Statistical Learning

Lecture 11a - Big Data - Causality

ANU - RSFAS

Last Updated: Wed May 18 09:17:28 2022

Considerations in High Dimensions

- Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting ($p << n$).
- This is do to the fact that historical scientific questions were based on the fact that small amounts of data could be collected.
- This lead to statistical methodologies being developed to answer those questions based on the available data!
- For example: Y is a patient's blood pressure, and $X = \{\text{age, gender, and body mass index}\}$.

$$Y_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Gender}_i + \beta_3 \text{BMI}_i + \epsilon_i$$

- New technologies have changed the way that data are collected in fields as diverse as finance, marketing, and medicine.

- It is now commonplace to collect an almost unlimited number of feature measurements.

- For example:

  1. Rather than predicting blood pressure on the basis of just age, gender, and BMI, one might also collect measurements for half a million single nucleotide polymorphisms (SNPs; these are individual DNA mutations that are relatively common in the population).

  2. A marketing analyst interested in understanding people's online shopping patterns could treat as features all of the search terms entered by users of a search engine. This is sometimes known as a "bag-of-words" models. For a given user, each of the p search terms is scored present (1) or absent (0), creating a large binary feature vector.

Identifying the Number of Seals from an Image

$X_i =$ 

- (Y_i): adult_males (7), subadult_males (4), adult_females (10), juveniles (1), pups (0)

```
#https://bioconductor.org/install/
#BiocManager::install("EBImage")
library("EBImage")
```

```
## Warning: package 'EBImage' was built under R version 4.1.1
```

```
img <- readImage("42.jpg")
dim(img)
```

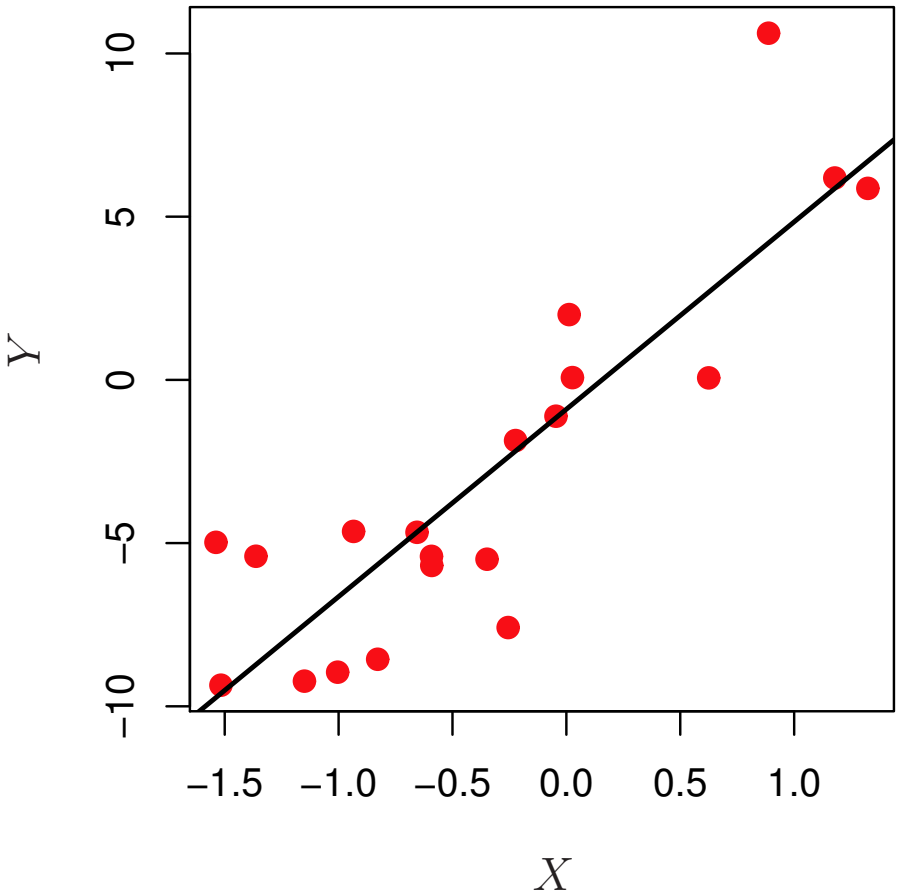
```
## [1] 4992 3328    3
```

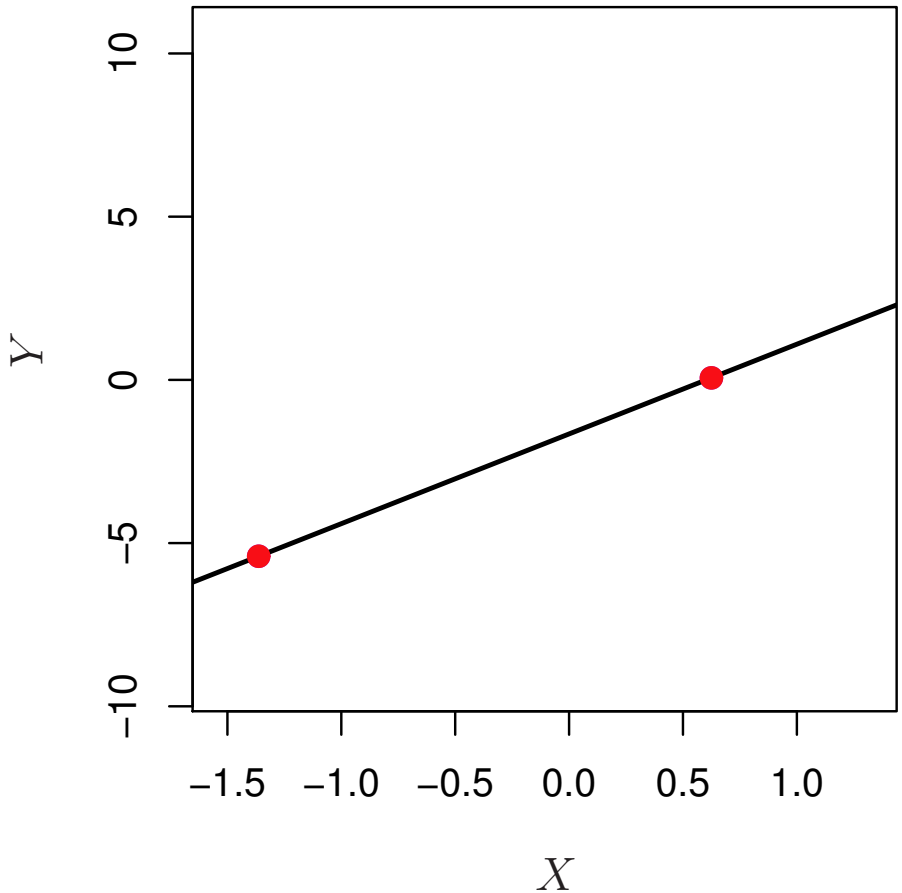
So X_i has $4992 \times 3328 \times 3 = 49,840,128$ pieces of information.

$p >> n$

- When p is close to n or larger, regression cannot or at least shouldn't be used.
- Regardless of whether or not there truly is a relationship between the features and the response, least squares will yield a set of coefficient estimates that result in a perfect fit to the data.

$p \approx n$

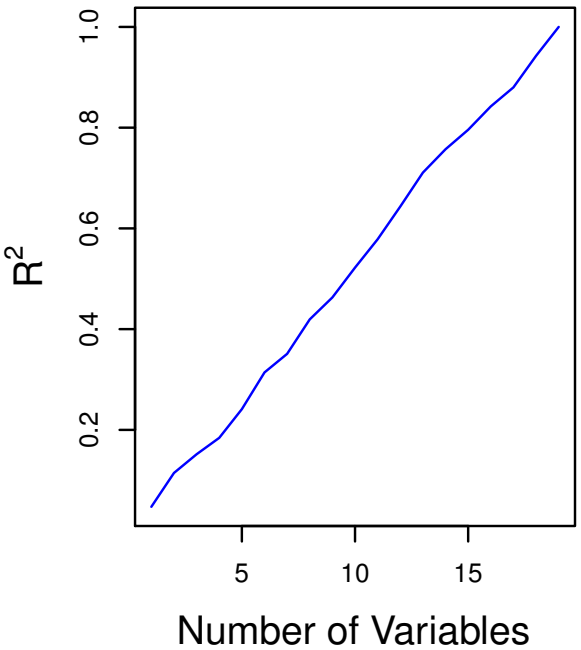


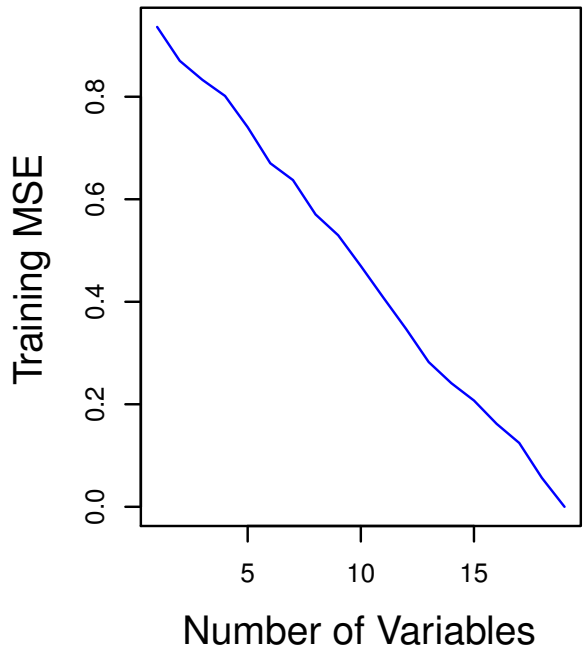


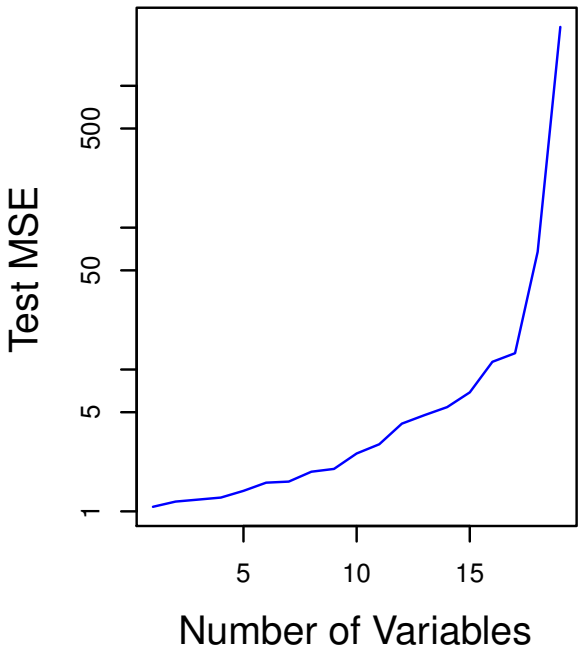
- First Case: $p = 1$ and $n = 20$.
- Second Case: $p = 1$ and $n = 2$.

$p \approx n$

- We can see when $p \approx n$ it is very easy to overfit the data!
- This suggests that standard regression is **too flexible** for these cases!







- Data are simulated for $n = 20$
- Regressions were performed for $p = 1, \ldots, 20$ covariates (X) that were unrelated to Y.

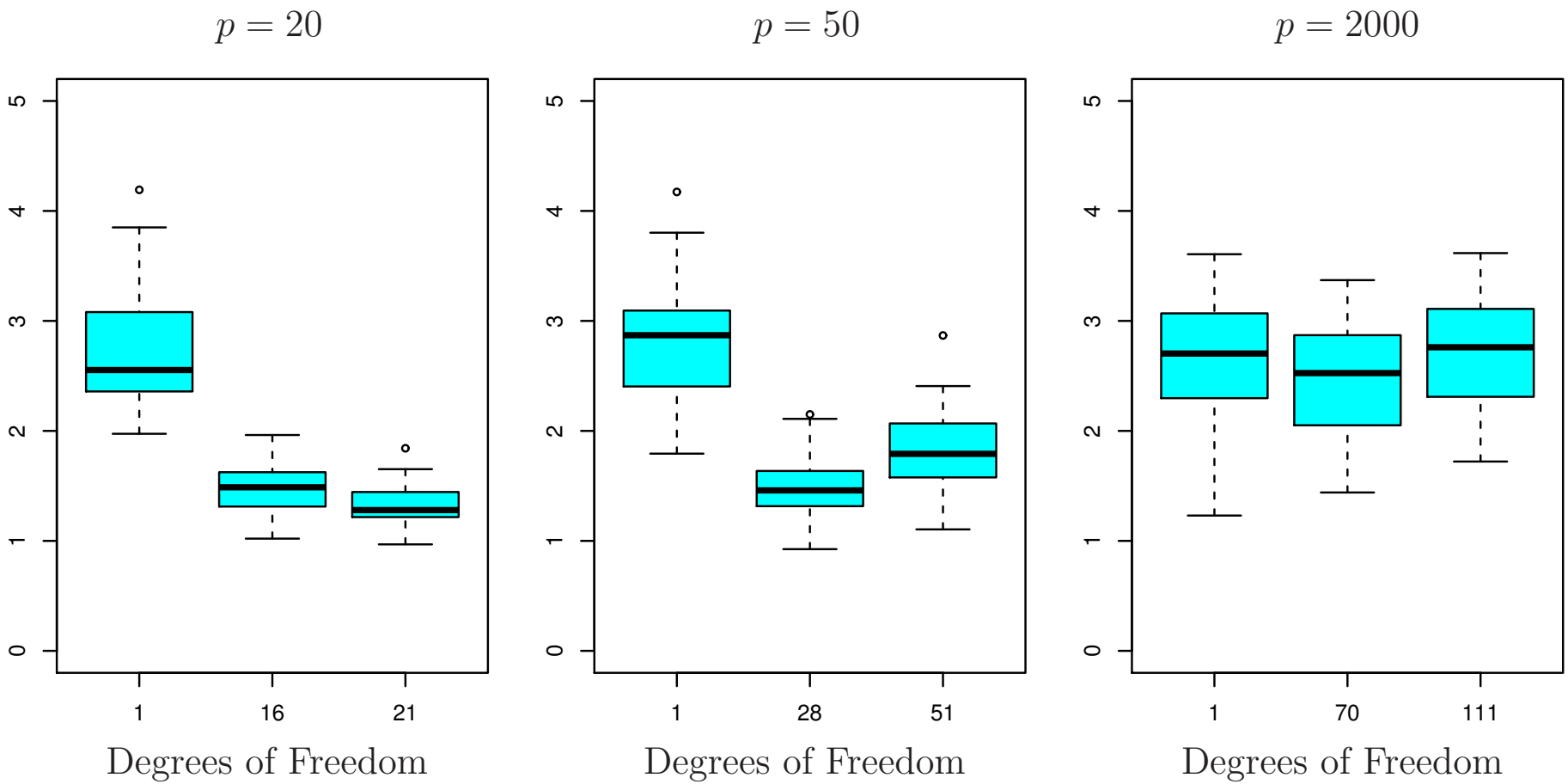
Adjustment?

- We learned a number of approaches to adjust the training RSS: C_p, AIC, and BIC.
  - These approaches are not appropriate in the high-dimensional setting, because estimating $\hat{\sigma}^2$ is poorly estimated or may be zero in many cases!
- What is needed is an approach to fit less flexible models:
  - forward selection
  - principle components regression
  - ridge regression
  - lasso regression

Lasso Example

- Data were generated for $n = 100$ training dataset.
- There were $p = 20, 50, \text{ or } 2,000$ covariates (X), of which 20 are truly associated with the outcome (Y).
- The lasso was performed and the MSE was evaluated on an independent test set.
- For ease of interpretation, rather than reporting λ, the degrees of freedom are reported; for the lasso this turns out to be simply the number of estimated non-zero coefficients.

MSE Test Set


$p = 20$
0
1
2
3
4
5
1
16
21
Degrees of Freedom
$p = 50$
0
1
2
3
4
5
1
28
51
Degrees of Freedom
$p = 2000$
0
1
2
3
4
5
1
70
111
Degrees of Freedom

- The figure highlights three important points:
  1. regularization or shrinkage plays a key role in high-dimensional problems
  2. appropriate tuning parameter selection is crucial for good predictive performance
  3. the test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response.

Interpreting Results in High Dimensions

- When we perform the lasso, ridge regression, or other regression procedures in the high-dimensional setting, we must be quite cautious in the way that we report the results obtained.
- Recall the idea of multicollinearity: the concept that the variables in a regression might be correlated with each other.
- In the high-dimensional setting, the multicollinearity problem is extreme: any variable in the model can be written as a linear combination of all of the other variables in the model.

SNPs Example

- Suppose that we are trying to predict blood pressure (Y) on the basis of half a million SNPs [single nucleotide polymorphisms] (X).
- We might consider using forward selection.
- We find that 17 of those SNPs lead to a good predictive model on the training data (via AIC, BIC, etc.).
- Due to multicollinearity and variability due to the small sample size, it would be incorrect to conclude that these 17 SNPs predict blood pressure more effectively than the other SNPs not included in the model.
- There are likely to be many sets of 17 SNPs that would predict blood pressure just as well as the selected model.

- If we were to obtain an independent data set and perform forward step-wise selection on that data set, we would likely obtain a model containing a different, and perhaps even non-overlapping, set of SNPs.
- This does not detract from the value of the model obtained.
- For instance, the model might turn out to be very effective in predicting blood pressure on an independent set of patients, and might be clinically useful for physicians.
- But we must be careful not to overstate the results obtained, and to make it clear that what we have identified is simply one of many possible models for predicting blood pressure, and that it must be further validated on independent data sets.

Interpretation of Regression - Again

- Load and attach the data set (King County House Sales Data - May 2014 to May 2015):

```
h.data <- read.csv("house_data.csv", header=TRUE)
```

Simple linear Regression

```
mod <- lm(log(price) ~ bedrooms, data=h.data)
summary(mod)$coef
```

```
##                Estimate  Std. Error   t value Pr(>|t|)
## (Intercept) 12.3918962 0.012653476 979.32741        0
## bedrooms     0.1946071 0.003618597  53.77973        0
```

- A unit change in x leads to a $\hat{\beta}_1$ change in y.
- Increase the number of bedrooms by 1, the log of the price increases by 0.195.

Multiple Linear Regression

```
mod <- lm(log(price) ~  bedrooms + log(sqft_living) + bathrooms,
          data=h.data)
summary(mod)$coef
```

```
##                     Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       6.79679886 0.067018833 101.41625 0.000000e+00
## bedrooms         -0.07360463 0.003594624 -20.47631 2.633478e-92
## log(sqft_living)  0.84136382 0.010397801  80.91748 0.000000e+00
## bathrooms         0.06933105 0.005248237  13.21035 1.089984e-39
```

- A unit change in x_1, **holding everything constant**, leads to a $\hat{\beta}_1$ change in y.
- Increase the number of bedrooms by 1, **holding the sqft living and number of bathrooms constant**, the log of the price **decreases** by 0.074.
- All of the covariates are **statistically significant**.

Multiple Linear Regression

- For observational data, intuitively, we know that if we increase the number of bedrooms, likely the *sqrft_living* and *bathrooms* will increase! The three variables are positively correlated!

```
cor( data.frame(h.data$bedrooms, log(h.data$sqft_living),
                h.data$bathrooms))
```

```
##                          h.data.bedrooms log.h.data.sqft_living.
## h.data.bedrooms                1.0000000                0.6208637
## log.h.data.sqft_living.        0.6208637                1.0000000
## h.data.bathrooms               0.5158836                0.7613035
##                          h.data.bathrooms
## h.data.bedrooms                 0.5158836
## log.h.data.sqft_living.         0.7613035
## h.data.bathrooms                1.0000000
```

- Interpretation of the model is tricky.

Causality?

- Thus far, our interpretation is about **predicting** a change. Not about causality, just correlation.
- Causality is an active area of research in statistics.
  "More has been learned about causal inference in the last few decades than the sum total of everything that has been learned about it in all prior recorded history" - *Gary King*
- Two main approaches:
  - Donald Rubin (re-balance data)
  - Judea Pearl (need additional structure conveyed by a graph)

Causality

- What do we mean? We take the view that the causal effect of an action is the difference between the outcomes where the action was or was not taken.

- To make things simple, consider $T = 0$ a control (placebo) and $T = 1$ a new treatment (eg. pain medication).

- The **causal effect** for person i is defined as:

$$\delta_i = Y_i^1 - Y_i^0$$

where y is a measured outcome (eg. resting heart rate).

- From the definition we can see that each patient either takes the placebo or the new treatment, but not both! **This is the fundamental problem with causal inference!**

Experiments

- In a designed experiment, we have control over T. For example, suppose we wish to **compare two physical exercise regimes**.
- The experimental units are the people we use for the study.
- There may be some other potential predictors which **we can control** such as the amount of time spent exercising or the type of equipment used.
- Some other predictors **might not be controlled**, but can be measured, such as the physical characteristics of the people.
- Still other predictors **may not be controlled or measured**. We may know about these predictors or we may be unaware of them.

Randomization is the Key to Success

Experiments - Simple Case

- Suppose we only vary T.
- We have a set of n patients.
- We randomize individuals in groups $T = 0$ and $T = 1$.
- We cannot estimate δ_i for each patient, but we can estimate $\bar{\delta}$ over the experimental units.
- What does randomization due for us?

  Achieves balance of the unmeasured and measured differences that we did not specifically set!

  Randomization removes the *arrow* from Z to T!

$$Z \dashrightarrow \not\dashrightarrow T \longrightarrow Y$$

Observational Data

- We have no control over the **assignment of** T. Individuals make their own choices!
- Example: $T = 0$ (placebo) vs $T = 1$ (vitamin). The measured outcome could be heart health measured through an exercise stress test.
- Perhaps we find **strong positive relationship** between taking vitamins and heart health.
- Should we alert the media? Not yet. Active individuals may be more likely to take vitamins than non-active people. This would create a "correlation" between vitamins and heart health.

$$Z \longrightarrow T \longrightarrow Y$$

Observational Example

- On the 8th January 2008, primaries to select US presidential candidates were held in New Hampshire.
- In the Democratic party primary, Hillary Clinton defeated Barack Obama.
- Two different voting technologies were used: paper vs digital.
- Obama had more votes on the paper ballots, while Clinton had more votes on the digital ballots.
- Since the method of voting should make no causal difference to the outcome, suspicions were raised regarding the integrity of the election.
- The data was derived from Herron et al. (2008) where a more detailed analysis may be found.

```
data(newhamp, package="faraway")
summary(newhamp)[, 1:3]
```

```
##  votesys     Obama           Clinton
##  D:174   Min.   :  34.0   Min.   :  30.0
##  H:102   1st Qu.: 164.2   1st Qu.: 153.0
##          Median : 285.0   Median : 319.5
##          Mean   : 374.2   Mean   : 403.5
##          3rd Qu.: 474.5   3rd Qu.: 480.2
##          Max.   :2779.0   Max.   :2869.0
```

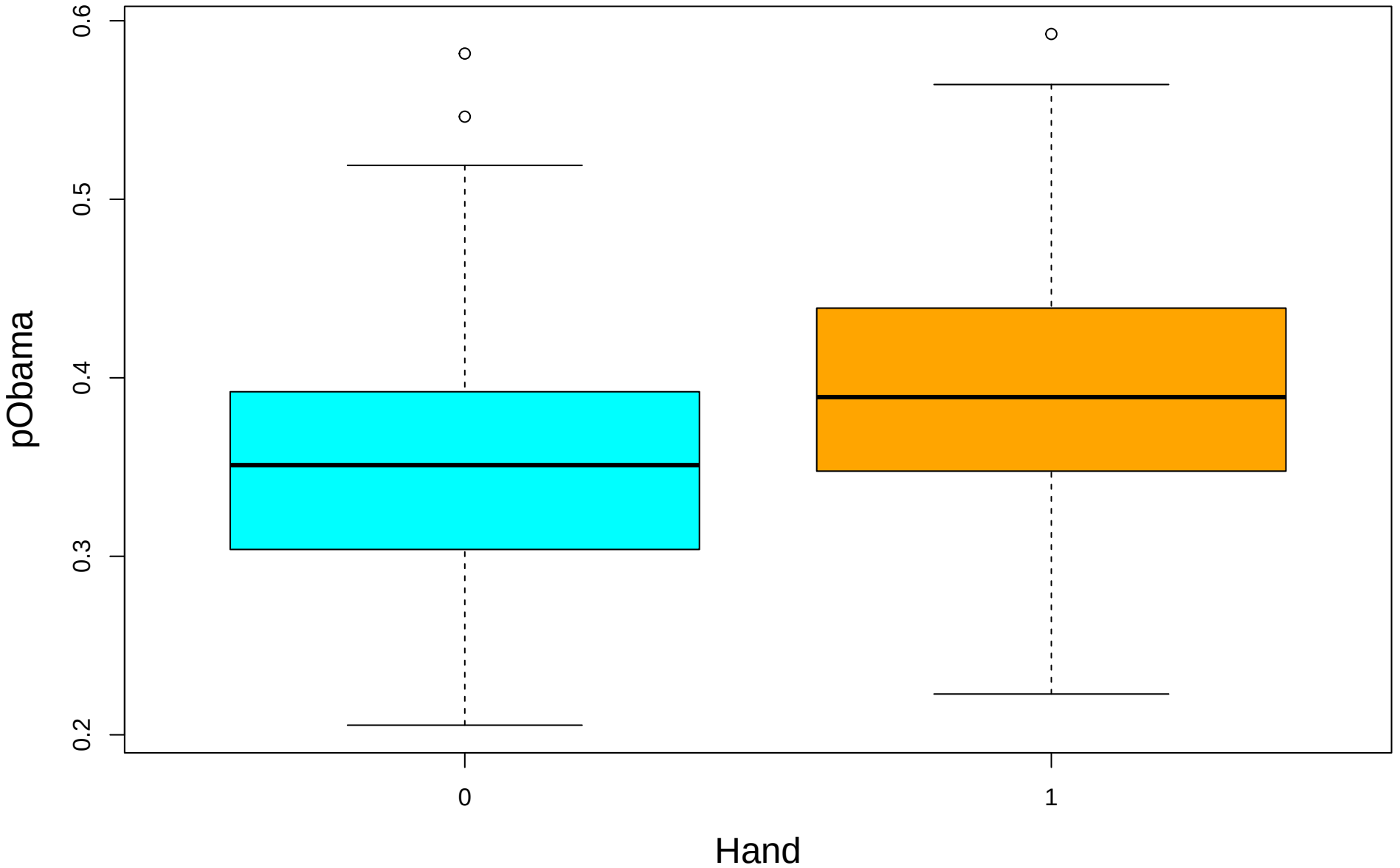
```
colSums(newhamp[newhamp$votesys == "D", 2:3])
```

```
##   Obama Clinton
##   86353   96890
```

```
colSums(newhamp[newhamp$votesys == "H", 2:3])
```

```
##   Obama Clinton
##   16926   14471
```

```
newhamp$trt <- factor(ifelse(newhamp$votesys == "H", 1, 0))
attach(newhamp)
plot(pObama ~ trt, col=c("cyan", "orange"), xlab="Hand",
     cex.lab=1.5)
```


pObama
0.2
0.3
0.4
0.5
0.6
0
1
Hand

- Let's consider the following model:

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

```r
mod <- lm(pObama ~ factor(trt))
summary(mod)
```

```
##
## Call:
## lm(formula = pObama ~ factor(trt))
##
## Residuals:
##       Min         1Q    Median         3Q        Max
## -0.172147 -0.047643 -0.004519  0.040363  0.229107
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.352517   0.005173  68.148  < 2e-16 ***
## factor(trt)1 0.042487   0.008509   4.993 1.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06823 on 274 degrees of freedom
## Multiple R-squared:  0.0834, Adjusted R-squared:  0.08006
## F-statistic: 24.93 on 1 and 274 DF,  p-value: 1.059e-06
```

- Suppose the correct model included another variable (Z):

$$Y_i = \beta_0^* + \beta_1^* T_i + \beta_2^* Z_i + \epsilon_i$$

where (writing the correlation between T and Z as a simple regression with T as our 'covariate' due to our interest in T on Y):

$$Z_i = \gamma_0 + \gamma_1 T_i + \epsilon_i'$$

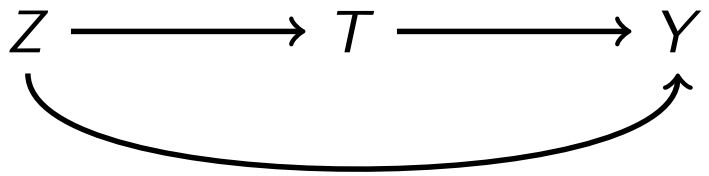
- Z is sometimes called a **confounder**. When the treatment and outcome share a common cause.

- Let's substitute the second equation into the first:

$$Y_i = \beta_0^* + \beta_1^* T_i + \beta_2^* (\gamma_0 + \gamma_1 T_i + \epsilon_i') + \epsilon_i$$

$$Y_i = (\beta_0^* + \beta_2^* \gamma_0) + (\beta_1^* + \beta_2^* \gamma_1) T_i + \epsilon_i''$$

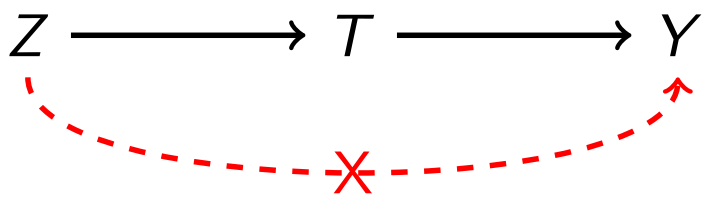
- Model 1: $Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$
- Model 2: $Y_i = (\beta_0^* + \beta_2^* \gamma_0) + (\beta_1^* + \beta_2^* \gamma_1) T_i + \epsilon_i''$
- To get the same results for T we need either:
  - Case 1: $\beta_2^* = 0$ (Z has no effect on y)
  - or Case 2: $\gamma_1 = 0$ (T has no effect on Z)
- Otherwise Z will have an effect on our conclusions and the initial model which excludes Z will provide a biased estimate of the treatment effect.
- Note: In a designed experiment, we have $\gamma_1 = 0$ by the randomization in the assignment of T.



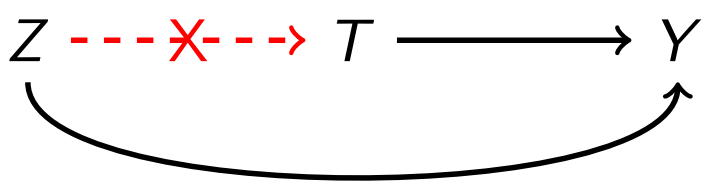
- For Model 1 to be correct (i.e. be used to estimate the average causal effect - average difference between hand (paper) and digital ballots)

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

- Case 1



- Case 2



- Does such a third variable Z exist for the New Hampshire voting example?
- Consider the proportion of votes for **Howard Dean**, a Democratic candidate in the previous presidential campaign in 2004.
- People vote for candidates based on political and character preferences.
- The proportion of voters choosing Howard Dean in the previous Democratic primary tells us something about the aggregate preferences of the voters in each ward in the 2008 primary.

```r
mod2 <- lm(pObama ~ factor(trt) + Dean)
summary(mod2)
```

```
##
## Call:
## lm(formula = pObama ~ factor(trt) + Dean)
##
## Residuals:
##       Min         1Q    Median         3Q        Max
## -0.179302 -0.035803 -0.003828  0.035325  0.213311
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.221119   0.011250  19.655   <2e-16 ***
## factor(trt)1 -0.004754   0.007761  -0.613    0.541
## Dean          0.522897   0.041650  12.555   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05443 on 273 degrees of freedom
## Multiple R-squared:  0.4189, Adjusted R-squared:  0.4146
## F-statistic:  98.4 on 2 and 273 DF,  p-value: < 2.2e-16
```

Matching

- Consider the experiment you would like to conduct.
- You should randomize T among the counties. Then the other variables would be roughly **matched**.
- For observational, we could create matched pairs - similar characteristics for $T_j = 0$ and $T_l = 1$.

```
## install.packages("Matching")
## install.packages("rgenoud")
library(Matching)
set.seed(123)
mm <- GenMatch(trt, Dean, ties=FALSE,
               caliper=0.025, pop.size=1000)
```

- Because *Dean* is a continuous variable, it is unlikely that we will find exact matches so we need to specify how a close a match is acceptable. A caliper of 0.025 means that we accept anything within 0.025 standard deviations of *Dean*.
- For simplicity, we also specify that no ties will be allowed so that each treatment ward will be matched to just one control.
- The matching method uses a genetic algorithm which has a random component.

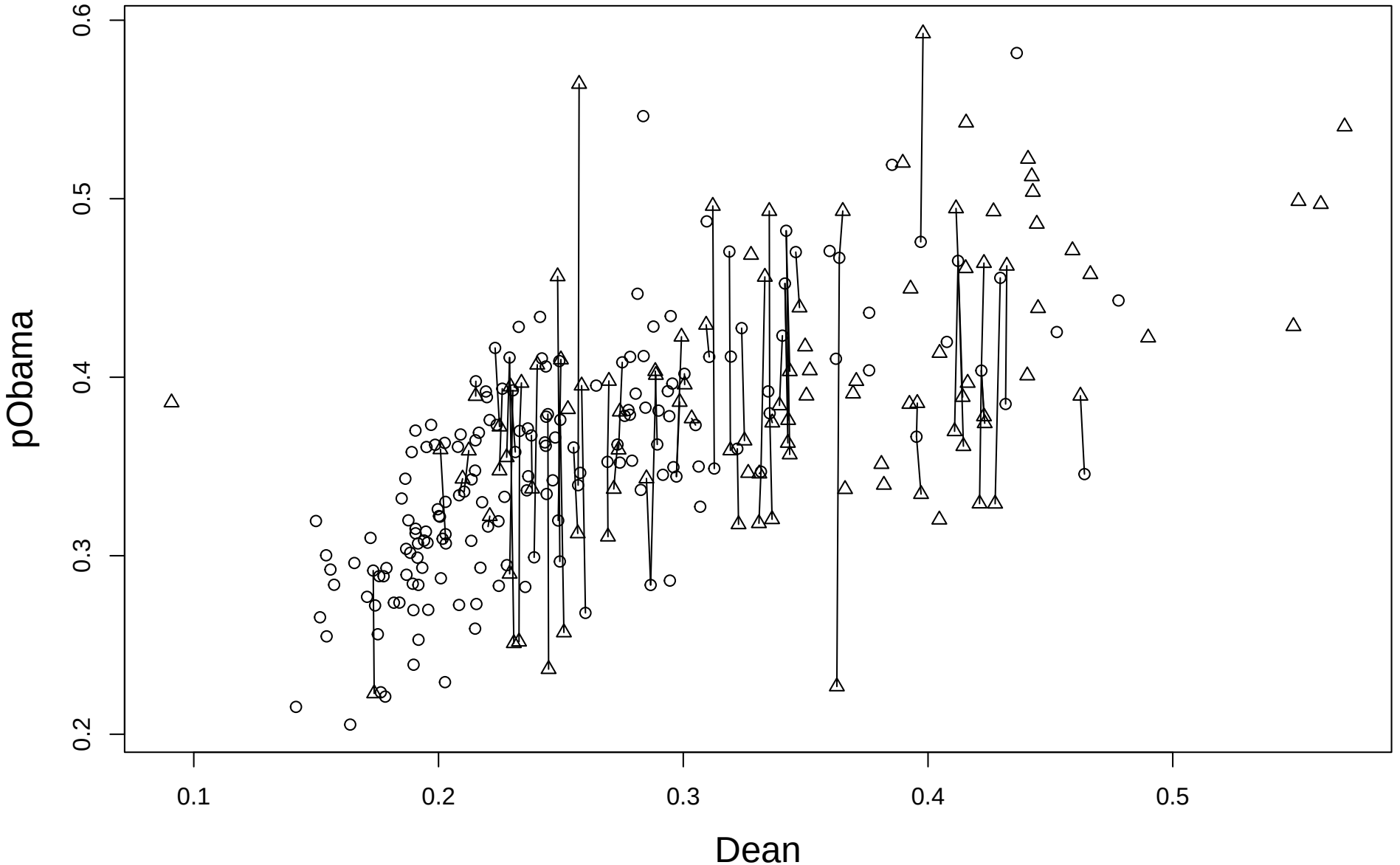
```
head(mm$matches[, 1:2])
```

```
##      [,1] [,2]
## [1,]    4   36
## [2,]   17  242
## [3,]   18    6
## [4,]   19   91
## [5,]   21  233
## [6,]   22  221
```

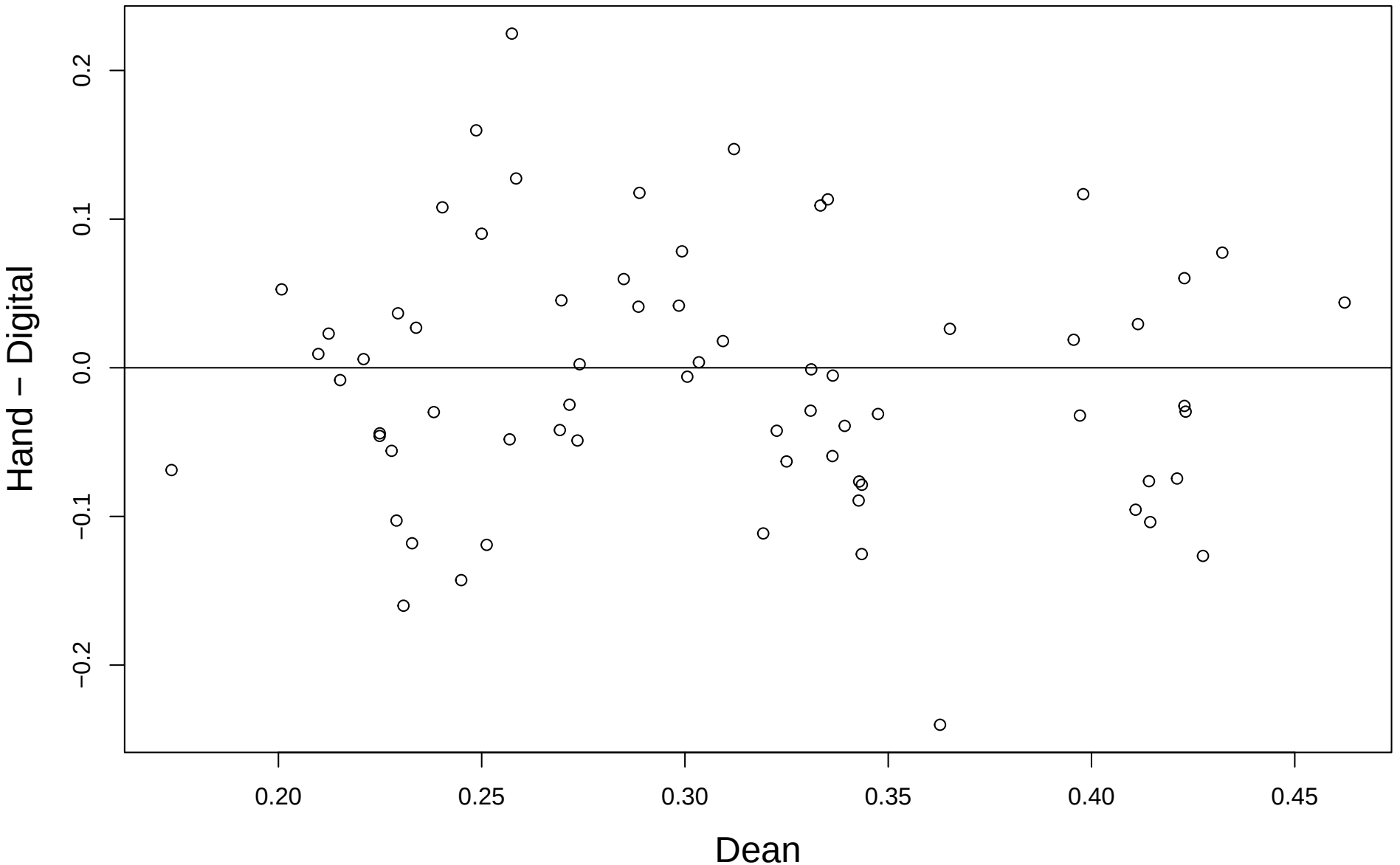
```
##
newhamp[c(4,36), c("Dean","pObama","trt")]
```

```
##                  Dean    pObama trt
## CenterHarbor 0.28495 0.3432836   1
## Fitzwilliam  0.28667 0.2836096   0
```

```
plot(pObama ~ Dean, newhamp, pch=trt+1, cex.lab=1.5)
with(newhamp,segments(Dean[mm$match[,1]],pObama[mm$match[,1]],
                      Dean[ mm$match[,2]],
                      pObama[mm$match[,2]]))
```



```
pdiff <- newhamp$pObama[mm$matches[,1]] - newhamp$pObama[mm$matches [,2]]
plot(pdiff ~ newhamp$Dean[mm$matches[,1]], xlab="Dean",
     ylab="Hand - Digital", cex.lab=1.5)
abline(h=0)
```



t-test Based on the Matched Pairs

```
t.test(pdiff)
```

```
##
##  One Sample t-test
##
## data:  pdiff
## t = -0.86201, df = 67, p-value = 0.3918
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02957785  0.01173578
## sample estimates:
##    mean of x
## -0.008921035
```

Covariate Adjustment

- We fitted two regression models - one with Dean and one without (covariate adjustment)
- We estimated the pair matched difference.
- Both approaches are trying to **estimate** the same quantity.
- The covariate adjustment method, is easier to use and extends well to multiple confounders. It require that we specify the functional form of the covariate in the model in an appropriate way.
- The matching approach is more robust in that it does not require we specify this functional form.
- See Faraway - Linear Models - Chapter 5 for more details.

- What do we do when we have more variables we need to match on?
- If a small number of characteristics: age, gender, ect. we may be able to find exact matches.
- More characteristics (continuous) - we might use a distance measure - mahalanobis
- Use a single measure such as the **probability of receiving treatment** - **propensity score** (eg. modeled through a logistic regression)

$$\text{logit}\left(Pr(T=1)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_P$$

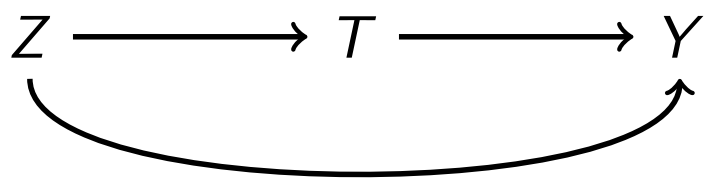
- The propensity score is $\widehat{Pr(T_i = 1)}$
- Match close value of $\widehat{Pr(T_i = 1)}_{[T_i=0]}$ and $\widehat{Pr(T_{i'} = 1)}_{[T_{i'}=1]}$

Qualitative Support for Causation - Faraway Sec 5.7

- **Strength:** We do not mean a high correlation or a small p-value but that $\hat{\beta}$ is large in practical terms.
- **Consistency:** A similar effect has been found for different subjects under different circumstances at different times and places.
- **Specificity:** The supposed causal factor is associated mostly with a particular response and not with a wide range of other possible responses.
- **Temporality:** The supposed causal factor is determined or fixed before the outcome or response is generated.
- **Gradient:** The response increases (or decreases) monotonically as the supposed causal variable increases.
- **Plausibility:** There is a credible theory suggesting a causal effect.
- **Natural Experiment:** A natural experiment exists where subjects have apparently been randomly assigned values of the causal variable.

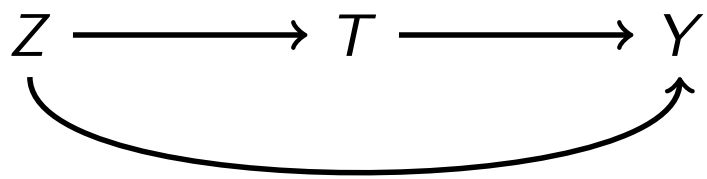
- Several points that Julian Faraway is making suggests that subject matter knowledge is key.

- Judea Pearl believes this subject matter knowledge must be represented by a **Directed Acyclic Graph**. Then from the graph, it can be determined whether a causal effect can be estimated.



DAGS

- Given the graph, can we estimate the causal effect of T on Y?



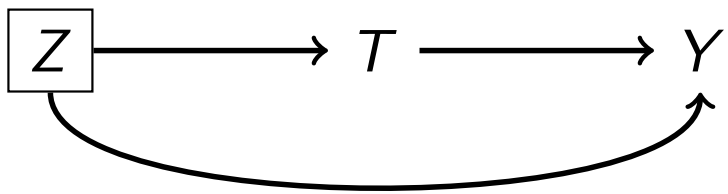
- What are the paths from T to Y?
- Path 1:

$$T \longrightarrow Y$$

- Path 2 (**Backdoor Path** - flow of information):

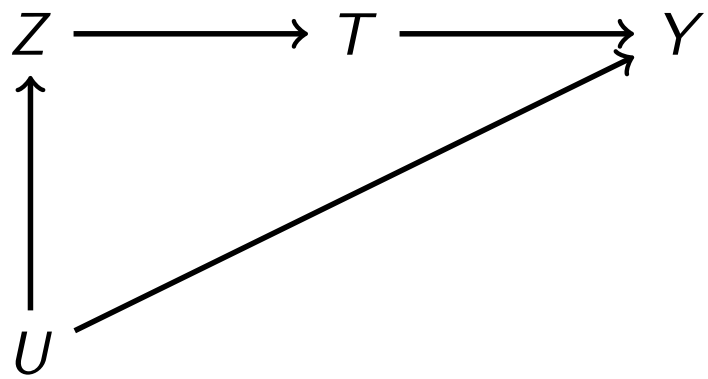
$$T \longleftarrow Z \longrightarrow Y$$

- So block the backdoor path - i.e. condition on Z in your model:

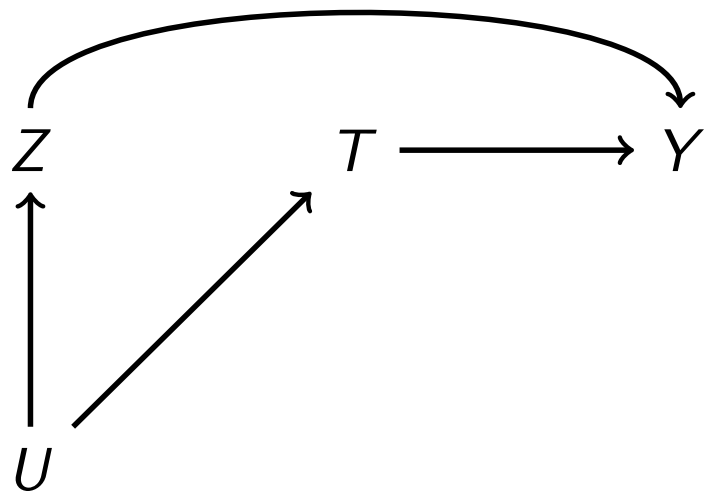


$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \epsilon_i$$

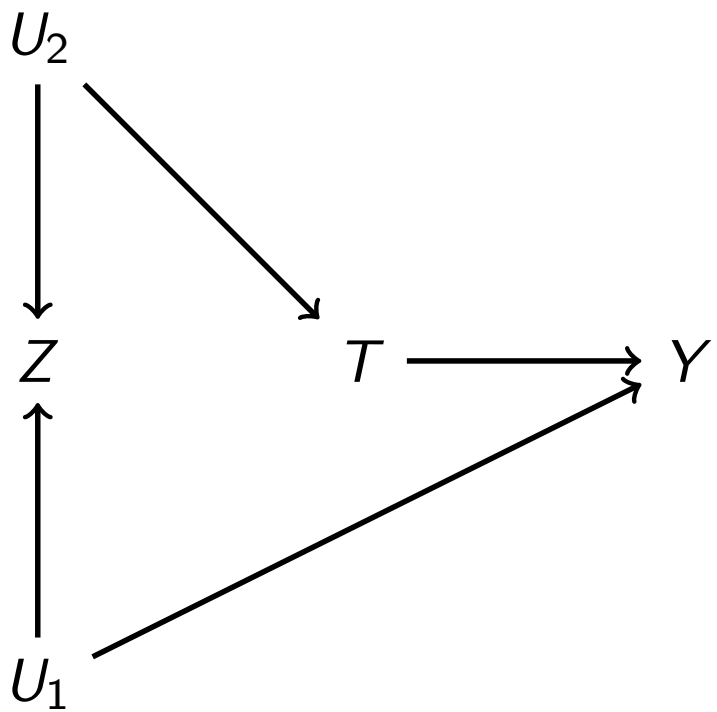
Unmeasured Variables



Unmeasured Variables

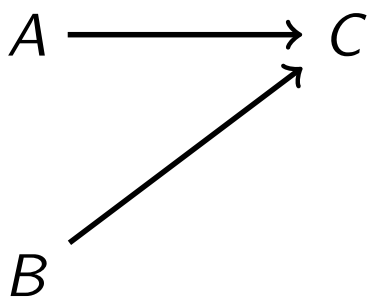


Unmeasured Variables



- Z is a **collider**.

Colliders




```
set.seed(2001)
n <- 10000
A <- rnorm(n)
B <- rnorm(n)
C <- rnorm(n, 3 + 2*A - 4*B, 1)
```

```
library(faraway)
sumary(lm(A ~ B))
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0021596  0.0099850 -0.2163  0.82877
## B            0.0166309  0.0100285  1.6584  0.09727
##
## n = 10000, p = 2, Residual SE = 0.99850, R-Squared = 0
```

```
sumary(lm(B ~ A))
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0033650  0.0099563 -0.3380  0.73539
## A            0.0165353  0.0099709  1.6584  0.09727
##
## n = 10000, p = 2, Residual SE = 0.99562, R-Squared = 0
```

```
sumary(lm(C ~ A + B))
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.9974627  0.0099722  300.58 < 2.2e-16
## A            2.0021981  0.0099882  200.46 < 2.2e-16
## B           -4.0028815  0.0100170 -399.61 < 2.2e-16
##
## n = 10000, p = 3, Residual SE = 0.99721, R-Squared = 0.95
```

```
sumary(lm(A ~ B + C))
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -1.1992631  0.0074517 -160.94 < 2.2e-16
## B            1.6042625  0.0090975  176.34 < 2.2e-16
## C            0.3999492  0.0019952  200.46 < 2.2e-16
##
## n = 10000, p = 3, Residual SE = 0.44569, R-Squared = 0.8
```

Reading - In Rough Order

- *Linear Models with R* [Chapter 5] by Julian Faraway
- *The Book of Why* by Judea Pearl and Dana Mackenzie (popular science book)
- *Causal Inference in Statistics - A Primer* by Judea Pearl, Madelyn Glymour, and Nicholas Jewell
- *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* by Guido Imbens and Donald Rubin
- *Causal Inference: What If* by Miguel Hernan and James Robbins
- *Causality: Models, Reasoning and Inference* by Judea Pearl

Final Thought: P-Hacking: "Hack Your Way To Scientific Glory"

- Interactive Example: "You're a social scientist with a hunch: The U.S. economy is affected by whether Republicans or Democrats are in office. Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are 'statistically significant' by achieving a low enough p-value."

https://projects.fivethirtyeight.com/p-hacking/

- **Be careful when reading the works of others and how you conduct your own research!**
- In your own work, discuss and justify the decisions you made to reach your conclusions. If possible provide data and the code used.