chance agreement on ‘A’) = (71+7)/100 x (71+6)/100

P(chance agreement on ‘B’) = (6+16)/100 x (7+16)/100

overall chance is sum of i. and ii. = 0.65

Cohen’s kappa statistic:

(actual – chance) / (100% - chance)

(0.87 – 0.65) / (1 – 0.65) = 0.63

agreement “beyond chance”

unstable when agreement is high

McNemar’s test:

matched pairs of binary data, supposedly agreeing

null hypothesis: they agree, and disagreements are random

null hypothesis: off-diagonal pairs (= discordant pairs) should be split 50-50 (i.e. A/B as often as B/A for the 2 raters)

test: how close are b & c to: 0.5(b+c) ?

test statistic is a chi-sq:

χ2 = sum(Observed – Expected)2 / Expected

= [ (b – 0.5(b+c))2 + (c – 0.5(b+c))2 ] / 0.5(b+c)

= (b – c)2 / (b+c)

|  |  |  |
| --- | --- | --- |
|  | Rater 1, rates ‘A’ | Rater 1, rates ‘B’ |
| Rater 2, rates ‘A’ | a | b |
| Rater 2, rates ‘B, | c | d |

McNemar’s test has low power (i.e. need fairly high disagreement before reject H0)

note that for ‘agreement’ we fail to reject H0

End of lecture 20.5, taking a break to go to ICA w/ MNS; resume tomorrow with 20.6

1. agreement with a binary gold standard (i.e. you know “the truth”)
   1. Agreement with a binary gold standard is measured by sensitivity, false negative rate, specificity, false positive rate, positive predictive value, and negative predictive value.
   2. Predictive value depends on prevalence.
   3. Using McNemar's Test with a binary gold standard, the null hypothesis is that the sensitivies are equal, and all disagreements are random.

|  |  |  |
| --- | --- | --- |
|  | Gold Standard = 1 | Gold Standard = 0 |
| y = 1 | 83 | 5 |
| y = 2 | 17 | 95 |

When GS is really 1:

Sensitivity = 83/100 = 0.83

False negative = 17/100 = 0.17

When GS is really 0:

Specificity = 95/100 = 0.95

False positive = 5/100 = 0.05

When you measure 1:

positive predictive value = 83/(83+5) = 0.94

When you measure 0:

negative predictive value = 95/(95+17) = 0.85

In the field (i.e. prevalence of disease (GS=1) is low)

|  |  |  |
| --- | --- | --- |
|  | Gold Standard = 1 | Gold Standard = 0 |
| y = 1 | 83 | 5,000 |
| y = 2 | 17 | 95,000 |

Sensitivity, false neg., specificity, false pos. do not change

**BUT** PPV = 83 / (83+5,000) = 0.016

Neg. PV = 95,000/ (95,000 + 17) = 0.9998

Moral: sensitivity and specificity are properties of the test, but PPV depends on prevalence (i.e. Bayesian prior or base rates)

Comparing sensitivity:

N patients, all with disease (D+):

|  |  |  |
| --- | --- | --- |
|  | Test 1: D+ | Test 1: D- |
| Test 2: D+ | a | b |
| Test 2: D- | c | d |

Est. sensitivity:

Test #1: (a+c) / N

Test #2: (a+b) / N

H0: sensitivity equal, and all the disagreements are random; in this case, the off-diagonals (b and c) should be equal

McNemar’s test: χ2 = (b – c)2 / (b+c)

small values of χ2 → accept H0

20.7/8: continuous measure vs. a binary gold standard: ROC analysis

* e.g. continuous measure of an enzyme activity vs. pt. did/did not have disease
* draw a threshold (criterion) to create a 2x2 table (i.e. ROC analysis)
* every time you do logistic regression, you get an ROC curve for free

20.9: Power calculations by way of simulation (Brian Healy)

1. As long as we can calculate the standard error of the effect estimate, we can compute our sample size calculation.
   1. generic formula for power: (z1-α/2 + z1-β) = Δ / SE(Δ)
   2. note that SE is related to both variability and sample size
2. Stata allows us to complete power calculations in several scenarios, and there are add-on packages to Stata that allow even more.
3. Power is the probability of rejecting the null hypothesis under a specific alternative.
4. Example #1: We want to compare MS patients to healthy controls in terms of an immunological factor
   1. continuous outcome w/ dichotomous predictor → t-test for comparison
   2. based on prelim. data, we have est. of mean (SD) for each group
      1. MS patients: 36.4 (15.2)
      2. Healthy controls: 30.2 (15.2)
   3. Sample size calculation by hypothesis testing approach:
      1. state H0 and HA:
         1. H0: μMS = μHC OR diff = 0
         2. HA: μMS - μHC = 6.2
      2. state SD of outcome:
         1. SDMS = 15.2, SDHC = 15.2
      3. state desired power and alpha level
         1. Power is usually 0.8 or 0.9
         2. alpha is usually 0.05 for 2-sided test
      4. state the test: t-test in this example
      5. use statistical package to calculate sample size:
         1. STATA menu: statistics → power and sample-size analysis → tests comparing means of two independent samples
         2. STATA command line: sampsi 36.4 30.2, sd1(15.2) sd2(15.2) power(0.8)
   4. Sample size testing by simulation:
      1. recall: Power = P(reject H0 | HA is true)
      2. step 1: simulate data assuming HA is true: STATA code
         1. set obs 2
         2. generate x=0
         3. replace x=1 in 2
         4. expand 100
         5. generate val=rnormal(36.4,15.2)
         6. replace val=rnormal(30.2,15.2) if x==1
      3. step 2: analyze “data” to determine if we reject H0
         1. regress val x
         2. look at p-value and decide whether we reject H0 or not
      4. step 3: repeat steps 1 and 2 many times (10,000)
         1. ‘for’ loop in STATA
         2. store p-values
      5. step 4: calculate how often we rejected H0

End of Lecture #20

Lecture #21: Factorial Design Henry Feldman viewed on 23 February 2016

1. Factor is the classical term for a **discrete independent variable**, e.g., sex (male, female), treatment (drug, placebo, counseling), weight class (underweight, normal, overweight, obese).
   1. data with the same factor level tend to be similar; vary randomly around a common mean
   2. a factor is a source of variance
2. Analysis of variance (ANOVA) is the classical statistical technique for determining whether a factor is influential, i.e., whether or not the levels have a common mean.
3. Factorial design is a special case of regression: all independent variables are indicator variables.
4. Factorial analysis can be applied to observational or experimental data.
5. Not to be confused with factor analysis, which is a technique for identifying underlying, unmeasured variables using multiple questions or measures; e.g. aptitude tests

Example #1: t-test: 6 subjects got placebo, 6 other subjects got drug and heart rate measured after 10 min.

* **Fixed factor** is the medication (independent variable)
* independent-sample t-test, H0: means are equal
* conclusion: fail to reject H0; “drug is not a source of variance”

Example #2: 1-way ANOVA: 6 subjects got placebo, 6 other subjects got drug, 6 others read a book and heart rate measured after 10 min.

* Fixed factor is experimental condition (3 levels of independent variable)
* 1-way ANOVA, H0: all means are equal
* ANOVA table: means of book keeping

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Experimental condition | 2 | 77.12 | 38.55 | 2.05 | 0.16 |
| Residual error | 15 | 282.25 | 18.82 |  |  |
| Total | 17 | 359.36 |  |  |  |

* “Residual error” is really just “residual variance”
* “degrees of freedom”: physical analogy: airplane in sky (3 Df) vs. moving on table (2 Df); always one less than the number of levels (because they have to add up to grand mean, once you know n-1 of them, you know the last one)
* “Sum Sq”: variance around each condition’s mean
* “Mean Sq”: Sum Sq divided by Df
* “F” is the ratio of the Mean Sq to the Residual

Example #3: 2-factor ANOVA: placebo vs. drug in normal adults vs. marathon runners

* Now have 2 factors: medication and fitness (each with 2 levels)
* 2-way ANOVA, testing 2 H0s:
  + equal drug means (within fitness group)
  + equal fitness means (for a given medication)
* sort of like getting 2 experiments for the price of one

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Medication | 1 | 7.92 | 7.92 | 0.40 | 0.53 |
| Fitness | 1 | 594.09 | 594.09 | 30.03 | <0.0001 |
| Residual error | 21 | 415.42 | 19.78 |  |  |
| Total | 23 | 1017.43 |  |  |  |

* Medication was not a significant source of variance for heart rate, but fitness is

Example #4: 2-factor ANOVA with interaction: i.e. does the drug have the same effect in the marathon runners as it does in the normal controls

* experimental design is identical to that of example #3
* testing 3 H0s:
  + equal drug means (within fitness group)
  + equal fitness means (for given medication)
  + same effect of medication regardless of fitness group

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Medication | 1 | 7.92 | 7.92 | 0.40 | 0.54 |
| Fitness | 1 | 594.09 | 594.09 | 29.88 | <0.0001 |
| Interaction | 1 | 17.79 | 17.79 | 0.89 | 0.36 |
| Residual error | 20 | 397.62 | 19.88 |  |  |
| Total | 23 | 1017.43 |  |  |  |

* since interaction term is NS, conclude that medication effect is same regardless of fitness
* ANOVA table tells us about sources of variance, but it does not tell us about actual differences between groups (e.g. How much slower is the HR of the marathon runners compared to the normal controls?)
* The differences between groups are called contrasts:

|  |  |  |  |
| --- | --- | --- | --- |
| Fixed-effect contrast | bpm | t | p |
| Drug – Placebo, Marathon | -0.6 | -0.22 | 0.83 |
| Drug – Placebo, Normal | 2.9 | 1.12 | 0.28 |
| Drug – Placebo, Marathon – Normal | -3.4 | -0.95 | 0.36 |

* last line of contrast table is testing same H0 as the interaction term in the ANOVA table
* NOTE: This is all actually done with a regression model

Example #5: 2-factor clinical trial: Vit. D and 3-omega FA trial for prevention of cancer and cardiovascular disease

**NOTE**: R. A. Fisher invented factorial analysis in 1930s; he was a plant geneticist interested in testing different fertilizers in different genetic strains of corn

lecture 21.5: fixed vs. random factors

1. Levels of a fixed factor have meaning; differences are interpretable scientifically.
2. Levels of a random factor differ 'just because'; more of a nuisance
3. Statistical agenda for fixed factor is inference (hypothesis test); for random factor, descriptive.

Example: 9 different radiologists make a single measurement on an image, 6 times each

* 1-way ANOVA with reader as the factor
* ‘reader’ is a **random** factor: the question is not **whether** the readers differ but by **how much**
* 1-way ANOVA, Model 2: reader means have a Gaussian distribution, read-to-reader SD describes how much they differ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Reader-to-reader | 8 | 133.54 | 16.69 | 15.74 | <0.0001 |
| Residual error (replicates) | 45 | 47.72 | 1.06 |  |  |
| Total | 53 | 181.26 |  |  |  |

So readers are different, but HOW different?

|  |  |  |
| --- | --- | --- |
| Source of variance | Variance | SD (mm) |
| Reader-to-reader | 2.61 | 1.61 |
| Residual error (replicates) | 1.06 | 1.03 |

In general, you want more replicates for sources of variance that have larger variance.

Conclusion: Reader is a significant source of variance . . . but we already knew that. Reader-to-reader SD is 1.61 mm.

Diameter = true diameter + N(0, SD2reader) + N(0, SD2error)

How does one distinguish a fixed from random factor:

* Levels of factor
  + fixed factor have meaning; differences are interpretable scientifically
  + random factor differ “just because” (e.g. people are different)
* Statistical agendas:
  + fixed factor: inference (H0 test)
  + random factor: descriptive
* Number of levels:
  + fixed: usually just a few
  + random: usually many
* Means of the factor:
  + fixed: constants of nature
  + random: normal distribution
* Labels of a fixed factor matter; random factor labels don’t
  + nice analogy with jars of data: if labels are different treatments, and you lose the labels, you are hosed; if the labels are measurements from different patients, you wouldn’t worry about whose was whose as long as the groupings in the different jars stay separate

21.6: Nested factors: relationship of factors to each other

Example: measurement from an x-ray performed by 4 different readers who are radiologists and 4 different readers who are neonatologists, and each reader measures a single image 6 times

* Design:
  + 4 radiologists, 4 neonatologists
  + Each reader measures a single image 6 times
* Analysis:
  + specialty is a fixed factor
  + reader is a random factor, **nested** within specialty
  + H0: specialty means are equal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Specialty | 1 | 49.46 | 49.46 | 3.28 | 0.12 |
| Reader-to-reader (within specialty) | 6 | 90.51 | 15.09 | 13.76 | <0.0001 |
| Residual error (replicates) | 40 | 43.86 | 1.10 |  |  |
| Total | 47 | 183.83 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Fixed-effect contrast | Estimate (mm) | t | p |
| Radiologists - Neonatologists | 2.03 | 1.81 | 0.12 |

|  |  |  |
| --- | --- | --- |
| Source of variance | Variance | SD (mm) |
| Reader-to-reader within specialty | 2.33 | 1.53 |
| Residual error (replicates) | 1.10 | 1.05 |

Diameter = true diameter + specialty correction + N(0, SD2reader) + N(0, SD2error)

21.7: Crossover design: each subject is measured under both levels of the fixed factor

In the crossover design example: each reader uses both machines to make 6 measurements. So reader is not nested within device, but rather crosses over to both levels of the fixed factor

1. Device (2 different machines) is a fixed factor.
2. Reader is a random factor, crossed with device.
3. The null hypothesis is that device means are equal.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Device | 1 | 49.46 | 49.46 | 23.01 | <0.0001 |
| Reader-to-reader | 3 | 41.94 | 13.98 | 6.50 | 0.001 |
| Residual error (replicates) | 43 | 92.43 | 2.15 |  |  |
| Total | 47 | 183.83 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Fixed-effect contrast | Estimate (mm) | t | p |
| Siemens - ORTEC | 2.03 | 4.80 | <0.0001 |

|  |  |  |
| --- | --- | --- |
| Source of variance | Variance | SD (mm) |
| Reader-to-reader | 0.99 | 0.99 |
| Residual error (replicates) | 2.15 | 1.47 |

Diameter = true diameter + device correction + N(0, SD2reader) + N(0, SD2error)

Fine points:

1. order of administration may matter:
   1. randomly permute the order across patients
   2. include order in the analysis
   3. put a ‘washout’ period between treatments

21.8: Repeated-measures design

* Design:
  + 2 drugs (metformin vs. placebo)
  + 6 patients in each group
  + measurements made at 4 time-points: BL, 3, 6 and 12 mos.
  + BMI measured at each time point
  + data points across time are connected (by patient identity)
* Analysis:
  + Treatment and time are fixed factors
  + Patient is a random factor, **nested** in treatment, **crossed** with time
  + H0: treatment x time interaction = 0; i.e. time-course is identical in metformin and control

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variance | Df | Sum Sq | Mean Sq | F | p |
| Treatment | 1 | 45.81 | 45.81 | 3.02 | 0.11 |
| Time | 3 | 21.00 | 7.00 | 5.83 | 0.003 |
| Treatment x Time | 3 | 34.65 | 11.55 | 9.62 | 0.0001 |
| Pt-to-Pt (within treatment) | 10 | 151.45 | 15.14 | 12.62 | <0.0001 |
| Residual error (replicates) | 30 | 36.00 | 1.20 |  |  |
| Total | 47 | 288.89 |  |  |  |

|  |  |  |
| --- | --- | --- |
| Source of variance | Variance | SD (mm) |
| Pt-to-Pt (within treatment) | 3.49 | 1.87 |
| Residual error (replicates) | 1.20 | 1.10 |

Conclusion: time course of BMI differed between drug and placebo

Mixed model (MM) notes:

1. include both fixed and random factors; may include covariates and interactions
2. MMs can be applied to normal or binary data . . . or other distributions
3. MMs are fitted by standard regression software
4. In MM analysis, H0 addresses the fixed factors
5. Random factors are included to account for random sources of variance
6. A random factor may be nested within another factor, or crossed with another factor
7. Repeated-measures designs include a random ‘subject’ factor to account for within-subject correlation and add precisions (like paired t-test, but for >2 correlated data points)

Software makes mixed modeling easy:

1. list the factors (discrete) and covariates (continuous)
2. specify any nesting or interaction
3. press ‘GO’

Great closing quote from Samuel Johnson to Sophie Thrale, 1783: “Never think that you have arithmetick enough: when you have exhausted your master, buy books. Nothing amuses more harmlessly than computation, and nothing is oftener applicable to real business or speculative enquiries. . . . A thousand stories which the ignorant tell, and believe, die away at once, when the computist take them in his gripe. I hope you will cultivate in yourself a disposition to numerical enquiries: they will give you entertainment in solitude by the practice, and reputation in publick by the effect.”

End of lecture #21

Lecture #22: Intro, features of longitudinal data Garrett Fitzmaurice, Sc.D. 29 Feb. 2016

(Beginning of **Unit #4**)

1. In a longitudinal study the same variable is measured repeatedly over time.
   1. extension of linear regression
2. The primary goal of a longitudinal study is to characterize the change in response and factors that influence change.
   1. modeling the mean: analysis of response profiles
   2. modeling the mean: parametric curves
   3. modeling the covariance
   4. linear mixed effect models
3. With repeated measures on individuals, one can capture within-individual change.
   1. repeated measures are usually positively correlated
   2. variance is often heterogeneous over time
   3. paired t-test is a special case (2 repeated measures) of long. analysis
4. Notation:
   1. Yi,j denotes the response variable of the ith individual at the jth time point
   2. ‘wide’ format: rows are subjects, columns are time points
   3. ‘long’ format: inverted
5. Covariance and correlation
   1. covariance between two variables, Yij and Yik:
      1. σjk = E[(Yij - μj)(Yik - μk)]
      2. a measure of linear dependence between Yij and Yik
      3. covar takes on the units of measurement of the 2 variables
   2. correlation, ρ:
      1. the covariance normalized by the product of the standard deviations
      2. ρ = E[(Yij - μj)(Yik - μk)] / (σjσk)
      3. measure of dependence free of scales of measurement
      4. must have value between -1 and 1
   3. covariance matrix (a.k.a. “variance-covariance matrix”)
      1. way of representing var/covar of all possible pairs of time-points
   4. in general, for longitudinal studies (observations):
      1. correlations are positive
      2. correlations decrease with increasing time separation
      3. correlations between repeated measures rarely ever approach zero
      4. correlation between a pair of repeated measures taken very closely together in time rarely approaches one

End of lecture #22

Lecture #23 Garrett Fitzmaurice viewed 07 March 2016

More on regression models for longitudinal data: **Analysis of Response Profiles**

* sample of N subjects measured repeatedly on n occasions
* We assume there are ni repeated measurements of the ith subject and each Yij is observed at time tij
  + The variable t is the timing of the measure of the outcome.
  + The index j isn't quite the timing of the measurements - it is really the ordering of the measurements.
* We can group (in order) the responses for the ith subject: (Yi1, Yi2, . . . Yini)
* Our goal is to relate the mean of Y to the covariates.
  + Yij = β1Xij1 + β2Xij2 + … + β1Xijp + eij, j = 1, 2 . . . , ni
  + The eij are random errors, with zero mean; they represent deviations of the Yij’s from their respective means, so we can re-write eqn:
  + E(Yij|Xij) = β1Xij1 + β2Xij2 + … + β1Xijp
  + Typically, Xij1 = 1 for all i and j, then β1 is the “intercept”
  + We assume the eij have a multivariate normal (MVN) distribution with zero mean and variance-covariance matrix denoted by Σ. ~ MVN(0, Σ)
* Estimation method: maximum likelihood
  + basic idea: get estimates of β’s and Σ the values that are most probable (or most “likely”) for the data that have actually been observed
    - in general, ML estimation requires iterative techniques
    - because ML estimation of Σ is biased in small samples, we use a variant on ML estimation known as restricted ML (REML)
* Longitudinal data present 2 aspects of the data that require modeling:
  + mean response over time
    - **analysis of response profiles** (lecture #23):
      * Is there a change in the *pattern* of mean responses over time in each of our groups
      * most useful for balanced data (a limitation)
      * Three main questions:
        + 1. Are the patterns parallel? (constant offset): lack of a group x time interaction effect

look at differences in the means over time

H0: Δ1 = Δ2 = . . . Δn  (n-1 df)

both group and time are regarded as categorical covariates (analogous to two-way ANOVA)

(still need to account for correlation and variability among repeated measures)

So we just code it up as a regression model using indicator variables for different levels of each factor:

for a factor with k levels, we need k-1 indicator variables

the omitted indicator variable determines what level of the factor is the “reference” level

**{Example**: treatment of lead-exposed children: two groups (placebo vs. succimer) and four measurement occasions (wk. 0, 1, 4, 6)

* + Coding of this problem:
    - X1 = 1 for all children at all occasions (intercept)
    - Group:
      * X2 = 1 if child randomized to succimer, X2 = 0 otherwise
    - Time:
      * X3 = 1 if measurement at week 1; 0 otherwise
      * X4 = 1 if measurement at week 4; 0 otherwise
      * X5 = 1 if measurement at week 6; 0 otherwise
    - Group x Time interaction:
      * X2\*X3, X2\*X4, X2\*X5
  + Regression model:
    - Y = β1 + β2X2 + β3X3 + β4X4 + β5X5 + β6X2\*X3 + β7X2\*X4 + β8X2\*X5
    - H0 for group x time interaction: β6 = β7 = β8 = 0
    - Analysis must also account for the correlation among repeated measures on the same child
    - Analysis of response profiles estimates separate variances for each occasion (4 variances) and six pairwise correlations
    - For this example, all 3 βs were significantly diff. from 0, so we reject H0
  + Interpretation of coefficients:
    - mean at baseline in Placebo group: β1
    - mean at week #1 in Placebo group: β1 + β3
    - change from baseline to wk. #1 in Placebo group: (β1 + β3) – β1 = β3
    - []
    - mean at baseline in Succimer group: β1 + β2
    - mean at week #1 in Succimer group: β1 + β2 + β3 + β6
    - change from baseline to wk. #1 in Placebo group: β3 + β6
    - β6 is the **additional** increase/decrease from baseline to wk. #1 in the Succimer group (when compare to the Placebo group)

NOTE: I found this lecture (#23) to be quite frustrating. On the one hand, it was a nice example of how to code factor levels into dummy variables so that the regression framework can be used, and in how to interpret regression coefficients in a more complex model. However, the thing I wanted was completely missing: namely, some intuition into HOW the regression model takes into account the correlations among measures over time and the changes in variance.

* strengths and weaknesses of the analysis of response profiles:
  + strengths:
    - allows arbitrary patterns in mean response over time and arbitrary patterns in covariance
    - robust, since potential risks of bias due to misspecification of models for mean and covariance are minimal
    - NOTE: Analysis of Response Profiles is a special case of “repeated measures MANOVA”
  + weaknesses:
    - limited to balanced designs (i.e. all subject tested at all time-points)
    - cannot incorporate mistimed measurements
    - time is treated categorically, so time-based trends are not captured
    - produces omnibus tests of effects that may have low power
    - number of estimated parameters, G x n mean parameters and n(n+1)/2 covariance parameters, grows rapidly with number of measurement occasions}
      * + 2. Are means constant over time? (effect of time)
        + 3. Are means at same level? (group effect)
    - parametric curves
  + covariance
    - “unstructured” or arbitrary pattern of covariance
    - covariance pattern models
    - random effects covariance structure

Lecture #24 Fitting Curves to Longitudinal Data Garrett Fitzmaurice viewed 15 March 2016

using a function—usually polynomial—to explicitly describe changes in the mean over time

* simplest model is linear: E(Yij) = β1 + β2Timeij + β3Groupi +β4Timeij\*Groupi
  + now we are treating time as continuous, not categorigcal
  + for control group (Group = 0): E(Yij) = β1 + β2Timeij
  + for treatment group (Group = 1): E(Yij) = (β1 + β3) + (β2 + β4)Timeij
* next simplest is quadratic (i.e. polynomial of order two; U-shaped):
  + E(Yij) = β1 + β2Timeij + β3Timeij2 + β4Groupi + β5Timeij\*Groupi + β6Timeij2\*Groupi

Case study: Dutch study on risk factors for COPD

FEV1 as a function of time for current vs. former smokers

24.6: Modeling the covariance

* “unstructured” or arbitrary pattern of covariance
  + appropriate when design is balanced and number of measurement occasions relatively small
  + makes no assumptions about pattern of variances and covariances
  + does assume homogeneity of covar across individuals
  + lots of parameters: for n measurement occasions: (n x (n+1) ) / 2
    - n variances
    - n-choose-2 correlations = n(n-1) / 2
  + Thus unstructured models only appealing when N large compared to n(n+1) / 2
* covariance pattern models
  + balance between bias and variance:
    - with too little structure, you have too many parameters to estimate
    - with too much structure, you run the risk of model misspecification, which can result in misleading inferences on βs
  + “compound symmetry” (or “exchangeable”):
    - assumes variance is constant across occasions
    - assumes all pairwise correlations are the same
    - thus there are only two parameters
    - same assumptions as repeated-measures ANOVA
    - These are strong assumptions that usually do not hold for longitudinal data
  + Toeplitz
    - assumes constant variance across occasions
    - assumes correlations among responses at adjacent measurement locations is constant (i.e. correlations only vary with distance in time between measurements)
    - a special case of Toeplitz is the (first-order) **autoregressive** covariance: builds in an assumption about how correlations change as a function of distance between them: ρk, where k is the distance between measurements
      * assumes correlations decay exponentially with increasing time separation
      * if measurements are not equally spaced in time, we can just model correlations as: ρ|tij – tik|
        + known as an **exponential** covariance model
* random effects covariance structure (lecture #25)

Lecture #25 Linear mixed-effects model Garrett Fitzmaurice viewed 8 April 2016

I skipped the lecture, took the quiz and got 100%

Here are some more mixed effects model examples:

[1] UCLA provides some mixed effects model examples, using Stata’s xtmixed command, from Chapter 4: Doing Data Analysis with the Multilevel Model for Change of “Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence” by Judith D. Singer and John B. Willett:

http://www.ats.ucla.edu/stat/stata/examples/alda/chapter4/aldastatach4.htm

[2] University of Michigan provides examples in Stata, SAS, SPSS, and R for mixed effects models:

http://www-personal.umich.edu/~bwest/chapter5.html

http://www-personal.umich.edu/~bwest/chapter6.html

[3] Chapter 26 of Applied Regression Analysis and Other Multivariable Methods by Kleinbaum, Kupper, Nizam, and Muller is on Analysis of Correlated Data Part 2: Random Effects and Other Issues. This book is available at the MGH Biostatistics Center statistical computing lab:

Kleinbaum, D., Kupper, L., Nizam, A., Muller, K. (2008). Analysis of Correlated Data Part 2: Random Effects and Other Issues. In Applied Regression Analysis and Other Multivariable Methods (pp. 735-787). Belmont, CA: Thomson Higher Education.

End of Lecture #25

Lecture #26 Multi-level models Brian Healy viewed Friday, April 08, 2016

**Idea for PiN student version of course:**

1. **BH intro lecture on regression/correlation (week #7)**
2. **Overview lecture on all of the branches of regression (GLM framework: logistic etc.; correlated/longitudinal data)**
3. **Detailed look at multiple linear regression: interpreting coefficients and regression diagnostics**
4. **Extension to classifiers**

Next few weeks are largely review of all of regression

1. Longitudinal data is a special type of correlated outcome because we observe people over time and we are usually interested in change over time.
   1. Longitudinal data
      1. handling correlated outcomes (observations within a patient are correlated)
      2. understanding covariance structures

Two approaches:

* 1. Marginal models: the covariance matrix between residuals
     1. e.g. wtloss.dta: effect of program and time (+ interaction)
     2. Yij = β1 + β2Timeij + β3Groupi +β4Timeij\*Groupi + εij
  2. Random effects models: An alternative way to account for the correlation among observations within a person is to include patient specific random effects.
     1. Yij = β1 + β2Timeij + β3Groupi +β4Timeij\*Groupi **+ b0,i** + εij
        1. i.e. In this model we allow each patient to have his own intercept (b0,i)
        2. Now the εij are considered independent, since we accounted for correlation within patient using the random effects (b0,i)
        3. We assume that the random effects have a specific distribution.
        4. Regression coefficients and variances are estimated using maximum likelihood.
     2. random intercept: patients start out at different points (e.g. wgt. loss Rx)
     3. random intercept **and slope**: change over time can also vary across patients
        1. Yij = β1 + β2Timeij + β3Groupi +β4Timeij\*Groupi **+ b0,i + b1,i\* Timeij** + εij
  3. Multi-level models (e.g. nesting: measurements of students clustered within schools)
     1. longitudinal data is not the only source of correlations: e.g. groups within data
     2. can think of longitudinal data as a type of multi-level data:
        1. patient is the top level
        2. observations within a patient over time is the lower level
        3. (sounds like repeated measures)
     3. Example #2: relationship to paired t-test
        1. two measurements for each patient
        2. paired t-test looks at differences within each patient
        3. multi-level model just includes a patient-specific intercept
           1. MOVij = β0 + β1\*timeij **+ b0,j** + εij
           2. We are treating time as a categorical variable
        4. A multilevel model allows flexibility to expand the paired t-test in terms of multiple time points, more complex models for the covariance as the number of time points increases, and additional effects that impact the change.
        5. Two Big Implications:
           1. When we have measurements both on and off the treatment on the same person, accounting for the correlation within a person often increases the power to detect an effect since we are removing a source of variability.
           2. When we have measurements only on one treatment for each person, accounting for the correlation within a person often decreases the power because we have less independent information. (But we really DO have less independent information, so this is the honest thing to do.)
     4. Example #3: We will assess if there is a significant effect of the mean socio-economic status (*meanSES*) of the school on math achievement (*mathAch*).
        1. We are modeling a school level effect, where j indexes schools (n=160), and i indexes students within schools.
        2. mathAchij = β0 + β1meanSESj + b0,j + εij
           1. *j* indexes schools; *i* indexes students within schools
           2. b0,j is the school-specific random effect (intercept), b0,j ~ N(0,σ2b)
           3. εij is the residual error, εi,j ~ N(0,σ2ε)
           4. We assume that the two errors are uncorrelated
           5. mixed mathAch meanSES, || school: , covariance(independent) variance
           6. H0: no effect of school SES: β1 = 0
           7. NOTE: ‘xtmixed’ is an old STATA command; ‘mixed’ is the modern command as of STATA ver. 14
        3. With a multilevel model, we have two sources of variability, and we can investigate the proportion of variance explained by each source.
     5. Example #4: three-level model: effects of a school curriculum (*curr*) intervention on knowledge of smoking as assessed by a pre- and post-intervention test score
        1. levels:
           1. students who take the test
           2. classrooms
           3. schools (curriculum intervention is at school level)
        2. postScoreijk = β0 + β1\*preScoreijk + β2\*currk c0,k + b0,jk + εijk
        3. *i* indexes students, *j* indexes classrooms, *k* indexes schools
        4. mixed postScore preScore curr, || schoolID: , covariance(independent) || classID: , covariance(independent)
        5. H0: no effect of curriculum while controlling for pre-intervention score: β2 = 0

End of Lecture #26

Lecture #27 Missing Data Brian Healy, PhD started viewing on 12 April 2016

* Missing data occurs commonly in clinical research, including longitudinal studies. (BH recommends article on course web site by Graham on general idea of missing data)
* Key question: Why are the data missing?
* Three types of missing data are: (Little and Rubin 1987)
  + 1. Missing completely at random (MCAR):
    - chance of missing an observation is independent of the observed and unobserved outcomes; best case scenario for missing data
    - observed data are a random sample of the full data set
    - mean and variance are unbiased estimates (same as for complete data set)
    - Example: A patient drops out of clinical trial for heart medication at the half-way point because the patient’s job requires moving to another state.
  + 2. Missing at random (MAR):
    - chance of missing an observation can depend on the observed outcomes (Yobs) but is independent of the unobserved (Ymiss) outcomes
    - considered ignorable missingness
    - Abstract example: There was missing data at time-point #3
      * The patients with specific values at time-points 1 and 2 who were observed at time-point 3 are a random sample of all patients with the same values at time-points 1 and 2
    - Concrete example: A patient with an observed disease severity worse than a specific value is forced to leave the study by design so that they can be placed on rescue therapy. (We observed a value of Y that led to their missingness.)
  + 3. Nonignorable missingness or nonignorable non response (NINR) or not missing at random (NMAR).
    - chance of missing an observation depends on both the observed and the unobserved outcomes
    - occurs when the chance of missing an observation is a function of the missed outcome variable: **people we are missing are different from the ones we observe in a way that we don’t know**
    - most challenging form of missingness to handle
    - Examples:
      * After the previous visit, a patient’s disease gets worse causing the patient to be unable to come to the clinic for the next scheduled visit
      * In a cancer trial with disease severity as the outcome, a patient dies due to cancer.
* In the presence of missing data, we will have Y's that we observe (Yobs) and Y's that are missing (Ymiss).
* “monotone” missingness: once we miss an observation, we never see that patient again (easier to handle than non-monotone missingness
* **Imputation**: the filling in of missing values:
  + single imputation: e.g. carry forward the last observed value
    - kind of cheesy, because it assumes we have information that we do not have
  + multiple imputation:
    - fill in missing value using some kind if model
    - complete the analysis as if you had complete data
      * obtain parameter estimate and standard error
    - repeat this procedure M times
      * get M estimates and M standard errors
    - calculate the final estimate and standard error
      * standard error combines the variance in each estimate and the variance across estimates

Stopping here for the day. Need to resume at 27.8 on Maximum likelihood techniques

resuming with 27.8 on 13 April 2016

* If you have missing data that you can reasonably assume is MAR, the maximum likelihood techniques that we have learned are a reasonable choice for the analysis.
* When the data are NMAR or you are unsure of the missing data mechanism, contact a statistician.
* If it is possible to contact the people who have dropped out, you can improve your analysis by incorporating this information into your model.
* **Really powerful simulation example** to get an intuitive handle on the 3 kinds of missingness:
  + Simulate data for 500 patients observed at 4 different time points. In all cases, we are going to eliminate 50% of the data for the last 2 time points:
    - MCAR: data selected to be eliminated from our full data set is chosen completely at random
    - MAR: We look at the data from the **first** 2 time points, then based on this information we determine who to eliminate
    - NMAR: We look at the data from the **last** 2 time points, then based on this (missing!) information, we decide who to eliminate
  + Generate 1000 data sets of the 500 patients at 4 time points
  + The true slope of the measured outcome over time is 0.25

Lecture 27.10: example

In the example, a model to predict disease progression in MS has complete clinical information, but only a small percentage of people have a vitamin D level (which is potentially related to disease progression). Three options for this analysis are:

* Build a model in only patients with vitamin D information even though sample size is reduced by more than 50%.
* Build a model using all the predictors other than vitamin D.
* Use multiple imputation of the missing vitamin D levels and analyze the data with the multiple imputation.

**Note re: PiN students**: Would be more useful to have a lecture on dealing with outliers. We rarely experience missing data in our world.

End of lecture #27

Lecture #28 Correlated Binary Outcomes using Generalized Estimating Equations 2 May 2016

Previous lectures, we looked at two approaches to handling correlated, continuous data:

1. model the mean and the covariance matrix (“Marginal model”)
   1. in this case, the covariance model accounts for the correlations
2. include random effects (subject specific/conditional model)
   1. in this case, the random effect(s) accounts for the individual correlations, and we then assume that the residuals are uncorrelated

Now we want to generalize these approaches to binary outcomes (logistic regression)

* The most common estimate of an effect from logistic regression is the odds ratio.
* Logistic regression also allows us to estimate the predicted probability of the event.
* Logistic regression also allows us to include multiple predictors in a model with a dichotomous outcome.

28.3 Quick diversion into the Generalized Linear Model (GLM)

To fit the GLM, you need to know 3 things: (STATA command: **glm**)

1. Relationship to predictors: What is the form of the predictors (X’s) to model the outcome (Y)?
2. Link function: What function do we use to relate the mean to the predictors?
3. Distribution family: What is the distribution of the Y’s?
   1. This specifies the relationship between the mean and the variance

e.g. for weight/smoking/sex logistic regression example:

NOTE: We dichotomized weight as either > or < 200 lbs.

>> **glm weight200 smoker male, family(binomial 1) link(logit)**

to do the equivalent, but using weight as a continuous variable:

>> **glm weight smoker male, family(gaussian) link(identity)**

NOTE: There is a subtle difference between using regress vs. GLM: regress uses the t-distribution for calculating p-values (more conservative), whereas the GLM uses the normal distribution (z).

28.4 Correlated dichotomous outcomes:

* A key assumption of logistic regression is that the patients are independent.
* A special case of a repeated dichotomous outcome is survival analysis.
* Currently, our focus are dichotomous outcomes that are correlated, but not determined by other observations.
* Generalized estimating equations (GEE) is an approach to estimate the parameters of a marginal model with correlated data. (Like the GLM, but we also include a model of the covariance between the correlated observations)
* This approach is used because in many cases we cannot write the marginal joint distribution of the outcome.
* For GEEs we need to know
  + 1. Relationship to the predictors
  + 2. Link function
  + 3. Relationship between mean and variance
  + 4. Within subject association among repeated measures.

28.6: GEE in STATA: menu looks very similar to the GLM menu

* The logit link for a binary outcome is the most commonly used because it forces the probability to stay between 0 and 1.
* For GEE models the default in Stata is to use model based variance, while the default in SAS is to use empirical variance.

28.9 Marginal vs. Conditional models

* In the logistic case the marginal and conditional model approaches are not the same.
* The estimated effect is larger in the mixed effects logistic model, but the standard error is similarly increased so p-values are similar.
* Each approach is answering a slightly different scientific question: e.g. Assume we are interested in the effect of a treatment on the probability of getting a lesion:
  + Marginal model (xtgee): On average in the population, does treatment B reduce the chance of a lesion?
  + Conditional model (xtmixed): In a specific patient, does treatment B reduce the chance of a lesion?

**Key graphic**: The red curves are what you would get from a conditional model (one for each patient; these are just 3 example patients), since in the mixed effects, we include a random effect, which allows each patient to have his own y-intercept (which appears as an x-axis offset in the logistic regression equation) but constrains the slope (β1) to be the same for each patient. The blue curve is the marginal model fit, which just gives the population average. The reason you don’t get the same result for the two methods (like you do get for the unity link-function case) is that the sigmoids don’t average in the same way that lines do.



End of lecture #28

Lecture #29: Time-to-event outcomes Paul Catalano, Sc.D.; DFCI/HSPH started 10 May 10, 2016

Survival analysis = time-to-event analysis; techniques for positive-valued random variables, such as:

time to death, time to onset, length of stay, duration of a strike, etc.

Two key differences with other kinds of data:

* 1. typically, survival date are not fully observed, but rather a subject to *censoring*:
     + Think of every subject as having 2 survival times: a “true” one, T, and another one that may be the only one we get to observe, U. If the ‘T’ comes before the ‘U’, then we get to observe the true time, but we might only get to observe out to ‘U’ in which case we say that it is right-censored. A random variable, X, is called a **censored failure time random variable** if X = min(T,U). Censoring can happen because the study was closed before the subject had the event, or because the subject left the study without experiencing the event.
  2. positive-valued random variables are NOT normally distributed

To discuss a failure time we need:

1. an unambiguous **time origin**
2. a **time scale** (could be real time or could be something like mileage of a car)
3. definition of the **event** (e.g. death; need a new set of tires)

Different kinds of censoring:

1. Right-censoring:
   1. Xi = min(Ti,Ui) is observed due to
      1. loss to follow-up
      2. drop-out
      3. study termination
   2. often use a binary code to indicate if patient was censored or not
      1. ‘δ’ codes ‘failure’ (1 if Ti < Ui)
      2. ‘c’ codes censoring (0 if Ti < Ui)
   3. Three types of right-censoring:
      1. Type I: all the Ui’s are the same: study closes at a fixed time
      2. Type II: Ui = T(r), the time of the rth failure: animal study that stops when 4/6 animals have tumors; NOTE: power of time-to-event studies depends on the number of events observed, not on the number of subjects entered into the study
      3. Type III: the Ui’s are random variables (a.k.a. ‘randomly censored’)
2. Left-censoring
   1. only observe Yi = max(Ti,Ui)
   2. example: study of age at which children learn a task: some already knew before study started (left-censored), some learned during study (exact), some had not yet learned by end of study (right-censored)
3. Interval-censoring
   1. observe (Li,Ri) where Ti ∈ (Li,Ri)
   2. i.e. we only observe times *around* the event, not the event itself
   3. ex. 1: time to prostate cancer, observe longitudinal PSA measurements
   4. ex. 2: detecting recurrence of colon cancer after surgery; follow patients every 3 months after resection of primary tumor

STATA: ‘**stset**’ command is key to setting up these analyses

29.5 Four key terms used in survival analysis; all are related to each other:

* density function: f(t)
  + probability that an event happens between t and Δt
  + tend to look exponential
* survivor function: S(t)
  + the cdf version of the density function, F(t) = P(T ≤ t)
  + S(t) = 1 – F(t) = P(T > t) = integral from t-to-infinity of f(u) du
* hazard function: λ(t)
  + probability that an event happens in an interval given that the person has already survived up to that point
  + a.k.a.: “instantaneous failure rate”, “force of mortality”, “age-specific failure rate”
  + λ(t) = f(t) / S(t)
  + examples of typical hazard functions:
    - increasing: e.g. mortality after age 65
    - decreasing: e.g. survival after surgery
    - bathtub: e.g. age-specific mortality (decreasing, then increasing)
    - constant: e.g. survival of patients with advanced chronic disease
* cumulative hazard function: Λ(t)
  + cumulative risk of failure: adding up the λ’s from 0 to t
  + S(t) = e-Λ(t)

Measuring central tendency in survival:

* mean survival,
* median survival, τ, defined by S(τ) = 0.5
  + time at which half of the patients have failed
  + in practice, we find the smallest time, τ, such that S(τ) ≤ 0.5

Estimating the survival or hazard function:

1. by specifying a parametric model for λ(t) based on a particular density function f(t)

* Some parametric survival distributions include:
  + **exponential**: f(t) = λe-λt for t ≥ 0
    - = e-λt
    - λ(t) = f(t) / S(t) = λ
    - i.e. the exponential distribution has a constant hazard rate
    - Λ(t) = λt
    - median, τ = -log(0.5) / λ = 0.69/λ
    - mean, μ = 1/λ
  + Weibull, Rayleigh, compound exponential, and log-normal, log-logistic
* One can usually distinguish between a one-parameter model (like the exponential) and two-parameter (like Weibull or log-normal) in terms of the adequacy of fit to a data set.
* Without a lot of data, it may be hard to distinguish between the various 2-parameter models (i.e., Weibull vs. log-normal).

1. by developing an empirical estimate of the survival function (i.e. non-parametric estimation)
   1. the empirical est. of S(t) is the proportion of individuals with event times greater than t
   2. The Kaplan-Meier (or KM) estimator is probably the most popular approach for estimating a survivorship function without resorting to parametric methods
      1. product of conditional probabilities
      2. example of 21 leukemia patients receiving control treatment, times to remission:
         1. 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23
         2. How would we estimate S(10), the probability that an individual survives to time 10 weeks or later?
            1. 8 survived past 10 weeks, so 8/21
         3. What about S(8)? is it 12/21 or 8/21?
         4. Construct a table of S(t):

|  |  |  |
| --- | --- | --- |
| Values of t | Fail | S(t) |
| t ≤ 1 | 0 | 21/21 = 1.000 |
| 1 < t ≤ 2 | 2 | 19/21 = 0.905 |
| 2 < t ≤ 3 | 2 | 17/21 = 0.809 |
| 3 < t ≤ 4 | 1 | 16/21 = 0.762 |
| 4 < t ≤ 5 | 2 | 14/21 = 0.667 |
| 5 < t ≤ 8 | 2 | 12/21 = 0.571 |
| 8 < t ≤ 11 | 4 | 8/21 = 0.381 |
| 11 < t ≤ 12 | 2 | 6/21 = 0.286 |
| 12 < t ≤ 15 | 2 | 4/21 = 0.190 |
| 15 < t ≤ 17 | 1 | 3/21 |
| 17 < t ≤ 22 | 1 | 2/21 |
| 22 < t ≤ 23 | 1 | 1/21 |
| 23 < t ≤ 24 | 1 | 0/21 |

in STATA:

. use leuken

. stset remiss status if trt==0

. sts list % generates above table

. sts graph % shows empirical survival function; K-M curve

[Pausing here; need to resume at lecture 29.10]

What do we do if there is censoring?

6\*,6,6,6,7,9\*,10\*,10,11\*,13,16,17\*,19\*,20\*,22,23,25\*,32\*,32\*,34\*,35\* (\* = right censored; n=21)

We know that S(6) = 21/21, because everyone survived until at least this time point. But we can’t say that S(7) = 17/21, because we don’t know the status at time 7 of the one person who was censored at time 6. Three failed and one was censored.

1958 paper in J. Amer. Stat. Assoc., Kaplan & Meier proposed a way to non-parametrically estimate S(t), even in the presence of censoring, using ideas based in conditional probability.

P(A∩B) = P(A|B)\*P(B), extended to more than two events

Suppose A1, A2 . . . Ak are k different events, then the probability of all k events happening together can be written as the product of conditional probabilities:

P(A1,∩A2 . . . ∩ Ak) = P(Ak|Ak-1 ∩ . . . ∩A1)

× P(Ak-1|Ak-2 ∩ . . . ∩A1)

. . .

× P(A2|A1)

× P(A1)

These conditional probabilities can be read off from the hazard function, as in:

where dj is the number of deaths at aj

and rj is the number **at risk** at aj

Key idea: you use the censored people for as long as you can; i.e. they appear in the “at risk” denominator for as long as they should (until they drop out)

As the intervals get finer and finer, the approximations made in estimating the probabilities of getting through each interval become smaller and smaller, so that the estimator converges to the true S(t). Hence an alternative name for the KM estimator is the “product limit estimator.”

[Interrupted: pick up at 29.11]

29.11: Calculating the KM: Cox & Oakes example

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| τj | dj | cj | rj | 1 – (dj/rj) | Ŝ(τ+j) |
| 6 | 3 | 1 | 21 | 1 – (3/21) = 0.8571 | 0.8571 |
| 7 | 1 | 0 | 17 | 1 – (1/17) = 0.941 | 0.8067 |
| 9 | 0 | 1 | 16 | 1 – (0/16) = 1.0 | 0.8067 |
| 10 | 1 | 1 | 15 | 1 – (1/15) = 0.933 | 0.753 |

KM survival function is at τj+, i.e. just **after** the time of the jth failure

End of Lecture #29

Lecture #30 Comparison of Survival Curves Paul Catalano, Sc.D. watched 12 May 2016

If NO censoring: Ŝ(t+) = (# individuals with T > t) / (total sample size)

If censoring, we drop individuals out of the denominator at the times they are censored

**Sum**: Only failures are informative about S(t); censored events just effect the denominator (at risk pool)

The basic calculation is just an estimated probability from a binomial distribution:

(recall: binomial, p, ≅ N[p, p(1-p)/n])), so, Ŝ(t) ≅ N(S(t), S(t)[1-S(t)] / n)

KM estimator is:

where λj = dj / rj

Each λj is just a binomial proportion, so we can apply standard formula for variance:

var(λj) = [λj(1- λj)] / rj

Use delta-method to get from var(λj) to var(Ŝ(t)), which is known as Greenwood’s Formula:

STATA example: leukem.dta

. use leukem

. describe % provides a description of the data set

. stset remiss status % declares data to be failure type

. stdes % gives a description of the survival data set

. stset remiss status if trt==1 % limit to only patients that received treatment

. sts list % KM in table form, with std. error and conf. intervals

. sts graph, ci % plot KM curve with confidence intervals

30.4: comparison of KM curves for two groups (e.g. treatment vs. control): logrank test, Wilcoxon test

General framework for survival analysis:

For each individual, we observe: (Xi, δi, Zi):

* Xi is a censored failure time random variable
* δi is the failure/censoring indicator
* Zi represents one or more covariates

General idea is to build a model that characterizes the relationship between survival and all of the covariates of interest. Simplest form of Zi is a single, binary variable (e.g. treatment vs. control).

* Two-group comparisons
* Multi-group and stratified comparisons
* Failure-time regression models
  + Cox proportional hazards

30.5 Mantel-Haenszel Logrank test (most common test for comparing 2 KM curves)

build on logic of 2 x 2 contingency table (Chi-square test)

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Event: yes | Event: no | Total |
| Control (0) | d0 | n0 – d0 | n0 |
| Treatment (1) | d1 | n1 – d1 | n1 |
| Total | d | n – d | n |

Under fixed margins, d0 follows a **hypergeometric distribution**. Under H0 of no association between event and group, it follows that:

E(d0) = n0d / n

Var(d0) = n0n1d(n-d) / n2(n-1)

Mantel-Haenszel Χ2 = [(observed d0) – (d0 expected under H0)]2 / Var(d0)

plug in above formulas = [d0 – E(d0)]2 / Var(d0)

Now, suppose we have K (2x2) tables: just sum up all the Χ2 values

For survival analysis, we just make a 2x2 table for each distinct failure time

[interrupted; resume at Lecture 30.6]

**Logrank test**: construct a 2x2 table at each distinct failure time, compare the failure rates between the two groups (conditional on number at risk in the groups), then combine the tables using the Cochran-Mantel-Haenszel (CMH) test. (i.e. just add a subscript *j* to each element in the 2x2 table, to represent the jth time)

In practice, calculate (0 – E) and Var(d0) for each time point, then just add them all up and compute test statistic and compare to Χ2 distribution with 1 DoF.

Lecture 30.7: Notes about logrank test, using STATA, linear rank tests

* The logrank statistic depends on ranks of event times ONLY (actual time-value does not matter)
* Numerator is classic (Observed – Expected), where expected number = #deaths x proportion at risk in group 0
* The (o – e) term can also be written as:
* It does not matter which group you choose to sum over
* logrank test is most powerful for proportional hazards

in STATA:

. stset remiss status

. sts list, by(trt)

. sts test trt

Generalization of logrank test → linear rank tests:

logrank and other tests can be derived by assigning scores to the ranks of the death times, and are members of a general class of linear rank tests.

First, define the estimated cumulative hazard:

where dj and rj are the numbers of deaths and the number at risk at the jth ordered death time.

Then assign these scores:

|  |  |
| --- | --- |
| **Event** | **Score** |
| Death at tj | wj = 1 – Λ(tj) |
| Censoring at tj | wj = 0 – Λ(tj) |

To calculate the logrank test, simply sum up the scores for group 0 (group is arbitrary).

Example: Group 0: 15,18,19,19,20; Group 1: 16\*,18\*,20\*,23,24\*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ordered Data | Group | dj | rj | Λ(tj) | score wj |
| 15 | 0 | 1 | 10 | 0.100 | **0.900** |
| 16\* | 1 | 0 | 9 | 0.100 | -0.100 |
| 18 | 0 | 1 | 8 | 0.225 | **0.775** |
| 18\* | 1 | 0 | 7 | 0.225 | -0.225 |
| 19 | 0 | 1 | 6 | 0.558 | **0.442** |
| 19 | 0 | 1 | 6 | 0.558 | **0.442** |
| 20 | 0 | 1 | 4 | 0.808 | **0.192** |
| 20\* | 1 | 0 | 3 | 0.808 | -0.808 |
| 23 | 1 | 1 | 2 | 1.308 | -0.301 |
| 24\* | 1 | 0 | 1 | 1.308 | -1.308 |

The logrank statistic, S, is the sum of score for group 0:

S = 0.900 + 0.775 + 0.442 + 0.192 = 2.75

Var(S) = 1.210, so

Z = 2.75 / sqrt(1.210) = 2.50 → Χ2logrank = (2.50)2 = 6.25

Consult the lecture notes (PDF) for the intuition about why the two approaches give the same answer, and why it makes sense to call it a “logrank” test—but should really be called the “log-survival via ranks” test.

30.8: Gehan’s generalized Wilcoxon test (related to Wilcoxon Rank Sum test and Mann-Whitney U test)

start with WRST: order combined sample of all observations – Consult lecture notes! This is great stuff, but two much notation to write into this MS Word doc. Lecture notes: **Catalyst\_Catalano\_2013\_lect2\_final.pdf**

in STATA:

. sts test varlist, [logrank|wilcoxon|cox]

. stset remiss status

. sts test trt, Wilcoxon

Which test should we use?

* Both tests have the right Type 1 error for testing HO of equal survival
* The choice therefore depends on the alternative hypothesis, which will drive the power of the test.
* Wilcoxon is sensitive to early differences between survival, while the logrank is sensitive to later ones. This can be seen by the relative weights they assign to the test statistic:
  + logrank numerator = Σ(oj – ej)
  + wilcoxon numerator = Σrj(oj – ej)
  + i.e. Wilcoxon is weighted by the total number of people in the risk pool, which is larger at the beginning of the study.
  + logrank is most powerful under the assumption of proportional hazards, which implies an alternative in terms of survival functions of Ha : S1(t) = [S2(t)]α
  + wilcoxon has high power when the failure times are log-normally distributed
  + Both tests will lack power if the survival curves “cross,” but that does not necessarily make them invalid!

30.9 P-sample and stratified logrank tests

* + - 1. What if there are more than two groups? (p samples, not just 2 samples)
         1. same approach, but use Px2 contingency tables
         2. STATA: . sts test group, logrank
         3. P-sample Wilcoxon test is also available
      2. Sometimes, even though we are interested in comparing two groups, we know there are other factors that also affect the outcome, and we want to adjust for these other factors. (e.g. patients treated at different hospitals). In this case, we want to **stratify** by these other factors.
         1. A stratified logrank allows the shapes of the hazards in the groups to differ across strata. It makes the assumption that the group 1 vs. group 2 hazard ratio is constant across strata.
         2. First, divide the data into S separate groups, and within each group we do our logrank test for the association of our variable, X, and survival.
         3. e.g. We want to test effect of age on discharge from nursing home, but we expect a difference due to sex.
         4. sum up all the (O – E) differences for each stratum separately, but then just add them all together to get the **stratified logrank** statistic

STATA example:

. use nurshome

. gen age1=0

. replace age1=1 if age>85

. sts test age1, strata(gender)

End of lecture #30

Lecture 31: Cox Proportional Hazards Regression Paul Catalano, Sc.D. 16 May 2016

goal: model the relationship between explanatory variables and survival (outcome)

* Proportional Hazards (PH) models:
  + λ(t,Z) = λ0(t)Ψ(Z), where Z is our vector of predictor variables
  + more commonly we right 2nd term as: Ψ(Z) = eβZ (because must be positive, since we are modeling an outcome variable which is always positive valued)
  + Suppose Z = 1 for treated subjects and Z = 0 for untreated subjects. Then our model say that the hazard function is modified by a factor of eβ for treated subjects versus untreated subjects. Note that eβ may be < 1.
  + This is an example of a **semi-parametric** model.
  + **Z** is a set of covariates and could include:
    - Continuous factors (age, blood pressure)
    - Discrete factors (gender, marital status)
      * same use of dummy (indicator) variables for different levels of a factor
      * main diff. is that β0 is implicit (embedded) in λ0
    - Possible interactions (age by sex interaction)
* Focus on Cox PH model:
  + What does the term λ0(t) mean?
    - It is the “baseline” hazard function
    - e.g. hazard rate for the control group
    - In general, reflects the underlying hazard for subjects with all covariates = 0
    - We can write the log of the hazard ratio for the ith individual to the reference group as:
      * And now we have standard linear regression, where we have just transformed our outcome variable as the log of the hazard ratio
      * This allows us to estimate the parameters, β, without having to make any assumptions about the form of the baseline hazard
  + What is “proportional” about the PH model?
    - Suppose Z = 1 for treated subjects and Z = 0 for untreated subjects. Then our model says that the hazard function is modified by a factor of eβ for treated subjects versus untreated subjects.
    - λ1(t) = λ0(t)exp(β), or:
    - λ1(t) / λ0(t) = exp(β) which is often written as:
    - , where phi (φ) is the hazard ratio and we see that this ratio is a constant that does NOT depend on time. In other words, this PH model assumes that the two groups remain proportional over time.
    - If we take the log of both sides , we see that β is the **log hazard ratio**.
    - In general, how do we think about βs in regression models?
      * linear regression: slopes (multipliers of the mean response)
      * logistic regression: log-odds ratios
      * PH model: log hazards ratio
  + How do we estimate parameters?
    - The basic idea is that under PH, information about βs can be obtained from the relative orderings (i.e. ranks) of the survival times, rather than the actual values.
    - Kalbfleisch and Prentice derived a likelihood involving only β and Z (not λ0(t)) based on the distribution of **ranks** of the observed failure times (in the absence of censoring)
    - D. R. Cox (1972) derived the same likelihood and generalized it for censoring, using the idea of **partial likelihood**
    - What is “partial likelihood”?
      * Intuitively, it is a product over the set of observed death times of the conditional probabilities of seeing the observed deaths, given the set of individuals at risk at those times (i.e. eliminate subjects who have been censored prior to that time)
    - Analogous to standard likelihood theory, it can be shown (with some work) that:
      * so now we can construct confidence intervals and test hypotheses
  + How do we interpret the estimated values?
    - e.g. in leukemia example (below), φhat=0.221, which means that the hazard for relapse among the treated patients is less than 25% of that for the control patients
    - NOTE that the Cox model does not give us the actual estimates of the survival function that we got from ‘**sts list**’, because we have factored this out of the analysis by using the ratio of the hazard functions—in this sense, the Kaplan-Meier curve and Cox regression are complementary and generally used together.
  + How can we construct tests of whether the covariates have a significant effect on the outcome?
    - so now we can construct confidence intervals and test hypotheses
  + How do these tests compare to the logrank test or the Wilcoxon test?
    - The two analyses (Cox & logrank) are very tightly connected to each other
    - the logrank test = “score” test from the Cox model
    - STATA: **sts test, cox**
    - Just as a Χ2 test for binary data can be derived from a logistic model, the logrank test can be derived as a special case of the Cox PH model.

Fitting a Cox PH model using STATA:

* uses the ‘**stcox**’ command
* basic form is the same as for any other kind of modeling (linear, logistic, etc.)
* as with previous tests, use ‘**stset**’ to set up the data
* e.g. leukemia data set:
  + control group has ‘trt’ (for “treatment”) = 0
  + if treatment works, we would expect the hazard rate for the treated group to be lower than that for the control group (i.e. hazard ratio < 1; log hazard ratio would be < 0)
  + . **stcox trt**
    - H0: hazard ratio = 1
  + if you want the actual βs: . **stcox trt nohr**
    - exact same test
    - H0: log hazard ratio = 0

Assessing the PH Assumption

* A major conceit of the PH model is that we can split the “real” hazard function, λ(t,Z), which depends on both time, t, and our covariates, Z, into two parts:
  + one of which depends on time, λ0(t), but not on the covariates
  + one which depends on the covariates, exp(βZ), but not on time
  + By looking at the **ratio** of hazard functions, we get rid of λ0(t)
* Several options for checking the assumption of PH:
  + Graphical:
    - plots of the survival curves for the two groups
    - plots of log(-log(S)) vs. log(t) for the two groups
      * recall that we can write the Survival function version of the PH model:
        + log(-log(S(t;Z)) = log(-log(S0(t)) + βZ

because S(t) = -logΛ

* + - * + This suggests a diagnostic: a plot of the one ‘log, minus log’ term vs. the other over time, we should see a constant offset

calculate Kaplan-Meier curves (S) for various levels of Z

compute log[-log(S(t;Z)], i.e. log cumulative hazard

plot vs. log-time to see if they are parallel

* + - * This is useful, because it is easier to assess a constant offset (visually) than to assess a constant proportionality
      * in STATA: . **stphplot, by(gender)**
      * NOTE that we do not have to fit the Cox model to do this—it is really just a comparison of a derived value taken directly from our estimate of the Survival function, S(t), i.e. the Kaplan-Meier curve
    - plots of weighted Schoenfeld residuals vs. time
    - plots of observed survival probabilities vs. expected under the PH model
  + Goodness-of-fit tests
  + Including interaction terms between a covariate and t (time-dependent covariates); i.e. explicitly include time in the exp(βZ) part of the model

31.8 What if PH assumptions fail?

* Do a stratified analysis?
  + Suppose we are happy with the proportionality assumption of Z1, but not for various levels of a second variable, Z2.
  + The stratification comes into play through the baseline hazard function:
    - λ(t;Z1,Z2) = λZ2(t)exp(βZ1)
  + in STATA: . **stcox z1, strata(z2)**
  + downside: no hazard ratio for Z2, because it’s not in the parametric part of the model
  + example: A new treatment might lead to a 50% decrease in hazard of death vs. the standard treatment, but the hazard for standard treatment might be different for each hospital
    - stratified model can be useful for both primary analysis *and for checking the PH assumption*
      * same trick as before: estimate the survival function, S(t), for each level of the stratified variable, Zk. Advantage is that we use the βZ part of the model to adjust for all of the other covariates.
      * Then we compute [log(–log(S(t))] for each level of Zk, controlling for the other covariates in the model, and graphically check whether the log cumulative hazards are parallel across stratification levels
      * STATA example: nursing home data; check PH assumption for gender, adjusted for other covariates:
        + include ‘married’ and ‘health’ as covariates in a Cox PH model, but stratify by ‘gender’
        + Calculate the baseline survival function for each level of ‘gender’
        + plot the log-cumulative hazards for males and females and evaluate whether the lines are parallel
    - by stratifying, we measure the treatment effect within each hospital separately
    - assumes that the hazard ratio is the same for each hospital
* Include a time-varying covariate to allow changing hazard ratios over time
* Include interactions between a predictor variable and time

End of lecture #31

Lecture #32 Power analysis / Study design with survival outcomes Brian Healy, PhD 16 May 2016

First half of lecture: example: survival analysis from beginning to end—I skipped these lectures and picked up at 32.7

32.7 Study Design

Big Picture

* In general, power/sample size calculations necessary to determine how many patients to enroll. Need to know four of these in order to calculate the fifth
  + Type I error rate (α)
  + Power = 1 – Type II error rate (β)
  + Sample size
  + Effect size (difference between groups)
  + Variability

in STATA: **sampsi** mu1 mu2, sd1(#) sd2(#) power(0.80)

* Studies with time-to-event outcome have some special design considerations:
  + censoring: need to know the probability of an event occurring
  + total follow-up: if patients withdraw, need to increase sample size proportionally
  + accrual (time over which we enroll patients → variability in observation times)
  + What do we use as our estimate of effect size?
    - hazard ratio (Freedman)
    - log of the hazard ratio (Schoenfeld)
    - These approaches calculate the number of **events** we need to observe, NOT the number of subjects we need to enroll. The latter depends on censoring, withdrawal, accrual.
  + STATA: . stpower logrank, hratio(2)
  + for data with *administrative* censoring, we need to take into account the probability of an event in order to calculate the number of patients we need to enroll:
    - stpower logrank 0.7 0.5
    - stpower logrand 0.7, hratio(2)
  + if some patients *withdraw*, need an estimate of this proportion, *w*, in order to adjust the sample size accordingly
    - stpower logrank 0.7 0.5, wdprob(0.1)
  + if we *accrue* patients over time but follow-up ends at a specific time; e.g. For a study of time-to-relapse over one year, we enroll subjects uniformly over two years.
    - This means that the first subject enrolled would have 3 years of follow-up and the last subject would have only one year of follow-up.
    - in STATA, we must specify the survival probability at each of 3 time-points: min, average and max follow-up
    - stpower logrank, hratio(2) simpson(0.7 0.49 0.34)

End of lecture #32

Lecture #33 Multiple Comparisons Brian Healy, PhD Monday, 23 May 2016

Key Questions:

* + - 1. What is the null hypothesis?
      2. How many tests have I done?
      3. Which comparisons am I interested in?
         * all pair-wise comparisons (Tukey)
         * all treatments to a specific control group (Dunnett’s correction)
         * special case: if only 3 groups and we reject H0 for the ANOVA, we can do the pairwise comparisons without further correction (‘closed testing procedure’; Bender & Lang)

Review of ANOVA (studies involving many groups: multiple comparisons between groups)

* Comparison wise error rate is the type I error rate for each comparison.
* Family wise error rate (FWER) is also called experiment error rate (EER).
* Family wise error rate is the type I error rate for the entire group of comparisons.
  + FWER = 1 – (1 - α)n, where n is # of comparisons

Studies involving multiple outcomes (e.g. genomics, MRI)

* key concept: p-values are uniformly distributed under H0
* difference between studies that are hypothesis-generating vs. hypothesis-testing
* Bonferroni always works, but at the expense of a large number of Type II errors (i.e. low power or too conservative)
* Alternative approaches:
  + Holm’s correction: Bonferroni step-down procedure
    - we compare our smallest p-value to α/n, if this is significant, we reject H0 for this test and then continue, comparing next smallest p-value to α/(n – 1)
    - continue until first fail-to-reject H0
  + Hochberg’s correction: Bonferroni step-up procedure
    - multiply highest p-value by 1, next highest by 2, etc.
    - slightly more powerful than Holm’s
    - similar to false discovery rate (FDR)
  + False Discovery Rate (Benjamini & Hochberg 1995):
    - sensitivity: P(T+/D+) vs. positive predictive value (PPV): P(D+/T+)
    - FDR is analogous to PPV: P(H0 is true / H0 was rejected)
    - Two handy features:
      1. FDR is calculated based on the p-values from the multiple tests
      2. more appropriate in cases where we expect H0 to be false for a fair number of features (e.g. genetic studies)
      3. If we choose an FDR of 0.05, it means that we would like no more than 5% of the features that we classify as significant to be false positives
         * This is not the same as choosing the FWER to be 5%
         * As with FWER, we can either use it to set a criterion or we can use it to adjust our p-values
    - FDRi = (# of false positives) / (# called significant)
      1. sort p-values and find the ith ordered p-value (= pi)
      2. if we chose this p-value as the cut-off for significance, we would have i significant p-values (by definition, due to sorting)
      3. if all H0’s are true (**uniform distribution of p-values**), then the expected # of false positives would be: n x pi
      4. Thus FDRhati = n x pi / i
      5. To control at a specific value of FDR (called q\*), we simply find the largest value of i (call this value ‘k’) for which P(i) ≤ (i/n) x q\*
         * then reject all H(i) i = 1, 2, . . . , k
      6. If all H0 are true, then FDR = FWER
      7. FDR is appropriate when a single falsely rejected H0 (i.e. false positive) is not too dangerous – see my MATLAB code: FDRdemo.m

**False Discovery Rate and the q-value:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Called significant  (H0 rejected) | Called not significant  (H0 accepted) | **Total** |
| H0 True | F  False Positive  "False Alarm"  Type I error | m0 – F  True Negative  "Correct Rejection" | **m0** |
| HA True  (H0 False) | T  True Positive  "Hit" | mA – T  False Negative  "Miss"  Type II error | **mA** |
| **Total** | **S** | **m – S** | **m** |

The table above helps us think about hypothesis testing in general. It gives all four possible outcomes of such a test, with the rows corresponding to the "ground truth" and the columns corresponding to how our test classifies the comparison. [**Note**: This same formalism is used in "Signal Detection Theory," which was developed by physicists and is now heavily used in the psychology of perception. I've included the corresponding terms in quotation marks.]

Most folks are comfortable with P-values, which give the probability of wrongly rejecting H0, also known as the *false positive rate (FPR)*. A related, but critically different, measure is called the *false discovery rate (FDR)*, which gives the proportion of tests labeled as "significant" when, in fact, H0 was true. In both cases, the numerator is the number of false positives, *F*, but, for the former, the denominator is the upper row total (m0), and, for the latter, it is the left column total (S).

**FPR** = F / (F + m0 – F) = F / m0 **FDR** = F / (F + T) = F/S

Related values that are also used to discuss this issue:

*sensitivity* = T / mA *specificity* = (m0 – F) / m0

**Q-value**:

* an extension of the original Benjamini & Hochberg approach
* rather than using the total # of tests to calculate the number of false positives, Storey estimates the number of null hypotheses using the data (Empirical Bayes!):
  + FDRhati = minp≥pi [(nhat0 x pi) / i]
  + a.k.a. q-value
  + The derivation is relatively straightforward and can be found in Storey (2002) or Storey & Tibshirani (2003)

The same table above using the annotation of Storey 2001 (classic on FDR)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Called significant  (H0 rejected) | Called not significant  (H0 accepted) | **Total** |
| H0 True | V | U | **m0** |
| HA True  (H0 False) | S | T | **m1** |
| **Total** | **R** | **W** | **m** |

33.10 Additional considerations

* Correlation among tests: How many comparisons do we truly have?
  + if tests are positively correlated, chances of a type I error **decrease**
  + intuition: imagine 20 tests on the height of two groups of subjects, where the 20 outcomes are identical, just measured in different units (in., cm., etc.). In this case, the outcomes are all perfectly positively correlated, and we are really only doing **one** test.
  + solutions:
    - reduce the number of hypotheses prior to analysis
    - resampling based approaches (**Nichols and Holmes article** on course web site)
      1. use a permutation test on the data to get the distribution of the test statistic under H0
      2. since the correlation within the observations is involved in the calculation of the maximum statistic, we gain power
* Clinical trials:
  + multiple endpoints
    - option #1: pick a primary end-point and test at p = 0.05; designate others as secondary (i.e. hypothesis generating)
    - option #2: create a summary measure (i.e. an index combining all outcomes)
      1. hard to interpret: What does the index mean?
    - option #3: test each end-point separately:
      1. version 1: only reject if ALL have p < 0.05 (very stringent)
      2. version 2: only reject if ANY have p < 0.05 (multiple comparison problem)
    - option #4: a global test, such as MANOVA
  + interim analyses
    - see Feldman lecture #19
    - alpha-spending rules: Pocock, etc.
* Genomics (or anything with ‘-omics’ or MRI): 3 common approaches
  + “genome-wide significance”: alpha-level that has been Bonferroni corrected
  + FDR correction for p-values
  + re-sampling based approaches of p-values
* Rothman (1990) in Epidemiology: one of most cited articles in statistics:
  + questions need to adjust for multiple comparisons
  + Are we ever really interested in the global H0?
  + Does chance truly cause the difference between groups?
  + Sounds whacky!

End of lecture #33

Lecture #34 Meta-analysis Brian Healy, PhD 31 May 2016

* Introduction to meta-analysis; STATA: **metan**
  + provides a statistically rigorous way to combine the results from multiple studies
  + gives a quantitative summary
  + less subject to bias
  + easy to interpret and replicate
    - requires reporting exactly how studies were identified
  + prevents unnecessary trials
    - Eliot Antman article in NEJM (see fig. 1 on beta-blocker trials)
  + Example: Are beta-blockers as effective as other antihypertensive treatments in preventing stroke, MI and mortality?
    - 13 randomized clinical trials provided the evidence
    - Mortality: y/n and beta-blocker: y/n → chi-square test (for a single trial)
    - How do we combine across trials? See below!
      1. just summing the cells across multiple 2x2 tables can lead to the wrong conclusion: Simpson’s paradox
      2. What we want:
         * an estimate that is a good summary of the tables
         * an estimate that gives more weight to tables with more information
         * some test of whether it is reasonable to combine at all
* **Fixed effects meta-analysis**
  + Mantel-Haenszel approach:
    - Step 1: calculate an estimate of the effect (odds ratios or risk ratio) in each group/strata
    - Step 2: test if estimates of effect are similar enough to combine
      1. test for homogeneity
         * H0: OR1 = OR2 = OR3 = . . . ORk
         * HA: at least one OR is different
      2. technically:
         * take log of OR
         * weight each OR by 1/variance
         * compute weighted average of OR’s (‘combined’ or ‘pooled’ estimate)
         * chi-square test: compare each OR to the weighted average
    - Step 3: if reasonable to combine, then combine:
      1. **weight OR’s by the inverse variance** (i.e. give more weight to more accurately estimated OR’s)
      2. Robins et al. formula for confidence interval for combined OR
    - I-squared statistic: alternative measure of the heterogeneity in effect size across studies (i.e. alternative to Step 2 above)
      1. estimate of the percentage of the variation in the effect estimate that is due to the heterogeneity across studies
      2. I2 = 100 x [(Q – df) / Q], where Q is the heterogeneity chi-squared
      3. rule of thumb: I2 < 25%: low, fixed effect is reasonable
    - **Forest plot**: graphical display of original data and the pooled data
  + Peto approach: slightly different method of weighting
* **Random effects meta-analysis**
  + Estimates from studies can differ due to sampling variability or between study differences.
    - If the estimates from studies differ only due to **sampling variability**, then a fixed effect works.
    - If there are between-study differences, then a random effects model is better.
  + Handling heterogeneity across studies
  + Der Simonian and Laird (1986) proposed an approach to combine effect size estimates based on a random effects model
  + This approach assumes there is a distribution of effect sizes rather than a single effect size
    - effect sizes across studies are normally distributed
    - we estimate the mean of the distribution of effects
    - two sources of variability in the effect size estimates:
      1. between-study difference
      2. sampling variability
    - as for the fixed-effects approach, we compute a weighted average of the effect sizes (i.e. weighted by inverse variance), but we change the way we calculate the weight:
      1. 1 / (D + Vi), where Vi is the sampling variability of the ith study, and D is the estimate of the between-study variability
  + because we allow for more variability, the confidence intervals for the random-effects model will be wider than those from fixed-effect approaches; thus the random-effects models are more conservative
* Fixed vs. Random: How do I choose?
  + fixed effect:
    - better if we believe each study is estimating the same underlying parameter
    - can give large studies very high weight
    - more powerful (vs. random-effect model)
    - all effect sizes, Yi, from same distribution: Yi ~ N(θ,Vi)



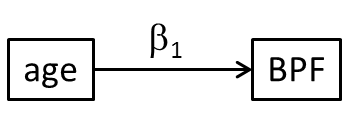
* + random effect:
    - best if we believe that there is variability in the effect estimate across studies
    - places more even weight across studies
    - more conservative
    - generally preferred
    - each study has its own θ: Yi ~ N(θi,Vi) and θi ~ N(θ,D)

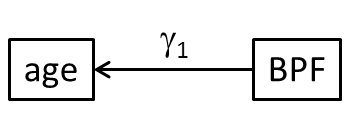
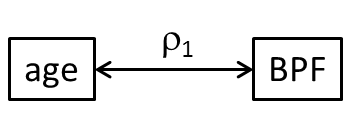


* + - see paper by Normand 1999 (I have the PDF.)
* Sensitivity analysis: How sensitive are the conclusions to the assumptions?
  + just playing with different ways of grouping different treatments, etc.
  + sounds like a recipe for p-hacking, but could also help describe sources of variability across studies
  + **Key point**: Important to try to understand the sources of heterogeneity across studies.
* other stuff: preparing for a meta-analysis (skipped); observational studies
  + observational studies:
    - since these are subject to confounding, we need to use adjusted estimates (i.e. controlling for other factors, such as age)
    - challenging if different studies adjusted for different potential confounders
  + meta-regression
    - effect size as dependent variable (outcome) and different study characteristics as predictor variables (weighted regression technique): STATA: **metareg**

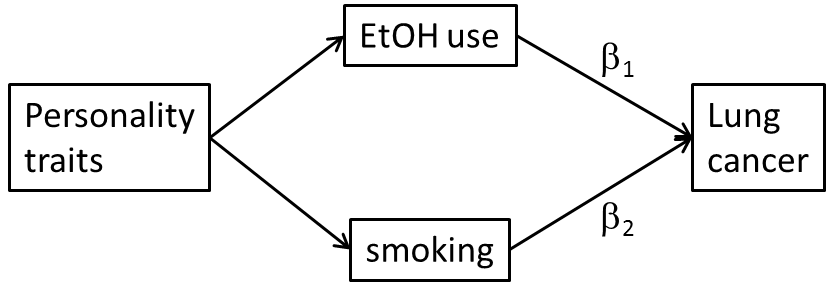
End of lecture #34

Lecture #35 Path analysis Brian Healy, PhD viewed 06 June 2016

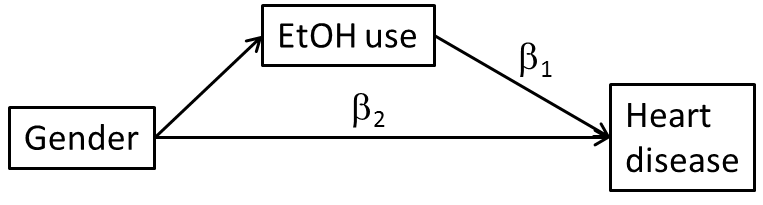
* One valuable approach for understanding the relationship between variables based on the proposed model is a **path diagram**
* Path diagrams can be used to represent many of the techniques that we have learned and others that you may have heard of, e.g., **Linear regression**.
  + Brief review of linear regression: BPFi = β0 + β1\*agei + ei
  + adding covariates allows us to control for different variables (e.g. sex): BPFi = β0 + β1\*agei + β1\*sexi + ei
    - assumes parallel lines
  + path diagram of simple model: 
  + But we could also reverse the direction of the arrow and do the regression in the opposite direction, or use a bidirectional arrow and simply look at the Pearson correlation

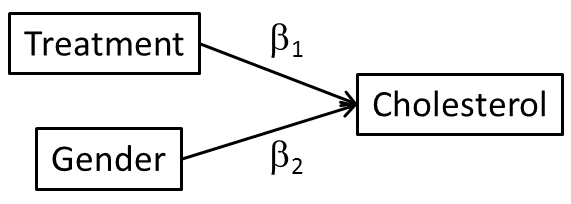
* + Association vs. Causation: How do we know the direction of the arrow?
    - In a randomized clinical trial, we can assess causation because we assign the treatment
    - Scientific knowledge: e.g. age/BPF
    - In path analysis, we assume as specific direction
    - Note: The diagrams in the Orav lecture (#15) on model selection were path diagrams
  + Directed acyclic graphs (DAGs) are a type of path diagram used in epidemiology to determine how to make causal inferences (see article on web site: **Hernan et al. 2002**; have PDF)
    - not allowed to have bi-directional arrows (hence, “acyclic”)
    - Example #1: relationship between EtOH and smoking and lung CA, we don’t say that EtOH/smoking are correlated (bi-directional arrow), we say that there is a **shared common cause**, such as “personality traits”, that lead to both:



* + - “shared common cause” is another definition of confounding
    - Example #2: a factor that causes both the predictor and the outcome:



* + - * gender is the common cause of both EtOH use and heart disease
      * regression model is the same (as for the correlation case)
    - Example #3: Two predictors are both associated with the outcome, but they are uncorrelated:



* + - * Since treatment is randomized, we know there is no relationship between gender and treatment
      * At the same time, we know that gender affects cholesterol, so including it in the model could reduce standard errors and improve our power.
  + Two extensions of multiple regression:
    - Interaction
      * implies that the impact of one predictor is changed by a second predictor
      * 2nd predictor a.k.a. “moderator variable”; in epidemiology: “effect modification”
      * e.g. lecture #14: relationship between age and weight might be different in males and females:
        + interaction term allows for different slopes for males and females
        + statistical model:

weighti = β0 + β1\*agei + β2\*sexi + β3\*sexi\*agei + ei

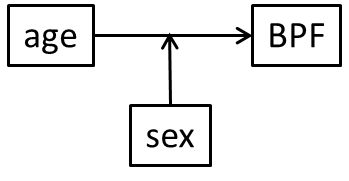
sex = 1 for males and 0 for females

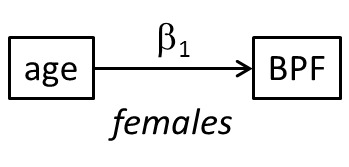
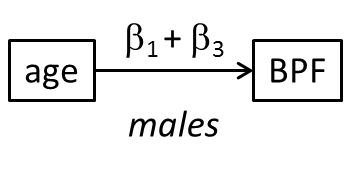
model for females: weighti = β0 + β1\*agei + ei

model for males: weighti = (β0 + β2)+ (β1 + β3)\*agei + ei

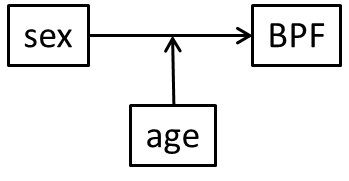
β3 is the effect of sex on the relationship between age and weight

* + - * + path diagram: two options for interactions:

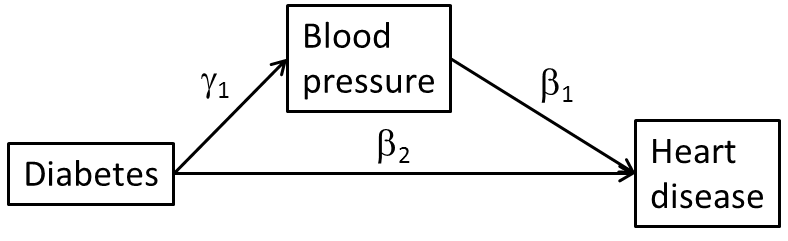


* + - * + In the present version, we state that the effect of age on weight is different in males and females, but we could put forward an alternative model stating that the BPF difference between males and females changes for different age values. (i.e. a main effect of gender that is influenced by age:



* + - * + for more on interactions, see articles by VanderWeele)
    - Intermediary variables
      * variables on the causal path between the predictor and the outcome
      * a.k.a. “mediator variable”



* + - * According to Orav, these should NOT be included in the model: If we are interested in the *total* effect of diabetes on heart disease, β1 would be stealing variance from β2. Put another way, we would be estimating the effect of diabetes on heart disease while controlling for blood pressure.
      * Two ways of estimating the full effect:
        + 1) fit the reduced model (per Orav)
        + 2) fit the full models, but combine the gamma and the two betas:

BPi = γ0 + γ1\*diabetesi

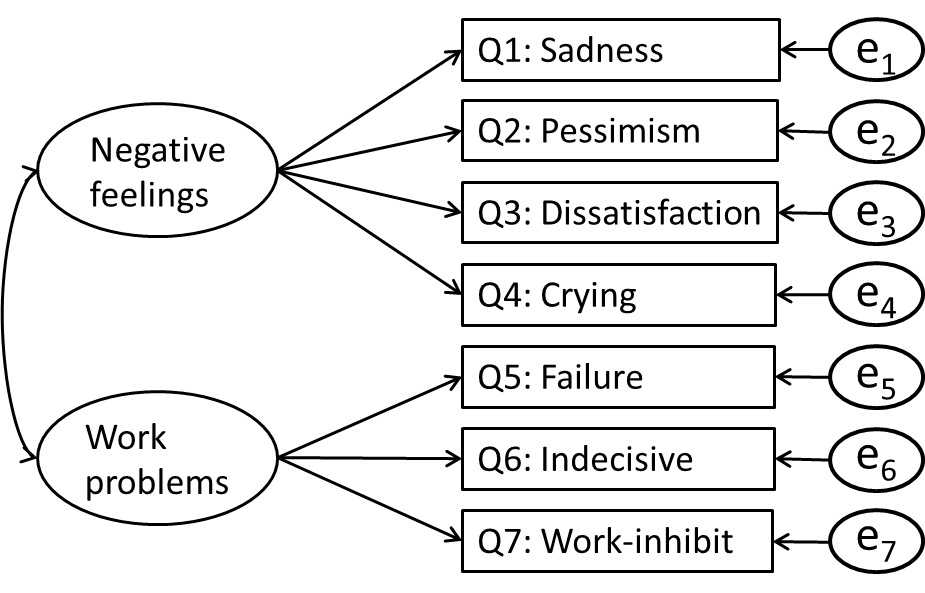
HDi = β0 + β1\*BPi + β2\*diabetesi

substitute for BPi in eqn. #2:

HDi = β0 + β1\*(γ0 + γ1\*diabetesi) + β2\*diabetesi

HDi = (β0 + β1\*γ0) + (β1\*γ1 + β2)\*diabetesi

* Path Analysis in STATA:
  + Structural Equation Model (SEM) builder: draw boxes and connect with arrows
  + command line: **sem** (HD <- diabetes) (HD <- BP) (BP <- diabetes)
* Two important extensions of models that we have learned are also investigated using path diagrams:
  + **mediation models**: assume the effect of one variable on the outcome is mediated by another variable
    - The diabetes example above is an example of mediation analysis:
      * diabetes can affect heart disease in two ways:
        + direct effect
        + indirect effect via blood pressure
    - Also common in psychology: How much of the effect of a predictor on an outcome is through the direct vs. indirect pathways? (see MacKinnon et al. on web site)
    - see also Baron & Kenny 1986
  + **structural equation models**: investigate the relationships between measured and latent factors
* **Factor analysis** (FA): an approach showing the relationship between unobserved variables and outcome variables
  + Exploratory FA: tries to identify factor structure from data:
    - trying to find factors underlying an index of depression
    - ten questions; look at correlation matrix to see which are grouped
    - slides for this on course web site
  + Confirmatory FA: proposes a specific model and allows us to test fit of proposed approach
    - based on previous exploratory FA, we found 3 latent factors related to the 10 questions
    - decide to validate results for the first 2 factors on a new sample of 252 subjects
      * we have a specific model in mind
      * goal is to confirm this proposed structure:



* + - * ovals represent “latent factors” – things we can’t measure
      * rectangles represent things we can measure
      * STATA: use ‘**sem**’ command to specify the model
* Goodness of fit for confirmatory factor analysis (CFA)
  + Several measures are available for CFA and SEM, in general
    - Tests for specific coefficients
      * STATA provides a z-statistic and p-value for each coefficient
      * p-values > 0.05 provide evidence that this part of the model is unnecessary
    - Global model fit statistics
      * Model chi-sq comparing model to saturated
        + rule-of-thumb: p-value > 0.05
      * RMS error of approximation
        + rule-of-thumb: p-value < 0.05
      * Comparative fit index (CFI)
        + rule-of-thumb: CFI > 0.9 or 0.95
* This part of the lecture was very rushed, but BH recommends the STATA manual on SEM
* Take home: ALWAYS DRAW THE DIAGRAM!

End of lecture #35

End of course