STAT 408
Applied Regression Analysis

Nan Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Linear Regression and Causal Inference

Two Levels of Model Interpretation

- Support we build a linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- After we estimate the model parameter $\hat{\beta}_1$, it has two levels of interpretation

- The first level is <u>association</u>

  One unit increase in X_1 with the other predictors held constant will change $\hat{\beta}_1$ in the response Y <u>on average</u>

- This interpretation may be unrealistic in some cases, and it does not provide <u>causal relation</u>

Causality

- Causal effect is the second level of model interpretation
  - The causal effect of an action is the difference between the outcomes where the action was or was not taken

- Suppose a study applied drug to patients
  - T = 0 for the control (placebo); T = 1 for the treatment (drug)
  - Let y_i^0 be the control response for patient i
  - Let y_i^1 be the treatment response for patient i

- The causal effect for patient i is then defined as

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

Causality

- The challenge in causal inference is that we can only apply treatment or control to patient i the same time,
  - Only y_i^0 or y_i^1 can be observed

- The unobserved outcome is called counterfactual outcome or potential outcome

- This challenge cannot be solved in real word

Randomly Controlled Experiment

- Pseudo-optimal solution is to conduct <u>randomly controlled experiment</u>
  - Randomly assign treatment and control groups
  - Calculate the average difference of response between two groups

- Randomization (almost) guarantees the balance of other variables in treatment and control groups
  - Control and treatment group are “similar” except the assignment

- To avoid unbalance due to randomness, we further use <u>stratified randomization</u>
  - Randomly assign treatment and control separately in male and female to balance the gender

Observational Study

- In most cases, it is not practical or ethical to conduct a controlled experiment
  - Cannot randomly assign smoking, education, …

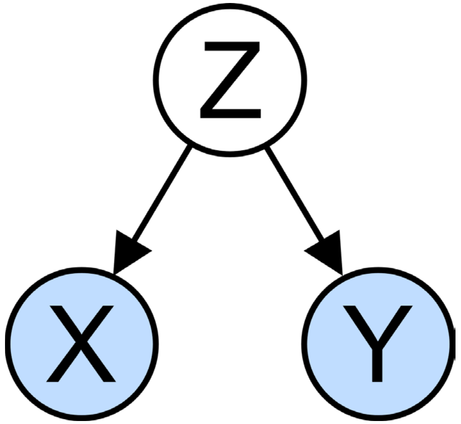
- We have to rely on observation study
  - Cannot control the assignment of treatment or control

- The balance of two groups are not guaranteed due to the possible existence of <u>confounder</u>
  - Exercise vs no-exercise: healthy people tend to do more exercise
  - College vs no-college: wealthy families more likely send children to college

Observational Study

- Confounder affects both treatment/control and response Y

- The change of Y is "caused" by Z, not X



Observational Study: Example

- Let's explore if different voting methods have causal effects on election result
- The data is about the 2008 Dem primary election in New Hampshire

	votesys	Obama	Clinton	dem	povrate	pci	Dean	Kerry	white	absentee	population	pObama
Hinsdale	H	256	331	759	0.0637	16611	0.36610	0.34915	0.97232	0.040836	4213.0	0.3372859
Jaffrey	D	460	462	1223	0.0784	21412	0.24975	0.40967	0.96896	0.070138	5573.0	0.3761243
KeeneWard1	D	416	233	891	0.1072	20544	0.36375	0.29250	0.97132	0.043137	4567.4	0.4668911
KeeneWard2	D	588	402	1433	0.1072	20544	0.36239	0.28073	0.97132	0.054213	4567.4	0.4103280
KeeneWard3	D	503	427	1283	0.1072	20544	0.33471	0.30062	0.97132	0.068720	4567.4	0.3920499
KeeneWard4	D	503	436	1330	0.1072	20544	0.29429	0.32857	0.97132	0.041597	4567.4	0.3781955
KeeneWard5	D	544	424	1347	0.1072	20544	0.37594	0.29041	0.97132	0.076056	4567.4	0.4038604
Marlborough	H	305	188	651	0.0354	19967	0.32768	0.29002	0.98059	0.049813	2064.0	0.4685100

- Each row is one district
- Votesys: ballot counted by hand (H); ballot counted by machine (D)
- We are interested in Obama vs. Clinton

Observational Study: Example

- Among hand-counted ballots, Obama had more votes

```
colSums(newhamp[newhamp$votesys == 'H', c('Obama','Clinton')])
Obama Clinton
16926   14471
```

- Among machine-counted ballots, Clinton had more votes

```
colSums(newhamp[newhamp$votesys == 'D', c('Obama','Clinton')])
Obama Clinton
86353   96890
```

- Let’s fit a linear model with voting system as the predictor, and Obama’s votes proportion as response

$$Y = \beta_0 + \beta_1 T + \epsilon$$

Observational Study: Example

lmod <- lm(pObama ~ votesys, data = newhamp)

summary(lmod)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.352517   0.005173  68.148  < 2e-16 ***
votesysH    0.042487   0.008509   4.993 1.06e-06 ***
```

- The hand voting increases Obama's vote share by 4% on average
- The result is significant
- Is it a causal effect?

Modeling the Confounder

- We suspect variable Z relates to both response (Obama votes) and treatment (machine or hand)
- We use linear model to model these two relations

$$Y = \beta_0^* + \beta_1^* T + \beta_2^* Z + \epsilon$$

$$Z = \gamma_0^* + \gamma_1^* T + \epsilon'$$

- If β_2^* is significant, β_1^* is insignificant, γ_1^* is significant, then Z is a confounder
    - Any change of T causes Z, and Z causes Y

Modeling the Confounder

- The identification of confounder Z relies on domain knowledge
- In this example political scientists propose Z as the variable Dean, the votes for Howard Dean in 2004 primary

$$Y = \beta_0^* + \beta_1^* T + \beta_2^* Z + \epsilon$$

```
lmod <- lm(pObama ~ votesys + Dean, newhamp)
summary(lmod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.221119   0.011250  19.655   <2e-16 ***
votesysH    -0.004754   0.007761  -0.613    0.541    
Dean         0.522897   0.041650  12.555   <2e-16 ***
```

- β_2^* is significant and treatment (counting method) is no longer significant

Modeling the Confounder

- Next, let's model the relationship between Z and treatment

$$Z = \gamma_0^* + \gamma_1^* T + \epsilon'$$

```
lmod <- lm(Dean ~ votesys, newhamp)
summary(lmod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.251289   0.005985  41.986   <2e-16 ***
votesysH    0.090345   0.009845   9.177   <2e-16 ***
```

- $\gamma_1 > 0$ is significant
- We show that Z is a confounder in this problem

Modeling the Confounder

- Dean's voters prefer hand-counted ballot
  - In $Z = \gamma_0^* + \gamma_1^* T + \epsilon'$, $\gamma_1^* > 0$

- Dean's voters also vote for Obama
  - In $Y = \beta_0^* + \beta_1^* T + \beta_2^* Z + \epsilon$, $\beta_2^* > 0$

- Without knowing confounder, we observe hand-counted "causes" more Obama votes

- But this is association: two variables happen to move simultaneously

Matching

- The previous linear model for causal inference is called covariate adjustment
  - With covariate Z fixed, how treatment affect response Y

$$Y = \beta_0^* + \beta_1^* T + \beta_2^* Z + \epsilon$$

- Covariate adjustment replies on the correct model specification

- A model-free way to infer causal effects is matching
  - Find observation pairs in treatment and control group with similar covariates, especially similar confounders
  - In clinical trials, match patients from two groups with same gender, age, income, health condition …

Matching

- In our election data, we try to match based on the Dean variable

```
library(Matching)
newhamp$trt <- ifelse(newhamp$votesys == 'H',1,0)
mm <- GenMatch(newhamp$trt, newhamp$Dean, ties=FALSE)
match <- mm$matches[,1:2]
```

- Match matrix save the indices of matched pairs

	V1	V2
1	4	213
2	17	20
3	18	6
4	19	91
5	21	246
6	22	221
7	23	166

Matching

- We compute the difference of Obama votes among matched pairs and perform a one sample t-test to test the mean of difference

```
pdiff <- newhamp$pObama[match[,1]] - newhamp$pObama[match[,2]]
t.test(pdiff)
```

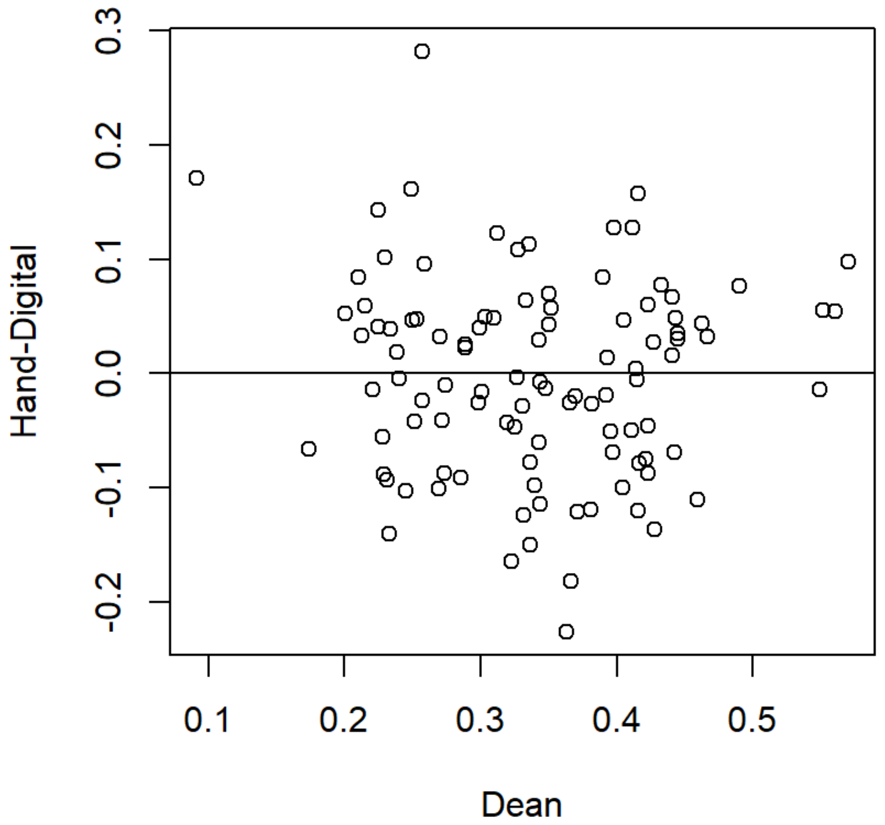
```
        One Sample t-test

data:  pdiff
t = -0.2337, df = 101, p-value = 0.8157
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01910950  0.01508153
sample estimates:
   mean of x
-0.002013984
```

- Statistically, there is no difference of Obama votes among matched pairs

Matching

- We plot the vote difference vs. matched variable Dean

```
pdiff <- newhamp$pObama[match[,1]] - newhamp$pObama[match[,2]]
t.test(pdiff)
```



- The matched pairs show no clear preference for hand or digital voting condition on Dean vote

Matching

- Matching is essentially a model-free covariate adjustment

- If matching seems very good, then why we still use linear regression?