Matching &

Conditional Logistic Regression

Mabel Carabali

EBOH, McGill University

(updated: 2025-10-19)

find a matching pair

Expected competencies

- Knows why we use "matching" in epidemiology.
- Knows advantages and disadvantages of matching

Objectives

- To discuss advantages and disadvantages of matching
- To illustrate the use of logistic regression in presence of paired data
- To illustrate and discuss the use of (conditional) logistic regression for the analysis of matched/paired data

- Slides 3 - 58: Main content
- Slides 59 - 73: Additional worked example

Paired Data vs Matched data

Paired data

- Expected to account for *"known & unknown"* potential confounders
- Correlated - by nature
  - E.g., twins, same people pre-post observations, eyes in the same individual, etc.
- Variable correlation type
  - $t_1 \neq t_0$,constant, exchangeable, lagged, etc.

Matching

- Making comparable a set of subjects:

> *“Matching refers to the selection of a reference series – unexposed subjects in a cohort study or controls in a case-control study – that is* ***identical, or nearly so, to the index series*** *with respect to the distribution of one or more potentially confounding factors.”* (RGL 2008, p. 171)
>
> *"When estimating causal effects using observational data, it is desirable to replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar covariate distributions. This goal can often be achieved by choosing* ***well-matched*** *samples of the original treated and control groups, thereby reducing bias due to the covariates."* (Stuart E, 2010) Matching methods for causal inference: A review and a look forward

Why do we match?

The main objective of matching is to make the comparison groups as similar as possible on everything **except** the variable of interest.

- Address confounding
  - Remember **Exchangeability** ?
    $Pr(Y^x|X = 1) = Pr(Y^x|X = 0)$ or $Y^x \perp\!\!\!\perp X$
  - The related term, Ignorability , assumes that there are no unobserved differences between the treatment and control groups, **conditional on the observed covariates**.
- Also called Conditional Exchangeability : $X \perp\!\!\!\perp Y^{(0)}, Y^{(1)}|Z$

Why do we match?

> "*Although causal assumptions are often invoked when using matching,* ***matching is simply an adjustment method*** *that can be used regardless of whether these assumptions are met; it is* ***the interpretation of the estimated effect after matching*** *as causal that requires these assumptions.*"

Matching Methods for Confounder Adjustment: An Addition to the Epidemiologist's Toolbox, Epidemiologic Reviews, Volume 43, Issue 1, 2021, Pages 118–129

Two settings

(1) Outcome values are not yet available:

- Matching is used to select subjects for follow-up.
- Relevant for high cost studies or logistics considerations preventing data collection for the full control group.
- The basis for original theoretical work and developments, comparing selecting matched versus random samples of the control group.

(2) Outcome data is already available:

- The goal of the matching is to reduce bias in the estimation of the treatment effect.

- Outcome values are not used in the matching process!
- The matching can be done multiple times
- Best balance –*the most similar treated and control groups*– the final matched samples.

Matching methods for causal inference: A review and a look forward

Claimed Advantages of Matching for Causal Inference

by Liz Stuart:

- "*1, matching methods should not be seen in conflict with regression adjustment and in fact the two methods are complementary and best used in combination.*

- "*2, matching methods highlight areas of the covariate distribution where there is not sufficient overlap between the treatment and control groups, ... treatment effect estimates would rely heavily on extrapolation.*

    - *Selection models and regression models perform poorly when there is insufficient overlap, ... standard diagnostics do not checking this.*
    - *Matching methods in part serve to make researchers aware of the quality of resulting inferences".*

- *3, matching methods have straightforward diagnostics by which their performance can be assessed".*

Matching methods for causal inference: A review and a look forward

Matching by study design

Matching is often thought about as analogous to physical control in randomized experiments

Case Controls: The effect of matching in a case-control study is to introduce bias into the crude association, which is accounted for only by adjusting for the matching factors in the analysis.

- If the matching factors are associated with exposure, then matching on such factors will introduce a confounding-like bias that needs to be accounted for in the analysis.
- Conditional logistic regression is a multivariate regression approach which treats each matched pair as a separate stratum and is the analytic control of choice for the matching introduced bias.

Cohort studies: Matching exposed to unexposed subjects according to some matching factors requires no additional [1] analytic control for these matching factors in cohort studies.

[1] Debatable for some, but in principle.

Case–control matching: effects, misconceptions, and recommendations

1. Matching, even for non-confounders, can create selection bias;
2. Matching distorts dose–response relations between matching variables and the outcome;
3. Unbiased estimation requires accounting for the actual matching protocol as well as for any residual confounding effects;
4. For efficiency, identically matched groups should be collapsed;
5. Matching may harm precision and power;
6. Matched analyses may suffer from sparse-data bias, even when using basic sparse-data methods.

"Supporting advice to limit case–control matching to a few strong well- measured confounders, which would devolve to no matching if no such confounders are measured".

"On the positive side, odds ratio modification by matched variables can be assessed in matched case–control studies without further data, and when one knows either the distribution of the matching factors or their relation to the outcome in the source population, one can estimate and study patterns in absolute rates."

Mansournia, M.A., Jewell, N.P. & Greenland, S. Case–control matching: effects, misconceptions, and recommendations. Eur J Epidemiol 33, 5–14 (2018).

If matching INDUCES confounding and/or selection bias in case-control studies, why do we do it?

Efficiency !!

- By selecting only the most relevant controls, one save time and money (except if data is already collected).
- Some argue that matching provides non-parametric control (e.g., Ho et al., Political Analysis 2007)

How do we match people?

- Matching can involve subset selection (i.e., selecting units from the sample to retain and dropping the rest) or,
- Stratification (i.e., assigning units to pairs or strata containing both exposed and unexposed units);
    - Some methods, like pair matching, involve both.

Matching steps

Matching methods have four key steps: #1 to #3 for “design” and #4 “analysis:”

1. Defining “closeness”: the distance measure used to determine whether an individual is a good match for another,
2. Implementing a matching method, given that measure of closeness,
3. Assessing the quality of the resulting matched samples , and perhaps iterating with Steps (1) and (2) until well-matched samples result, and
4. Analysis of the outcome and estimation of the treatment effect, given the matching done in Step (3).

Matching methods for causal inference: A review and a look forward

1) Closeness (i)

Two main aspects to determining the measure of distance (or “closeness”) to use in matching:

A) Determining which covariates to include

- The key concept is strong ignorability.
- The assumption: no unobserved differences between the treatment and control groups, conditional on the observed covariates.
- Include all variables known to be related to both treatment assignment and the outcome.
- Should not include variables that may have been affected by the treatment of interest (Rosenbaum, 1984; Frangakis and Rubin, 2002; Greenland, 2003).

B) Combining those covariates into one distance measure.

- “Distance:” is a measure of the similarity between two individuals

Matching methods for causal inference: A review and a look forward

1) Closeness (ii)

Four primary ways to define the distance D_{ij} between individuals i and j for matching:

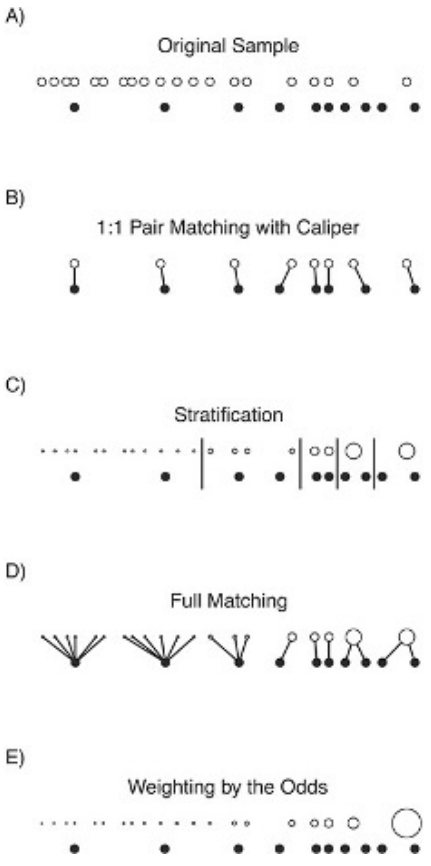
1. Exact: $D_{ij} = 0,\ if\ X_i = X_j;\ D_{ij} = \infty,\ if\ X_i \neq X_j$
2. Mahalanobis (Distance between a point and a distribution): distance of the vector from the mean divided by the covariance matrix to account for correlation
3. Propensity score [1]
4. Linear propensity score [1]

- These measures can be combined, e.g., exact matching on key covariates followed by propensity score matching within groups.
- When exact matching is not possible (e.g., sample size limitations), “fine balance” methods may be a good alternative (Rosenbaum et al., 2007).
- Exact matches often leads to many individuals not being matched, which can result in larger bias than if the matches are inexact but more individuals remain in the analysis.

[1] We have a lecture on this!

Matching methods for causal inference: A review and a look forward

Matching and weighting for the average exposure effect in the exposed



Exposed units (filled circles) and unexposed units (unfilled circles) are aligned horizontally by their propensity score. The size of the dots corresponds to the value of the resulting matching weights for the matching methods and propensity score weights for weighting by the odds.

Matching Methods for Confounder Adjustment: An Addition to the Epidemiologist's Toolbox, Epidemiologic Reviews, Volume 43, Issue 1, 2021, Pages 118–129

2) Implementing matching

- Type: “Individually matched” or “Group matched”
- The most commonly used design is 1:1 matched
  - Nearest neighbor matching
  - The order of matching for "treated" may change the quality of the matches

Selecting the number of matches involves a bias : variance trade-off.

- Multiple controls for each treated individual will generally increase bias since the 2nd, 3rd, and 4th closest matches are, by definition, further away from the treated individual than is the 1st closest match.
- Multiple matches can decrease variance due to the larger matched sample size.

With or without replacement: controls that look similar to many treated individuals can be used multiple times and replacement can decrease bias. Also, the order in which the treated individuals are matched does not matter.

- **Recall:** Once we match on certain factors, we are forfeiting estimating their effect

Impletment matching

Option	Benefits	Cautions
Matching on the covariates directly (e.g., Mahalanobis distance matching)	Can better balance the joint distribution of covariates; does not require an exposure model	May not perform well with many covariates, due to curse of dimensionality
Matching on the propensity score	Requires matching only on a single dimension; has theoretical balancing properties; tends to perform well empirically	Relies on specification of exposure model, pairs may not be close on covariates
Restrictions on closeness of matches	Can improve balance; yields close pairs; improves robustness to assumptions about outcome model	Dropping units decreases precision and can change the target population/estimand
Matching with replacement	Better balance than without replacement; good with small unexposed samples or when ratio of exposed to unexposed is high	Reusing units decreases precision; increases reliance on a few units
k:1 matching	Retains more units, thereby increasing precision	Balance can be worse

3) Assess quality of the matching (i)

Covariate balance and effective sample size.

- **Balance:** The "standardized bias" or "standardized difference in means" (SDM)
  - The difference in means of each covariate, divided by the standard deviation in the full treated group.
  - Similar to an effect size and is compared before and after matching (Rosenbaum and Rubin, 1985b).
  - The SDM should be computed for each covariate, as well as two-way interactions and squares.
- For regression adjustment to be trustworthy, the absolute SDM should be < 0.25 and the variance ratios should be between 0.5 and 2 (Rubin 1973, Cochran and Rubin 1973 & Rubin 2001).

Standardized Difference in Means (SDM)

Data

```
library(stddiff); library("MatchIt") #?matchIt
set.seed(7042025); treat<- rbinom(100, 1, .45); outc<- rbinom(100, 1, .25);
numeric1<-round(abs(rnorm(100)+1)*10,0); binary1<- rbinom(100, 1, .55);
numeric2<-round(abs(rnorm(100)+1)*10,0); binary2<- rbinom(100, 1, .25)
data<-data.frame(outc, treat, numeric1, binary1, numeric2, binary2) #;summary(data)
```

Estimation of the SDM

```
##the std difference using the package
stddiff.numeric(data=data, gcol=2,vcol=c(3, 5)) #;stddiff.binary(data=data, gcol=2,vcol=c(2,
```

```
##           mean.c  sd.c mean.t  sd.t missing.c missing.t stddiff stddiff.l
## numeric1 10.288 8.137 13.471 7.633         0         0   0.403    -0.014
## numeric2 11.379 7.217 12.088 7.025         0         0   0.100    -0.314
##           stddiff.u
## numeric1      0.821
## numeric2      0.514
```

```
##the std difference in means by hand
(mean(data$numeric1[data$treat==1])- mean(data$numeric1[data$treat==0]))/
  sd(data$numeric1[data$treat==1])
```

```
## [1] 0.4169885
```

Assess quality of the matching (ii)

Hypothesis tests and p-values that incorporate information on the sample size (e.g., t-tests) should not be used as measures of balance (Austin, 2007; Imai et al., 2008).

1. Balance is inherently an in-sample property, without reference to any broader population or super-population.
2. NHST can be misleading as measures of balance, because they often conflate changes in balance with changes in statistical power. E.g., randomly discarding control individuals seemingly leads to increased balance, simply because of the reduced power.

- *Hypothesis tests should not be used as part of a stopping rule to select a matched sample when those samples have varying sizes (or effective sample sizes).*
- *Some argue that NHST are OK for testing balance due to the reduced power for estimating the treatment effect (Hansen, 2008), but that argument requires trading off Type I and Type II errors. The cost of those two types of errors may differ for balance checking and treatment effect estimation.*

What about the "observations are *independent*" assumption?

Matching methods for causal inference: A review and a look forward

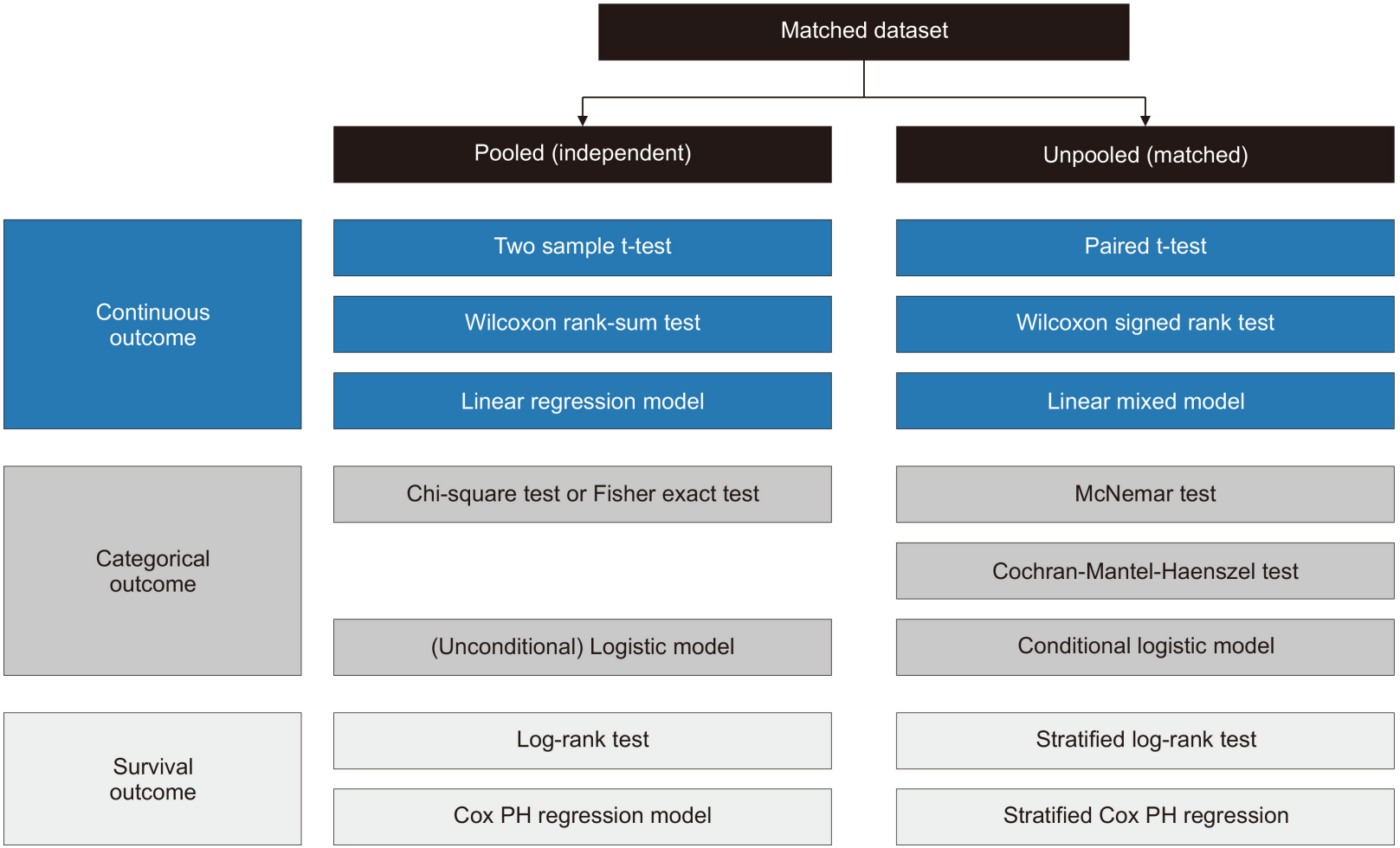
4) Analyze

When satisfactory matching (i.e., good covariate balance and a reasonable effective sample size) is NOT achieved after repeated specification and assessment of the quality, maybe the sample is fundamentally so different that no effect can be robustly estimated.

When satisfactory matching is achieved, one can estimates the exposure effect and its uncertainty (i.e., its standard error, confidence interval, and *p-value*).

- Stratification and Regression.
- Similar to the idea of "double robustness", and the intuition is that regression adjustment is used to "clean up" small residual covariate imbalance between the groups.
- Matching methods should also make the treatment effect estimates less sensitive to particular outcome model specifications (Ho et al., 2007).

4) Analyze



Lee B, Kim N, Won S, Gim J. Propensity score matching for comparative studies: a tutorial with R and Rex. J Minim Invasive Surg 2024;27:55-71. https://doi.org/10.7602/jmis.2024.27.2.55

Stratification for Matching

Assume case-control data: Consider 100 matched pairs (i.e., 100 cases and 100 controls, each paired by matching factors)

There are 100 2x2 tables, each containing the two observations in the matched pair, which can be grouped by their combination of exposure and cases vs controls as: W, X, Y, and Z pairs.

W pairs

	E	**-E**
D	1	0
-D	1	0

Y pairs

	E	**-E**
D	0	1
-D	1	0

X pairs

	E	**-E**
D	1	0
-D	0	1

Z pairs

	E	**-E**
D	0	1
-D	0	1

Stratification for Matching

The 100 2x2 tables can be summarized further as follows:

Example: W=30, X=30, Y=10, Z=30 $\rightarrow$ W + X + Y + Z = 100 TOTAL PAIRS

			Disease Present
		Exposure(+)	**Exposure (-)**
~Disease absent	**Exposure(+)**	W = 30	X = 30
	Exposure(-)	Y = 10	Z = 30

For all tables, from $i = 1$ to 100, $N + i = 2$.

Then, how to analyze this?

Two stratified analysis options for Matching:

1) Mantel-Haenszel Odds Ratio (ME3 2008, p. 287, eq 16-8)

$$OR_{MH} = \left(\frac{\sum_i A_{1i} B_{0i} / N + i}{\sum_i A_{0i} B_{1i} / N + i} \right)$$

2) McNemar Test and McNemar Odds Ratio (ME3 2008, p. 286-288)

$X^2 = \left(\frac{(X-Y)^2}{X+Y} \right), df = 1$, with $OR_{McN} = X/Y$

1) Mantel-Haenszel Odds Ratio for Matching

$$OR_{MH} = \left(\frac{\sum_i A_{1i} B_{0i} / N + i}{\sum_i A_{0i} B_{1i} / N + i} \right)$$

- Type W tables (case and control are exposed), $A_{1i} = B_{1i} = 1$ and $A_{0i} = B_{0i} = 0$.
- Type X tables (case is exposed, control is unexposed), $A_{1i} = B_{0i} = 1$ and $A_{0i} = B_{1i} = 0$
- Type Y tables (case is unexposed, control is exposed), $A_{1i} = B_{0i} = 0$ and $A_{0i} = B_{1i} = 1$
- Type Z tables (case and control are unexposed), $A_{1i} = B_{1i} = 0$ and $A_{0i} = B_{0i} = 1$.

1) Mantel-Haenszel Odds Ratio

- Type W tables (case and control are exposed), $A_{1i} = B_{1i} = 1$ and $A_{0i} = B_{0i} = 0$.
- Type X tables (case is exposed, control is unexposed), $A_{1i} = B_{0i} = 1$ and $A_{0i} = B_{1i} = 0$
- Type Y tables (case is unexposed, control is exposed), $A_{1i} = B_{0i} = 0$ and $A_{0i} = B_{1i} = 1$
- Type Z tables (case and control are unexposed), $A_{1i} = B_{1i} = 0$ and $A_{0i} = B_{0i} = 1$.

- The Type W and Type Z tables have values of zero for all products of A_{1i} and B_{0i} as well as all products of A_{0i} and B_{1i}.

- W and Z are concordant pairs, cases and controls have the same exposure level and in the OR_{MH} do not contribute to the OR estimation.

1) Mantel-Haenszel Odds Ratio

$$OR_{MH} = \left(\frac{\sum_i A_{1i} B_{0i} / N + i}{\sum_i A_{0i} B_{1i} / N + i} \right)$$

			D
		E	-E
~D	E	W = 30	X = 30
	-E	Y = 10	Z = 30

- Estimating $OR_{MH} = (30/2)/(10/2)$ using tables X and Y Type tables, the OR estimate is 15/5 = 3.0.

- Can be obtained using `mantelhaen.test` function of the `stats` package, the `cmh.test` function of the `lawstat`, `stratastats`, or `epi.2by2`. Input data must be either a list of 2x2 tables or a 3Dimensional array (e.g. 3 levels or 2x2x2 table).

Two stratified analysis options:

2) McNemar Test and McNemar Odds Ratio

Sum across all matched pair tables to form a single summary table:

			D
		E	-E
~D	E	W (30)	X (30)
	-E	Y (10)	Z (30)

$$X^2 = \left(\frac{(X-Y)^2}{X+Y}\right), df = 1$$

$$OR_{McN} = X/Y = 30/10 = 3$$

$$SE = \sqrt{(1/X + 1/Y)} = \sqrt{(1/30 + 1/10)}$$

= 0.365

The McNemar $\chi 2$ test the null hypothesis of no association between exposure and outcome.

- Using the numbers above, $\chi 2 = (30 - 10)2/(30 + 10) = 400/40 = 10.$

$OR_{McN} =$ **3** and the 95% CI is exp(1.099 ± 1.96(0.365)) = exp(0.383, 1.814) = **(1.467, 6.14)**

2) McNemar Test and McNemar Odds Ratio

```
X <- cbind(c(30,10), c(30,30));
mn_test <- mcNemar(X, alpha = 0.05)
mn_test
```

```
##
## Matched Pairs Analysis: McNemar's Chi^2 Statistic and Odds Ratio
##
## McNemar's Chi^2 Statistic (corrected for continuity) = 9.025 which has a p-value of: 0.003
##
## McNemar's Odds Ratio (b/c): 3
## 95% Confidence Limits for the OR are: [1.521, 8.68]
```

2) McNemar Test and McNemar Odds Ratio

```
kable(mn_test$X)
```

	Exposed Person: Disease Present	Exposed Person: Disease Absent
Control Person: Disease Present	30	30
Control Person: Disease Absent	10	30

Can use this code to obtain details

```
summary(mn_test)
```

What to do if I have more data, other covariates to account for?

As long as you have gone through steps 1 to 3, one can move to the analysis step using Regressions...

Regression *Adjustment* for Matching

Straightforward way: Fit a regression model including the matching weights in the estimation and using the coefficient on exposure as the exposure effect estimate; which is equivalent to computing a (weighted) difference in means.

- A binary regression model with a log link can be used to estimate the risk ratio.
- ***g-computation*** methods and targeted minimum loss-based estimation, can be used to ensure the resulting effect estimate is interpretable as marginal rather than conditional when the effect measure is non-collapsible.
- The coefficient on exposure in stratified, conditional, and covariate-adjusted models for odds or hazard ratios corresponds to a conditional effect; thus, these models should be avoided after matching, which is best suited for estimating marginal effects.

It's not matching *or* regression, it's matching *and* regression.

Posted on June 22, 2014 1:36 PM by Andrew

A colleague writes:

> *Why do people keep praising matching over regression for being non parametric? Isn't it f'ing parametric in the matching stage, in effect, given how many types of matching there are... you're making structural assumptions about how to deal with similarities and differences.... the likelihood two observations are similar based on something quite similar to parametric assumptions... you're just hiding the parametric part..*

My reply: It's not matching *or* regression, it's matching *and* regression. Matching is a way to discard some data so that the regression model can fit better. Trying to do matching without regression is a fool's errand or a mug's game or whatever you want to call it. Jennifer and I discuss this in chapter 10 of our book, also it's in Don Rubin's PhD thesis from 1970!

This entry was posted in Causal Inference by Andrew. Bookmark the permalink.

It's not matching or regression, it's matching and regression.

Example: Infertility after Spontaneous and Induced Abortion: This is a matched case-control study dating from before the availability of conditional logistic regression. There are 83-strata indicated by the variable **`stratum`**.

```
data("infert"); set.seed(7042025); infert$treat <- rbinom(length(infert$case), 1, .40) #crea
summary(infert[, c("case", "treat","education", "parity", "induced",
                   "age", "spontaneous", "stratum")]); dim(infert) #?infert
```

```
##       case            treat          education       parity
##  Min.   :0.0000   Min.   :0.000   0-5yrs : 12   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:0.000   6-11yrs:120   1st Qu.:1.000
##  Median :0.0000   Median :0.000   12+ yrs:116   Median :2.000
##  Mean   :0.3347   Mean   :0.371                 Mean   :2.093
##  3rd Qu.:1.0000   3rd Qu.:1.000                 3rd Qu.:3.000
##  Max.   :1.0000   Max.   :1.000                 Max.   :6.000
##     induced            age         spontaneous        stratum
##  Min.   :0.0000   Min.   :21.00   Min.   :0.0000   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.:0.0000   1st Qu.:21.00
##  Median :0.0000   Median :31.00   Median :0.0000   Median :42.00
##  Mean   :0.5726   Mean   :31.50   Mean   :0.5766   Mean   :41.87
##  3rd Qu.:1.0000   3rd Qu.:35.25   3rd Qu.:1.0000   3rd Qu.:62.25
##  Max.   :2.0000   Max.   :44.00   Max.   :2.0000   Max.   :83.00

## [1] 248   9
```

One case with two prior spontaneous abortions and two prior induced abortions is omitted. Source: Trichopoulos et al (1976) Br. J. of Obst. and Gynaec. 83, 645–650. R-datasets packages

Assessing the dataset, closeness and quality of the matching Example The "standardized difference in means" (SDM) using `stddiff` package

```
#(mean(infert$age[infert$treat==1])- mean(infert$age[infert$treat==0]))/
#                         sd(infert$age[infert$treat==1])
stddiff.numeric(data=infert,gcol=9,vcol=c(2,3 ))
```

```
##        mean.c  sd.c mean.t  sd.t missing.c missing.t stddiff stddiff.l
## age    31.712 5.463 31.152 4.881         0         0   0.108     -0.15
## parity  2.051 1.206  2.163 1.328         0         0   0.088     -0.17
##        stddiff.u
## age        0.366
## parity     0.346
```

```
# stddiff.category(data=infert,gcol=9,vcol=c(4,6, 1))
stddiff.binary(data=infert,gcol=9,vcol=c(4,6 ))
```

```
##                p.c   p.t missing.c missing.t stddiff stddiff.l stddiff.u
## induced      0.545 0.620         0         0   0.152    -0.106     0.410
## spontaneous 0.571 0.587         0         0   0.033    -0.224     0.291
```

This is the step to check the balance across variables

Implementing the matching Example The “standardized difference in means” (SDM) using `matchIt` package **Assessing the matching**

```
m.out<- matchit(treat ~  age+ parity + spontaneous + induced + education,
                  data = infert, method = NULL, distance = "glm")
m.out
```

```
## A `matchit` object
##  - method: None (no matching)
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 248 (original)
##  - target estimand: ATT
##  - covariates: age, parity, spontaneous, induced, education
```

This is the step to ask the software to match/check individuals treated and not treated by the given covariates

Assessing the matching using `matchIt` package: Example

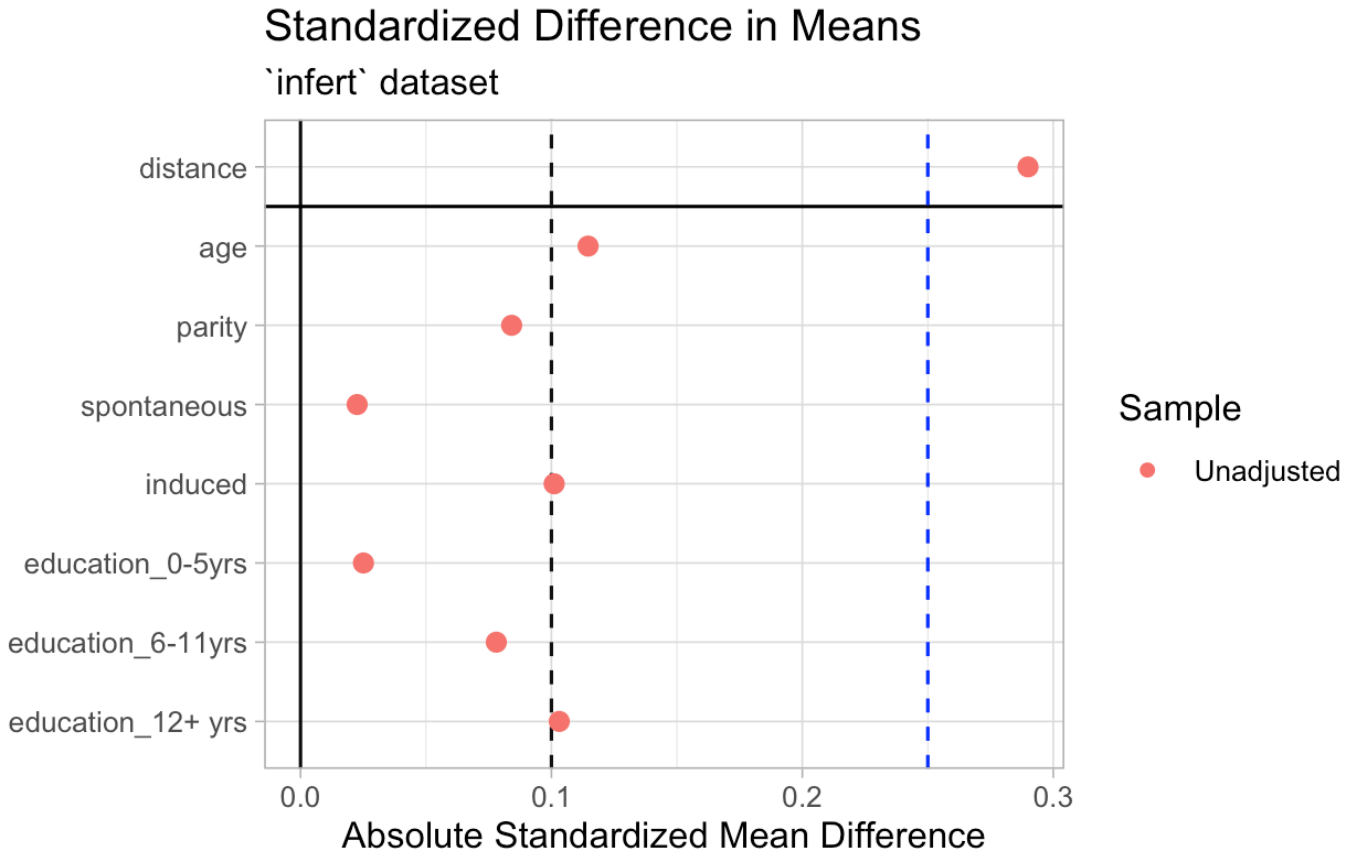
```
round(summary(m.out)$"sum.all"[,1:4], 3) #summary(m.out)
```

```
##                  Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                 0.383          0.364            0.290      1.038
## age                     31.152         31.712           -0.115      0.798
## parity                   2.163          2.051            0.084      1.212
## spontaneous              0.587          0.571            0.023      0.977
## induced                  0.620          0.545            0.101      0.999
## education0-5yrs          0.033          0.058           -0.141         NA
## education6-11yrs         0.435          0.513           -0.157         NA
## education12+ yrs         0.533          0.429            0.207         NA
```

This is the step to check the balance across variables before matching. Recall in this data individuals were already matched

Assessing the quality of the matching using `matchIt` package: Example

Plot of "standardized difference in means" (SDM)



This plot helps with the visualization of the (un)balance across variables

Regression model for Matched Data

The model for a matched data with k = 1, ..., K strata is

$$logit[\pi_k(X)] = \alpha_k + \beta_1 X_1 + \ldots + \beta_p X_p$$

Where D is "Disease/Outcome, $\pi_k(X) = Pr(D_{ik} = 1|X)$, α_k is log-odds in the k_{th} stratum

- Unless the number of subjects in each stratum is large, fitting these models using the unconditional ML does not work well.
- In individually matched there is only one case in each stratum and hence we need some way of getting rid of the nuisance parameters.

Conditional likelihood: the probability of the observed data conditional on the stratum total and the number of cases observed is the conditional likelihood for the *k* the stratum.

Considerations for Regression Adjustment with Matched Data

- Using standard logistic regression model to analyze the matched data, the effect estimates (i.e., exponentiated slope coefficients) will generally be overestimates.

- If the data are matched 1-to-1 in pairs, the OR estimate obtained from a standard logistic model will be the square of the correct value.

- To complete the matched data analysis, one needs a set of indicator variable that records that matched strata.

    - The coefficients for the $k-1$ indicator variables (i.e. stratum-specific intercepts) are "nuisance parameters" in the sense that they have no epidemiologic interpretation.

Standard Logistic Regression for matching: Regression Example using the `infert` data set

```
mod.logistic <- glm(case ~ treat + age + parity + education + spontaneous +induced,
                    family = binomial(), data = infert)
```

```
##                   Coeff 2.5 % 97.5 %
## (Intercept)       -1.08 -3.89   1.69
## treat             -0.25 -0.89   0.37
## age                0.04 -0.02   0.10
## parity            -0.81 -1.22  -0.44
## education6-11yrs -1.00 -2.57   0.59
## education12+ yrs -1.34 -3.01   0.32
## spontaneous        2.03  1.45   2.67
## induced            1.28  0.71   1.90
```

```
##                     OR 2.5 % 97.5 %
## (Intercept)       0.34  0.02   5.41
## treat             0.78  0.41   1.45
## age               1.04  0.98   1.11
## parity            0.44  0.30   0.64
## education6-11yrs 0.37  0.08   1.81
## education12+ yrs 0.26  0.05   1.38
## spontaneous       7.63  4.28  14.46
## induced           3.61  2.03   6.66
```

Although the model 'ran' and produced results, we know this model is wrong because does not account for the matched structure

Standard Logistic Regression Example using the `infert` data set **Adding the stratum (n=83) variable**

```
mod.logistic1 <- glm(case ~ treat + age + parity+ education + spontaneous +induced +
                       factor(stratum),
                     family = binomial(), data = infert)
```

```
##                    Coeff  2.5 % 97.5 %
## (Intercept)        -3.32 -32.34  23.34
## treat              -0.35  -1.24   0.52
## age                 0.00  -0.62   0.62
## parity             -0.59  -6.78   6.43
## education6-11yrs    1.44 -29.33  36.03
## education12+ yrs    0.12 -30.60  34.29
## spontaneous         3.21   2.36   4.19
## induced             2.18   1.33   3.15
## factor(stratum)2    2.51 -36.68  46.70
## factor(stratum)3    1.93  -7.10  11.03
## factor(stratum)4    0.21 -16.84  19.22
## factor(stratum)5   -1.23 -10.67   7.62
## factor(stratum)6   -3.03 -18.36  10.79
## factor(stratum)7    1.89  -4.70   8.79
## factor(stratum)8   -0.60  -6.56   5.41
## factor(stratum)9   -0.91  -8.18   6.68
## factor(stratum)10 -0.22  -7.60   7.27
## factor(stratum)11  0.40  -6.56   7.47
## factor(stratum)12 -2.17 -17.05  11.33
## factor(stratum)13  1.17  -5.27   7.65
## factor(stratum)14 -1.90 -14.54   9.85
## factor(stratum)15 -1.05  -7.52   5.37
```

```
##                      OR 2.5 %       97.5 %
## (Intercept)        0.04  0.00 1.369945e+10
## treat              0.70  0.29 1.680000e+00
## age                1.00  0.54 1.850000e+00
## parity             0.55  0.00 6.215800e+02
## education6-11yrs   4.22  0.00 4.464663e+15
## education12+ yrs   1.13  0.00 7.782848e+14
## spontaneous       24.73 10.63 6.599000e+01
## induced            8.82  3.76 2.329000e+01
## factor(stratum)2  12.32  0.00 1.917583e+20
## factor(stratum)3   6.87  0.00 6.156937e+04
## factor(stratum)4   1.23  0.00 2.224256e+08
## factor(stratum)5   0.29  0.00 2.029810e+03
## factor(stratum)6   0.05  0.00 4.848055e+04
## factor(stratum)7   6.63  0.01 6.559530e+03
## factor(stratum)8   0.55  0.00 2.246300e+02
## factor(stratum)9   0.40  0.00 7.987500e+02
## factor(stratum)10  0.80  0.00 1.432570e+03
## factor(stratum)11  1.49  0.00 1.751440e+03
## factor(stratum)12  0.11  0.00 8.323751e+04
## factor(stratum)13  3.21  0.01 2.090240e+03
## factor(stratum)14  0.15  0.00 1.895340e+04
## factor(stratum)15  0.35  0.00 2.146000e+02
```

Standard Logistic Regression Example using the `infert` data set **Adding the stratum variable (n=83)**

- Then the OR formula is just the usual logistic regression formula for exposure E, confounder C, but adding in 82 indicator variables for the 83 strata of matched pairs.

This doesn't seem *Efficient*...

Conditional Logistic Regression is the righ type for matched data

- We can make use of a **conditional** maximum likelihood method to estimate the exposure effect in this design, rather than the usual unconditional model.
    - The “conditional” part refers to "conditioned on the strata of matched pairs".
- The *k* stratum-specific conditional likelihood is obtained as the probability of the observed data conditioned on the number of observations in stratum k and the number of these that are cases.
- The probability of the observed data relative to the probability of the data under all other possible assignments of the n_{1k} cases and n_{0k} controls to $nk(= n_{1k} + n_{0k})$ subjects.

Considerations for the Conditional Logistic Regression

- This conditional likelihood is complex (see Hosmer & Lemeshow 2000, pp. 225-226)
- For 1-to-1 matching there are only 2 subjects per stratum, and the conditional likelihood for stratum k is: $l_k(\beta) = \left(\frac{e^{\beta^{x_{1k}}}}{e^{\beta^{x_{1k}}} + e^{\beta^{x_{0k}}}} \right)$,

where x_{1k} is the data vector for the case and x_{0k} is the data vector for the control.

- Given values for $\beta,\ x_{1k}\ and\ x_{0k}$, the expression above is interpreted as the modeled probability that an exposed subject is a case, assuming the 1-to-1 matched design (so one of the two observations in the stratum must be a case).
- For any stratum in which $x_{1k} = x_{0k}$ the prob. of each observation being a case is 0.5, regardless the value of β, and therefore the stratum is uninformative.
- Checking on the frequency of the 2 types of discordant pairs, recognizing that if one or the other doesn't occur that the conditional estimator is undefined.

Conditional Logistic Regression for matched data, Example using the `infert` data

```
modelclogit <- clogit(case ~ treat + spontaneous + induced + strata(stratum), data = infert)
cbind(Coeff=round(coef(modelclogit), 2), round(confint(modelclogit), 2)) #summary(modelclogi
```

```
##             Coeff 2.5 % 97.5 %
## treat       -0.21 -0.91   0.49
## spontaneous  1.97  1.28   2.66
## induced      1.40  0.69   2.10
```

This shows that only contributing parameters are used in the estimation

```
clogit(case ~ treat + spontaneous + induced +
       age + parity+ education +
         strata(stratum), data = infert)
```

```
## Call:
## clogit(case ~ treat + spontaneous + induced + age + parity +
##     education + strata(stratum), data = infert)
##
##                      coef exp(coef) se(coef)      z        p
## treat             -0.2091    0.8113   0.3553 -0.589 0.556107
## spontaneous        1.9705    7.1741   0.3527  5.587 2.32e-08
## induced            1.3979    4.0465   0.3607  3.875 0.000107
## age                    NA        NA   0.0000     NA       NA
## parity                 NA        NA   0.0000     NA       NA
## education6-11yrs       NA        NA   0.0000     NA       NA
## education12+ yrs       NA        NA   0.0000     NA       NA
##
## Likelihood ratio test=53.5  on 3 df, p=1.431e-11
```

Conditional Logistic Regression for matched data, Example using the `infert` data

Using Standard (Unconditional) Logistic Regression

```
##                       OR 2.5 % 97.5 %
## (Intercept)         0.34  0.02   5.41
## treat               0.78  0.41   1.45
## age                 1.04  0.98   1.11
## parity              0.44  0.30   0.64
## education6-11yrs    0.37  0.08   1.81
## education12+ yrs    0.26  0.05   1.38
## spontaneous         7.63  4.28  14.46
## induced             3.61  2.03   6.66
```

Using Conditional Logistic Regression

```
##                  OR 2.5 % 97.5 %
## treat          0.81  0.40   1.63
## spontaneous 7.17  3.59  14.32
## induced     4.05  2.00   8.21
```

Wait, what happened to the intercept?

There is no intercept estimated in this *conditional* model because the likelihood function conditions on each matched set of *n* observations, and a baseline average outcome cannot be estimated within each of these strata

Interpretations for the Conditional Logistic Regression β coefficient?

Standard Matched Case-control

- Estimated β coefficient is the average log(odds) of the exposure/ variable of interest on the outcome.
- Exponentiated β coefficient is the average OR for the exposure/ variable of interest on the outcome

Nested Case-control with Incidence Density Sampling

- Estimated β coefficient is the log of the average incidence rate ratio for the exposure/ variable of interest on the outcome.
- Exponentiated β coefficient is the average incidence rate ratio for the exposure/ variable of interest on the outcome.

WHY? Here we have time-matched controls, sampled when the cases occur.

Recall, this is comparing exposed vs non exposed holding other variables constant [consider parameterization !!].

Matching and Conditional Logistic Regression: Example using the `infert` data and using the `matchIt` package.

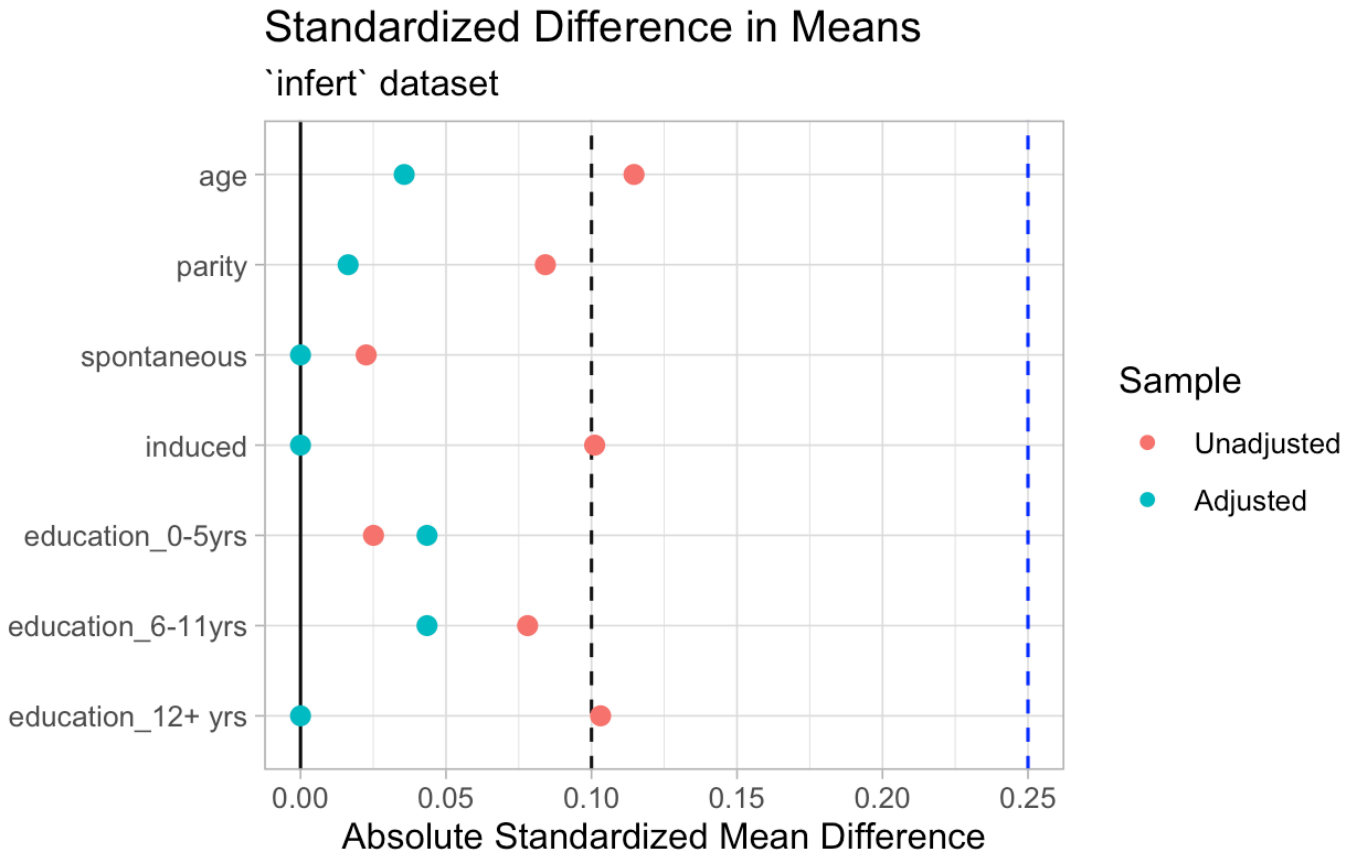
```
m.out2 <- matchit(treat  ~ age+ parity + spontaneous + induced + education,
                  data = infert, method ="cem", cutpoints = list(parity=3),
                  grouping = list( education= list(c("0-5yrs","6-11yrs"), "12+ yrs")),
                  k2k = TRUE, k2k.method = "mahalanobis")
```

```
## A `matchit` object
##  - method: Coarsened exact matching
##  - number of obs.: 248 (original), 92 (matched)
##  - target estimand: ATT
##  - covariates: age, parity, spontaneous, induced, education
```

Sample Sizes:		
	Control	Treated
All	99	149
Matched	52	52
Unmatched	47	97
Discarded	0	0

For illustration ONLY, here we changed the matching structure but the dataset was already matched

Matching and Conditional Logistic Regression: Example using the `infert` data and the `matchIt` package to plot of balance



This plot illustrates the balance before (unadjusted) and after (adjusted) matching

Matching and Conditional Logistic Regression: Example using the `infert` data and the `matchIt` package to provide a **Summary of Balance Before-Matching**

```
##                   Means Treated Means Control Std. Mean Diff. Var. Ratio
## age                       31.152        31.712           -0.115      0.798
## parity                     2.163         2.051            0.084      1.212
## spontaneous                0.587         0.571            0.023      0.977
## induced                    0.620         0.545            0.101      0.999
## education0-5yrs            0.033         0.058           -0.141         NA
## education6-11yrs           0.435         0.513           -0.157         NA
## education12+ yrs           0.533         0.429            0.207         NA
```

Summary of Balance After-Matching

```
##                   Means Treated Means Control Std. Mean Diff. Var. Ratio
## age                       31.065        31.239           -0.036      0.906
## parity                     1.609         1.630           -0.016      0.923
## spontaneous                0.391         0.391            0.000      1.000
## induced                    0.413         0.413            0.000      1.000
## education0-5yrs            0.065         0.022            0.245         NA
## education6-11yrs           0.391         0.435           -0.088         NA
## education12+ yrs           0.543         0.543            0.000         NA
```

Matching and Conditional Logistic Regression: Example using the `infert` data and the `matchIt` package

With weights

```
match.data1 <- match.data(m.out2)
#;head(match.data1)
mod.logistic2 <- glm(case ~ treat + age + p
                     family = binomial(),
                     data = match.data1, we
#summary(mod.logistic2); coeftest(mod.logis
cbind(Coeff= round(coefficients(mod.logisti
```

```
##                    Coeff 2.5 % 97.5 %
## (Intercept)         4.12 -1.94  10.96
## treat              -0.06 -1.14   1.01
## age                -0.01 -0.13   0.10
## parity             -1.24 -2.33  -0.35
## education6-11yrs   -4.05 -8.72  -0.47
## education12+ yrs   -4.79 -9.81  -1.02
## spontaneous         2.33  1.11   3.76
## induced             1.15 -0.25   2.56
```

Without weights

```
mod.logistic2a <- glm(case ~ treat + age +
                      data = match.data1)
cbind(round(coefficients(mod.logistic2a), 2
```

```
##                          2.5 % 97.5 %
## (Intercept)       4.12 -1.94  10.96
## treat            -0.06 -1.14   1.01
## age              -0.01 -0.13   0.10
## parity           -1.24 -2.33  -0.35
## education6-11yrs -4.05 -8.72  -0.47
## education12+ yrs -4.79 -9.81  -1.02
## spontaneous       2.33  1.11   3.76
## induced           1.15 -0.25   2.56
```

```
summary(match.data1$weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

Coefficients are identical because all have same weight=1

Using `clogit` function

```
mod.clog3 <- clogit(case ~ treat + age + parity+ education + spontaneous + induced,
                    data = match.data1)
cbind(Coeff=round(coefficients(mod.clog3), 2), round(confint(mod.clog3),2))
```

```
##                   Coeff 2.5 % 97.5 %
## treat             -0.06 -1.12   0.99
## age               -0.01 -0.13   0.10
## parity            -1.22 -2.20  -0.25
## education6-11yrs -4.00 -8.02   0.02
## education12+ yrs -4.73 -9.01  -0.45
## spontaneous        2.30  1.00   3.59
## induced            1.13 -0.23   2.50
```

GLM using weights and the `strata`

```
cbind(Coeff=round(coefficients(mod.logistic2), 2), round(confint(mod.logistic2),2))
```

```
##                   Coeff 2.5 % 97.5 %
## (Intercept)        4.12 -1.94  10.96
## treat             -0.06 -1.14   1.01
## age               -0.01 -0.13   0.10
## parity            -1.24 -2.33  -0.35
## education6-11yrs -4.05 -8.72  -0.47
## education12+ yrs -4.79 -9.81  -1.02
## spontaneous        2.33  1.11   3.76
## induced            1.15 -0.25   2.56
```

What is the quantity estimated in presence of Matching?

- The estimand matching is most often used for is the average exposure effect among those who were exposed, also known as the average treatment effect on the treated **(ATT)**,
  - I.e., the average difference between the observed outcomes for those exposed and their counterfactual outcomes had they not been exposed.
  - This is the same quantity estimated using weighting by the odds (if such).
- Some matching methods allow estimation of the average exposure effect in the population, e.g., estimated with inverse probability weights. [1]
  - The choice of estimand depends on the desired target population of interest, which should be specified before the analysis, and matching methods appropriate for that estimand should be used.

[1] More on this on the propensity score lecture!

The Bayesian way???

```
infert1 <- infert[order(infert$stratum), ]
post <- stan_clogit(case ~ spontaneous + induced + (1 | parity), strata = stratum,
                    data = infert1, # order necessary subset = parity <= 2,
                       QR = TRUE, cores = 2, seed = 7042025)
post
PPD <- posterior_predict(post) #; summary(PPD)
post1 <- stan_clogit(case ~  treat + spontaneous + induced + (1 |education),
                     data = infert[order(infert$stratum), ], strata = stratum,
                     QR = TRUE, cores = 4, seed = 7042025)
post1
```

```
> post
stan_clogit
 family:       binomial [clogit]
 formula:      case ~ spontaneous + induced + (1 | parity)
 observations: 248
------
            Median MAD_SD
spontaneous 2.0    0.3
induced     1.4    0.4

Error terms:
 Groups Name        Std.Dev.
 parity (Intercept) 1.4
Num. levels: parity 6
```

```
> post1
stan_clogit
 family:       binomial [clogit]
 formula:      case ~ treat + spontaneous + induced + (1 | education)
 observations: 248
------
            Median MAD_SD
treat       -0.9    0.4
spontaneous  2.1    0.4
induced      1.4    0.4

Error terms:
 Groups    Name        Std.Dev.
 education (Intercept) 1.4
Num. levels: education 3
```

Interpretation: Frequentist vs Bayesian

- The Bayesian approach provides a **complete posterior** distribution of the log (average Odss or IRR)

- The frequentist approach provides a fixed parameter estimate of the log (average Odds or IRR) and an estimate of its sampling variance.

QUESTIONS?

COMMENTS?

RECOMMENDATIONS?

Details of the McNemar Test

```
summary(mn_test)
```

```
##
## Matched Pairs Analysis: McNemar's Statistic and Odds Ratio (Detailed Summary):
##
##                                  Exposed Person: Disease Present
## Control Person: Disease Present                                30
## Control Person: Disease Absent                                 10
##                                  Exposed Person: Disease Absent
## Control Person: Disease Present                               30
## Control Person: Disease Absent                                30
##
## Entries in above matrix correspond to number of pairs.
##
## McNemar's Chi^2 Statistic (corrected for continuity) = 9.025 which has a p-value of: 0.003
## Note: The p.value for McNemar's Test corresponds to the hypothesis test: H0: OR = 1 vs. HA: OR !=
## McNemar's Odds Ratio (b/c): 3
## 95% Confidence Limits for the OR are: [1.521, 8.68]
## The risk difference is: 0.2
## 95% Confidence Limits for the rd are: [0.072, 0.328]
```

Code for the Plot of "standardized difference in means" (SDM)

```
#plot(summary(m.out)) #this provides a series of Q-Q plots
cobalt::love.plot(m.out, thresholds = c(m = .1), abs= T)+ #this provide the line at 0.1
 labs(title = 'Standardized Difference in Means', subtitle = "`infert` dataset",
      x="Absolute Standardized\ Mean Difference", y=" ") +
  geom_vline(xintercept = 0.25, color= "blue", linetype =2)+ #this provide the line at 0.25
   theme_light() +
    theme( panel.spacing = unit(0.5, "lines"),
          strip.text.x = element_text(size = 14),
           strip.text.y = element_text(size = 16))
```

Additonal worked example

Matching and Conditional Logistic Regression

Using the `simualted` example and the `matchIt` package.

```
m.out.sim <- matchit(treat ~ numeric1 + binary1 + numeric2 + binary2,
                     data = data,  method = NULL, distance = "glm")
m.out.sim
```

```
## A `matchit` object
##  - method: None (no matching)
##  - distance: Propensity score
##              - estimated with logistic regression
##  - number of obs.: 100 (original)
##  - target estimand: ATT
##  - covariates: numeric1, binary1, numeric2, binary2

##
## Call:
## matchit(formula = treat ~ numeric1 + binary1 + numeric2 + binary2,
##     data = data, method = NULL, distance = "glm")
##
## Summary of Balance for All Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.3868        0.3159          0.5459     1.1985    0.1620
## numeric1        13.4706       10.2879          0.4170     0.8799    0.1219
## binary1          0.4118        0.6061         -0.3948          .    0.1943
## numeric2        12.0882       11.3788          0.1010     0.9476    0.0483
## binary2          0.1765        0.2424         -0.1730          .    0.0660
```

Matching and Conditional Logistic Regression

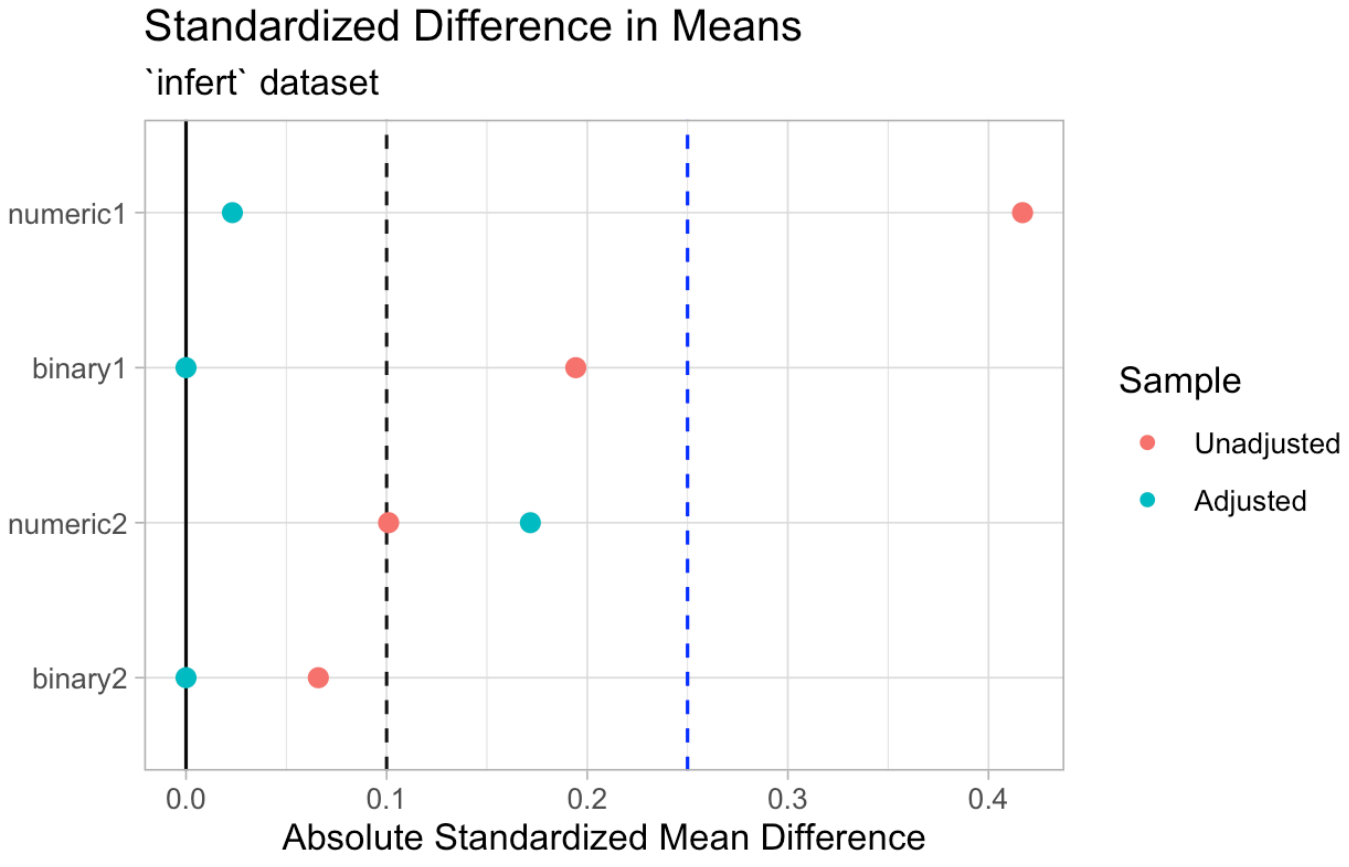
Using the `simualted` example

```
m.out.sim1 <- matchit(treat ~ numeric1 + binary1 + numeric2 + binary2,  data = data,
                  distance = "mahalanobis", replace = TRUE)
m.out.sim1
```

```
## A `matchit` object
##  - method: 1:1 nearest neighbor matching with replacement
##  - distance: Mahalanobis - number of obs.: 100 (original), 58 (matched)
##  - target estimand: ATT
##  - covariates: numeric1, binary1, numeric2, binary2
```

Matching and Conditional Logistic Regression

Using the `simualted` example, Plot of balance `simualted` Example



Matching and Conditional Logistic Regression

Summary of Balance for All Data (Pre-Matching) Using the `simualted` example

```
##
## Call:
## matchit(formula = treat ~ numeric1 + binary1 + numeric2 + binary2,
##     data = data, distance = "mahalanobis", replace = TRUE)
##
## Summary of Balance for All Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## numeric1       13.4706       10.2879          0.4170     0.8799    0.1219
## binary1         0.4118        0.6061         -0.3948          .    0.1943
## numeric2       12.0882       11.3788          0.1010     0.9476    0.0483
## binary2         0.1765        0.2424         -0.1730          .    0.0660
##          eCDF Max
## numeric1   0.2986
## binary1    0.1943
## numeric2   0.1738
## binary2    0.0660
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## numeric1       13.4706       13.2941          0.0231     1.0999    0.0347
## binary1         0.4118        0.4118          0.0000          .    0.0000
## numeric2       12.0882       10.8824          0.1716     1.0685    0.0647
## binary2         0.1765        0.1765          0.0000          .    0.0000
##          eCDF Max Std. Pair Dist.
```

Matching and Conditional Logistic Regression

Summary of Balance for Matched Data Using the `simualted` example

```
##
## Call:
## matchit(formula = treat ~ numeric1 + binary1 + numeric2 + binary2,
##     data = data, distance = "mahalanobis", replace = TRUE)
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## numeric1       13.4706       13.2941          0.0231     1.0999    0.0347
## binary1         0.4118        0.4118          0.0000          .    0.0000
## numeric2       12.0882       10.8824          0.1716     1.0685    0.0647
## binary2         0.1765        0.1765          0.0000          .    0.0000
##          eCDF Max Std. Pair Dist.
## numeric1   0.0882          0.2929
## binary1    0.0000          0.0000
## numeric2   0.2059          0.3391
## binary2    0.0000          0.0000
##
## Sample Sizes:
##                Control Treated
## All              66.        34
## Matched (ESS)    18.06      34
## Matched          24.        34
## Unmatched        42.         0
## Discarded         0.         0
```

Matching and Conditional Logistic Regression

Using the `simualted` example

Estimation without weights

```
sim.match.data1 <- match.data(m.out.sim1)
mod.sim1 <- glm(outc ~ treat + numeric1 + b
                      family = binomial(), d
#summary(mod.sim1)
cbind(Coeff=round(mod.sim1$coefficients, 2)
```

```
##             Coeff 2.5 % 97.5 %
## (Intercept) -1.98 -4.29   0.09
## treat       -0.51 -1.83   0.77
## numeric1     0.09  0.01   0.19
## binary1     -0.01 -1.40   1.36
## numeric2     0.00 -0.10   0.10
## binary2     -0.52 -2.60   1.21
```

```
summary(sim.match.data1 $weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7059  0.7059  1.0000  1.0000  1.0000  2.8235
```

Estimation using matching weights

```
mod.sim2 <- glm(outc ~ treat + numeric1 + b
                      family = binomial(), d
#summary(mod.sim2)
round(coeftest(mod.sim2, vcov. = vcovCL), 2
```

```
##
## z test of coefficients:
##
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.04       1.18   -1.73     0.08
## treat          -0.01       0.67   -0.02     0.99
## numeric1        0.07       0.04    1.77     0.08
## binary1        -0.01       0.71   -0.01     0.99
## numeric2        0.00       0.05   -0.09     0.93
## binary2        -0.48       0.98   -0.49     0.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05
```

Matching and Conditional Logistic Regression

Estimation using Conditional Logistic Regression

`clogit` Regression using the `simualted` example

```
modelclogit2 <- clogit(outc ~ treat + numer
                       weights = weights, d
cbind(Coeff= round(coef(modelclogit2), 2),
```

```
##          Coeff 2.5 % 97.5 %
## treat    -0.50 -1.77   0.77
## numeric1  0.09  0.00   0.18
## binary1  -0.01 -1.36   1.33
## numeric2  0.00 -0.09   0.10
## binary2  -0.51 -2.32   1.31
```

Sample Sizes:		
	Control	Treated
All	55.	45
Matched (ESS)	18.58	45
Matched	28.	45
Unmatched	27.	0
Discarded	0.	0

Matching and Conditional Logistic Regression

Modification of the matching using the `simualted` example

```
m.out.sim2<- matchit(treat ~ numeric1 + binary1 + numeric2 + binary2, data = data,
                     method = "cem", cutpoints = list(numeric1 = 5),
                     grouping = list(binary1 = list(c(0, 1)) ))
summary(m.out.sim2, un=F)
```

```
##
## Call:
## matchit(formula = treat ~ numeric1 + binary1 + numeric2 + binary2,
##     data = data, method = "cem", cutpoints = list(numeric1 = 5),
##     grouping = list(binary1 = list(c(0, 1))))
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## numeric1        11.1429       10.9143           0.0299     1.2470    0.0438
## binary1          0.5714        0.6683          -0.1967          .    0.0968
## numeric2         9.7143        9.3968           0.0452     1.0664    0.0229
## binary2          0.1429        0.1429           0.0000          .    0.0000
##           eCDF Max Std. Pair Dist.
## numeric1    0.1587          0.2237
## binary1     0.0968          0.7929
## numeric2    0.1667          0.1458
## binary2     0.0000          0.0000
##
## Sample Sizes:
##                Control Treated
## All              66.        34
```

Matching and Conditional Logistic Regression

Modification of the matching using the `simualted` example

```
match.data2 <- match.data(m.out.sim2)
```

```
##    outc treat numeric1 binary1 numeric2 binary2
## 3     0     0         4       1        19       1
## 5     1     0         7       0        17       0
## 6     0     0        17       1         3       0
## 7     0     1         4       1         1       1
## 8     0     0        21       1         2       0
## 11    0     1         7       1        11       0

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3333  0.5556  1.0000  1.0000  1.0000  3.3333

##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
##  2  4  2  4  4  3  4  6  3  3  3  4  3  4  2  3  2
```

outc	treat	numeric1	binary1	numeric2	binary2	weights	subclass
0	0	8	0	6	0	0.9565217	1
0	1	8	0	6	0	1.0000000	1
0	0	13	0	7	0	0.9565217	1
0	0	10	0	5	0	0.9565217	1
0	1	11	0	5	0	1.0000000	1
0	0	12	0	6	0	0.9565217	1
1	1	14	0	8	0	1.0000000	1
0	1	12	0	8	0	1.0000000	1
0	0	14	0	15	0	0.9565217	2
0	1	12	0	16	0	1.0000000	2
0	0	13	1	6	0	2.8695652	3
1	1	14	1	6	0	1.0000000	3
0	1	13	1	7	0	1.0000000	3
0	1	13	1	7	0	1.0000000	3

The Bayesian Way

```
dat2 <- match.data2[order(match.data2$subclass), ] # order by strata
post3 <- stan_clogit(outc ~  treat +   numeric1 +  numeric2 +binary1+ (1 | binary2), #
                     strata = subclass,
                     data = dat2,
                     chains = 2, iter = 100)
post3
post4 <- stan_clogit(outc ~  treat +  numeric1 +  numeric2 + binary1+ (1 | binary2),
                     data = dat2[order(dat2$subclass), ], # order necessary
                     strata = subclass, QR = TRUE,
                     cores = 2, seed = 704)

post4
```

Regression equation

Consider the 1:1 matched design (simplest case) with $k = 1$, ..., *K* strata and *p* covariates $logit(\pi_k(X)) = \alpha_k + \beta' X$

Where $\pi_k(X) = Pr(D_{ik} = 1|X)$, α_k is log-odds in the k_{th} stratum; X_{0k} be the data vector for the control and X_{1k} be the data vector for the case. $S_k = D_{0k} + D_{1k}$.

$$L_k(\beta) = Pr(D_{1k} = 1, D_{0k} = 0|X_{1k}, X_{0k}, S_k = 1, n_k = 2)$$