Data Science for Biological, Medical and Health Research: Notes for 432

Thomas E. Love, Ph.D.

Built 2018-01-16 21:09:22

Contents

Introduction

These Notes provide a series of examples using R to work through issues that are likely to come up in PQHS/CRSP/MPHP 432.

While these Notes share some of the features of a textbook, they are neither comprehensive nor completely original. The main purpose is to give students in 432 a set of common materials on which to draw during the course. In class, we will sometimes:

- reiterate points made in this document,
- amplify what is here,
- simplify the presentation of things done here,
- use new examples to show some of the same techniques,
- refer to issues not mentioned in this document,

but what we don't (always) do is follow these notes very precisely. We assume instead that you will read the materials and try to learn from them, just as you will attend classes and try to learn from them. We welcome feedback of all kinds on this document or anything else. Just email us at `431-help at case dot edu`, or submit a pull request. Note that we still use `431-help` even though we're now in 432.

What you will mostly find are brief explanations of a key idea or summary, accompanied (most of the time) by R code and a demonstration of the results of applying that code.

Everything you see here is available to you as HTML or PDF. You will also have access to the R Markdown files, which contain the code which generates everything in the document, including all of the R results. We will demonstrate the use of R Markdown (this document is generated with the additional help of an R package called bookdown) and R Studio (the "program" which we use to interface with the R language) in class.

To download the data and R code related to these notes, visit the Data and Code section of the 432 course website.

R Packages used in these notes

Here, we'll load in the packages used in these notes.

```
library(tableone)
library(skimr)
library(broom)
library(tidyverse)
```

Data used in these notes

Here, we'll load in the data sets used in these notes.

```
fakestroke <- read.csv("data/fakestroke.csv") %>% tbl_df
bloodbrain <- read.csv("data/bloodbrain.csv") %>% tbl_df
smartcle1 <- read.csv("data/smartcle1.csv") %>% tbl_df
```

Chapter 1

Building Table 1

Many scientific articles involve direct comparison of results from various exposures, perhaps treatments. In 431, we studied numerous methods, including various sorts of hypothesis tests, confidence intervals, and descriptive summaries, which can help us to understand and compare outcomes in such a setting. One common approach is to present what's often called Table 1. Table 1 provides a summary of the characteristics of a sample, or of groups of samples, which is most commonly used to help understand the nature of the data being compared.

1.1 Two examples from the *New England Journal of Medicine*

1.1.1 A simple Table 1

Table 1 is especially common in the context of clinical research. Consider the excerpt below, from a January 2015 article in the *New England Journal of Medicine* (Tolaney et al., 2015).

Table 1. Baseline Characteristics of the Patients.*

Characteristic	Patients (N=406)
	no. (%)
Age group	
<50 yr	132 (32.5)
50–59 yr	137 (33.7)
60–69 yr	96 (23.6)
≥70 yr	41 (10.1)
Sex	
Female	405 (99.8)
Male	1 (0.2)
Race†	
White	351 (86.5)
Black	28 (6.9)
Asian	11 (2.7)
Other	16 (3.9)

This (partial) table reports baseline characteristics on age group, sex and race, describing 406 patients with

HER2-positive[1] invasive breast cancer that began the protocol therapy. Age, sex and race (along with severity of illness) are the most commonly identified characteristics in a Table 1.

In addition to the measures shown in this excerpt, the full Table also includes detailed information on the primary tumor for each patient, including its size, nodal status and histologic grade. Footnotes tell us that the percentages shown are subject to rounding, and may not total 100, and that the race information was self-reported.

1.1.2 A group comparison

A more typical Table 1 involves a group comparison, for example in this excerpt from Roy et al. (2008). This Table 1 describes a multi-center randomized clinical trial comparing two different approaches to caring for patients with heart failure and atrial fibrillation[2].

Table 1. Baseline Characteristics of the Patients.*

Variable	Rhythm-Control Group (N=682)	Rate-Control Group (N=694)
Male sex (%)	78	85
Age (yr)	66±11	67±11
Body-mass index†	27.8±5.4	28.0±5.1
Nonwhite race (%)‡	16	13
NYHA class III or IV (%)		
At baseline	32	31
During previous 6 mo	76	76
Predominant cardiac diagnosis (%)§		
Coronary artery disease	48	48
Valvular heart disease	5	5
Nonischemic cardiomyopathy	36	39
Congenital heart disease	1	1
Hypertensive heart disease	10	7

The article provides percentages, means and standard deviations across groups, but note that it does not provide p values for the comparison of baseline characteristics. This is a common feature of NEJM reports on randomized clinical trials, where we anticipate that the two groups will be well matched at baseline. Note that the patients in this study were *randomly* assigned to either the rhythm-control group or to the rate-control group, using blocked randomizations stratified by study center.

1.2 The MR CLEAN trial

Berkhemer et al. (2015) reported on the MR CLEAN trial, involving 500 patients with acute ischemic stroke caused by a proximal intracranial arterial occlusion. The trial was conducted at 16 medical centers in the Netherlands, where 233 were randomly assigned to the intervention (intraarterial treatment plus usual care) and 267 to control (usual care alone.) The primary outcome was the modified Rankin scale score at 90 days; this categorical scale measures functional outcome, with scores ranging from 0 (no symptoms) to 6 (death). The fundamental conclusion of Berkhemer et al. (2015) was that in patients with acute ischemic stroke

[1] HER2 = human epidermal growth factor receptor type 2. Over-expression of this occurs in 15-20% of invasive breast cancers, and has been associated with poor outcomes.

[2] The complete Table 1 appears on pages 2668-2669 of Roy et al. (2008), but I have only reproduced the first page and the footnote in this excerpt.

caused by a proximal intracranial occlusion of the anterior circulation, intraarterial treatment administered within 6 hours after stroke onset was effective and safe.

Here's the Table 1 from Berkhemer et al. (2015).

Table 1. Baseline Characteristics of the 500 Patients.*

Characteristic	Intervention (N = 233)	Control (N = 267)
Age — yr		
Median	65.8	65.7
Interquartile range	54.5–76.0	55.5–76.4
Male sex — no. (%)	135 (57.9)	157 (58.8)
NIHSS score†		
Median (interquartile range)	17 (14–21)	18 (14–22)
Range	3–30	4–38
Location of stroke in left hemisphere — no. (%)	116 (49.8)	153 (57.3)
History of ischemic stroke — no. (%)	29 (12.4)	25 (9.4)
Atrial fibrillation — no. (%)	66 (28.3)	69 (25.8)
Diabetes mellitus — no. (%)	34 (14.6)	34 (12.7)
Prestroke modified Rankin scale score — no. (%)‡		
0	190 (81.5)	214 (80.1)
1	21 (9.0)	29 (10.9)
2	12 (5.2)	13 (4.9)
>2	10 (4.3)	11 (4.1)
Systolic blood pressure — mm Hg§	146±26.0	145±24.4
Treatment with IV alteplase — no. (%)	203 (87.1)	242 (90.6)
Time from stroke onset to start of IV alteplase — min		
Median	85	87
Interquartile range	67–110	65–116
ASPECTS — median (interquartile range)¶	9 (7–10)	9 (8–10)
Intracranial arterial occlusion — no./total no. (%)‖		
Intracranial ICA	1/233 (0.4)	3/266 (1.1)
ICA with involvement of the M1 middle cerebral artery segment	59/233 (25.3)	75/266 (28.2)
M1 middle cerebral artery segment	154/233 (66.1)	165/266 (62.0)
M2 middle cerebral artery segment	18/233 (7.7)	21/266 (7.9)
A1 or A2 anterior cerebral artery segment	1/233 (0.4)	2/266 (0.8)
Extracranial ICA occlusion — no./total no. (%)‖**	75/233 (32.2)	70/266 (26.3)
Time from stroke onset to randomization — min††		
Median	204	196
Interquartile range	152–251	149–266
Time from stroke onset to groin puncture — min		
Median	260	NA
Interquartile range	210–313	

The Table was accompanied by the following notes.

* The intervention group was assigned to intraarterial treatment plus usual care, and the control group was assigned to usual care alone. Plus–minus values are means ±SD. ICA denotes internal carotid artery, IV intravenous, and NA not applicable.
† Scores on the National Institutes of Health Stroke Scale (NIHSS) range from 0 to 42, with higher scores indicating more severe neurologic deficits. The NIHSS is a 15-item scale, and values for 30 of the 7500 items were missing (0.4%). The highest number of missing items for a single patient was 6.
‡ Scores on the modified Rankin scale of functional disability range from 0 (no symptoms) to 6 (death). A score of 2 or less indicates functional independence.
§ Data on systolic blood pressure at baseline were missing for one patient assigned to the control group.
¶ The Alberta Stroke Program Early Computed Tomography Score (ASPECTS) is a measure of the extent of stroke. Scores ranges from 0 to 10, with higher scores indicating fewer early ischemic changes. Scores were not available for four patients assigned to the control group: noncontrast computed tomography was not performed in one patient, and three patients had strokes in the territory of the anterior cerebral artery.
‖ Vessel imaging was not performed in one patient in the control group, so the level of occlusion was not known.
** Extracranial ICA occlusions were reported by local investigators.
†† Data were missing for two patients in the intervention group.

1.3 Simulated `fakestroke` data

Consider the simulated data, available on the Data and Code page of our course website in the `fakestroke.csv` file, which I built to let us mirror the Table 1 for MR CLEAN (Berkhemer et al., 2015). The `fakestroke.csv` file contains the following 18 variables for 500 patients.

Variable	Description
`studyid`	Study ID # (z001 through z500)
`trt`	Treatment group (Intervention or Control)
`age`	Age in years
`sex`	Male or Female
`nihss`	NIH Stroke Scale Score (can range from 0-42; higher scores indicate more severe neurological deficits)
`location`	Stroke Location - Left or Right Hemisphere
`hx.isch`	History of Ischemic Stroke (Yes/No)
`afib`	Atrial Fibrillation (1 = Yes, 0 = No)
`dm`	Diabetes Mellitus (1 = Yes, 0 = No)
`mrankin`	Pre-stroke modified Rankin scale score (0, 1, 2 or > 2) indicating functional disability - complete range is 0 (no symptoms) to 6 (death)
`sbp`	Systolic blood pressure, in mm Hg
`iv.altep`	Treatment with IV alteplase (Yes/No)
`time.iv`	Time from stroke onset to start of IV alteplase (minutes) if iv.altep=Yes
`aspects`	Alberta Stroke Program Early Computed Tomography score, which measures extent of stroke from 0 - 10; higher scores indicate fewer early ischemic changes
`ia.occlus`	Intracranial arterial occlusion, based on vessel imaging - five categories[3]
`extra.ica`	Extracranial ICA occlusion (1 = Yes, 0 = No)
`time.rand`	Time from stroke onset to study randomization, in minutes
`time.punc`	Time from stroke onset to groin puncture, in minutes (only if Intervention)

Here's a quick look at the simulated data in `fakestroke`.

[3]The five categories are Intracranial ICA, ICA with involvement of the M1 middle cerebral artery segment, M1 middle cerebral artery segment, M2 middle cerebral artery segment, A1 or A2 anterior cerebral artery segment

```
fakestroke
```

```
# A tibble: 500 x 18
   studyid trt        age sex   nihss location hx.isch  afib    dm mrankin
   <fct>   <fct>    <dbl> <fct> <int> <fct>    <fct>   <int> <int> <fct>
 1 z001    Control   53.0 Male     21 Right    No          0     0 2
 2 z002    Interve~  51.0 Male     23 Left     No          1     0 0
 3 z003    Control   68.0 Fema~    11 Right    No          0     0 0
 4 z004    Control   28.0 Male     22 Left     No          0     0 0
 5 z005    Control   91.0 Male     24 Right    No          0     0 0
 6 z006    Control   34.0 Fema~    18 Left     No          0     0 2
 7 z007    Interve~  75.0 Male     25 Right    No          0     0 0
 8 z008    Control   89.0 Fema~    18 Right    No          0     0 0
 9 z009    Control   75.0 Male     25 Left     No          1     0 2
10 z010    Interve~  26.0 Fema~    27 Right    No          0     0 0
# ... with 490 more rows, and 8 more variables: sbp <int>, iv.altep <fct>,
#   time.iv <int>, aspects <int>, ia.occlus <fct>, extra.ica <int>,
#   time.rand <int>, time.punc <int>
```

1.4 Building Table 1 for `fakestroke`: Attempt 1

Our goal, then, is to take the data in `fakestroke.csv` and use it to generate a Table 1 for the study that compares the 233 patients in the Intervention group to the 267 patients in the Control group, on all of the other variables (except study ID #) available. I'll use the `tableone` package of functions available in R to help me complete this task. We'll make a first attempt, using the `CreateTableOne` function in the `tableone` package. To use the function, we'll need to specify:

- the `vars` or variables we want to place in the rows of our Table 1 (which will include just about everything in the `fakestroke` data except the `studyid` code and the `trt` variable for which we have other plans, and the `time.punc` which applies only to subjects in the Intervention group.)
  - A useful trick here is to use the `dput` function, specifically something like `dput(names(fakestroke))` can be used to generate a list of all of the variables included in the `fakestroke` tibble, and then this can be copied and pasted into the `vars` specification, saving some typing.
- the `strata` which indicates the levels want to use in the columns of our Table 1 (for us, that's `trt`)

```
fs.vars <- c("age", "sex", "nihss", "location",
          "hx.isch", "afib", "dm", "mrankin", "sbp",
          "iv.altep", "time.iv", "aspects",
          "ia.occlus", "extra.ica", "time.rand")

fs.trt <- c("trt")

att1 <- CreateTableOne(data = fakestroke,
                       vars = fs.vars,
                       strata = fs.trt)
print(att1)
```

```
                       Stratified by trt
                        Control        Intervention   p      test
  n                        267            233
  age (mean (sd))        65.38 (16.10)  63.93 (18.09)  0.343
  sex = Male (%)           157 (58.8)     135 (57.9)   0.917
  nihss (mean (sd))      18.08 (4.32)   17.97 (5.04)   0.787
  location = Right (%)     114 (42.7)     117 (50.2)   0.111
```

```
                                                                    
hx.isch = Yes (%)          25 ( 9.4)      29 (12.4)   0.335
afib (mean (sd))         0.26 (0.44)    0.28 (0.45)   0.534
dm (mean (sd))           0.13 (0.33)    0.12 (0.33)   0.923
mrankin (%)                                           0.922
   > 2                     11 ( 4.1)      10 ( 4.3)
   0                      214 (80.1)     190 (81.5)
   1                       29 (10.9)      21 ( 9.0)
   2                       13 ( 4.9)      12 ( 5.2)
sbp (mean (sd))        145.00 (24.40) 146.03 (26.00)  0.647
iv.altep = Yes (%)        242 (90.6)     203 (87.1)   0.267
time.iv (mean (sd))     87.96 (26.01)  98.22 (45.48)  0.003
aspects (mean (sd))      8.65 (1.47)    8.35 (1.64)   0.033
ia.occlus (%)                                         0.795
   A1 or A2                 2 ( 0.8)       1 ( 0.4)
   ICA with M1             75 (28.2)      59 (25.3)
   Intracranial ICA         3 ( 1.1)       1 ( 0.4)
   M1                     165 (62.0)     154 (66.1)
   M2                      21 ( 7.9)      18 ( 7.7)
extra.ica (mean (sd))    0.26 (0.44)    0.32 (0.47)   0.150
time.rand (mean (sd)) 213.88 (70.29) 202.51 (57.33)  0.051
```

1.4.1 Some of this is very useful, and other parts need to be fixed.

1. The 1/0 variables (`afib`, `dm`, `extra.ica`) might be better if they were treated as the factors they are, and reported as the Yes/No variables are reported, with counts and percentages rather than with means and standard deviations.
2. In some cases, we may prefer to re-order the levels of the categorical (factor) variables, particularly the `mrankin` variable, but also the `ia.occlus` variable. It would also be more typical to put the Intervention group to the left and the Control group to the right, so we may need to adjust our `trt` variable's levels accordingly.
3. For each of the quantitative variables (`age`, `nihss`, `sbp`, `time.iv`, `aspects`, `extra.ica`, `time.rand` and `time.punc`) we should make a decision whether a summary with mean and standard deviation is appropriate, or whether we should instead summarize with, say, the median and quartiles. A mean and standard deviation really only yields an appropriate summary when the data are least approximately Normally distributed. This will make the p values a bit more reasonable, too. The `test` column in the first attempt will soon have something useful to tell us.
4. If we'd left in the `time.punc` variable, we'd get some warnings, having to do with the fact that `time.punc` is only relevant to patients in the Intervention group.

1.4.2 `fakestroke` Cleaning Up Categorical Variables

Let's specify each of the categorical variables as categorical explicitly. This helps the `CreateTableOne` function treat them appropriately, and display them with counts and percentages. This includes all of the 1/0, Yes/No and multi-categorical variables.

```
fs.factorvars <- c("sex", "location", "hx.isch", "afib", "dm",
                   "mrankin", "iv.altep", "ia.occlus", "extra.ica")
```

Then we simply add a `factorVars = fs.factorvars` call to the `CreateTableOne` function.

We also want to re-order some of those categorical variables, so that the levels are more useful to us. Specifically, we want to:

- place Intervention before Control in the `trt` variable,
- reorder the `mrankin` scale as 0, 1, 2, > 2, and

- rearrange the `ia.occlus` variable to the order[4] presented in Berkhemer et al. (2015).

To accomplish this, we'll use the `fct_relevel` function from the `forcats` package (loaded with the rest of the core `tidyverse` packages) to reorder our levels manually.

```
fakestroke <- fakestroke %>%
    mutate(trt = fct_relevel(trt, "Intervention", "Control"),
           mrankin = fct_relevel(mrankin, "0", "1", "2", "> 2"),
           ia.occlus = fct_relevel(ia.occlus, "Intracranial ICA",
                                   "ICA with M1", "M1", "M2",
                                   "A1 or A2")
           )
```

1.5 `fakestroke` Table 1: Attempt 2

```
att2 <- CreateTableOne(data = fakestroke,
                       vars = fs.vars,
                       factorVars = fs.factorvars,
                       strata = fs.trt)
print(att2)
```

```
                       Stratified by trt
                        Intervention   Control        p      test
  n                        233            267
  age (mean (sd))         63.93 (18.09)  65.38 (16.10)  0.343
  sex = Male (%)           135 (57.9)     157 (58.8)    0.917
  nihss (mean (sd))       17.97 (5.04)   18.08 (4.32)   0.787
  location = Right (%)     117 (50.2)     114 (42.7)    0.111
  hx.isch = Yes (%)         29 (12.4)      25 ( 9.4)    0.335
  afib = 1 (%)              66 (28.3)      69 (25.8)    0.601
  dm = 1 (%)                29 (12.4)      34 (12.7)    1.000
  mrankin (%)                                           0.922
     0                     190 (81.5)     214 (80.1)
     1                      21 ( 9.0)      29 (10.9)
     2                      12 ( 5.2)      13 ( 4.9)
     > 2                    10 ( 4.3)      11 ( 4.1)
  sbp (mean (sd))        146.03 (26.00) 145.00 (24.40)  0.647
  iv.altep = Yes (%)       203 (87.1)     242 (90.6)    0.267
  time.iv (mean (sd))     98.22 (45.48)  87.96 (26.01)  0.003
  aspects (mean (sd))      8.35 (1.64)    8.65 (1.47)   0.033
  ia.occlus (%)                                         0.795
     Intracranial ICA        1 ( 0.4)       3 ( 1.1)
     ICA with M1            59 (25.3)      75 (28.2)
     M1                     154 (66.1)     165 (62.0)
     M2                      18 ( 7.7)      21 ( 7.9)
     A1 or A2                1 ( 0.4)       2 ( 0.8)
  extra.ica = 1 (%)         75 (32.2)      70 (26.3)    0.179
  time.rand (mean (sd)) 202.51 (57.33) 213.88 (70.29)  0.051
```

The categorical data presentation looks much improved.

[4] We might also have considered reordering the `ia.occlus` factor by its frequency, using the `fct_infreq` function

1.5.1 What summaries should we show?

Now, we'll move on to the issue of making a decision about what type of summary to show for the quantitative variables. Since the `fakestroke` data are just simulated and only match the summary statistics of the original results, not the details, we'll adopt the decisions made by Berkhemer et al. (2015), which were to use medians and interquartile ranges to summarize the distributions of all of the continuous variables **except** systolic blood pressure.

- Specifying certain quantitative variables as *non-normal* causes R to show them with medians and the 25th and 75th percentiles, rather than means and standard deviations, and also causes those variables to be tested using non-parametric tests, like the Wilcoxon signed rank test, rather than the t test. The `test` column indicates this with the word `nonnorm`.
  - In real data situations, what should we do? The answer is to look at the data. I would not make the decision as to which approach to take without first plotting (perhaps in a histogram or a Normal Q-Q plot) the observed distributions in each of the two samples, so that I could make a sound decision about whether Normality was a reasonable assumption. If the means and medians are meaningfully different from each other, this is especially important.
  - To be honest, though, if the variable in question is a relatively unimportant covariate and the p values for the two approaches are nearly the same, I'm not sure that further investigation is especially important,
- Specifying *exact* tests for certain categorical variables (we'll try this for the `location` and `mrankin` variables) can be done, and these changes will be noted in the `test` column, as well.
  - In real data situations, I would rarely be concerned about this issue, and often choose Pearson (approximate) options across the board. This is reasonable so long as the number of subjects falling in each category is reasonably large, say above 10. If not, then an exact test may be an improvement.

To accomplish the Table 1, then, we need to specify which variables should be treated as non-Normal in the `print` statement - notice that we don't need to redo the `CreateTableOne` for this change.

```
print(att2,
      nonnormal = c("age", "nihss", "time.iv", "aspects", "time.rand"),
      exact = c("location", "mrankin"))
```

```
                        Stratified by trt
                         Intervention            Control
  n                         233                     267
  age (median [IQR])       65.80 [54.50, 76.00]    65.70 [55.75, 76.20]
  sex = Male (%)            135 (57.9)              157 (58.8)
  nihss (median [IQR])     17.00 [14.00, 21.00]    18.00 [14.00, 22.00]
  location = Right (%)      117 (50.2)              114 (42.7)
  hx.isch = Yes (%)          29 (12.4)               25 ( 9.4)
  afib = 1 (%)               66 (28.3)               69 (25.8)
  dm = 1 (%)                 29 (12.4)               34 (12.7)
  mrankin (%)
     0                      190 (81.5)              214 (80.1)
     1                       21 ( 9.0)               29 (10.9)
     2                       12 ( 5.2)               13 ( 4.9)
     > 2                     10 ( 4.3)               11 ( 4.1)
  sbp (mean (sd))          146.03 (26.00)          145.00 (24.40)
  iv.altep = Yes (%)        203 (87.1)              242 (90.6)
  time.iv (median [IQR])    85.00 [67.00, 110.00]   87.00 [65.00, 116.00]
  aspects (median [IQR])     9.00 [7.00, 10.00]      9.00 [8.00, 10.00]
  ia.occlus (%)
     Intracranial ICA          1 ( 0.4)                3 ( 1.1)
     ICA with M1              59 (25.3)               75 (28.2)
```

```
   M1                        154 (66.1)              165 (62.0)
   M2                         18 ( 7.7)               21 ( 7.9)
   A1 or A2                    1 ( 0.4)                2 ( 0.8)
extra.ica = 1 (%)             75 (32.2)               70 (26.3)
time.rand (median [IQR]) 204.00 [152.00, 249.50] 196.00 [149.00, 266.00]
                        Stratified by trt
                         p      test
n
age (median [IQR])        0.579 nonnorm
sex = Male (%)            0.917
nihss (median [IQR])      0.453 nonnorm
location = Right (%)      0.106 exact
hx.isch = Yes (%)         0.335
afib = 1 (%)              0.601
dm = 1 (%)                1.000
mrankin (%)               0.917 exact
   0
   1
   2
   > 2
sbp (mean (sd))           0.647
iv.altep = Yes (%)        0.267
time.iv (median [IQR])    0.596 nonnorm
aspects (median [IQR])    0.075 nonnorm
ia.occlus (%)             0.795
   Intracranial ICA
   ICA with M1
   M1
   M2
   A1 or A2
extra.ica = 1 (%)         0.179
time.rand (median [IQR])  0.251 nonnorm
```

1.6 Obtaining a more detailed Summary

If this was a real data set, we'd want to get a more detailed description of the data to make decisions about things like potentially collapsing categories of a variable, or whether or not a normal distribution was useful for a particular continuous variable, etc. You can do this with the `summary` command applied to a created Table 1, which shows, among other things, the effect of changing from normal to non-normal p values for continuous variables, and from approximate to "exact" p values for categorical factors.

Again, as noted above, in a real data situation, we'd want to plot the quantitative variables (within each group) to make a smart decision about whether a t test or Wilcoxon approach is more appropriate.

Note in the summary below that we have some missing values here. Often, we'll present this information within the Table 1, as well.

```
summary(att2)
```

```
     ### Summary of continuous variables ###

trt: Intervention
            n miss p.miss mean sd median p25 p75 min max  skew  kurt
```

```
age       233    0    0.0   64 18     66  54  76  23  96 -0.34 -0.52
nihss     233    0    0.0   18  5     17  14  21  10  28  0.48 -0.74
sbp       233    0    0.0  146 26    146 129 164  78 214 -0.07 -0.22
time.iv   233   30   12.9   98 45     85  67 110  42 218  1.03  0.08
aspects   233    0    0.0    8  2      9   7  10   5  10 -0.56 -0.98
time.rand 233    2    0.9  203 57    204 152 250 100 300  0.01 -1.16
------------------------------------------------------------
trt: Control
            n miss p.miss mean sd median p25 p75 min max   skew  kurt
age       267    0    0.0   65 16     66  56  76  24  94 -0.296 -0.28
nihss     267    0    0.0   18  4     18  14  22  11  25  0.017 -1.24
sbp       267    1    0.4  145 24    145 128 161  82 231  0.156  0.08
time.iv   267   25    9.4   88 26     87  65 116  44 130  0.001 -1.32
aspects   267    4    1.5    9  1      9   8  10   5  10 -1.071  0.36
time.rand 267    0    0.0  214 70    196 149 266 120 360  0.508 -0.93

p-values
              pNormal pNonNormal
age       0.342813660 0.57856976
nihss     0.787487252 0.45311695
sbp       0.647157646 0.51346132
time.iv   0.003073372 0.59641104
aspects   0.032662901 0.07464683
time.rand 0.050803672 0.25134327

Standardize mean differences
              1 vs 2
age       0.08478764
nihss     0.02405390
sbp       0.04100833
time.iv   0.27691223
aspects   0.19210662
time.rand 0.17720957

================================================================================

     ### Summary of categorical variables ###

trt: Intervention
       var   n miss p.miss            level freq percent cum.percent
       sex 233    0    0.0           Female   98    42.1        42.1
                                      Male  135    57.9       100.0

  location 233    0    0.0             Left  116    49.8        49.8
                                      Right  117    50.2       100.0

   hx.isch 233    0    0.0               No  204    87.6        87.6
                                        Yes   29    12.4       100.0

      afib 233    0    0.0                0  167    71.7        71.7
                                          1   66    28.3       100.0

        dm 233    0    0.0                0  204    87.6        87.6
                                          1   29    12.4       100.0
```

```
   mrankin 233    0    0.0               0  190    81.5        81.5
                                         1   21     9.0        90.6
                                         2   12     5.2        95.7
                                       > 2   10     4.3       100.0

  iv.altep 233    0    0.0              No   30    12.9        12.9
                                       Yes  203    87.1       100.0

 ia.occlus 233    0    0.0 Intracranial ICA    1     0.4         0.4
                           ICA with M1   59    25.3        25.8
                                   M1  154    66.1        91.8
                                   M2   18     7.7        99.6
                              A1 or A2    1     0.4       100.0

 extra.ica 233    0    0.0               0  158    67.8        67.8
                                         1   75    32.2       100.0

-------------------------------------------------------
trt: Control
       var   n miss p.miss           level freq percent cum.percent
       sex 267    0    0.0          Female  110    41.2        41.2
                                     Male  157    58.8       100.0

  location 267    0    0.0            Left  153    57.3        57.3
                                     Right  114    42.7       100.0

   hx.isch 267    0    0.0              No  242    90.6        90.6
                                       Yes   25     9.4       100.0

      afib 267    0    0.0               0  198    74.2        74.2
                                         1   69    25.8       100.0

        dm 267    0    0.0               0  233    87.3        87.3
                                         1   34    12.7       100.0

   mrankin 267    0    0.0               0  214    80.1        80.1
                                         1   29    10.9        91.0
                                         2   13     4.9        95.9
                                       > 2   11     4.1       100.0

  iv.altep 267    0    0.0              No   25     9.4         9.4
                                       Yes  242    90.6       100.0

 ia.occlus 267    1    0.4 Intracranial ICA    3     1.1         1.1
                           ICA with M1   75    28.2        29.3
                                   M1  165    62.0        91.4
                                   M2   21     7.9        99.2
                              A1 or A2    2     0.8       100.0

 extra.ica 267    1    0.4               0  196    73.7        73.7
                                         1   70    26.3       100.0
```

```
p-values
            pApprox    pExact
sex       0.9171387 0.8561188
location  0.1113553 0.1056020
hx.isch   0.3352617 0.3124683
afib      0.6009691 0.5460206
dm        1.0000000 1.0000000
mrankin   0.9224798 0.9173657
iv.altep  0.2674968 0.2518374
ia.occlus 0.7945580 0.8189090
extra.ica 0.1793385 0.1667574

Standardize mean differences
               1 vs 2
sex       0.017479025
location  0.151168444
hx.isch   0.099032275
afib      0.055906317
dm        0.008673478
mrankin   0.062543164
iv.altep  0.111897009
ia.occlus 0.117394890
extra.ica 0.129370206
```

In this case, I have simulated the data to mirror the results in the published Table 1 for this study. In no way have I captured the full range of the real data, or any of the relationships in that data, so it's more important here to see what's available in the analysis, rather than to interpret it closely in the clinical context.

1.7 Exporting the Completed Table 1 from R to Excel or Word

Once you've built the table and are generally satisfied with it, you'll probably want to be able to drop it into Excel or Word for final cleanup.

1.7.1 Approach A: Save and open in Excel

One option is to **save the Table 1** to a `.csv` file, which you can then open directly in Excel. This is the approach I generally use. Note the addition of some `quote`, `noSpaces` and `printToggle` selections here.

```
fs.table1save <- print(att2,
      nonnormal = c("age", "nihss", "time.iv", "aspects", "time.rand"),
      exact = c("location", "mrankin"),
      quote = FALSE, noSpaces = TRUE, printToggle = FALSE)

write.csv(fs.table1save, file = "fs-table1.csv")
```

When I then open the `fs-table1.csv` file in Excel, it looks like this:

	A	B	C	D	E
1		Intervention	Control	p	test
2	n	233	267		
3	age (median [IQR])	65.80 [54.50, 76.00]	65.70 [55.75, 76.20]	0.579	nonnorm
4	sex = Male (%)	135 (57.9)	157 (58.8)	0.917	
5	nihss (median [IQR])	17.00 [14.00, 21.00]	18.00 [14.00, 22.00]	0.453	nonnorm
6	location = Right (%)	117 (50.2)	114 (42.7)	0.111	
7	hx.isch = Yes (%)	29 (12.4)	25 (9.4)	0.335	
8	afib = 1 (%)	66 (28.3)	69 (25.8)	0.601	
9	dm = 1 (%)	29 (12.4)	34 (12.7)	1	
10	mrankin (%)			0.922	
11	0	190 (81.5)	214 (80.1)		
12	1	21 (9.0)	29 (10.9)		
13	2	12 (5.2)	13 (4.9)		
14	> 2	10 (4.3)	11 (4.1)		
15	sbp (mean (sd))	146.03 (26.00)	145.00 (24.40)	0.647	
16	iv.altep = Yes (%)	203 (87.1)	242 (90.6)	0.267	
17	time.iv (median [IQR])	85.00 [67.00, 110.00]	87.00 [65.00, 116.00]	0.596	nonnorm
18	aspects (median [IQR])	9.00 [7.00, 10.00]	9.00 [8.00, 10.00]	0.075	nonnorm
19	ia.occlus (%)			0.795	
20	Intracranial ICA	1 (0.4)	3 (1.1)		
21	ICA with M1	59 (25.3)	75 (28.2)		
22	M1	154 (66.1)	165 (62.0)		
23	M2	18 (7.7)	21 (7.9)		
24	A1 or A2	1 (0.4)	2 (0.8)		
25	extra.ica = 1 (%)	75 (32.2)	70 (26.3)	0.179	
26	time.rand (median [IQR])	204.00 [152.00, 249.50]	196.00 [149.00, 266.00]	0.251	nonnorm
27	time.punc (median [IQR])	260.00 [212.00, 313.00]	NA [NA, NA]	NA	nonnorm
28					

And from here, I can either drop it directly into Word, or present it as is, or start tweaking it to meet formatting needs.

1.7.2 Approach B: Produce the Table so you can cut and paste it

```
print(att2,
      nonnormal = c("age", "nihss", "time.iv", "aspects", "time.rand"),
      exact = c("location", "mrankin"),
      quote = TRUE, noSpaces = TRUE)
```

This will look like a mess by itself, but if you:

1. copy and paste that mess into Excel
2. select Text to Columns from the Data menu
3. select Delimited, then Space and select Treat consecutive delimiters as one

you should get something usable again.

Or, in Word,

1. insert the text

2. select the text with your mouse
3. select Insert … Table … Convert Text to Table
4. place a quotation mark in the "Other" area under Separate text at …

After dropping blank columns, the result looks pretty good.

1.8 A Controlled Biological Experiment - The Blood-Brain Barrier

My source for the data and the following explanatory paragraph is page 307 from Ramsey and Schafer (2002). The original data come from Barnett et al. (1995).

> The human brain (and that of rats, coincidentally) is protected from the bacteria and toxins that course through the bloodstream by something called the blood-brain barrier. After a method of disrupting the barrier was developed, researchers tested this new mechanism, as follows. A series of 34 rats were inoculated with human lung cancer cells to induce brain tumors. After 9-11 days they were infused with either the barrier disruption (BD) solution or, as a control, a normal saline (NS) solution. Fifteen minutes later, the rats received a standard dose of a particular therapeutic antibody (L6-F(ab')2. The key measure of the effectiveness of transmission across the brain-blood barrier is the ratio of the antibody concentration in the brain tumor to the antibody concentration in normal tissue outside the brain. The rats were then sacrificed, and the amounts of antibody in the brain tumor and in normal tissue from the liver were measured. The study's primary objective is to determine whether the antibody concentration in the tumor increased when the blood-barrier disruption infusion was given, and if so, by how much?

1.9 The `bloodbrain.csv` file

Consider the data, available on the Data and Code page of our course website in the `bloodbrain.csv` file, which includes the following variables:

Variable	Description
`case`	identification number for the rat (1 - 34)
`brain`	an outcome: Brain tumor antibody count (per gram)
`liver`	an outcome: Liver antibody count (per gram)
`tlratio`	an outcome: tumor / liver concentration ratio
`solution`	the treatment: BD (barrier disruption) or NS (normal saline)
`sactime`	a design variable: Sacrifice time (hours; either 0.5, 3, 24 or 72)
`postin`	covariate: Days post-inoculation of lung cancer cells (9, 10 or 11)
`sex`	covariate: M or F
`wt.init`	covariate: Initial weight (grams)
`wt.loss`	covariate: Weight loss (grams)
`wt.tumor`	covariate: Tumor weight (10^{-4} grams)

And here's what the data look like in R.

```
bloodbrain
```

```
# A tibble: 34 x 11
    case  brain   liver tlratio solution sactime postin sex   wt.init
   <int>  <int>   <int>   <dbl> <fct>      <dbl>  <int> <fct>   <int>
 1     1  41081 1456164  0.0282 BD         0.500     10 F         239
```

```
 2     2  44286 1602171  0.0276 BD         0.500     10 F         225
 3     3 102926 1601936  0.0642 BD         0.500     10 F         224
 4     4  25927 1776411  0.0146 BD         0.500     10 F         184
 5     5  42643 1351184  0.0316 BD         0.500     10 F         250
 6     6  31342 1790863  0.0175 NS         0.500     10 F         196
 7     7  22815 1633386  0.0140 NS         0.500     10 F         200
 8     8  16629 1618757  0.0103 NS         0.500     10 F         273
 9     9  22315 1567602  0.0142 NS         0.500     10 F         216
10    10  77961 1060057  0.0735 BD         3.00      10 F         267
# ... with 24 more rows, and 2 more variables: wt.loss <dbl>, wt.tumor
#   <int>
```

1.10 A Table 1 for `bloodbrain`

Barnett et al. (1995) did not provide a Table 1 for these data, so let's build one to compare the two `solutions` (`BD` vs. `NS`) on the covariates and outcomes, plus the natural logarithm of the tumor/liver concentration ratio (`tlratio`). We'll opt to treat the sacrifice time (`sactime`) and the days post-inoculation of lung cancer cells (`postin`) as categorical rather than quantitative variables.

```
bloodbrain <- bloodbrain %>%
    mutate(logTL = log(tlratio))

dput(names(bloodbrain))
```

```
c("case", "brain", "liver", "tlratio", "solution", "sactime",
"postin", "sex", "wt.init", "wt.loss", "wt.tumor", "logTL")
```

OK - there's the list of variables we'll need. I'll put the outcomes at the bottom of the table.

```
bb.vars <- c("sactime", "postin", "sex", "wt.init", "wt.loss",
             "wt.tumor", "brain", "liver", "tlratio", "logTL")

bb.factors <- c("sactime", "sex", "postin")

bb.att1 <- CreateTableOne(data = bloodbrain,
                          vars = bb.vars,
                          factorVars = bb.factors,
                          strata = c("solution"))
summary(bb.att1)
```

```
     ### Summary of continuous variables ###

solution: BD
          n miss p.miss   mean    sd median    p25   p75    min   max
wt.init  17    0      0    243 3e+01  2e+02  2e+02 3e+02  2e+02 3e+02
wt.loss  17    0      0      3 5e+00  4e+00  1e+00 6e+00 -5e+00 1e+01
wt.tumor 17    0      0    157 8e+01  2e+02  1e+02 2e+02  2e+01 4e+02
brain    17    0      0  56043 3e+04  5e+04  4e+04 8e+04  6e+03 1e+05
liver    17    0      0 672577 7e+05  6e+05  2e+04 1e+06  2e+03 2e+06
tlratio  17    0      0      2 3e+00  1e-01  6e-02 3e+00  1e-02 9e+00
logTL    17    0      0     -1 2e+00 -2e+00 -3e+00 1e+00 -4e+00 2e+00
          skew kurt
wt.init  -0.39  0.7
wt.loss  -0.10  0.2
```

```
wt.tumor  0.53  1.0
brain     0.29 -0.6
liver     0.35 -1.7
tlratio   1.58  1.7
logTL     0.08 -1.7
------------------------------------------------------------
solution: NS
          n miss p.miss   mean    sd median    p25    p75    min   max
wt.init  17    0      0    240 3e+01  2e+02  2e+02  3e+02  2e+02 3e+02
wt.loss  17    0      0      4 4e+00  3e+00  2e+00  7e+00 -4e+00 1e+01
wt.tumor 17    0      0    209 1e+02  2e+02  2e+02  3e+02  3e+01 5e+02
brain    17    0      0  23887 1e+04  2e+04  1e+04  3e+04  1e+03 5e+04
liver    17    0      0 664975 7e+05  7e+05  2e+04  1e+06  9e+02 2e+06
tlratio  17    0      0      1 2e+00  5e-02  3e-02  9e-01  1e-02 7e+00
logTL    17    0      0     -2 2e+00 -3e+00 -3e+00 -7e-02 -5e+00 2e+00
          skew  kurt
wt.init   0.33 -0.48
wt.loss  -0.09  0.08
wt.tumor  0.63  0.77
brain     0.30 -0.35
liver     0.40 -1.56
tlratio   2.27  4.84
logTL     0.27 -1.61

p-values
             pNormal  pNonNormal
wt.init  0.807308940 0.641940278
wt.loss  0.683756156 0.876749808
wt.tumor 0.151510151 0.190482094
brain    0.001027678 0.002579901
liver    0.974853609 0.904045603
tlratio  0.320501715 0.221425879
logTL    0.351633525 0.221425879

Standardize mean differences
             1 vs 2
wt.init  0.08435244
wt.loss  0.14099823
wt.tumor 0.50397184
brain    1.23884159
liver    0.01089667
tlratio  0.34611465
logTL    0.32420504

================================================================================

     ### Summary of categorical variables ###

solution: BD
     var  n miss p.miss level freq percent cum.percent
 sactime 17    0    0.0   0.5    5    29.4        29.4
                          3    4    23.5        52.9
                         24    4    23.5        76.5
                         72    4    23.5       100.0
```

```
  postin 17    0    0.0     9    1     5.9         5.9
                            10   14    82.4        88.2
                            11    2    11.8       100.0

     sex 17    0    0.0     F   13    76.5        76.5
                            M    4    23.5       100.0

------------------------------------------------------------
solution: NS
     var  n miss p.miss level freq percent cum.percent
 sactime 17    0    0.0   0.5    4    23.5        23.5
                             3    5    29.4        52.9
                            24    4    23.5        76.5
                            72    4    23.5       100.0

  postin 17    0    0.0     9    2    11.8        11.8
                            10   13    76.5        88.2
                            11    2    11.8       100.0

     sex 17    0    0.0     F   13    76.5        76.5
                            M    4    23.5       100.0


p-values
          pApprox pExact
sactime 0.9739246      1
postin  0.8309504      1
sex     1.0000000      1

Standardize mean differences
           1 vs 2
sactime 0.1622214
postin  0.2098877
sex     0.0000000
```

Note that, in this particular case, the decisions we make about normality vs. non-normality (for quantitative variables) and the decisions we make about approximate vs. exact testing (for categorical variables) won't actually change the implications of the p values. Each approach gives similar results for each variable. Of course, that's not always true.

1.10.1 Generate final Table 1 for `bloodbrain`

I'll choose to treat `tlratio` and its logarithm as non-Normal, but otherwise, use t tests, but admittedly, that's an arbitrary decision, really.

```
print(bb.att1, nonnormal = c("tlratio", "logTL"))
```

```
                       Stratified by solution
                        BD                      NS
  n                            17                         17
  sactime (%)
     0.5                        5 (29.4)                   4 (23.5)
     3                          4 (23.5)                   5 (29.4)
     24                         4 (23.5)                   4 (23.5)
```

```
   72                          4 (23.5)               4 (23.5)
  postin (%)
   9                           1 ( 5.9)               2 (11.8)
   10                         14 (82.4)              13 (76.5)
   11                          2 (11.8)               2 (11.8)
  sex = M (%)                  4 (23.5)               4 (23.5)
  wt.init (mean (sd))        242.82 (27.23)         240.47 (28.54)
  wt.loss (mean (sd))          3.34 (4.68)            3.94 (3.88)
  wt.tumor (mean (sd))       157.29 (84.00)         208.53 (116.68)
  brain (mean (sd))        56043.41 (33675.40)    23887.18 (14610.53)
  liver (mean (sd))       672577.35 (694479.58)  664975.47 (700773.13)
  tlratio (median [IQR])       0.12 [0.06, 2.84]      0.05 [0.03, 0.94]
  logTL (median [IQR])        -2.10 [-2.74, 1.04]    -2.95 [-3.41, -0.07]
                        Stratified by solution
                         p      test
  n
  sactime (%)             0.974
   0.5
   3
   24
   72
  postin (%)              0.831
   9
   10
   11
  sex = M (%)             1.000
  wt.init (mean (sd))     0.807
  wt.loss (mean (sd))     0.684
  wt.tumor (mean (sd))    0.152
  brain (mean (sd))       0.001
  liver (mean (sd))       0.975
  tlratio (median [IQR])  0.221 nonnorm
  logTL (median [IQR])    0.221 nonnorm
```

Or, we can get an Excel-readable version, using

```
bb.t1 <- print(bb.att1, nonnormal = c("tlratio", "logTL"), quote = FALSE,
               noSpaces = TRUE, printToggle = FALSE)

write.csv(bb.t1, file = "bb-table1.csv")
```

which, when dropped into Excel, will look like this:

	A	B	C	D	E
1		BD	NS	p	test
2	n	17	17		
3	sex = M (%)	4 (23.5)	4 (23.5)	1	
4	sactime (%)			0.974	
5	0.5	5 (29.4)	4 (23.5)		
6	3	4 (23.5)	5 (29.4)		
7	24	4 (23.5)	4 (23.5)		
8	72	4 (23.5)	4 (23.5)		
9	postin (%)			0.831	
10	9	1 (5.9)	2 (11.8)		
11	10	14 (82.4)	13 (76.5)		
12	11	2 (11.8)	2 (11.8)		
13	wt.init (mean (sd))	242.82 (27.23)	240.47 (28.54)	0.807	
14	wt.loss (mean (sd))	3.34 (4.68)	3.94 (3.88)	0.684	
15	wt.tumor (mean (sd))	157.29 (84.00)	208.53 (116.68)	0.152	
16	brain (mean (sd))	56043.41 (33675.40)	23887.18 (14610.53)	0.001	
17	liver (mean (sd))	672577.35 (694479.58)	664975.47 (700773.13)	0.975	
18	tlratio (median [IQR])	0.12 [0.06, 2.84]	0.05 [0.03, 0.94]	0.221	nonnorm
19	logTL (median [IQR])	-2.10 [-2.74, 1.04]	-2.95 [-3.41, -0.07]	0.221	nonnorm
20					

One thing I would definitely clean up here, in practice, is to change the presentation of the p value for `sex` from 1 to > 0.99, or just omit it altogether. I'd also drop the `computer-ese` where possible, add units for the measures, round **a lot**, identify the outcomes carefully, and use notes to indicate deviations from the main approach.

1.10.2 A More Finished Version (after Cleanup in Word)

Table 1. Comparing Rats Receiving BD to those Receiving NS on Available Covariates and Design Variables, and Key Outcomes

	Barrier Disruption (BD: treatment)	Normal Saline (NS: control)	p
# of Rats	17	17	
Sex = Male	4 (23.5)	4 (23.5)	-
Sacrifice Time (hours)			0.97
0.5	5 (29.4)	4 (23.5)	
3	4 (23.5)	5 (29.4)	
24	4 (23.5)	4 (23.5)	
72	4 (23.5)	4 (23.5)	
Days post-inoculation of lung cancer cells			0.83
9	1 (5.9)	2 (11.8)	
10	14 (82.4)	13 (76.5)	
11	2 (11.8)	2 (11.8)	
Initial Weight (g)	243 (27)	240 (29)	0.81
Weight Loss (g)	3.3 (4.7)	3.9 (3.9)	0.68
Tumor Weight (10^{-4} g)	157.3 (84.0)	208.5 (116.7)	0.15
Key Outcomes: mean (sd) unless otherwise indicated			
Brain Tumor Antibody Count (per g)	56,043 (33,675)	23,887 (14,611)	0.001
Liver Antibody Count (per g)	672,577 (694,480)	664,975 (700,773)	0.98
Tumor/Liver Ratio (median [Q25, Q75])	0.12 [0.06, 2.84]	0.05 [0.03, 0.94]	0.22
Natural Log of Tumor/Liver Ratio (median [Q25, Q75])	-2.10 [-2.74, 1.04]	-2.95 [-3.41, -0.07]	0.22

Table 1 Notes:

- Categorical variables are summarized with counts, percentages and p values based on approximate chi-square tests.
- Continuous variables, unless otherwise indicated, are summarized with means, standard deviations and p values based on t tests.
- The Tumor / Liver ratio and its natural logarithm are summarized with the median and quartiles and a p value from a non-parametric (Wilcoxon signed rank) test.

Chapter 2

Linear Regression on a small SMART data set

2.1 BRFSS and SMART

The Centers for Disease Control analyzes Behavioral Risk Factor Surveillance System (BRFSS) survey data for specific metropolitan and micropolitan statistical areas (MMSAs) in a program called the Selected Metropolitan/Micropolitan Area Risk Trends of BRFSS (SMART BRFSS.)

In this work, we will focus on data from the 2016 SMART, and in particular on data from the Cleveland-Elyria, OH, Metropolitan Statistical Area. The purpose of this survey is to provide localized health information that can help public health practitioners identify local emerging health problems, plan and evaluate local responses, and efficiently allocate resources to specific needs.

2.1.1 Key resources

- the full data are available in the form of the 2016 SMART BRFSS MMSA Data, found in a zipped SAS Transport Format file. The data were released in August 2017.
- the MMSA Variable Layout PDF which simply lists the variables included in the data file
- the Calculated Variables PDF which describes the risk factors by data variable names - there is also an online summary matrix of these calculated variables, as well.
- the lengthy 2016 Survey Questions PDF which lists all questions asked as part of the BRFSS in 2016
- the enormous Codebook for the 2016 BRFSS Survey PDF which identifies the variables by name for us.

Later this term, we'll use all of those resources to help construct a more complete data set than we'll study today. I'll also demonstrate how I built the `smartcle1` data set that we'll use in this Chapter.

2.2 The `smartcle1` data: Cookbook

The `smartcle1.csv` data file available on the Data and Code page of our website describes information on 11 variables for 1036 respondents to the BRFSS 2016, who live in the Cleveland-Elyria, OH, Metropolitan Statistical Area. The variables in the `smartcle1.csv` file are listed below, along with (in some cases) the BRFSS items that generate these responses.

Variable	Description
`SEQNO`	respondent identification number (all begin with 2016)

Variable	Description
physhealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
menthealth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
poorhealth	During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
genhealth	Would you say that in general, your health is … (five categories: Excellent, Very Good, Good, Fair or Poor)
bmi	Body mass index, in kg/m^2
female	Sex, 1 = female, 0 = male
internet30	Have you used the internet in the past 30 days? (1 = yes, 0 = no)
exerany	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise? (1 = yes, 0 = no)
sleephrs	On average, how many hours of sleep do you get in a 24-hour period?
alcdays	How many days during the past 30 days did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor?

```
str(smartcle1)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':   1036 obs. of  11 variables:
 $ SEQNO     : num  2.02e+09 2.02e+09 2.02e+09 2.02e+09 2.02e+09 ...
 $ physhealth: int  0 0 1 0 5 4 2 2 0 0 ...
 $ menthealth: int  0 0 5 0 0 18 0 3 0 0 ...
 $ poorhealth: int  NA NA 0 NA 0 6 0 0 NA NA ...
 $ genhealth : Factor w/ 5 levels "1_Excellent",..: 2 1 2 3 1 2 3 3 2 3 ...
 $ bmi       : num  26.7 23.7 26.9 21.7 24.1 ...
 $ female    : int  1 0 0 1 0 0 1 1 0 0 ...
 $ internet30: int  1 1 1 1 1 1 1 1 1 1 ...
 $ exerany   : int  1 1 0 1 1 1 1 1 1 0 ...
 $ sleephrs  : int  6 6 8 9 7 5 9 7 7 7 ...
 $ alcdays   : int  1 4 4 3 2 28 4 2 4 25 ...
```

2.3 smartcle2: Omitting Missing Observations: Complete-Case Analyses

For the purpose of fitting our first few models, we will eliminate the missingness problem, and look only at the *complete cases* in our smartcle1 data.

To inspect the missingness in our data, we might consider using the skim function from the skimr package. We'll exclude the respondent identifier code (SEQNO) from this summary as uninteresting.

```
smartcle1 %>%
    skim(-SEQNO)
```

```
Skim summary statistics
 n obs: 1036
 n variables: 11
```

```
Variable type: factor
  variable missing complete    n n_unique
 genhealth       3     1033 1036        5
                           top_counts ordered
 2_V: 350, 3_G: 344, 1_E: 173, 4_F: 122   FALSE

Variable type: integer
   variable missing complete    n mean   sd p0 p25 median p75 p100
    alcdays      46      990 1036 4.65 8.05  0   0      1   4   30
    exerany       3     1033 1036 0.76 0.43  0   1      1   1    1
     female       0     1036 1036 0.6  0.49  0   0      1   1    1
 internet30       6     1030 1036 0.81 0.39  0   1      1   1    1
 menthealth      11     1025 1036 2.72 6.82  0   0      0   2   30
 physhealth      17     1019 1036 3.97 8.67  0   0      0   2   30
 poorhealth     543      493 1036 4.07 8.09  0   0      0   3   30
   sleephrs       8     1028 1036 7.02 1.53  1   6      7   8   20
     hist
 <U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2582><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2587>
 <U+2585><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2587>
 <U+2582><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2587>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
 <U+2581><U+2581><U+2587><U+2581><U+2581><U+2581><U+2581><U+2581>

Variable type: numeric
 variable missing complete    n  mean   sd    p0  p25 median   p75  p100
      bmi      84      952 1036 27.89 6.47 12.71 23.7  26.68 30.53 66.06
     hist
 <U+2581><U+2587><U+2587><U+2582><U+2581><U+2581><U+2581><U+2581>
```

Now, we'll create a new tibble called `smartcle2` which contains every variable except `poorhealth`, and which includes all respondents with complete data on the variables (other than `poorhealth`). We'll store those observations with complete data in the `smartcle2` tibble.

```
smartcle2 <- smartcle1 %>%
    select(-poorhealth) %>%
    filter(complete.cases(.))

smartcle2
```

```
# A tibble: 896 x 10
     SEQNO physhealth menthealth genhealth   bmi female internet30 exerany
     <dbl>      <int>      <int> <fct>     <dbl>  <int>      <int>   <int>
 1  2.02e9          0          0 2_VeryGo~  26.7      1          1       1
 2  2.02e9          0          0 1_Excell~  23.7      0          1       1
 3  2.02e9          1          5 2_VeryGo~  26.9      0          1       0
 4  2.02e9          0          0 3_Good     21.7      1          1       1
 5  2.02e9          5          0 1_Excell~  24.1      0          1       1
 6  2.02e9          4         18 2_VeryGo~  27.6      0          1       1
 7  2.02e9          2          0 3_Good     25.7      1          1       1
 8  2.02e9          2          3 3_Good     28.5      1          1       1
 9  2.02e9          0          0 2_VeryGo~  28.6      0          1       1
```

```
10  2.02e9          0          0 3_Good     23.1      0          1       0
# ... with 886 more rows, and 2 more variables: sleephrs <int>, alcdays
#   <int>
```

Note that there are only 896 respondents with **complete** data on the 10 variables (excluding `poorhealth`) in the `smartcle2` tibble, as compared to our original `smartcle1` data which described 1036 respondents and 11 variables, but with lots of missing data.

2.4 A Small Study

We'll begin by investigating the problem of predicting `physhealth`, at first with just two predictor variables: `exerany` and `bmi`, in our new `smartcle2` data set.

- The outcome of interest is `physhealth`.
- Inputs to the regression model are:
  - `exerany` = 1 if the subject exercised in the past 30 days, and 0 if they didn't
  - `bmi` = body mass index (treated as qualitative and continuous)

2.4.1 Some exploratory data analysis

Counting things can be amazingly useful.

2.4.1.1 How many respondents had exercised in the past 30 days?

```
smartcle2 %>% count(exerany)
```

```
# A tibble: 2 x 2
  exerany     n
    <int> <int>
1       0   209
2       1   687
```

This counting approach works for quantitative variables with discrete sets of possible values, like `physhealth`, which must be an integer between 0 and 30.

2.4.1.2 What's the distribution of `physhealth`?

```
smartcle2 %>% count(physhealth)
```

```
# A tibble: 19 x 2
   physhealth     n
        <int> <int>
 1          0   591
 2          1    35
 3          2    55
 4          3    22
 5          4    12
 6          5    25
 7          6     4
 8          7    20
 9          8     1
10          9     1
```

```
11          10    18
12          12     3
13          14    10
14          15    14
15          18     1
16          20     8
17          21     1
18          25     1
19          30    74
```

2.4.1.3 How many of the respondents have a BMI below 30?

```
smartcle2 %>% count(bmi < 30)
```

```
# A tibble: 2 x 2
  `bmi < 30`     n
  <lgl>      <int>
1 F            253
2 T            643
```

2.4.1.4 How many of the respondents who have a BMI < 30 exercised?

```
smartcle2 %>% count(bmi < 30, exerany)
```

```
# A tibble: 4 x 3
  `bmi < 30` exerany     n
  <lgl>        <int> <int>
1 F                0    88
2 F                1   165
3 T                0   121
4 T                1   522
```

2.4.1.5 Comparing physhealth summaries by obesity status

Can we compare the **physhealth** means, medians and 75th percentiles for respondents whose BMI is below 30 to the respondents whose BMI is not?

```
smartcle2 %>%
    group_by(bmi < 30) %>%
    summarize(mean(physhealth), median = median(physhealth),
              q75 = quantile(physhealth, 0.75))
```

```
# A tibble: 2 x 4
  `bmi < 30` `mean(physhealth)` median   q75
  <lgl>                   <dbl>  <int> <dbl>
1 F                        5.95      0  7.00
2 T                        3.22      0  2.00
```

2.4.1.6 The skim function within a pipe

The **skim** function works within pipes and with the other tidyverse functions.

```
smartcle2 %>%
    group_by(exerany) %>%
    skim(bmi, physhealth)
```

```
Skim summary statistics
 n obs: 896
 n variables: 10
 group variables: exerany

Variable type: integer
 exerany   variable missing complete   n mean    sd p0 p25 median p75 p100
       0 physhealth       0      209 209 7.95 11.68  0   0      0  15   30
       1 physhealth       0      687 687 2.78  7.06  0   0      0   2   30
     hist
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2582>
 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>

Variable type: numeric
 exerany variable missing complete   n  mean   sd    p0   p25 median   p75
       0      bmi       0      209 209 29.57 7.46 18    24.11  28.49 33.13
       1      bmi       0      687 687 27.35 5.84 12.71 23.7   26.52 29.81
  p100     hist
 66.06 <U+2586><U+2587><U+2586><U+2582><U+2581><U+2581><U+2581><U+2581>
 60.95 <U+2581><U+2586><U+2587><U+2582><U+2581><U+2581><U+2581><U+2581>
```

2.4.1.7 The usual summary for a data frame

Of course, we can use the usual `summary` to get some basic information about the data, too.

```
summary(smartcle2)
```

```
     SEQNO             physhealth      menthealth          genhealth
 Min.   :2.016e+09   Min.   : 0.00   Min.   : 0.000   1_Excellent:155
 1st Qu.:2.016e+09   1st Qu.: 0.00   1st Qu.: 0.000   2_VeryGood :306
 Median :2.016e+09   Median : 0.00   Median : 0.000   3_Good     :295
 Mean   :2.016e+09   Mean   : 3.99   Mean   : 2.693   4_Fair     :102
 3rd Qu.:2.016e+09   3rd Qu.: 2.00   3rd Qu.: 2.000   5_Poor     : 38
 Max.   :2.016e+09   Max.   :30.00   Max.   :30.000
      bmi            female         internet30        exerany
 Min.   :12.71   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:23.70   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:1.0000
 Median :26.80   Median :1.0000   Median :1.0000   Median :1.0000
 Mean   :27.87   Mean   :0.5848   Mean   :0.8147   Mean   :0.7667
 3rd Qu.:30.53   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :66.06   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
    sleephrs         alcdays
 Min.   : 1.000   Min.   : 0.000
 1st Qu.: 6.000   1st Qu.: 0.000
 Median : 7.000   Median : 1.000
 Mean   : 7.022   Mean   : 4.834
 3rd Qu.: 8.000   3rd Qu.: 5.000
 Max.   :20.000   Max.   :30.000
```

2.4.1.8 The `describe` function in `Hmisc`

Or we can use the `describe` function from the `Hmisc` package.

```
Hmisc::describe(smartcle2)
```

```
smartcle2

 10  Variables      896  Observations
--------------------------------------------------------------------------
SEQNO
       n   missing  distinct      Info      Mean       Gmd       .05
     896         0       896         1 2.016e+09     345.7 2.016e+09
     .10       .25       .50       .75       .90       .95
2.016e+09 2.016e+09 2.016e+09 2.016e+09 2.016e+09 2.016e+09

lowest : 2016000001 2016000002 2016000003 2016000004 2016000005
highest: 2016001031 2016001032 2016001033 2016001034 2016001036
--------------------------------------------------------------------------
physhealth
       n  missing distinct     Info     Mean      Gmd      .05      .10
     896        0       19    0.712     3.99    6.664        0        0
     .25      .50      .75      .90      .95
       0        0        2       15       30

Value          0     1     2     3     4     5     6     7     8     9
Frequency    591    35    55    22    12    25     4    20     1     1
Proportion 0.660 0.039 0.061 0.025 0.013 0.028 0.004 0.022 0.001 0.001

Value         10    12    14    15    18    20    21    25    30
Frequency     18     3    10    14     1     8     1     1    74
Proportion 0.020 0.003 0.011 0.016 0.001 0.009 0.001 0.001 0.083
--------------------------------------------------------------------------
menthealth
       n  missing distinct     Info     Mean      Gmd      .05      .10
     896        0       17    0.645    2.693    4.652        0        0
     .25      .50      .75      .90      .95
       0        0        2        8       20

Value          0     1     2     3     4     5     6     7     8    10
Frequency    634    25    56    27    15    30     4    13     4    18
Proportion 0.708 0.028 0.062 0.030 0.017 0.033 0.004 0.015 0.004 0.020

Value         14    15    18    20    23    29    30
Frequency      2    20     1     9     1     1    36
Proportion 0.002 0.022 0.001 0.010 0.001 0.001 0.040
--------------------------------------------------------------------------
genhealth
       n  missing distinct
     896        0        5

Value      1_Excellent  2_VeryGood      3_Good      4_Fair      5_Poor
Frequency          155         306         295         102          38
Proportion       0.173       0.342       0.329       0.114       0.042
--------------------------------------------------------------------------
```

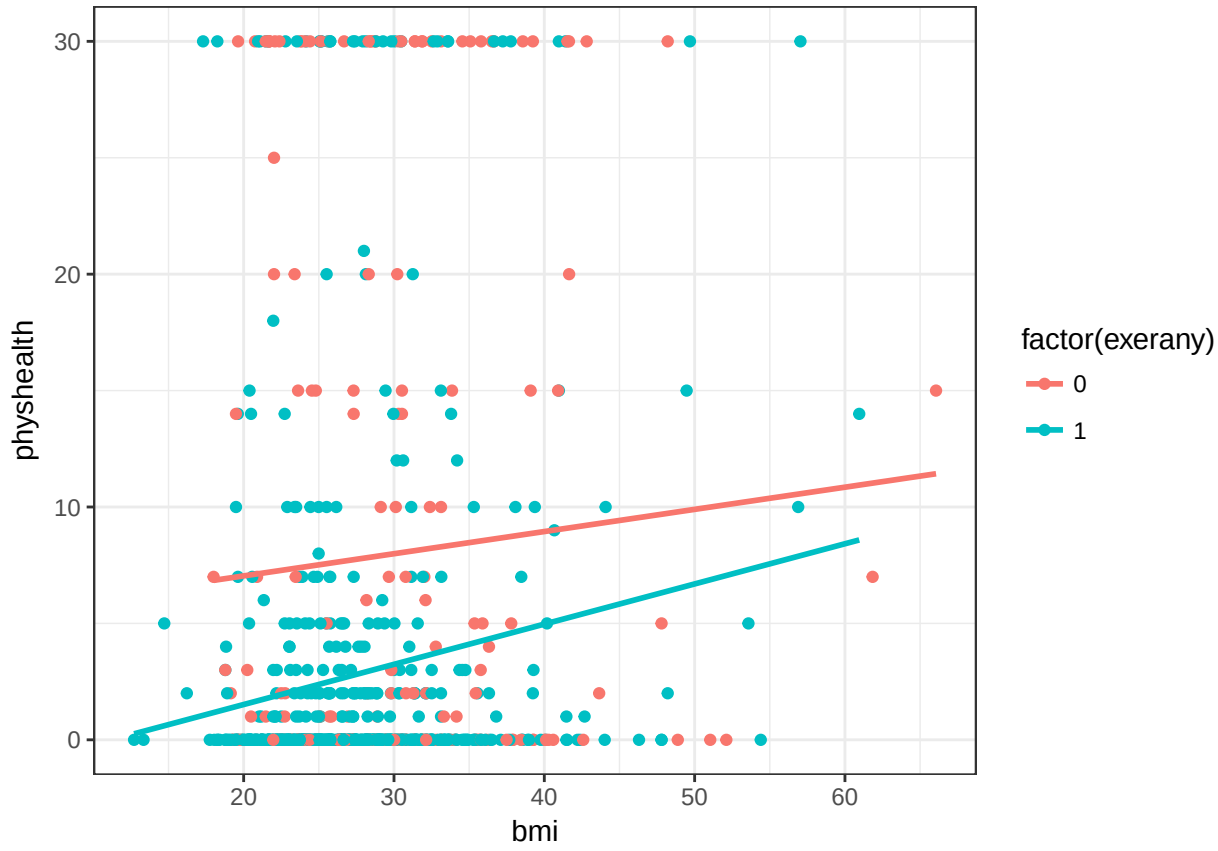
```
bmi
       n  missing distinct     Info     Mean      Gmd      .05      .10
     896        0      467        1    27.87    6.572    20.06    21.23
     .25      .50      .75      .90      .95
   23.70    26.80    30.53    35.36    39.30

lowest : 12.71 13.34 14.72 16.22 17.30, highest: 56.89 57.04 60.95 61.84 66.06
---------------------------------------------------------------------------
female
       n  missing distinct     Info      Sum     Mean      Gmd
     896        0        2    0.728      524   0.5848   0.4862

---------------------------------------------------------------------------
internet30
       n  missing distinct     Info      Sum     Mean      Gmd
     896        0        2    0.453      730   0.8147   0.3022

---------------------------------------------------------------------------
exerany
       n  missing distinct     Info      Sum     Mean      Gmd
     896        0        2    0.537      687   0.7667   0.3581

---------------------------------------------------------------------------
sleephrs
       n  missing distinct     Info     Mean      Gmd      .05      .10
     896        0       14    0.934    7.022    1.477        5        5
     .25      .50      .75      .90      .95
       6        7        8        8        9

Value          1     2     3     4     5     6     7     8     9    10
Frequency      5     1     6    20    63   192   276   266    38    22
Proportion 0.006 0.001 0.007 0.022 0.070 0.214 0.308 0.297 0.042 0.025

Value         11    12    16    20
Frequency      2     2     2     1
Proportion 0.002 0.002 0.002 0.001
---------------------------------------------------------------------------
alcdays
       n  missing distinct     Info     Mean      Gmd      .05      .10
     896        0       22    0.909    4.834    7.189        0        0
     .25      .50      .75      .90      .95
       0        1        5       17       30

lowest :  0  1  2  3  4, highest: 25 26 27 28 30
---------------------------------------------------------------------------
```

2.4.2 Graphing The Data

We'll build an exploratory figure (or several) to show the relationship between `bmi` and `physhealth` within each of the two `exerany` groups.

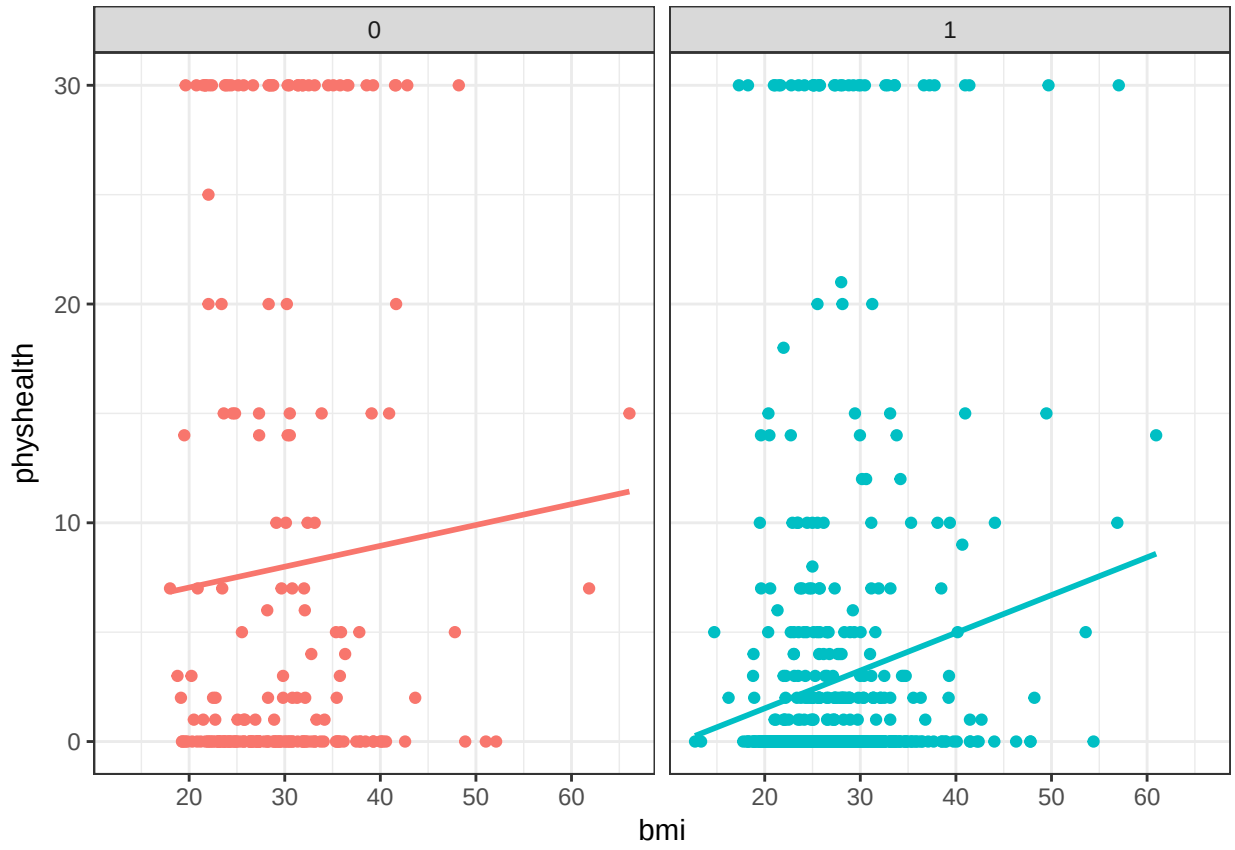
```
ggplot(smartcle2, aes(x = bmi, y = physhealth,
                      group = exerany, color = factor(exerany))) +
    geom_point() +
```

```
    geom_smooth(method = "lm", se = FALSE) +
    theme_bw()
```



The figure *could* be improved by separating the two groups into facets.

```
ggplot(smartcle2, aes(x = bmi, y = physhealth,
                      group = exerany, color = factor(exerany))) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    theme_bw() +
    guides(color = FALSE) +
    facet_wrap(~ exerany)
```



Now, what can we learn from these plots?

1. Does `physhealth` look like a good candidate for a linear model?
2. Does there seem to be a meaningful difference in the slopes of the two fitted lines?
3. In what BMI range can we make a reasonable prediction of `physhealth`?

2.5 Model A: Predicting `physhealth`

2.5.1 Building Model A

First, we'll fit a simple model describing the main effects but not the interaction of `exerany` and `bmi`, without doing any of the exploratory ground work we should do in advance.

- The outcome of interest is `physhealth`.
- Inputs to the regression model are:
  - `exerany` = 1 if the subject exercised in the past 30 days, and 0 if they didn't
  - `bmi` = body mass index (treated as qualitative and continuous)

```
modA <- lm(physhealth ~ exerany + bmi, data = smartcle2)
glance(modA)
```

```
   r.squared adj.r.squared    sigma statistic      p.value df    logLik
1 0.07538333    0.07331252 8.316983  36.40282 6.337345e-16  3 -3167.863
       AIC      BIC deviance df.residual
1 6343.726 6362.917 61770.78         893
```

```
tidy(modA)
```

```
         term   estimate  std.error statistic      p.value
1 (Intercept)  3.6042411 1.43482193  2.511978 1.218103e-02
2     exerany -4.8400260 0.66442519 -7.284531 7.091236e-13
3         bmi  0.1470196 0.04444619  3.307811 9.778491e-04
```

```
summary(modA)
```

```
Call:
lm(formula = physhealth ~ exerany + bmi, data = smartcle2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2654  -3.3284  -2.4368  -0.8999  28.6923

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.60424    1.43482   2.512 0.012181 *
exerany     -4.84003    0.66443  -7.285 7.09e-13 ***
bmi          0.14702    0.04445   3.308 0.000978 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.317 on 893 degrees of freedom
Multiple R-squared:  0.07538,   Adjusted R-squared:  0.07331
F-statistic:  36.4 on 2 and 893 DF,  p-value: 6.337e-16
```

What conclusions can we draw here?

Bibliography

Barnett, P. A., Roman-Golstein, S., Ramsey, F., et al. (1995). Differential permeability and quantitative mr imaging of a human lung carcinoma brain xenograft in the nude rat. *American Journal of Pathology*, 146(2):436–449.

Berkhemer, O. A., Fransen, P. S. S., Buemer, D., et al. (2015). A randomized trial of intraarterial treatment for acute ischemic stroke. *New England Journal of Medicine*, 372:11–20.

Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury, Pacific Grove, CA, second edition.

Roy, D., Talajic, M., Nattel, S., et al. (2008). Rhythm control versus rate control for atrial fibrillation and heart failure. *New England Journal of Medicine*, 358:2667–2677.

Tolaney, S. M., Barry, W. T., Chau, T. D., et al. (2015). Adjuvant paclitaxel and trastuzumab for node-negative, her2-positive breast cancer. *New England Journal of Medicine*, 372:134–141.