

Introduction to Causal Machine learning

Igor Asanov

February 5, 2024



This handout summarizes the lectures slides. Please note that the handout is not very useful if you do not attend the class. The handout is also not a substitution for the book. The course is built around Martin Huber book: “Causal analysis: Impact evaluation and Causal Machine Learning with applications in R”.

Homepage: <http://www.igorasanov.com/teaching.html>

Literature:

- ! Huber, M., 2023. Causal analysis: Impact evaluation and Causal Machine Learning with applications in R. MIT Press.
- Efron, B. and Hastie, T., 2021. Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press.
- Tibshirani, R.J. and Efron, B., 1993. An introduction to the bootstrap. Monographs on statistics and applied probability, 57(1).
- Course: Causal Inference with Machine Learning by V. Chernozhukov.

Software:

We will use **R** for most of the exercises: Here is the link on the homepage of **R**.

In the class I use **RStudio** as a front end and I would recommend you to install it too - it greatly simplifies workflow in **R**.

Contents

1	Introduction to Machine Learning	3
1.1	Motivation	3
1.1.1	Motivation: Spam Example	3
1.2	In and Out of Sample Predictions	4
1.3	Prediction Study Design	6
1.4	Random Forest	8
1.5	Casuality and Machine Learning?	8
1.6	Exercises	10
2	Recap: Statistics	11
2.1	Sample Variation	11
2.2	Hypothesis Testing	16
3	Casual Inference	17
3.1	Fundamental Problem of Casual Inference	17
3.2	Statistical solution	18
3.2.1	Statistical solution: Observational Data	19
3.2.2	Statistical solution: Experiment	20
3.3	Summary	24
3.4	Exercises	25
4	Modern High Dimensional Linear Regression	27
4.1	Framework	27
4.2	Double Lasso Regression	29
4.3	Debiased Machine Learning (DML)	30
4.4	Exercises	33
5	Effect Heterogeneity and Statistical Power	33
5.1	Effect Heterogeneity and Statistical Power	33
5.2	Hypothesis Testing	34
5.3	Determinants of Power	34
5.4	Algebra of Determinants of Power	39
5.4.1	Individual Level Randomization	39
5.4.2	Group Level Randomization	40
5.5	Interaction or subgroup analysis	40
5.5.1	Effect Heterogeneity with Machine Learning	40
5.6	Exercises	40
6	Exploring Effect Heterogeneity with Machine Learning	41
6.1	Motivation	41
6.2	Casual Forest	41
6.3	Generic Machine Learning	42
6.3.1	Generic Machine Learning Inference	42
6.3.2	Example of Generic Machine Learning Inference	44
6.4	(Optimal) Policy Learning	46
6.4.1	Policy Learning	46
6.4.2	Example of Optimal Policy Learning	47
6.5	Summary	50
6.6	Exercises	50

7	Course Summary	51
7.1	Machine Learning	51
7.2	Statistical Inference	51
7.3	Modern High Dimensional Linear Regression	51
7.4	Exploring effect Heterogeneity with Machine Learning	52
7.5	Exercises	53

1 Introduction to Machine Learning

1.1 Motivation

- Econometrics, inferential statistics
 - $\bar{X} = 239$
 - Reject the the null-hypothesis that $\mu = 0$ and calculating the confidence intervals based on our estimations.
 - β is good estimator of relation between x and y
- However, what we often want is to predict certain event.
 - We want to predict that the letter is spam
 - What will person look next on Youtube
 - There will be a flu outbreak soon
 - Will be this teacher good based on his CV

1.1.1 Motivation: Spam Example

Dear Valentine,
you won this year.
Your money is waiting for you.
Just send us your bank account.
Best,
Valencia.
Dear|| Valentine||,
you|| won|| this|| year||.
Your|| money || is || waiting || for || you || .
Just|| send|| us|| your|| bank || account||.
Best||,
Valencia||.

Frequency of ||your||

$$f = 2/20 = 0.1$$

```
library(kernlab)

data(spam)

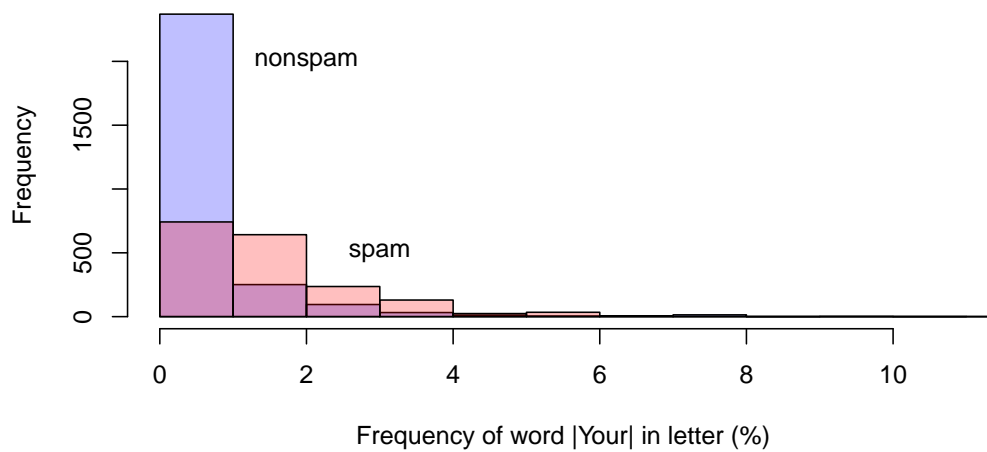
require(stargazer)

stargazer(head(spam[, c(1, 2, 21, 24, 58)]), summary = FALSE)
```

Table 1

	make	address	your	money	type
1	0	0.640	0.960	0	spam
2	0.210	0.280	1.590	0.430	spam
3	0.060	0	0.510	0.060	spam
4	0	0	0.310	0	spam
5	0	0	0.310	0	spam
6	0	0	0	0	spam

Histogram of word |your| based on type



```
prediction <- ifelse(spam[, "your"] > 1, "spam", "nonspam")
t <- table(prediction, spam[, "type"])/length(prediction)
require(xtable)
xtable(t * 100)
```

	nonspam	spam
nonspam	51.51	16.13
spam	9.08	23.28

Accuracy-correct predictions

$\approx 52+23=75\%$

1.2 In and Out of Sample Predictions

- In sample errors – the error rate on the same data set you used.
- Out of Sample error – the error rate you get on new data set.

- Key points:
 - We care about out of sample prediction
 - In sample error < out of sample error
 - In sample overfitting can occur
- Let's take a small sample from our data set.

```

set.seed(239)

require(dplyr)

s100 <- sample_n(spam, 100)

prediction <- ifelse(s100[, "your"] > 1, "spam", "nonspam")

t <- table(prediction, s100[, "type"])/length(prediction)
require(xtable)

xtable(t * 100)

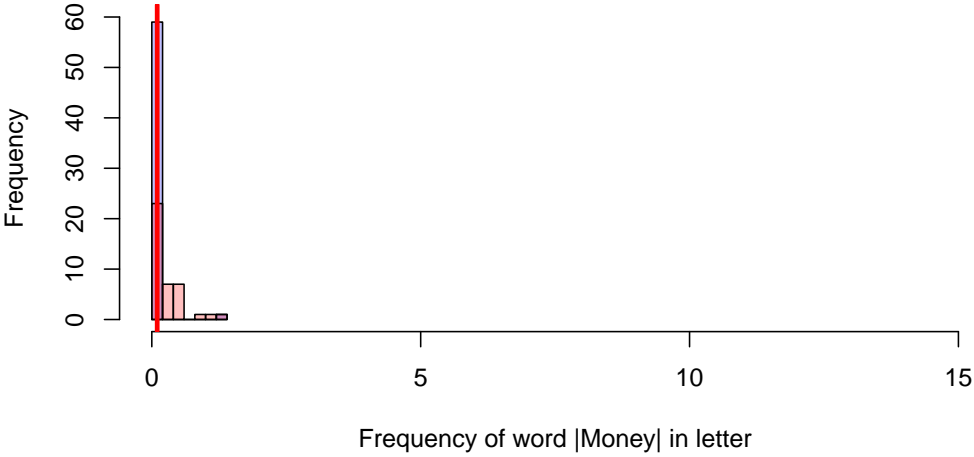
```

	nospam	spam
nospam	48.00	14.00
spam	12.00	26.00

Accuracy-correct predictions
 $\approx 48 + 26 = 74\%$

- What about other variables?

Histogram of word |Money| based on type



```
prediction<-ifelse(s100[,"your"]>1,"spam",
                  ifelse(s100[,"money"]>0,
                          "spam","nonspam"))

t<-table(prediction, s100[,"type"])/length(prediction)
require(xtable)

xtable(t*100)
```

	nonspam	spam
nonspam	48.00	9.00
spam	12.00	31.00

Accuracy-correct predictions

$\approx 48 + 31 = 79\%$

- But what about out applying this rule to the whole dataset?

```
prediction<-ifelse(spam[,"your"]>1,"spam",
                  ifelse(spam[,"money"]>0.2, "spam","nonspam"))

t<-table(prediction, spam[,"type"])/length(prediction)
require(xtable)

xtable(t*100)
```

	nonspam	spam
nonspam	50.88	13.98
spam	9.72	25.43

Accuracy-correct predictions

$\approx 51 + 25 = 76\%$

1.3 Prediction Study Design

- Define error rate e.g. Accuracy
- Split data into:
 - Training, Test, Validation data set
- Pick features (variables) and model on training set
 - Use Cross-validation
- Apply to test data set (once)
- Apply on validation data set
- Sensitivity – $TP / (TP+FN)$

Table 2: Error Types

		Predicted	
		Spam	Not spam
Actual	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- Specificity – $TN / (FP+TN)$
- Positive Predictive Value – $TP / (TP+FP)$
- Negative Predictive Value – $TN / (FN+TN)$
- Accuracy – $(TP+TN) / (TP+FN+TN+FP)$
- We estimate the model on the training set e.g. random 60% of sample
- We test the model on test data set and refine it e.g. calculate sensitivity of the model
- We apply model model on validation data set.

Can we do this multiple times?

1. Use training set
 2. Split it again into training/test set
 3. Build a model on training set
 4. Evaluate on test set
 5. Repeat from 1 to 4 n times.
 6. Estimate average errors e.g. Mean Squared Error = $\frac{\sum (y_i - \hat{y}_i)^2}{n}$
- It allows:
 - Pick up features(variables)
 - Select model
 - Parameters of Prediction function

How to split training/test set?

- Random sampling without replacement.
 - Alternative: Random sampling without replacement, bootstrap. (needs correction)
- K -fold cross validation.
 - Divide data set in K parts. (Needs to carefully choose K)

How to build a model and select variables?

- OLS
- Lasso Regression
- Trees
- Random Forest
- Neural networks
- ...

1.4 Random Forest

- Recall Spam example
 - We can base our prediction on word `||your||`
 - Then we can use word `||money||`
 - ...
 - We build a decision tree till the groups are very small

```
prediction <- ifelse(s100[, "your"] > 1, "spam", "nonspam")
```

```
prediction<-ifelse(spam[,"your"]>1,"spam",  
  
ifelse(spam[,"money"]>0.2, "spam","nonspam"))
```

- We find the variable that best separate our outcome
- Divide the data on into two groups based on this variable
- Within each split, find the variable that separates outcomes the best
- Continue until the groups are too small

Why not to do this multiple times?

- Bootstrap sample
- At each split randomly select n variables
- Grow multiple trees and vote
- Pros: Extremely high accuracy
- Cons: Computationally intensive, interpretability, overfit

1.5 Casuality and Machine Learning?

- ML focus on prediction out of “black-box”
- → Causality is not necessary there!
 - Prediction of crime recidivism: Ethnicity or Income? see “Weapon of math destruction” by Cathy O’Neil
 - Google flu prediction algorithm: Sensitivity to words, time variance. See ”The Parable of Google Flu ...”

I run out of examples. Let’s ask ChatGPT

Bring examples when machine learning fails to establish casual relationships with ciation, please

Certainly, here are examples illustrating when machine learning fails to establish causal relationships, along with citations:

- 1. Ice Cream Sales and Drowning Incidents:**
 - Example: Machine learning might reveal a correlation between the increase in ice cream sales and the number of drowning incidents. However, it cannot establish a causal relationship, as both events are influenced by the temperature during summertime.
 - Citation: Phillips, N. (1985). The relationship between drowning and the sale of ice cream and iced drinks in New Zealand. *The Medical Journal of Australia*, 143(11-12), 610-612.
- 2. Crime Rates and Police Car Numbers:**
 - Example: Machine learning might find a correlation between the number of police cars patrolling an area and crime rates. However, it cannot confirm that adding more police cars directly reduces crime; other factors, such as community engagement, may play a role.
 - Citation: Klick, J., & Tabarrok, A. (2005). Using terror alert levels to estimate the effect of police on crime. *Journal of Law and Economics*, 48(1), 267-279.
- 3. Obesity and Diet Soda Consumption:**
 - Example: Machine learning could reveal a correlation between diet soda consumption and obesity. However, it cannot establish causality, as people who are already obese might consume diet soda to manage their weight.
 - Citation: Fowler, S. P., Williams, K., Resendez, R. G., Hunt, K. J., Hazuda, H. P., & Stern, M. P. (2008). Fueling the obesity epidemic? Artificially sweetened beverage use and long-term weight gain. *Obesity*, 16(8), 1894-1900.
- 4. Educational Attainment and Income:**
 - Example: Machine learning might discover a strong correlation between educational attainment and higher income. However, it cannot determine if more education directly causes higher income or if factors like networking and job opportunities are equally significant.
 - Citation: Goldin, C., & Katz, L. F. (2008). *The Race between Education and Technology*. Harvard University Press.
- 5. Social Media Usage and Loneliness:**
 - Example: Machine learning might uncover a connection between heavy social media use and feelings of loneliness. However, it cannot establish whether social media use causes loneliness, or if lonely individuals are more drawn to social media for connection.
 - Citation: Primack, B. A., Shensa, A., Sidani, J. E., Whaite, E. O., Lin, L. Y., Rosen, D., ... & Miller, E. (2017). Social media use and perceived social isolation among young adults in the U.S. *PLoS ONE*, 12(8), e0182145.

Regenerate

- We intuitively feel that
 - Machine learning predictions in the cases above are kind of funny
 - Machine learning predictions may have dangerous (policy) implications e.g. less cars to the police?
- To understand how to avoid this, we will have to understand causal relationship.

1.6 Exercises

1. **In and Out of Sample Predictions I:** What is in sample errors?
2. **In and Out of Sample Predictions II:** What is out of Sample error?
3. **Prediction study design I:** Describe simple prediction study design
4. **Prediction study design II:** Describe Cross-Validation
5. **Install R and Rstudio.**

Download R from <https://www.r-project.org/> and install it.

Download RStudio <https://www.rstudio.com/> and install it.

6. **Simple linear model: Crime rate.**

```
# Install and call the package 'Ecdat'.
install.packages("Ecdat")
require(Ecdat)

# Use the data 'Crime'
data(Crime)
`?`(Crime)

# Plot the crime rate with respect to number of policemen
# per capita.
plot(crmrte ~ polpc, ylab = "Crime Rate", xlab = "Number of Policemen per capita",
     pch = 18, data = Crime)

# Estimate simple linear relation between crime rate and
# number of policemen per capita.
lmcrime <- lm(crmrte ~ polpc, data = Crime)
summary(lmcrime)

# Plot the crime rate with respect to number of policemen
# per capita.
plot(crmrte ~ polpc, ylab = "Crime Rate", xlab = "Number of Policemen per capita",
     pch = 18, data = Crime)
abline(lmcrime, col = "red")
```

How do you interpret this result? What are the policy implications?

7. **Let's use Machine Learning to predict Crime rate**

```
require(Ecdat)

# Install and call the package 'randomForest' and
# 'datasets'
install.packages("randomForest")
install.packages("datasets")
library(randomForest)
library(datasets)

set.seed(239) # set seed to reproduce results

ind <- sample(2, nrow(Crime), replace = TRUE, prob = c(0.5, 0.5))
train <- Crime[ind == 1, ] # training data set
test <- Crime[ind == 2, ] # testing data set
nrow(train)
nrow(test)

rf1cr <- randomForest(crmrte ~ ., ntree = 50, data = train, importance = TRUE)
rf1cr
```

```

varImpPlot(rf1cr, main = "Variable Importance")

lm1 <- lm(crmrte ~ density + polpc + smsa + prbarr + pctmin +
         wfed, data = train)

summary(lm1)

# Out of Sample

predictedcrmrate <- predict(lm1, test)

compare <- cbind(test, predictedcrmrate)

plot(compare$crmrate, compare$predictedcrmrate, xlab = "Actual Crime Rate",
      ylab = "Predicated Crime Rate")
hist(compare$crmrate - compare$predictedcrmrate)
sum((compare$crmrate - compare$predictedcrmrate)^2)/length(compare$crmrate)

```

8. Your Turn

- Call the package `Ecdat` and use the data `DoctorAUS`. What is the data set about?
- What is the influence of number of doctor visits (`doctorco`) on the number of illness in past 2 weeks (`illness`)? Visualize this relation.
- Estimate a simple linear relation between the number of doctor visits and the number of illness in past 2 weeks. Visualize regressions output.
- Make training and testing dataset based on `DoctorAUS`.
- Use Random Forest `randomForest` from package `randomForest` to build regression predicting the number of doctor visits on your training set.
- “Test” your model on the testing set. Plot Actual number of doctor visits against predicted one.
- Calculate mean square error (MSE) for your testing set.

2 Recap: Statistics

2.1 Sample Variation

- Population – the set of all entities which could theoretically be observed.
- Sample – a part of the population that we observe.
- Typically we observe A but want to know something about B .

Example

We know working hours Y of person x in year t , but we want to know it for the same (other) person next year (time) $t + 1$.

- We can do it, use A to *infer* something about B , if we assume that they come from the same population.
- We can look at the first observation and make conclusion.
- Perhaps, it is not best idea.
- It is better to draw a random sample from the same population.
 - ⇒ Values are identically and independently distributed .i.i.d.

Law of Large Numbers

If Y_1, \dots, Y_n i.i.d. and $0 < \sigma_Y^2 < \infty$ and , then \bar{Y} is a consistent estimator of Y .

Central Limit Theorem

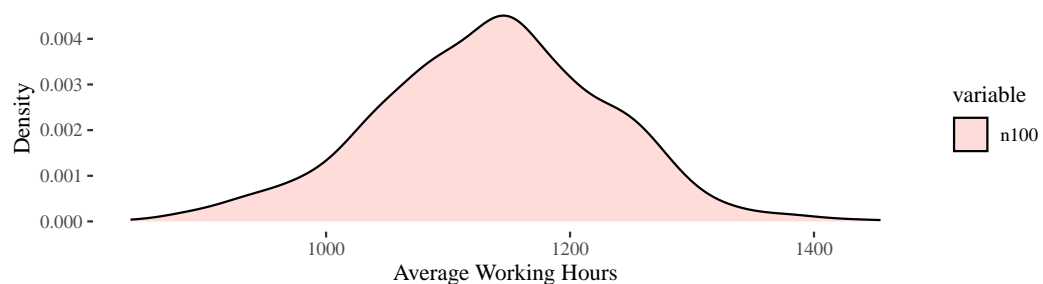
If Y_1, \dots, Y_n i.i.d. and $0 < \sigma_Y^2 < \infty$ and n is large, then the distribution of \bar{Y} approximates a normal distribution

```
library(Ecdat)
data(Workinghours)
# Make sample of 'hours'
s1 <- sample(Workinghours$hours, 100, replace = TRUE)
str(s1)

## int [1:100] 735 0 0 1312 1928 1348 2783 920 911 0 ...

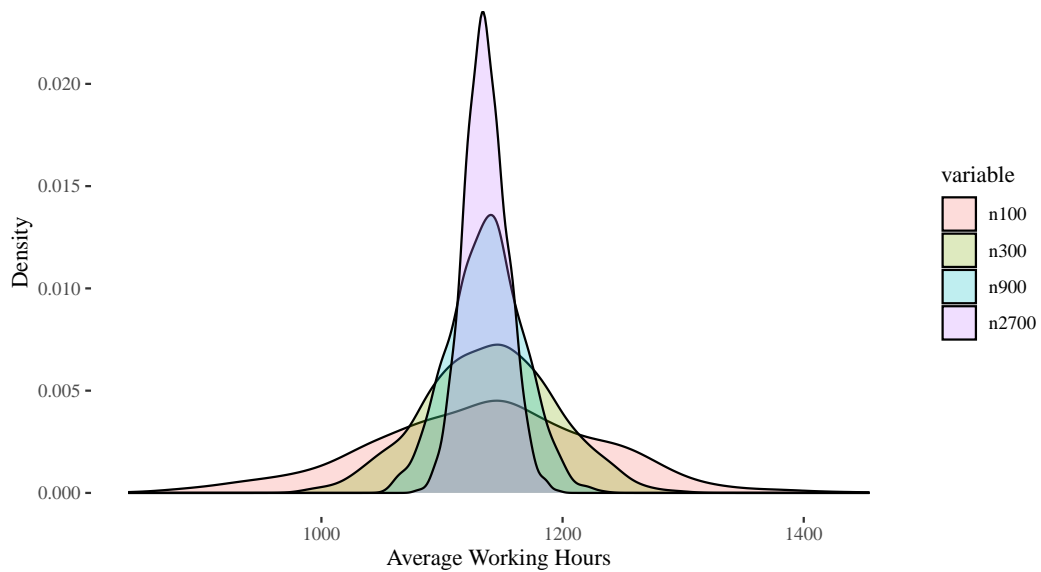
# Make 1000 samples of 'hours' of size 100
n100 <- replicate(1000, mean(sample(Workinghours$hours, 100,
  replace = TRUE)))
```

```
require(ggplot2)
require(ggthemes)
require(reshape2)
ggplot(melt(data.frame(n100)), aes(x = value, fill = variable)) +
  geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Working Hours") +
  ylab("Density")
```



```
n300 <- replicate(1000, mean(sample(Workinghours$hours, 300,
  replace = TRUE)))
n900 <- replicate(1000, mean(sample(Workinghours$hours, 900,
  replace = TRUE)))
n2700 <- replicate(1000, mean(sample(Workinghours$hours, 2700,
  replace = TRUE)))
```

```
ggplot(melt(data.frame(n100, n300, n900, n2700)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Working Hours") + ylab("Density")
```



Standard Deviation

$$sd = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

```
s100 <- sample(Workinghours$hours, 100, replace = TRUE)
s300 <- sample(Workinghours$hours, 300, replace = TRUE)
s900 <- sample(Workinghours$hours, 900, replace = TRUE)
s2700 <- sample(Workinghours$hours, 2700, replace = TRUE)
sqrt(sum((s100 - mean(s100))^2)/99)

## [1] 841.4578

sd(s100)

## [1] 841.4578

sd(s300)

## [1] 888.9162

sd(s900)

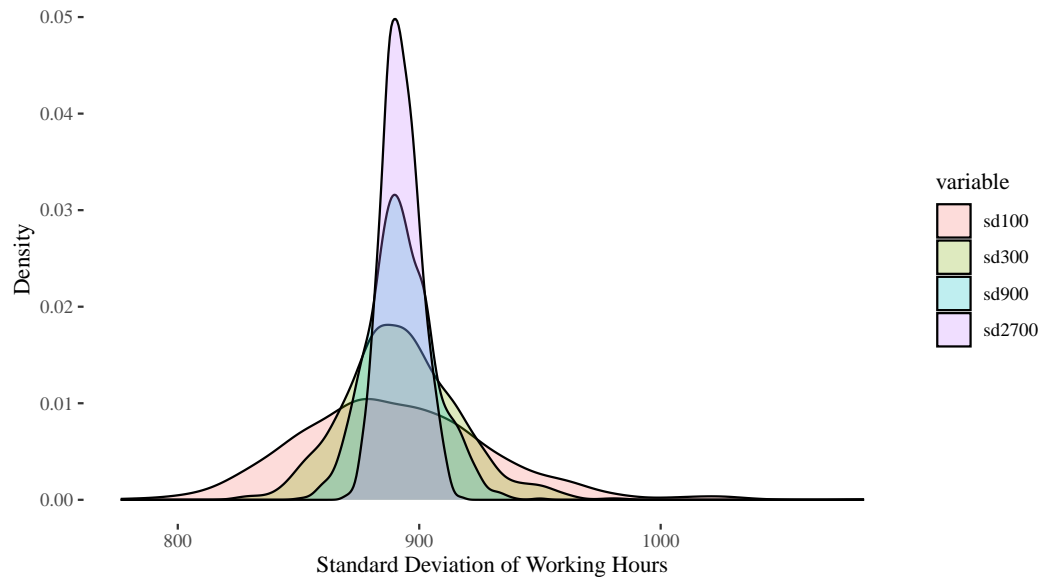
## [1] 874.7947

sd(s2700)

## [1] 896.1815
```

```
sd100 <- replicate(1000, sd(sample(Workinghours$hours, 100, replace = TRUE)))
sd300 <- replicate(1000, sd(sample(Workinghours$hours, 300, replace = TRUE)))
sd900 <- replicate(1000, sd(sample(Workinghours$hours, 900, replace = TRUE)))
sd2700 <- replicate(1000, sd(sample(Workinghours$hours, 2700,
  replace = TRUE)))
```

```
ggplot(melt(data.frame(sd100, sd300, sd900, sd2700)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Standard Deviation of Working Hours") + ylab("Density")
```



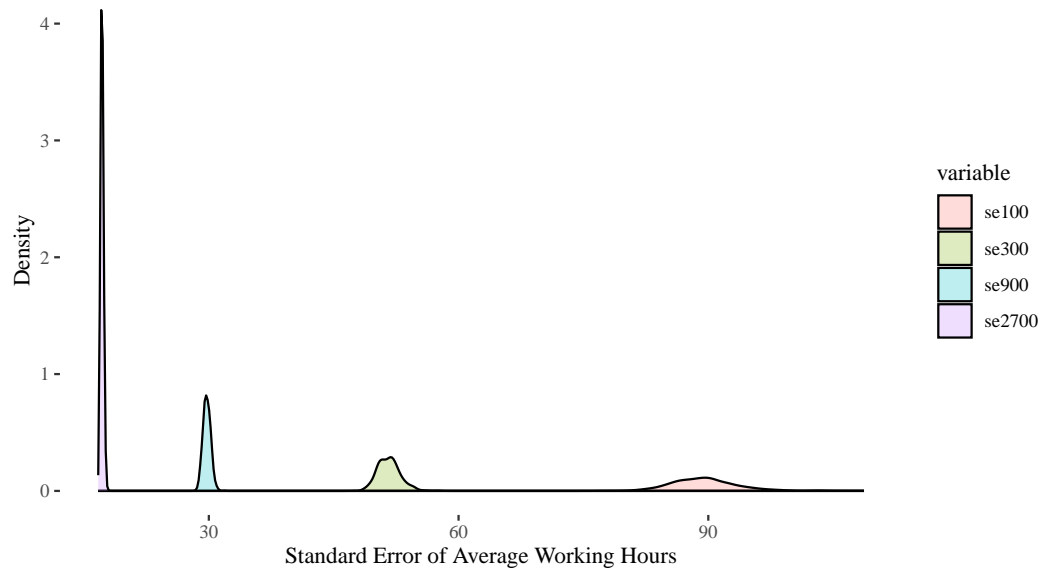
Standard Error of the Mean

$$se = \frac{sd}{\sqrt{n-1}}$$

```
sd(s100)/sqrt(length(s100) - 1)
## [1] 84.56969
sd(s300)/sqrt(length(s300) - 1)
## [1] 51.40735
sd(s900)/sqrt(length(s900) - 1)
## [1] 29.17604
sd(s2700)/sqrt(length(s2700) - 1)
## [1] 17.25022
```

```
se100 <- replicate(1000, sd(sample(Workinghours$hours, 100, replace = TRUE))/sqrt(99))
se300 <- replicate(1000, sd(sample(Workinghours$hours, 300, replace = TRUE))/sqrt(299))
se900 <- replicate(1000, sd(sample(Workinghours$hours, 900, replace = TRUE))/sqrt(899))
se2700 <- replicate(1000, sd(sample(Workinghours$hours, 2700,
  replace = TRUE))/sqrt(2699))
```

```
ggplot(melt(data.frame(se100, se300, se900, se2700)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Standard Error of Average Working Hours") + ylab("Density")
```

Central Limit Theorem

If Y_1, \dots, Y_n i.i.d. and $0 < \sigma_Y^2 < \infty$ and n is large, then the distribution of \bar{Y} approximates a normal distribution

Confidence intervals

Upper 95% limit = $\bar{x} + 1.96 \cdot se$

Lower 95% limit = $\bar{x} - 1.96 \cdot se$

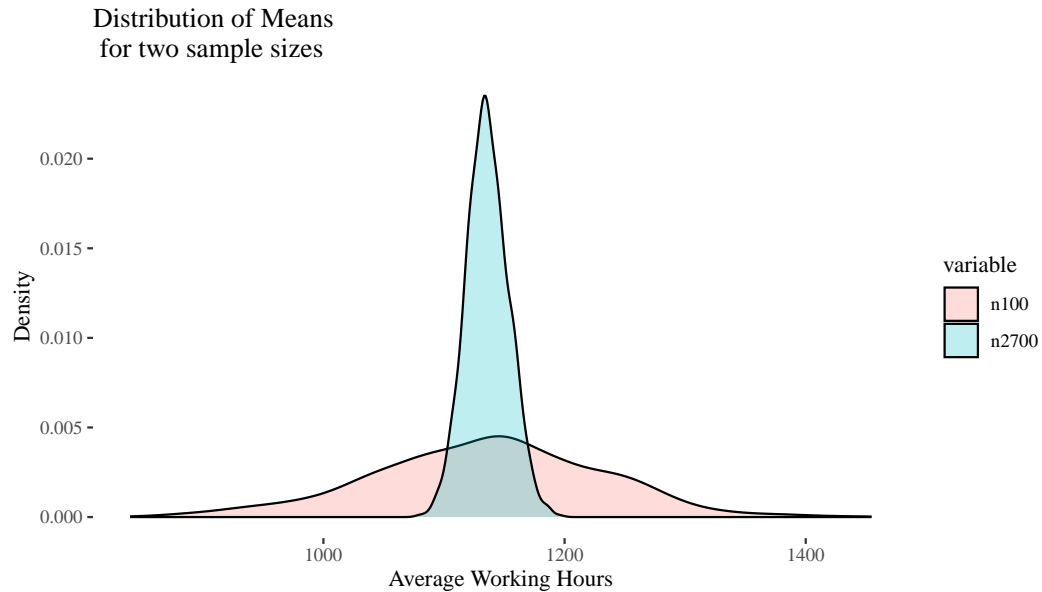
```

mean(s100) - 1.96 * sd(s100)/sqrt(length(s100) - 1)
## [1] 1251.433
mean(s100) + 1.96 * sd(s100)/sqrt(length(s100) - 1)
## [1] 1582.947
mean(s2700) - 1.96 * sd(s2700)/sqrt(length(s2700) - 1)
## [1] 1084.962
mean(s2700) + 1.96 * sd(s2700)/sqrt(length(s2700) - 1)
## [1] 1152.583

```

Table 3: Error Types

		Null Hypothesis, H_0	
		Valid	Invalid
Judgement of H_0	Reject	False Positive (Type Error I)	Correct Inference
	Fail to Reject	Correct Inference	False Negative (Type Error II)



Note: The exact confidence interval is slightly different since we have to correct for the fact that we sample from larger sample.

2.2 Hypothesis Testing

1. Formulate Null-Hypothesis e.g. means is zero. $H_0 : E(Y) = \mu_{Y,0}$
2. Formulate alternative hypothesis:
 - $H_1 : E(Y) \neq \mu_{Y,0}$. (Two-sided test)
 - $H_1 : E(Y) > \mu_{Y,0}$. (One-sided test)
 - $H_1 : E(Y) < \mu_{Y,0}$. (One-sided test)

What can happen then?

- α , significance level – predefined probability of rejecting H_0 despite it being true, typically 5%.
- k – predefined probability of failing to reject H_0 when it is false.
- $(1-k)$, power – predefined probability of rejecting H_0 when it is false, typically 80%

P-Value

- **Level of significance** – Predefined probability of rejecting the null hypothesis, despite it being true.
- **P-value** – Probability of drawing a sample that is at least as averse to the null hypothesis as our data given that the null hypothesis is true.

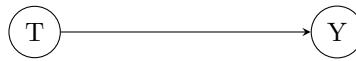


Figure 1: The casual effect of treatment on outcomes.

Example

- H_0 : Coin is fair, $q_T = q_H = 0.5$
- H_1 : Heads are more likely, $q_H > 0.5$
- We observe 3 heads out of 4 throws.
- $p = \frac{4!}{3! \cdot 1!} \cdot (0.5)^3 \cdot 0.5 = 0.25$
- and more averse case $p = \frac{4!}{4! \cdot 0!} \cdot (0.5)^4 = 0.0625$
- p -value= 0.3125

```

binom.test(3, 4, p = 0.5, alternative = c("greater"))

##
## Exact binomial test
##
## data: 3 and 4
## number of successes = 3, number of trials = 4, p-value = 0.3125
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.2486046 1.0000000
## sample estimates:
## probability of success
##                0.75

```

3 Casual Inference

3.1 Fundamental Problem of Casual Inference

Road Not Taken

“Two roads diverged in a yellow wood,
 And sorry **I could not travel both**
 And be one traveler, long I stood
 And looked down one as far as I could
 To where it bent in the undergrowth;
 ...
 Two roads diverged in a wood, and
 I took the one less traveled by,
 And that has made all the difference.”

Robert Frost, 1920

- T - treatment e.g. educational program
- Y - outcome e.g. reading level
- $T_i = 1$ – if person i is under treatment
- $T_i = 0$ – if person i is not under treatment

- $Y_i(1)$ – (potential) outcome if person i is under treatment
- $Y_i(0)$ – (potential) outcome if person i is not under treatment

$$Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i) \quad (1)$$

$$Y_i = Y_i(0) + \underbrace{(Y_i(1) - Y_i(0))}_{\text{Casual Effect}} T_i \quad (2)$$

Holland, (1986) Statistics and Causal Inference

Casual effect of $T_i = 1$ on unit i (relative to $T_i = 0$):

$$Y_i(1) - Y_i(0)$$

The problem: It is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ on the same unit, therefore, it is impossible to observe the effect of $T_i = 1$ on unit i .

Any solutions?

- “Scientific solution” - requires the unbelievably strong assumption of homogeneity of units, e.g., **identical** twins that are put under treatment and no treatment condition, and the **only** difference in outcomes between them is induced by treatment.
- Statistical solution – we can hope to estimate **averages** and the distribution of $Y(T)$ at the population level to compute **Average Treatment Effect (ATE)**:

$$\Delta = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \quad (3)$$

3.2 Statistical solution

Knowing Average Treatment Effect (ATE) can be extremely useful as we can learn that treatment, on average, improved outcomes:

- Visits to doctors reduce the number of sick people in the population (unfortunately, not necessarily all will recover, but still good for society).
- Educational programs improve the student’s reading ability (not necessarily all will read equally, but still good for society).
- Training increased income on average (not necessary for all, but still welfare might improve).

Now, we can think about how to calculate the average treatment effect:

1. Observational data. We can use observational data on the population or use samples to infer the average treatment effect.
2. Experimental data. We can treat randomly selected people and compare them to non-treated to infer the average treatment effect.

3.2.1 Statistical solution: Observational Data

- Let's think about calculating the average treatment effect if we have the population data at hand or sample.
- For instance, we know if the person is sick and if this patient is treated by a doctor in the population.
- We can naïvely calculate the average of the outcome conditional on the **observed** "treatment" status:

$$E[Y_i|T = 1] = E[Y_i(1)|T = 1] \quad (4)$$

$$E[Y_i|T = 0] = E[Y_i(0)|T = 0] \quad (5)$$

- and subtract 4 from 5 to get **Average Predictive Effect (APE)**:

$$\pi = E[Y_i|T = 1] - E[Y_i|T = 0] = E[Y_i(1)|T = 1] - E[Y_i(0)|T = 0] \quad (6)$$

```
library(Ecdat)

data(DoctorAUS)
attach(DoctorAUS)

T <- ifelse(doctorco > 0, 1, 0) # 1 if treated by a doctor or specialist in the past 2 weeks
Y <- illness #number of illness in past 2 weeks

mean(Y[T == 1]) # average number of illnesses conditional on observed treatment by a doctor or
[1] 2.125834

mean(Y[T == 0]) # average number of illnesses conditional on observed no treatment by the doctor
[1] 1.256218

mean(Y[T == 1]) - mean(Y[T == 0]) #The average predictive effect of doctor visits
[1] 0.8696158
```

→ The average predictive effect is positive ($\pi \approx 0.87 > 0$): Doctor or specialist visit is associated with more illness.

```
library(Ecdat)
data("VietNamI")
attach(VietNamI)

T <- ifelse(educ > 0, 1, 0) # 1 if a person got at least two years of higher education
Y <- illness #number of illnesses experienced in past 12 months

mean(Y[T == 1])
[1] 0.6150017
```

```
mean(Y[T == 0])
```

```
[1] 0.7094727
```

```
mean(Y[T == 1]) - mean(Y[T == 0])
```

```
[1] -0.09447091
```

→ The average predictive effect is negative ($\pi \approx -0.09 < 0$): Higher education is associated with less illness.

- Why do we observe it?
- We calculate this:

$$\pi = E[Y_i(1)|T = 1] - E[Y_i(0)|T = 0] \quad (7)$$

- Yet,

$$E[Y_i(1)|\mathbf{T}=\mathbf{1}] \neq E[Y_i(1)] \quad (8)$$

$$E[Y_i(0)|\mathbf{T}=\mathbf{0}] \neq E[Y_i(0)] \quad (9)$$

3.2.2 Statistical solution: Experiment

- Now, I can decide to assign people to the treatment and control groups randomly.
- Random assignment guarantees that treatment T is independent from potential outcome Y . $T \perp\!\!\!\perp Y$
- $T \perp\!\!\!\perp Y \Rightarrow E[Y_i(1)|\mathbf{T}=\mathbf{1}] = E[Y_i(1)] \Rightarrow$

$$\Delta = E[Y_i(1)] - E[Y_i(0)] = E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \quad (10)$$

$$\Delta = \pi \quad (11)$$

STAR Experiment: Does reducing class size improve educational outcomes?

Over 7,000 students in 79 schools were randomly assigned to one of three interventions:

- regular class (22 to 25 students per teacher)
- small class (13 to 17 students per teacher),
- regular-with-aide class (22 to 25 students with a full-time teacher's aide)

- For simplicity, we will focus on regular classes and small class intervention

STAR Experiment: Does reducing class size improve educational outcomes?

0 regular class (22 to 25 students per teacher)

1 small class (13 to 17 students per teacher),

Table 4: Control Group: Regular Class

Statistic	N	Mean	St. Dev.
Girls	4,085	0.479	0.500
Caucasian	4,075	0.645	0.479
African-American	4,075	0.351	0.477
Other Ethnicity	4,075	0.004	0.064

Table 5: Treatment Group: Small Class

Statistic	N	Mean	St. Dev.
Girls	1,915	0.500	0.500
Caucasian	1,913	0.685	0.465
African-American	1,913	0.308	0.462
Other Ethnicity	1,913	0.006	0.079

- Let's say we are interested in the reading abilities of students in grade 3
- Specifically, we want to know if reducing class improves reading.
- But before, let's look at how basic demographic characteristics of students are distributed across groups.

```
library(AER)
data(STAR)

STAR3 <- subset(STAR, star3 != c(" regular+aide") & !is.na(STAR$read3) &
  !is.na(STAR$star3))

T <- ifelse(STAR3$star3 == "small", 1, 0) # 1 if student is in the small class treatment group
Y <- STAR3$read3 #total reading scaled score in 3rd grade.

mean(Y[T == 1])
```

```
[1] 621.0757
```

```
mean(Y[T == 0])
```

```
[1] 612.7576
```

```
mean(Y[T == 1]) - mean(Y[T == 0])
```

```
[1] 8.318068
```

Now, we can assess significance of this difference with t-test:

```
t.test(Y ~ T)

##
## Welch Two Sample t-test
```

```
##
## data: Y by T
## t = -7.7229, df = 3621.6, p-value = 1.461e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
## 95 percent confidence interval:
## -10.429778 -6.206358
## sample estimates:
## mean in group 0 mean in group 1
## 612.7576 621.0757
```

Assumptions of Regression (OLS):

1. Strict exogeneity, $E(u_i|T_i = t) = 0$
2. (T_i, Y_i) are identically independently distributed (i.i.d.)
3. Large outliers are rare
4. $var(u|T = t)$ is constant
5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$

Assumptions of Regression (OLS):

1. Strict exogeneity, $E(u_i|T_i = t) = 0$ ✓
2. (T_i, Y_i) are identically independently distributed (i.i.d.)
3. Large outliers are rare
4. $var(u|T = t)$ is constant
5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$

✓ by randomization, n is large

Similarly with regression:

```
summary(lm(Y ~ T))

##
## Call:
## lm(formula = Y ~ T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.758  -27.758   -1.076    26.242   162.242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  612.7576     0.6004 1020.584 < 2e-16 ***
## T              8.3181     1.0628   7.827 5.87e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.37 on 5998 degrees of freedom
## Multiple R-squared:  0.01011, Adjusted R-squared:  0.009945
## F-statistic: 61.26 on 1 and 5998 DF, p-value: 5.869e-15
```


- Let's recall that we know the characteristics (W) of students e.g. gender. Can we include them?

Assumptions:

1. Strict exogeneity, $E(u_i|T_i = t) = 0$
2. (T_i, Y_i) are identically independently distributed (i.i.d.)
3. Large outliers are rare
4. $var(u|T = t)$ is constant
5. u is normally distributed, $u \sim \mathcal{N}(0, \sigma^2)$
6. No perfect multicollinearity ✓

✓ by randomization. Moreover, $T \perp\!\!\!\perp Y, W \Rightarrow$

$$E[Y_i(T)|W] = E[Y_i|T_i, W], T_i \in \{0, 1\} \Rightarrow$$

$$\Delta(W) = \pi(W)$$

Table 6: Outcomes of STAR project in Grade 3

	<i>Dependent variable:</i>			
	Y - Reading score			
	(1)	(2)	(3)	(4)
T - Small Class	8.318*** (1.063)	8.156*** (1.058)	7.982*** (1.062)	7.477*** (1.031)
Girl		7.567*** (0.987)	7.525*** (0.992)	7.598*** (0.962)
Teachers Experience (Years)			-0.013 (0.057)	0.099* (0.056)
Sub-urban School				20.294*** (1.434)
Rural School				24.702*** (1.287)
Urban				20.165*** (2.056)
Constant	612.758*** (0.600)	609.134*** (0.762)	609.500*** (1.113)	589.820*** (1.480)
Observations	6,000	6,000	5,951	5,951
R ²	0.010	0.020	0.019	0.078
Adjusted R ²	0.010	0.019	0.019	0.077

Note:

*p<0.1; **p<0.05; ***p<0.01

I. Shall we include these characteristics (W) – covariates?

- Let's note that

- Standard errors in model 4 are slightly lower than in model 1.
- R^2 or Adjusted R^2 is larger in the model 4 than in model 1.

- Covariates in randomized experiments is “denoising”, improving the precision of estimation of Average Treatment Effect.
 - Yet, which one to include and which not, especially if I have a lot more than people e.g. $W \gg i$?
- Here, Machine Learning tools will help us. We will study this in the first part of “Modern High Dimensional Linear Regression”

II. Treatment Heterogeneity.

- Let’s note that
 - Treatment (small class) increases the reading score by 8 points on average.
 - * Average treatment effect (ATE) ≈ 8 .
 - It is comparable to the gender gap in reading: Girls get 8 points more than boys.
 - * Average predictive effect (APE) ≈ 8 .
- Nice, could it be that the treatment helps to close the reading gender gap by being particularly helpful to boys?
- We can check it by estimating the interaction effect of treatment on boys (B):
- $Y = \beta_0 + \beta_T T + \beta_B B + \beta_{T \cdot B} T \cdot B + u_i$
- But could it be that while it is particularly useful for boys in rural areas
- $Y = \beta_0 + \beta_T T + \beta_B B + \beta_R R + \beta_{T \cdot B} T \cdot B + \beta_{R \cdot B} R \cdot B + \beta_{T \cdot B \cdot R} T \cdot B \cdot R + u_i$
- What about the income level of the parents, and the education of the mother, low achieving girls in grade 3?
- ...
- We can continue this exercise till infinity (or sample size).
- Increasing the number of hypotheses till infinity (or sample size) and shrinking our analysis to a subgroup of one person i .
- What can we do about it?
- We can explore heterogeneity in the systematic agnostic way with the help of machine learning.
- We will study this in the second part of “Modern High-Dimensional Linear Regression”

3.3 Summary

- Fundamental problem of Casual Inference: We can not observe the same person under treatment and not in the same time
- Statistical solution: Look at the groups of people and compare averages to calculate Average Treatment Effect (Δ).
 - Observational. Average Predictive Effect (π), yet, $\pi \neq \Delta$
 - Experimental. Average Treatment Effect (Δ) and $\pi = \Delta$
- We collect a lot of characteristics (W) about people (i) so that $W \gg i$
 - Here machine learning can help us to learn more.

3.4 Exercises

1. **Sample Variation I:** What is population?
2. **Sample Variation II:** What is sample?
3. **Sample Variation III:** What is the difference between standard deviation and standard error of the mean?
4. **Sample Variation IV:** Here are two tables of error types:

Table 7: Inferential Statistics: Error Types

		Null Hypothesis, H_0	
		Valid	Invalid
Judgement of H_0	Reject	False Positive	Correct Inference
	Fail to Reject	Correct Inference	False Negative

Table 8: Prediction Study: Error Types

		Predicted Value, \hat{Y}	
		Positive	Negative
Actual Value, Y	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- What is the key differences between them? Do they contradict each other?
- When is it useful to use one approach, and when the other?
- Give examples to use each way of thinking for economic problems.

5. Hypothesis Testing I:

- The average interview callback rate is 8% e.g., Bertrand, M. and Mullainathan, S., 2004.
- You send ten resumes and applications and get one callback.
- How likely are you to observe this outcome given that your resume is not worse than average?
- How likely are you to observe this outcome or worse, given that your resume is not worse than average?
- Calculate the p-value by formula and with simple `binom.test()`.
- Formula of binomial probability: $p = \frac{n!}{n!x!}(q)^x q^{n-x}$, where q is the probability of success on a single trial, x - number of success, n - number of trials.

6. Hypothesis Testing II:

- Call the package `AER` and use the data `ResumeNames`. What is it about?
- Calculate the average rate of callback.
- Calculate the rate of callback for applicants with afro-american sounding names. Use `subset()` function.
- Test with `binom.test()` that applicant with afro-american sounding names is less likely to receive the callback than the average applicant.
- Test with `binom.test()` that applicant with afro-american sounding names is less or more likely to receive the callback than the average applicant.

- Turn `call` variable to binary variable that takes 1 if "yes" with `ifelse()` function.
 - Use regression (`lm()`) to test if an afro-american applicant is equally likely to receive the call as a caucasian applicant.
1. **Causal Inference I:** What is the fundamental problem of causal inference?
 2. **Causal Inference II:** What are the two solutions proposed?
 3. **Causal Inference III:** What is an average predictive effect (APE)?
 4. **Causal Inference IV:** What is the average treatment effect (ATE)?
 5. **Causal Inference V:** When APE and ATE is equal?
6. **Statistical Solution I:**
 - Call the package `AER` and use the data `ResumeNames`.
 - Check the variables. For which variables you can assume that ATE is equal to APE in predicting callback?
 - Turn `call` variable to binary variable that takes 1 if "yes" with `ifelse()` function.
 - Use regression (`lm()`) to test how ethnicity, gender, and industry is associated with callback. Interpret your results.
 7. **Statistical Solution II: What predicts callback? or How to make a fictitious CV to get a callback?;**
 - Call the package `AER` and use the data `ResumeNames`.
 - Split the dataset in training and testing set.
 - Build a model on your training set with regression.
 - Use function `predict(...type="response")` to predict callback on your testing set.
 - Calculate your accuracy rate.
 - Calculate your mean square error.
 - Find a way to improve the prediction.
 - Hint 1: Use `ifelse()` function to turn probabilities into binary outcomes.
 - Hint 2: You can find your threshold level based on predictions.
 - Hint 3: You can re-run the code many times.
 - Hint 4: Consider using `randomForest` for variable selection once you setup the model. Re-run the code many times.

4 Modern High Dimensional Linear Regression

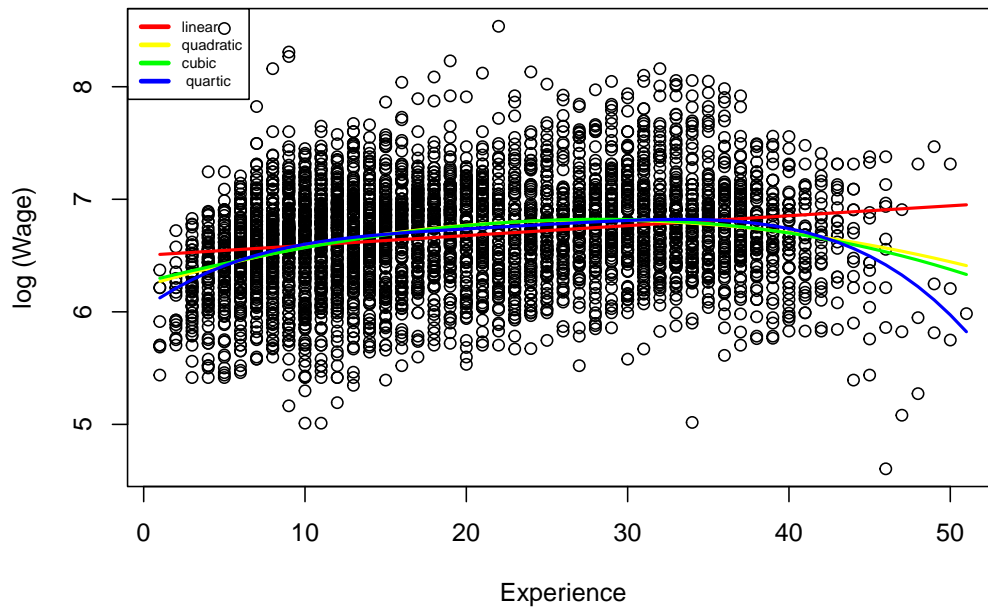
4.1 Framework

- Consider the following regression framework for outcome Y :

$$Y_i = \beta' X + u_i,$$

- Where X - is a vector of regressors x , $X = (x_j)_{j=1}^p$.
 - Data have large dimensional features (many variables)
 - We construct many technical regressors from raw regressors e.g. x_j^2 , $\log x_j$, $x_k \cdot x_m$ etc.
 - or both
- In other words: $p \gg i$
- X can be country characteristics, product characteristics, individual health information in electronic health records, participants of experimental study characteristics at the baseline etc.
- With non-linear regressions and more variables we have more flexibility:
 - We improve prediction
- For example, consider higher degree polynomial.

```
require(Ecdat)
data(Wages)
attach(Wages)
exp2 <- exp * exp
exp3 <- exp * exp * exp
exp4 <- exp * exp * exp * exp
or <- order(exp)
plot(lwage ~ exp, ylab = "log (Wage)", xlab = "Experience")
lines(exp[or], fitted(lm(lwage ~ exp))[or], col = "red", lwd = 2)
lines(exp[or], fitted(lm(lwage ~ exp + exp2))[or], col = "yellow",
      lwd = 2)
lines(exp[or], fitted(lm(lwage ~ exp + exp2 + exp3))[or], col = "green",
      lwd = 2)
lines(exp[or], fitted(lm(lwage ~ exp + exp2 + exp3 + exp4))[or],
      col = "blue", lwd = 2)
legend("topleft", c("linear", "quadratic", "cubic", "quartic"),
      lwd = 3, col = c("red", "yellow", "green", "blue"), cex = 0.6)
```



```

library(AER)

data("ResumeNames")

ResumeNames$callb <- ifelse(ResumeNames$call == "yes", 1, 0)

lm(callb ~ name + ethnicity, data = ResumeNames)

##
## Call:
## lm(formula = callb ~ name + ethnicity, data = ResumeNames)
##
## Coefficients:
## (Intercept)      nameAnne      nameCarrie      nameEmily      nameJill
## 0.094828      -0.012183      0.036125      -0.015532      -0.011084
## nameLaurie      nameKristen      nameMeredith      nameSarah      nameBrad
## 0.002608      0.036628      0.006777      0.003618      0.063903
## nameBrendan      nameGeoffrey      nameGreg      nameBrett      nameJay
## -0.017905      -0.027031      -0.016396      -0.027031      0.039501
## nameMatthew      nameNeil      nameTodd      nameAisha      nameEbony
## -0.005275      -0.029038      -0.036004      -0.072605      0.001326
## nameKeisha      nameKenya      nameLakisha      nameLatonya      nameLatoya
## -0.056576      -0.008093      -0.039828      -0.003523      -0.010757
## nameTamika      nameTanisha      nameDarnell      nameHakim      nameJamal
## -0.040140      -0.036857      -0.047209      -0.040282      -0.029254
## nameJermaine      nameKareem      nameLeroy      nameRasheed      nameTremayne
## 0.001326      -0.047953      -0.001078      -0.064977      -0.051349
## nameTyrone      ethnicityafam
## -0.041494      NA

```

```

library(AER)

data("ResumeNames")

ResumeNames$callb <- ifelse(ResumeNames$call == "yes", 1, 0)

lm(callb ~ ethnicity + name, data = ResumeNames)

##
## Call:
## lm(formula = callb ~ ethnicity + name, data = ResumeNames)
##
## Coefficients:
## (Intercept) ethnicityafam nameAnne nameCarrie nameEmily
## 0.094828 -0.041494 -0.012183 0.036125 -0.015532
## nameJill nameLaurie nameKristen nameMeredith nameSarah
## -0.011084 0.002608 0.036628 0.006777 0.003618
## nameBrad nameBrendan nameGeoffrey nameGreg nameBrett
## 0.063903 -0.017905 -0.027031 -0.016396 -0.027031
## nameJay nameMatthew nameNeil nameTodd nameAisha
## 0.039501 -0.005275 -0.029038 -0.036004 -0.031111
## nameEbony nameKeisha nameKenya nameLakisha nameLatonya
## 0.042821 -0.015082 0.033401 0.001667 0.037971
## nameLatoya nameTamika nameTanisha nameDarnell nameHakim
## 0.030737 0.001354 0.004638 -0.005714 0.001212
## nameJamal nameJermaine nameKareem nameLeroy nameRasheed
## 0.012240 0.042821 -0.006458 0.040417 -0.023483
## nameTremayne nameTyrone
## -0.009855 NA

```

- Classical linear regression or least squares fails in these high-dimensional settings.
- It overfits the data
- What to do?
- Lasso regression estimate β parameter:

$$\text{minimize } \sum_{i=1}^i (y_i - \sum_j x_{i,j} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (12)$$

- λ - penalty parameter
- Note we can find λ with cross-validation

4.2 Double Lasso Regression

- We know how to answer this question:
 - How does the predicted value of Y change if a regressor T increases by a unit, while other regressors W remain unchanged?
- We want to answer this question:
 - How does the predicted value of Y change if intervene and increase T by a unit, while other regressors W remain unchanged?

- Let's consider the effect of class size on test scores (performance of students on knowledge tests):

$$TestScore = \beta_0 + \beta_C ClassSize + \varepsilon$$

	Class Size (T)	Test Score (Y)
Percent of English learners	X	X
Time of day of the test		X
Number of parking lots	X	

How can we learn large that the variable is potentially a confounder?

Double Lasso

1. Run a lasso regression to select variables associated with Y .
 2. Run a lasso regression to select variables associated with T .
 3. Include the union of them in the regression.
- Double lasso Selection allows us to Neyman orthogonality.
 - Double lasso learns both strong and weak predictors of outcome Y .
 - These weak predictors of the outcome can also be strong predictors of variable of interest T . Therefore shall be included to reduce omitted variable bias.
 - Note that If we naively follow only one step lasso regression on Y or simply select the variable that highly correlates with Y , we would drop weak predictor of Y that can be strong predictors T

4.3 Debiased Machine Learning (DML)

The Frisch-Waugh-Lovell (FWL) Theorem allows to isolate the effect of a specific variable in a multiple regression analysis.

- Let's suppose we have next regression $\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ and we are interested in $\hat{\beta}_1$. Let's
 - Regress: $\hat{y}^* = \hat{\gamma}_1 X_2$ and $\hat{X}_1 = \hat{\gamma}_2 X_2$
 - Obtain residuals: $\tilde{X}_1 = X_1 - \hat{X}_1$ and $\tilde{y}_1 = y - \hat{y}^*$
 - Regress residuals: $\tilde{y} = \hat{\beta}_1 \tilde{X}_1$
 - We will get isolated coefficient of $\hat{\beta}_1$.

```
x1 = rnorm(100)
x2 = rnorm(100)
y1 = 1 + x1 - x2 + rnorm(100)
r1 = residuals(lm(y1 ~ x2))
r2 = residuals(lm(x1 ~ x2))

# ols
coef(lm(y1 ~ x1 + x2))
```

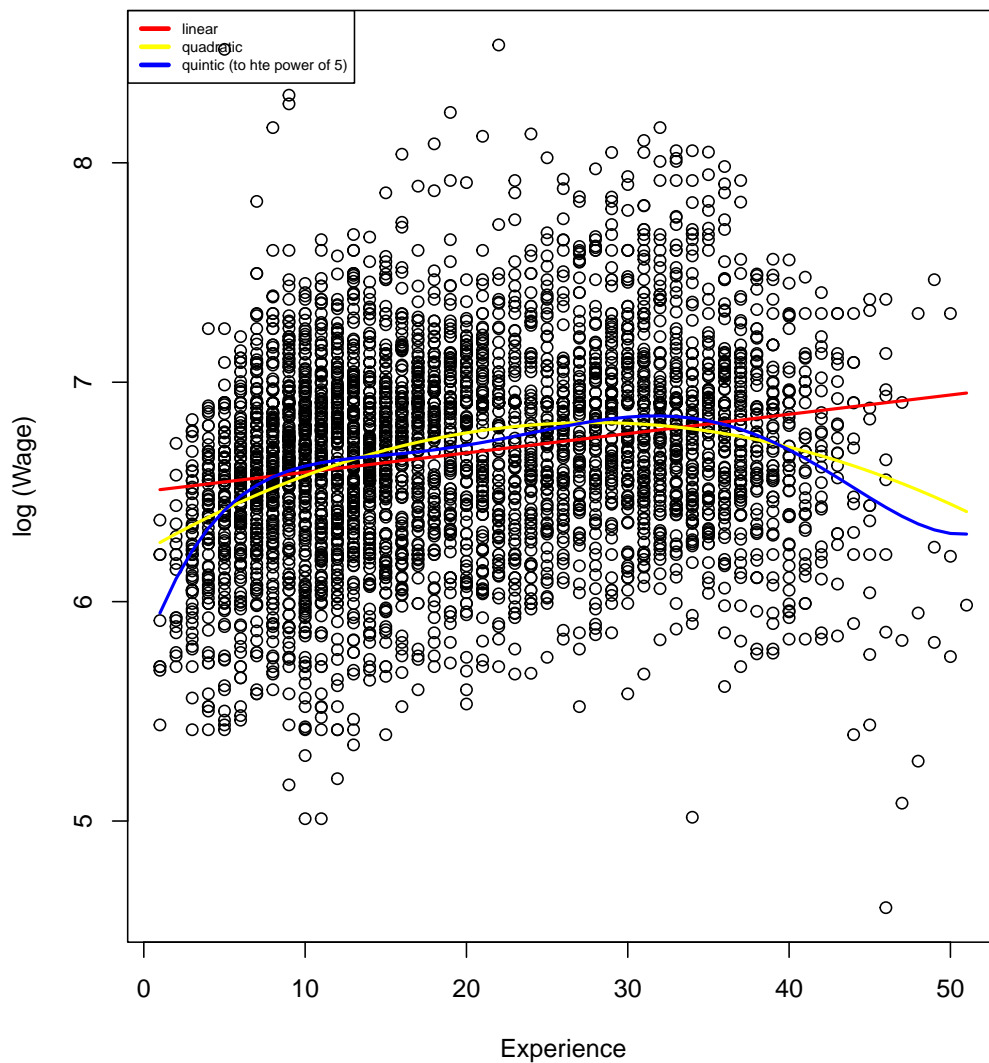


```
## (Intercept)      x1      x2
##  1.0062377  0.9290603 -1.0211978

# full ols
coef(lm(r1 ~ -1 + r2))

##      r2
## 0.9290603
```

- Now, recall that we might suspect non-linear relationships in our data
- But among a large set of variables we under risk of overfitting
- Yet, we want to model this **nuisance parameter** to properly estimate the relationship between our variable of interest and outcomes
- Moreover, we want to model it not only for one variable but as many as could



- Here machine learning helps.
- Machine learning allows:
 - Flexible modeling e.g. capturing complicated interactions etc.
 - "well" predicting outcomes
- Thus we can use this feature to estimate the relationship of interest with help of the Frisch-Waugh-Lovell Theorem
- Let's suppose we are interested in $\hat{Y} = \hat{\beta}_T T$ and we are interested in $\hat{\beta}_T$ having set of controls X . We can
 - Estimate the relationship between Y and X using Machine Learning regression model M_Y .
 - Estimate the relationship between T and X using Machine Learning regression model M_T .
 - Obtain residuals: $\tilde{T} = T - M_T(X)$ and $\tilde{y} = y - M_Y(X)$
 - Regress the residuals of the outcome on the residuals of the treatment: $\tilde{Y} = \alpha + \tau \tilde{T}$
 - where τ is an approximation of our casual effect ($\hat{\beta}_T$).
- Using Debiased machine learning (DML) we in simplified terms combine the next things
 - Flexible modeling with Machine Learning allows us to account for the nuisance parameter
 - Frisch-Waugh-Lovell Theorem allows to use of this flexible modelling
 - Double estimation reduces risk omitted variables and overfitting
- We want to estimate the relationship between (two) variables having flexibility in modeling but avoid overfitting and reduce the risk of omitted variable bias
- Modern regression techniques - Double lasso, debiased machine learning - allows us to do it in disciplined manner
- Key literature:
 - Belloni, A., Chernozhukov, V. and Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), pp.608-650.
 - Chernozhukov, V., Chetverikov, D., Demirer, M., Dufo, E., Hansen, C., Newey, W. and Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters.

4.4 Exercises

1. **Modern High-Dimensional Regression I:** Why do we have to consider Machine Learning in regression analysis when we want to estimate casual effect?
2. **Modern High-Dimensional Regression II:** What is lasso regression?
3. **Modern High-Dimensional Regression III:** Can we use lasso regression selecting only on the outcome to estimate the effect of variable interest on the outcome? Why?
1. **Double selection procedure I:** Explain post double (lasso) selection procedure.
2. **Double selection procedure II:** What is the advantage of a double selection procedure if we want to estimate the average treatment effect?
3. **Double selection procedure III:** Suppose you can not (data does not allow, do not know how etc.) use lasso. How can you use the knowledge about the double lasso procedure to estimate the average treatment effect?
1. **Double Debiased Machine Learning I:** Explain Frisch-Waugh-Lovell Theorem?
2. **Double Debiased Machine Learning II:** Explain double debiased machine learning procedure?
3. **Double Debiased Machine Learning II:**What are the advantages of ouble debiased machine learning procedure compared to post-double lasso?
1. **Summary:** Please, give examples where you can REALISTICALLY use post-double lasso procedure or double debiased machine learning procedure.
2. **Homework I:** Check for a paper in economics from a top journal that is interesting for you. For instance: American Economic Review, Econometric, Quarterly Journal of Economics, Journal of Monetary Economics, Review of Economic Studies
3. **Homework II:** Check if they have data next to the paper or allow to download data from the open repository
4. **Homework III:** Restimate (some) analysis from this paper using post-double lasso procedure and/or double debiased machine learning procedure.

5 Effect Heterogeneity and Statistical Power

5.1 Effect Heterogeneity and Statistical Power

- Consider that the effect of the treatment can be especially effective for some groups of people
- It is good to know as we can target the treatment to these people
- We can do subgroup analysis or use interaction to find out if the treatment T works differently for some group G e.g $Y = \beta_0 + \beta_T T + \beta_{GT} G * T + u$
- If we use this approach how does it affect our chances to find true difference?

→ Recap Statistical power

Table 9: Error Types

		Null Hypothesis, H_0	
		Valid	Invalid
Judgement of H_0	Reject	False Positive (Type Error I)	Correct Inference
	Fail to Reject	Correct Inference	False Negative (Type Error II)

5.2 Hypothesis Testing

1. Formulate Null-Hypothesis e.g. Difference in means is zero. $H_0 : E(Y) = \mu_{Y,0}$
2. Formulate alternative hypothesis:
 - $H_1 : E(Y) \neq \mu_{Y,0}$. (Two-sided test)
 - $H_1 : E(Y) > \mu_{Y,0}$. (One-sided test)
 - $H_1 : E(Y) < \mu_{Y,0}$. (One-sided test)

What can happen then?

- α , significance level – predefined probability of rejecting H_0 despite it being true, typically 5%.
- k – predefined probability of failing to reject H_0 when it is false.
- $(1-k)$, power – predefined probability of rejecting H_0 when it is false, typically 80%

5.3 Determinants of Power

Returning to the STAR experiment.

Let's estimate the distribution of the mean from the same population twice with different sample sizes:

```
require(AER)
data("STAR")

n100s1 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 ==
  "small" | star1 == "regular")$math1, 100, replace = TRUE)))
n100s2 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 ==
  "small" | star1 == "regular")$math1, 100, replace = TRUE)))

n300s1 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 ==
  "small" | star1 == "regular")$math1, 300, replace = TRUE)))
n300s2 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 ==
  "small" | star1 == "regular")$math1, 300, replace = TRUE)))

n900s1 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 ==
  "small" | star1 == "regular")$math1, 900, replace = TRUE)))
n900s2 <- replicate(1000, mean(sample(subset(na.omit(STAR), star1 ==
  "small" | star1 == "regular")$math1, 900, replace = TRUE)))
```

Now from different populations:

```
n100s1sr <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "small")$math1, 100, replace = TRUE)))
n100s2sr <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular")$math1, 100, replace = TRUE)))
```

```

n300s1sr <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "small")$math1, 300, replace = TRUE)))
n300s2sr <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular")$math1, 300, replace = TRUE)))

n900s1sr <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "small")$math1, 900, replace = TRUE)))
n900s2sr <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular")$math1, 900, replace = TRUE)))

```

```

plot100 <- ggplot(melt(data.frame(n100s1, n100s2)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density") + ylim(0, 0.32) +
  xlim(535, 560) + ggtitle("Same Population")
plot300 <- ggplot(melt(data.frame(n300s1, n300s2)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density") + ylim(0, 0.32) +
  xlim(535, 560)
plot900 <- ggplot(melt(data.frame(n900s1, n900s2)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density") + ylim(0, 0.32) +
  xlim(535, 560)

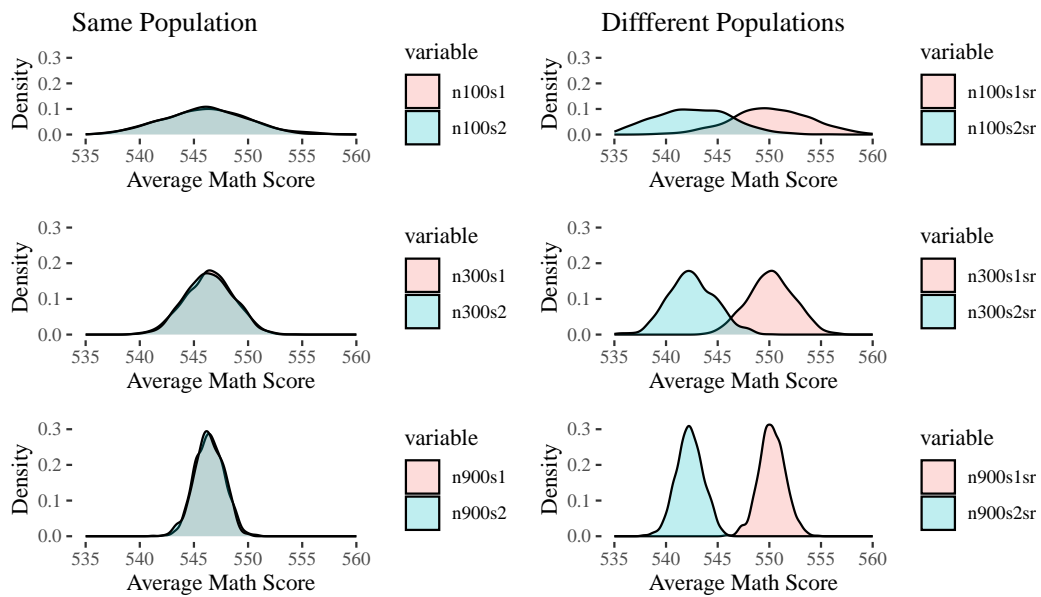
plot100sr <- ggplot(melt(data.frame(n100s1sr, n100s2sr)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density") + ylim(0, 0.32) +
  xlim(535, 560) + ggtitle("Different Populations")
plot300sr <- ggplot(melt(data.frame(n300s1sr, n300s2sr)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density") + ylim(0, 0.32) +
  xlim(535, 560)
plot900sr <- ggplot(melt(data.frame(n900s1sr, n900s2sr)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density") + ylim(0, 0.32) +
  xlim(535, 560)

```

```

library(gridExtra)
grid.arrange(plot100, plot100sr, plot300, plot300sr, plot900,
  plot900sr, ncol = 2, nrow = 3)

```

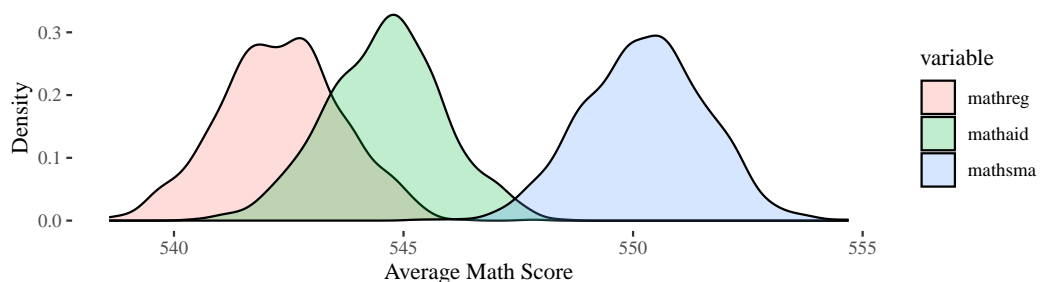


```
tapply(na.omit(STAR)$math1, na.omit(STAR)$star1, mean)
```

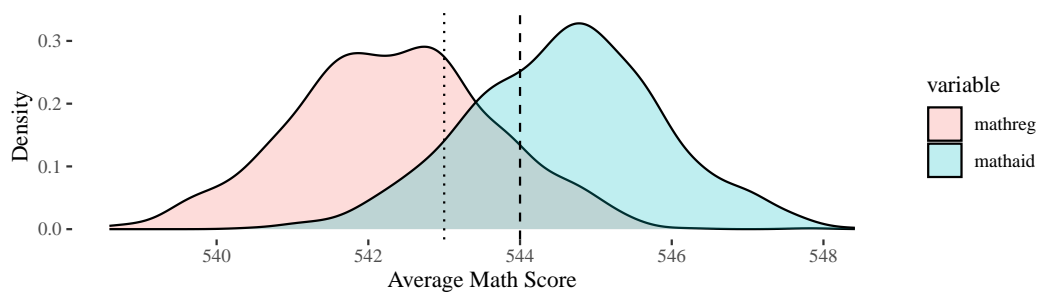
```
##      regular      small regular+aide
## 542.3272  550.2633  544.4835
```

```
mathreg <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular")$math1, 900, replace = TRUE)))
mathaid <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular+aide")$math1, 900, replace = TRUE)))
mathsma <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "small")$math1, 900, replace = TRUE)))
```

```
ggplot(melt(data.frame(mathreg, mathaid, mathsma)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density")
```



```
ggplot(melt(data.frame(mathreg, mathaid)), aes(x = value, fill = variable)) +
  geom_density(alpha = 0.25) + theme_tufte() + xlab("Average Math Score") +
  ylab("Density") + geom_vline(xintercept = 543, linetype = 3) +
  geom_vline(xintercept = 544, linetype = 2)
```



```

var(subset(na.omit(STAR))$math1)

## [1] 1555.032

var(subset(na.omit(STAR))$read1)

## [1] 2731.149

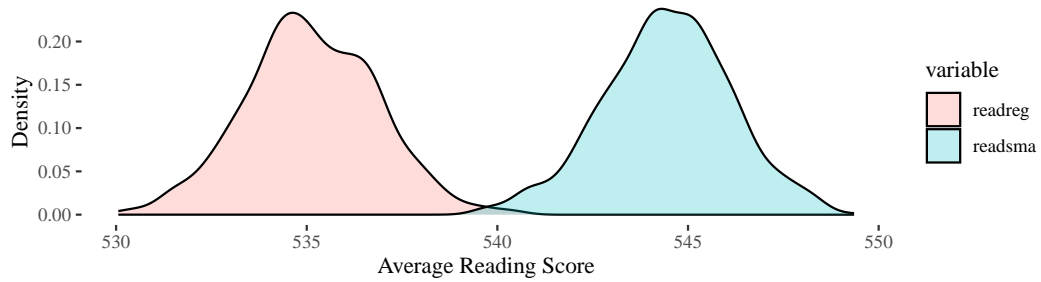
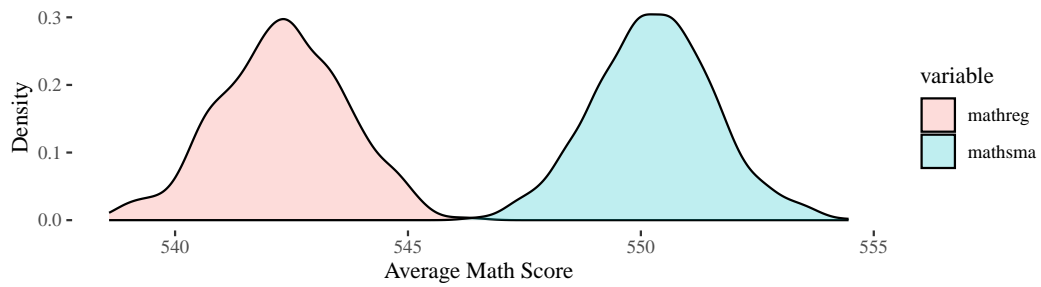
mathreg <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular")$math1, 900, replace = TRUE)))
mathsma <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "small")$math1, 900, replace = TRUE)))

readreg <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "regular")$read1, 900, replace = TRUE)))
readsma <- replicate(1000, mean(sample(subset(na.omit(STAR),
  star1 == "small")$read1, 900, replace = TRUE)))

plotmath <- ggplot(melt(data.frame(mathreg, mathsma)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Math Score") + ylab("Density")
plotread <- ggplot(melt(data.frame(readreg, readsma)), aes(x = value,
  fill = variable)) + geom_density(alpha = 0.25) + theme_tufte() +
  xlab("Average Reading Score") + ylab("Density")

grid.arrange(plotmath, plotread)

```

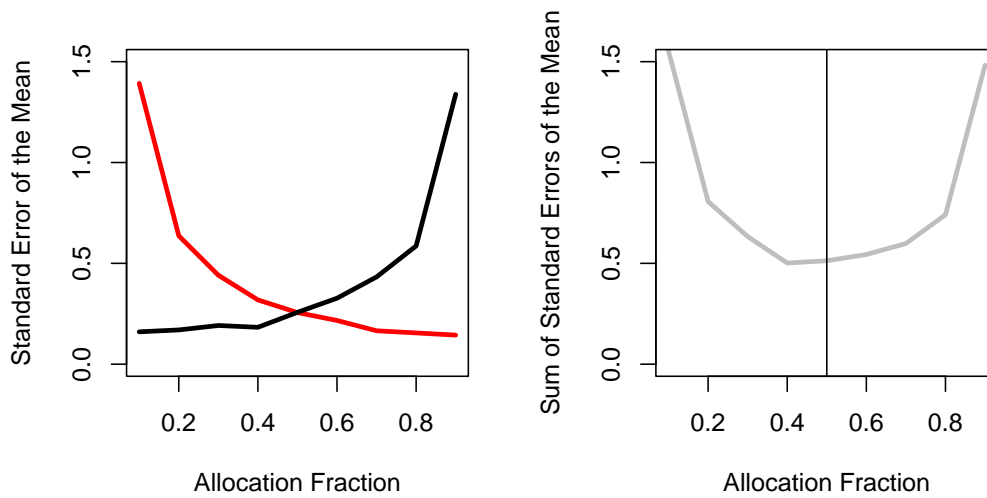


```
powerfraction <- function(P) {
  n <- 300
  sesma <- sd(replicate(100, mean(sample(subset(na.omit(STAR),
    star1 == "small")$math1, n * P, replace = TRUE))))/sqrt(n *
    P - 1)
  sereg <- sd(replicate(100, mean(sample(subset(na.omit(STAR),
    star1 == "regular")$math1, n * (1 - P), replace = TRUE))))/sqrt(n *
    (1 - P) - 1)
  as.data.frame(cbind(sesma, sereg, sesma + sereg))
}
powerfraction(0.4)

##      sesma      sereg      V3
## 1 0.3221651 0.2145788 0.5367439

dd <- sapply(c((1:9)/10), powerfraction)
```

```
par(mfrow = c(1, 2))
plot(c((1:9)/10), dd[1, ], type = "l", col = "red", lwd = 3,
  xlab = "Allocation Fraction", ylab = "Standard Error of the Mean",
  ylim = c(0, 1.5))
lines(c((1:9)/10), dd[2, ], lwd = 3)
plot(c((1:9)/10), dd[3, ], type = "l", lwd = 3, xlab = "Allocation Fraction",
  ylab = "Sum of Standard Errors of the Mean", col = "grey",
  ylim = c(0, 1.5))
abline(v = 0.5)
```

5.4 Algebra of Determinants of Power

5.4.1 Individual Level Randomization

$$Y = \beta_0 + \beta_T T + \sum_{i=1}^k \beta_{C_i} C_i + u$$

Variance

$$\text{Variance}(\hat{\beta}_T) = \frac{1}{P(1-P)} \frac{\sigma^2}{N}$$

, where N – number of observations; σ – standard error term; P – allocation fraction

Critical Value, $t_{\alpha/2}$

$$\Phi(t_{\alpha/2}) = 1 - \alpha/2$$

, where Φ is standard normal cumulative distribution function; α – significance level.

We can reject the Null-hypothesis if $\frac{|\hat{\beta}_T|}{SE(\hat{\beta}_T)} > t_{\alpha/2}$

Minimum Detectable Effect Size

$$MDE = (t_{(1-k)} + t_{\alpha/2}) \times \sqrt{\frac{1}{P(1-P)}} \times \sqrt{\frac{\sigma^2}{N}}$$

Power

$$1 - k = \Phi([MDE \times \sqrt{P(1-P)} \times \sqrt{\frac{N}{\sigma^2}}] - t_{\alpha/2})$$

So, we see:

- More observations(N) more power
- Higher MDE gives higher power
- Lower $t_{\alpha/2}$ or significance level gives higher power
- Smaller variance gives higher power
- Power is maximized if $P = 0.5$

5.4.2 Group Level Randomization

$$Y_{i,j} = \beta_0 + \beta_T T + \sum_{i=1}^k \beta_{C_i} C_i + v_j + w_{i,j}$$

, where v_j – shocks at cluster(group) level which we assume to be i.i.d. with variance τ^2 , $w_{i,j}$ – shocks at individual level within a cluster $w_{i,j}$, which we assume to be i.i.d. with variance σ^2

If $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$ – intercluster correlation \Rightarrow

Minimum Detectable Effect Size

$$MDE = \frac{t_{(1-k)} + t_{\alpha/2}}{\sqrt{P(1-P)}} \times \sqrt{\rho + \frac{1-\rho}{n} \sigma}$$

,where n – cluster size.

5.5 Interaction or subgroup analysis

- We often want to investigate interaction effects e.g. effect is stronger for women than men etc. However, it is natural to assume that the effect of the interaction is smaller than the main effect.
- Indeed, the median absolute magnitude of the interactions in their re-analysis is **0.065 s.d.**, or 37% of the size of the main treatment effects. (Muralidharan et al., 2019)
- Thus, we have to have even larger dataset to do subgroup analysis.
- Moreover, we increased number of hypothesis further reducing statistical power as we have to correct for multiple hypotheses testing
- How Machine Learning can help?

5.5.1 Effect Heterogeneity with Machine Learning

- A. Use machine learning tools e.g. Random Forest to explore treatment heterogeneity in systematic way.
- B. Use predictions from ML to estimate the treatment effect by groups.

\rightarrow Generic Machine Learning Chernozhukov et al. (2018), Policy Tree (Athey and Wager, 2021) See next section.

5.6 Exercises

1. **Hypothesis Testing:** Explain type error I and type error II. Explain significance level and power. Why typically in empirical analysis we talk about significance level but not power?
2. **Power I:** What is power of the statistical test? What is the typical power level one should choose? What are the advantages and disadvantages of higher power level?
3. **Power II:** What are the key five factors that determines the power of the study? In which direction?
4. **Power III:** If you use group level randomization, is it better to have small or big clusters?
5. **Power IV:** One can determine Minimum Detectable Effect Size. Which way would you choose?
6. **Power V:** How to choose residual variance for your study? Can you use data at hand?

7. **Power and interactions I:** Do you think it is required to have larger sample size to provide heterogeneity analysis? Why?
8. **Power and interactions II:** Do you think typically the effect size for interaction is larger or smaller than the main treatment effect?
9. **Power and interactions III:** You study in a randomized control trial if response to request about donation vary across people of different ethnicity: In the treatment group you emphasize the importance of donation for others; In the control group you just ask to donate. You provide statistical analysis of the whole sample and do not find any difference between control and treatment group. When, however, you provide an subgroup analysis of 100 ethnic groups, you find that some group is less likely to donate if they were treated and you can reject the null-hypothesis at 5% level providing simple t-test for this group (p -value= 0.0467). What will be your conclusion and how would you summarize the results?

6 Exploring Effect Heterogeneity with Machine Learning

6.1 Motivation

- We often want to see if the treatment effect varies by characteristics such as gender, age, income, etc.
- Understanding subgroup-specific effects aids in targeting treatments more effectively.
- However, we saw that analyzing sub-groups in a disciplined manner can be demanding from a statistical power perspective (small effect size, large samples, but we have sub-sample, etc.)
- Moreover, having many hypotheses tested increases the false positive rate (which can also decrease power) and has to be accounted for systematically.
- We cover next ML methods to adress it: (1) Casual Forest; (2) Generic Machine Learning; (3) Optimal Policy Tree

6.2 Casual Forest

- Let's suppose we have a randomly assigned treatment, T .
- We get used to estimate $\hat{y}(T, x) = E[Y_i|T_i, X_i]$ using Machine Learning or other methods
- Idea!
- Instead of predicting $\hat{y}(T, x)$, we predict the mean difference τ_i in Y across treatment groups T , building decision trees for different variables X .
- Algorithm aims at generating variables value-specific subgroups in a way that minimizes the sum of squared residuals in effect estimation (rather than outcome Y prediction).
- We aim to maximize within-group effect homogeneity or inter-group effect heterogeneity based on values of X .
- Honest prediction: use one set of observations to build the tree structure and another to generate predictions of effects within subsets.

Conditional Average Treatment Effect (CATE):

$$\begin{aligned} E[\tau_i|X_i] &:= E[Y_i(1) - Y_i(0)|X_i] \\ &= E[Y_i|X_i, T_i = 1] - E[Y_i|X_i, T_i = 0] \end{aligned}$$

Adjust the criterion:

$$\min \sum_{i=1}^n (\tau(X_i) - \tau_i)^2 \iff \max \sum_{i=1}^n \tau(X_i)^2$$

Predicting outcomes vs. treatment effects

Typically, we use ML to Predict Outcome	Now we want to predicting treatment effects
---	---

Target: $\hat{y}(x) = E[Y_i X_i = x]$	Target: $\tau(x) = E[\tau_i X_i = x]$
---------------------------------------	---------------------------------------

\Rightarrow	\Rightarrow
---------------	---------------

Criterion:	Criterion:
------------	------------

$\min E[(\hat{y}(x) - Y_i)^2 X_i = x]$	$\min E[(\tau(x) - \tau_i)^2 X_i = x]$
--	--

$\{Y_i, X_i\}_{i=1}^n$	$\{\tau_i, X_i\}_{i=1}^n$
Causal (Random) Forest Algorithm	

1. We predict using the Random Forest Algorithm:

- Y with a set of variables of X
- T with a set of variables of X

2. We calculate residuals using prediction from 1

3. We use a random forest to predict the effect of T's residuals on Y's residuals, based on X. This involves averaging over numerous causal trees, which model effect heterogeneity with a tree structure and estimate treatment effects within subsets.

In short: We use (a) splitting and paratilling out to control for important confounders (see previous Modern High Dimensional Linear Regression) and (b) the causal tree approach for finding effect heterogeneity.

6.3 Generic Machine Learning

6.3.1 Generic Machine Learning Inference

Instead of estimating the Conditional Average Treatment Effect, we can:

1. Test if heterogeneity is present in treatment effect τ
2. If heterogeneity is present, we shall see the predicted outcomes interact with the treatment effect
3. See the average characteristics of the units in the most and least affected groups.

In other words, we estimate:

1. Best linear predictor (BLP) of Conditional Average Treatment Effect using ML
2. Group Average Treatment Effects (GATES) sorted by the groups induced by ML prediction

3. Classification Analysis (CLAN): Average characteristics of the units in most and least affected groups

The approach is generic in the sense we are flexible which Machine Learning approach we use.

Generic Machine Learning Inference Element 1: Best linear predictor (BLP)

- To estimate the Best linear predictor (BLP), we estimate:

$$Y = a_0 + aB(X) + \beta_1(T - p(X)) + \beta_2(T - p(X))(S(X) - E[S(X)]) + \varepsilon$$

- $B(X)$ - baseline effect conditional on the covariates X .
- T - treatment assignment, $p(X)$ propensity score, the weights are $\frac{1}{p(X)(1-p(X))}$,
- $S(X)$ - treatment effect - the difference between the predicted control group outcome and the outcome in the treated group.
- *Note* Recall simple regression with interaction term to get intuition about interpretation of coefficients: $Y = \beta_0 + \beta_T T + \beta_{GT} G * T + u$
- β_2 - is the key coefficient we are interested for heterogeneity
- $S(X)$ is perfect proxy/predictor $\rightarrow \beta_2 = 1$
- $S(X)$ is complete noise \rightarrow No (observed) heterogeneity, $\beta_2 = 0$,
- $S(X)$ is its relevant predictor $\rightarrow \beta_2 \neq 0$, we see heterogeneity
 - Thus, empirically rejecting the null hypothesis that $\beta_2 = 0$ suggests heterogeneity and $S(X)$ is a relevant predictor

Generic Machine Learning Inference Element 2: Group Average Treatment Effects (GATES)

- Now, we can estimate the treatment effect (on the hold-out sample)
- The sample is divided into k -quantiles based on the predicted treatment effect $S(X)$ distribution.
- K -quantiles correspond to the $N\%$ units sorted by the predicted outcome, typically quartiles (four groups).

$$Y = a_0 + \alpha_1 B(X) + \sum_{k=1}^4 \gamma_k (T - p(X)) \mathbf{1}(G_k) + \varepsilon$$

- γ_k - quartiles of the predicted treatment effect $S(X)$
- G_k group defined by the quartiles of the predicted treatment effect $S(X)$
- For instance, for the most affected G_4 , the coefficient γ_4 indicates the average treatment effect for the outcome Y .

Generic Machine Learning Inference Element 3: Classification Analysis (CLAN)

- Now, since we have the most affected and least affected groups e.g., for quartiles G_1 an G_4

$$\delta_1 = E[g(Y, X)|G_1] \quad \text{and} \quad \delta_k = E[g(Y, X)|G_k].$$

- We can compare δ_1 and δ_k based on X e.g. Gender, Age, Income
- Significant differences in covariates e.g. Gender, Age, Income correspond to influences of these covariates on the treatment mechanism
- For instance, average income in the least affected group is significantly lower than in the most affected group

→ income may influence how the treatment impacts an individual.

6.3.2 Example of Generic Machine Learning Inference

- Where are nudges for immunization for measles shot the most effective?
- Analyzing part of the experiment
 - Treatment: information-hub ambassador about immunization, reminders, and increased incentives for immunization
 - Control: No intervention
- The treatment is expensive. Whom to target?

TABLE 3. BLP of Immunization Incentives

Elastic Net		Neural Network	
ATE (β_1)	HET (β_2)	ATE (β_1)	HET (β_2)
2.812	0.876	2.530	1.059
(0.867,4.774)	(0.656,1.105)	(0.984,4.079)	(0.724,1.401)
[0.008]	[0.000]	[0.003]	[0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Figure 2: Best Linear Predictor. Chernozhukov et al. (2020)

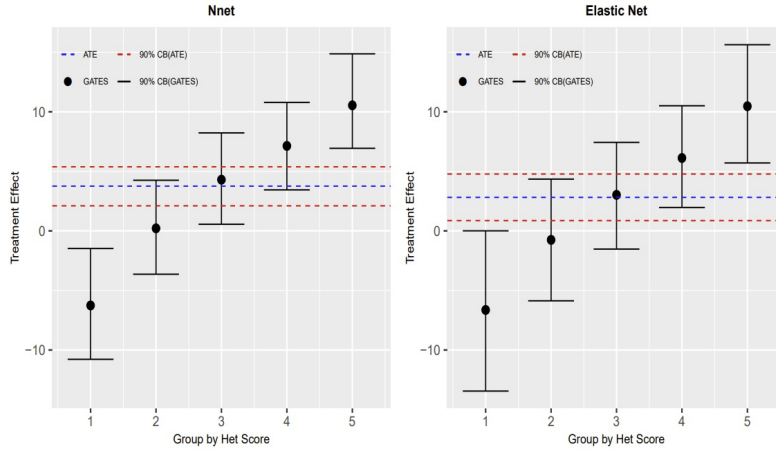


Figure 3: Group Average Treatment Effects (GATES). Chernozhukov et al. (2020)

TABLE 5. CLAN of Immunization Incentives

	Elastic Net			Nnet		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Number of vaccines to pregnant mother	2.161 (2.110,2.212)	2.288 (2.237,2.337)	-0.128 (-0.200,-0.055) [0.001]	2.164 (2.107,2.221)	2.328 (2.273,2.385)	-0.160 (-0.245,-0.082) [0.000]
Number of vaccines to child since birth	4.230 (4.100,4.369)	4.714 (4.573,4.860)	-0.513 (-0.710,-0.311) [0.000]	3.995 (3.816,4.165)	4.670 (4.507,4.835)	-0.690 (-0.937,-0.454) [0.000]
Fraction of children received polio drops	1.000 (1.000,1.000)	1.000 (1.000,1.000)	0.000 (0.000,0.000) [0.000]	0.998 (0.996,1.000)	1.000 (0.998,1.002)	-0.002 (-0.005,0.001) [0.485]
Number of polio drops to child	2.964 (2.954,2.975)	2.998 (2.987,3.007)	-0.033 (-0.047,-0.019) [0.000]	2.956 (2.940,2.971)	2.994 (2.980,3.008)	-0.038 (-0.059,-0.016) [0.001]
Fraction of children received immunization card	0.899 (0.878,0.922)	0.932 (0.908,0.956)	-0.036 (-0.065,-0.004) [0.000]	0.804 (0.765,0.842)	0.930 (0.895,0.966)	-0.125 (-0.178,-0.070) [0.006]
Fraction of children received Measles vaccine by 15 months of age	0.127 (0.100,0.155)	0.255 (0.230,0.282)	-0.131 (-0.167,-0.094) [0.052]	0.125 (0.098,0.152)	0.254 (0.229,0.279)	-0.134 (-0.169,-0.098) [0.000]
Fraction of children received Measles at credible locations	0.290 (0.252,0.327)	0.435 (0.400,0.470)	-0.152 (-0.198,-0.097) [0.000]	0.275 (0.236,0.315)	0.426 (0.391,0.461)	-0.151 (-0.203,-0.100) [0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
Notes: P-values for the hypothesis that the parameter is equal to zero in brackets.

Figure 4: Classification Analysis. Chernozhukov et al. (2020)

- If we use Generic Machine Learning to estimate (1) the Best linear predictor (BLP), (2) Group Average Treatment Effects (GATES), and (3) provide Classification Analysis (CLAN)
 - We test agnostically if (1) heterogeneity is present, (2) see how large it is, and (3) what potential drivers can be.
 - We are flexible on which Machine Learning tools we use to discover heterogeneity.
- Note that a single estimation may not be stable; thus, we repeat the process many times and store parameters of interest.

6.4 (Optimal) Policy Learning

6.4.1 Policy Learning

- Let's suppose we have to allocate policy/treatment across different groups of people.
- We might give treatment not to all but to a group of people due to budget constraints, especially if the treatment has some costs. For instance,
 - Medical treatment works e.g. Vaccine, but whom to give it (first) to maximize protection given that Vaccine is scarce
 - Programs to fight unemployment can be beneficial, but whom to select from to maximize employment chances under budget constraints.
 - Having a centralized management online system improves the learning process and outcomes in schools on average (Asanov et al., 2023), but we need to know if the optimal policy implies establishing this system for all schools or only some specific schools to maximize the effect under budget constraints.
- We will focus on learning how to improve average results in scenarios with two treatment options based on different background factors.
- Define treatment policy $\pi(X)$ as a function of covariates X .
- Example policy: medical treatment assignment, e.g., vaccination, denoted by T .
- For age 65 or older, set $T = 1$ (vaccination); otherwise $D = 0$ (no vaccination).
- Policy rule example: $\pi(X) = I\{age \geq 65\}$, so $\pi(age = 30) = 0$ and $\pi(age = 80) = 1$.
- The policy may depend on one covariate (age) or several covariates X .
- Average effect of policy $\pi(X)$, denoted by $\Delta(\pi(X))$.
- $\Delta(\pi(X))$ corresponds to the difference in mean potential outcomes:

$$\begin{aligned}
 \Delta(\pi(X)) &= E[Y(\pi(X)) - Y(0)] \\
 &= E[\pi(X) \cdot (Y(1) - Y(0))] \\
 &= E[\pi(X)] \cdot E[Y(1) - Y(0)] \\
 &= E[\pi(X) \cdot \Delta X].
 \end{aligned}$$

- This reflects the difference between treatment and non-treatment for everyone.
- The optimal policy $\pi^*(X)$ maximizes the average effect among all feasible policies, denoted by Π , assuming a countable number of policies in this set Π :

$$\pi^*(X) = \max_{\pi \in \Pi} \Delta(\pi(X)). \quad (13)$$

- We define the regret function associated with treatment policy $\pi(X)$, which corresponds to the reduction in the average policy effect when a suboptimal policy is implemented instead of the optimal one:

$$R(\pi(X)) = \Delta(\pi^*(X)) - \Delta(\pi(X)). \quad (14)$$

- Finding the optimal policy among the set of feasible policies Π means that the average policy effect is maximized and regret R is zero. This is equivalent to solving the following maximization problem:

$$\pi^*(X) = \max_{\pi \in \Pi} E[2(\pi(X) - 1) \cdot \phi(X)]. \quad (15)$$

- $\pi^*(X) = \max_{\pi \in \Pi} E[2(\pi(X) - 1) \cdot \phi(X)]$ aligns with effect heterogeneity analysis, utilizing the efficient influence function $\phi(X)$, beneficial for optimal policy learning.
 - Optimal policy maximization requires a balance in Conditional Average Treatment Effect (CATE) among treated and non-treated when choosing the treatment policy among all possible policies.
 - The term $(2\pi(X) - 1)$ indicates that Conditional Average Treatment Effect CATEs for treated and non-treated subjects affect expectations positively and negatively, respectively.
 - Now, how do we estimate $\phi(X)$ using ML?
 - We can estimate $\phi(X)$ by cross-validation. Recall, for instance, how we estimated λ for lasso regression in modern high-dimensional linear regression.
 - What if we have unobserved confounders in the observational study or non-perfect compliance in the experiment?
- Learning policies via Double Machine Learning (DML) for estimating the optimal policy $\hat{\pi}^*(X)$
- The regret $R(\hat{\pi}^*(X)) = \Delta(\pi^*(X)) - \Delta(\hat{\pi}^*(X))$ for the estimated optimal policy versus the true optimal policy can approach zero at an optimal \sqrt{n} -rate. (Athey and Wager 2021)
 - A necessary condition for these properties is that the policy space Π is not overly complex. Namely, number of policies is finite.

6.4.2 Example of Optimal Policy Learning

- How can we improve remote online learning and educational outcomes?
- We tested light-touch interventions in 1,151 schools in Ecuador, covering more than 45,000 students in their final years of high school (Asanov et al., 2023).
- As part of the experiment, we randomly assigned schools to
 - Self-Management (Control)
 - Centralized Management (We add ministry personnel to the online management system and monitoring information)
- Centralized monitoring improved subject knowledge by 0.13 SD relative to decentralized management.

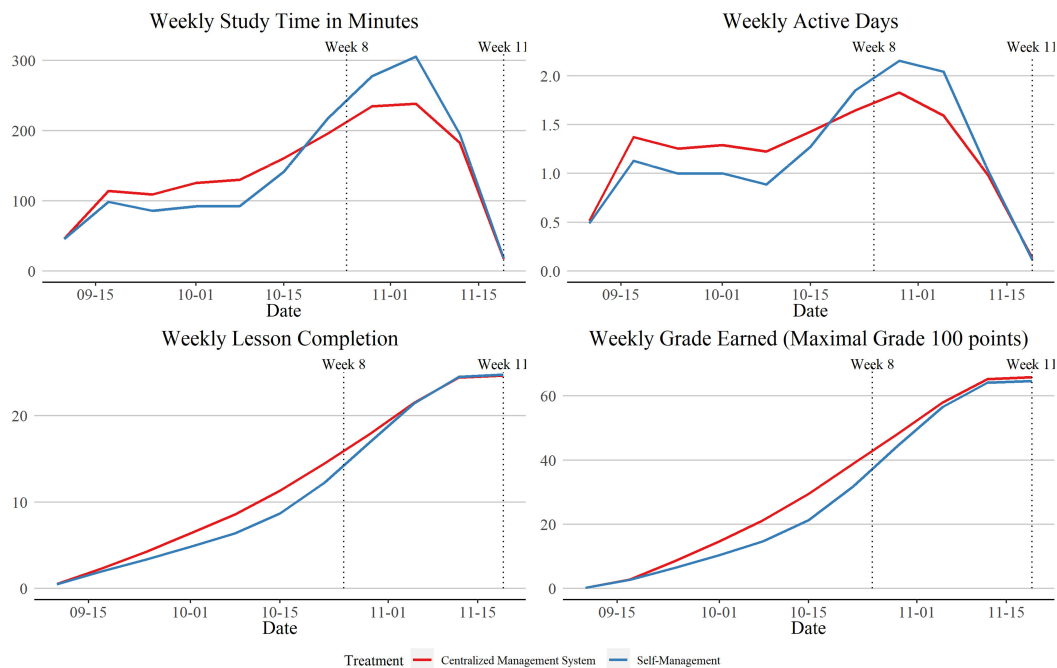


Figure 5: Students learning pattern (Asanov et al., 2023).

- The size of the improvement is comparable to how much a student might learn in 71 percent of a year of business-as-usual schooling in grade 12.
- The intervention is cheap (below 60 cents per student), yet which schools shall we target under budget constraints to maximize the policy effect?
- We run a policy tree algorithm feeding administrative data that the Ministry of Education collects regularly and can thus use for decision-making.
- The algorithm selects a decision rule based on the average score in the school in the national end-of-school examinations.
- The algorithm suggests using a centralized Management system for schools that score below national end-of-school examinations.

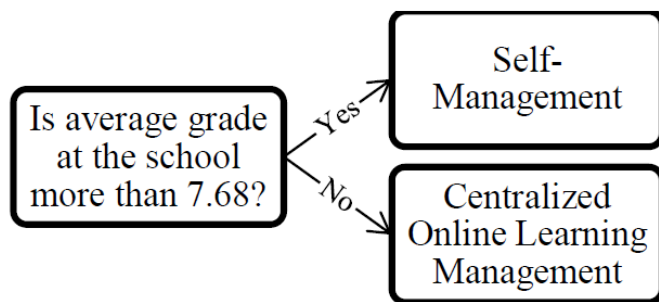


Figure 6: Policy tree results. Asanov et al.(2023)

- We want to allocate policy/treatment across different groups of people.
- Policy tree agnostically considers a set of targeting policies and selects the one that maximizes the effect.
- Policy tree brings us simple and easy-to-communicate decision rules on whom to target.

6.5 Summary

- We often want to investigate if the treatment effect varies by characteristics of interest.
 - **Casual trees** can help us find variables based on which we observe different treatment effect
 - **Generic Machine Learning** can help us to see if heterogeneity exists (BLP), how large it is (GATES), and what potentially drives it (CLAN)
 - **Policy learning** can help us to find simple and easy-to-communicate targeting policies/policy trees among several policies under budget constraints.

References:

- Asanov, I., Asanov, A.M., Åstebro, T., Buenstorf, G., Crépon, B., McKenzie, D., Flores T, F.P., Mensmann, M. and Schulte, M., 2023. System-, teacher-, and student-level interventions for improving participation in online learning at scale in high schools. *Proceedings of the National Academy of Sciences*, 120(30), p.e2216686120.
- Athey, S. and Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), pp.7353-7360.
- Athey, S. and Wager, S., 2021. Policy learning with observational data. *Econometrica*, 89(1), pp.133-161.
- Chernozhukov et al. (2020) Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India

6.6 Exercises

1. **Causal Forest I:** What is our target function in Causal Forest?
2. **Causal Forest II:** Which Criterion we use? Why?
3. **Causal Forest III:** What additionally we do in Casual Random Forest? What is the advantage?
4. **Generic Machine Learning I:** Explain Best Linear Predictor, Group Average Treatment Effects (GATES), Classification Analysis (CLAN).
5. **Generic Machine Learning II:** If hypothesis $\beta_2 = 0$ is rejected in Generic Machine Learning procedure. What does it imply?
6. **Generic Machine Learning III:** Why do we use BPL, GATES, CLAN,?
7. **Generic Machine Learning IV:** What do variables selected in Classification Analysis (CLAN) mean? Can we interpret variables selected in Classification Analysis (CLAN) in causal way?
8. **Generic Machine Learning V:** What are the advantages of Generic Machine Learning?

9. **Generic Machine Learning VI:** What is the disadvantage of Generic Machine Learning?
10. **Applications I:** Give examples of using the Causal Forest in research or for a company.
11. **Applications II:** Give examples of using Generic Machine Learning in research or for a company.

7 Course Summary

7.1 Machine Learning

- We can solve many prediction tasks using cross-validation.
 - We can split the sample to train the model and validate the model on testing set (many times).
 - We are flexible which model (OLS, decision trees etc.) we use, but cross-validation helps to see performance of the model in sense prediction.
- Yet, even though predictions can be useful, it is not necessary that we causal relationship.

7.2 Statistical Inference

- We assume that observed sample comes from some population
- Test hypothesise a s aiming to reject the Null-hypothesis
- Thus making inference about the population
- **Fundamental Problem of Causal Inference** implies that we can not observe the same unit under different conditions in the same time
- We aim to apply statistical inference - statistical solution - to circumvent this problem on average
 - Calculate the **Average Treatment effect (ATE)**
- Unfortunately, in non-experimental data, we often observe **Average Predictive effect (APE)**.
- $APE \neq ATE$ due to the endogeneity e.g. selection of participants to the program, unobserved factors
 - Consult chapter 2 and 3 in "Causal Analysis..." by M. Huber

7.3 Modern High Dimensional Linear Regression

- We can account for confounders in regression analysis e.g. OLS
- Yet, which variables to select is not known.
- Moreover, in presence of many variables (many dimensions) we might not able to properly estimate the regression at all (overfit)
- We can use **double lasso procedure:**

- With the help of lasso, select the variables that strongly **predict** the outcome and variable of interest.
- In this way we reduce the risk to overfit, but also account for the strong confounders directly (or through proxies)
- We can do more by explicitly modelling non-linearities e.g. polynomials, interactions
- We can use double debiased machine learning (DML) approach
 - We use the flexibility of Machine learning to predict the outcome and variable of interest.
 - Then, with the help of the Frisch–Waugh–Lovell theorem estimate coefficient of interest through residuals based on ML predictions

Consult chapter 4 and 5 in "Causal Analysis..." by M. Huber

7.4 Exploring effect Heterogeneity with Machine Learning

- We often want to know if the effect is particularly strong for certain group
- Yet, subgroup analysis likely to undermine the statistical power e.g. we split the sample, interaction effect likely to be smaller than main effect, testing many hypotheses
- Testing many hypotheses can be particularly dangerous as we increase false discovery rate
- We might use a Machine learning approach to explore heterogeneity in systematic way
- We might use machine learning to search for the predictors of effect heterogeneity - random casual forest
- We can test if the heterogeneity is generally present, see how large is that and what potentially drives it -generic machine learning
- We can develop predictions whom to target to maximize the effect of the policy- Policy learning/policy tree

Consult chapter 5 in "Causal Analysis..." by M. Huber

- Causal Machine learning is growing field
- It embraces the predictive power of machine learning to get close to causal estimates
- In certain cases, the casual machine learning works wonders
- Yet, one still have to remember that wonders are built on predictions based on observable factors
- So, one shall be still cautious in interpretation of results.

7.5 Exercises

12. For the exam:
 - (a) Please, revisit handout
 - (a) Please, consult chapter 1-5 of "Causal Analysis..." by M. Huber
 - (b) Attempt to make double lasso estimations
1. Describe steps in standard prediction study design?
2. Is out of sample error typically equal than in sample error in prediction study?
3. What are the limitations of Machine Learning?
4. What is the fundamental problem of causal inference?
5. You get the data from the , identical twins that are put under treatment and no treatment condition, and the only difference in outcomes between them is induced by treatment. You find that the treatment - private school type that one of the twins attended - is associated with higher income. Do you think you can apply these results to general population?
6. On average, countries that spend more on education have higher GDP. Can you conclude that education drives economic growth?
1. Double lasso procedure selects the number of times that person takes a shower when you study the effect of education on income. Shall you keep this variable in regression?
2. You estimate $Y = \beta_0 + \beta_T T + \beta_X X + u$ regression and get $\beta_T = 2.6$ for your variable of interest T. Colleague estimated β_T using Frisch–Waugh–Lovell theorem. What value of β_T she will get?
3. A Colleague of yours says that there is a highly powerful method for prediction - a generative neural network. Do you think you can apply it in double-biased machine learning estimations?
4. Double lasso procedure selects the number of times that person takes a shower when you study the effect of education on income. Shall you keep this variable in regression?
5. You estimate $Y = \beta_0 + \beta_T T + \beta_X X + u$ regression and get $\beta_T = 2.6$ for your variable of interest T. Colleague estimated β_T using Frisch–Waugh–Lovell theorem. What value of β_T she will get?
6. A colleague of yours says that there is a highly powerful method for prediction - a generative neural network. Do you think you can apply it in double-biased machine learning estimations?
7. Tell 5 determinants of statistical power.
8. Colleague of you has a theory that women with higher education and whose parents have higher education will have higher income than man. How can you test it? What alternative approach can you choose?
9. Casual forest selects $\text{age} > 30$ as a predictor that entrepreneurial program will result in higher self-employed revenue. You, however, can not reject the null hypothesis that the Best Linear predictor coefficient is different from the zero based on the Generic Machine learning approach. Would you recommend targeting entrepreneurial programs for people above 30?