Homework 3

Anna Cook

10/13/2021

Number 1a

```r
library(readxl)
MI <- read_xlsx("MI_HW1.xlsx")

MI <- as.data.frame(lapply(MI, rep, MI$N))

fit1 <- glm(Statins ~ factor(RiskScore), data = MI, family = binomial(link = "logit"))

MI <- MI %>%
  mutate(prob = predict(fit1, type = 'response')) %>%
  mutate(invwt = Statins/prob + (1-Statins)/(1-prob))

fit2 <- glm(MI ~ Statins + factor(RiskScore), data = subset(MI, Statins == 0), weights = invwt, family =

pred_risk_no_statins <- mean(predict(fit2,
                                     newdata = MI, type = "response"))

print(pred_risk_no_statins)
```

```
## [1] 0.003999952
```

Number 1b

```r
fit3 <- glm(MI ~ Statins + factor(RiskScore), data = subset(MI, Statins == 1), weights = invwt,
            family = binomial(link = "logit"))

pred_risk_statins <- mean(predict(fit3,
                                  newdata = MI, type = "response"))

print(pred_risk_statins)
```

```
## [1] 0.002599897
```

Number 1c

0.003 - 0.004 = -0.001

Number 1d

This is exactly the same values that I found in Homework 2. This is because IPTW and conditioning arguments are two different methods of carrying out the same analysis, so it makes sense that the answers

would be the same, or at least very very close such that by the time you round to the nearest thousandth, the values are the same.

Number 1e

In number 9 of Homework 1, I found that if the risk score is not taken into account, statins have a negative effect (-0.0006). This finding is very biased, because the higher the risk score, the more likely someone is to take Statins in the first place, and also the more likely they are to have an MI.

-0.0006 - (-0.001) = 0.0004

Number 1f

We have no unmeasured confounding because the treatment assignment is independent of the possible outcomes given the risk score. In this case, we calculated the predicted risk had everyone received the same treatment, so the treatment assignment was all the same in each case, and the thing that varied was the risk score (X), and the outcomes (Y1 and Y0)

Number 2a

```
data <- read_excel("hw2_1000.xlsx")

# predict probability of treatment and calculate weights
logit_model <- glm(treatment ~ factor(creatininehigh) +
                    factor(infectiontype) + factor(Pitt_less4), data = data,
                family = binomial(link = "logit"))

summary(logit_model)
```

```
##
## Call:
## glm(formula = treatment ~ factor(creatininehigh) + factor(infectiontype) +
##     factor(Pitt_less4), family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3331  -0.7401  -0.6627   1.0903   1.8688
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.3055     0.1403  -9.304   <2e-16 ***
## factor(creatininehigh)1   1.5139     0.1426  10.617   <2e-16 ***
## factor(infectiontype)2    0.1211     0.2092   0.579    0.563
## factor(infectiontype)3    0.1504     0.1554   0.968    0.333
## factor(Pitt_less4)1      -0.2491     0.1451  -1.717    0.086 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1284.7  on 999  degrees of freedom
## Residual deviance: 1161.3  on 995  degrees of freedom
## AIC: 1171.3
##
## Number of Fisher Scoring iterations: 4
```

```r
data <- data %>%
  mutate(prob = predict(logit_model, type = 'response')) %>%
  mutate(invwt = treatment/prob + (1-treatment)/(1-prob))

# use weights in logit model with only treatment = 1
iptw_model_1 <- glm(hospitaldeath ~ treatment + factor(creatininehigh) +
                      factor(infectiontype) + factor(Pitt_less4), data = subset(data, treatment == 1),
                    weights = invwt, family = binomial(link = "logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
pred_risk_cazavi <- mean(predict(iptw_model_1,
                                 newdata = data, type = "response"))
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
print(pred_risk_cazavi)
```

```
## [1] 0.08682059
```

Number 2b

```r
# use weights in logit model with only treatment = 0
iptw_model_0 <- glm(hospitaldeath ~ treatment + factor(creatininehigh) + factor(infectiontype) +
                      factor(Pitt_less4), data = subset(data, treatment == 0), weights = invwt,
                    family = binomial(link = "logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
pred_risk_colistin <- mean(predict(iptw_model_0,
                                   newdata = data, type = "response"))
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
print(pred_risk_colistin)
```

```
## [1] 0.3569986
```

Number 2c

In parts 2a and 2b, we made the assumptions of consistency and asymptotic normality.

Number 2d

0.087 - 0.357 = -0.27

Number 2e

The results for parts a, b, and d are almost exactly the same as in homework 2. This makes sense because IPTW and conditioning arguments are two different methods for carrying out the same analysis.

Number 2f

Because the two different methods gave almost the exact same results, I would guess that means that it isn't super sensitive to model specification. However, I did use the same predictors in each model, with no

interaction terms. So I think it would be interesting to test out what the results would be like, and how they might differ, if included one or more interactions in the model.