Homework 2

Anna Cook

10/5/2021

Number 1a

Probability of MI if none had used statins = 0.004 (see attached sheet for work)

Number 1b

Probability of MI if all had used statins = 0.003 (see attached sheet for work)

Number 1c

Average effect = 0.004 - 0.003 = 0.001

People are .1% less likely to experience MI if they use statins, on average

Number 1d

This roughly matches what I had found in Homework 1, which was that statins do help, but there seems to be a pretty small difference in the low risk men. The high risk men showed a bigger difference.

Number 1e

In number 9, I concluded that it is very important to know the risk status in determining whether statins help or not. Low risk men are not very likely to have MI regardless of whether they take statins or not, so if these men are included in the analysis, there may be a bias which leans toward indicating that statins do not help as much. But if we only look at high risk men, it appears that statins make a much bigger difference. Here, I found the average effect of statins is 0.001, but in question 9 on homework 1, I found that if you ignore risk status, the effect of statins is -0.001. So this is a difference of 0.002.

Number 1f

I do think that the results are sensitive to model specification. As we saw in the previous homework, including or not including interaction terms can change whether the predictors are statistically significant or not. We also know that risk plays a role in how likely the statins are to make a difference, so if we did not include risk as a baseline covariate, our predictions would be different.

Number 2a

```
hw2_2 <- read_excel("hw2_1000.xlsx")
hw2_2a <- read_excel("hw2_1000.xlsx")

fit <- glm(factor(hospitaldeath) ~ factor(creatininehigh) + factor(infectiontype) + factor(Pitt_less4),

summary(fit)
```

```
##
## Call:
## glm(formula = factor(hospitaldeath) ~ factor(creatininehigh) +
##     factor(infectiontype) + factor(Pitt_less4), family = binomial(link = "logit"),
##     data = subset(hw2_2a, treatment == 1))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9035  -0.6048  -0.4361  -0.2387   2.6730
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.2536     0.5533  -5.881 4.09e-09 ***
## factor(creatininehigh)1    1.9380     0.5422   3.574 0.000351 ***
## factor(infectiontype)2     0.6305     0.4397   1.434 0.151558
## factor(infectiontype)3    -0.2904     0.3962  -0.733 0.463670
## factor(Pitt_less4)1       -0.6991     0.3816  -1.832 0.066985 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 254.78  on 341  degrees of freedom
## Residual deviance: 228.45  on 337  degrees of freedom
## AIC: 238.45
##
## Number of Fisher Scoring iterations: 6
```

```
hw2_2a$pred <- predict(fit, newdata = hw2_2)
hw2_2a$prob <- invlogit(hw2_2a$pred)
hw2_2a$pred_hosp_death <- ifelse(hw2_2a$prob >= 0.5, 1, 0)

sum(hw2_2a$pred_hosp_death) / nrow(hw2_2a)
```

```
## [1] 0
```

Number 2b

You can also embed plots, for example:

```
hw2_2b <- read_excel("hw2_1000.xlsx")

fit2 <- glm(factor(hospitaldeath) ~ factor(creatininehigh) + factor(infectiontype) + factor(Pitt_less4)



hw2_2b$pred <- predict(fit2, newdata = hw2_2)
hw2_2b$prob <- invlogit(hw2_2b$pred)
hw2_2b$pred_hosp_death <- ifelse(hw2_2b$prob >= 0.5, 1, 0)

sum(hw2_2b$pred_hosp_death) / nrow(hw2_2b)
```

```
## [1] 0.258
```

Number 2c

We are assuming that there are no unmeasured confounders; that is, that the treatment assignment is independent of the possible outcomes, given the baseline coviariates. We are also assuming consistency (that if A = 1, we observe Y = Y(1), and if A = 0, we observe Y = Y(0)). We also assume positivity, which says that the probabilities we are working with are positive.

Number 2d

0 - 0.26 = -0.26; caz-avi reduces changes of hospital death by 26%, on average

Number 2e

I do think the specific results may be sensitive to model specifications, although I think even if the model changed a bit, the caz-avi treatment would still yield better results than the other treatment. So the probabilities would likely vary if I included interactions or different combinations of variables in the model, but I think the punchline would still be the same.

Number 2f

I think there could be some confounding. First, I think it's suspicious that the probability of hospital death under treatment 1 was estimated to be 0. This makes me think that maybe patients who were expected to not die were given this treatment. Also, many more patients were given treatment 0 than treatment 1, and it's possible that this choice in treatment was partly determined by some unmeasured variable.

Number 2g

$E(Y \mid A = 1) = 0.123$

$E(Y \mid A = 0) = 0.312$

```
treat1 <- data.frame(hw2_2 %>% filter(hw2_2$treatment == 1))

print(sum(treat1$hospitaldeath)/nrow(treat1))
```

```
## [1] 0
```

```
treat0 <- data.frame(hw2_2 %>% filter(hw2_2$treatment == 0))

print(sum(treat0$hospitaldeath)/nrow(treat0))
```

```
## [1] 0
```

Number 2h

The estimates are biased because they are slightly different from the estimates I found had everyone been treated the same way. For treatment 1, the difference between the two estimates I found was 0.123 - 0 = 0.123. For treatment 0, the difference between the estimates is 0.312 - 0.258 = 0.054.