R Assignment 3

Alvaro J. Castro Rivadeneira

November 12, 2021

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

2 Implement IPTW for a binary exposure.

1. Read in and explore the data set `RAssign3.csv`.

```
# Read in data
ObsData <- read.csv('RAssign3.csv')
head(ObsData)
```

```
##   W1 W2    W3    W4 A Y
## 1  1 67 -0.95 -2.81 0 1
## 2  1 52 -0.06  0.35 0 0
## 3  1 73  0.92 -0.60 1 0
## 4  0 62 -1.56  0.64 0 1
## 5  0 62 -0.70  1.62 0 1
## 6  1 72  1.32 -0.34 0 1
```

```
summary(ObsData)
```

```
##        W1               W2              W3                  W4
##  Min.   :0.0000   Min.   :50.00   Min.   :-3.69000   Min.   :-3.49000
##  1st Qu.:0.0000   1st Qu.:56.00   1st Qu.:-0.71250   1st Qu.:-0.66000
##  Median :1.0000   Median :62.00   Median :-0.02000   Median : 0.03000
##  Mean   :0.5028   Mean   :62.07   Mean   :-0.03393   Mean   : 0.01644
##  3rd Qu.:1.0000   3rd Qu.:68.00   3rd Qu.: 0.63000   3rd Qu.: 0.68000
##  Max.   :1.0000   Max.   :75.00   Max.   : 3.78000   Max.   : 3.61000
##        A                Y
##  Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.000
##  Mean   :0.3452   Mean   :0.292
##  3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.000
```

```
# 2,500 participants aged 50-75
# CV health range: -3.69 to 3.78, SES range: -3.49 to 3.61
table(ObsData[,c('W1', 'W2', 'A')])
```

```
## , , A = 0
## 
##    W2
## W1  50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
##   0 15 27 40 27 23 32 36 32 33 28 36 30 29 37 26 24 24 27 31 28 33 33 31 26 24
##   1 19 43 42 29 36 25 44 47 43 29 28 42 35 31 31 23 32 42 35 35 31 37 33 41 41
##    W2
## W1  75
##   0 15
##   1 16
## 
## , , A = 1
## 
##    W2
## W1  50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
##   0  7 31 20 20 26 18 25 21 18 16 24 23 18 21 23 13 23 22 27 16 18 14 14 12 17
##   1  8 15 24 16 19 14 14 19 20 12  8 16 14 13 19 20 16 13 15 16  9  7 13 13  7
##    W2
## W1  75
##   0  9
##   1  7
```

2. Estimate the propensity score $\mathbb{P}_0(A = 1|W)$, which is the conditional probability of owning a dog given the participants characteristics. Use the following *a priori*-specified parametric regression model:

$$\mathbb{P}_0(A = 1|W) = logit^{-1}[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4]$$

In practice, we would generally use a machine learning algorithm, such as Super Learner (coming next).

```
# Run a logistic regression to estimate the treatment mechanism P(A|W)
prob.AW.reg <- glm(A ~ W1 +W2 + W3 + W4, family="binomial", data=ObsData)
prob.AW.reg
```

```
## 
## Call:  glm(formula = A ~ W1 + W2 + W3 + W4, family = "binomial", data = ObsData)
## 
## Coefficients:
## (Intercept)           W1           W2           W3           W4  
##      0.9736      -0.7395      -0.0272       1.5555       1.5503  
## 
## Degrees of Freedom: 2499 Total (i.e. Null);  2495 Residual
## Null Deviance:       3222 
## Residual Deviance: 2017  AIC: 2027
```

3. Predict each participant's probability of having and not having a dog, given their covariates: $\hat{\mathbb{P}}(A = 1|W_i)$ and $\hat{\mathbb{P}}(A = 0|W_i)$.

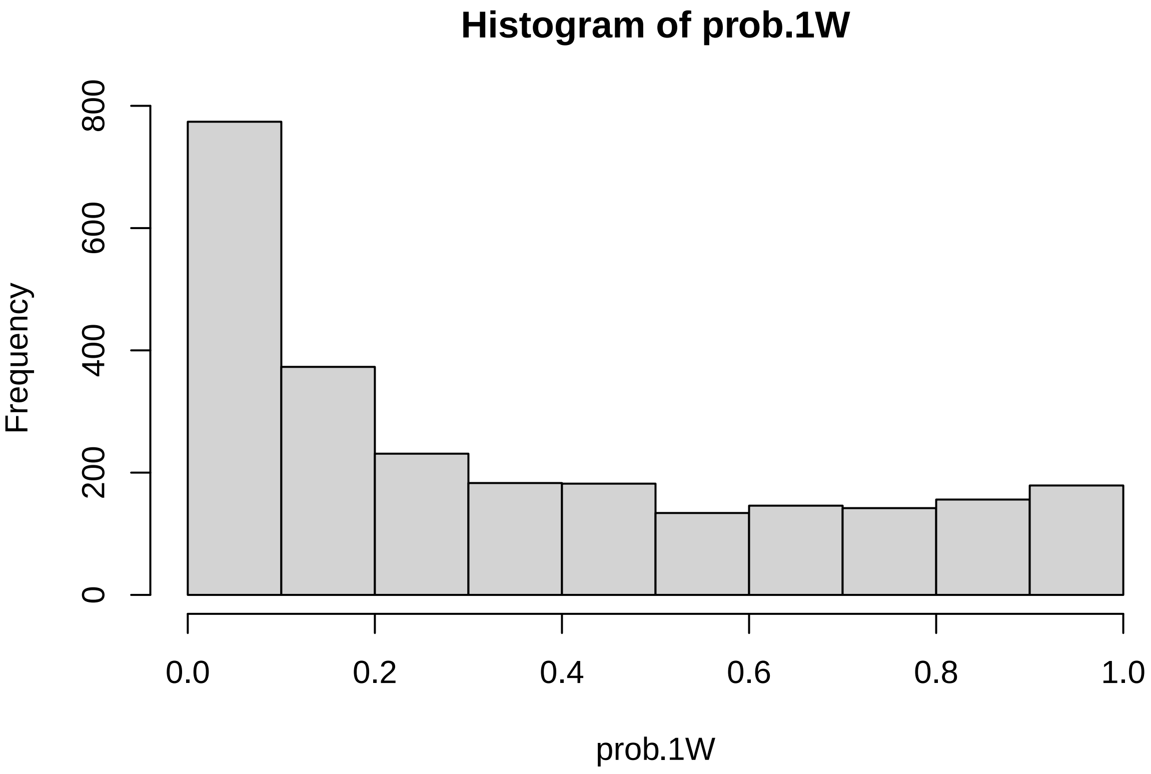
```
# Predicted probability of having a dog, given the obs cov P(A=1|W)
prob.1W <- predict(prob.AW.reg, type= "response")
# Predicted probability of not having a dog, given the obs cov P(A=0|W)
prob.0W <- 1 - prob.1W
```

4. Use the `summary` function to examine the distribution of the predicted probabilities $\hat{\mathbb{P}}(A = 1|W_i)$ and $\hat{\mathbb{P}}(A = 0|W_i)$. Any cause for concern?

```
# look at the distribution of predicted probabilities
summary(prob.1W)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0002953 0.0705011 0.2388129 0.3452000 0.5984808 0.9993384
```

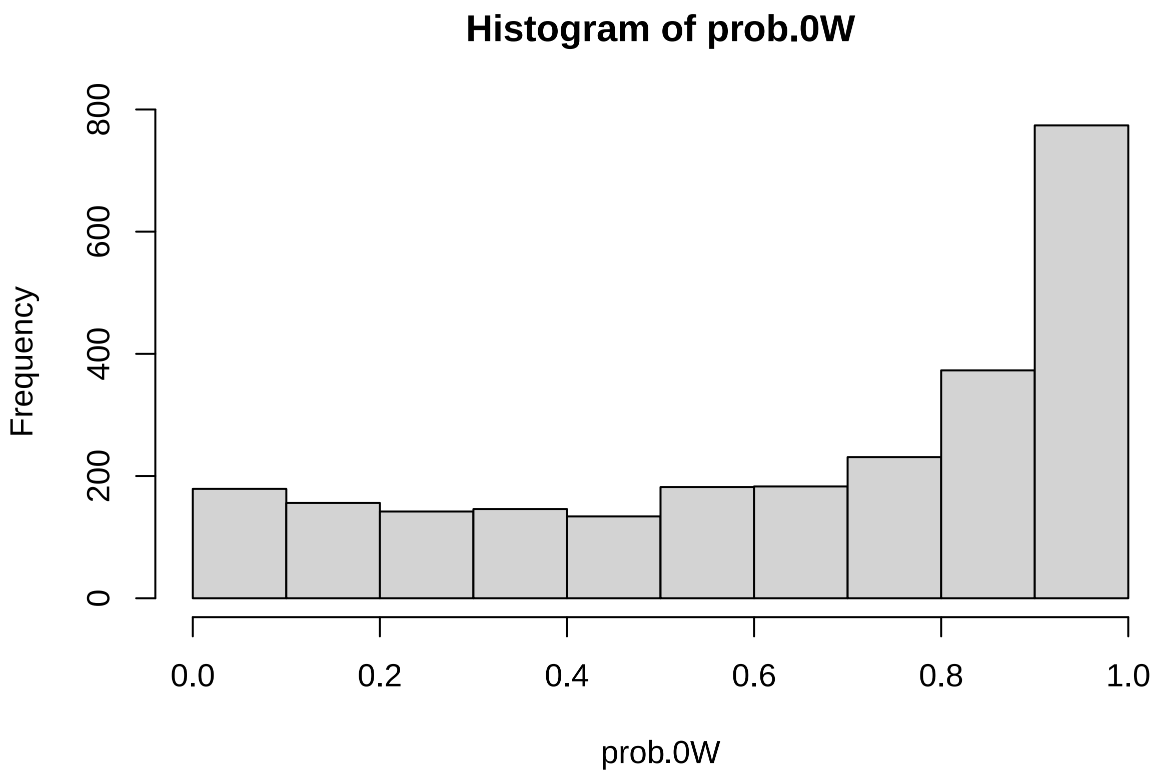
```
hist(prob.1W)
```



```
summary(prob.0W)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0006616 0.4015192 0.7611871 0.6548000 0.9294989 0.9997047
```

```
hist(prob.0W)
```



There are no evident positivity violations in our data set (as was already seen in question 1), although results indicate that given the covariates, it is more likely a participant does not have a dog. No major cause for concern.

5. Create the weights, and comment on the distribution of the weights.

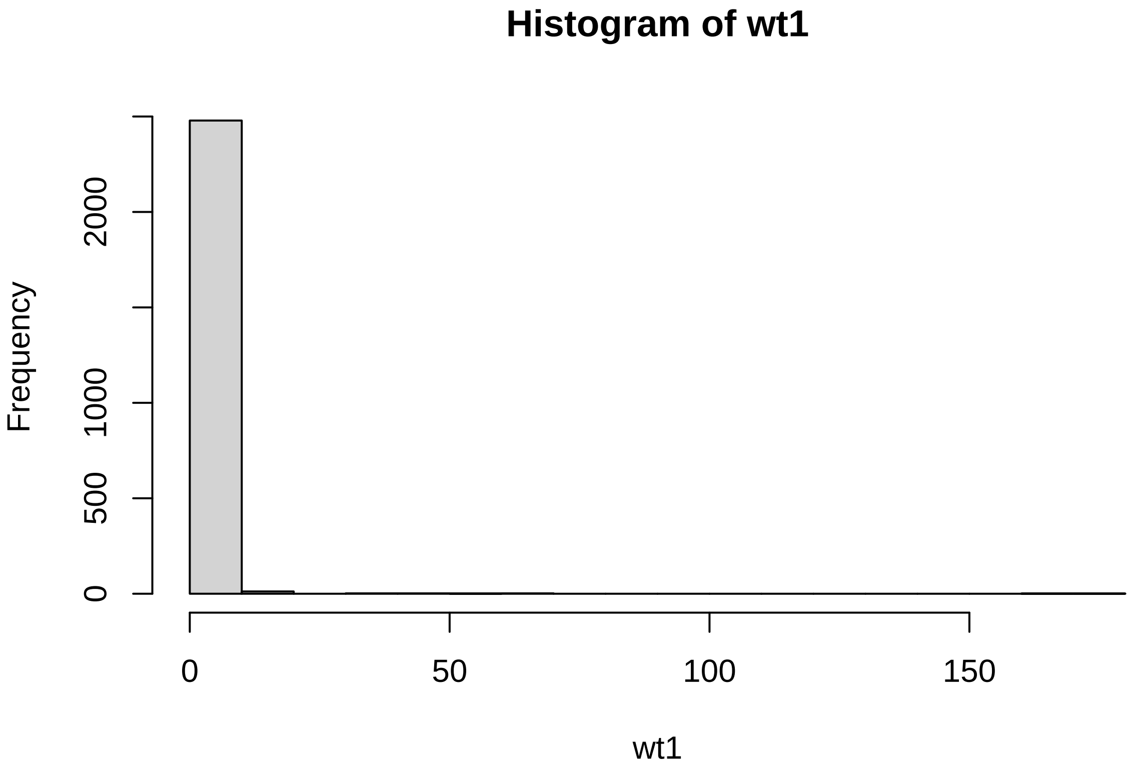
```
# Create the weights
wt1 <- as.numeric(ObsData$A==1)/prob.1W
wt0 <- as.numeric(ObsData$A==0)/prob.0W
summary(wt1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.046   1.188 177.719
```

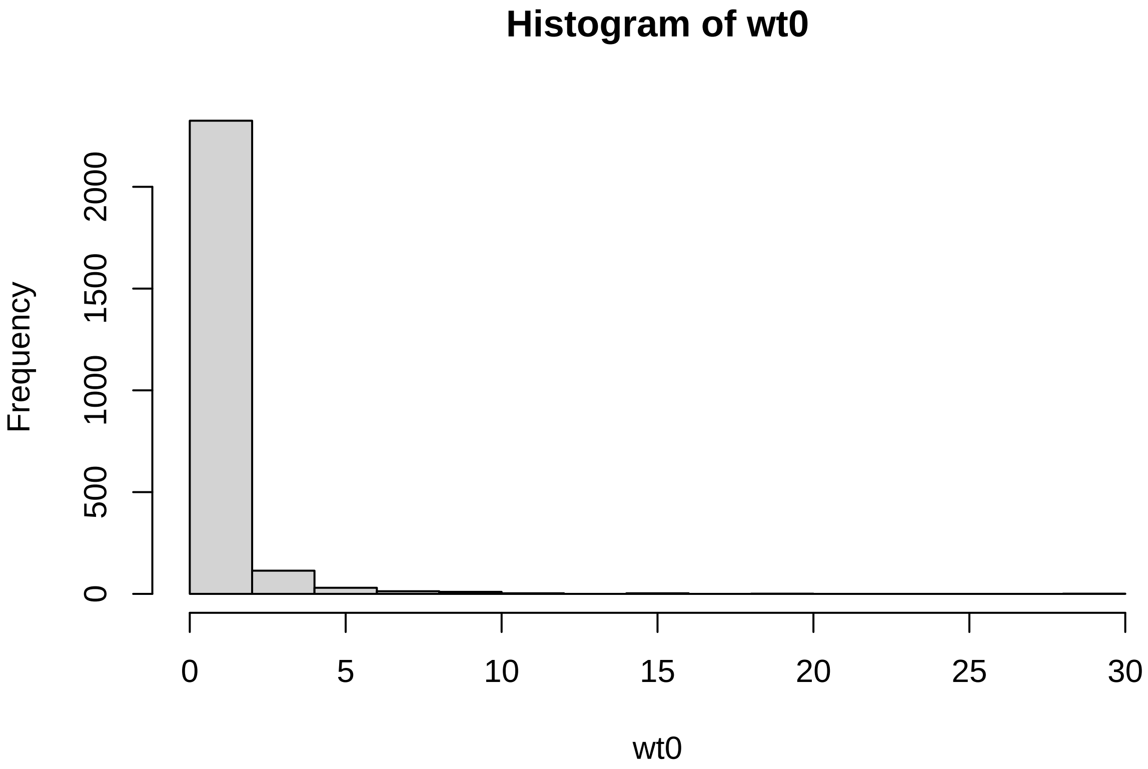
```
summary(wt0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.036   1.005   1.211  28.619
```

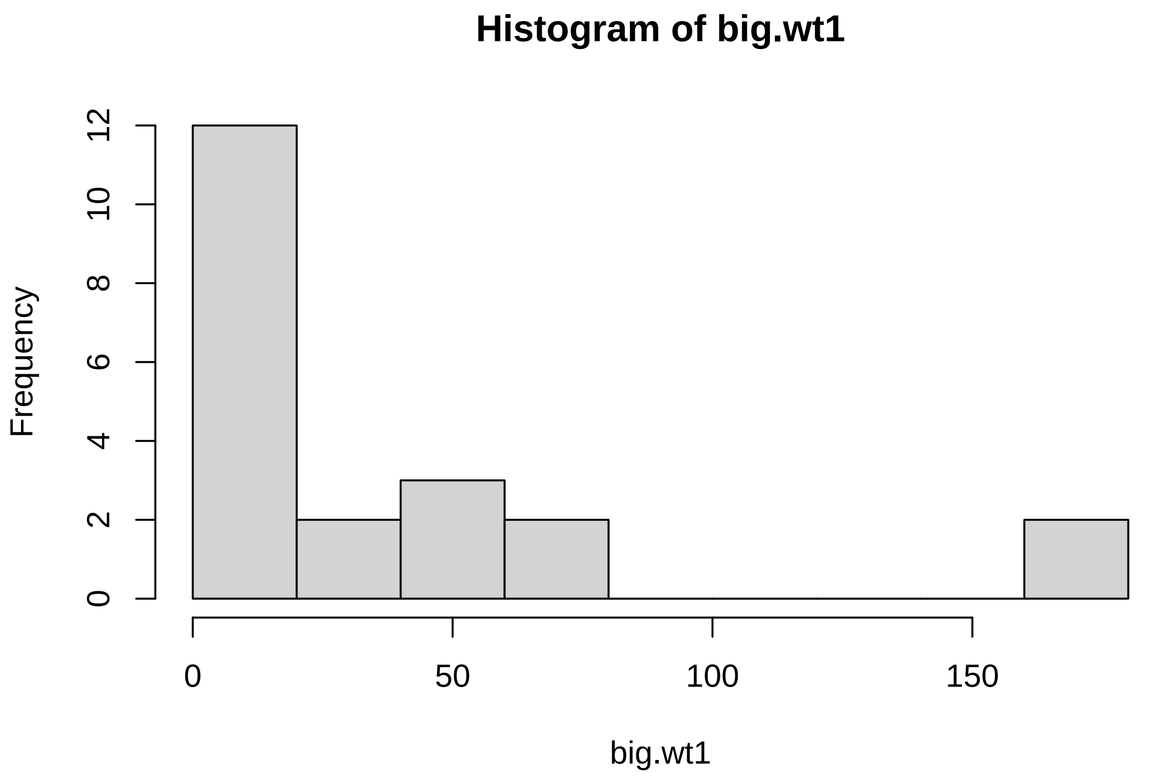
```
hist(wt1)
```



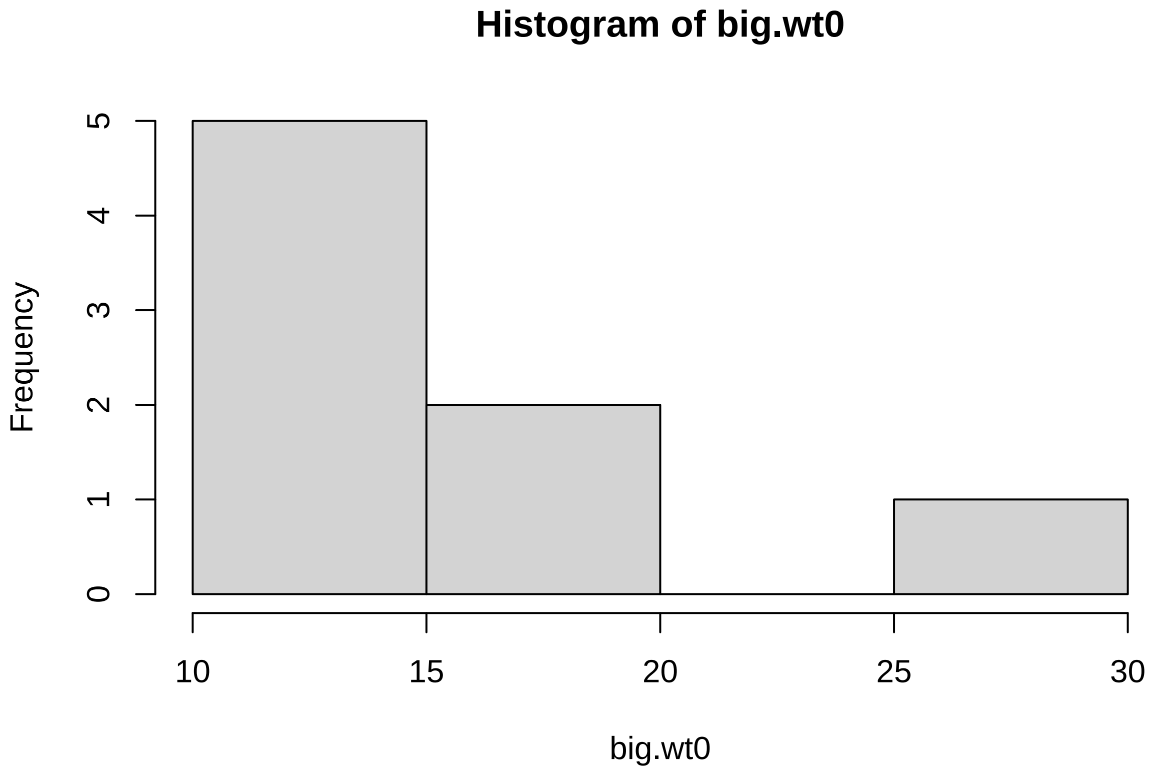
```
hist(wt0)
```



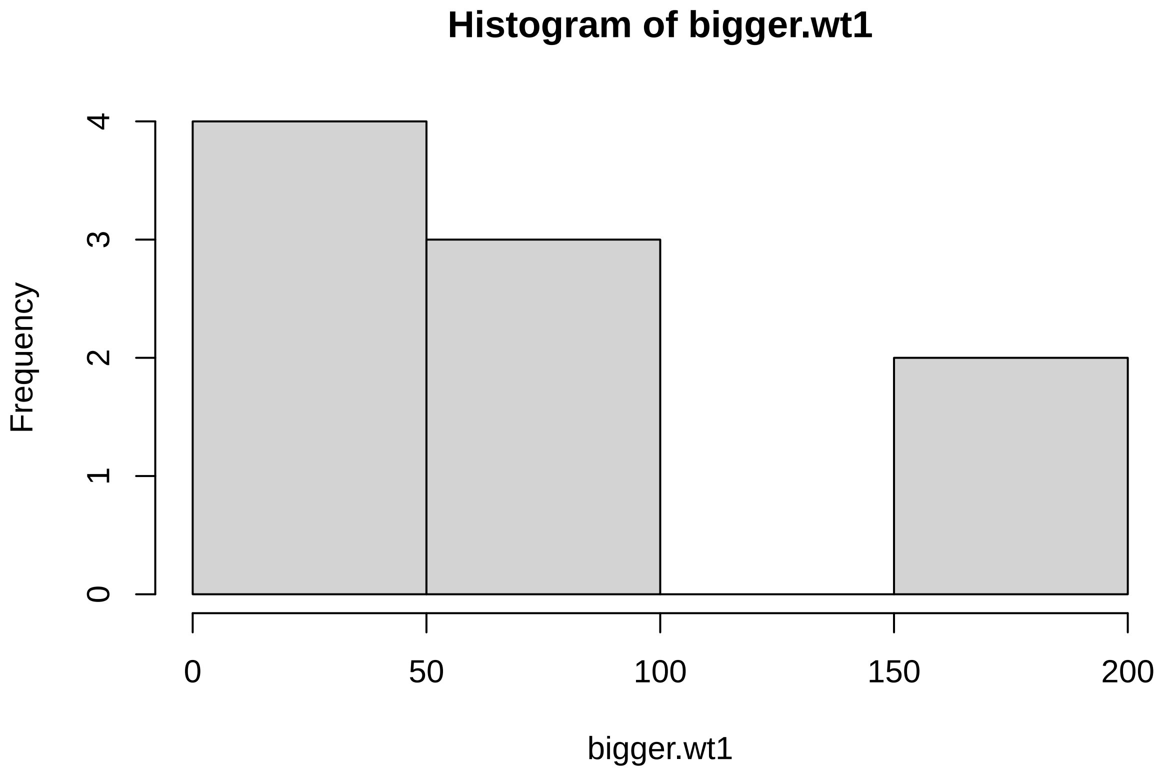
```
big.wt1 <- wt1[wt1 > 10]
big.wt0 <- wt0[wt0 > 10]
hist(big.wt1)
```



```
hist(big.wt0)
```



```
bigger.wt1 <- wt1[wt1 > 20]
hist(bigger.wt1)
```



The weights for participants without a dog seem mostly reasonable, with the highest weight being ~29, in a set of 2,500. Moreover, 75% of weights are under 1.22. There are only eight participants with weights greater than ten, with only one very high weight ~29. Nonetheless, for participants with a dog, there are more extreme weights. Although 75% of weights are under 1.19, there are twenty-one participants who have weights greater than ten, and nine with weights greater than twenty. At least one participant is being upweighted by ~178, which is very big, and may lead to poor finite sample performance.

6. Evaluate the IPTW estimand by taking the difference of the empirical means of the weighted outcomes:

$$\hat{\Psi}_{IPTW}(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1|W_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0|W_i)}Y_i$$

```
IPTW <- mean(wt1*ObsData$Y) - mean(wt0*ObsData$Y)
IPTW
```

```
## [1] -0.2416146
```

```
iptw <- mean((wt1-wt0)*ObsData$Y)
iptw
```

```
## [1] -0.2416146
```

The IPTW estimate of $\Psi(\mathbb{P}_0)$ is -24.2%, which can be interpreted as the estimated marginal differences in the mortality risk during 12 years, associated with having a dog, after controlling for the covariates.

7. Arbitrarily truncate the weights at 10 and re-evaluate the IPTW estimand.

```
# I already found that there are 21 weights under the exposure, and 8 weights under no exposure, which
wt1.trunc <- wt1
wt1.trunc[wt1.trunc > 10] <- 10
wt0.trunc <- wt0
wt0.trunc[wt0.trunc > 10] <- 10
# evaluate the IPTW estimand with the truncated weights
mean(wt1.trunc*ObsData$Y) - mean(wt0.trunc*ObsData$Y)
```

```
## [1] -0.3195715
```

The IPTW estimate of $\Psi(\mathbb{P}_0)$ is now -32.0%. By bounding the predicted probabilities, our estimator of the propensity score $\mathbb{P}_0(A = 1|W)$ is not consistent, and thus the IPTW will be biased.

8. Implement the stabilized IPTW estimator (a.k.a. the modified Horvitz-Thompson estimator):

$$\hat{\Psi}_{St.IPTW}(\mathbb{P}_n) = \frac{\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1|W_i)}Y_i}{\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1|W_i)}} - \frac{\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0|W_i)}Y_i}{\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0|W_i)}}$$

```
# Stabilized IPTW estimator - Modified Horvitz-Thompson estimator
mean(wt1*ObsData$Y)/mean(wt1) - mean(wt0*ObsData$Y)/mean(wt0)
```

```
## [1] -0.244704
```

```
# this is equivalent to
sum(wt1*ObsData$Y)/sum(wt1) - sum(wt0*ObsData$Y)/sum(wt0)
```

```
## [1] -0.244704
```

The stabilized IPTW estimate of $\Psi(\mathbb{P}_0)$ is now -24.5%.

9. For comparision, also implement the unadjusted estimator.

$$\begin{aligned}\hat{\Psi}_{unadj}(\mathbb{P}_n) &= \hat{\mathbb{E}}(Y|A=1) - \hat{\mathbb{E}}(Y|A=0) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0)}Y_i\end{aligned}$$

```
ttx1 <- filter(ObsData, A == 1)
ttx0 <- filter(ObsData, A == 0)
unadj <- mean(ttx1$Y) - mean(ttx0$Y)
unadj
```

```
## [1] -0.4105452
```

The unadjusted estimator is -41.1%, which is considerably larger in magnitude than what was estimated with IPTW.

10. *Bonus:* Implement a simple substitution estimator (a.k.a. parameteric G-computation) of $\Psi(\mathbb{P}_0)$ using the following parametric regression to estimate $\mathbb{E}_0(Y|A, W1, W2, W3, W4)$:

$$\mathbb{E}(Y|A, W1, W2, W3, W4) = logit^{-1}[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4 + \beta_5 A]$$

```
# Estimate the conditional mean of Y given the treatment A and covariates W
reg.model<- glm(Y ~ A + W1 + W2 + W3 + W4, family="binomial", data=ObsData)
reg.model
```

```
##
## Call:  glm(formula = Y ~ A + W1 + W2 + W3 + W4, family = "binomial",
##     data = ObsData)
##
## Coefficients:
## (Intercept)            A           W1           W2           W3           W4
##    -3.14223     -2.81428     -2.14327      0.05248     -0.99526     -1.01621
##
## Degrees of Freedom: 2499 Total (i.e. Null);  2494 Residual
## Null Deviance:       3020
## Residual Deviance: 1777  AIC: 1789
```

```
# Copy the original dataset O into two new dataframes txt and control
txt <- control <- ObsData
# set A=1 in the txt dataframe and A=0 in control dataframe
txt$A <- 1
control$A <- 0
# Predict the mean outcome for each individual in the sample under the treatment
predictY.txt <- predict(reg.model, newdata = txt, type='response')
```

```
# Predict the mean outcome for each individual in the sample under the control
predictY.control <- predict(reg.model, newdata = control, type='response')
# Observe results
head(cbind(ObsData, predictY.txt, predictY.control))
```

```
##   W1 W2    W3    W4 A Y predictY.txt predictY.control
## 1  1 67 -0.95 -2.81 0 1  0.313700036       0.88405532
## 2  1 52 -0.06  0.35 0 0  0.003446294       0.05454091
## 3  1 73  0.92 -0.60 1 0  0.010201439       0.14670411
## 4  0 62 -1.56  0.64 0 1  0.141755515       0.73370432
## 5  0 62 -0.70  1.62 0 1  0.025268841       0.30189176
## 6  1 72  1.32 -0.34 0 1  0.005017639       0.07759499
```

```
tail(cbind(ObsData, predictY.txt, predictY.control))
```

```
##      W1 W2    W3    W4 A Y predictY.txt predictY.control
## 2495  0 55  0.03  0.82 1 0  0.019199323       0.24615781
## 2496  0 72 -0.21  3.16 1 0  0.005593914       0.08578807
## 2497  0 74 -0.43 -0.80 0 1  0.303141292       0.87888359
## 2498  0 62 -0.25 -0.05 0 1  0.082914338       0.60130154
## 2499  1 67  1.22 -1.48 1 0  0.013464488       0.18544862
## 2500  1 63  1.69 -0.14 1 0  0.001772600       0.02876945
```

```
# Take the mean of the predicted outcomes to average over the distribution of Ws
mean(predictY.txt - predictY.control)
```

```
## [1] -0.2875429
```

11. Comment on the results.

The estimated difference in the mortality risk during 12 years, associated with having a dog, averaged with respect to the distribution of the covariates, is -28.8%. This is similar to, although slightly larger in magnitude, than the difference found using IPTW.

3 Extensions to handle missingness

1. Import and explore the modified data set `RAssign3.missing.csv`.

```
# Read in data
MissData <- read.csv('RAssign3.missing.csv')
head(MissData)
```

```
##   W1 W2    W3    W4 A Delta Y
## 1  1 67 -0.95 -2.81 0     1 1
## 2  1 52 -0.06  0.35 0     1 0
## 3  1 73  0.92 -0.60 1     1 0
## 4  0 62 -1.56  0.64 0     1 1
## 5  0 62 -0.70  1.62 0     1 1
## 6  1 72  1.32 -0.34 0     1 1
```

```
summary(MissData)
```

```
##        W1               W2              W3                  W4          
##  Min.   :0.0000   Min.   :50.00   Min.   :-3.69000   Min.   :-3.49000  
##  1st Qu.:0.0000   1st Qu.:56.00   1st Qu.:-0.71250   1st Qu.:-0.66000  
##  Median :1.0000   Median :62.00   Median :-0.02000   Median : 0.03000  
##  Mean   :0.5028   Mean   :62.07   Mean   :-0.03393   Mean   : 0.01644  
##  3rd Qu.:1.0000   3rd Qu.:68.00   3rd Qu.: 0.63000   3rd Qu.: 0.68000  
##  Max.   :1.0000   Max.   :75.00   Max.   : 3.78000   Max.   : 3.61000  
##        A              Delta              Y         
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :1.0000   Median :0.0000  
##  Mean   :0.3452   Mean   :0.8432   Mean   :0.2464  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
```

```
# 2,500 participants aged 50-75
# CV health range: -3.69 to 3.78, SES range: -3.49 to 3.61
table(MissData[,c('W1', 'W2', 'A')])
```

```
## , , A = 0
## 
##    W2
## W1   50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
##   0 15 27 40 27 23 32 36 32 33 28 36 30 29 37 26 24 24 27 31 28 33 33 31 26 24
##   1 19 43 42 29 36 25 44 47 43 29 28 42 35 31 31 23 32 42 35 35 31 37 33 41 41
##    W2
## W1   75
##   0 15
##   1 16
## 
## , , A = 1
## 
##    W2
## W1   50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
##   0  7 31 20 20 26 18 25 21 18 16 24 23 18 21 23 13 23 22 27 16 18 14 14 12 17
##   1  8 15 24 16 19 14 14 19 20 12  8 16 14 13 19 20 16 13 15 16  9  7 13 13  7
##    W2
## W1   75
##   0  9
##   1  7
```

2. Estimate the propensity score $\mathbb{P}_0(A = 1|W)$, which is the conditional probability of owning a dog, given the participant's characteristics. Use the following *a priori*-specified parametric regression model:

$$\mathbb{P}_0(A = 1|W) = logit^{-1}[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4]$$

```
# Run a logistic regression to estimate the treatment mechanism P(A|W)
prob.AW.reg <- glm(A ~ W1 +W2 + W3 + W4, family="binomial", data=MissData)
prob.AW.reg
```

```
## 
## Call:  glm(formula = A ~ W1 + W2 + W3 + W4, family = "binomial", data = MissData)
## 
## Coefficients:
## (Intercept)           W1           W2           W3           W4  
##      0.9736      -0.7395      -0.0272       1.5555       1.5503  
## 
## Degrees of Freedom: 2499 Total (i.e. Null);  2495 Residual
## Null Deviance:       3222 
## Residual Deviance: 2017  AIC: 2027
```

3. Predict each participant's probability of having and not having a dog, given their covariates: $\hat{\mathbb{P}}(A = 1|W_i)$ and $\hat{\mathbb{P}}(A = 0|W_i)$.

```
# Predicted probability of having a dog, given the obs cov P(A=1|W)
prob.1W <- predict(prob.AW.reg, type= "response")
# Predicted probability of not having a dog, given the obs cov P(A=0|W)
prob.0W <- 1 - prob.1W
```

4. Estimate the probability of being measured, given the exposure, sex, age, baseline cardiovascular health, and SES: $\mathbb{P}_0(\Delta = 1|A, W)$. Use the following *a priori*-specified parametric regression model:

$$\mathbb{P}_0(\Delta = 1|A, W1, W2) = logit^{-1}[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4 + \beta_5 A]$$

```
# Run a logistic regression to estimate the treatment mechanism P(D|A,W)
prob.DAW.reg <- glm(Delta ~ A + W1 + W2 + W3 + W4, family="binomial", data=MissData)
prob.DAW.reg
```

```
## 
## Call:  glm(formula = Delta ~ A + W1 + W2 + W3 + W4, family = "binomial", 
##     data = MissData)
## 
## Coefficients:
## (Intercept)            A           W1           W2           W3           W4  
##     2.88026      2.05728      0.78319     -0.02436     -1.83831      0.38926  
## 
## Degrees of Freedom: 2499 Total (i.e. Null);  2494 Residual
## Null Deviance:       2172 
## Residual Deviance: 1529  AIC: 1541
```

5. Predict each participant's probability of being measured, given their observed past $\hat{\mathbb{P}}(\Delta = 1|A_i, W_i)$.

```
# Predicted probability of having a dog, given the obs cov P(A=1|W)
prob.1AW <- predict(prob.DAW.reg, type= "response")
# Predicted probability of not having a dog, given the obs cov P(A=0|W)
prob.0AW <- 1 - prob.1AW
```

6. Create the weights - now accounting for confounding and incomplete measurement.

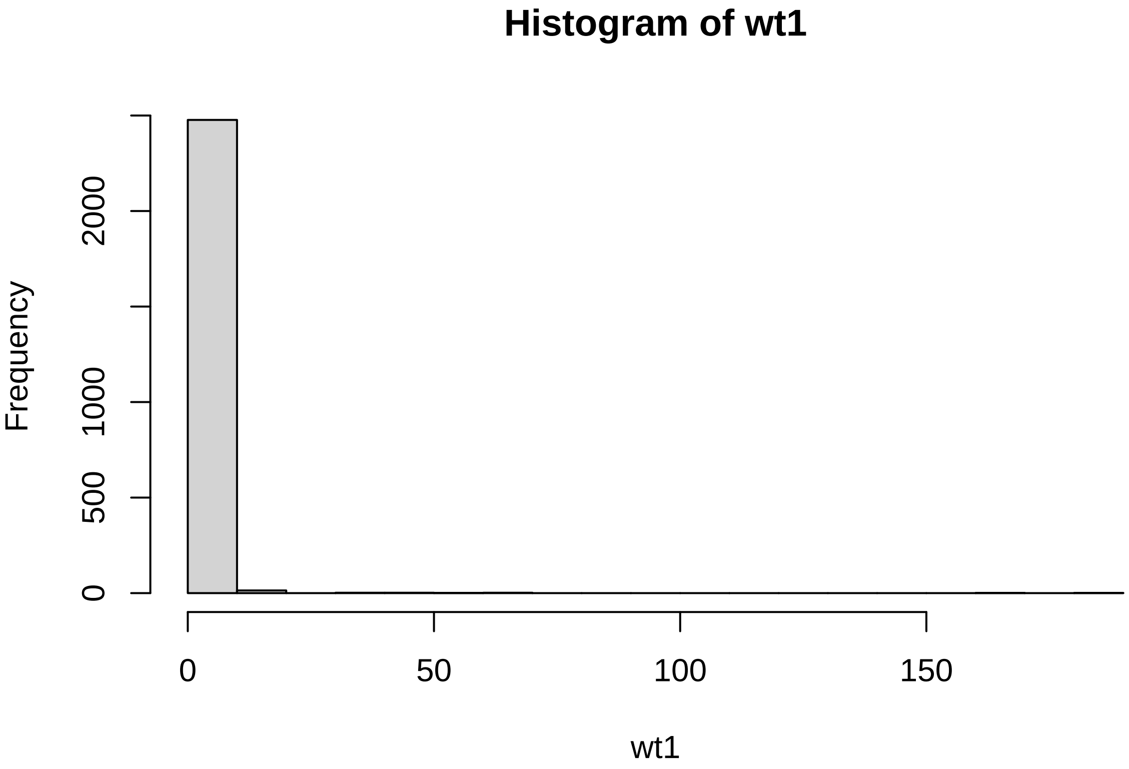
(a) Create a vector `wt1` with numerator as indicator of having a dog and being measured, and with denominator as the estimated probability of having a dog, given the adjustment set, times the estimated probability of being measured, given the observed past:

$$wt1_i = \frac{\mathbb{I}(A_i = 1, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 1|W_i) \times \hat{\mathbb{P}}(\Delta = 1|A_i, W_i)}$$

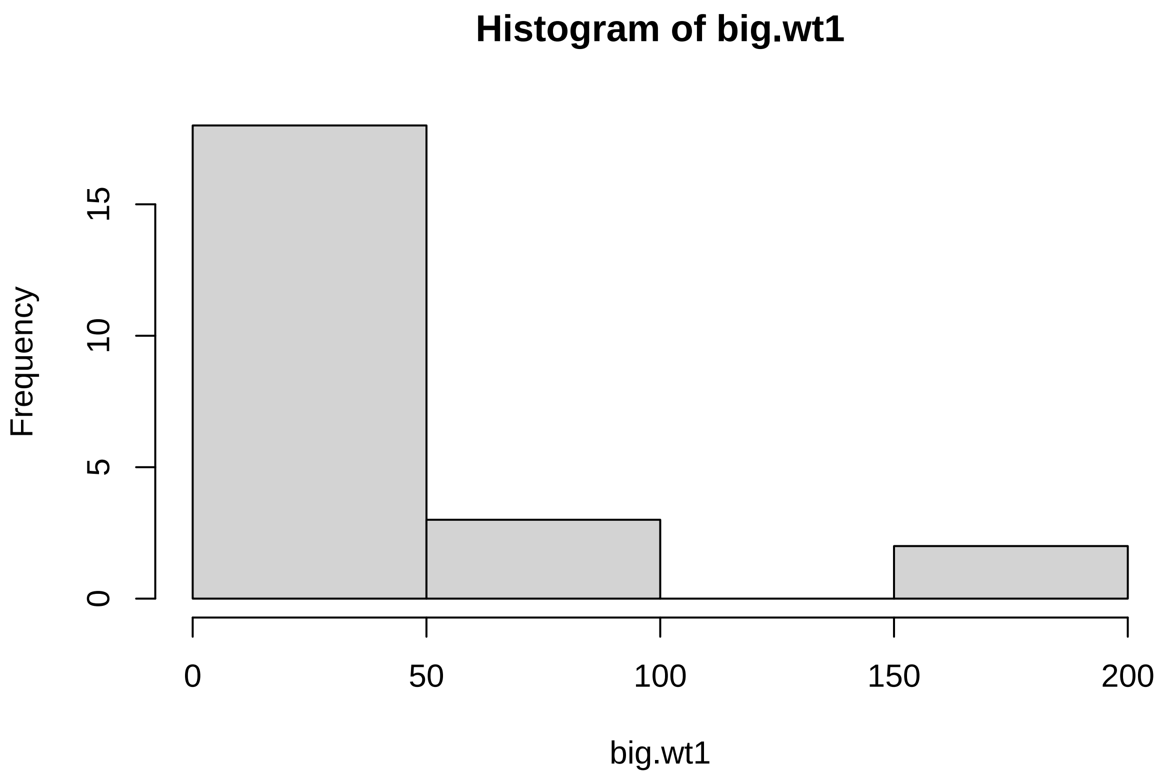
```
wt1 <- as.numeric(MissData$A==1 & MissData$Delta==1)/(prob.1W*prob.1AW)
summary(wt1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.051   1.256 180.654
```

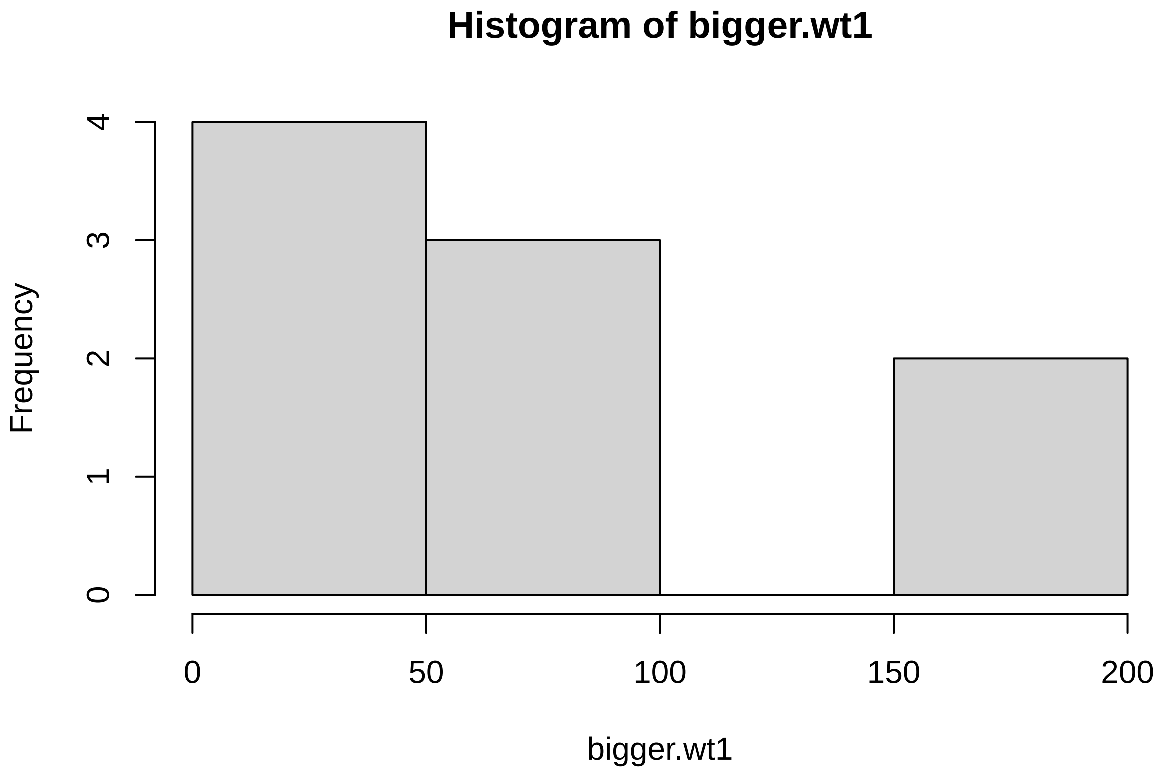
```
hist(wt1)
```



```
big.wt1 <- wt1[wt1 > 10]
hist(big.wt1)
```



```
bigger.wt1 <- wt1[wt1 > 20]
hist(bigger.wt1)
```



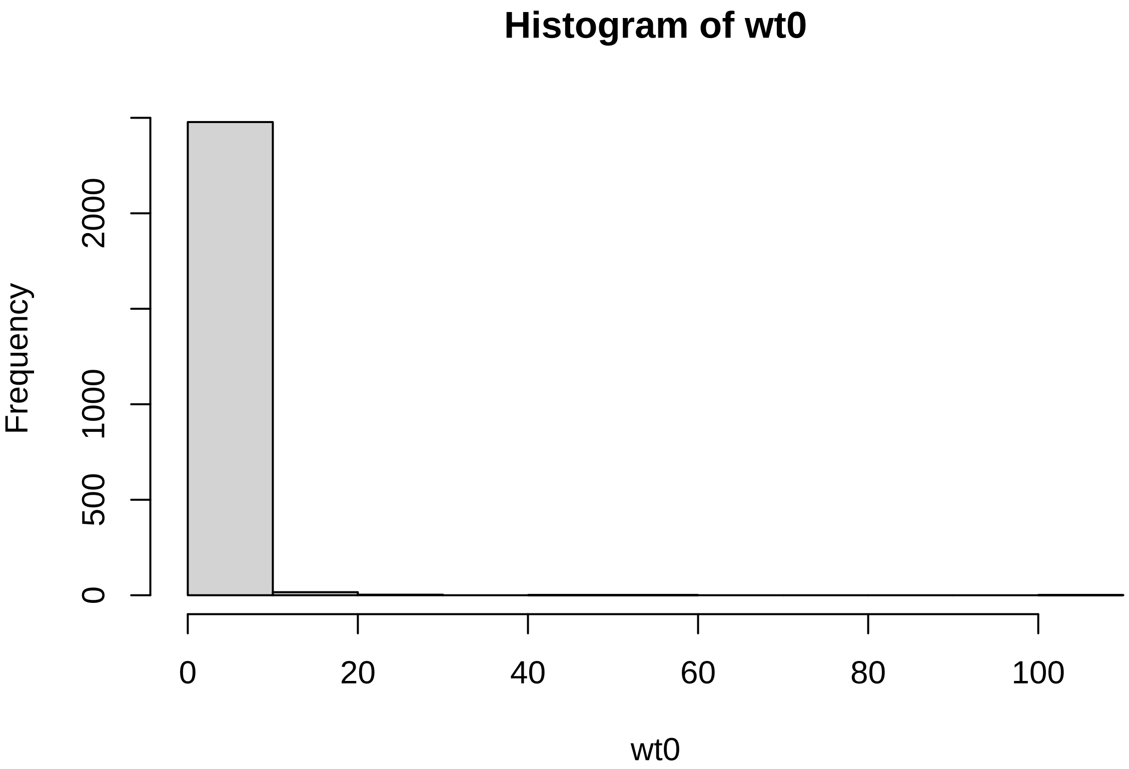
(b) Create a vector `wt0` with numerator as indicator of not having a dog and being measured, and with denominator as the estimated probability of not having a dog, given the adjustment set, times the estimated probability of being measured, given the observed past:

$$wt0_i = \frac{\mathbb{I}(A_i = 0, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 0|W_i) \times \hat{\mathbb{P}}(\Delta = 1|A_i, W_i)}$$

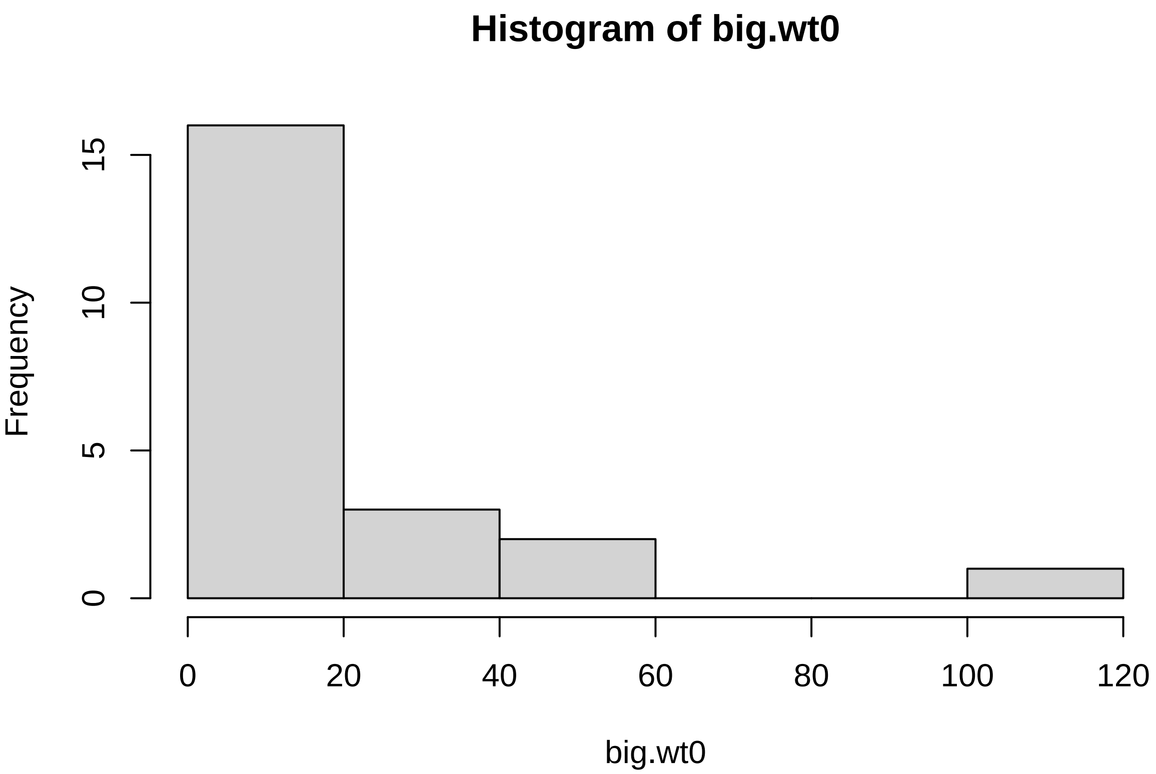
```
wt0 <- as.numeric(MissData$A==0 & MissData$Delta==1)/(prob.0W*prob.1AW)
summary(wt0)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    1.027    1.049    1.273  102.405
```

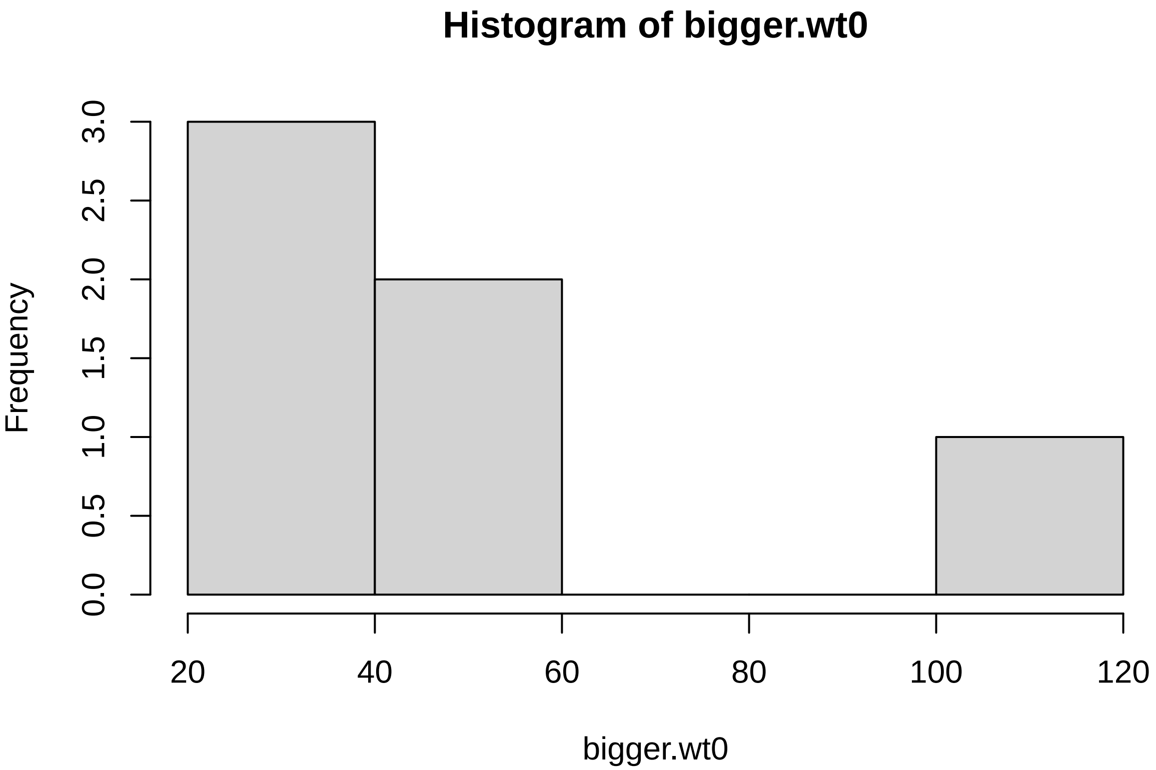
```
hist(wt0)
```



```
big.wt0 <- wt0[wt0 > 10]
hist(big.wt0)
```



```
bigger.wt0 <- wt0[wt0 > 20]
hist(bigger.wt0)
```



(c) Comment on the distribution of the weights.

The weights for participants without and with a dog are highly skewed to the right, with the highest weight being 102.4 for the former and 180.7 for the latter. The vast majority of weights are relatively small, with >75% being less than 1.3, and only 22 and 23 (without and with a dog) being greater than ten (and only 6 and 9 being greater than twenty). The large sample weights may lead to poor finite sample performance.

7. Evaluate the IPTW estimand by taking the difference of the empirical means of the weighted outcomes:

$$\hat{\Psi}_{IPTW}(\mathbb{P}_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 1, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 1|W_i)\hat{\mathbb{P}}(\Delta = 1|A_i, W_i)} Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 0, \Delta_i = 1)}{\hat{\mathbb{P}}(A = 0|W_i)\hat{\mathbb{P}}(\Delta = 1|A_i, W_i)} Y_i$$

```
IPTW <- mean(wt1*MissData$Y) - mean(wt0*MissData$Y)
IPTW
```

```
## [1] -0.2445379
```

```
iptw <- mean((wt1-wt0)*MissData$Y)
iptw
```

```
## [1] -0.2445379
```

The IPTW estimate of $\Psi(\mathbb{P}_0)$ is -24.5%, which can be interpreted as the estimated marginal differences in the mortality risk during 12 years, associated with having a dog and being measured, after controlling for the covariates.

8. Arbitrarily truncate the weights at 10 and evaluate the IPTW estimand.

```
# I already found that there are 23 weights under the exposure, and 22 weights under no exposure, which
wt1.trunc <- wt1
wt1.trunc[wt1.trunc > 10] <- 10
wt0.trunc <- wt0
wt0.trunc[wt0.trunc > 10] <- 10
# evaluate the IPTW estimand with the truncated weights
mean(wt1.trunc*ObsData$Y) - mean(wt0.trunc*ObsData$Y)
```

```
## [1] -0.3177973
```

The IPTW estimate of $\Psi(\mathbb{P}_0)$ is now -31.8%. By bounding the predicted probabilities, our estimator of the propensity score $\mathbb{P}_0(A = 1|W)$ is not consistent, and thus the IPTW will be biased.

9. Implement the stabilized IPTW estimator (a.k.a. the modified Horvitz-Thompson estimator).

```
# Stabilized IPTW estimator - Modified Horvitz-Thompson estimator
mean(wt1*MissData$Y)/mean(wt1) - mean(wt0*MissData$Y)/mean(wt0)
```

```
## [1] -0.2331161
```

```
# this is equivalent to
sum(wt1*MissData$Y)/sum(wt1) - sum(wt0*MissData$Y)/sum(wt0)
```

```
## [1] -0.2331161
```

The stabilized IPTW estimate of $\Psi(\mathbb{P}_0)$ is now -23.3%.

10. For comparison, also implement the unadjusted estimator.

$$\begin{aligned}\hat{\Psi}_{unadj}(\mathbb{P}_n) &= \hat{\mathbb{E}}(Y|A=1,\Delta=1) - \hat{\mathbb{E}}(Y|A=0,\Delta=1)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=1,\Delta_i=1)}{\hat{\mathbb{P}}(A=1,\Delta=1)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}(A_i=0,\Delta_i=1)}{\hat{\mathbb{P}}(A=0,\Delta=1)}Y_i\end{aligned}$$

```
ttx1 <- filter(MissData, A == 1)
ttx0 <- filter(MissData, A == 0)
unadj <- mean(ttx1$Y) - mean(ttx0$Y)
unadj
```

```
## [1] -0.3409057
```

The unadjusted estimator is -34.1%, which is considerably larger in magnitude than what was estimated with IPTW.

11. ***Bonus:*** **Implement a simple substitution estimator (a.k.a. parametric G-computation) of $\Psi(\mathbb{P}_{\nvdash})$, where in the first step the following parametric regression is used to estimate $\mathbb{E}_0(Y|A,W1,W2,W3,W4)$ -** ***among those who are measured:***

$$\mathbb{E}(Y|A,\Delta=1,W1,W2,W3,W4) = logit^{-1}[\beta_0 + \beta_1 W1 + \beta_2 W2 + \beta_3 W3 + \beta_4 W4 + \beta 5A]$$

```
outcomereg <- glm(Y   A + W1 + W2 + W3 + W4, data=ObsData[ObsData$Delta==1,], family="binomial")
```

```
# Estimate the conditional mean of Y given the treatment A and covariates W
outcomereg <- glm(Y ~ A + W1 + W2 + W3 + W4, data=MissData[MissData$Delta==1,], family="binomial")
outcomereg
```

```
##
## Call:  glm(formula = Y ~ A + W1 + W2 + W3 + W4, family = "binomial",
##     data = MissData[MissData$Delta == 1, ])
##
## Coefficients:
## (Intercept)            A           W1           W2           W3           W4
##    -3.45050     -2.76768     -2.17666      0.05816     -0.96136     -1.06888
##
## Degrees of Freedom: 2107 Total (i.e. Null);  2102 Residual
## Null Deviance:       2547
## Residual Deviance: 1472  AIC: 1484
```

```
# Copy the original dataset O into two new dataframes txt and control
txt <- control <- MissData
# set A=1 in the txt dataframe and A=0 in control dataframe
txt$A <- 1
control$A <- 0
# Predict the mean outcome for each individual in the sample under the treatment
predictY.txt <- predict(outcomereg, newdata = txt, type='response')
# Predict the mean outcome for each individual in the sample under the control
predictY.control <- predict(outcomereg, newdata = control, type='response')
# Observe results
head(cbind(MissData, predictY.txt, predictY.control))
```

```
##   W1 W2    W3    W4 A Delta Y predictY.txt predictY.control
## 1  1 67 -0.95 -2.81 0     1 1  0.358570754       0.89899507
## 2  1 52 -0.06  0.35 0     1 0  0.003377687       0.05119801
## 3  1 73  0.92 -0.60 1     1 0  0.012217387       0.16452725
## 4  0 62 -1.56  0.64 0     1 1  0.142228500       0.72527459
## 5  0 62 -0.70  1.62 0     1 1  0.024815148       0.28833347
## 6  1 72  1.32 -0.34 0     1 1  0.005980755       0.08742183
```

```
tail(cbind(MissData, predictY.txt, predictY.control))
```

```
##      W1 W2    W3    W4 A Delta Y predictY.txt predictY.control
## 2495  0 55  0.03  0.82 1     1 0  0.019360942       0.23916448
## 2496  0 72 -0.21  3.16 1     1 0  0.005449548       0.08024099
## 2497  0 74 -0.43 -0.80 0     1 1  0.343851529       0.89297574
## 2498  0 62 -0.25 -0.05 0     1 1  0.089581920       0.61038495
## 2499  1 67  1.22 -1.48 1     1 0  0.016474540       0.21054452
## 2500  1 63  1.69 -0.14 1     0 0  0.002013051       0.03111646
```

```
# Take the mean of the predicted outcomes to average over the distribution of Ws
mean(predictY.txt - predictY.control)
```

```
## [1] -0.2871167
```

12. Comment on the results.

The estimated difference in the mortality risk during 12 years, associated with having a dog, averaged with respect to the distribution of the covariates, is -28.7%. This is similar to, although slightly larger in magnitude, than the difference found using IPTW.

4 Improving IPTW

```
# SECTION 4 of R Assign 3

set.seed(1)

# The true value of the conditional mean outcome E_0[Y|A,W]
true.meanY.AW <- function(A,W){
  1000 + plogis(W*A)
}
# The true value of propensity score Pr(A=1|W)
true.prob.AW <- function(W){
  0.2 + 0.6*W
}

# A function which returns a data frame with n i.i.d. observations from P_0
gen.data <- function(n){
  # note this is a shortcut way of coding that skips generating the Us
  # first and then generating the endogenous variables deterministically
    W <- rbinom(n, 1, 1/2)
    A <- rbinom(n, 1, true.prob.AW(W=W))
    Y <- 1000 + rbinom(n, 1, true.meanY.AW(A=A,W=W) - 1000)
    return(data.frame(W=W,A=A,Y=Y))
}

# samples size
n<- 1000
# Number of Monte Carlo draws
R <- 2000
# Matrix of estimates from IPTW, modified Horvitz-Thompson, and my.est
est <- matrix(NA,nrow=R,ncol=3)
colnames(est) <- c('IPTW','Modifed HT','my.est')
for(r in 1:R){
    # Generate data with sample size
    ObsData <- gen.data(n)
    W <- ObsData$W
    A <- ObsData$A
    Y <- ObsData$Y
    # True propensity score P_0(A=1|W)
    pscore <- true.prob.AW(W=W)
    # IPTW estimate
    IPTW.est <- mean(A/pscore*Y)
    # Modified Horvitz-Thompson estimate
    HT.est <- mean(A/pscore*Y)/mean(A/pscore)
    # You should replace the NA below with your own estimator
    my.est <- mean(A/pscore*(Y-1000))+1000
    # Put the estimates into the est matrix
    est[r,] <- c(IPTW.est, HT.est, my.est)
}

# Calculate the true value of sum_w E[Y|A=1,W=w) P(W=w)
truth <- .5*true.meanY.AW(A=1, W=0) + .5*true.meanY.AW(A=1,W=1)
# note: we know P_0(W=1) = 0.5
truth
```

```
## [1] 1000.616
```

```
# Calculate the estimated bias, variance, and MSE
est.bias <- colMeans(est) - truth
est.var <- apply(est,2,var)
est.mse <- est.bias^2 + est.var

# The estimators have (estimated) bias:
est.bias
```

```
##         IPTW   Modifed HT       my.est
## 0.4990738357 0.0002180333 0.0003238357
```

```
# The estimators have (estimated) variance:
est.var
```

```
##         IPTW   Modifed HT       my.est
## 2.179920e+03 7.772966e-04 1.354703e-03
```

```
# The estimators have (estimated) MSE:
est.mse
```

```
##         IPTW   Modifed HT       my.est
## 2.180170e+03 7.773441e-04 1.354808e-03
```

1. Run the code given in Rassign3_modifiedIPTW.R and report how the standard IPTW and modified Horvitz-Thompson estimators perform in terms of bias, variance, and MSE over 2000 simulations each with sample size 1000. Which estimator would you use in practice?

```
# The estimators have (estimated) bias:
est.bias
```

```
##         IPTW   Modifed HT       my.est
## 0.4990738357 0.0002180333 0.0003238357
```

```
# The estimators have (estimated) variance:
est.var
```

```
##         IPTW   Modifed HT       my.est
## 2.179920e+03 7.772966e-04 1.354703e-03
```

```
# The estimators have (estimated) MSE:
est.mse
```

```
##         IPTW   Modifed HT       my.est
## 2.180170e+03 7.773441e-04 1.354808e-03
```

Clearly, the modofied HT estimator performs best with significantly smaller bias, variance, and MSE. It would make sense to use the Modified HT in this scenario even with the increased computational cost, since the performance is thousands of times better.

2. Look at the IPTW column in the `est` matrix. What do you notice about the IPTW estimates across these 2000 Monte Carlo draws? `Hint:` Recall the values that Y can take.

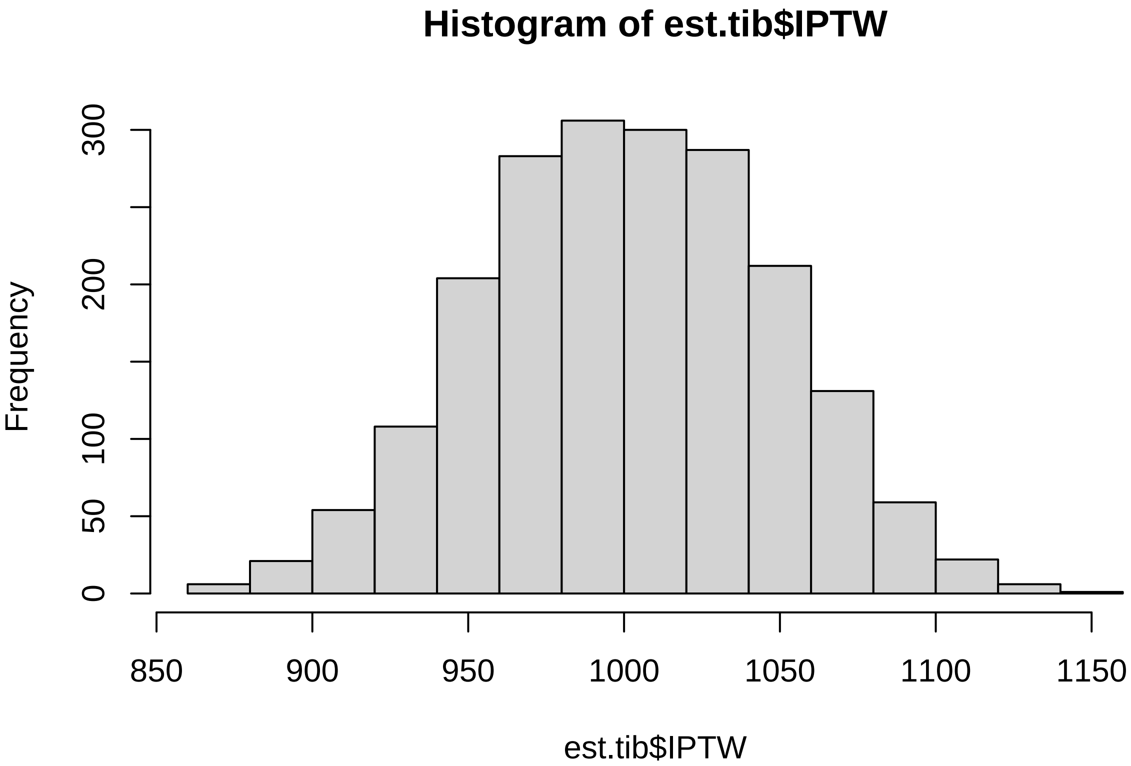
```
head(est, n=10)
```

```
##            IPTW Modifed HT   my.est
##  [1,]  963.1050   1000.629 1000.605
##  [2,] 1049.3750   1000.596 1000.625
##  [3,] 1024.3563   1000.592 1000.606
##  [4,]  965.6175   1000.640 1000.617
##  [5,]  989.4325   1000.690 1000.683
##  [6,] 1074.4150   1000.619 1000.665
##  [7,]  990.6275   1000.634 1000.628
##  [8,] 1016.8737   1000.614 1000.624
##  [9,]  979.3325   1000.595 1000.582
## [10,]  965.5925   1000.614 1000.592
```

```
est.tib <- as_tibble(est)
summary(est.tib)
```

```
##        IPTW           Modifed HT        my.est
##  Min.   : 863.0   Min.   :1001   Min.   :1000
##  1st Qu.: 968.1   1st Qu.:1001   1st Qu.:1001
##  Median :1000.6   Median :1001   Median :1001
##  Mean   :1001.1   Mean   :1001   Mean   :1001
##  3rd Qu.:1034.4   3rd Qu.:1001   3rd Qu.:1001
##  Max.   :1145.7   Max.   :1001   Max.   :1001
```

```
hist(est.tib$IPTW)
```



What is especially noteworthy is that the IPTW estimates are dispersed widely from 863 to 1146, even though Y can only take values of 1,000 and 1,001. In other words, the values generally don't make sense, as they should be binary, and instead are distributed in gaussian fashion.

3. What is the variance of a random variable X with $\mathbb{P}(X = 0) = 1/2$ and $\mathbb{P}(X = 1) = 1/2$?

The variance of the binomial distribution is $var = Np(1 - p)$ which in this case would be $0.25N$

4. What is the variance of a random variable $X2$ with $Pr(X2 = 0) = 1/2$ and $Pr(X2 = 1000) = 1/2$?

`Hint:` $X2 = 1000 \times X$

As explained by Josh in the lab, when a random variable is multiplied by a constant, the variance $Var(c{\cdot}X) = c^2Var(X)$, which in this case would mean the variance becomes $250,000N$

5. How are the above two calculations relevant to improving the IPTW estimator in this problem? We currently have an estimator that is an empirical mean of variables like those in Question 4. What transformations of the outcome Y would make your estimator behave more like an empirical mean of variables like those in Question 3?

My intuition is to simply subtract 1,000 from the outcome Y so that it is a simple binary variable (and then add it again to get the correct result).

6. *Graded leniently:* Write down an estimator $\hat{\Psi}_{my.est}$ which applies the ideas of the previous three questions into an estimator. There's no need to give the best possible estimator, but you should give an estimator that outperforms the IPTW estimator by a significant margin (i.e., does as or almost as well as the modified Horvitz-Thompson estimator in terms of bias/variance/MSE).

Using my intuition from the question above I created the following estimator:
my.est <- mean(A/pscore*(Y-1000))+1000
This gives an estimator that outperforms the IPTW by a significant margin and does almost as well as the modified Horvitz-Thompson estimator in terms of bias/variance/MSE.

7. *Graded leniently:* Code your estimator and replace the NA on the line `my.est = NA` with the estimator you defined in the previous question. Report the bias/variance/MSE of your estimator over the 2000 Monte Carlo draws.

```
est.bias
```

```
##         IPTW   Modifed HT       my.est
## 0.4990738357 0.0002180333 0.0003238357
```

```
est.var
```

```
##         IPTW   Modifed HT       my.est
## 2.179920e+03 7.772966e-04 1.354703e-03
```

```
est.mse
```

```
##         IPTW   Modifed HT       my.est
## 2.180170e+03 7.773441e-04 1.354808e-03
```

Thank you!