Assignment 2 A Matching Exercise

Husayn Jessa and Aidan Bodner

11/11/2022

```
library(here) # To set directory
```

```
## here() starts at /Users/aidanbodner/Documents/GitHub/HAD5744
```

```
library(haven) # To import data
library(dplyr) # For piping
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gtsummary) # For some balance tables
```

```
## #Uighur
```

```
library(MatchIt) # For matching
library(modelsummary) # For some balance tables
library(ggplot2) # To create plots
library(vtable) #To output summary tables
```

```
## Loading required package: kableExtra
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(purrr) # To output reference list
```

```
name <- Sys.info()
name[7]
```

```
##          user
## "aidanbodner"
```

```
here()
```

```
## [1] "/Users/aidanbodner/Documents/GitHub/HAD5744"
```

```
assign_data <- read_dta("/Users/aidanbodner/Documents/GitHub/HAD5744/Assignment 2/Dataset2a_Claims (1).d

assign_data %>%
  summarise_all(funs(sum(is.na(.)))) # Returns no NAs, can probably remove
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## i Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
## # A tibble: 1 x 14
##    year enrolid famid   oop   pay netpay  hdhp   age   sex famsize policyholder
##   <int>   <int> <int> <int> <int>  <int> <int> <int> <int>   <int>        <int>
## 1     0       0     0     0     0      0     0     0     0       0            0
## # ... with 3 more variables: num_hospitalizations <int>, num_scrips <int>,
## #   num_chronicconditions <int>
```

Question 1a

```
assign_data$sex <- as.factor(assign_data$sex) # Convert sex from a character to a factor
assign_data$sex <- as.numeric(assign_data$sex) # Convert sex from a character to a numeric


unmatched_table <- tbl_summary(data = assign_data, by = "hdhp",
            type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
```

```
                              all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")

unmatched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 227,517	**1**, N = 2,655	**p-value**
Age of Patient	48.29 (14.28)	48.19 (14.93)	0.7
sex			0.041
1	102,867.00 (45.21%)	1,148.00 (43.24%)	
2	124,650.00 (54.79%)	1,507.00 (56.76%)	
famsize	2.38 (1.41)	2.55 (1.53)	<0.001
policyholder			0.007
0	90,018.00 (39.57%)	1,119.00 (42.15%)	
1	137,499.00 (60.43%)	1,536.00 (57.85%)	
num_hospitalizations	1.49 (1.01)	1.46 (0.97)	0.2
num_hccs	1.96 (1.20)	1.95 (1.15)	0.5
num_scrips	11.89 (6.15)	12.56 (6.52)	<0.001

From the p-values obtained from performing the t-test, we can see that there is no statistical differences between people in HDHP compared to other individuals in terms of age, number of hospitalizations, and number of chronic conditions. However, the groups are not comparable in terms of sex,if they are a policyholder, family size, and number of prescriptions.

We might expect to see significant differences in the distribution of HDHP by family size as people with larger families may be incentivized to find innovative ways to save costs. In terms of the number of prescriptions, people who already need more prescriptions will likely enroll in a plan that allows them to save money on future prescriptions. In terms of sex, females are more likely to engage in care seeking or plan for a future medical event than males making it conceivable that they would be statistically likely to be in an HDHP. In terms of being a policy holder, the policy holder is someone who owns the insurance policy and as such are more inclined to use the insurance they have and as such when enrolled in an HDHP they take greater pride in using it.

Question 1b

```
one_b_reg_naive <- lm(pay ~ hdhp + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                  data = assign_data)

msummary(list("naive_regression" = one_b_reg_naive))
```

The results from (a) suggest that for characteristics such as age, number of hospitalizations, and number of chronic conditions, we can likely assume that since there is no statistical difference in the means, that the fit of the regression is fairly good. However, for characteristics such as sex, if they are a policyholder,

	naive_regression
(Intercept)	45 480.553
	(1534.878)
hdhp	608.807
	(2054.570)
age	−303.184
	(18.465)
sex	−13 141.403
	(449.069)
famsize	106.075
	(188.289)
policyholder	−743.930
	(505.578)
num_hospitalizations	38 793.507
	(224.936)
num_chronicconditions	−341.172
	(182.767)
num_scrips	1933.702
	(38.771)
Num.Obs.	230 172
R2	0.149
R2 Adj.	0.149
AIC	5 976 590.0
BIC	5 976 693.4
Log.Lik.	−2 988 284.991
F	5023.036
RMSE	105 229.19

family size, and number of prescriptions, there are statistical differences in the means indicating that the regression line likely is not able to fit the observations as well. The differences in these groups shown in (a) show more females than males, higher means family size, more likely to be a policyholder, and higher number of prescriptions for those who enroll in high deductibale health plans (HDHPs). Thus when looking at the reported regression coefficients they do not accurately show the true effect these variables have on the outcome variable as there is a statistically significant difference within these variables for those who are in a HDHP.

There is an underestimated effect for these coefficients as a result, as we expect people enrolled in HDHPs to spend more as they are more likely to engage in healthcare seeking due to a higher likelihood of being female, having higher average family sizes, being a policyholder, and having higher average number of prescriptions. We also expect that a HDHP exposes these participants to higher fees and thus through higher deductibles these groups will be expected to spend more on health services than those who are not enrolled in these forms of insurance plans. Thus these variables underestimate the true causal effect.

Question 1c

```
question1c_match <- matchit(hdhp ~ num_hospitalizations,
                            data = assign_data,
                            method = "exact")
summary(question1c_match)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations, data = assign_data,
##     method = "exact")
##
## Summary of Balance for All Data:
##                       Means Treated Means Control Std. Mean Diff. Var. Ratio
## num_hospitalizations        1.4637        1.4906         -0.0278     0.9219
##                       eCDF Mean eCDF Max
## num_hospitalizations     0.0016   0.0171
##
##
## Summary of Balance for Matched Data:
##                       Means Treated Means Control Std. Mean Diff. Var. Ratio
## num_hospitalizations        1.4637        1.4637               0     1.0004
##                       eCDF Mean eCDF Max Std. Pair Dist.
## num_hospitalizations         0        0               0
##
## Sample Sizes:
##                 Control Treated
## All           227517.     2655
## Matched (ESS) 226275.3    2655
## Matched       227429.     2655
## Unmatched         88.        0
## Discarded          0.        0
```

```
regdata <-  match.data(question1c_match)
```

	Unmatched Regression	Exact Matching Regression	Exact Matching Regression w/ Covariate
(Intercept)	45 480.553	89 347.731	108 811.629
	(1534.878)	(234.935)	(1579.900)
hdhp	608.807	1770.654	−1430.768
	(2054.570)	(2187.046)	(2178.290)
age	−303.184		−672.951
	(18.465)		(19.447)
sex	−13 141.403		−16 733.179
	(449.069)		(475.684)
famsize	106.075		−58.665
	(188.289)		(199.676)
policyholder	−743.930		−3434.044
	(505.578)		(535.854)
num_hospitalizations	38 793.507		
	(224.936)		
num_chronicconditions	−341.172		−330.564
	(182.767)		(193.801)
num_scrips	1933.702		3597.377
	(38.771)		(39.797)
Num.Obs.	230 172	230 084	230 084
R2	0.149	0.000	0.038
R2 Adj.	0.149	0.000	0.038
AIC	5 976 590.0	6 003 827.3	6 001 219.0
BIC	5 976 693.4	6 003 858.3	6 001 312.1
Log.Lik.	−2 988 284.991	−3 001 910.651	−3 000 600.491
F	5023.036	0.655	1306.641
RMSE	105 229.19	113 767.95	111 567.79

```
question1c_unmatched_reg <- lm(pay ~ hdhp + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips, data = assign_data)


question1c_reg <- lm(pay ~ hdhp, data = regdata, weights = weights) #using matched data with no covaria

question1c_reg_with_covariates <- lm(pay ~ hdhp + age + sex + famsize + policyholder  +
                               + num_chronicconditions + num_scrips , data = regdata)
#^using matched data with covariates, did not include num_hospitalizations because it was matched on al

msummary(list("Unmatched Regression" = question1c_unmatched_reg,
              "Exact Matching Regression" = question1c_reg,
              "Exact Matching Regression w/ Covariates" = question1c_reg_with_covariates),
         )
```

```
onlyhospitalizations_matched_balance_table <- tbl_summary(data = regdata, by = "hdhp",
            type = list(num_hospitalizations ~ "continuous"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c(
                  "num_hospitalizations")) %>%
  add_p(test = everything() ~ "t.test")
```

```
onlyhospitalizations_matched_balance_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 227,429	**1**, N = 2,655	**p-value**
num_hospitalizations	1.49 (0.99)	1.46 (0.97)	0.2

```
qa_qb_balance_tables <- tbl_merge(tbls = list(unmatched_table,
                                              onlyhospitalizations_matched_balance_table),
                                  tab_spanner = c("UnMatched Balance Table",
                                                  "Only Hospitalizations Balance Table"))

qa_qb_balance_tables
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 227,517	**1**, N = 2,655	**p-value**	**0**, N = 227,429	**1**, N = 2,655
Age of Patient	48.29 (14.28)	48.19 (14.93)	0.7		
sex			0.041		
1	102,867.00 (45.21%)	1,148.00 (43.24%)			
2	124,650.00 (54.79%)	1,507.00 (56.76%)			
famsize	2.38 (1.41)	2.55 (1.53)	<0.001		
policyholder			0.007		
0	90,018.00 (39.57%)	1,119.00 (42.15%)			
1	137,499.00 (60.43%)	1,536.00 (57.85%)			
num_hospitalizations	1.49 (1.01)	1.46 (0.97)	0.2	1.49 (0.99)	1.46 (0.97)
num_hccs	1.96 (1.20)	1.95 (1.15)	0.5		
num_scrips	11.89 (6.15)	12.56 (6.52)	<0.001		

The back door being closed by matching the HDHP group with the non-HDHP group is the back door associated with inpatient hospitalizations. Through matching both groups on the # of hospitalizations, it is possible to say that both groups are similar with regards to this variable. This means that between the two groups there is no variations with regards to hospitalizations and as a result, we can more accurately depict the relationship between HDHP and spending. With regards to the quality of our match, we can note that there were not many members of the control group that went unmatched (88 people) and there were 0 people discarded from the match. As we noted from 1a. their are no significant differences between the comparison groups with the number of hospitalizations and thus as expected there were many matches for the treatment group. The quality of this match is poor because most of the control matched to the treatment group however we only matched on one variable so there are many back doors between pay and HDHP still open.

For regression analyses, we compare an unmatched sample with all covariates, a matched sample with no covariates, and a matched sample with all covariates.

The regression results show that after matching on inpatient hospitalizations occurred (but not including all covariates), the regression coefficient of HDHP when no covariates were added to the regression was 1770.65. Thus, for individuals who are enrolled in the HDHP they had an increased spending of $1770.65 in comparison to those who were not enrolled in an HDHP. However, this regression does not tell the whole story as it does not include the other covariates that effect how much is being spent on health care.

When the matched regression included the other covariates a large change in the regression coefficient was observed. The regression value for HDHP became -1430.77 when all the other covariates were controlled for. These results were surprising as it suggests that superutilizers enrolled in an HDHP spend less in comparison to those who were not in an HDHP. It appears that the number of hospitalizations play a large role in how much a person in an HDHP will spend on healthcare. When all the variation in the number of hospitilizations is removed (through matching) and the other covariates are controlled for (through regression) we can see that superutilizers enrolled in HDHP spend less. This suggests to us that hospitalizations contribute to large amounts of spending for people in HDHP because they perform a larger array of services and as a result the deductibale a person pays at the hospital is much larger than anywhere else. When we compare the results of this regression with the regression that used the unmatched data and controlled for all the variables, including number of hospitalizations, we see that the regression coefficient for HDHP is 608.81 and the regression coefficient for number of hospitalizations is 38793.51. This indicates that belonging to a HDHP contributes to higher spending and number of hospitalizations has a very large effect on spending. When the variation in hospitalizations is removed by the match it thus explains why the regression coefficient for HDHP becomes negative as number of hospitalizations had a large effect on how much was being spent on healthcare.

Question 1d

```
question1d_match <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                  replace = TRUE,
                  distance = "scaled_euclidean",
                  verbose = TRUE)
```

```
## Nearest neighbor matching...
## Calculating matching weights... Done.
```

```
summary(question1d_match)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
##     num_scrips, data = assign_data, method = "nearest", distance = "scaled_euclidean",
##     replace = TRUE, verbose = TRUE)
##
## Summary of Balance for All Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## num_hospitalizations           1.4637        1.4906         -0.0278     0.9219
## age                           48.1936       48.2925         -0.0066     1.0932
## sex                            1.5676        1.5479          0.0398     0.9912
## famsize                        2.5454        2.3845          0.1053     1.1728
## policyholder                   0.5785        0.6043         -0.0523          .
## num_chronicconditions          1.9484        1.9620         -0.0118     0.9178
## num_scrips                    12.5593       11.8947          0.1020     1.1223
##                         eCDF Mean eCDF Max
## num_hospitalizations       0.0016   0.0171
```

```
## age                       0.0089   0.0229
## sex                       0.0099   0.0197
## famsize                   0.0115   0.0500
## policyholder              0.0258   0.0258
## num_chronicconditions     0.0027   0.0055
## num_scrips                0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## num_hospitalizations           1.4637        1.4637          0.0000     1.0016
## age                           48.1936       48.2316         -0.0025     1.0124
## sex                            1.5676        1.5676          0.0000     1.0000
## famsize                        2.5454        2.5394          0.0039     1.0235
## policyholder                   0.5785        0.5785          0.0000          .
## num_chronicconditions          1.9484        1.9476          0.0007     1.0052
## num_scrips                    12.5593       12.5424          0.0026     1.0151
##                         eCDF Mean eCDF Max Std. Pair Dist.
## num_hospitalizations       0.0000   0.0004          0.0008
## age                        0.0018   0.0045          0.0330
## sex                        0.0000   0.0000          0.0000
## famsize                    0.0005   0.0019          0.0044
## policyholder               0.0000   0.0000          0.0000
## num_chronicconditions      0.0004   0.0015          0.0046
## num_scrips                 0.0011   0.0053          0.0218
##
## Sample Sizes:
##                   Control Treated
## All            227517.      2655
## Matched (ESS)    2416.53    2655
## Matched          2534.      2655
## Unmatched      224983.         0
## Discarded           0.         0
```

```r
q_d_match_data <-  match.data(question1d_match)


q_d_match_data$sex <- as.factor(q_d_match_data$sex) # Convert sex from a character to a factor
q_d_match_data$sex <- as.numeric(q_d_match_data$sex) # Convert sex from a character to a numeric

q_d_matched_table <- tbl_summary(data = q_d_match_data, by = "hdhp",
                        type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")
```

	nearest_neighbour_match
(Intercept)	87 087.406
	(2110.643)
hdhp	4030.979
	(2950.696)
Num.Obs.	5189
R2	0.000
R2 Adj.	0.000
AIC	134 877.0
BIC	134 896.7
Log.Lik.	−67 435.498
F	1.866
RMSE	106 901.75

```
q_d_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,534	**1**, N = 2,655	**p-value**
Age of Patient	47.91 (14.95)	48.19 (14.93)	0.5
sex			0.5
1	1,122.00 (44.28%)	1,148.00 (43.24%)	
2	1,412.00 (55.72%)	1,507.00 (56.76%)	
famsize	2.58 (1.52)	2.55 (1.53)	0.4
policyholder			0.6
0	1,086.00 (42.86%)	1,119.00 (42.15%)	
1	1,448.00 (57.14%)	1,536.00 (57.85%)	
num_hospitalizations	1.48 (0.99)	1.46 (0.97)	0.7
num_hccs	1.96 (1.16)	1.95 (1.15)	0.8
num_scrips	12.50 (6.54)	12.56 (6.52)	0.8

```
question1d_reg <- lm(pay ~ hdhp, data = q_d_match_data, weights = weights)

msummary(list("nearest_neighbour_match" = question1d_reg))
```

```
qa_qd_balance_tables <- tbl_merge(tbls = list(unmatched_table,
                                              q_d_matched_table),
                                  tab_spanner = c("UnMatched Balance Table",
                                                  "Nearest Neighbour Matched Balance Table"))

qa_qd_balance_tables
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 227,517	**1**, N = 2,655	**p-value**	**0**, N = 2,534	**1**, N = 2,655
Age of Patient	48.29 (14.28)	48.19 (14.93)	0.7	47.91 (14.95)	48.19 (14.93)
sex			0.041		
1	102,867.00 (45.21%)	1,148.00 (43.24%)		1,122.00 (44.28%)	1,148.00 (43.24%)
2	124,650.00 (54.79%)	1,507.00 (56.76%)		1,412.00 (55.72%)	1,507.00 (56.76%)
famsize	2.38 (1.41)	2.55 (1.53)	<0.001	2.58 (1.52)	2.55 (1.53)
policyholder			0.007		
0	90,018.00 (39.57%)	1,119.00 (42.15%)		1,086.00 (42.86%)	1,119.00 (42.15%)
1	137,499.00 (60.43%)	1,536.00 (57.85%)		1,448.00 (57.14%)	1,536.00 (57.85%)
num_hospitalizations	1.49 (1.01)	1.46 (0.97)	0.2	1.48 (0.99)	1.46 (0.97)
num_hccs	1.96 (1.20)	1.95 (1.15)	0.5	1.96 (1.16)	1.95 (1.15)
num_scrips	11.89 (6.15)	12.56 (6.52)	<0.001	12.50 (6.54)	12.56 (6.52)

Looking at the number of observations included in the balance table we can see that there has been a large reduction in the number of people included in our comparison group. It can be observed that the comparison group (non-HDHP) now contains 2,534 subjects after matching and there are a total of 5189 people who have been matched. When all covariates were included and nearest neighbour matching occurred, 224,983 people in the control group were unmatched with the treated group. In comparison to the results of (c), only 88 people in the control group were unmatched. This shows that when all covariates are matched on and all the variation associated with these covariates is removed the amount of people who look like the HDHP group is a lot smaller. The regression results for this newly matched sample brings us closer to the true causal effect of HDHP on spending. After this more comprehensive match in comparison to that of (c), we can see the regression coefficient for HDHP has increased to 4030.98, thus showing that being enrolled in HDHP has a much larger effect on spending than earlier noted.

When all variables are matched upon and all the variation is removed with regards to the covariates that influence spending, the regression coefficient for HDHP flips from the negative value observed in (c). This may occur because now instead of looking at the effects of the covariates independently when all other covariates are controlled for, the match removes the variation between control and treatment group together. When this occurs we do actually see that HDHP has a positive influence on spending. This further confirms that the quality of the match in (c) was too poor and left many of back doors open. As expected HDHP superutilizers pay a larger amount for health care than those who are not enrolled in HDHP and this is seen when the match includes all covariates.

When looking at the balance table we can see that the mean age difference between the two groups is 0.28. The difference in mean of sex is 0.01. The difference in mean family size is 0.03, the difference in mean of policyholder is 0.01, the difference in mean of number of hospitilizations is 0.02, the difference in mean for number of chronic conditions is 0.01 and the difference in number of prescriptions is 0.06. As we can see there is now very little difference between the two groups and thus we know that the variation for these variables has been removed. Moreover, looking at the p-value for the t-tests, we can further see that there is no statistical difference for any of the covariates.

Question 1e

```
question1e_match <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                   distance='glm', # generalized linear model for propensity,
                  replace = TRUE
                  )
summary(question1e_match)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
##     num_scrips, data = assign_data, method = "nearest", distance = "glm",
##     replace = TRUE)
##
## Summary of Balance for All Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                       0.0119        0.0115          0.1775     1.0901
## num_hospitalizations           1.4637        1.4906         -0.0278     0.9219
## age                           48.1936       48.2925         -0.0066     1.0932
## sex                            1.5676        1.5479          0.0398     0.9912
## famsize                        2.5454        2.3845          0.1053     1.1728
## policyholder                   0.5785        0.6043         -0.0523          .
## num_chronicconditions          1.9484        1.9620         -0.0118     0.9178
## num_scrips                    12.5593       11.8947          0.1020     1.1223
##                         eCDF Mean eCDF Max
## distance                   0.0495   0.0835
## num_hospitalizations       0.0016   0.0171
## age                        0.0089   0.0229
## sex                        0.0099   0.0197
## famsize                    0.0115   0.0500
## policyholder               0.0258   0.0258
## num_chronicconditions      0.0027   0.0055
## num_scrips                 0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                       0.0119        0.0119          0.0000     1.0002
## num_hospitalizations           1.4637        1.3872          0.0787     1.2800
## age                           48.1936       49.1571         -0.0645     1.1596
## sex                            1.5676        1.5710         -0.0068     1.0019
## famsize                        2.5454        2.5119          0.0219     1.0665
## policyholder                   0.5785        0.5827         -0.0084          .
## num_chronicconditions          1.9484        1.8923          0.0488     1.1851
## num_scrips                    12.5593       12.3797          0.0276     1.0370
##                         eCDF Mean eCDF Max Std. Pair Dist.
## distance                   0.0000   0.0008          0.0000
## num_hospitalizations       0.0041   0.0271          0.3670
## age                        0.0148   0.0260          0.3553
## sex                        0.0017   0.0034          0.2805
## famsize                    0.0025   0.0169          0.2925
## policyholder               0.0041   0.0041          0.2937
## num_chronicconditions      0.0074   0.0188          0.4011
## num_scrips                 0.0047   0.0181          0.3352
##
## Sample Sizes:
##                   Control Treated
## All            227517.      2655
## Matched (ESS)    2393.56    2655
## Matched          2522.      2655
## Unmatched      224995.         0
```

```
## Discarded          0.         0
```

```
qe_psmmatch_data <- match.data(question1e_match)

qe_psmmatch_data$sex <- as.factor(qe_psmmatch_data$sex) # Convert sex from a character to a factor
qe_psmmatch_data$sex <- as.numeric(qe_psmmatch_data$sex) # Convert sex from a character to a numeric


q_e_matched_table <- tbl_summary(data = qe_psmmatch_data, by = "hdhp",
                         type = list(age ~ "continuous",
                         famsize ~ "continuous",
                         policyholder ~ "categorical",
                         num_hospitalizations ~ "continuous",
                         num_chronicconditions ~ "continuous",
                         num_scrips ~ "continuous",
                         sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")

q_e_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,522	**1**, N = 2,655	**p-value**
Age of Patient	48.87 (13.96)	48.19 (14.93)	0.094
sex			0.7
1	1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.55 (1.48)	2.55 (1.53)	>0.9
policyholder			>0.9
0	1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011
num_hccs	1.90 (1.07)	1.95 (1.15)	0.10
num_scrips	12.35 (6.48)	12.56 (6.52)	0.2

```
qa_qe_balance_tables <- tbl_merge(tbls = list(unmatched_table,
                                      q_e_matched_table),
                          tab_spanner = c("UnMatched Balance Table",
                                          "Propensity Score Matched Balance Table"))

qa_qe_balance_tables
```

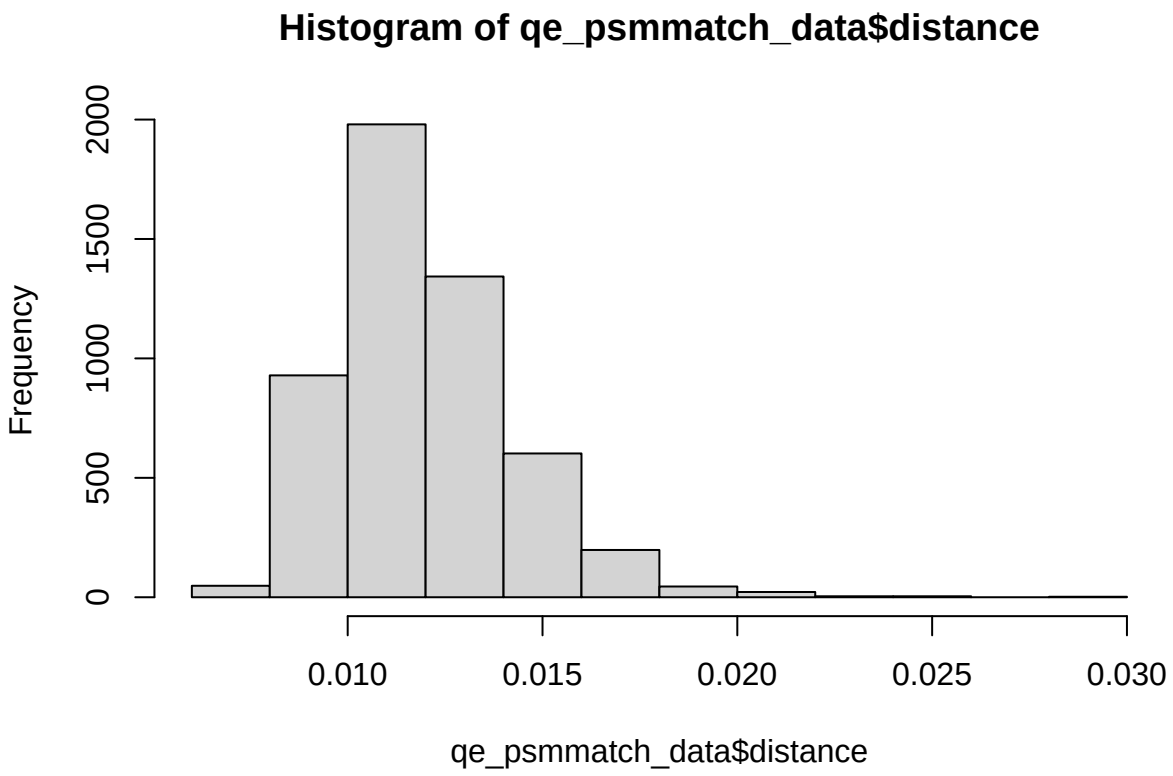
```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

	psm_match
(Intercept)	85 362.873
	(2170.633)
hdhp	4518.841
	(3030.542)
Num.Obs.	5177
R2	0.000
R2 Adj.	0.000
AIC	134 876.5
BIC	134 896.1
Log.Lik.	−67 435.241
F	2.223
RMSE	107 447.60

Characteristic	**0**, N = 227,517	**1**, N = 2,655	**p-value**	**0**, N = 2,522	**1**, N = 2,655
Age of Patient	48.29 (14.28)	48.19 (14.93)	0.7	48.87 (13.96)	48.19 (14.93)
sex			0.041		
1	102,867.00 (45.21%)	1,148.00 (43.24%)		1,104.00 (43.77%)	1,148.00 (43.24%)
2	124,650.00 (54.79%)	1,507.00 (56.76%)		1,418.00 (56.23%)	1,507.00 (56.76%)
famsize	2.38 (1.41)	2.55 (1.53)	<0.001	2.55 (1.48)	2.55 (1.53)
policyholder			0.007		
0	90,018.00 (39.57%)	1,119.00 (42.15%)		1,066.00 (42.27%)	1,119.00 (42.15%)
1	137,499.00 (60.43%)	1,536.00 (57.85%)		1,456.00 (57.73%)	1,536.00 (57.85%)
num_hospitalizations	1.49 (1.01)	1.46 (0.97)	0.2	1.40 (0.87)	1.46 (0.97)
num_hccs	1.96 (1.20)	1.95 (1.15)	0.5	1.90 (1.07)	1.95 (1.15)
num_scrips	11.89 (6.15)	12.56 (6.52)	<0.001	12.35 (6.48)	12.56 (6.52)

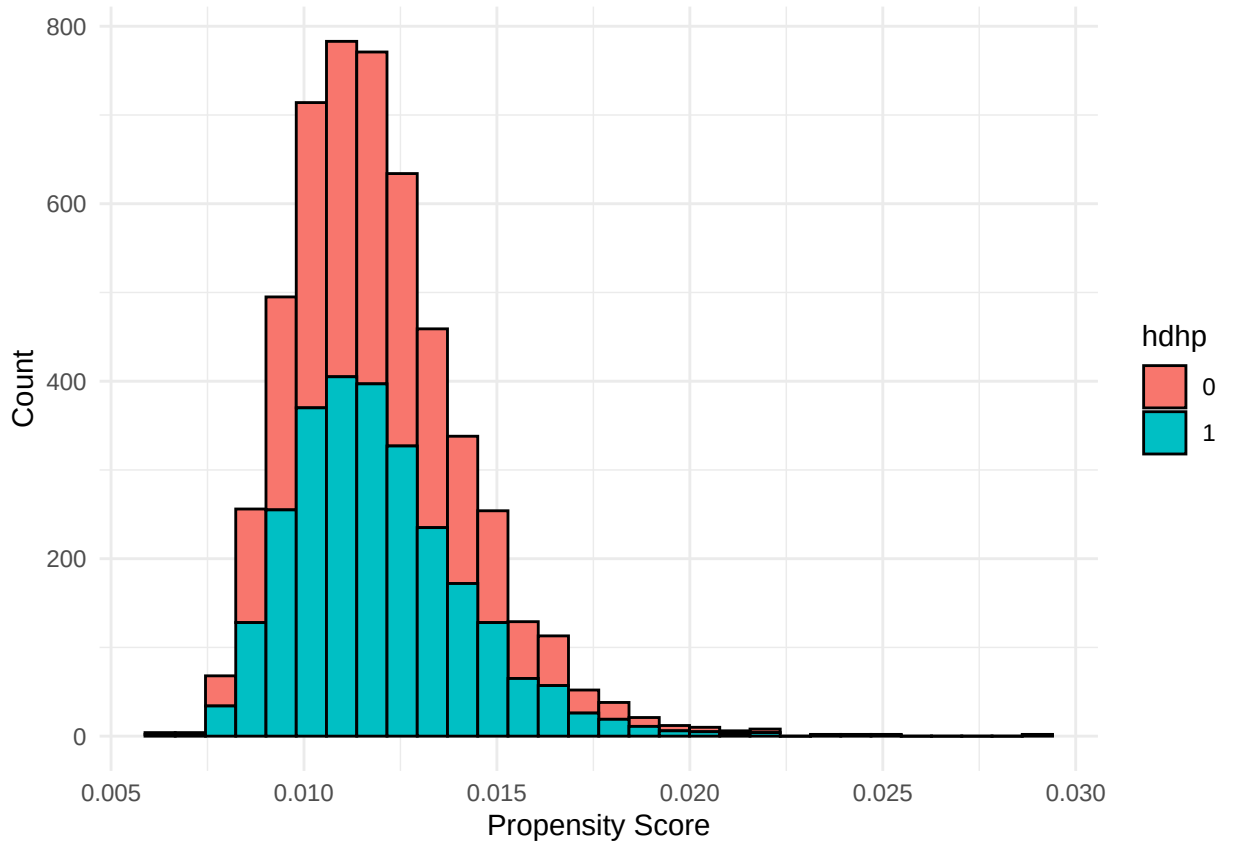
```
question1e_reg <- lm(pay ~ hdhp, data = qe_psmmatch_data, weights = 1/distance)

msummary(list("psm_match" = question1e_reg))
```

```
hist(qe_psmmatch_data$distance) # Look at overall distribution of propensity scores
```




```
ggplot(qe_psmmatch_data, aes(x=distance, fill = factor(hdhp))) +
  geom_histogram(color='black') +
  theme_minimal() +
  labs(x="Propensity Score", y="Count",fill="hdhp")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



With a propensity score approach we see similar results with regards to how many members of our control group got matched to the treatment group, 2522 members. Like the nearest neighbour approach in question (d), the match results using the propensity approach are much better than that of (c). Under this approach, the HDHP regression coefficient is now 4518.85 which is close to the regression coefficient of nearest neighbour (4030.98) but a bit higher.

In the balance table we can see that there is a 0.68 difference in the mean age, a 0.01 difference in the mean sex, no difference between the mean famsize or mean policyholder, there is a 0.06 difference in the mean num_hospitilizations, a 0.05 difference in the num_chronnicconditions, and a 0.21 difference in the mean num_scrips. The results are similar to that of nearest neighbour matching however we can see that the differences for mean age and mean num_scrips have increased in comparison to those for nearest neighbour. As with the nearest neighbour matching, there are no statistical differences for any of the covariates apart from the num_hospitalizations. This indicates that while the propensity score approach is able to attain similar results to the number of participants matched, it is not able to do so in as a sophisticated a manner as nearest neighbour matching.

Conditional independence assumption. The variables chosen to match our data on is considered to be sufficient to meet this assumption because it is of our belief that these variables close most of the backdoors that are associated with the HDHP and spending relationship. There may be open backdoors associated with geographic location and or education status which may also influence who enrolls in an HDHP and who does not. Data for these variables however were not provided and as a result the backdoors for these variables may still exist. So, it can be assummed that most of the backdoors have been closed based on the matched variables we have however there may be some backdoors left open due to a lack of data to match on.

Common support assumption. No values are centred around 0 or 1 which is good. To meet the assumption of common support, there must be substantial overlap in the distribution of the propensity score. This is mainly met as the propensity score distribution has a slight right tail indicating that there is quite a bit of

overlap between propensity score values of 0.01 and 0.015. Moreover, 95% of the treated observations were able to find a match indicating that common support is likely met.

Balance assumption. Looking at the balance table there is sufficient balance given that there are no statistical differences between any of the covariates being matched on apart from the number of prescriptions. This indicates that the balance assumption is mainly met.

The PSM assumptions are therefore met.

Question 1f

```
# Caliper set to 0.01
question1f_one <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                   distance='glm', # generalized linear model for propensity,
                  caliper = c(0.01),
                  replace = TRUE
                  )
summary(question1f_one)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
##     num_scrips, data = assign_data, method = "nearest", distance = "glm",
##     replace = TRUE, caliper = c(0.01))
##
## Summary of Balance for All Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                       0.0119        0.0115          0.1775     1.0901
## num_hospitalizations           1.4637        1.4906         -0.0278     0.9219
## age                           48.1936       48.2925         -0.0066     1.0932
## sex                            1.5676        1.5479          0.0398     0.9912
## famsize                        2.5454        2.3845          0.1053     1.1728
## policyholder                   0.5785        0.6043         -0.0523          .
## num_chronicconditions          1.9484        1.9620         -0.0118     0.9178
## num_scrips                    12.5593       11.8947          0.1020     1.1223
##                         eCDF Mean eCDF Max
## distance                   0.0495   0.0835
## num_hospitalizations       0.0016   0.0171
## age                        0.0089   0.0229
## sex                        0.0099   0.0197
## famsize                    0.0115   0.0500
## policyholder               0.0258   0.0258
## num_chronicconditions      0.0027   0.0055
## num_scrips                 0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
```

```
## distance                        0.0119        0.0119          0.0000     1.0000
## num_hospitalizations            1.4640        1.3875          0.0787     1.2799
## age                            48.2054       49.1557         -0.0637     1.1576
## sex                             1.5677        1.5711         -0.0068     1.0019
## famsize                         2.5375        2.5081          0.0192     1.0397
## policyholder                    0.5786        0.5824         -0.0076          .
## num_chronicconditions           1.9476        1.8922          0.0482     1.1837
## num_scrips                     12.5665       12.3683          0.0304     1.0431
##                          eCDF Mean eCDF Max Std. Pair Dist.
## distance                    0.0000   0.0008          0.0000
## num_hospitalizations        0.0041   0.0271          0.3673
## age                         0.0146   0.0256          0.3546
## sex                         0.0017   0.0034          0.2807
## famsize                     0.0023   0.0170          0.2900
## policyholder                0.0038   0.0038          0.2931
## num_chronicconditions       0.0073   0.0185          0.4007
## num_scrips                  0.0049   0.0188          0.3326
##
## Sample Sizes:
##                   Control Treated
## All            227517.      2655
## Matched (ESS)    2391.58    2653
## Matched          2520.      2653
## Unmatched      224997.         2
## Discarded           0.         0
```

```
# Caliper set to 0.1
question1f_two <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                   distance='glm', # generalized linear model for propensity,
                  caliper = c(0.1),
                  replace = TRUE
                  )
summary(question1f_two)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
##     num_scrips, data = assign_data, method = "nearest", distance = "glm",
##     replace = TRUE, caliper = c(0.1))
##
## Summary of Balance for All Data:
##                          Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                        0.0119        0.0115          0.1775     1.0901
## num_hospitalizations            1.4637        1.4906         -0.0278     0.9219
## age                            48.1936       48.2925         -0.0066     1.0932
## sex                             1.5676        1.5479          0.0398     0.9912
## famsize                         2.5454        2.3845          0.1053     1.1728
## policyholder                    0.5785        0.6043         -0.0523          .
## num_chronicconditions           1.9484        1.9620         -0.0118     0.9178
## num_scrips                     12.5593       11.8947          0.1020     1.1223
```

```
##                         eCDF Mean eCDF Max
## distance                   0.0495   0.0835
## num_hospitalizations       0.0016   0.0171
## age                        0.0089   0.0229
## sex                        0.0099   0.0197
## famsize                    0.0115   0.0500
## policyholder               0.0258   0.0258
## num_chronicconditions      0.0027   0.0055
## num_scrips                 0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                         Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                       0.0119        0.0119          0.0000     1.0002
## num_hospitalizations           1.4637        1.3872          0.0787     1.2800
## age                           48.1936       49.1571         -0.0645     1.1596
## sex                            1.5676        1.5710         -0.0068     1.0019
## famsize                        2.5454        2.5119          0.0219     1.0665
## policyholder                   0.5785        0.5827         -0.0084          .
## num_chronicconditions          1.9484        1.8923          0.0488     1.1851
## num_scrips                    12.5593       12.3797          0.0276     1.0370
##                         eCDF Mean eCDF Max Std. Pair Dist.
## distance                   0.0000   0.0008          0.0000
## num_hospitalizations       0.0041   0.0271          0.3670
## age                        0.0148   0.0260          0.3553
## sex                        0.0017   0.0034          0.2805
## famsize                    0.0025   0.0169          0.2925
## policyholder               0.0041   0.0041          0.2937
## num_chronicconditions      0.0074   0.0188          0.4011
## num_scrips                 0.0047   0.0181          0.3352
##
## Sample Sizes:
##                   Control Treated
## All            227517.      2655
## Matched (ESS)    2393.56    2655
## Matched          2522.      2655
## Unmatched      224995.         0
## Discarded           0.         0
```

```
# Caliper set to 0.2
question1f_three <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                   distance='glm', # generalized linear model for propensity,
                  caliper = c(0.2),
                  replace = TRUE
                  )
summary(question1f_three)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
```

```
##     num_scrips, data = assign_data, method = "nearest", distance = "glm",
##     replace = TRUE, caliper = c(0.2))
##
## Summary of Balance for All Data:
##                       Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                     0.0119        0.0115          0.1775     1.0901
## num_hospitalizations         1.4637        1.4906         -0.0278     0.9219
## age                         48.1936       48.2925         -0.0066     1.0932
## sex                          1.5676        1.5479          0.0398     0.9912
## famsize                      2.5454        2.3845          0.1053     1.1728
## policyholder                 0.5785        0.6043         -0.0523          .
## num_chronicconditions        1.9484        1.9620         -0.0118     0.9178
## num_scrips                  12.5593       11.8947          0.1020     1.1223
##                       eCDF Mean eCDF Max
## distance                 0.0495   0.0835
## num_hospitalizations     0.0016   0.0171
## age                      0.0089   0.0229
## sex                      0.0099   0.0197
## famsize                  0.0115   0.0500
## policyholder             0.0258   0.0258
## num_chronicconditions    0.0027   0.0055
## num_scrips               0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                       Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                     0.0119        0.0119          0.0000     1.0002
## num_hospitalizations         1.4637        1.3872          0.0787     1.2800
## age                         48.1936       49.1571         -0.0645     1.1596
## sex                          1.5676        1.5710         -0.0068     1.0019
## famsize                      2.5454        2.5119          0.0219     1.0665
## policyholder                 0.5785        0.5827         -0.0084          .
## num_chronicconditions        1.9484        1.8923          0.0488     1.1851
## num_scrips                  12.5593       12.3797          0.0276     1.0370
##                       eCDF Mean eCDF Max Std. Pair Dist.
## distance                 0.0000   0.0008          0.0000
## num_hospitalizations     0.0041   0.0271          0.3670
## age                      0.0148   0.0260          0.3553
## sex                      0.0017   0.0034          0.2805
## famsize                  0.0025   0.0169          0.2925
## policyholder             0.0041   0.0041          0.2937
## num_chronicconditions    0.0074   0.0188          0.4011
## num_scrips               0.0047   0.0181          0.3352
##
## Sample Sizes:
##                 Control Treated
## All           227517.      2655
## Matched (ESS)   2393.56    2655
## Matched         2522.      2655
## Unmatched     224995.         0
## Discarded          0.         0
```

```
# Caliper set to 0.5
question1f_four <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
```

```
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                   distance='glm', # generalized linear model for propensity,
                  caliper = c(0.5),
                  replace = TRUE
                  )
summary(question1f_four)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
##     num_scrips, data = assign_data, method = "nearest", distance = "glm",
##     replace = TRUE, caliper = c(0.5))
##
## Summary of Balance for All Data:
##                        Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                      0.0119        0.0115          0.1775     1.0901
## num_hospitalizations          1.4637        1.4906         -0.0278     0.9219
## age                          48.1936       48.2925         -0.0066     1.0932
## sex                           1.5676        1.5479          0.0398     0.9912
## famsize                       2.5454        2.3845          0.1053     1.1728
## policyholder                  0.5785        0.6043         -0.0523          .
## num_chronicconditions         1.9484        1.9620         -0.0118     0.9178
## num_scrips                   12.5593       11.8947          0.1020     1.1223
##                        eCDF Mean eCDF Max
## distance                  0.0495   0.0835
## num_hospitalizations      0.0016   0.0171
## age                       0.0089   0.0229
## sex                       0.0099   0.0197
## famsize                   0.0115   0.0500
## policyholder              0.0258   0.0258
## num_chronicconditions     0.0027   0.0055
## num_scrips                0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                        Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                      0.0119        0.0119          0.0000     1.0002
## num_hospitalizations          1.4637        1.3872          0.0787     1.2800
## age                          48.1936       49.1571         -0.0645     1.1596
## sex                           1.5676        1.5710         -0.0068     1.0019
## famsize                       2.5454        2.5119          0.0219     1.0665
## policyholder                  0.5785        0.5827         -0.0084          .
## num_chronicconditions         1.9484        1.8923          0.0488     1.1851
## num_scrips                   12.5593       12.3797          0.0276     1.0370
##                        eCDF Mean eCDF Max Std. Pair Dist.
## distance                  0.0000   0.0008          0.0000
## num_hospitalizations      0.0041   0.0271          0.3670
## age                       0.0148   0.0260          0.3553
## sex                       0.0017   0.0034          0.2805
## famsize                   0.0025   0.0169          0.2925
```

```
## policyholder             0.0041   0.0041          0.2937
## num_chronicconditions    0.0074   0.0188          0.4011
## num_scrips               0.0047   0.0181          0.3352
##
## Sample Sizes:
##                 Control Treated
## All           227517.      2655
## Matched (ESS)   2393.56    2655
## Matched         2522.      2655
## Unmatched     224995.         0
## Discarded          0.         0
```

```
# Caliper set to 1
question1f_five <- matchit(hdhp ~ num_hospitalizations + age + sex + famsize + policyholder +
                  num_hospitalizations + num_chronicconditions + num_scrips,
                            data = assign_data,
                            method = "nearest",
                   distance='glm', # generalized linear model for propensity,
                  caliper = c(1),
                  replace = TRUE
                  )
summary(question1f_five)
```

```
##
## Call:
## matchit(formula = hdhp ~ num_hospitalizations + age + sex + famsize +
##     policyholder + num_hospitalizations + num_chronicconditions +
##     num_scrips, data = assign_data, method = "nearest", distance = "glm",
##     replace = TRUE, caliper = c(1))
##
## Summary of Balance for All Data:
##                        Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                      0.0119        0.0115          0.1775     1.0901
## num_hospitalizations          1.4637        1.4906         -0.0278     0.9219
## age                          48.1936       48.2925         -0.0066     1.0932
## sex                           1.5676        1.5479          0.0398     0.9912
## famsize                       2.5454        2.3845          0.1053     1.1728
## policyholder                  0.5785        0.6043         -0.0523          .
## num_chronicconditions         1.9484        1.9620         -0.0118     0.9178
## num_scrips                   12.5593       11.8947          0.1020     1.1223
##                        eCDF Mean eCDF Max
## distance                  0.0495   0.0835
## num_hospitalizations      0.0016   0.0171
## age                       0.0089   0.0229
## sex                       0.0099   0.0197
## famsize                   0.0115   0.0500
## policyholder              0.0258   0.0258
## num_chronicconditions     0.0027   0.0055
## num_scrips                0.0140   0.0479
##
##
## Summary of Balance for Matched Data:
##                        Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                      0.0119        0.0119          0.0000     1.0002
```

```
## num_hospitalizations         1.4637        1.3872          0.0787     1.2800
## age                         48.1936       49.1571         -0.0645     1.1596
## sex                          1.5676        1.5710         -0.0068     1.0019
## famsize                      2.5454        2.5119          0.0219     1.0665
## policyholder                 0.5785        0.5827         -0.0084          .
## num_chronicconditions        1.9484        1.8923          0.0488     1.1851
## num_scrips                  12.5593       12.3797          0.0276     1.0370
##                        eCDF Mean eCDF Max Std. Pair Dist.
## distance                  0.0000   0.0008          0.0000
## num_hospitalizations      0.0041   0.0271          0.3670
## age                       0.0148   0.0260          0.3553
## sex                       0.0017   0.0034          0.2805
## famsize                   0.0025   0.0169          0.2925
## policyholder              0.0041   0.0041          0.2937
## num_chronicconditions     0.0074   0.0188          0.4011
## num_scrips                0.0047   0.0181          0.3352
##
## Sample Sizes:
##                  Control Treated
## All           227517.      2655
## Matched (ESS)   2393.56    2655
## Matched         2522.      2655
## Unmatched     224995.         0
## Discarded          0.         0
```

```r
match_data_one <- match.data(question1f_one)
match_data_two <- match.data(question1f_two)
match_data_three <- match.data(question1f_three)
match_data_four <- match.data(question1f_four)
match_data_five <- match.data(question1f_five)


match_data_one$sex <- as.factor(match_data_one$sex) # Convert sex from a character to a factor
match_data_one$sex <- as.numeric(match_data_one$sex) # Convert sex from a character to a numeric

match_data_two$sex <- as.factor(match_data_two$sex) # Convert sex from a character to a factor
match_data_two$sex <- as.numeric(match_data_two$sex) # Convert sex from a character to a numeric

match_data_three$sex <- as.factor(match_data_three$sex) # Convert sex from a character to a factor
match_data_three$sex <- as.numeric(match_data_three$sex) # Convert sex from a character to a numeric

match_data_four$sex <- as.factor(match_data_four$sex) # Convert sex from a character to a factor
match_data_four$sex <- as.numeric(match_data_four$sex) # Convert sex from a character to a numeric

match_data_five$sex <- as.factor(match_data_five$sex) # Convert sex from a character to a factor
match_data_five$sex <- as.numeric(match_data_five$sex) # Convert sex from a character to a numeric


#Balance Tables for each caliper size

q_f_1_matched_table <- tbl_summary(data = match_data_one, by = "hdhp",
                        type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
```

```
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")


q_f_2_matched_table <- tbl_summary(data = match_data_two, by = "hdhp",
                        type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")

q_f_3_matched_table <- tbl_summary(data = match_data_three, by = "hdhp",
                        type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")

q_f_4_matched_table <- tbl_summary(data = match_data_four, by = "hdhp",
                        type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
```

```
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")

q_f_5_matched_table <- tbl_summary(data = match_data_five, by = "hdhp",
                        type = list(age ~ "continuous",
                        famsize ~ "continuous",
                        policyholder ~ "categorical",
                        num_hospitalizations ~ "continuous",
                        num_chronicconditions ~ "continuous",
                        num_scrips ~ "continuous",
                        sex ~ "categorical"),
            statistic = list(all_continuous() ~ "{mean} ({sd})",
                             all_categorical() ~ "{n} ({p}%)"),
            digits = everything() ~ 2,
            include = c("age", "sex", "famsize", "policyholder",
                  "num_hospitalizations", "num_chronicconditions", "num_scrips")) %>%
  add_p(test = everything() ~ "t.test")


q_f_1_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,520	**1**, N = 2,653	**p-value**
Age of Patient	48.87 (13.96)	48.21 (14.92)	0.10
sex			0.7
1	1,103.00 (43.77%)	1,147.00 (43.23%)	
2	1,417.00 (56.23%)	1,506.00 (56.77%)	
famsize	2.54 (1.47)	2.54 (1.50)	0.9
policyholder			>0.9
0	1,066.00 (42.30%)	1,118.00 (42.14%)	
1	1,454.00 (57.70%)	1,535.00 (57.86%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011
num_hccs	1.90 (1.07)	1.95 (1.15)	0.11
num_scrips	12.34 (6.46)	12.57 (6.51)	0.2

```
q_f_2_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,522	**1**, N = 2,655	**p-value**
Age of Patient	48.87 (13.96)	48.19 (14.93)	0.094
sex			0.7
1	1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.55 (1.48)	2.55 (1.53)	>0.9
policyholder			>0.9
0	1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011
num_hccs	1.90 (1.07)	1.95 (1.15)	0.10
num_scrips	12.35 (6.48)	12.56 (6.52)	0.2

```
q_f_3_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,522	**1**, N = 2,655	**p-value**
Age of Patient	48.87 (13.96)	48.19 (14.93)	0.094
sex			0.7
1	1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.55 (1.48)	2.55 (1.53)	>0.9
policyholder			>0.9
0	1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011
num_hccs	1.90 (1.07)	1.95 (1.15)	0.10
num_scrips	12.35 (6.48)	12.56 (6.52)	0.2

```
q_f_4_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,522	**1**, N = 2,655	**p-value**
Age of Patient	48.87 (13.96)	48.19 (14.93)	0.094
sex			0.7
1	1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.55 (1.48)	2.55 (1.53)	>0.9
policyholder			>0.9
0	1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011
num_hccs	1.90 (1.07)	1.95 (1.15)	0.10
num_scrips	12.35 (6.48)	12.56 (6.52)	0.2

```
q_f_5_matched_table
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,522	**1**, N = 2,655	**p-value**
Age of Patient	48.87 (13.96)	48.19 (14.93)	0.094
sex			0.7
1	1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.55 (1.48)	2.55 (1.53)	>0.9
policyholder			>0.9
0	1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011
num_hccs	1.90 (1.07)	1.95 (1.15)	0.10
num_scrips	12.35 (6.48)	12.56 (6.52)	0.2

```
caliper_tbls_descr <- tbl_merge(tbls = list(q_f_1_matched_table,
                                            q_f_2_matched_table,
                                            q_f_3_matched_table,
                                            q_f_4_matched_table,
                                            q_f_5_matched_table),
                                tab_spanner = c("**Caliper Size - 0.01**",
                                                "**Caliper Size - 0.1**",
                                                "**Caliper Size - 0.2**",
                                                "**Caliper Size - 0.5**",
                                                "**Caliper Size - 1**"))
caliper_tbls_descr
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,520	**1**, N = 2,653	**p-value**	**0**, N = 2,522	**1**, N = 2,655	*
Age of Patient	48.87 (13.96)	48.21 (14.92)	0.10	48.87 (13.96)	48.19 (14.93)	
sex			0.7			
1	1,103.00 (43.77%)	1,147.00 (43.23%)		1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,417.00 (56.23%)	1,506.00 (56.77%)		1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.54 (1.47)	2.54 (1.50)	0.9	2.55 (1.48)	2.55 (1.53)	
policyholder			>0.9			
0	1,066.00 (42.30%)	1,118.00 (42.14%)		1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,454.00 (57.70%)	1,535.00 (57.86%)		1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011	1.40 (0.87)	1.46 (0.97)	
num_hccs	1.90 (1.07)	1.95 (1.15)	0.11	1.90 (1.07)	1.95 (1.15)	
num_scrips	12.34 (6.46)	12.57 (6.51)	0.2	12.35 (6.48)	12.56 (6.52)	

```
question1f_one_reg <- lm(pay ~ hdhp, data = match_data_one, weights = 1/weights)
question1f_two_reg <- lm(pay ~ hdhp, data = match_data_two, weights = 1/weights)
question1f_three_reg <- lm(pay ~ hdhp, data = match_data_three, weights = 1/weights)
```

	Caliper Size - 0.01	Caliper Size - 0.1	Caliper Size - 0.2	Caliper Size - 0.5	Caliper Size - 1
(Intercept)	84 926.122***	85 046.736***	85 046.736***	85 046.736***	85 046.736***
	(2132.572)	(2133.574)	(2133.574)	(2133.574)	(2133.574)
hdhp	6236.011**	6071.649**	6071.649**	6071.649**	6071.649**
	(2997.154)	(2998.577)	(2998.577)	(2998.577)	(2998.577)
Num.Obs.	5173	5177	5177	5177	5177
R2	0.001	0.001	0.001	0.001	0.001
R2 Adj.	0.001	0.001	0.001	0.001	0.001
AIC	134 597.6	134 710.5	134 710.5	134 710.5	134 710.5
BIC	134 617.2	134 730.2	134 730.2	134 730.2	134 730.2
Log.Lik.	−67 295.787	−67 352.274	−67 352.274	−67 352.274	−67 352.274
F	4.329	4.100	4.100	4.100	4.100
RMSE	107 353.55	107 443.08	107 443.08	107 443.08	107 443.08

* p < 0.1, ** p < 0.05, *** p < 0.01

```
question1f_four_reg <- lm(pay ~ hdhp, data = match_data_four, weights = 1/weights)
question1f_five_reg <- lm(pay ~ hdhp, data = match_data_five, weights = 1/weights)

caliper_tbls_reg <- msummary(list("Caliper Size - 0.01" = question1f_one_reg,
                                  "Caliper Size - 0.1" = question1f_two_reg,
                                  "Caliper Size - 0.2" = question1f_three_reg,
                                  "Caliper Size - 0.5" = question1f_four_reg,
                                  "Caliper Size - 1" = question1f_five_reg),
                             stars=c('*' = .1, '**' = .05, '***' = .01))


caliper_tbls_descr
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**0**, N = 2,520	**1**, N = 2,653	**p-value**	**0**, N = 2,522	**1**, N = 2,655	*
Age of Patient	48.87 (13.96)	48.21 (14.92)	0.10	48.87 (13.96)	48.19 (14.93)	
sex			0.7			
1	1,103.00 (43.77%)	1,147.00 (43.23%)		1,104.00 (43.77%)	1,148.00 (43.24%)	
2	1,417.00 (56.23%)	1,506.00 (56.77%)		1,418.00 (56.23%)	1,507.00 (56.76%)	
famsize	2.54 (1.47)	2.54 (1.50)	0.9	2.55 (1.48)	2.55 (1.53)	
policyholder			>0.9			
0	1,066.00 (42.30%)	1,118.00 (42.14%)		1,066.00 (42.27%)	1,119.00 (42.15%)	
1	1,454.00 (57.70%)	1,535.00 (57.86%)		1,456.00 (57.73%)	1,536.00 (57.85%)	
num_hospitalizations	1.40 (0.87)	1.46 (0.97)	0.011	1.40 (0.87)	1.46 (0.97)	
num_hccs	1.90 (1.07)	1.95 (1.15)	0.11	1.90 (1.07)	1.95 (1.15)	
num_scrips	12.34 (6.46)	12.57 (6.51)	0.2	12.35 (6.48)	12.56 (6.52)	

```
caliper_tbls_reg
```

As the caliper size increases, the sample size also increases. However, this is only true when moving between a caliper size of 0.01 and 0.1; when increasing the caliper size after 0.1, the effective sample size stays the same. The tradeoff between larger and smaller caliper sizes is that when the caliper size is larger a greater

number of matches can occur, however, these matches are less exact while when the caliper size is smaller more exact matches can be formed although not as many matches will occur. As narrower calipers lead to matching of more similar groups, it is expected that this leads to less systematic differences between treated and untreated subjects. However, the lower number of matches as a result of a narrower caliper may cause higher variance in the estimated treatment effect. Having a wider caliper size has the opposite effects and thus choice of caliper size has effects on the systematic differences between treatment groups as well as effects on the variance of the estimated treatment effect. In reference to the match above, it can be seen that after the caliper size increases to 0.1 more matches are included however increasing the caliper size any further has no effect as that caliper size is sufficient for finding complete matches with the data provided.

Looking at the balance table comparing the five caliper sizes, we can see that after increasing the caliper size from 0.01 onwards, there is no change in the balance of the covariates. The means of all the covariates apart from number of hospitalizations remain not statistically different.

When decreasing the caliper size from 0.01 to 0.1, there is a decrease in total spending for those enrolled in a HDHP from 6236.011 to 6071.649 units. This makes sense because as the caliper increases in size, the more bias we introduce into our matches and estimates. There is no further change in the estimated treatment effect when the caliper size increases after 0.1.

Question 1g

If we had data on all spenders this would allow us to apply a causal statement about the effect of being enrolled in a HDHP on total spending for everyone that uses health services. However, as HDHPs are more advantageous for people who anticipate primarily requiring preventive services, we may expect individuals who are not in the top 1% of spending to be generally enrolled in these types of plans. Therefore, by including all spenders in our matching, we would not be able to isolate the effect of HDHP enrollment on being a superutilizer as we would now be making a more general conclusion about all utilizers of HDHP enrollment on total cost. Due to this reason of not being able to isolate the variation caused from being a superutilizer on costs, I would choose not to utilize all spenders in the study.

Question 1h

From the analysis done in the previous sections, I conclude that if a person is enrolled in an HDHP it results in higher total costs than if the person is not enrolled in an HDHP. This matches my intuition as HDHPs are beneficial to total cost if the person does not anticipate using a lot of healthcare services. If a person does use many healthcare services, the deductible is much higher for HDHP type plans compared to other plans leading to higher total costs which is reflected in our analysis.

```
print("==============================Works Cited==============================")
```

```
## [1] "==============================Works Cited=============================="
```

```
loadedNamespaces() %>%
  map(citation) %>%
print(style = "text") # Adds citations for each package to end of .rmd file
```

```
## [[1]]
## Eddelbuettel D, François R (2011). "Rcpp: Seamless R and C++
## Integration." _Journal of Statistical Software_, *40*(8), 1-18.
## doi:10.18637/jss.v040.i08 <https://doi.org/10.18637/jss.v040.i08>.
```

##
Eddelbuettel D (2013). _Seamless R and C++ Integration with Rcpp_.
Springer, New York. doi:10.1007/978-1-4614-6868-4
<https://doi.org/10.1007/978-1-4614-6868-4>, ISBN 978-1-4614-6867-7.
##
Eddelbuettel D, Balamuta JJ (2018). "Extending extitR with extitC++: A
Brief Introduction to extitRcpp." _The American Statistician_, *72*(1),
28-36. doi:10.1080/00031305.2017.1375990
<https://doi.org/10.1080/00031305.2017.1375990>.
##
[[2]]
Wickham H, Henry L, Pedersen T, Luciani T, Decorde M, Lise V (2022).
svglite: An 'SVG' Graphics Device. R package version 2.1.0,
<https://CRAN.R-project.org/package=svglite>.
##
[[3]]
Müller K (2020). _here: A Simpler Way to Find Your Files_. R package
version 1.0.1, <https://CRAN.R-project.org/package=here>.
##
[[4]]
Sarkar D (2008). _Lattice: Multivariate Data Visualization with R_.
Springer, New York. ISBN 978-0-387-75968-5,
<http://lmdvr.r-forge.r-project.org>.
##
[[5]]
Wickham H, Girlich M (2022). _tidyr: Tidy Messy Data_. R package
version 1.2.1, <https://CRAN.R-project.org/package=tidyr>.
##
[[6]]
Zeileis A, Grothendieck G (2005). "zoo: S3 Infrastructure for Regular
and Irregular Time Series." _Journal of Statistical Software_, *14*(6),
1-27. doi:10.18637/jss.v014.i06
<https://doi.org/10.18637/jss.v014.i06>.
##
[[7]]
Wickham H (2019). _assertthat: Easy Pre and Post Assertions_. R package
version 0.2.1, <https://CRAN.R-project.org/package=assertthat>.
##
[[8]]
Müller K (2022). _rprojroot: Finding Files in Project Subdirectories_.
R package version 2.0.3,
<https://CRAN.R-project.org/package=rprojroot>.
##
[[9]]
Lucas DEwcbA, Tuszynski J, Bengtsson H, Urbanek S, Frasca M, Lewis B,
Stokely M, Muehleisen H, Murdoch D, Hester J, Wu W, Kou Q, Onkelinx T,
Lang M, Simko V, Hornik K, Neal R, Bell K, de Queljoe M, Suruceanu I,
Denney B, Schumacher D, Chang. aW (2021). _digest: Create Compact Hash
Digests of R Objects_. R package version 0.6.29,
<https://CRAN.R-project.org/package=digest>.
##
[[10]]
Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression
Relationships." _R News_, *2*(3), 7-10.

<https://CRAN.R-project.org/doc/Rnews/>.
##
[[11]]
Perry PO (2021). _utf8: Unicode Text Processing_. R package version
1.2.2, <https://CRAN.R-project.org/package=utf8>.
##
[[12]]
Chang W (2021). _R6: Encapsulated Classes with Reference Semantics_. R
package version 2.5.1, <https://CRAN.R-project.org/package=R6>.
##
[[13]]
Lang M, R Core Team (2021). _backports: Reimplementations of Functions
Introduced Since R-3.0.0_. R package version 1.4.1,
<https://CRAN.R-project.org/package=backports>.
##
[[14]]
Ho DE, Imai K, King G, Stuart EA (2011). "MatchIt: Nonparametric
Preprocessing for Parametric Causal Inference." _Journal of Statistical
Software_, *42*(8), 1-28. doi:10.18637/jss.v042.i08
<https://doi.org/10.18637/jss.v042.i08>.
##
[[15]]
Huntington-Klein N (2022). _vtable: Variable Table for Variable
Documentation_. R package version 1.3.4,
<https://CRAN.R-project.org/package=vtable>.
##
[[16]]
Wickham H, Xie Y (2022). _evaluate: Parsing and Evaluation Tools that
Provide More Details than the Default_. R package version 0.16,
<https://CRAN.R-project.org/package=evaluate>.
##
[[17]]
Xie Y, Qiu Y (2021). _highr: Syntax Highlighting for R Source Code_. R
package version 0.9, <https://CRAN.R-project.org/package=highr>.
##
[[18]]
Wickham H (2022). _httr: Tools for Working with URLs and HTTP_. R
package version 1.4.4, <https://CRAN.R-project.org/package=httr>.
##
[[19]]
Wickham H (2016). _ggplot2: Elegant Graphics for Data Analysis_.
Springer-Verlag New York. ISBN 978-3-319-24277-4,
<https://ggplot2.tidyverse.org>.
##
[[20]]
Müller K, Wickham H (2022). _pillar: Coloured Formatting for Columns_.
R package version 1.8.1, <https://CRAN.R-project.org/package=pillar>.
##
[[21]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[22]]

Henry L, Wickham H (2022). _rlang: Functions for Base Types and Core R
and 'Tidyverse' Features_. R package version 1.0.6,
<https://CRAN.R-project.org/package=rlang>.
##
[[23]]
Ushey K, Allaire J, Wickham H, Ritchie G (2022). _rstudioapi: Safely
Access the RStudio API_. R package version 0.14,
<https://CRAN.R-project.org/package=rstudioapi>.
##
[[24]]
Lüdecke D, Ben-Shachar M, Patil I, Waggoner P, Makowski D (2021).
"performance: An R Package for Assessment, Comparison and Testing of
Statistical Models." _Journal of Open Source Software_, *6*(60), 3139.
doi:10.21105/joss.03139 <https://doi.org/10.21105/joss.03139>.
##
[[25]]
Lang M (2017). "checkmate: Fast Argument Checks for Defensive R
Programming." _The R Journal_, *9*(1), 437-445.
doi:10.32614/RJ-2017-028 <https://doi.org/10.32614/RJ-2017-028>.
##
[[26]]
Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham
H, Cheng J, Chang W, Iannone R (2022). _rmarkdown: Dynamic Documents
for R_. R package version 2.16, <https://github.com/rstudio/rmarkdown>.
##
Xie Y, Allaire J, Grolemund G (2018). _R Markdown: The Definitive
Guide_. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338,
<https://bookdown.org/yihui/rmarkdown>.
##
Xie Y, Dervieux C, Riederer E (2020). _R Markdown Cookbook_. Chapman
and Hall/CRC, Boca Raton, Florida. ISBN 9780367563837,
<https://bookdown.org/yihui/rmarkdown-cookbook>.
##
[[27]]
Talbot, J (2020). _labeling: Axis Labeling_. R package version 0.4.2,
<https://CRAN.R-project.org/package=labeling>.
##
[[28]]
Chang W (2022). _webshot: Take Screenshots of Web Pages_. R package
version 0.5.4, <https://CRAN.R-project.org/package=webshot>.
##
[[29]]
Wickham H, Hester J, Bryan J (2022). _readr: Read Rectangular Text
Data_. R package version 2.1.3,
<https://CRAN.R-project.org/package=readr>.
##
[[30]]
Wickham H (2022). _stringr: Simple, Consistent Wrappers for Common
String Operations_. R package version 1.4.1,
<https://CRAN.R-project.org/package=stringr>.
##
[[31]]
Forner K (2020). _RcppProgress: An Interruptible Progress Bar with
OpenMP Support for C++ in R Packages_. R package version 0.4.2,

<https://CRAN.R-project.org/package=RcppProgress>.
##
[[32]]
Wickham C (2018). _munsell: Utilities for Using Munsell Colours_. R
package version 0.5.0, <https://CRAN.R-project.org/package=munsell>.
##
[[33]]
Robinson D, Hayes A, Couch S (2022). _broom: Convert Statistical
Objects into Tidy Tibbles_. R package version 1.0.1,
<https://CRAN.R-project.org/package=broom>.
##
[[34]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[35]]
Xie Y (2022). _xfun: Supporting Functions for Packages Maintained by
'Yihui Xie'_. R package version 0.33,
<https://CRAN.R-project.org/package=xfun>.
##
[[36]]
Csárdi G (2019). _pkgconfig: Private Configuration for 'R' Packages_. R
package version 2.0.3, <https://CRAN.R-project.org/package=pkgconfig>.
##
[[37]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[38]]
Pedersen T, Ooms J, Govett D (2022). _systemfonts: System Native Font
Finding_. R package version 1.0.4,
<https://CRAN.R-project.org/package=systemfonts>.
##
[[39]]
Lüdecke D, Ben-Shachar M, Patil I, Makowski D (2020). "Extracting,
Computing and Exploring the Parameters of Statistical Models using R."
Journal of Open Source Software, *5*(53), 2445.
doi:10.21105/joss.02445 <https://doi.org/10.21105/joss.02445>.
##
[[40]]
Cheng J, Sievert C, Schloerke B, Chang W, Xie Y, Allen J (2022).
htmltools: Tools for HTML. R package version 0.5.3,
<https://CRAN.R-project.org/package=htmltools>.
##
[[41]]
Lüdecke D, Waggoner P, Makowski D (2019). "insight: A Unified Interface
to Access Information from Model Objects in R." _Journal of Open Source
Software_, *4*(38), 1412. doi:10.21105/joss.01412
<https://doi.org/10.21105/joss.01412>.
##
[[42]]
Henry L, Wickham H (2022). _tidyselect: Select from a Set of Strings_.

R package version 1.1.2,
<https://CRAN.R-project.org/package=tidyselect>.
##
[[43]]
Müller K, Wickham H (2022). _tibble: Simple Data Frames_. R package
version 3.1.8, <https://CRAN.R-project.org/package=tibble>.
##
[[44]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[45]]
Gaslam B (2022). _fansi: ANSI Control Sequence Aware String Functions_.
R package version 1.0.3, <https://CRAN.R-project.org/package=fansi>.
##
[[46]]
Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro A, Sciaini,
Marco, Scherer, Cédric (2022). _viridis - Colorblind-Friendly Color
Maps for R_. doi:10.5281/zenodo.4679424
<https://doi.org/10.5281/zenodo.4679424>, R package version 0.4.1,
<https://sjmgarnier.github.io/viridis/>.
##
[[47]]
Wickham H, François R, Henry L, Müller K (2022). _dplyr: A Grammar of
Data Manipulation_. R package version 1.0.10,
<https://CRAN.R-project.org/package=dplyr>.
##
[[48]]
Vaughan D (2022). _tzdb: Time Zone Database Information_. R package
version 0.3.0, <https://CRAN.R-project.org/package=tzdb>.
##
[[49]]
Hester J, Henry L, Müller K, Ushey K, Wickham H, Chang W (2022).
withr: Run Code 'With' Temporarily Modified Global State. R package
version 2.5.0, <https://CRAN.R-project.org/package=withr>.
##
[[50]]
Murdoch D (2020). _tables: Formula-Driven Table Generation_. R package
version 0.9.6, <https://CRAN.R-project.org/package=tables>.
##
[[51]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[52]]
Wickham H, Pedersen T (2022). _gtable: Arrange 'Grobs' in Tables_. R
package version 0.3.1, <https://CRAN.R-project.org/package=gtable>.
##
[[53]]
Henry L, Wickham H (2022). _lifecycle: Manage the Life Cycle of your
Package Functions_. R package version 1.0.3,
<https://CRAN.R-project.org/package=lifecycle>.

##
[[54]]
R Special Interest Group on Databases (R-SIG-DB), Wickham H, Müller K
(2022). _DBI: R Database Interface_. R package version 1.1.3,
<https://CRAN.R-project.org/package=DBI>.
##
[[55]]
Bache S, Wickham H (2022). _magrittr: A Forward-Pipe Operator for R_. R
package version 2.0.3, <https://CRAN.R-project.org/package=magrittr>.
##
[[56]]
Makowski D, Ben-Shachar M, Lüdecke D (2019). "bayestestR: Describing
Effects and their Uncertainty, Existence and Significance within the
Bayesian Framework." _Journal of Open Source Software_, *4*(40), 1541.
doi:10.21105/joss.01541 <https://doi.org/10.21105/joss.01541>,
<https://joss.theoj.org/papers/10.21105/joss.01541>.
##
[[57]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[58]]
Wickham H, Seidel D (2022). _scales: Scale Functions for
Visualization_. R package version 1.2.1,
<https://CRAN.R-project.org/package=scales>.
##
[[59]]
Patil I, Makowski D, Ben-Shachar M, Wiernik B, Bacher E, Lüdecke D
(2022). "datawizard: An R Package for Easy Data Preparation and
Statistical Transformations." _CRAN_. R package,
<https://easystats.github.io/datawizard/>.
##
[[60]]
Csárdi G (2022). _cli: Helpers for Developing Command Line Interfaces_.
R package version 3.4.1, <https://CRAN.R-project.org/package=cli>.
##
[[61]]
Gagolewski M (2022). "stringi: Fast and portable character string
processing in R." _Journal of Statistical Software_, *103*(2), 1-59.
doi:10.18637/jss.v103.i02 <https://doi.org/10.18637/jss.v103.i02>.
##
[[62]]
Pedersen T, Nicolae B, François R (2022). _farver: High Performance
Colour Space Manipulation_. R package version 2.1.1,
<https://CRAN.R-project.org/package=farver>.
##
[[63]]
Larmarange J, Sjoberg D (2022). _broom.helpers: Helpers for Model
Coefficients Tibbles_. R package version 1.9.0,
<https://CRAN.R-project.org/package=broom.helpers>.
##
[[64]]
Arel-Bundock V (2022). "modelsummary: Data and Model Summaries in R."

Journal of Statistical Software, *103*(1), 1-23.
doi:10.18637/jss.v103.i01 <https://doi.org/10.18637/jss.v103.i01>.
##
[[65]]
Wickham H, Hester J, Ooms J (2021). _xml2: Parse XML_. R package
version 1.3.3, <https://CRAN.R-project.org/package=xml2>.
##
[[66]]
Wickham H (2021). _ellipsis: Tools for Working with ..._. R package
version 0.3.2, <https://CRAN.R-project.org/package=ellipsis>.
##
[[67]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[68]]
Wickham H, Kuhn M, Vaughan D (2022). _generics: Common S3 Generics not
Provided by Base R Methods Related to Model Fitting_. R package version
0.1.3, <https://CRAN.R-project.org/package=generics>.
##
[[69]]
Wickham H, Henry L, Vaughan D (2022). _vctrs: Vector Helpers_. R
package version 0.4.2, <https://CRAN.R-project.org/package=vctrs>.
##
[[70]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[71]]
Zhu H (2021). _kableExtra: Construct Complex Table with 'kable' and
Pipe Syntax_. R package version 1.3.4,
<https://CRAN.R-project.org/package=kableExtra>.
##
[[72]]
R Core Team (2022). _R: A Language and Environment for Statistical
Computing_. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>.
##
[[73]]
Wickham H (2022). _forcats: Tools for Working with Categorical
Variables (Factors)_. R package version 0.5.2,
<https://CRAN.R-project.org/package=forcats>.
##
[[74]]
Hester J, Bryan J (2022). _glue: Interpreted String Literals_. R
package version 1.6.2, <https://CRAN.R-project.org/package=glue>.
##
[[75]]
Henry L, Wickham H (2022). _purrr: Functional Programming Tools_. R
package version 0.3.5, <https://CRAN.R-project.org/package=purrr>.
##
[[76]]

Müller K (2022). _hms: Pretty Time of Day_. R package version 1.1.2,
<https://CRAN.R-project.org/package=hms>.
##
[[77]]
Chang W (2021). _fastmap: Fast Data Structures_. R package version
1.1.0, <https://CRAN.R-project.org/package=fastmap>.
##
[[78]]
Garbett SP, Stephens J, Simonov K, Xie Y, Dong Z, Wickham H, Horner J,
reikoch, Beasley W, O'Connor B, Warnes GR, Quinn M, Kamvar ZN (2022).
yaml: Methods to Convert R Data to YAML and Back. R package version
2.3.5, <https://CRAN.R-project.org/package=yaml>.
##
[[79]]
Zeileis A, Fisher JC, Hornik K, Ihaka R, McWhite CD, Murrell P,
Stauffer R, Wilke CO (2020). "colorspace: A Toolbox for Manipulating
and Assessing Colors and Palettes." _Journal of Statistical Software_,
96(1), 1-49. doi:10.18637/jss.v096.i01
<https://doi.org/10.18637/jss.v096.i01>.
##
Zeileis A, Hornik K, Murrell P (2009). "Escaping RGBland: Selecting
Colors for Statistical Graphics." _Computational Statistics \& Data
Analysis_, *53*(9), 3259-3270. doi:10.1016/j.csda.2008.11.033
<https://doi.org/10.1016/j.csda.2008.11.033>.
##
Stauffer R, Mayr GJ, Dabernig M, Zeileis A (2009). "Somewhere over the
Rainbow: How to Make Effective Use of Colors in Meteorological
Visualizations." _Bulletin of the American Meteorological Society_,
96(2), 203-216. doi:10.1175/BAMS-D-13-00155.1
<https://doi.org/10.1175/BAMS-D-13-00155.1>.
##
[[80]]
Sjoberg D, Whiting K, Curry M, Lavery J, Larmarange J (2021).
"Reproducible Summary Tables with the gtsummary Package." _The R
Journal_, *13*, 570-580. doi:10.32614/RJ-2021-053
<https://doi.org/10.32614/RJ-2021-053>,
<https://doi.org/10.32614/RJ-2021-053>.
##
[[81]]
Iannone R, Cheng J, Schloerke B, Hughes E (2022). _gt: Easily Create
Presentation-Ready Display Tables_. R package version 0.7.0,
<https://CRAN.R-project.org/package=gt>.
##
[[82]]
Wickham H (2022). _rvest: Easily Harvest (Scrape) Web Pages_. R package
version 1.0.3, <https://CRAN.R-project.org/package=rvest>.
##
[[83]]
Xie Y (2022). _knitr: A General-Purpose Package for Dynamic Report
Generation in R_. R package version 1.40, <https://yihui.org/knitr/>.
##
Xie Y (2015). _Dynamic Documents with R and knitr_, 2nd edition.
Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1498716963,
<https://yihui.org/knitr/>.

```
##
## Xie Y (2014). "knitr: A Comprehensive Tool for Reproducible Research in
## R." In Stodden V, Leisch F, Peng RD (eds.), _Implementing Reproducible
## Computational Research_. Chapman and Hall/CRC. ISBN 978-1466561595,
## <http://www.crcpress.com/product/isbn/9781466561595>.
##
## [[84]]
## Wickham H, Miller E, Smith D (2022). _haven: Import and Export 'SPSS',
## 'Stata' and 'SAS' Files_. R package version 2.5.1,
## <https://CRAN.R-project.org/package=haven>.
##
## [[85]]
## R Core Team (2022). _R: A Language and Environment for Statistical
## Computing_. R Foundation for Statistical Computing, Vienna, Austria.
## <https://www.R-project.org/>.
```

```
knitr::write_bib(file = 'packages.bib') # Constructs a citation file for all packages used in this file
```