Final Project

Ruihan Jiang

Due date = 12/16/2022

Final Project

As the final project of this class, you will either

1) analyze a data set using some of the methods for causal inference that we discuss or

The findings will be written-up in a short paper not to exceed six pages including reference.

Data

The data applied here is obtained from the book Machine Learning with R by Brett Lantz. Since the US Census Bureau's demographic figures were used to compile the data originally, it roughly reflects actual situations. It contains 1,338 samples of beneficiaries who are presently enrolled in the insurance plan, with attributes that describe the patient's features and the total amount of medical costs billed to the plan for the calendar year. The features are:

- `age`: An integer indicating the age of the primary beneficiary (excluding those above 64 years, as they are generally covered by the government).
- `sex` (treatment, Z): The policy holder's gender: either male or female.
- `bmi`: The body mass index (BMI), which provides a sense of how over or underweight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
- `children`: An integer indicating the number of children/dependents covered by the insurance plan.
- `smoker`: A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
- `region`: The beneficiary's place of residence in the US, divided into four geographic regions: northeast (4), southeast (2), southwest (1), or northwest (3).
- `charges` (outcome, Y): medical cost per year in dollars.

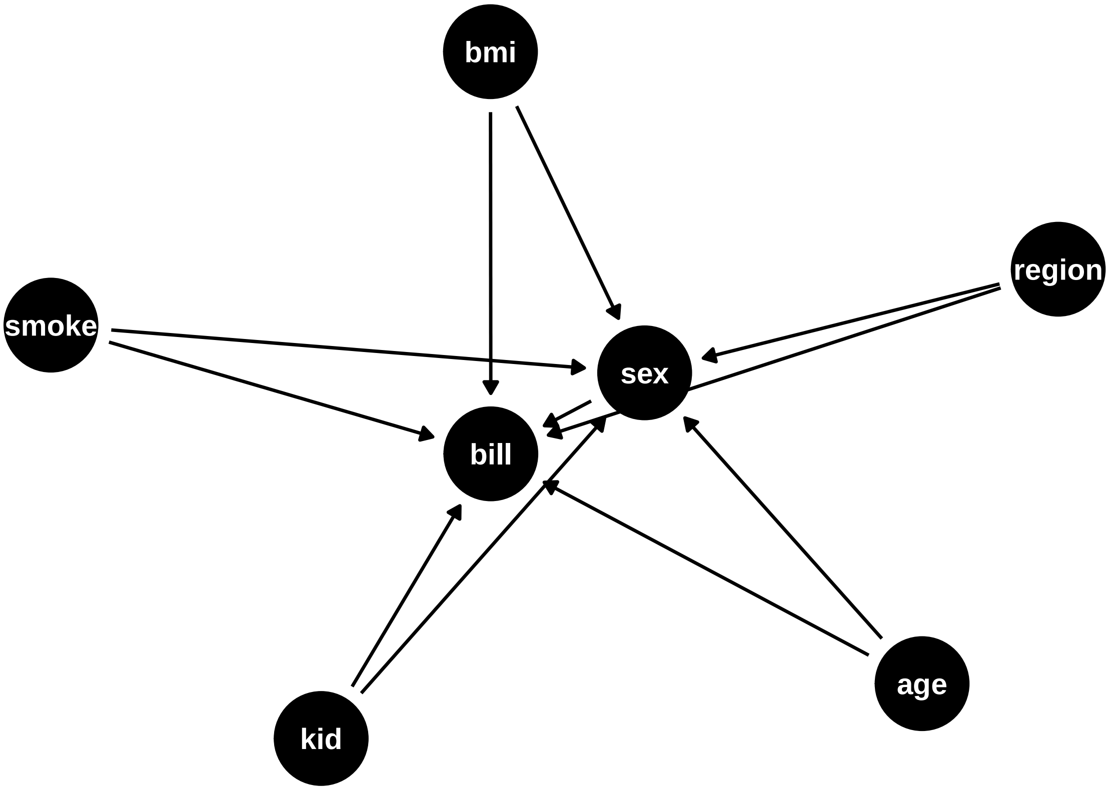
```r
library(tibble)
library(tidyverse)
library(MatchIt)
library(modelsummary)
library(dagitty)
library(ggdag)
library(ggplot2)

data <- read.csv("insurance.csv") %>% mutate(sex = ifelse(sex == "female", 1, 0),
                                              smoker = ifelse(smoker == "yes", 1, 0),
                                             region = ifelse(region == "southwest", 1,
                                                             ifelse(region == "southeast", 2,
                                                                    ifelse(region == "northwest", 3, 4)
head(data)
```

```
##   age sex    bmi children smoker region   charges
## 1  19   1 27.900        0      1      1 16884.924
## 2  18   0 33.770        1      0      2  1725.552
## 3  28   0 33.000        3      0      2  4449.462
## 4  33   0 22.705        0      0      3 21984.471
## 5  32   0 28.880        0      0      3  3866.855
## 6  31   1 25.740        0      0      2  3756.622
```

Directed Acyclic Graph (DAG)

```
dag <- dagify(bill ~ bmi + kid + smoke + region + age + sex,
              sex ~ bmi + kid + smoke + region + age)
ggdag(dag, edge_type = "link") + theme_dag()
```



I first focus on the outcome which is the insurance bill, I think all observed features have influence on that. Age, whether is a smoker, bmi and number of children covered belong to the personal physical information that affects the insurance price. Region is important too because different regions have different tax rates. Then, I look at the treatment gender. female tend to have more kids covered (Less children per man than per woman 2016), relatively higher bmi (ooper & Gupta, Moustafa & Chao, 2021) and be less likely a smoker (Are there gender differences in tobacco smoking? 2021). What's more, young single women may pay the health insurance early than men (Romaine, 2022), therefore the age is also related to gender. To sum up, all observed variable except bill and sex are confounders.

Assumptions

1. Causal Ordering

It is obvious to see that the cost of health insurance can't affect individuals' gender so the causal ordering assumption is followed.

2. Stable Unit-treatment Variation Assumption (SUTVA)

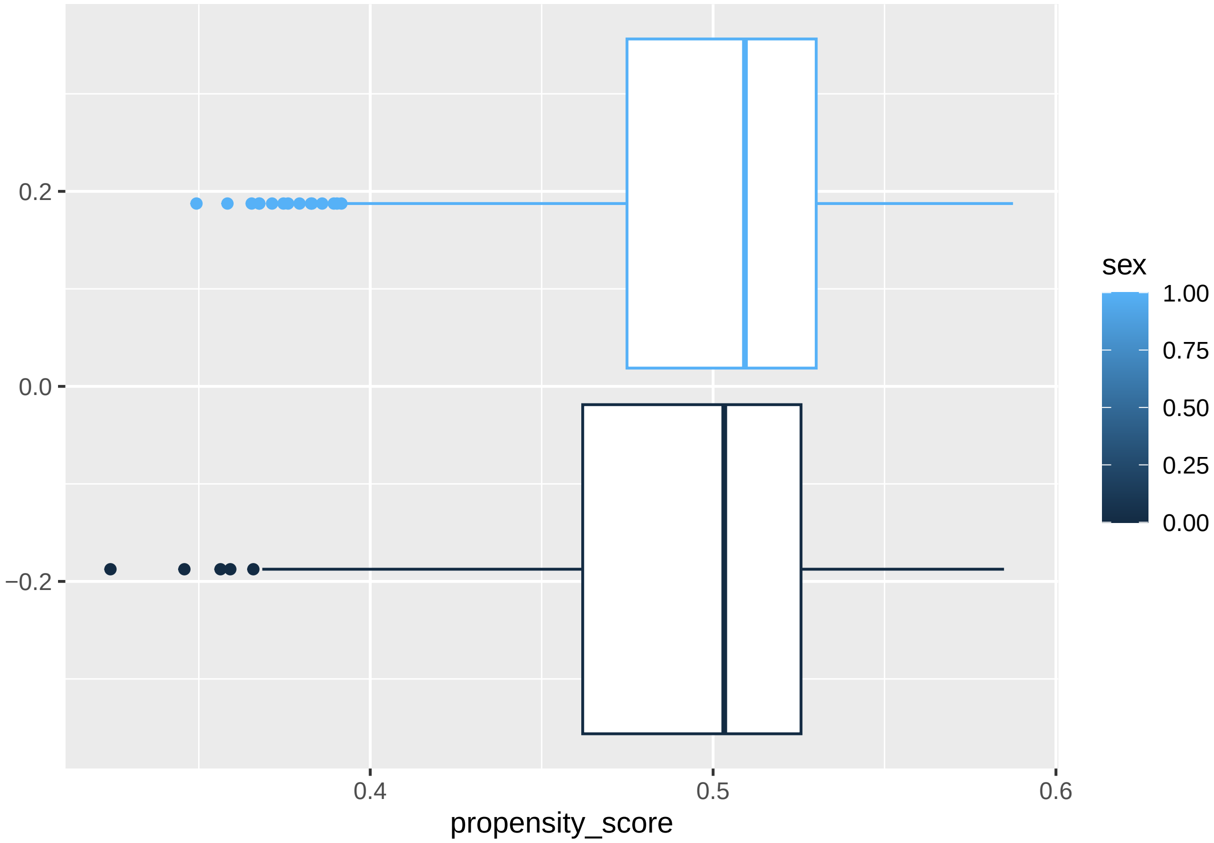
Because each unit represents one individual so there's no interference between each unit. As for the treatment gender, the data recorded natural biological gender which only contains male and female, thus there's no hidden version of treatment.

3. Overlap Assumption

I did this via propensity score. Look at the propensity score distributions in both groups, it shows that both groups are overlapped but with three outliers. Since there is only a few of it, I don't remove them.

```
#Compute Propensity Score
model <- glm(sex ~ age + bmi + children + smoker + region, data = data,
             family = binomial(link = "logit"))
data$propensity_score <- predict(model, data, type = "response")

# Propensity Score Distribution Plot
data %>% ggplot() + geom_boxplot(mapping = aes(x=propensity_score, group = sex, color = sex))
```



```
#Outliers
range_e <- data %>% group_by(sex) %>% summarise(min = min(propensity_score),
                                                max = max(propensity_score))

outliers <- rbind(data %>% filter(sex == 0, !between(propensity_score, range_e[2,2], range_e[2,3])),
                  data %>% filter(sex == 1, !between(propensity_score, range_e[1,2], range_e[1,3]))
                  )
nrow(outliers)
```

```
## [1] 3
```

4. Unconfoundness Assumption

The treatment assignment of gender conditioned on covariates does not depend on any potential outcomes which means that the gender factor is independent of the health insurance charges that female or male individual would need to pay. Therefore, this assumption is fulfilled.

Various Estimators via Different Methods and Summary of Estimation Results

1. Check Balance

I am really happy with these balances since the data has similar sample size for each group (male has 14 more data) and similar means for every pre-treatment characteristics floating within approximately 0.05.

```
balance <- data %>%
  group_by(sex) %>%
  summarise(sample_size = n(),
            proportion = n()/nrow(data),
            ave_age = mean(age),
            ave_bmi = mean(bmi),
            ave_children = mean(children),
            ave_smoker = mean(smoker),
            ave_region = mean(region),
            ave_charges = mean(charges))
balance
```

```
## # A tibble: 2 x 9
##     sex sample_size proportion ave_age ave_bmi ave_chi~1 ave_s~2 ave_r~3 ave_c~4
##   <dbl>       <int>      <dbl>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
## 1     0         676      0.505    38.9    30.9      1.12   0.235    2.48  13957.
## 2     1         662      0.495    39.5    30.4      1.07   0.174    2.49  12570.
## # ... with abbreviated variable names 1: ave_children, 2: ave_smoker,
## #   3: ave_region, 4: ave_charges
```

```
diff(balance$ave_charges) #female pay more, male pay less
```

```
## [1] -1387.172
```

2. Get Various Estimators via Different Methods

```
#1 Naive difference in means
naive <- lm(charges ~ sex, data = data)

#2 Adjustment with Mahalanobis nearest-neighbor matching
adjusted <- matchit(sex ~ age + bmi + children + smoker + region, data = data)
Mnnmatching <- lm(charges ~ sex, data = match.data(adjusted), weights = weights)

#3 Adjustment with inverse probability weighting
block <- quantile(data$propensity_score, probs = seq(0, 1, 0.1))
data$block <- cut(data$propensity_score, breaks = block, labels = 1:10, include.lowest=TRUE)
data$ipw <- (data$sex/data$propensity_score)+((1-data$sex)/(1-data$propensity_score))
```

	Naive	Matching	Weighting(W.)
sex	est:−1387.172	est:−970.270	est:114.031
Num.Obs.	1338	1324	1338
R2	0.003	0.002	0.000 02
R2 Adj.	0.003	0.0009	−0.0007
AIC	28 956.9	28 615.8	28 962.5
BIC	28 972.5	28 631.4	28 978.1
Log.Lik.	−14 475.432	−14 304.914	−14 478.262
F	4.400	2.192	0.030
RMSE	12 085.60	11 912.90	12 108.98

```
weighting <- lm(charges ~ sex, data = data, weights = ipw)

#4 regression adjustment (ANCOVA) using all the confounders
rgs_allCon <- lm(charges ~ sex + age + bmi + children + smoker + region, data = data)

#5 regression adjustment using the estimated propensity score and the treatment status as the predictor
rgs_psTreatment <- lm(charges ~ sex + propensity_score, data = data)

#6 regression adjustment using the estimated propensity score, in addition to the treatment status and
rgs_psAll <- lm(charges ~ sex + propensity_score + age + bmi + children + smoker + region, data = data)

#7 Weighted average treatment effect estimator using the overlap weight
data$overlap_weight <- data$sex * (1-data$propensity_score) + (1-data$sex) * data$propensity_score
overlapW <- lm(charges ~ sex + age + bmi + children + smoker + region, data = data, weights = overlap_w

#8 Regression on clever covariate (inverse probability as an additional covariate)
rgs_ipwAll <- lm(charges ~ sex + ipw + age + bmi + children + smoker + region, data = data)

#9 Weighted Regression (inverse probability as the weights)
Wrgs <- lm(charges ~ sex + age + bmi + children + smoker + region, data = data, weights = ipw)

# Compare Various Estimators
modelsummary(list("Naive" = naive,
                  "Matching" = Mnnmatching,
                  "Weighting(W.)" = weighting),
             estimate = "est:{estimate}",
             statistic = NULL,
             coef_omit = 1)
```

```
modelsummary(list("Overlap W."=overlapW,
                  "Inverse Probability W.: all+ipw" = rgs_ipwAll,
                  "W. Regression"=Wrgs,
                  "ANCOVA(A.): all" = rgs_allCon,
                  "A.: all+ps" = rgs_psAll,
                  "A.: ps+sex" = rgs_psTreatment),
             estimate = "est:{estimate}",
             statistic = NULL,
             coef_omit = c(1,3:9))
```

```
#10 stratification using propensity score
stratification <- data %>%
```

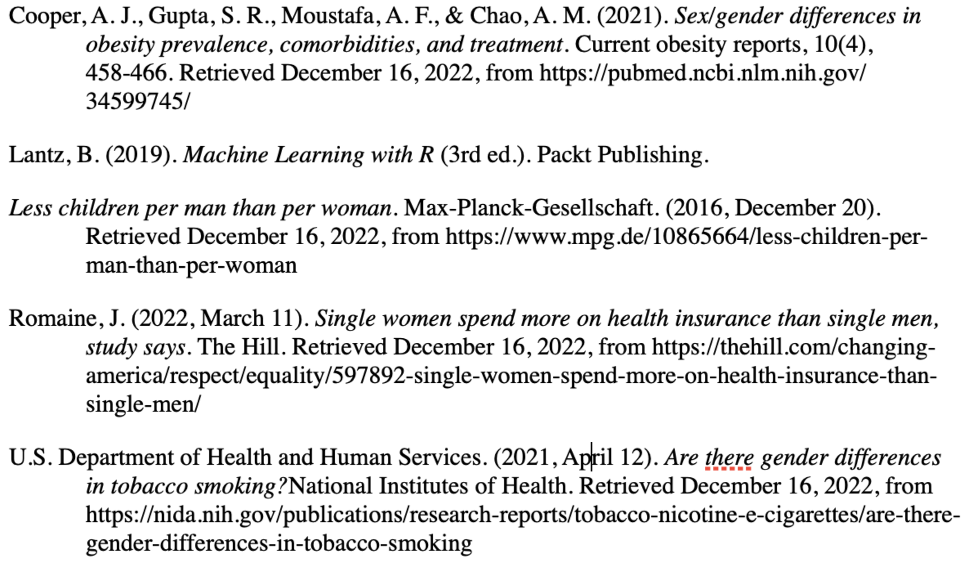
	Overlap W.	Inverse Probability W.: all+ipw	W. Regression	ANCOVA(A.): all	A.: all+ps	A.: ps+
sex	est:127.494	est:155.756	est:127.616	est:131.111	est:300.431	est:122
Num.Obs.	1338	1338	1338	1338	1338	1338
R2	0.746	0.751	0.749	0.751	0.790	0.44
R2 Adj.	0.744	0.750	0.747	0.750	0.789	0.44
AIC	27 116.7	27 113.9	27 125.2	27 112.5	26 883.7	28 179
BIC	27 158.3	27 160.7	27 166.8	27 154.0	26 930.5	28 199
Log.Lik.	−13 550.367	−13 547.964	−13 554.610	−13 548.225	−13 432.844	−14 085
F	650.026	572.545	660.522	668.124	715.729	532.0
RMSE	6044.31	6042.63	6044.00	6043.81	5544.47	9030.

```
  group_by(block, sex) %>%
  summarise(mean = mean(charges),
            count = n()) %>%
  group_by(block) %>%
  summarise(meandiff = diff(mean),
            count = sum(count)) %>%
  mutate(ATE = meandiff*(count/nrow(data)))
   #ATE
sum(stratification$ATE)
```

```
## [1] 523.1409
```

Based on R^2 and R^2_{adj}, I think "overlap weighting" model, "weighting regression" model, "ANCOVA: all" model, and "ANCOVA: all+ps" model is not that fit for the data comparing with other models. So I pay more atttention on other methods. Except "ANCOVA: all+ps" model's estimator and the estimator by "stratification using propensity score" are more than $200 per year, all other estimator have a common range from approximately $122 to $131 per year. Since they all show a positive result, it can be concluded that the overall effects on the insurance bill is that female pay more than male per year.

Reference



Cooper, A. J., Gupta, S. R., Moustafa, A. F., & Chao, A. M. (2021). *Sex/gender differences in obesity prevalence, comorbidities, and treatment*. Current obesity reports, 10(4), 458-466. Retrieved December 16, 2022, from https://pubmed.ncbi.nlm.nih.gov/34599745/

Lantz, B. (2019). *Machine Learning with R* (3rd ed.). Packt Publishing.

Less children per man than per woman. Max-Planck-Gesellschaft. (2016, December 20). Retrieved December 16, 2022, from https://www.mpg.de/10865664/less-children-per-man-than-per-woman

Romaine, J. (2022, March 11). *Single women spend more on health insurance than single men, study says*. The Hill. Retrieved December 16, 2022, from https://thehill.com/changing-america/respect/equality/597892-single-women-spend-more-on-health-insurance-than-single-men/

U.S. Department of Health and Human Services. (2021, April 12). *Are there gender differences in tobacco smoking?*National Institutes of Health. Retrieved December 16, 2022, from https://nida.nih.gov/publications/research-reports/tobacco-nicotine-e-cigarettes/are-there-gender-differences-in-tobacco-smoking