Causal Inference Notes

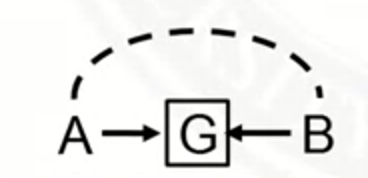
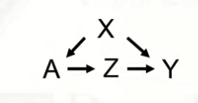
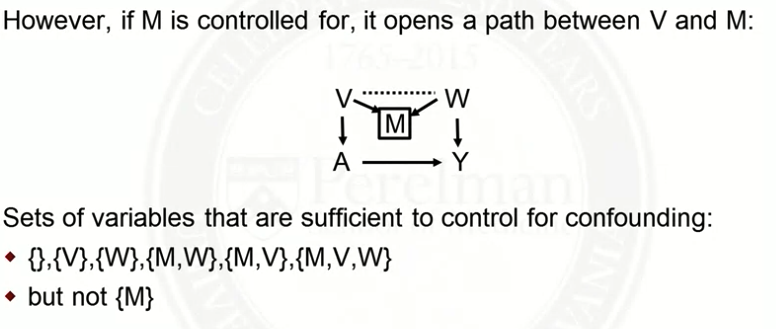
[A Crash Course in Causality: Inferring Causal Effects from Observational Data](https://www.coursera.org/learn/crash-course-in-causality)

* [Confluence page with notes](https://joinroot.atlassian.net/wiki/spaces/DS/pages/2104918017/Data+Science+Book+Clubs)

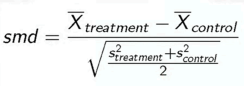
**Introduction to Causal Effects**

* Fundamental Problem of Causal Inference: We can only observe one potential outcome for each person
* Want to know relationship between means of difference potential outcomes: e.g. E(Y1-Y0)
  + Make several assumptions:
    - **Consistency**: Ya is equal to the observed outcome if the actual treatment received is A=a
    - **Positivity**: For every set of values for X, P(A=a|X=x)>0
    - **Ignorability**: treatment assignment is independent of potential outcomes conditional on some set of covariates X
      * Violation example: sicker patients are more likely to be treated (suppose X are measures of health)
      * Within levels of X (i.e. same age, same weight, same history of smoking), perhaps sicker patients are not more likely to get treatment -> ignorability
      * Want to find set of variables X to make ignorability assumption hold

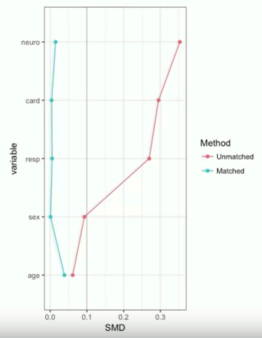
**Confounding and Directed Acyclic Graphs**

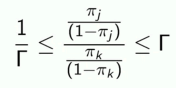
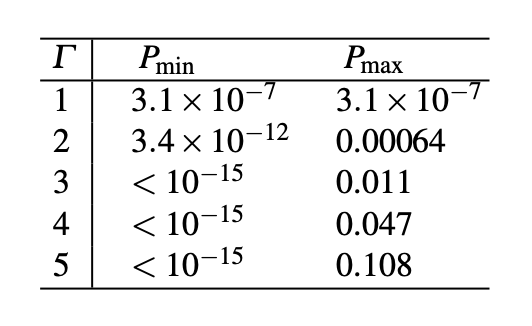
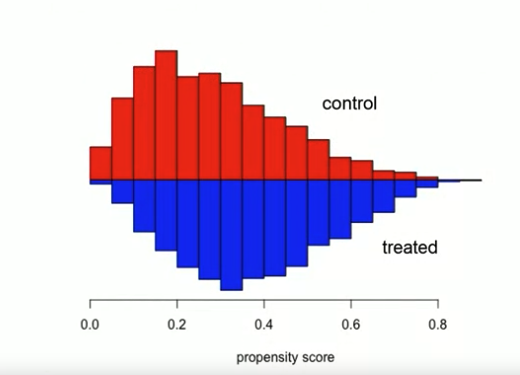
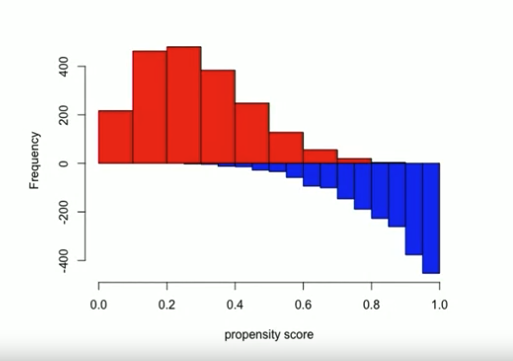
* **Conditional Independence (d-separation)**
  + ***Blocking***: conditioning on nodes in the path
    - A → G → B
    - Conditioning on G will block path between A and B
    - This will cause independence between A and B
    - A: temperature, G: icy sidewalks, B: someone falls
  + Conditioning on ***collider*** creates association between marginally unrelated nodes
    - A, B: two light switches -> independent
    - G: whether light is on, lit if both A and B are on
      * If we know G is off, then A must be off if B on and vice versa (now dependent)
    - 
* **Frontdoor/Backdoor paths**
  + 
  + A -> Z -> Y is frontdoor path from A to Y: don’t need to control for Z, it captures the effects of treatment
  + A <- X -> Y is a backdoor path: affects both treatment and outcome, need to control for X
  + What about if there’s a collider:
    - 

**Matching and Propensity Scores**

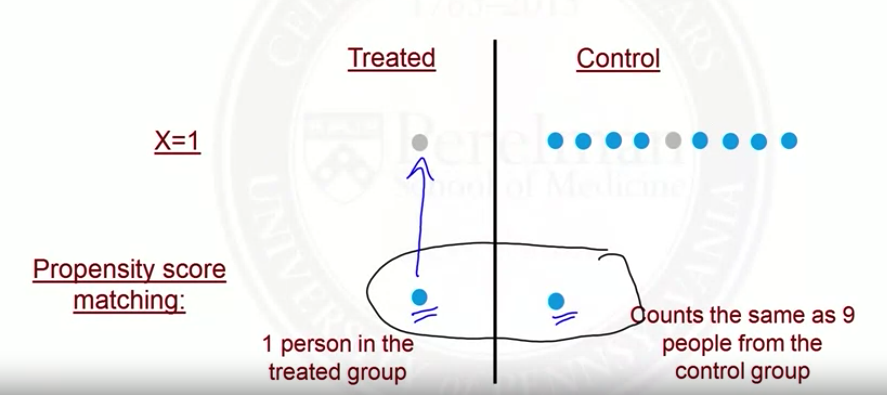
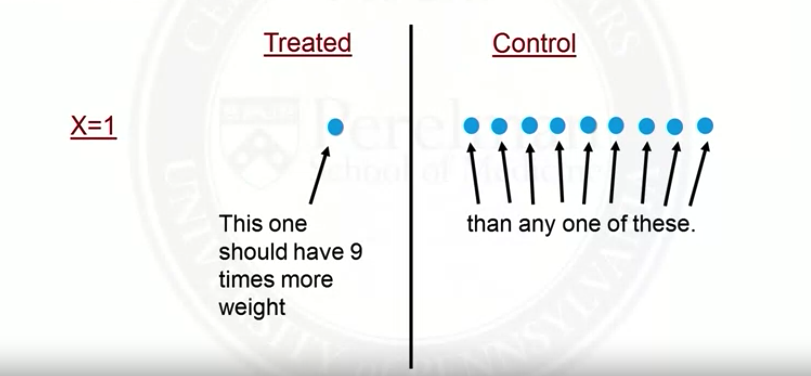
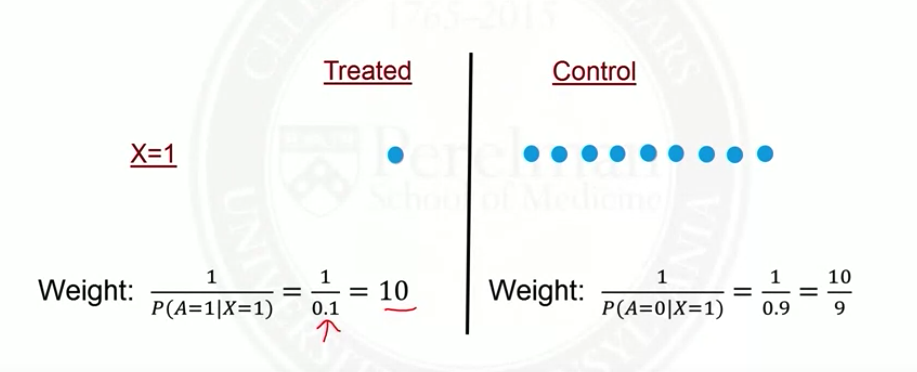
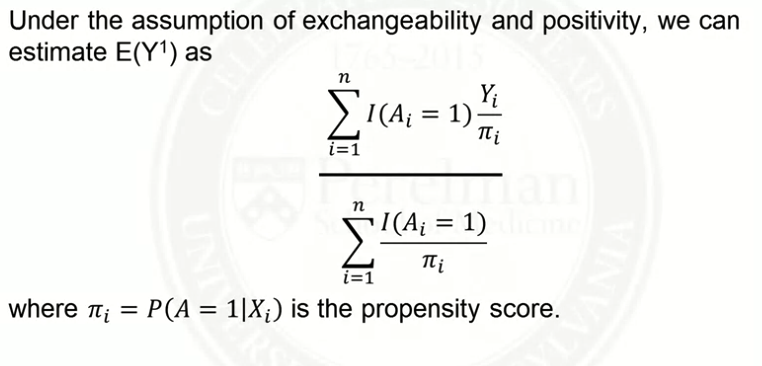
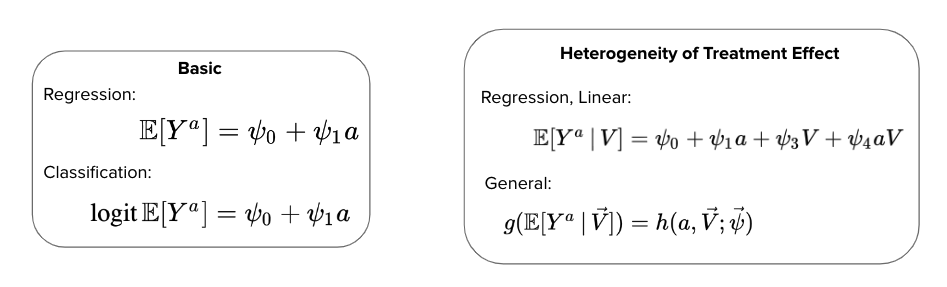
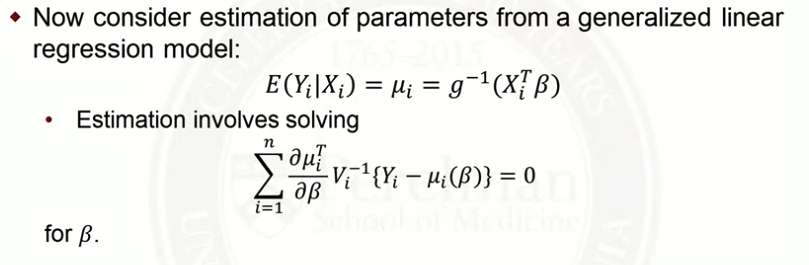
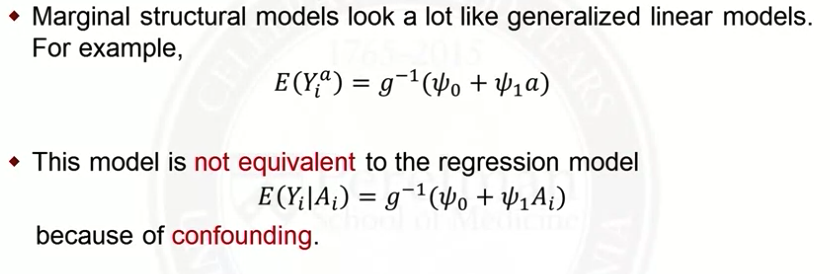
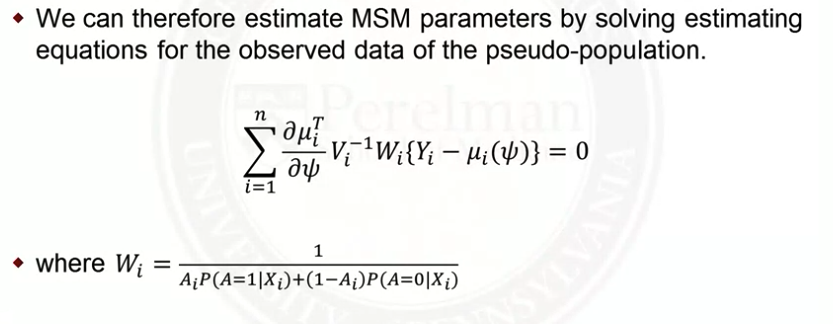
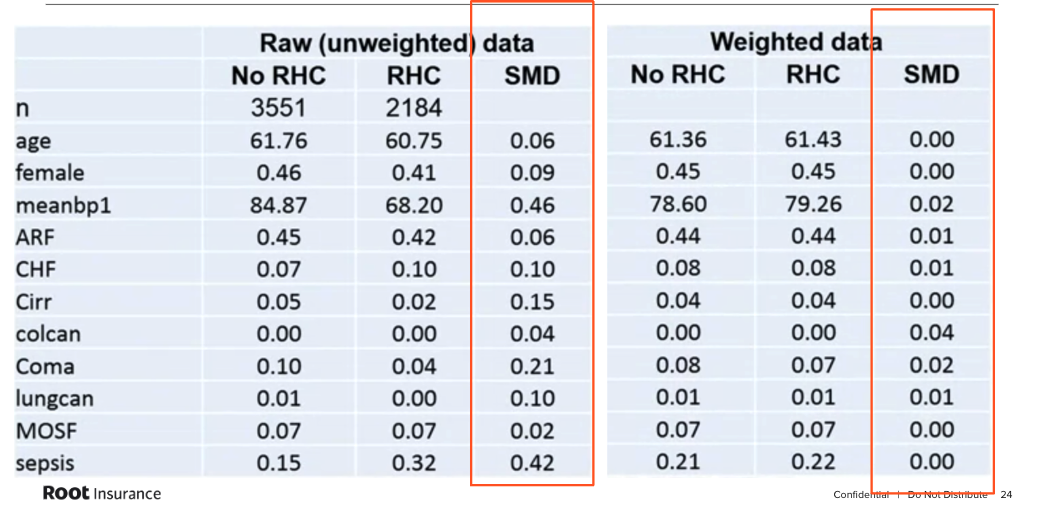
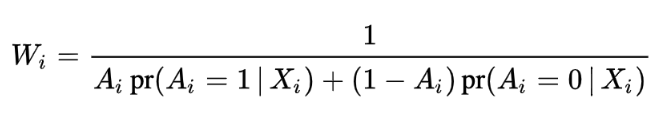
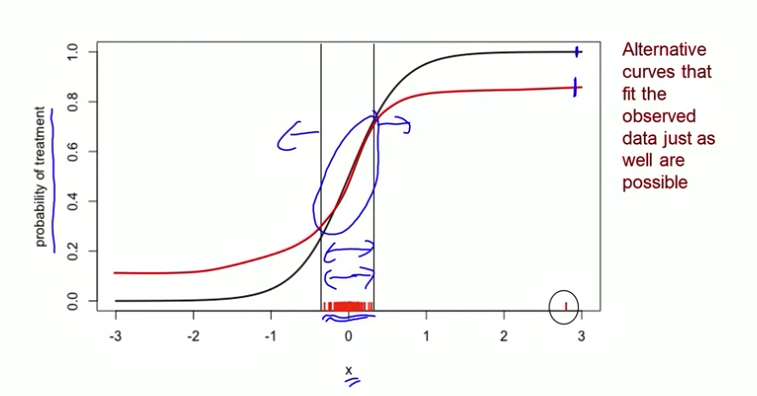
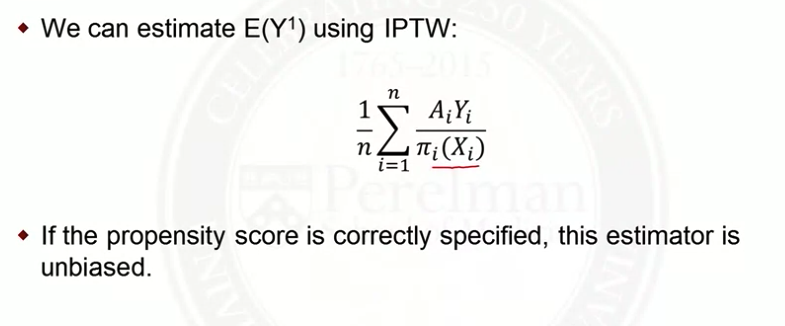
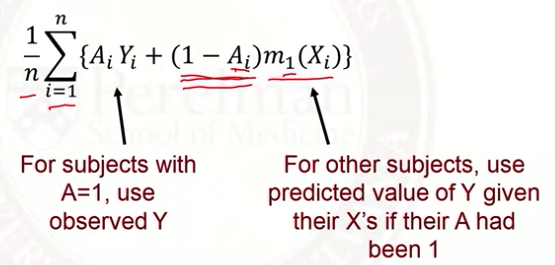
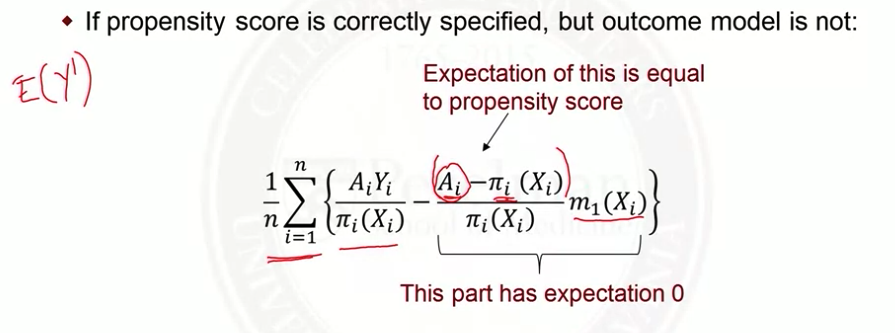
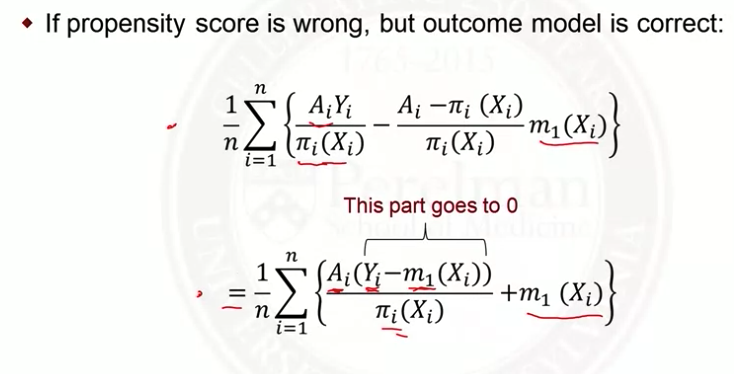
* **Matching:** attempt to make an observational study more like a randomized trial
  + Match individuals in treated group (A=1) to individuals in the control group (A=0) on covariates X
* **Metric of closeness:**
  + Mahalanobis
    - **
    - S describes the COV(X)
    - This variance estimates that squared distance between two numerical instances standardized by the respective covariances
  + Robust Mahalanobis
    - Use ranks of data-points as distances
    - Prevents high impact of outliers, however, it may be problematic for binary data
* **Greedy matching**
  + Calculate distance between each treated subject with every control subject
  + Randomly order the data, move through the list matching each treated to the control with the smallest distance, remove the matched control from available matches
  + Advantages: intuitive, fast
  + Disadvantages: not invariant to initial order of list, not optimal
  + 1 to k matching:
    - Circle through the list k times removing controls that have been matched
    - Larger sample size, higher bias, lower variance
  + Can also consider defining a caliper: maximum acceptable distance, to avoid bad matches
* **Optimal matching**
  + Minimize total distance across all pairs
  + Computationally expensive
* **Assessing balance**
  + Standardized mean difference, values <0.1 adequate balance
    - 
  + Example





* **Analyzing data after matching**
  + Randomization tests (permutation tests, exact tests)
    - Compute test statistic, permute treated and control labels, compare statistics
    - Equivalent to McNemar’s Test
  + Other outcome models:
    - Conditional logistic regression (binary outcome)
    - Stratified Cox model (time to event)
    - Generalized estimating equations (clusters by match id)
* **Sensitivity Analysis**
  + Overt bias: imbalance exists and is caused by observed confounders
  + Hidden bias: Imbalance is caused by an unobserved confounder
  + Goal
    - Determine sensitivity of hidden bias
    - Estimate direction of hidden bias effect
    - Set bounds to hidden bias impact
  + πj probability of treatment j
    - For 2 matched instances j and k, if they are a perfect match and πj = πk then there is no hidden bias
  + Odds of treatment for person j over odds of treatment for person k:
  + 
    - If Γ= 1, then no overt bias -> two individuals with the same covariates X have the same probability of treatment
    - Γ > 1 implies hidden bias
      * E.g. Γ = 2 -> one individual twice as likely to receive treatment even with same covariates X because of unmeasured covariate u
  + Suppose we have evidence of a treatment effect (under assumption that Γ = 1)
    - Increase Γ until evidence of treatment goes away
    - E.g. if happens when Γ = 1.1, then very sensitive to hidden bias
    - If happens when Γ = 5, then not very sensitive
  + Example: Male welder and male control matched for age and smoking, table of one-sided p-value for departures from random assignment:
    - 
    - Calculate p value for various Γ
      * More info starting [page 76 of paper](http://www.stewartschultz.com/statistics/books/Design%20of%20observational%20studies.pdf)
* **Propensity Scores**
  + Probability of receiving treatment (A=1) given covariates X (e.g. score of 0.3 means 30% chance of treatment)
    - Propensity score for subject i: πi=P(A=1|Xi)
    - E.g. age is the only X variable and older people more likely to get treatment, then P(A=1|age=60) > P(A=1|age=30)
  + Need to estimate P(A=1|X)
    - Estimate using logistic regression: outcome A, covariates X
    - Predicted probability is the estimated propensity score
* **Propensity Score Matching**
  + Matching on propensity scores should achieve balance
    - P(X=x|π(X)=p, A=1) = P(X=x|π(X)=p, A=0)
    - Propensity score is a scalar -> matching problem is simplified so we’re only matching on one variable
  + Compare overlap:
  + 
  + 
    - If lack of overlap: can trim tails, makes positivity assumption more reasonable
  + Matching
    - Use nearest neighbor or optimal matching (as before)
      * Logit of propensity score often used
      * Propensity score is bound between 0 and 1, making many values seem similar
      * Logit is unbound
    - Can use caliper (maximum tolerable distance) to avoid bad matches
      * Typical choice: 0.2 times the standard deviation of logit of the propensity score
      * Smaller caliper - less bias, more variance

**Inverse Probability of Treatment Weighting (IPTW)**

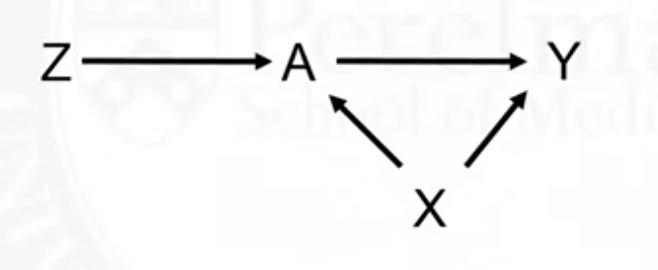
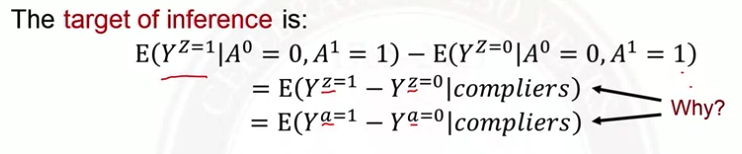
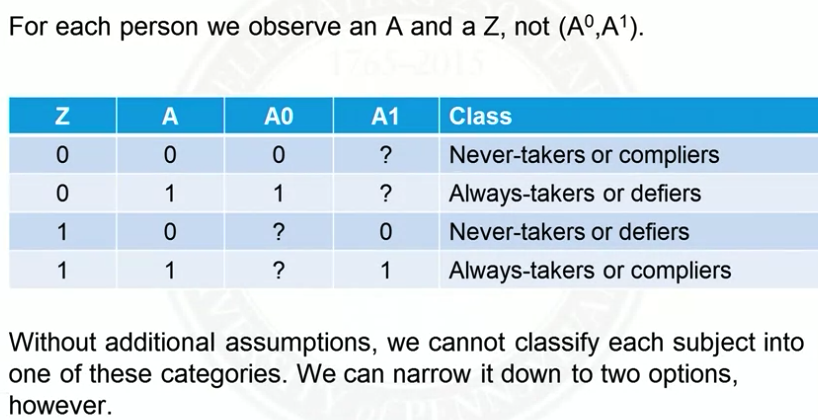
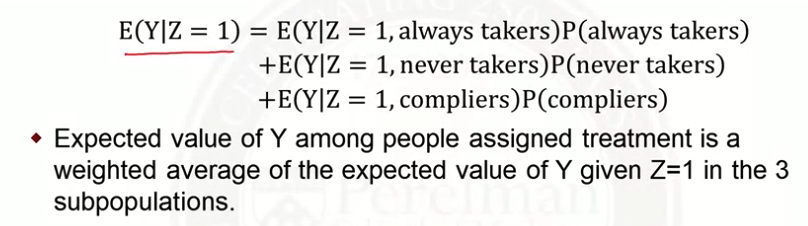
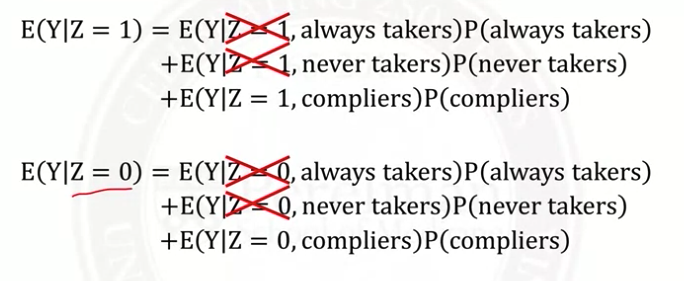
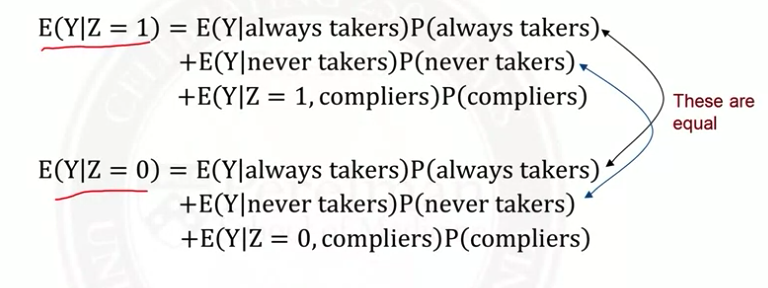
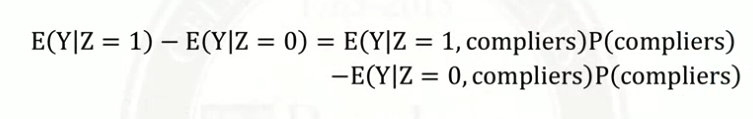
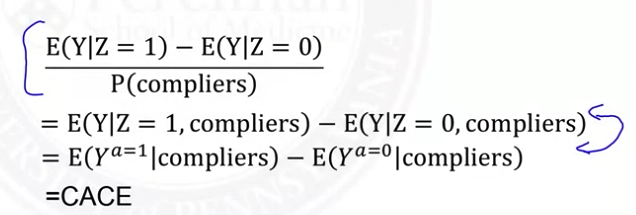
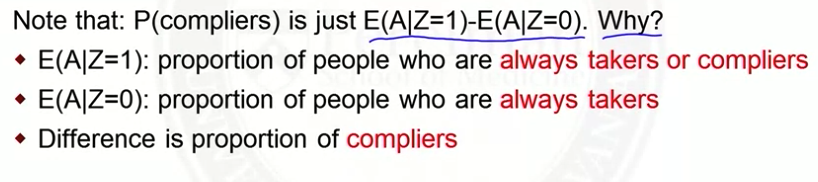
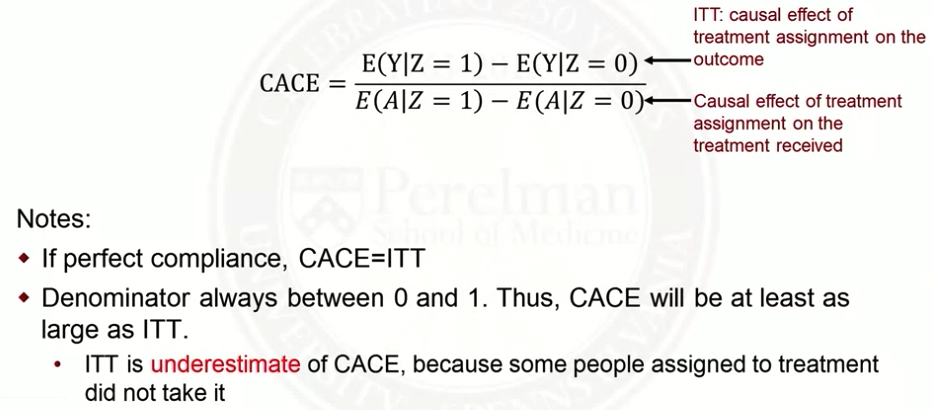
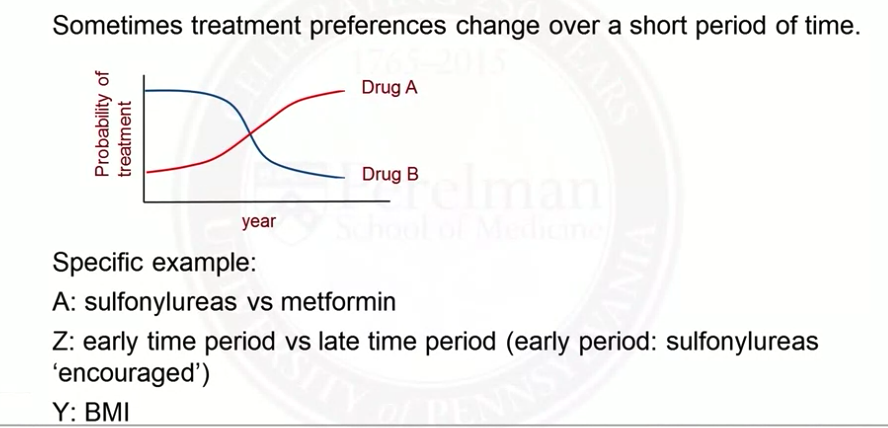
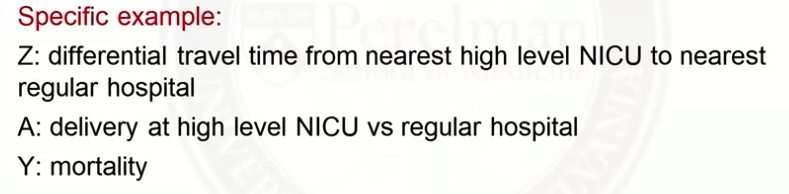
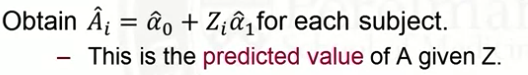
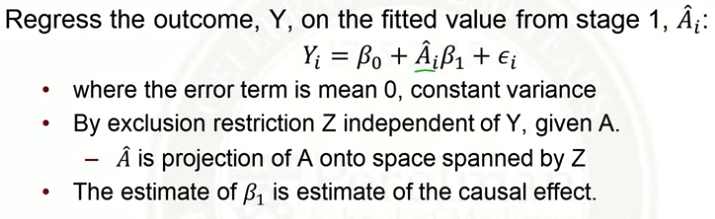
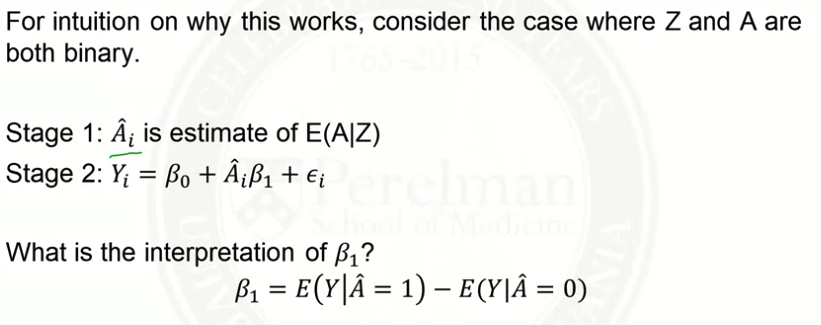
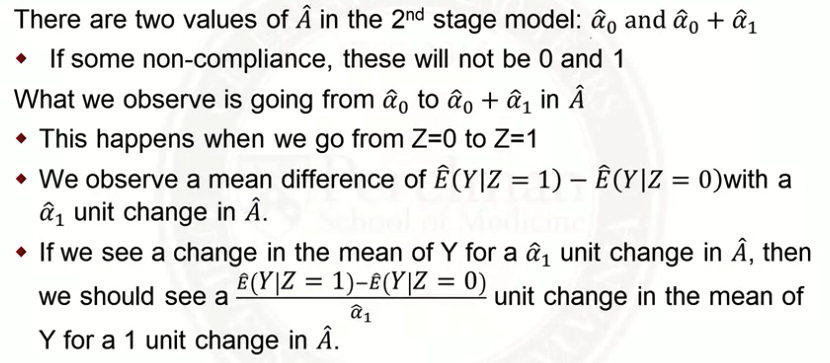
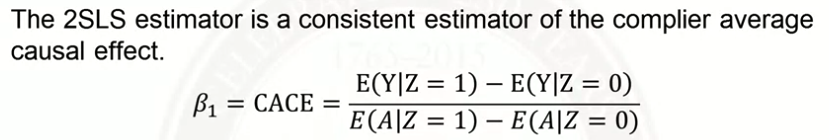
* **Intuition:**
  + 
  + 
  + Weighting:
    - Treated subjects, weight by the inverse of P(A=1|X)
    - Control subjects, weight by the inverse of P(A=0|X)
    - 
  + Creates pseudo-population where confounding is removed (treatment assignment no longer depends on X)
* **More intuition**
  + In surveys it’s common to over-sample some groups relative to the population (e.g. a minority group, older adults)
  + 
* **Marginal Structural Models**
  + Marginal: model that is not conditional on the confounders
  + Structural: model for potential outcomes, not observed outcomes
  + 
  + Want the expected value of potential outcomes (a = 0 or 1), estimate ψ0 and ψ1 
    - In basic regression case: ψ1 is the average causal effect E(Y1) - E(Y0)
    - In basic logistic case: exp(ψ1) is the causal odds ratio
  + Could also have variable V that modifies the effect of a (e.g. sex, race)
    - E(Y1|V) - E(Y0|V) = ψ1 + ψ4V
  + General case: g() is a link function, h() is a function specifying parametric form of a and V
* **IPTW Estimation**
  + For a generalized linear regression model: 
  + 
  + But, pseudo-population (from IPTW) is free from confounding: 
  + MSM Steps:
    - Estimate propensity score
    - Create weights
      * For treated: 1/(propensity score)
      * For control: 1/(1-propensity score)
    - Specify the MSM of interest
    - Fit weighted generalized linear model
    - Use asymptotic (sandwich) variance estimator (or bootstrapping) to determine standard errors, which account for the fact that the pseudo-population might be larger than the sample size.
* **Assessing Balance**
  + Use standardized mean difference (from before), but use weighted means/variance
  + Example: 
  + If imbalance after weighting:
    - Refine propensity score model (maybe add interactions, non-linearity)
* **Distribution of Weights**
  + Larger weights lead to noisier estimates of causal effects
    - E.g. 1 person with a weight of 10,000 represents 10,000 people
  + Can estimate standard errors using bootstrapping
    - Someone with a large weight will be included in some samples but not others
      * Whether they’re included will have a large impact on the parameter estimates
  + A large weight means the probability of that treatment was very small (for treated subjects)
  + 
    - Large weights indicate near violations of the positivity assumption
      * People with certain values of the covariates are unlikely to get one of the treatments
* **Remedies for Large Weights**
  + Investigate why the weights are large:
    - What’s unusual about the subjects with large weights?
    - Is there a problem with their data?
    - Is there a problem with the propensity score model?
  + Can I adjust the propensity score model to fit the majority of the data just as well, and be less extreme for outliers?
  + Trimming: could trim the tails to eliminate extreme weights
    - E.g. remove treated subjects with propensity score above 98th percentile of the distribution of controls
    - Remove control subjects with propensity score below the 2nd percentile of the distribution of treated
    - Trimming tails changes the population
  + Truncate the large weights: Determine maximum allowable weight
    - E.g. specific value, or based on percentile
    - Truncation: bias but smaller variance, no truncation: unbiased but larger variance
* **Doubly Robust Estimators** (also known as augmented IPTW estimators -> AIPTW)
  + Option 1: Estimate using IPTW (using propensity score)
  + Option 2: Estimate by specifying an outcome model m1(X)=E(Y|A=1,X) and average over the distribution of X → haven’t discussed outcome models yet:
  + 
  + Instead: Doubly robust estimator is unbiased if either the propensity score model or the outcome regression model are correctly specified
    - 
    - 

**PSM vs IPTW**

* PSM:
  + Advantages: more familiar/intuitive
  + Disadvantages:
    - The imperfect balance of covariates is often ignored (two individuals with the same propensity score are considered equal while they may have a strongly different set of covariates)
    - A measure of proximity between propensity scores may be arbitrary
    - Some subjects have to be excluded because no match can be found and the analysis is therefore restricted to a sub-population that is not explicitly described.
* IPTW (using propensity score weighting)
  + Advantages:
    - Uses entire population (unless we trim)
    - Can include more than two comparisons
  + Disadvantages:
    - Extreme weights at tails of propensity score distribution increase the variance between covariates
    - Less intuitive
* Doubly robust estimator becoming standard practice

Sources: [stack exchange thread](https://stats.stackexchange.com/questions/424223/advantages-and-disadvantages-of-iptw-vs-propensity-score-matching), [article](https://www.futuremedicine.com/doi/10.2217/cer-2020-0013#:~:text=The%20main%20conceptual%20difference%20between,for%20the%20treated%20(ATT).)

**Instrumental Variables Methods**

* **Introduction**
  + What is there’s unmeasured confounding that we can’t control for?
    - Instrumental Variable: affects the treatment but doesn’t (directly) affect the outcome -> think of like an *encouragement*
    - 
  + Example:
    - A = smoking during pregnancy
    - Y = birthweight
    - X = mother’s age, weight, etc.
    - Challenge: unethical to assign smoking to some pregnant women
    - Z = randomize to either receive encouragement to stop smoking (Z = 1) or receive usual care (Z = 0)
  + Other IV examples:
    - Quarter of birth, geographic distance to specialty care provider
* **Randomized Trials with Noncompliance**
  + Randomized trial:
    - Z = randomization to treatment
    - A = treatment received
      * Not everyone assigned treatment will actually receive treatment (non-compliance)
      * AZ=1 = A1 = value of treatment if randomized to Z = 1
    - Y = outcome
  + Average causal effect of treatment assignment on treatment received
    - E(A1 - A0)
    - Proportion treated if everyone was assigned to receive treatment - proportion treated if no one had been assigned to receive treatment
    - Estimable from observed data: E(A1) = E(A|Z = 1)
  + Average causal effect of treatment assignment on the outcome (intention-to-treat effect)
    - E(YZ=1 - YZ=0)
    - Average values of the outcome if everyone had been assigned to receive treatment - average outcome if no one had been assigned to receive the treatment
    - Estimable from observed data: E(YZ=1) = E(Y|Z=1)
  + What about the causal effect of treatment received on the outcome?
* **Compliance Classes**
  + Classify based on potential treatment 
    - Never and always takers: no variation in treatment received, no information about causal effect
  + IV methods focus on local average treatment effect (not average causal effect for the population)
  + 
    - Since restricting to compliers, we know Yz=1 = Ya=1
    - Contrasts counterfactuals in a common population, known as compiler average causal effective (CACE)
  + 
* **Assumptions**
  + Variable is an instrumental variable if:
    - It’s associated with the treatment
    - It affects the outcome only through its effect on treatment -> exclusion restriction
  + Monotonicity assumption: there are no defiers
    - Probability of treatment should increase with more encouragement
    - 
* **Causal Effect Identification and Estimation**
  + Want to estimate E(Ya=1 - Ya=0|compliers) -> average effect of treatment received among compliers
    - Start with intention to treat (ITT) effect: E(Yz=1 - Yz=0) = E(Y|Z=1)-E(Y|Z=0)
  + 
    - But, for always/never takers, Z does nothing:
    - 
  + Now consider the difference: 
  + 
  + Divide both sides by P(compliers):
  + 
  + 
* **IVs in Observational Studies**
  + Z (instrument) can be thought of as encouragement:
    - Binary: encourage yes or no
    - Continuous: ‘dose’ of encouragement
  + Hard to find variable that affects treatment but not outcome
    - Can check with data if it affects treatment
    - Rely on subject matter knowledge to know if it affects outcome
  + Examples:
    - Calendar time
    - Distance:
      * Distance to a specialty care center, shorter distance is stronger encouragement
* **Two Stage Least Squares**
  + In ordinary least squares, we assume the error term and covariate (A) are independent
    - If there is confounding, these will be correlated
  + Two stage least squares is a method for estimating a causal effect in the instrumental variables setting
  + Stage 1: Regress treatment received A on the instrumental variable Z
    - By randomization, Zi and are independent
    - 
  + Stage 2: 
  + 
  + 
  + The denominator below is just 
  + Sensitivity analysis for IV assumptions:
    - Exclusion restriction: If Z does directly affect Y by an amount p, would my conclusions change? Vary p
    - Monotonicity: If the proportion of defiers was π, would my conclusions change?
* **Weak Instruments**
  + The strength of an IV is how well it predicts treatment received
    - Strong instrument: highly predictive of treatment, encouragement greatly increases the probability of treatment
    - Weak instrument: weakly predictive of treatment, encouragement barely increases the probability of treatment
  + Measure the strength of an instrument:
    - Estimate the proportion of compilers: E(A|Z=1) - E(A|Z=0)
  + Weak instruments (small proportion of compliers): lead to large variance estimates
  + Alternatives: near/far matching, match so covariates are similar but the instrument is very different