ADA1: Cumulative project file

How Drink Dependency Correlates to Depression and Mania

Advanced Data Analysis 1, Stat 427/527, Fall 2022, Prof. Erik Erhardt, UNM

AUTHOR
Ryan Riner

PUBLISHED
November 21, 2022

1 Document overview

> **Important**
>
> Please don't let the initial size and detail of this document intimidate you. You got this!

This document is organized by Week and Class number. The worksheet assignments are indicated by the Class numbers.

Consider your readers (graders):

- organize the document clearly (use this document as an example)
- label minor sections under each day (use this document as an example)
- For each thing you do, always have these three parts:
  1. Say what you're going to do and why.
  2. Do it with code, and document your code.
  3. Interpret the results.

1.1 Document

1.1.1 Naming

> **Note**
>
> Each class save this file with a new name, updating the last two digits to the class number. Then, you'll have a record of your progress, as well as which files you turned in for grading.

- `ADA1_ALL_05.qmd`
- `ADA1_ALL_06.qmd`
- `ADA1_ALL_07.qmd` ...

A version that I prefer is to use a date using Year-Month-Day, `YYYYMMDD`:

- `ADA1_ALL_20220903.qmd`

- ADA1_ALL_20220905.qmd
- ADA1_ALL_20220910.qmd ...

1.1.2 Structure

We will include all of our assignments together in this document to retain the relevant information needed for subsequent assignments since our analysis is cumulative. You will also have an opportunity to revisit previous parts to make changes or improvements, such as updating your codebook, recoding variables, and improving tables and plots. I've provided an initial predicted organization of our sections and subsections using the # and ## symbols. A table of contents is automatically generated using the "toc: true" in the yaml and can headings in the table of contents are clickable to jump down to each (sub)section.

1.1.3 Classes not appearing in this document

Some assignments are in a separate worksheet and are indicated with "(separate worksheet)". For these, I'll provide a dataset I want you to analyze. Typically, you'll then return to this document and repeat the same type of analysis with your dataset.

2 Research Questions

2.1 Class 02, Personal Codebook

Rubric

1. (1 p) Is there a topic of interest?
2. (2 p) Are the variables relevant to a set of research questions?
3. (4 p) Are there at least 2 categorical and 2 numerical variables (at least 4 "data" variables)?
    - 1 categorical variable with only 2 levels
    - 1 categorical variable with at least 3 levels
    - 2 numerical variables with many possible unique values
    - More variables are welcome and you're likely to add to this later in the semester
4. (3 p) For each variable, is there a variable description, a data type, and coded value descriptions?
5. Compile this qmd file to an html, print/save to pdf, and upload to UNM Canvas.

2.1.1 Topic and research questions

Topic:

Heavy alcohol use and its respective relationships to manic and depressive symptoms.

Research questions:

1. Do depressive symptoms (Depression) predict heavy drinking (DrinkExperience/DrinkQuantity/DrinkDependence)?
2. Do manic symptoms (Mania) predict heavy drinking (DrinkExperience/DrinkQuantity /DrinkDependence)?
3. Is there an income bracket (Income) with higher correlations between heavy drinking and mania or depression?

3 Codebook

National Epidemiologic Survey on Alcohol and Related Conditions-III (NESARC-III)

- Codebook: https://statacumen.com/teach/ADA1/PDS_data/NESARC_W1_CodeBook.pdf
- Official site: https://www.niaaa.nih.gov/research/nesarc-iii
- Introduction: https://pubs.niaaa.nih.gov/publications/arh29-2/74-78.htm

```
Dataset: NESARC
Primary association: alcohol abuse and manic/depressive disorders

Key:
RenamedVarName
  VarName original in dataset
  Variable description
  Data type (Continuous, Discrete, Nominal, Ordinal)
  Frequency ItemValue Description

ID
  IDNUM
  UNIQUE ID NUMBER WITH NO ALPHABETICS
  Nominal
  43093 1-43093. Unique Identification number

Sex
  SEX
  SEX
  Nominal
```

```
  18518 1. Male
  24575 2. Female

Age
  AGE
  AGE
  Continuous
  43079 18-97. Age in years
     14 98. 98 years or older

Income
  S1Q11B
  TOTAL FAMILY INCOME IN LAST 12 MONTHS
  Categorical
  1718 1.  Less than $5,000
  2338 2.  $5,000 to $7,999
  1373 3.  $8,000 to $9,999
  2559 4.  $10,000 to $12,999
  1343 5.  $13,000 to $14,999
  3317 6.  $15,000 to $19,999
  3368 7.  $20,000 to $24,999
  2941 8.  $25,000 to $29,999
  3052 9.  $30,000 to $34,999
  2565 10. $35,000 to $39,999
  4301 11. $40,000 to $49,999
  3428 12. $50,000 to $59,999
  2644 13. $60,000 to $69,999
  2008 14. $70,000 to $79,999
  1363 15. $80,000 to $89,999
  977  16. $90,000 to $99,999
  1130 17. $100,000 to $109,999
  428  18. $110,000 to $119,999
  914  19. $120,000 to $149,999
  725  20. $150,000 to 199,999
  601  21. $200,000 or more

DrinkExperience
  S2AQ20
  DURATION (YEARS) OF PERIOD OF HEAVIEST DRINKING
  Continuous
  33377 1-80. Number of years
  1450 99. Unknown
  8266 BL. NA, lifetime abstainer

DrinkQuantity
```

```
  S2AQ21B
  NUMBER OF DRINKS OF ANY ALCOHOL USUALLY CONSUMED ON DAYS WHEN DRANK ALCOHOL
  DURING PERIOD OF HEAVIEST DRINKING
  Discrete
  33683 1-98. Number of drinks
  1144 99. Unknown
  8266 BL. NA, lifetime abstainer

DrinkDependence
  S2BQ1A4
  EVER INCREASE DRINKING BECAUSE AMOUNT FORMERLY CONSUMED NO LONGER
  GAVE DESIRED EFFECT
  Categorical
  3048 1. Yes
  31467 2. No
  312 9. Unknown
  8266 BL. NA, lifetime abstainer

Depression
  S4AQ1
  EVER HAD 2-WEEK PERIOD WHEN FELT SAD, BLUE, DEPRESSED, OR DOWN MOST OF TIME
  Categorical
  12785 1. Yes
  29416 2. No
  892 9. Unknown

Mania
  S5Q3
  HAD 1+ WEEK PERIOD IRRITABLE/EASILY ANNOYED THAT CAUSED YOU TO SHOUT/BREAK
  THINGS/START FIGHTS OR ARGUMENTS
  Categorical
  3402 1. Yes
  38620 2. No
  1071 9. Unknown
```

Additional variables were created from the original variables:

```
CREATED VARIABLES

Height_inches
  Total height in inches
  Height_ft * 12 + Height_in


ADD MORE HERE
ADD MORE HERE
```

```
ADD MORE HERE
ADD MORE HERE (If you think you'll combine or transform any variables)
ADD MORE HERE
ADD MORE HERE
ADD MORE HERE
```

4 Data Management

4.1 Class 03, Data subset and numerical summaries

Rubric

1. (4 p) The data are loaded and a data.frame subset is created by selecting only the variables in the personal codebook.
    - Scroll down to sections labeled "(Class 03)".
2. (1 p) Output confirms the subset is correct (e.g., using `dim()` and `str()`).
3. (3 p) Rename your variables to descriptive names (e.g., from "S3AQ3B1" to "SmokingFreq").
    - Scroll down to sections labeled "(Class 03)".
4. (2 p) Provide numerical summaries for all variables (e.g., using `summary()`).
    - Scroll down to sections labeled "(Class 03)".

4.1.1 Data subset (Class 03)

First, the data is placed on the search path.

```
# data analysis packages
library(erikmisc)   # Helpful functions
```

```
── Attaching packages ──────────────────────────────────── erikmisc 0.1.16 ──

✓ tibble 3.1.8     ✓ dplyr  1.0.9

── Conflicts ─────────────────────────────────────── erikmisc_conflicts() ──
✕ dplyr::filter() masks stats::filter()
✕ dplyr::lag()    masks stats::lag()
```

```
erikmisc, solving common complex data analysis workflows
  by Dr. Erik Barry Erhardt <erik@StatAcumen.com>
```

```
library(tidyverse)  # Data manipulation and visualization suite
```

```
── Attaching packages
─────────────────────────────────────
tidyverse 1.3.2 ──

✓ ggplot2 3.3.6     ✓ purrr   0.3.4
✓ tidyr   1.2.0     ✓ stringr 1.4.1
✓ readr   2.1.2     ✓ forcats 0.5.2
── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)  # Dates
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
  ## 1. Download the ".RData" file for your dataset into your ADA Folder.
  ## 2. Use the load() statement for the dataset you want to use.
# read data example
#load("NESARC.RData")

#dim(NESARC)
```

```
load("NESARC.RData")
dim(NESARC)
```

```
[1] 43093  3008
```

4.1.2 Renaming Variables (Class 03)

```
nesarc_sub <-
  NESARC %>%
  dplyr::select(
    IDNUM
```

```
  , SEX
  , AGE
  , S1Q11B
  , S2AQ20
  , S2AQ21B
  , S2BQ1A4
  , S4AQ1
  , S5Q3
  , ETHRACE2A
  , S2AQ22
  )

dim(nesarc_sub)
```

```
[1] 43093    11
```

```
str(nesarc_sub)
```

```
'data.frame':   43093 obs. of  11 variables:
 $ IDNUM    : Factor w/ 43093 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ SEX      : Factor w/ 2 levels "1","2": 1 2 2 1 1 2 1 2 2 2 ...
 $ AGE      : num  23 28 81 18 36 34 19 84 29 18 ...
 $ S1Q11B   : Factor w/ 21 levels "1","2","3","4",..: 11 10 2 11 15 12 7 7 13 10 ...
 $ S2AQ20   : num  NA 1 NA 1 1 1 1 54 8 1 ...
 $ S2AQ21B  : num  NA 1 NA 2 1 4 1 1 1 6 ...
 $ S2BQ1A4  : Factor w/ 3 levels "1","2","9": NA 2 NA 2 2 1 2 2 1 2 ...
 $ S4AQ1    : Factor w/ 3 levels "1","2","9": 2 2 2 2 2 2 1 2 1 2 ...
 $ S5Q3     : Factor w/ 3 levels "1","2","9": 2 2 2 2 2 2 1 2 1 2 ...
 $ ETHRACE2A: Factor w/ 5 levels "1","2","3","4",..: 5 5 5 5 2 2 2 1 1 5 ...
 $ S2AQ22   : Factor w/ 12 levels "1","2","3","4",..: NA 11 NA 11 11 9 11 11 11 8 ...
```

```
nesarc_sub <-
  nesarc_sub %>%
  dplyr::rename(
    ID                = IDNUM
  , Sex               = SEX
  , Age               = AGE
  , Income            = S1Q11B
  , DrinkExperience   = S2AQ20
  , DrinkQuantity     = S2AQ21B
  , DrinkDependence   = S2BQ1A4
  , Depression        = S4AQ1
  , Mania             = S5Q3
  , Ethnicity         = ETHRACE2A
  )
```

```
summary(nesarc_sub)
```

```
      ID        Sex            Age            Income      DrinkExperience
1      :    1   1:18518   Min.   :18.0   11     : 4301   Min.   : 1.00
2      :    1   2:24575   1st Qu.:32.0   12     : 3428   1st Qu.: 1.00
3      :    1             Median :44.0   7      : 3368   Median : 4.00
4      :    1             Mean   :46.4   6      : 3317   Mean   :12.26
5      :    1             3rd Qu.:59.0   9      : 3052   3rd Qu.:12.00
6      :    1             Max.   :98.0   8      : 2941   Max.   :99.00
(Other):43087                           (Other):22686   NA's   :8266
DrinkQuantity    DrinkDependence Depression Mania     Ethnicity
Min.   : 1.000   1   : 3048      1:12785    1: 3402   1:24507
1st Qu.: 1.000   2   :31467      2:29416    2:38620   2: 8245
Median : 2.000   9   :  312      9:  892    9: 1071   3:  701
Mean   : 6.583   NA's: 8266                           4: 1332
3rd Qu.: 5.000                                        5: 8308
Max.   :99.000
NA's   :8266
    S2AQ22
11     :20698
1      : 2090
5      : 1908
4      : 1856
3      : 1764
(Other): 6511
NA's   : 8266
```

4.1.3 Coding missing values (Class 04)

There are two steps. The first step is to recode any existing `NA`s to actual values, if necessary. The method for doing this differs for numeric and categorical variables. The second step is to recode any coded missing values, such as 9s or 99s, as actual `NA`.

4.1.3.1 Coding `NA`s as meaningful "missing"

First step: the existing blank values with `NA` mean "never", and "never" has a meaning different from "missing". For each variable we need to decide what "never" means and code it appropriately.

4.1.3.1.1 `NA`s recoded as numeric

```
table(nesarc_sub$DrinkExperience)
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
8960 4467 2780 1855 2463  833  621  585  250 2729  200  407  171  190  997  143
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
 154  202  117 1618  132  145  106  103  493   72   77   99   35  765   38   58
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
  43   44  137   41   47   36   18  416   20   28   23   25   67   22   32   29
  49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64
  12  229    9   17   16   12   28   11   12   14    6   68    4    2    3    1
  65   66   67   68   69   70   71   72   75   77   80   99
  11    1    1    5    2   10    3    1    4    1    1 1450
```

```
nesarc_sub <-
  nesarc_sub %>%
  replace_na(
    list(
      DrinkExperience = 0
    )
  )

table(nesarc_sub$DrinkExperience)
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
8266 8960 4467 2780 1855 2463  833  621  585  250 2729  200  407  171  190  997
  16   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31
 143  154  202  117 1618  132  145  106  103  493   72   77   99   35  765   38
  32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47
  58   43   44  137   41   47   36   18  416   20   28   23   25   67   22   32
  48   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63
  29   12  229    9   17   16   12   28   11   12   14    6   68    4    2    3
  64   65   66   67   68   69   70   71   72   75   77   80   99
   1   11    1    1    5    2   10    3    1    4    1    1 1450
```

```
table(nesarc_sub$DrinkQuantity)
```

```
    1     2     3     4     5     6     7     8     9    10    11    12    13
11397  7445  4413  2785  1688  2454   468   679   149   690    31   758    30
   14    15    16    17    18    19    20    21    22    24    25    26    27
   28   180    23     7    64     3   148     1     2   107    28     3     1
   28    30    32    33    34    35    36    38    40    42    43    44    48
    3    32     6     1     1     5     8     1    11     2     1     1    11
   50    56    60    64    70    80    84    98    99
    6     1     2     1     1     1     1     5  1144
```

```
nesarc_sub <-
  nesarc_sub %>%
  replace_na(
    list(
      DrinkQuantity = 0
    )
  )

table(nesarc_sub$DrinkQuantity)
```

```
   0    1    2    3    4    5    6    7    8    9   10   11   12
8266 11397 7445 4413 2785 1688 2454  468  679  149  690   31  758
  13   14   15   16   17   18   19   20   21   22   24   25   26
  30   28  180   23    7   64    3  148    1    2  107   28    3
  27   28   30   32   33   34   35   36   38   40   42   43   44
   1    3   32    6    1    1    5    8    1   11    2    1    1
  48   50   56   60   64   70   80   84   98   99
  11    6    1    2    1    1    1    1    5 1144
```

4.1.3.1.2 NA s recoded as categorical

```
table(nesarc_sub$DrinkDependence)
```

```
   1    2    9
3048 31467  312
```

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
    DrinkDependence = as.character(DrinkDependence)
  ) %>%
  replace_na(
    list(
      DrinkDependence = "3"
    )
  )

table(nesarc_sub$DrinkDependence)
```

```
   1    2    3    9
```

```
3048 31467  8266   312
```

4.1.3.2 Coding 9s and 99s as NAs

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
      DrinkExperience  = replace(DrinkExperience,  DrinkExperience  %in% c( 99),
         NA)
    , DrinkQuantity    = replace(DrinkQuantity,    DrinkQuantity    %in% c( 99),
         NA)
    , DrinkDependence  = replace(DrinkDependence,  DrinkDependence  %in% c("9"),
         NA)
    , Depression       = replace(Depression,       Depression       %in% c("9"),
         NA)
    , Mania            = replace(Mania,            Mania            %in% c("9"),
         NA)
  )

summary(nesarc_sub)
```

```
      ID        Sex            Age           Income      DrinkExperience
1      :    1   1:18518   Min.   :18.0   11     : 4301   Min.   : 0.000
2      :    1   2:24575   1st Qu.:32.0   12     : 3428   1st Qu.: 1.000
3      :    1             Median :44.0   7      : 3368   Median : 2.000
4      :    1             Mean   :46.4   6      : 3317   Mean   : 6.808
5      :    1             3rd Qu.:59.0   9      : 3052   3rd Qu.:10.000
6      :    1             Max.   :98.0   8      : 2941   Max.   :80.000
(Other):43087                           (Other):22686   NA's   :1450
DrinkQuantity    DrinkDependence    Depression    Mania       Ethnicity
Min.   : 0.000   Length:43093       1   :12785   1   : 3402   1:24507
1st Qu.: 1.000   Class :character   2   :29416   2   :38620   2: 8245
Median : 2.000   Mode  :character   9   :    0   9   :    0   3:  701
Mean   : 2.765                      NA's:  892   NA's: 1071   4: 1332
3rd Qu.: 3.000                                                5: 8308
Max.   :98.000
NA's   :1144
```

4.1.4 Labeling Categorical variable levels (Class 04)

```
summary(nesarc_sub)
```

```
      ID        Sex            Age           Income      DrinkExperience
1      :    1   1:18518   Min.   :18.0   11     : 4301   Min.   : 0.000
2      :    1   2:24575   1st Qu.:32.0   12     : 3428   1st Qu.: 1.000
3      :    1             Median :44.0   7      : 3368   Median : 2.000
```

```
4      :    1             Mean   :46.4   6      : 3317   Mean   : 6.808
5      :    1             3rd Qu.:59.0   9      : 3052   3rd Qu.:10.000
6      :    1             Max.   :98.0   8      : 2941   Max.   :80.000
(Other):43087                              (Other):22686   NA's   :1450
DrinkQuantity    DrinkDependence    Depression    Mania       Ethnicity
Min.   : 0.000   Length:43093       1   :12785   1   : 3402   1:24507
1st Qu.: 1.000   Class :character   2   :29416   2   :38620   2: 8245
Median : 2.000   Mode  :character   9   :    0   9   :    0   3:  701
Mean   : 2.765                      NA's:  892   NA's: 1071   4: 1332
3rd Qu.: 3.000                                                5: 8308
Max.   :98.000
NA's   :1144
```

```
nesarc_sub$Sex <-
  factor(
    nesarc_sub$Sex
    , labels = c("Male"
                 , "Female"
                 )
  )

nesarc_sub$DrinkDependence <-
  factor(
    nesarc_sub$DrinkDependence
    , labels = c("Yes Dependence Symptoms"
                 , "No Dependence Symptoms"
                 , "Unsure Dependence Symptoms"
                 )
  )

nesarc_sub$Depression <-
  factor(
    nesarc_sub$Depression
    , labels = c("Yes Depression"
                 , "No Depression"
                 )
  )

nesarc_sub$Mania <-
  factor(
    nesarc_sub$Mania
    , labels = c("Yes Manic Symptoms"
                 , "No Manic Symptoms"
                 )
  )
```

```
nesarc_sub$Ethnicity <-
  factor(
    nesarc_sub$Ethnicity
    , levels = c( 1
                , 2
                , 3
                , 4
                , 5
                )
    , labels = c( "Cauc"
                , "AfAm"
                , "NaAm"
                , "Asia"
                , "Hisp")
  )


nesarc_sub$Income <-
  factor(nesarc_sub$Income
         , levels = c( 1
                     , 2
                     , 3
                     , 4
                     , 5
                     , 6
                     , 7
                     , 8
                     , 9
                     , 10
                     , 11
                     , 12
                     , 13
                     , 14
                     , 15
                     , 16
                     , 17
                     , 18
                     , 19
                     , 20
                     , 21
                     )
         , labels = c("Less than $5,000"
                     , "$5,000 to $7,999"
                     , "$8,000 to $9,999"
```

```
                  , "$10,000 to $12,999"
                  , "$13,000 to $14,999"
                  , "$15,000 to $19,999"
                  , "$20,000 to $24,999"
                  , "$25,000 to $29,999"
                  , "$30,000 to $34,999"
                  , "$35,000 to $39,999"
                  , "$40,000 to $49,999"
                  , "$50,000 to $59,999"
                  , "$60,000 to $69,999"
                  , "$70,000 to $79,999"
                  , "$80,000 to $89,999"
                  , "$90,000 to $99,999"
                  , "$100,000 to $109,999"
                  , "$110,000 to $119,999"
                  , "$120,000 to $149,999"
                  , "$150,000 to 199,999"
                  , "$200,000 or more"
                  )
  )


summary(nesarc_sub)
```

```
      ID            Sex             Age                      Income
1      :    1   Male  :18518   Min.   :18.0   $40,000 to $49,999: 4301
2      :    1   Female:24575   1st Qu.:32.0   $50,000 to $59,999: 3428
3      :    1                  Median :44.0   $20,000 to $24,999: 3368
4      :    1                  Mean   :46.4   $15,000 to $19,999: 3317
5      :    1                  3rd Qu.:59.0   $30,000 to $34,999: 3052
6      :    1                  Max.   :98.0   $25,000 to $29,999: 2941
(Other):43087                                 (Other)           :22686
DrinkExperience  DrinkQuantity                      DrinkDependence
Min.   : 0.000   Min.   : 0.000   Yes Dependence Symptoms   : 3048
1st Qu.: 1.000   1st Qu.: 1.000   No Dependence Symptoms    :31467
Median : 2.000   Median : 2.000   Unsure Dependence Symptoms: 8266
Mean   : 6.808   Mean   : 2.765   NA's                      :  312
3rd Qu.:10.000   3rd Qu.: 3.000
Max.   :80.000   Max.   :98.000
NA's   :1450     NA's   :1144
         Depression                    Mania       Ethnicity
Yes Depression:12785   Yes Manic Symptoms: 3402   Cauc:24507
No Depression :29416   No Manic Symptoms :38620   AfAm: 8245
NA's          :  892   NA's              : 1071   NaAm:  701
                                                  Asia: 1332
```

```
Hisp: 8308
```

4.1.5 Creating new variables (Class 04+)

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
    Age_log = log(Age)
  )
```

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
    DrinkQuantity_Drinkers = ifelse(DrinkQuantity > 1, DrinkQuantity, NA)
  , DrinkQuantity_Drinkers_log2 = log2(DrinkQuantity_Drinkers)
  )
```

4.1.5.1 From categories to numeric

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
    Average_Earnings =
      case_when(
        Income == 1  ~ 5000                                  # 1.  Less than $5,000
      , Income == 2  ~ (((7999 - 5000)     / 2) + 5000)      # 2.  $5,000 to $7,999
      , Income == 3  ~ (((9999 - 8000)     / 2) + 8000)      # 3.  $8,000 to $9,999
      , Income == 4  ~ (((12999 - 10000)   / 2) + 10000)     # 4.  $10,000 to $12,999
      , Income == 5  ~ (((14999 - 13000)   / 2) + 13000)     # 5.  $13,000 to $14,999
      , Income == 6  ~ (((19999 - 15000)   / 2) + 15000)     # 6.  $15,000 to $19,999
      , Income == 7  ~ (((24999 - 20000)   / 2) + 20000)     # 7.  $20,000 to $24,999
      , Income == 8  ~ (((29999 - 25000)   / 2) + 25000)     # 8.  $25,000 to $29,999
      , Income == 9  ~ (((34999 - 30000)   / 2) + 30000)     # 9.  $30,000 to $34,999
      , Income == 10 ~ (((39999 - 35000)   / 2) + 35000)     # 10. $35,000 to $39,999
      , Income == 11 ~ (((49999 - 40000)   / 2) + 40000)     # 11. $40,000 to $49,999
      , Income == 12 ~ (((59999 - 50000)   / 2) + 50000)     # 12. $50,000 to $59,999
      , Income == 13 ~ (((69999 - 60000)   / 2) + 60000)     # 13. $60,000 to $69,999
      , Income == 14 ~ (((79999 - 70000)   / 2) + 70000)     # 14. $70,000 to $79,999
      , Income == 15 ~ (((89999 - 80000)   / 2) + 80000)     # 15. $80,000 to $89,999
      , Income == 16 ~ (((99999 - 90000)   / 2) + 90000)     # 16. $90,000 to $99,999
      , Income == 17 ~ (((100000 - 109999) / 2) + 100000) # 17. $100,000 to
         $109,999
```

```
      , Income == 18 ~ (((119999 - 110000) / 2) + 110000) # 18. $110,000 to
         $119,999
      , Income == 19 ~ (((149999 - 120000) / 2) + 120000) # 19. $120,000 to
         $149,999
      , Income == 20 ~ (((199999 - 150000) / 2) + 150000) # 20. $150,000 to 199,999
      , Income == 21 ~ 200000                             # 21. $200,000 or more
      )
  )

summary(nesarc_sub$Average_Earnings)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
     NA      NA      NA     NaN      NA      NA   43093
```

4.1.5.2 From numeric to numeric

`Intoxication_Units` is the approximation of intoxicating quantities of alcohol consumed, derived form `DrinkQuantity`. This is the average number of drinks across men and women that produces a blood alcohol content of .08, legally recognized as a state of intoxication.

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
    Intoxication_Units = (DrinkQuantity / 3)
  )

summary(nesarc_sub$Intoxication_Units)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0000  0.3333  0.6667  0.9217  1.0000 32.6667    1144
```

4.1.5.3 From numeric to categories based on quantiles

`Heavy_Drinking` categorizes the longest period of heavy drinking reported, derived from `DrinkExperience`.

```
nesarc_sub <-
  nesarc_sub %>%
  mutate(
    Heavy_Drinking =
      case_when(
        DrinkExperience == 0                              ~ "No Heavy Drinking"
      , (DrinkExperience > 0)  & (DrinkExperience <= 5)   ~ "Short Heavy Drinking
         Period (5 years or less)"
      , (DrinkExperience > 5)  & (DrinkExperience <= 20)  ~ "Notable Heavy Drinking
         Period (5 to 20 years)"
```

```
      , (DrinkExperience > 20) & (DrinkExperience <= 40)  ~ "Extended Heavy
         Drinking Period (20 to 40 years)"
      , (DrinkExperience > 40)                            ~ "Lifetime Drinker (40
         years or more)"
      )
    , Heavy_Drinking =
      factor(
        Heavy_Drinking
        , levels = c(
            "No Heavy Drinking"
          , "Short Heavy Drinking Period (5 years or less)"
          , "Notable Heavy Drinking Period (5 to 20 years)"
          , "Extended Heavy Drinking Period (20 to 40 years)"
          , "Lifetime Drinker (40 years or more)"
        )
      )
  )

summary(nesarc_sub$Heavy_Drinking)
```

```
                             No Heavy Drinking
                                          8266
  Short Heavy Drinking Period (5 years or less)
                                         20525
  Notable Heavy Drinking Period (5 to 20 years)
                                          9217
Extended Heavy Drinking Period (20 to 40 years)
                                          2905
            Lifetime Drinker (40 years or more)
                                           730
                                          NA's
                                          1450
```

```
table(nesarc_sub$Heavy_Drinking)
```

```
                             No Heavy Drinking
                                          8266
  Short Heavy Drinking Period (5 years or less)
                                         20525
  Notable Heavy Drinking Period (5 to 20 years)
                                          9217
Extended Heavy Drinking Period (20 to 40 years)
                                          2905
            Lifetime Drinker (40 years or more)
```

730

4.1.5.4 From many categories to a few

4.1.5.5 Working with Dates

4.1.5.6 Review results of new variables

4.1.6 Data subset rows

4.2 Data is complete (Class 04)

4.2.1 Plot entire dataset, show missing values

4.2.2 Numerical summaries to assess correctness (Class 03)

5 Graphing and Tabulating

5.1 Class 04, Plotting univariate

Rubric

1. (3 p) For one categorical variable, a barplot is plotted with axis labels and a title. Interpret the plot: describe the relationship between categories you observe.
2. (3 p) For one numerical variable, a histogram or boxplot is plotted with axis labels and a title. Interpret the plot: describe the distribution (shape, center, spread, outliers).
3. (2 p) Code missing variables, remove records with missing values, indicate with R output that this was done correctly (e.g., `str()`, `dim()`, `summary()`).
    - Scroll up to sections labeled "(Class 04)".
4. (2 p) Label levels of factor variables.
    - Scroll up to sections labeled "(Class 04)".

5.2 Categorical variables

5.2.1 Tables for categorical variables

```
table(nesarc_sub$DrinkDependence)
```

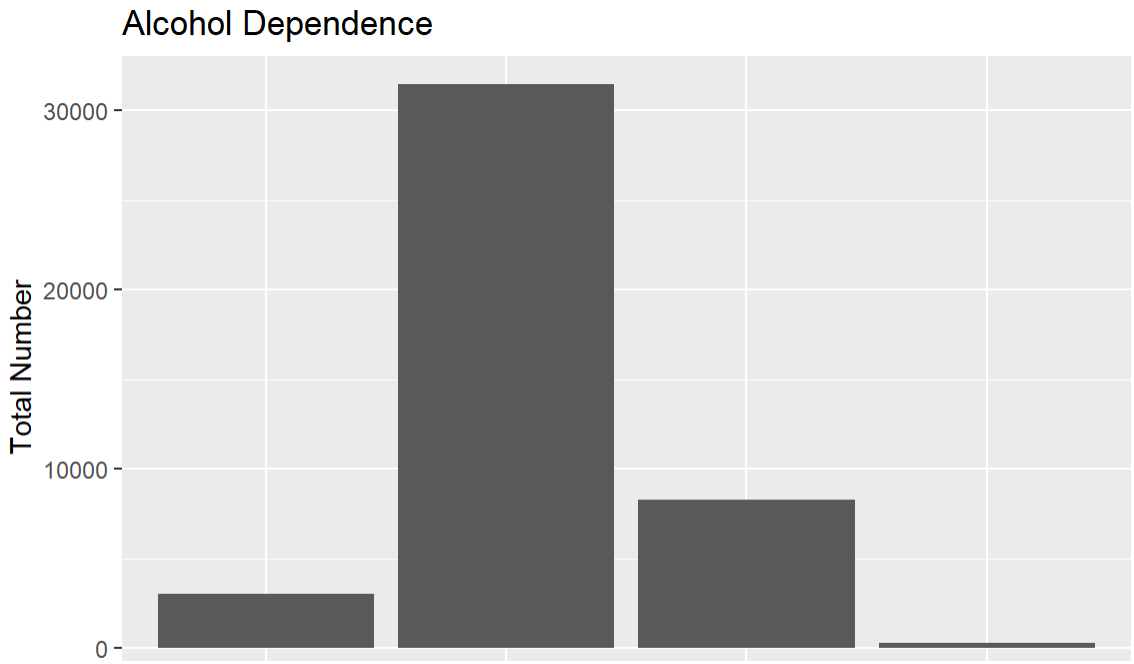
```
   Yes Dependence Symptoms     No Dependence Symptoms
                      3048                      31467
Unsure Dependence Symptoms
                      8266
```

```
table(nesarc_sub$DrinkDependence) %>% prop.table()
```

```
   Yes Dependence Symptoms     No Dependence Symptoms
                0.07124658                0.73553680
Unsure Dependence Symptoms
                0.19321661
```

5.2.2 Graphing frequency tables

```
library(ggplot2)
pl <- ggplot(data = nesarc_sub, aes(x = DrinkDependence))
pl <- pl + geom_bar()
pl <- pl + labs(x     = ""
              , y   = "Total Number"
              , title = "Alcohol Dependence"
              )
print(pl)
```



Interpretation: There are far fewer (9%) participants that reported increasing the amount they drank to achieve their desired effect than those who did not (90%).

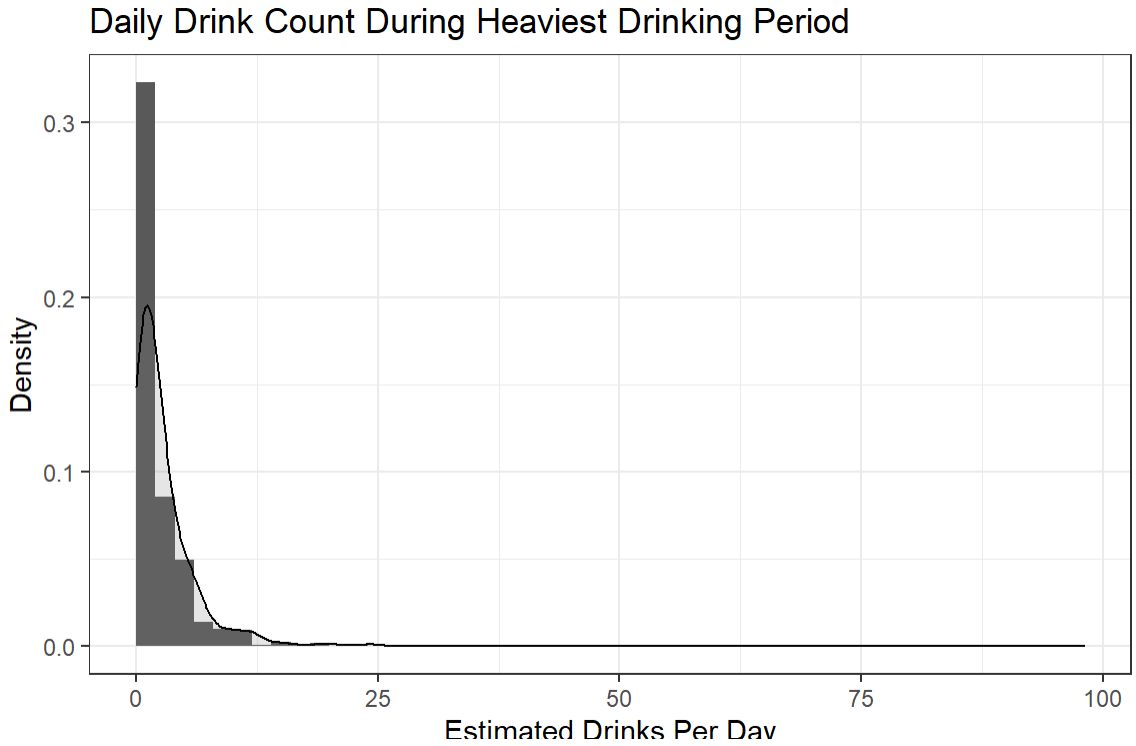
5.3 Numeric variables

5.3.1 Graphing numeric variables

```
library(ggplot2)
p <- ggplot(data = nesarc_sub, aes(x = DrinkQuantity))
p <- p + theme_bw()
p <- p + geom_histogram(aes(y = ..density..), boundary = 0, binwidth = 2)
#p <- p + geom_rug()
p <- p + geom_density(alpha = 0.2, fill = "gray50", adjust = 7)
p <- p + labs(x = "Estimated Drinks Per Day"
              , y = "Density"
              , title = "Daily Drink Count During Heaviest Drinking Period"
              )
print(p)
```

```
Warning: Removed 1144 rows containing non-finite values (stat_bin).
```

```
Warning: Removed 1144 rows containing non-finite values (stat_density).
```



```
summary(nesarc_sub$DrinkQuantity)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.000   2.000   2.765   3.000  98.000    1144
```

Interpretation: Most participants (about 44%) consumed 1-6 drinks per day during their period of heaviest drinking. Few participants (about 5%) consumed 6-12 drinks per day during their period of heaviest drinking.

5.3.2 Creating Density Plots

5.4 Class 05-1, Plotting bivariate, numeric response

Rubric

1. Each of the following (2 p for plot, 2 p for labelled axes and title, 1 p for interpretation):
    1. Scatter plot (for regression): x = numerical, y = numerical, include axis labels and a title. Interpret the plot: describe the relationship.
    2. Box plots (for ANOVA): x = categorical, y = numerical, include axis labels and a title. Interpret the plot: describe the relationship.

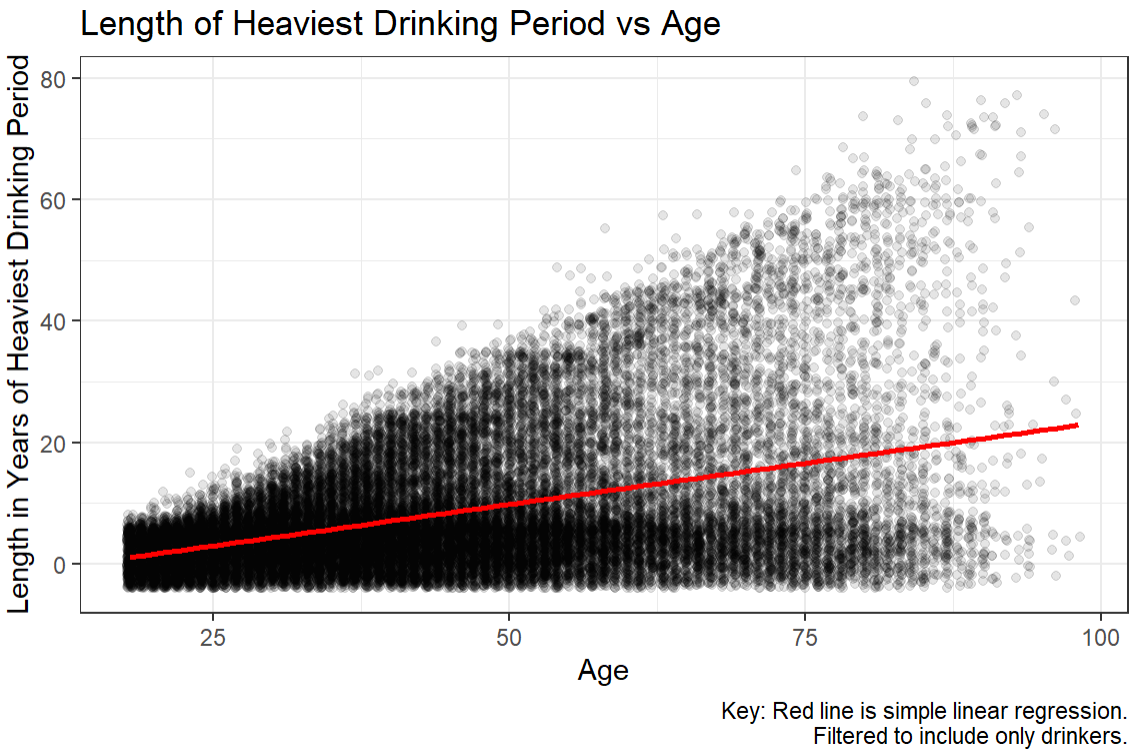
5.4.1 Scatter plot (for regression): x = numerical, y = numerical

Interpretation: There are two distinct relationships between `Age` and `DrinkExperience` among those who drink. One correlation is nearly 0, probably representing those whose heaviest drinking period was in college (4-10 years). The other correlation is nearly 1, probably representing those who have spent their whole lives drinking heavily. The former relationship dominates the linear regression.

```
library(ggplot2)
p <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
        DrinkExperience))
p <- p + theme_bw()
p <- p + geom_jitter(width = 0.25, height = 5, alpha = 1/10)
p <- p + stat_smooth(method = lm, colour = "red")
p <- p + labs(
    title = "Length of Heaviest Drinking Period vs Age"
  , y = "Length in Years of Heaviest Drinking Period"
  , caption = "Key: Red line is simple linear regression.\nFiltered to include
```

```
        only drinkers."
    )
print(p)
```

```
`geom_smooth()` using formula 'y ~ x'
```

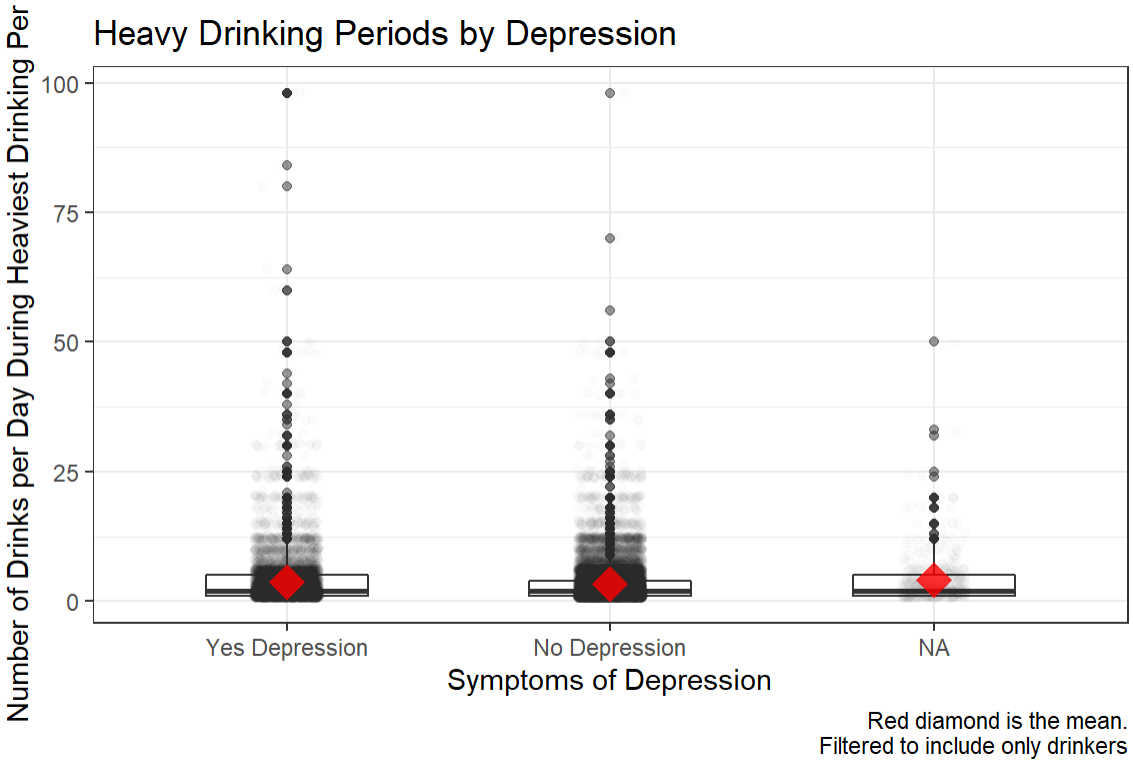


5.4.2 Box plots (for ANOVA): x = categorical, y = numerical

Interpretation There is very little difference in the number of drinks consumed per day between those reporting depressive symptoms and those reporting no depressive symptoms.

```
library(ggplot2)
p <- ggplot(nesarc_sub %>% filter(DrinkQuantity > 0), aes(x = Depression, y =
        DrinkQuantity))
p <- p + theme_bw()
p <- p + geom_boxplot(width = 0.5, alpha = 0.5)
p <- p + geom_jitter(position = position_jitter(width = 0.1), alpha = 1/100)
p <- p + stat_summary(fun = mean, geom = "point", shape = 18, size = 6,
                      colour = "red", alpha = 0.8)
p <- p + labs(
             title = "Heavy Drinking Periods by Depression"
           , x = "Symptoms of Depression"
           , y = "Number of Drinks per Day During Heaviest Drinking Period"
           , caption = "Red diamond is the mean.\nFiltered to include only
        drinkers"
)
```

```
print(p)
```



5.5 Class 05-2, Plotting bivariate, categorical response

Rubric

1. Each of the following (2 p for plot, 2 p for labelled axes and title, 1 p for interpretation):
    1. Mosaic plot or bivariate bar plots (for contingency tables): x = categorical, y = categorical, include axis labels and a title. Interpret the plot: describe the relationship.
    2. Logistic scatter plot (for logistic regression): x = numerical, y = categorical (binary), include axis labels and a title. Interpret the plot: describe the relationship.

5.5.1 Mosaic plot or bivariate bar plots (for contingency tables): x = categorical, y = categorical

```
tab_dep_alc <-
  nesarc_sub %>%
  group_by(
    Depression, DrinkDependence
  ) %>%
  summarize(
```

```
    n = n()
  ) %>%
  mutate(
    prop = round(n / sum(n), 3)
  ) %>%
  ungroup() %>%
  na.omit()
```

```
`summarise()` has grouped output by 'Depression'. You can override using the
`.groups` argument.
```

```
tab_dep_alc
```

```
# A tibble: 6 × 4
  Depression     DrinkDependence                n  prop
  <fct>          <fct>                      <int> <dbl>
1 Yes Depression Yes Dependence Symptoms     1362 0.107
2 Yes Depression No Dependence Symptoms      9634 0.754
3 Yes Depression Unsure Dependence Symptoms  1736 0.136
4 No Depression  Yes Dependence Symptoms     1627 0.055
5 No Depression  No Dependence Symptoms     21417 0.728
6 No Depression  Unsure Dependence Symptoms  6232 0.212
```

```
tab_dep_alc <-
  nesarc_sub %>%
  group_by(
    DrinkDependence, Depression
  ) %>%
  summarize(
    n = n()
  ) %>%
  mutate(
    prop = round(n / sum(n), 3)
  ) %>%
  ungroup() %>%
  na.omit()
```

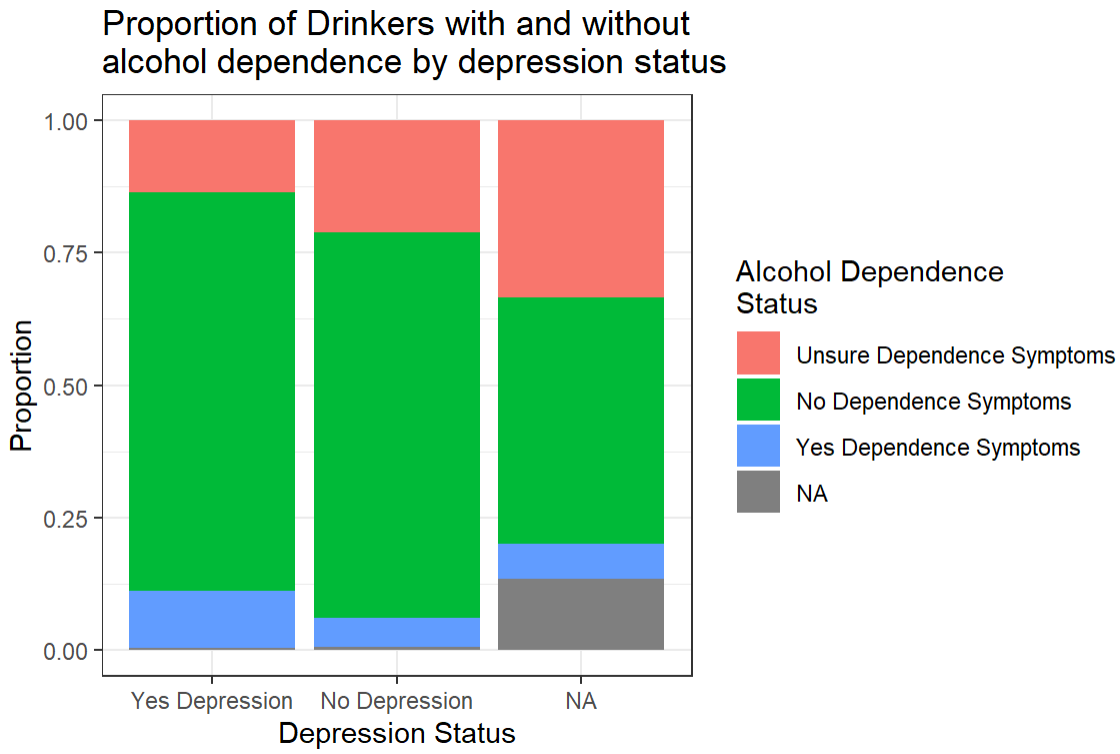
```
`summarise()` has grouped output by 'DrinkDependence'. You can override using
the `.groups` argument.
```

```
tab_dep_alc
```

```
# A tibble: 6 × 4
  DrinkDependence            Depression         n  prop
```

	<fct>	<fct>	<int>	<dbl>
1	Yes Dependence Symptoms	Yes Depression	1362	0.447
2	Yes Dependence Symptoms	No Depression	1627	0.534
3	No Dependence Symptoms	Yes Depression	9634	0.306
4	No Dependence Symptoms	No Depression	21417	0.681
5	Unsure Dependence Symptoms	Yes Depression	1736	0.21
6	Unsure Dependence Symptoms	No Depression	6232	0.754

```
library(ggplot2)
p <- ggplot(data = nesarc_sub, aes(x = Depression, fill =
         forcats::fct_rev(DrinkDependence)))
p <- p + theme_bw()
p <- p + geom_bar(position = "fill")
p <- p + labs(x = "Depression Status"
            , y = "Proportion"
            , title = "Proportion of Drinkers with and without\nalcohol dependence
          by depression status"
            )
p <- p + scale_fill_discrete(name = "Alcohol Dependence\nStatus")

print(p)
```



Interpretation: Those who exhibit symptoms of depression are about twice as likely to exhibit symptoms of alcohol dependence (10.7%) as those without symptoms of depression (5.5%).

5.5.2 Logistic scatter plot (for logistic regression): x = numerical, y = categorical (binary)

```
table(nesarc_sub$DrinkDependence)
```

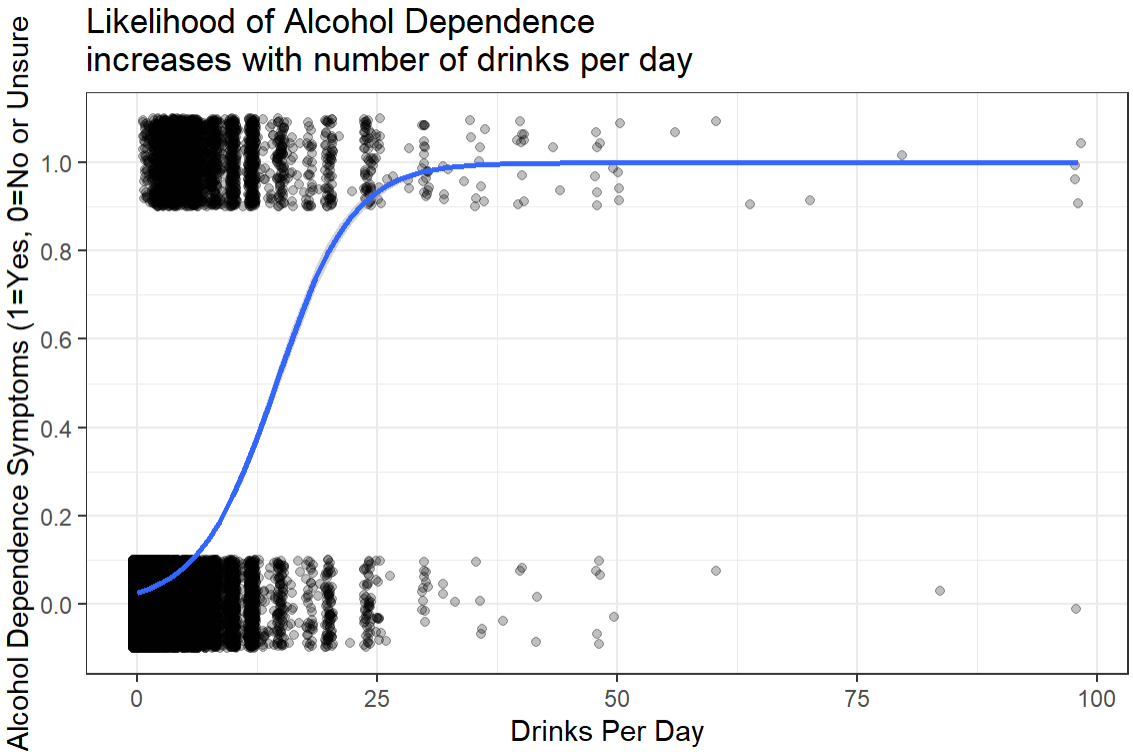
```
   Yes Dependence Symptoms     No Dependence Symptoms
                      3048                      31467
Unsure Dependence Symptoms
                      8266
```

```
nesarc_sub$DrinkDependence01 <- ifelse(nesarc_sub$DrinkDependence == "Yes
         Dependence Symptoms", 1, 0)

table(nesarc_sub$DrinkDependence01)
```

```
    0     1
39733  3048
```

```
library(ggplot2)
p <- ggplot(nesarc_sub, aes(x = DrinkQuantity, y = DrinkDependence01))
p <- p + theme_bw()
p<- p + geom_jitter(position = position_jitter(height = 0.1), alpha = 1/4)

binomial_smooth <- function(...) {
  geom_smooth(method = "glm", method.args = list(family = "binomial"), ...)
}
p <- p + binomial_smooth()
p <- p + scale_y_continuous(breaks = seq(0, 1, by=0.2), minor_breaks = seq(0, 1,
         by=0.1))
p <- p + labs(x = "Drinks Per Day"
            , y = "Alcohol Dependence Symptoms (1=Yes, 0=No or Unsure"
            , title = "Likelihood of Alcohol Dependence\nincreases with number of
         drinks per day"
            )

print(p)
```

```
`geom_smooth()` using formula 'y ~ x'

Warning: Removed 1255 rows containing non-finite values (stat_smooth).

Warning: Removed 1255 rows containing missing values (geom_point).
```



Interpretation: The likelihood of exhibiting symptoms of alcohol dependence increases with the number of drinks consumed per day.

5.6 Class 06, Figure arrangement, captions, cross-referencing

Rubric

1. Reorganize your Class 05 bivariate plots above using `plot_grid()` and `quarto`, creating captions and cross-referencing them from the text.
    1. (3 p) For your numeric response plots, use `cowplot::plot_grid()` to create a single figure with separate plot panels.
    2. (3 p) For your categorical response plots, use quarto chunk options `fig-cap`, `fig-subcap`, and `layout-ncol` to create a single figure with separate plot panels.
    3. (2 p) Use captions to describe (not interpret) both sets of plots so a reader understands what is being plotted.
    4. (2 p) Use cross-referencing from the text to refer to the plots when you interpret them.

> **Note**
>
> Go above and reformat your plots and update your interpretations with cross-referencing

Go above and reformat your plots and update your interpretations with cross referencing.

5.6.1 Scatter plot (for regression): x = numerical, y = numerical

```
library(ggplot2)
p <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
        DrinkExperience))
p <- p + theme_bw()
p <- p + geom_jitter(width = 0.25, height = 5, alpha = 1/10)
p <- p + stat_smooth(method = lm, colour = "red")
p <- p + labs(
     title = "Heavy Drinking vs Age"
   , y = "Length in Years of Heaviest Drinking Period"
   , caption = "Key: Red line is simple linear regression.\nFiltered to include
        only drinkers."
   )

p1 <- p
```

5.6.2 Box plots (for ANOVA): x = categorical, y = numerical

```
library(ggplot2)
p <- ggplot(nesarc_sub %>% filter(DrinkQuantity > 0), aes(x = Depression, y =
        DrinkQuantity))
p <- p + theme_bw()
p <- p + geom_boxplot(width = 0.5, alpha = 0.5)
p <- p + geom_jitter(position = position_jitter(width = 0.1), alpha = 1/100)
p <- p + stat_summary(fun = mean, geom = "point", shape = 18, size = 6,
                      colour = "red", alpha = 0.8)
p <- p + labs(
              title = "Heavy Drinking by Depression"
            , x = "Symptoms of Depression"
            , y = "Drinks per Day During Heaviest Drinking Period"
            , caption = "Red diamond is the mean.\nFiltered to include only
        drinkers"
)

p2 <- p
```

```
p_arranged <-
  cowplot::plot_grid(
      plotlist = list(p1, p2)
    , nrow   = 1
    , ncol   = NULL
    , labels = "AUTO"
```

```
  )
```

```
`geom_smooth()` using formula 'y ~ x'
```

```
print(p_arranged)
```

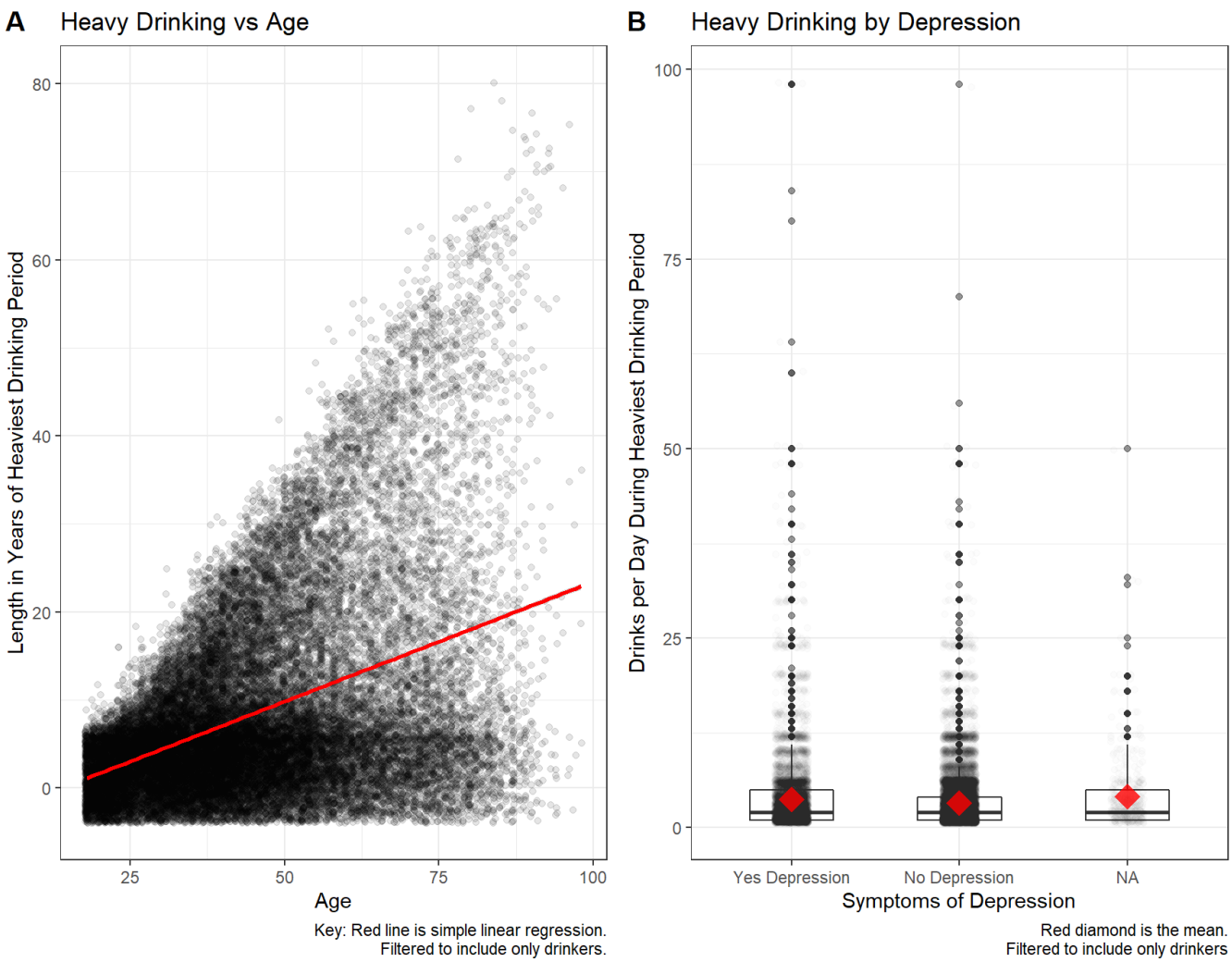


Figure 1: Numeric response bivariate plots. (A) Heavy Drinking vs Age. (B) Heavy Drinking by Depression

Interpretation:

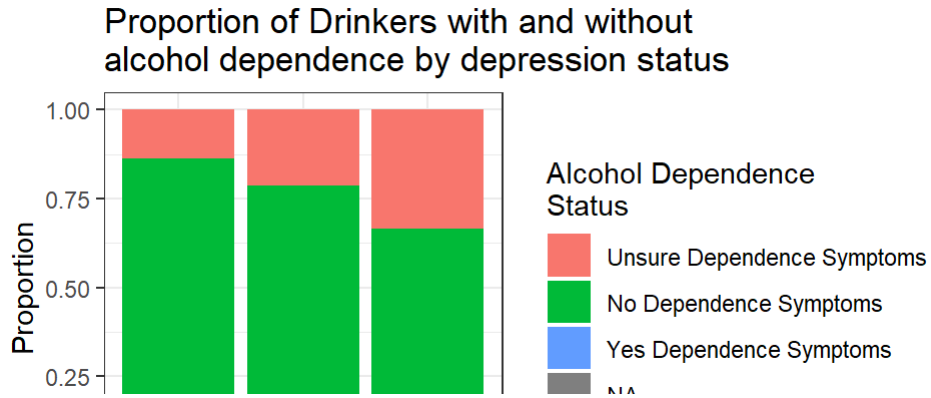
In Figure 1 A, the relationship between `Age` and `DrinkExperience` is distinctly positive.

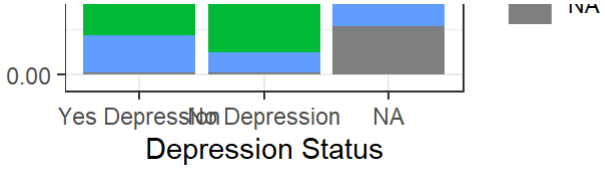
In Figure 1 B, there is little correlation between `DrinkQuantity` and `Depression`.

```
# Mosaic plot or bivariate bar plots (for contingency tables): x = categorical, y =
        categorical


library(ggplot2)
p <- ggplot(data = nesarc_sub, aes(x = Depression, fill =
        forcats::fct_rev(DrinkDependence)))
```

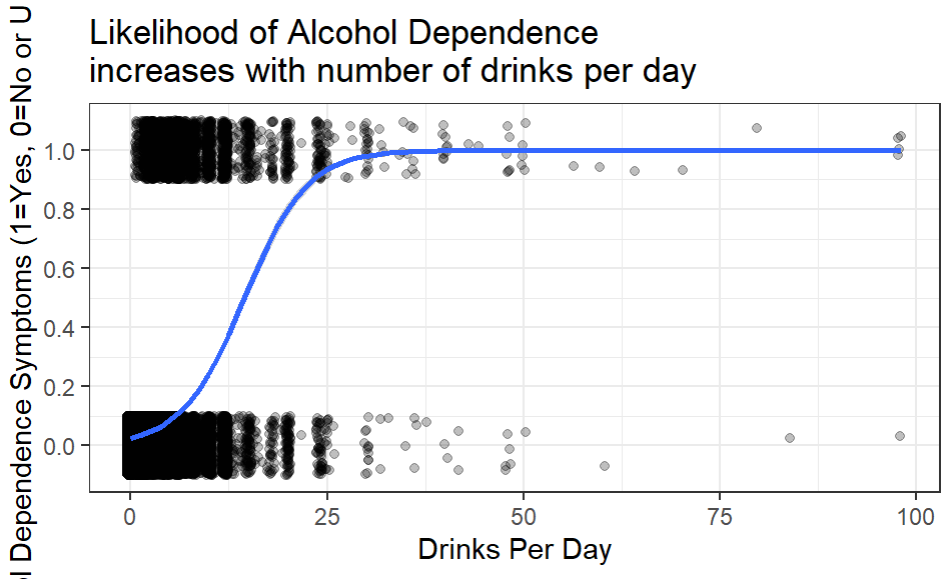
```
p <- p + theme_bw()
p <- p + geom_bar(position = "fill")
p <- p + labs(x = "Depression Status"
              , y = "Proportion"
              , title = "Proportion of Drinkers with and without\nalcohol dependence
            by depression status"
              )
p <- p + scale_fill_discrete(name = "Alcohol Dependence\nStatus")
print(p)

# Logistic scatter plot (for logistic regression): x = numerical, y = categorical
          (binary)

library(ggplot2)
p <- ggplot(nesarc_sub, aes(x = DrinkQuantity, y = DrinkDependence01))
p <- p + theme_bw()
p<- p + geom_jitter(position = position_jitter(height = 0.1), alpha = 1/4)

binomial_smooth <- function(...) {
  geom_smooth(method = "glm", method.args = list(family = "binomial"), ...)
}
p <- p + binomial_smooth()
p <- p + scale_y_continuous(breaks = seq(0, 1, by=0.2), minor_breaks = seq(0, 1,
          by=0.1))
p <- p + labs(x = "Drinks Per Day"
              , y = "Alcohol Dependence Symptoms (1=Yes, 0=No or Unsure"
              , title = "Likelihood of Alcohol Dependence\nincreases with number of
            drinks per day"
              )
print(p)
```

```
`geom_smooth()` using formula 'y ~ x'

Warning: Removed 1255 rows containing non-finite values (stat_smooth).

Warning: Removed 1255 rows containing missing values (geom_point).
```





(a) Proportion of Drinkers with and without alcohol dependence by depression status



(b) Likelihood of Alcohol Dependence increases with number of drinks per day

Figure 2: Categorical response bivariate plots.

Interpretation:

Categorical response bivariate plots are in Figure 2.

Those who exhibit symptoms of depression are about twice as likely to exhibit symptoms of alcohol dependence (10.7%) as those without symptoms of depression (5.5%) (Figure 2 (a)).

The likelihood of exhibiting symptoms of alcohol dependence increases with the number of drinks consumed per day (Figure 2 (b)).

6 Statistical methods

6.1 Class 07-1, Simple linear regression (separate worksheet)

6.2 Class 07-2, Simple linear regression

6.2 Class 07: 2, Simple linear regression

Rubric

1. With your previous (or new) bivariate scatter plot, add a regression line.
   - (2 p) plot with regression line,
   - (1 p) label axes and title.

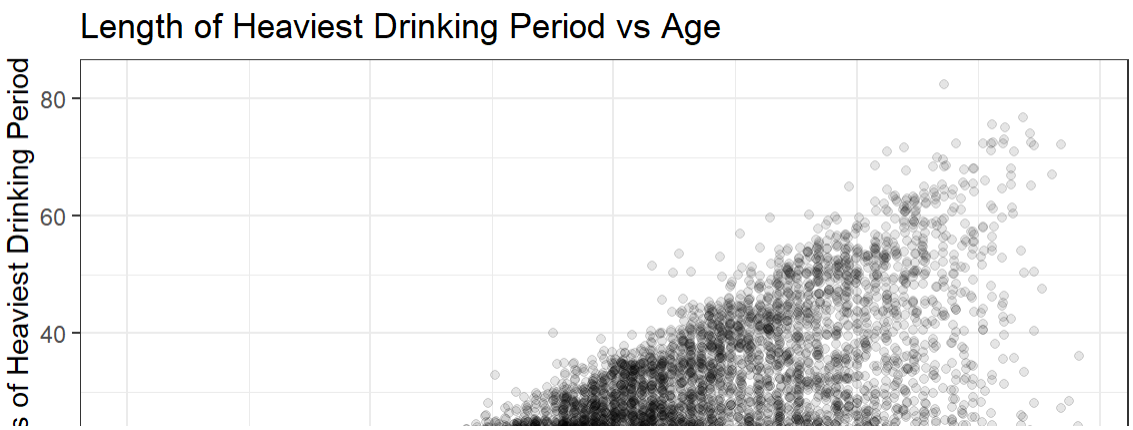
2. Use `lm()` to fit the linear regression and interpret slope and R^2 (R-squared) values.
   - (2 p) lm `summary()` table is presented,
   - (2 p) slope is interpreted with respect to a per-unit increase of the x variable in the context of the variables in the plot,
   - (2 p) R^2 is interpretted in a sentence.

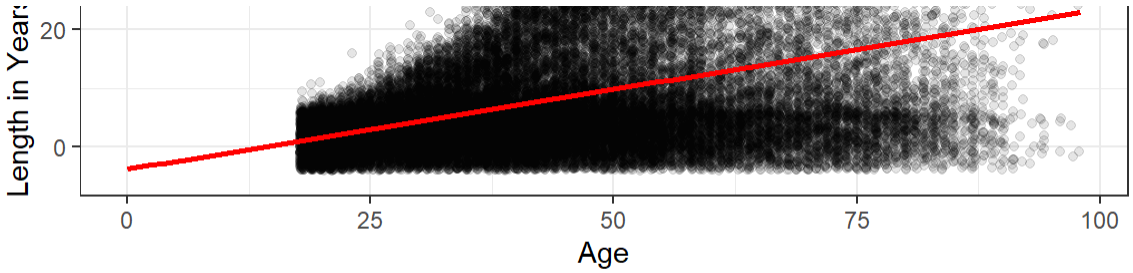
3. (1 p) Interpret the intercept. Does it make sense in the context of your study?

6.2.1 1. Scatter plot, add a regression line.

```
library(ggplot2)
p <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
        DrinkExperience))
p <- p + theme_bw()
p <- p + geom_jitter(width = 0.25, height = 5, alpha = 1/10)
p <- p + stat_smooth(method = lm, colour = "red", fullrange = TRUE)
p <- p + labs(
     title = "Length of Heaviest Drinking Period vs Age"
   , y = "Length in Years of Heaviest Drinking Period"
   , caption = "Key: Red line is simple linear regression.\nFiltered to include
        only drinkers."
   )
p <- p + xlim(c(0, NA))
print(p)
```

```
`geom_smooth()` using formula 'y ~ x'
```





Key: Red line is simple linear regression.
Filtered to include only drinkers.

6.2.2 2. Fit the linear regression, interpret slope and R^2 (R-squared) values

```
lm_DE_Age <-
  lm(
      formula = DrinkExperience ~ Age
    , data = nesarc_sub
  )
summary(lm_DE_Age)
```

```
Call:
lm(formula = DrinkExperience ~ Age, data = nesarc_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-16.059  -5.579  -2.154   2.134  66.434

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.395594   0.131660  -10.60   <2e-16 ***
Age          0.178113   0.002661   66.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.807 on 41641 degrees of freedom
  (1450 observations deleted due to missingness)
Multiple R-squared:  0.09711,   Adjusted R-squared:  0.09709
F-statistic:  4479 on 1 and 41641 DF,  p-value: < 2.2e-16
```

Slope: For every 1 year increase in Age, we expect an increase of 0.178 years in the period of heaviest drinking.

R^2: The proportion of variance explained by the regression model, compared to the grand mean, is R^2 = 0.097, which is very small.

6.2.3 3. Interpret the intercept. Does it make sense?

Intercept: For someone 0 years old, the expected number of years of heaviest drinking is -1.4 years. This does not make sense since you cannot have a negative span of heaviest drinking.

6.3 Class 08-1, Logarithm transformation (separate worksheet)

6.4 Class 08-2, Logarithm transformation

Rubric

1. Try plotting the data on a logarithmic scale
   - (6 p) Each of the logarithmic relationships is plotted, axes are labelled with scale.

   1. original scales
   2. $\log(x)$-only
   3. $\log(y)$-only
   4. both $\log(x)$ and $\log(y)$

2. What happened to your data when you transformed it?
   - (2 p) Describe what happened to the relationship after each log transformation (compare transformed scale to original scale; is the relationship more linear, more curved?).
   - (1 p) Choose the best scale for a linear relationship and explain why.
   - (1 p) Does your relationship benefit from a logarithmic transformation? Say why or why not.

6.4.1 1. Try plotting on log scale (original scale, $\log(x)$-only, $\log(y)$ -only, both $\log(x)$ and $\log(y)$)

```
library(ggplot2)
p1 <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
        DrinkExperience))
p1 <- p1 + theme_bw()
p1 <- p1 + geom_point(alpha = 1/20)
p1 <- p1 + stat_smooth(method = lm)
p1 <- p1 + labs(
            title = "Original scale"
          , x = "Age"
```

```
          , y = "Heaviest drinking period in years"
          )
#print(p1)

library(ggplot2)
p2 <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
         DrinkExperience))
p2 <- p2 + theme_bw()
p2 <- p2 + geom_point(alpha = 1/20)
p2 <- p2 + stat_smooth(method = lm)
p2 <- p2 + scale_x_log10()
p2 <- p2 + labs(
            title = "Log(x)"
          , x = "Age"
          , y = "Heaviest drinking period in years"
          )
#print(p2)

library(ggplot2)
p3 <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
         DrinkExperience))
p3 <- p3 + theme_bw()
p3 <- p3 + geom_point(alpha = 1/20)
p3 <- p3 + stat_smooth(method = lm)
p3 <- p3 + scale_y_log10()
p3 <- p3 + labs(
            title = "Log(y)"
          , x = "Age"
          , y = "Heaviest drinking period in years"
          )
#print(p3)

library(ggplot2)
p4 <- ggplot(nesarc_sub %>% filter(DrinkExperience > 0), aes(x = Age, y =
         DrinkExperience))
p4 <- p4 + theme_bw()
p4 <- p4 + geom_point(alpha = 1/20)
p4 <- p4 + stat_smooth(method = lm)
p4 <- p4 + scale_x_log10()
p4 <- p4 + scale_y_log10()
p4 <- p4 + labs(
            title = "Log(x) and Log(y)"
          , x = "Age"
          , y = "Heaviest drinking period in years"
          )
#print(p4)
```

6.4.2 2. What happened to your data when you transformed it?

- Describe what happened to the relationship after each log transformation (compare transformed scale to original scale).
- Choose the best scale for a linear relationship and explain why.
- Does your relationship benefit from a logarithmic transformation? Say why or why not.

6.5 Class 09, Correlation (separate worksheet)

6.6 Class 10, Categorical contingency tables (separate worksheet)

6.7 Class 11, Correlation and Categorical contingency tables

Rubric

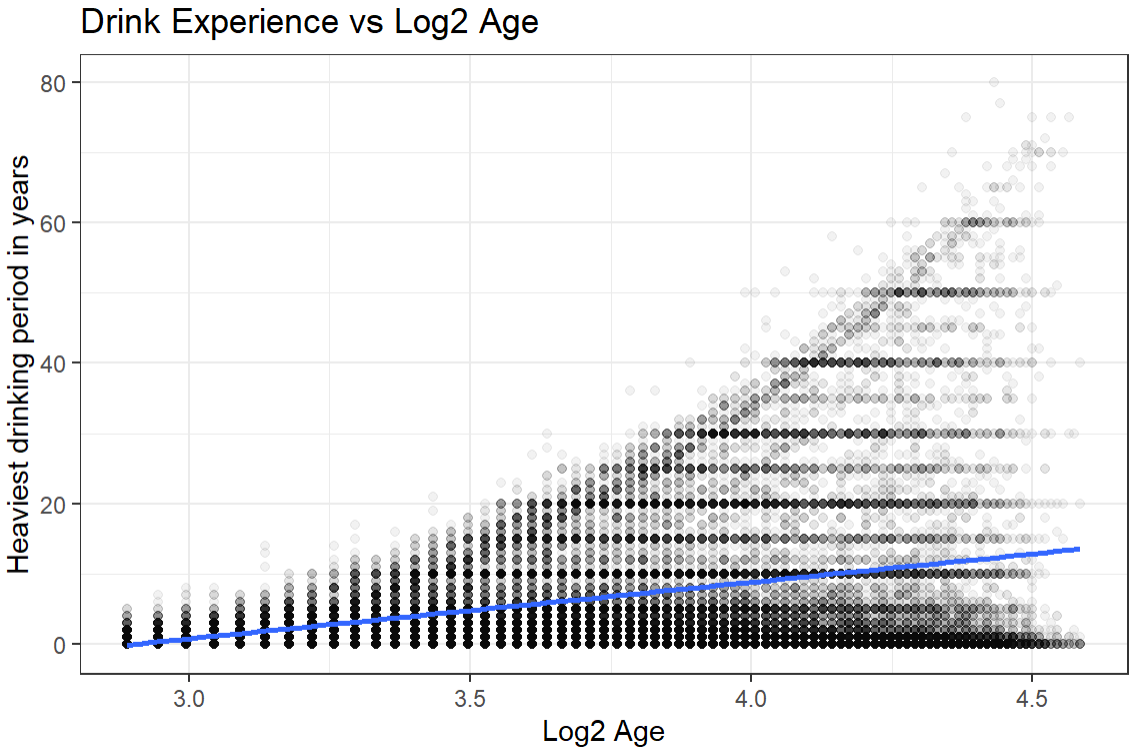
1. With your previous (or a new) bivariate scatter plot, calculate the correlation and interpret.
   - (1 p) plot is repeated here or the plot is referenced an easy to find from a plot above,
   - (1 p) correlation is calculated,
   - (2 p) correlation is interpretted (direction, strength of LINEAR relationship).
2. With your previous (or a new) two- or three-variable categorical plot, calculate conditional proportions and interpret.
   - (1 p) frequency table of variables is given,
   - (2 p) conditional proportion tables are calculated of the outcome variable conditional on one or two other variables,
   - (1 p) a well-labelled plot of the proportion table is given,
   - (2 p) the conditional proportions are interpretted and compared between conditions.

6.7.1 Correlation

```
library(ggplot2)
p2 <- ggplot(nesarc_sub, aes(x = Age_log, y = DrinkExperience))
p2 <- p2 + theme_bw()
p2 <- p2 + geom_point(alpha = 1/20)
p2 <- p2 + stat_smooth(method = lm)
```

```
p2 <- p2 + labs(
            title = "Drink Experience vs Log2 Age"
          , x = "Log2 Age"
          , y = "Heaviest drinking period in years"
          )
print(p2)
```

```
`geom_smooth()` using formula 'y ~ x'

Warning: Removed 1450 rows containing non-finite values (stat_smooth).

Warning: Removed 1450 rows containing missing values (geom_point).
```




```
cor_A_D <-
  cor(
      nesarc_sub$Age_log
    , nesarc_sub$DrinkExperience
    , use = "complete.obs"
  )
cor_A_D
```

```
[1] 0.3198571
```

6.7.2 Interpretation of correlation

The correlation is 0.32. This is a positive, moderately weak linear relationship between

`Age_log` and `DrinkExperience`, meaning that as people get older, their longest period of heaviest drinking tends to increase slightly.

6.7.3 Contingency table

```
tab_S_D_D <-
  nesarc_sub %>%
  group_by(
    Sex
  , Depression
  , DrinkDependence
  ) %>%
  summarise(
    Frequency = n()
  ) %>%
  mutate(
    Proportion = round(Frequency / sum(Frequency), 3)
  ) %>%
  ungroup()
```

```
`summarise()` has grouped output by 'Sex', 'Depression'. You can override using
the `.groups` argument.
```

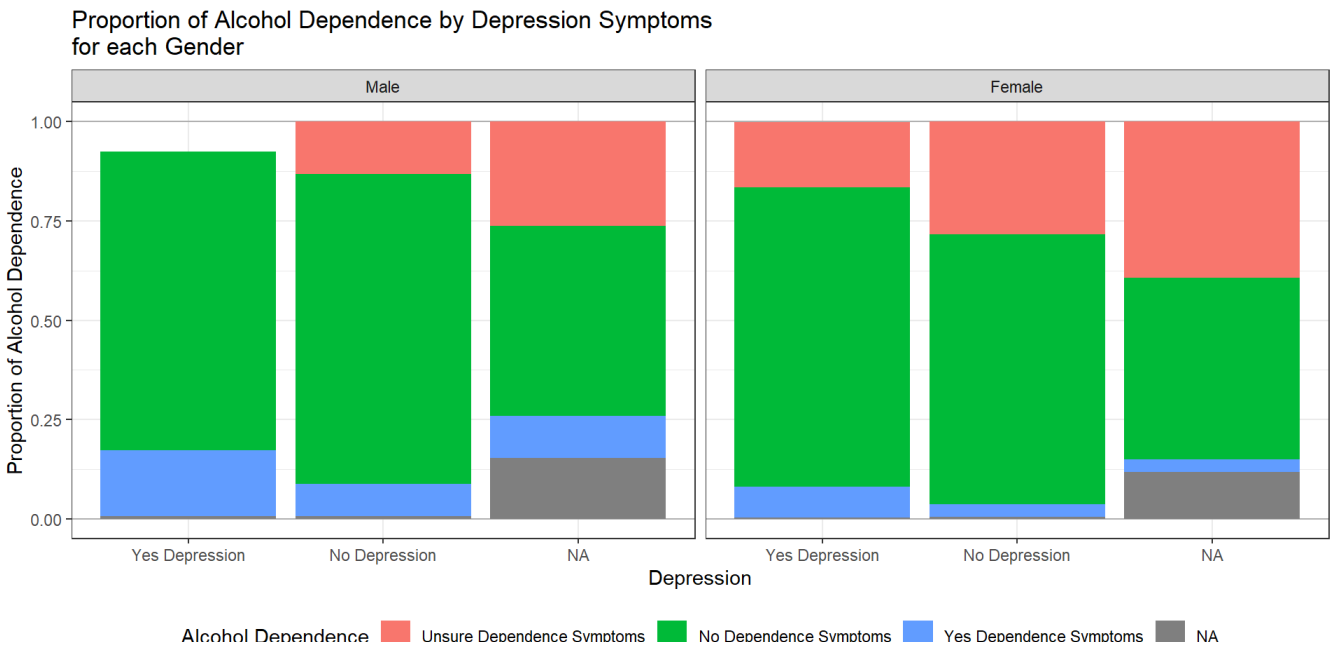
```
tab_S_D_D %>%
  knitr::kable()
```

Sex	Depression	DrinkDependence	Frequency	Proportion
Male	Yes Depression	Yes Dependence Symptoms	704	0.165
Male	Yes Depression	No Dependence Symptoms	3211	0.752
Male	Yes Depression	Unsure Dependence Symptoms	328	0.077
Male	Yes Depression	NA	28	0.007
Male	No Depression	Yes Dependence Symptoms	1127	0.081
Male	No Depression	No Dependence Symptoms	10825	0.782
Male	No Depression	Unsure Dependence Symptoms	1810	0.131
Male	No Depression	NA	80	0.006
Male	NA	Yes Dependence Symptoms	43	0.106
Male	NA	No Dependence Symptoms	194	0.479
Male	NA	Unsure Dependence Symptoms	106	0.262
Male	NA	NA	62	0.153
Female	Yes Depression	Yes Dependence Symptoms	658	0.077
Female	Yes Depression	No Dependence Symptoms	6423	0.754
Female	Yes Depression	Unsure Dependence Symptoms	1408	0.165

Female	Yes Depression	NA	25	0.003
Female	No Depression	Yes Dependence Symptoms	500	0.032
Female	No Depression	No Dependence Symptoms	10592	0.680
Female	No Depression	Unsure Dependence Symptoms	4422	0.284
Female	No Depression	NA	60	0.004
Female	NA	Yes Dependence Symptoms	16	0.033
Female	NA	No Dependence Symptoms	222	0.456
Female	NA	Unsure Dependence Symptoms	192	0.394
Female	NA	NA	57	0.117

```
library(ggplot2)
p <- ggplot(data = tab_S_D_D, aes(x = Depression, y = Proportion, fill =
          forcats::fct_rev(DrinkDependence)))
p <- p + theme_bw()
p <- p + geom_hline(yintercept = c(0, 1), alpha = 1/4)
p <- p + geom_bar(stat="identity")
p <- p + labs(
    title = "Proportion of Alcohol Dependence by Depression Symptoms\nfor each
         Gender"
  , y = "Proportion of Alcohol Dependence"
  , fill = "Alcohol Dependence"
  )
p <- p + scale_y_continuous(limits = c(0, 1))
p <- p + facet_wrap(~ Sex)
p <- p + theme(legend.position = "bottom")
print(p)
```

```
Warning: Removed 1 rows containing missing values (geom_bar).
```



6.7.4 Interpretation of conditional proportions

- For Males with depression, the proportion who exhibit alcohol dependence is 0.165. For those without depression, the proportion is 0.081 (about half).
- For Females with depression, the proportion who exhibit alcohol dependence is 0.077. For those without depression, the proportion is 0.032 (less than half).

Both Males and Females are about twice as likely to develop alcohol dependence if they are depressed compared to if they are not depressed.

6.8 Class 12-1, Parameter estimation (one-sample) (separate worksheet)

6.9 Class 12-2, Inference and Parameter estimation (one-sample)

Rubric

1. Using a numerical variable, calculate and interpret a confidence interval for the population mean.
   - (1 p) Identify and describe the variable,
   - (1 p) use `t.test()` to calculate the mean and confidence interval, and
   - (1 p) interpret the confidence interval.
   - (2 p) Using plotting code from the last two classes, plot the data, estimate, and confidence interval in a single well-labelled plot.
2. Using a two-level categorical variable, calculate and interpret a confidence interval for the population proportion.
   - (1 p) Identify and describe the variable,
   - (1 p) use `binom.test()` to calculate the mean and confidence interval, and
   - (1 p) interpret the confidence interval.
   - (2 p) Using plotting code from the last two classes, plot the data, estimate, and confidence interval in a single well-labelled plot.

6.9.1 Numeric variable confidence interval for mean μ

6.9.2 Categorical variable confidence interval for proportion p

6.10 Class 13, Hypothesis testing (one- and two-sample) (separate worksheet)

6.11 Class 14, Paired data, assumption assessment (separate worksheet)

6.12 Class 15, Hypothesis testing (one- and two-sample)

6.12.1 Mechanics of a hypothesis test (review)

1. Set up the **null and alternative hypotheses** in words and notation.
   - In words: ``The population mean for [what is being studied] is different from [value of μ_0].'' (Note that the statement in words is in terms of the alternative hypothesis.)
   - In notation: $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$ (where μ_0 is specified by the context of the problem).
2. Choose the **significance level** of the test, such as $\alpha = 0.05$.
3. Compute the **test statistic**, such as $t_s = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}}$, where $SE_{\bar{Y}} = s/\sqrt{n}$ is the standard error.
4. Determine the **tail(s)** of the sampling distribution where the p**-value** from the test statistic will be calculated (for example, both tails, right tail, or left tail). (Historically, we would compare the observed test statistic, t_s, with the **critical value** $t_{\text{crit}} = t_{\alpha/2}$ in the direction of the alternative hypothesis from the t-distribution table with degrees of freedom $df = n - 1$.)
5. State the **conclusion** in terms of the problem.
   - Reject H_0 in favor of H_A if p-value $< \alpha$.
   - Fail to reject H_0 if p-value $\geq \alpha$. (Note: We DO NOT *accept* H_0.)
6. **Check assumptions** of the test (for now we skip this).

6.12.2 What do we do about “significance”?

Adapted from **Significance Magazine**.

Recent calls have been made to abandon the term “statistical significance”. The American Statistical Association (ASA) issued its statement and recommendation on p-values (see the

Statistical Association (ASA) issued its statement and recommendation on p-values (see the special issue of p-values for more).

In summary, the problem of "significance" is one of misuse, misunderstanding, and misinterpretation. The recommendation in this class is that it is no longer sufficient to say that a result is "statistically significant" or "non-significant" depending on whether a p-value is less than a threshold. Instead, we will be looking for wording as in the following paragraph.

"The difference between the two groups turns out to be small (8%), while the probability (p) of observing a result at least as extreme as this under the null hypothesis of no difference between the two groups is $p = 0.003$ (that is, 0.3%). This p-value is statistically significant as it is below our pre-defined threshold ($p < 0.05$). However, the p-value tells us only that the 8% difference between the two groups is somewhat unlikely given our hypothesis and null model's assumptions. More research is required, or other considerations may be needed, to conclude that the difference is of practical importance and reproducible."

6.12.3 Two-sample t-test

Rubric

1. Using a numerical response variable and a two-level categorical variable (or a categorical variable you can reduce to two levels), specify a two-sample t-test associated with your research questions.
    - (2 p) Specify the hypotheses in words and notation (either one- or two-sided test),
    - (0 p) use `t.test()` to calculate the mean, test statistic, and p-value,
    - (3 p) state the significance level, test statistic, and p-value, and
    - (2 p) state the conclusion in the context of the problem.
    - (1 p) Given your conclusion, could you have committed at Type-I or Type-II error?
    - (2 p) Provide an appropriate plot of the data and sample estimates in a well-labelled plot.

1. Null and Alternative Hypotheses

- ``The true difference in means of drink consumption (DrinkExperience) between group Yes Manic Symptoms and group No Manic Symptoms is less than 0."
- $H_0 : \mu_{YM} - \mu_{NM} \geq 0$ versus $H_A : \mu_{YM} - \mu_{NM} < 0$

```
t_summary_DE_M <-
  t.test(
      DrinkExperience ~ Mania
    , data = nesarc_sub
    , alternative = "less"
  )
```

```
t_summary_DE_M
```

```
    Welch Two Sample t-test

data:  DrinkExperience by Mania
t = -11.104, df = 4519, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Yes Manic Symptoms and
group No Manic Symptoms is less than 0
95 percent confidence interval:
      -Inf -1.376622
sample estimates:
mean in group Yes Manic Symptoms  mean in group No Manic Symptoms
                        5.357612                          6.973675
```

2. Significance level

- Let $\alpha = 0.05$, the significance level of the test and the Type-I error probability if the null hypothesis is true.

3. Test Statistic

- $t_s = -11.1$

4. p-value

- $p = 1.39 \times 10^{-28}$, this is the observed significance of the test.

5. Conclusion

- Reject H_0 in favor of H_A, concluding that the population mean total of drink consumption is higher among those with mania than those without.
- Because we rejected the null hypothesis, we could have made a Type-I error.

6. Plot and Sample Estimates

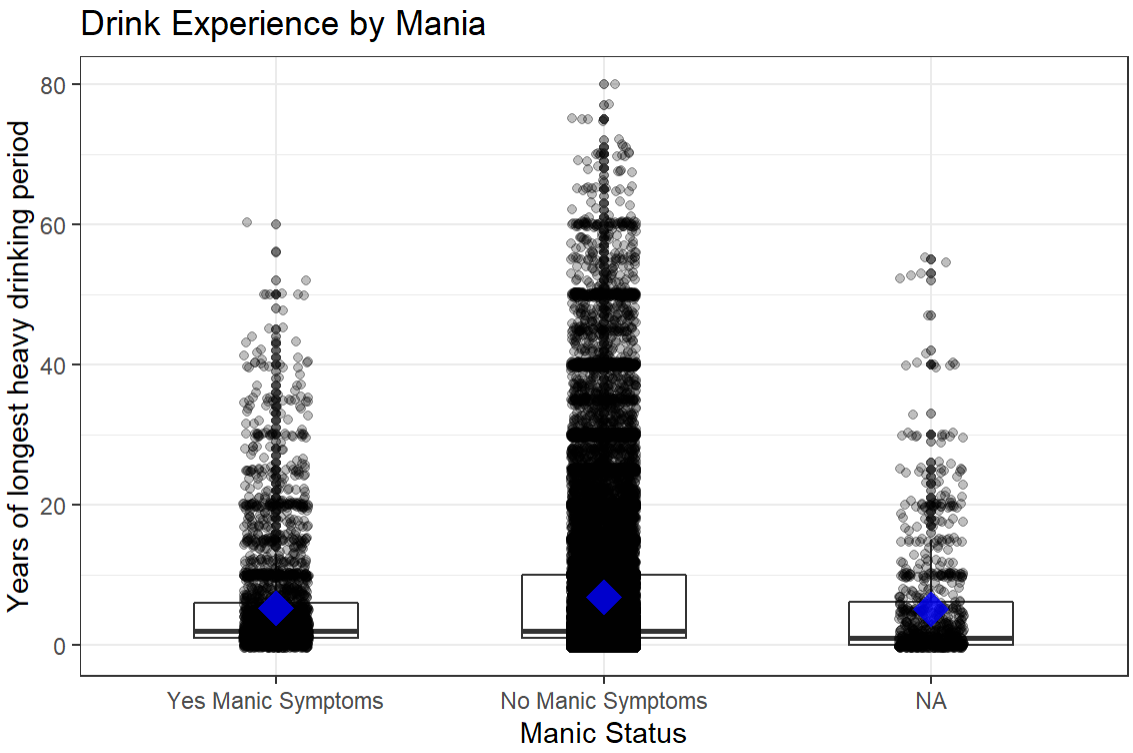
```
est_mean_DE_M <-
  nesarc_sub %>%
  group_by(Mania) %>%
  summarise(DrinkExperience = mean(DrinkExperience, na.rm = TRUE)) %>%
  ungroup()
est_mean_DE_M
```

```
# A tibble: 3 × 2
  Mania              DrinkExperience
```

```
  Mania                 DrinkExperience
  <fct>                           <dbl>
1 Yes Manic Symptoms               5.36
2 No Manic Symptoms                6.97
3 <NA>                             5.27
```

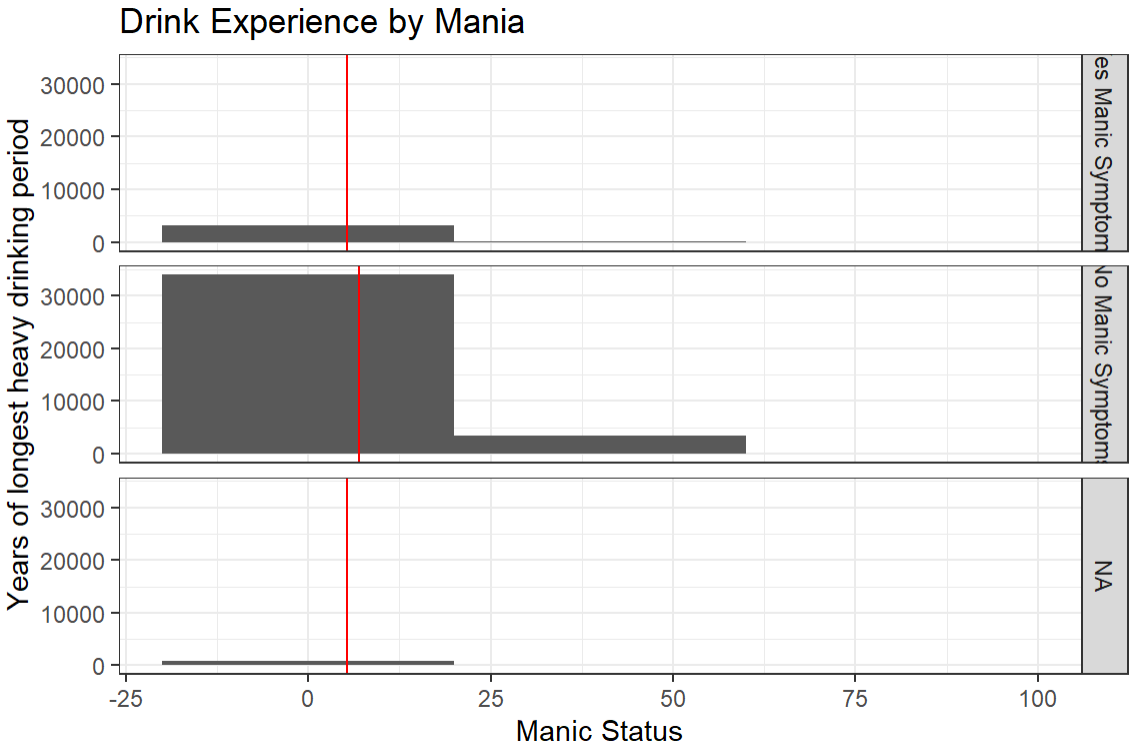
```
library(ggplot2)
p <- ggplot(nesarc_sub, aes(x = Mania, y = DrinkExperience))
p <- p + theme_bw()
p <- p + geom_boxplot(width = 0.5, alpha = 0.5)
p <- p + geom_jitter(position = position_jitter(width = 0.1), alpha = 1/4)
p <- p + stat_summary(fun = mean, geom = "point", shape = 18, size = 6, colour =
          "blue", alpha = 0.8)
p <- p + labs(
               title = "Drink Experience by Mania"
             , x = "Manic Status"
             , y = "Years of longest heavy drinking period"
             )
print(p)
```

```
Warning: Removed 1450 rows containing non-finite values (stat_boxplot).

Warning: Removed 1450 rows containing non-finite values (stat_summary).

Warning: Removed 1450 rows containing missing values (geom_point).
```

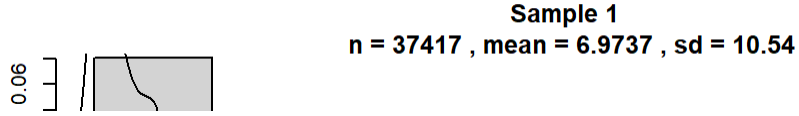


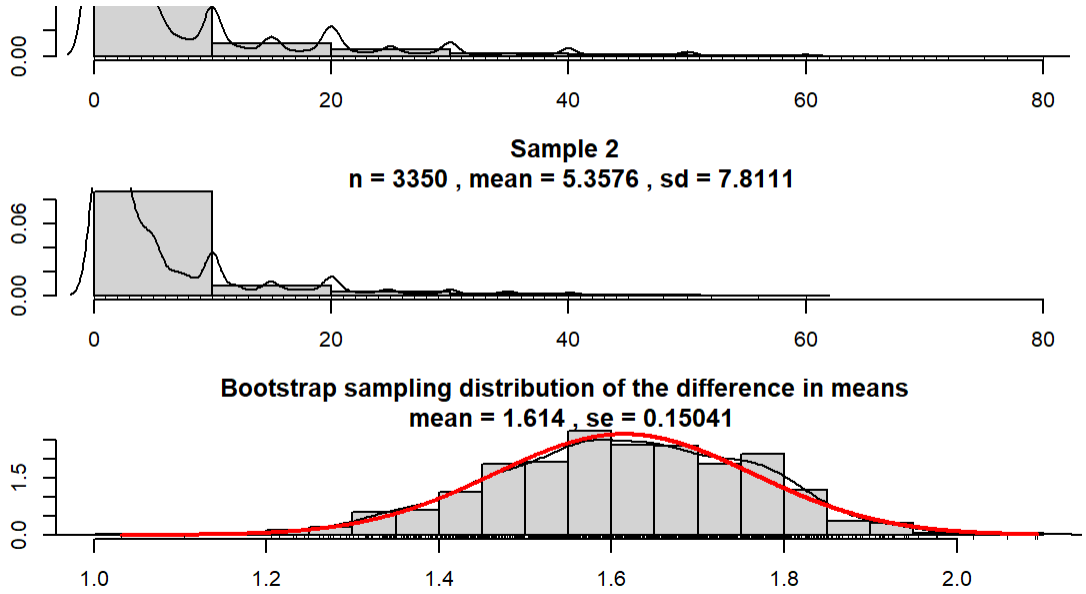
```
p <- ggplot(nesarc_sub, aes(x = DrinkExperience))
p <- p + theme_bw()
p <- p + geom_histogram(binwidth = 40)
p <- p + geom_vline(data = est_mean_DE_M, aes(xintercept = DrinkExperience), colour
        = "red")
p <- p + facet_grid(Mania ~ .)
p <- p + labs(
             title = "Drink Experience by Mania"
           , x = "Manic Status"
           , y = "Years of longest heavy drinking period"
           )
print(p)
```

```
Warning: Removed 1450 rows containing non-finite values (stat_bin).
```




```
e_plot_bs_two_samp_diff_dist(
  nesarc_sub %>% filter(Mania == "No Manic Symptoms") %>% drop_na(DrinkExperience)
         %>% pull(DrinkExperience)
, nesarc_sub %>% filter(Mania == "Yes Manic Symptoms") %>% drop_na(DrinkExperience)
         %>% pull(DrinkExperience)
, N = 1000
)
```





6.13 Class 16, ANOVA, Pairwise comparisons (separate worksheet)

6.14 Class 17, ANOVA and Assessing Assumptions

Rubric

1. Using a numerical response variable and a categorical variable with three to five levels (or a categorical variable you can reduce to three to five levels), specify an ANOVA hypothesis associated with your research questions.
    - (1 p) Specify the ANOVA hypotheses in words and notation,
    - (1 p) plot the data in a way that is consistent with hypothesis test (comparing means, assess equal variance assumption),
    - (1 p) use `aov()` to calculate the hypothesis test statistic and p-value,
    - (1 p) state the significance level, test statistic, and p-value,
    - (1 p) state the conclusion in the context of the problem,
    - (2 p) assess the normality assumption of the residuals using appropriate methods (QQ-plot and Anderson-Darling test), and
    - (1 p) assess the assumption of equal variance between your groups using an appropriate test (also mention standard deviations of each group).
    - (2 p) If you rejected the ANOVA null hypothesis, perform follow-up pairwise comparisons using Tukey's HSD to indicate which groups have statistically different means and summarize the results.

6.14.1 Hypothesis and plot

Let μ_j = population mean longest consecutive years of heaviest drinking for the five ethnicities identified in the dataset: Caucasian, African American, Native American, Asian, and Hispanic, numbered $(j = 1, 2, 3, 4, 5)$ respectively. We wish to test $H_0 : \mu_1 = \cdots = \mu_5$ versus H_A : not H_0 (at least one pair of means differ).

```
summary(nesarc_sub$Ethnicity)
```

```
 Cauc  AfAm  NaAm  Asia  Hisp
24507  8245   701  1332  8308
```

```
summary(nesarc_sub$DrinkExperience)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.000   2.000   6.808  10.000  80.000    1450
```

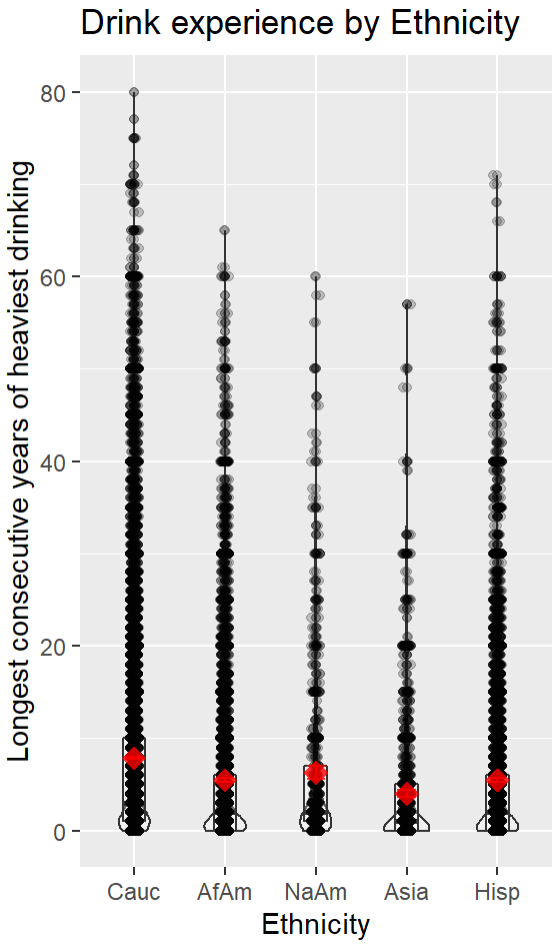
Plot the data in a way that compares the means. Error bars are 95% confidence intervals of the mean.

```
# Plot the data using ggplot
library(ggplot2)
p <- ggplot(nesarc_sub, aes(x = Ethnicity, y = DrinkExperience))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(yintercept = mean(nesarc_sub$c_bwt),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
```

```
Warning in mean.default(nesarc_sub$c_bwt): argument is not numeric or logical:
returning NA
```

```
# boxplot, size=.75 to stand out behind CI
p <- p + geom_violin(width = 0.5, alpha = 0.25)
p <- p + geom_boxplot(width = 0.25, alpha = 0.25)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.2)
# diamond at mean for each group
p <- p + stat_summary(fun = mean, geom = "point", shape = 18, size = 4,
                      colour = "red", alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, colour = "red", alpha = 0.8)
p <- p + labs(title = "Drink experience by Ethnicity")
p <- p + ylab("Longest consecutive years of heaviest drinking")
print(p)
```

```
Warning: Removed 1450 rows containing non-finite values (stat_ydensity).

Warning: Removed 1450 rows containing non-finite values (stat_boxplot).

Warning: Removed 1450 rows containing non-finite values (stat_summary).
Removed 1450 rows containing non-finite values (stat_summary).

Warning: Removed 1 rows containing missing values (geom_hline).

Warning: Removed 1450 rows containing missing values (geom_point).
```



###(NOT USED) Transform the response variable to satisfy assumptions

6.14.2 ANOVA Hypothesis test

1. Set up the **null and alternative hypotheses** in words and notation.
    - In words: ``The population mean drink experience is different between ethnic groups.''
    - In notation: $H_0 : \mu_1 = \cdots = \mu_3$ versus H_A : not H_0 (at least one pair of means differ).

2. Let the significance level of the test be $\alpha = 0.05$.

3. Compute the **test statistic**.

```
aov_summary <-
  aov(
    DrinkExperience ~ Ethnicity
  , data = nesarc_sub
  )
summary(aov_summary)
```

```
              Df  Sum Sq Mean Sq F value Pr(>F)
Ethnicity      4   65561   16390   156.2 <2e-16 ***
Residuals  41638 4369934     105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1450 observations deleted due to missingness
```

The F-statistic for the ANOVA is $F = 156$.

4. Compute the p-**value** from the test statistic.

The p-value for testing the null hypothesis is $p = 7.05 \times 10^{-133}$.

5. State the **conclusion** in terms of the problem.

Because $p = 7.05 \times 10^{-133} < 0.05$, we must reject H_0 in favor of H_A and conclude that at least one pair of means differ.

6.14.3 Check assumptions

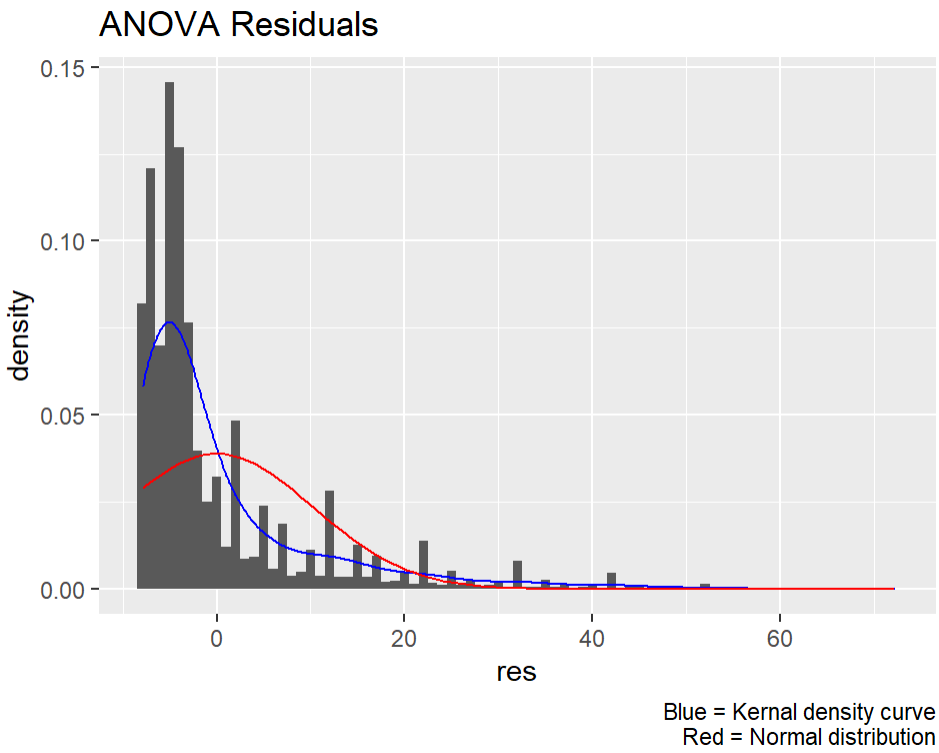
6. **Check assumptions** of the test.

a. Residuals are normal
b. Populations have equal variances.

- Check whether residuals are normal.
- Plot the residuals and assess whether they appear normal.

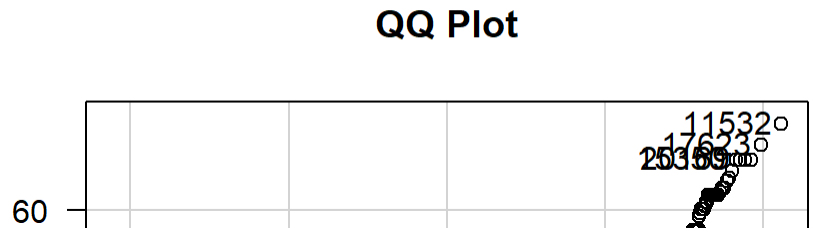
```
# Plot the data using ggplot
df_res <- data.frame(res = aov_summary$residuals)
library(ggplot2)
p <- ggplot(df_res, aes(x = res))
p <- p + geom_histogram(aes(y = ..density..), binwidth = 1)
p <- p + geom_density(colour = "blue", adjust = 5)
#p <- p + geom_rug()
```

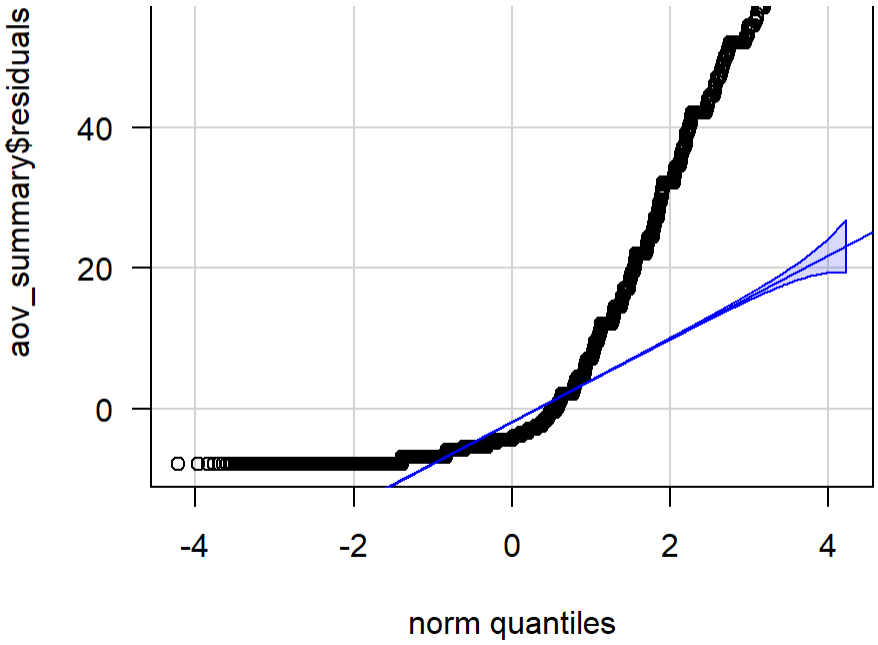
```
#p <- p + geom_rug()
p <- p + stat_function(fun = dnorm, colour = "red", args = list(mean =
        mean(df_res$res), sd = sd(df_res$res)))
p <- p + labs(title = "ANOVA Residuals"
            , caption = "Blue = Kernal density curve\nRed = Normal distribution")
print(p)
```



The residuals plot is highly right skewed, and thus not normal.

```
# QQ plot
par(mfrow=c(1,1))
#library(car)
car::qqPlot(
    aov_summary$residuals
  , las = 1
  , id = list(n = 4, cex = 1)
  , lwd = 1
  , main="QQ Plot"
  )
```






```
11532 17623 15350 20169
11146 17046 14852 19500
```

The QQ-plot of the residuals versus normal quantiles is U-shaped (very right skewed), and thus not normal.

- A formal test of normality on the residuals tests the hypothesis H_0 : The distribution is Normal vs H_1 : The distribution is not Normal. We can test the distribution of the residuals.

```
#shapiro.test(aov_summary$residuals)
#library(nortest)
nortest::ad.test(aov_summary$residuals)
```

```
    Anderson-Darling normality test

data:  aov_summary$residuals
A = 4136.3, p-value < 2.2e-16
```

```
nortest::cvm.test(aov_summary$residuals)
```

```
Warning in nortest::cvm.test(aov_summary$residuals): p-value is smaller than
7.37e-10, cannot be computed more accurately
```

```
    Cramer-von Mises normality test
```

```
data:  aov_summary$residuals
W = 785.23, p-value = 7.37e-10
```

The formal normality tests of the residuals reject H_0 in favor of H_A, concluding that the data are not normal.

- Check whether populations have equal variances.
- Look at the numerical summaries below.

```
# calculate summaries
dat_EthDExp_summary <-
  nesarc_sub %>%
  group_by(Ethnicity) %>%
  summarize(
    m = mean(DrinkExperience, na.rm = TRUE)
  , s = sd(DrinkExperience, na.rm = TRUE)
  , n = n()
  , .groups = "drop_last"
  ) %>%
  ungroup()

dat_EthDExp_summary
```

```
# A tibble: 5 × 4
  Ethnicity     m     s     n
  <fct>     <dbl> <dbl> <int>
1 Cauc       7.87 11.3  24507
2 AfAm       5.48  8.72  8245
3 NaAm       6.29  9.77   701
4 Asia       3.99  7.08  1332
5 Hisp       5.48  8.83  8308
```

The standard deviations appear different between groups; in particular, the Caucasian group has a significantly higher standard deviation than the other groups.

- Formal tests for equal variances. We can test whether the variances are equal between our three groups. This is similar to the ANOVA hypothesis, but instead of testing means we're tesing variances. $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ versus H_A : not H_0 (at least one pair of variances differ).

```
## Test equal variance
# assumes populations are normal
bartlett.test(DrinkExperience ~ Ethnicity, data = nesarc_sub)
```

```
    Bartlett test of homogeneity of variances

data:  DrinkExperience by Ethnicity
Bartlett's K-squared = 1433.8, df = 4, p-value < 2.2e-16
```

```
# does not assume normality, requires car package
#library(car)
car::leveneTest(DrinkExperience ~ Ethnicity, data = nesarc_sub)
```

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     4  113.18 < 2.2e-16 ***
      41638
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# nonparametric test
fligner.test(DrinkExperience ~ Ethnicity, data = nesarc_sub)
```

```
    Fligner-Killeen test of homogeneity of variances

data:  DrinkExperience by Ethnicity
Fligner-Killeen:med chi-squared = 979.61, df = 4, p-value < 2.2e-16
```

Since the data were not normal, we will look at the Levene test. Since the p-value is less than 0.05, we reject H_0 (equal variances) in favor of H_A (unequal variances).

6.14.4 Post Hoc pairwise comparison tests

EMM plot interpretation

This **EMM plot (Estimated Marginal Means, aka Least-Squares Means)** is only available when conditioning on one variable. The blue bars are confidence intervals for the EMMs; don't ever use confidence intervals for EMMs to perform comparisons – they can be very misleading. The red arrows are for the comparisons among means; the degree to which the "comparison arrows" overlap reflects as much as possible the significance of the comparison of the two estimates. If an arrow from one mean overlaps an arrow from another group, the difference is not significant, based on the adjust setting (which defaults to "tukey").

```
## CHDS
# Tukey 95% Individual p-values
```

```
#TukeyHSD(fit_c)

## Contrasts
adjust_method <- c("none", "tukey", "bonferroni")[2]

library(emmeans)
emm_cont <-
  emmeans::emmeans(
    aov_summary
  , specs = "Ethnicity"
  )

# means and CIs
emm_cont %>% print()
```

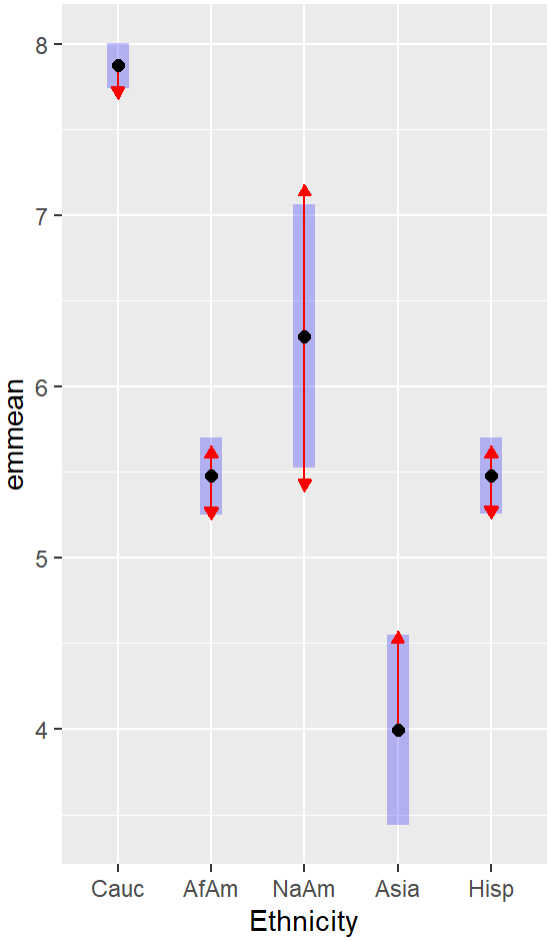
```
 Ethnicity emmean     SE    df lower.CL upper.CL
 Cauc        7.87 0.0665 41638     7.74     8.00
 AfAm        5.48 0.1154 41638     5.25     5.70
 NaAm        6.29 0.3917 41638     5.52     7.06
 Asia        3.99 0.2837 41638     3.44     4.55
 Hisp        5.48 0.1141 41638     5.26     5.70

Confidence level used: 0.95
```

```
# pairwise differences
emm_cont %>% pairs(adjust = adjust_method) %>% print()
```

```
 contrast    estimate    SE    df t.ratio p.value
 Cauc - AfAm  2.39561 0.133 41638  17.987  <.0001
 Cauc - NaAm  1.57962 0.397 41638   3.976  0.0007
 Cauc - Asia  3.87892 0.291 41638  13.312  <.0001
 Cauc - Hisp  2.39229 0.132 41638  18.115  <.0001
 AfAm - NaAm -0.81599 0.408 41638  -1.998  0.2667
 AfAm - Asia  1.48331 0.306 41638   4.843  <.0001
 AfAm - Hisp -0.00332 0.162 41638  -0.020  1.0000
 NaAm - Asia  2.29930 0.484 41638   4.754  <.0001
 NaAm - Hisp  0.81267 0.408 41638   1.992  0.2698
 Asia - Hisp -1.48663 0.306 41638  -4.862  <.0001

P value adjustment: tukey method for comparing a family of 5 estimates
```

```
# plot of means, CIs, and comparison arrows
plot(
    emm_cont
```

```
  , comparisons = TRUE
  , adjust = adjust_method
  , horizontal = FALSE
  , ylab = "Ethnicity"
  )
```



African American, Native American, and Hispanic ethnic groups do not significantly differ from one another, but Caucasian and Asian ethnic groups both differ significantly from all other ethnic groups.

- `Cauc - AfAm`: means are significantly different
- `Cauc - NaAm`: means are significantly different
- `Cauc - Asia`: means are significantly different
- `Cauc - Hisp`: means are significantly different
- `AfAm - NaAm`: means are NOT significantly different
- `AfAm - Asia`: means are significantly different
- `AfAm - Hisp`: means are NOT significantly different
- `NaAm - Asia`: means are significantly different
- `NaAm - Hisp`: means are NOT significantly different
- `Asia - Hisp`: means are significantly different

```
dat_EthDExp_summary %>% arrange(m)
```

```
# A tibble: 5 × 4
  Ethnicity     m     s     n
  <fct>     <dbl> <dbl> <int>
1 Asia       3.99  7.08  1332
2 AfAm       5.48  8.72  8245
3 Hisp       5.48  8.83  8308
4 NaAm       6.29  9.77   701
5 Cauc       7.87 11.3  24507
```

Summarize results by ordering the means and grouping pairs that do not differ.

	Ethnicity	m	s	n
	<fctr>	<dbl>	<dbl>	<int>
|	Asia	3.993098	7.080555	1332
|	AfAm	5.476408	8.723156	8245
|	Hisp	5.479727	8.832371	8308
|	NaAm	6.292398	9.767241	701
|	Cauc	7.872016	11.269414	24507

6.15 Class 18, Nonparametric methods (separate worksheet)

6.16 Class 19, Binomial and Multinomial tests (separate worksheet)

6.17 Class 20-1, Two-way categorical tables (separate worksheet)

6.18 Class 20-2, Simple linear regression (separate worksheet)

6.19 Class 21, Two-way categorical and simple linear regression

Rubric

Rubric

6.19.1 Two-way categorical analysis

Using two categorical variables with two to five levels each, specify a hypothesis test for homogeneity of proportions associated with your research questions.

- `Mania` and `Ethnicity`

6.19.1.1 (1 p) Specify the hypotheses in words and notation.

- In words: "There is an association between `Mania` and `Ethnicity`."
- In notation: $H_0 : p(i \text{ and } j) = p(i)p(j)$ for all row categories i and column categories j versus $H_A : p(i \text{ and } j) \neq p(i)p(j)$, for at least one row category i and column category j.

6.19.1.2 (1 p) State the conclusion of the test in the context of the problem.

```
# Row: Ethnicity Column: Mania

# Tabulate by two categorical variables:
tab_Mania_Eth <-
  xtabs(
    ~ Ethnicity + Mania
  , data = nesarc_sub
  )
tab_Mania_Eth
```

```
         Mania
Ethnicity Yes Manic Symptoms No Manic Symptoms
     Cauc                2080             21853
     AfAm                 582              7400
     NaAm                 105               574
     Asia                  65              1239
     Hisp                 570              7554
```

```
# column proportions
prop.table(
    tab_Mania_Eth
  , margin = 1
  ) %>%
  signif(2)
```

```
         Mania
Ethnicity Yes Manic Symptoms No Manic Symptoms
     Cauc               0.087             0.910
```

```
Cauc            0.057           0.910
AfAm            0.073           0.930
NaAm            0.150           0.850
Asia            0.050           0.950
Hisp            0.070           0.930
```

```
# chi^2 test
chisq_me <-
  chisq.test(
    tab_Mania_Eth
  , correct = FALSE
  )
chisq_me
```

```
    Pearson's Chi-squared test

data:  tab_Mania_Eth
X-squared = 97.567, df = 4, p-value < 2.2e-16
```

- The test statistic is $X^2 = 97.57$.
- The p-value $= 3.24 \times 10^{-20}$.
- Because 3.24^{-20} < 0.05, we must reject H_0 in favor f H_A and conclude that there is an association between Ethnicity and Mania.

```
chisq_me$expected
```

```
         Mania
Ethnicity Yes Manic Symptoms No Manic Symptoms
     Cauc         1937.55809        21995.4419
     AfAm          646.20351         7335.7965
     NaAm           54.97021          624.0298
     Asia          105.56870         1198.4313
     Hisp          657.69949         7466.3005
```

```
min(chisq_me$expected)
```

```
[1] 54.97021
```

- The model assumptions are met since the expected count for each cell is at least 5.

6.19.1.3 (1 p) Plot a mosaic plot of the data and Pearson residuals.

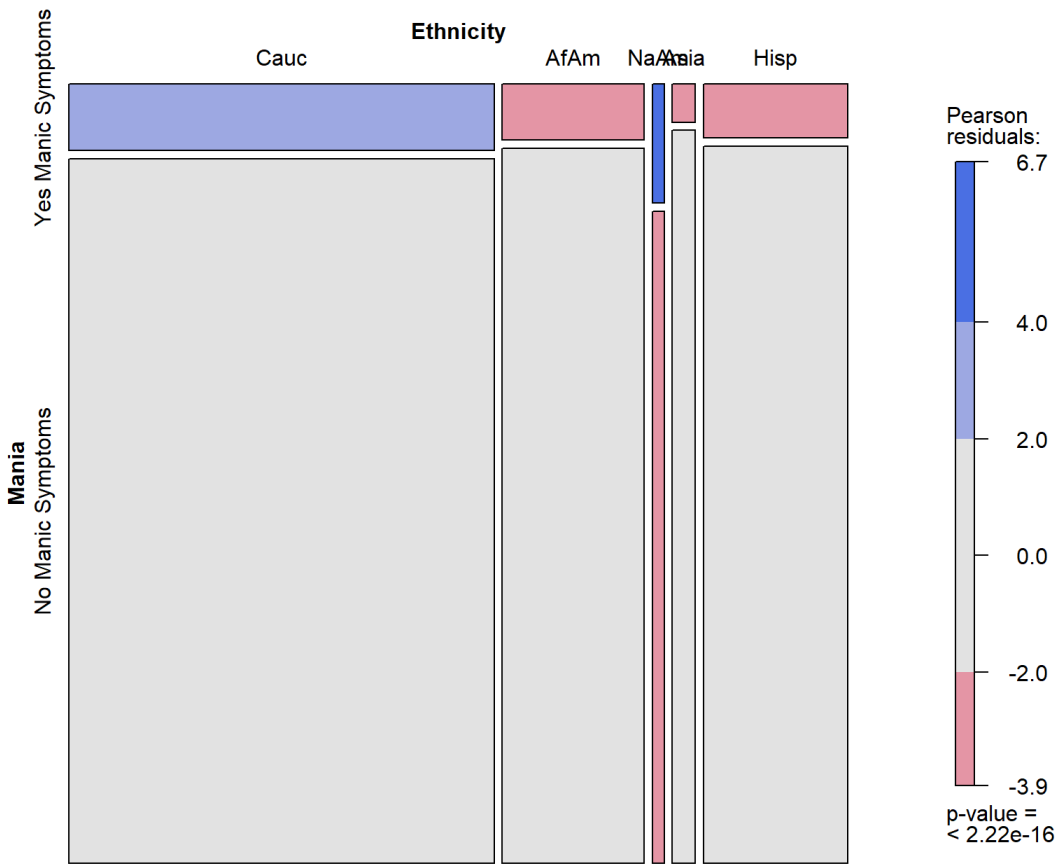
```
chisq_me$residuals
```

```
        Mania
Ethnicity Yes Manic Symptoms No Manic Symptoms
     Cauc          3.2360143        -0.9604427
     AfAm         -2.5256556         0.7496096
     NaAm          6.7478440        -2.0027468
     Asia         -3.9484216         1.1718838
     Hisp         -3.4196631         1.0149493
```

```
# mosaic plot
library(vcd)
```

```
Loading required package: grid
```

```
vcd::mosaic(
    tab_Mania_Eth
  , shade     = TRUE
  , legend    = TRUE
  , direction = "v"
  )
```



6.19.1.4 (1 p) Interpret the mosaic plot with reference to the Pearson residuals.

- Along the Yes Manic symptoms row, we see that the Caucasian group is slightly higher than expected, while the African American, Hispanic, and Asian groups are all slightly lower than expected. The Native American group has a very high positive Pearson residual, however. We may thus conclude that the Native American group is the primary cause for rejecting H_0.

6.19.2 Simple linear regression

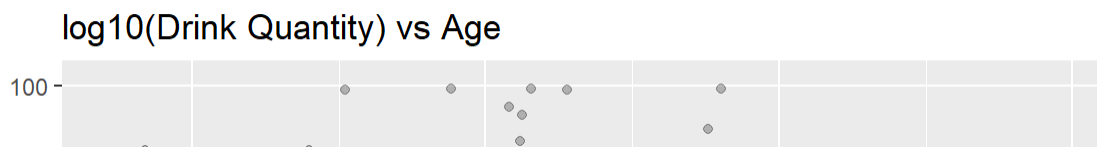
Select two numerical variables.

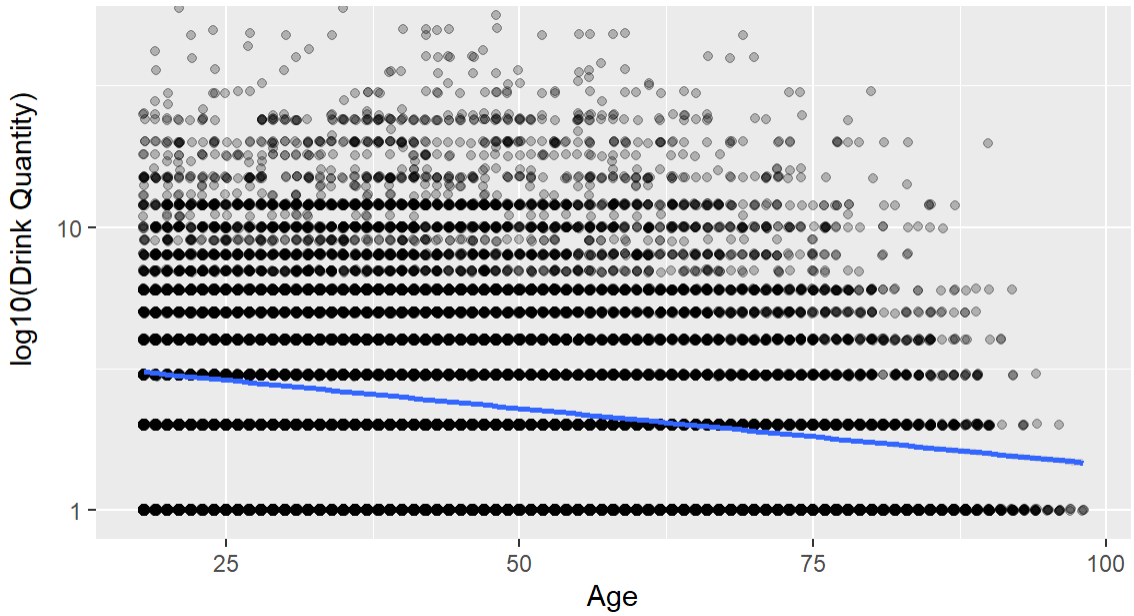
- `DrinkExperience` and `DrinkQuantity`

6.19.2.1 (1 p) Plot the data and, if required, transform the variables so a roughly linear relationship is observed. All interpretations will be done on this scale of the variables.

```
library(ggplot2)
p <- ggplot(nesarc_sub, aes(x = Age, y = DrinkQuantity))
p <- p + geom_jitter(position = position_jitter(width = 0.1), alpha = 1/4)
p <- p + stat_smooth(method = lm)
p <- p + scale_y_log10()
p <- p + labs(title = "log10(Drink Quantity) vs Age")
p <- p+ labs(y = "log10(Drink Quantity)")
print(p)
```

```
Warning: Transformation introduced infinite values in continuous y-axis
Transformation introduced infinite values in continuous y-axis

`geom_smooth()` using formula 'y ~ x'

Warning: Removed 9410 rows containing non-finite values (stat_smooth).

Warning: Removed 9410 rows containing missing values (geom_point).
```



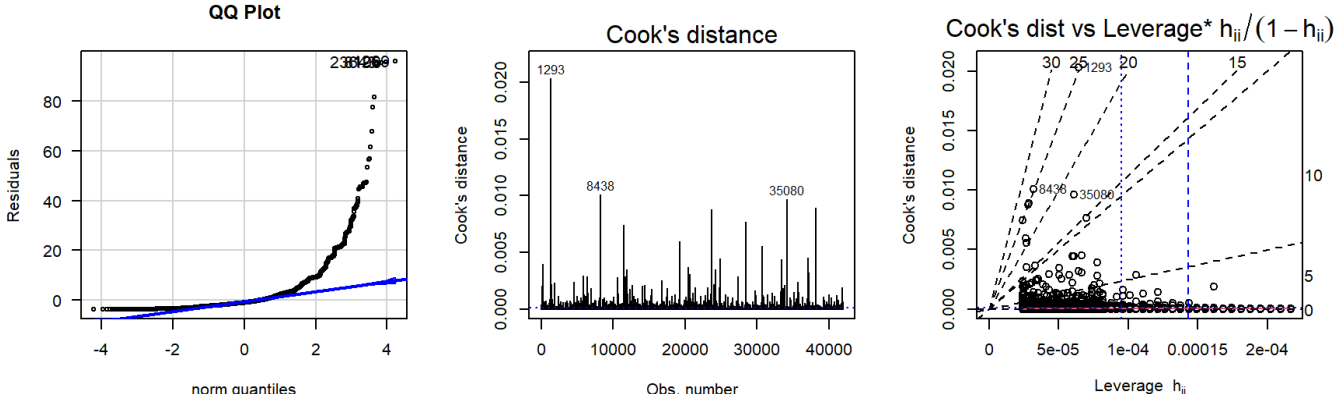


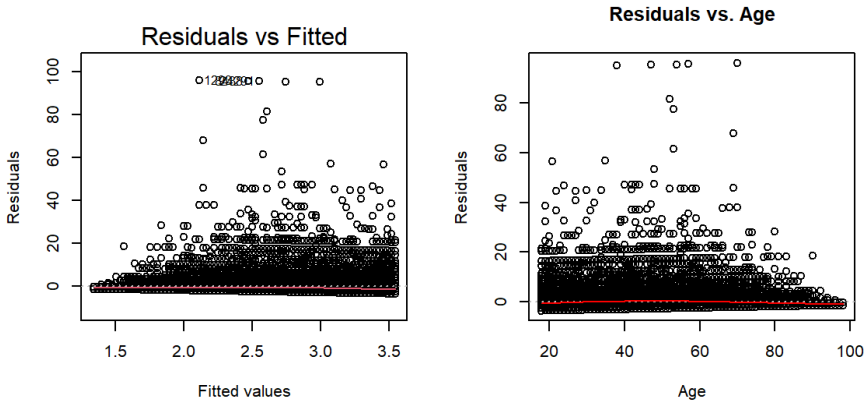
6.19.2.2 (0 p) Fit the simple linear regression model.

```
# fit model
lm_fit <-
  lm(
    DrinkQuantity ~ Age
  , data = nesarc_sub
  )
```

6.19.2.3 (1 p) Assess the residuals for lack of fit (interpret plots of residuals vs fitted and x-value).

```
e_plot_lm_diagostics(
    lm_fit
  #, rc_mfrow    = c(1, 2)
  , sw_plot_set = "simple"
  )
```





6.19.2.4 (1 p) Assess the residuals for normality (interpret QQ-plot and histogram).

6.19.2.5 (1 p) Assess the relative influence of points.

6.19.2.6 (1 p) Test whether the slope is different from zero, $H_A : \beta_1 \neq 0$.

6.19.2.7 (1 p) Interpret the R^2 value.

6.20 Class 22, Logistic regression (separate worksheet)

6.21 Class 23, Logistic regression

Rubric

1. Logistic regression.

6.21.1 Select a binary response and continue explanatory/predictor variable.

6.21.2 (1 p) Plot the data.

- See below.

6.21.3 (1 p) Summarize the $\hat{p}$ values for each value of the x-variable. Also, calculate the empirical logits.

Summarize observed probability of dependency for each quantity of drinks consumed per day.

```r
dat_drinkexp_depend_sum <-
  nesarc_sub %>%
  group_by(
    DrinkQuantity_Drinkers_log2
  ) %>%
  summarize(
    Success = sum(DrinkDependence01)
  , Total   = n()
  # estimated proportion of preg for each age group
  , p_hat   = Success / Total
  , .groups = "drop_last"
  ) %>%
  ungroup() %>%
  na.omit()

dat_drinkexp_depend_sum %>% print(n=Inf)
```

```
# A tibble: 34 × 4
   DrinkQuantity_Drinkers_log2 Success Total p_hat
                         <dbl>   <dbl> <int> <dbl>
 1                        3.17      39   149 0.262
 2                        3.46       7    31 0.226
 3                        3.70       9    30 0.3
 4                        3.81      13    28 0.464
 5                        4         13    23 0.565
 6                        4.09       4     7 0.571
 7                        4.17      29    64 0.453
 8                        4.25       1     3 0.333
 9                        4.39       1     1 1
10                        4.46       1     2 0.5
11                        4.64      15    28 0.536
12                        4.70       1     3 0.333
13                        4.75       1     1 1
14                        4.81       3     3 1
15                        4.91      20    32 0.625
16                        5          4     6 0.667
17                        5.04       0     1 0
18                        5.09       1     1 1
19                        5.13       4     5 0.8
20                        5.17       5     8 0.625
21                        5.25       0     1 0
22                        5.32       9    11 0.818
23                        5.39       0     2 0
24                        5.43       1     1 1
25                        5.46       1     1 1
```

```
26                      5.58       6    11 0.545
27                      5.64       5     6 0.833
28                      5.81       1     1 1
29                      5.91       1     2 0.5
30                      6          1     1 1
31                      6.13       1     1 1
32                      6.32       1     1 1
33                      6.39       0     1 0
34                      6.61       4     5 0.8
```

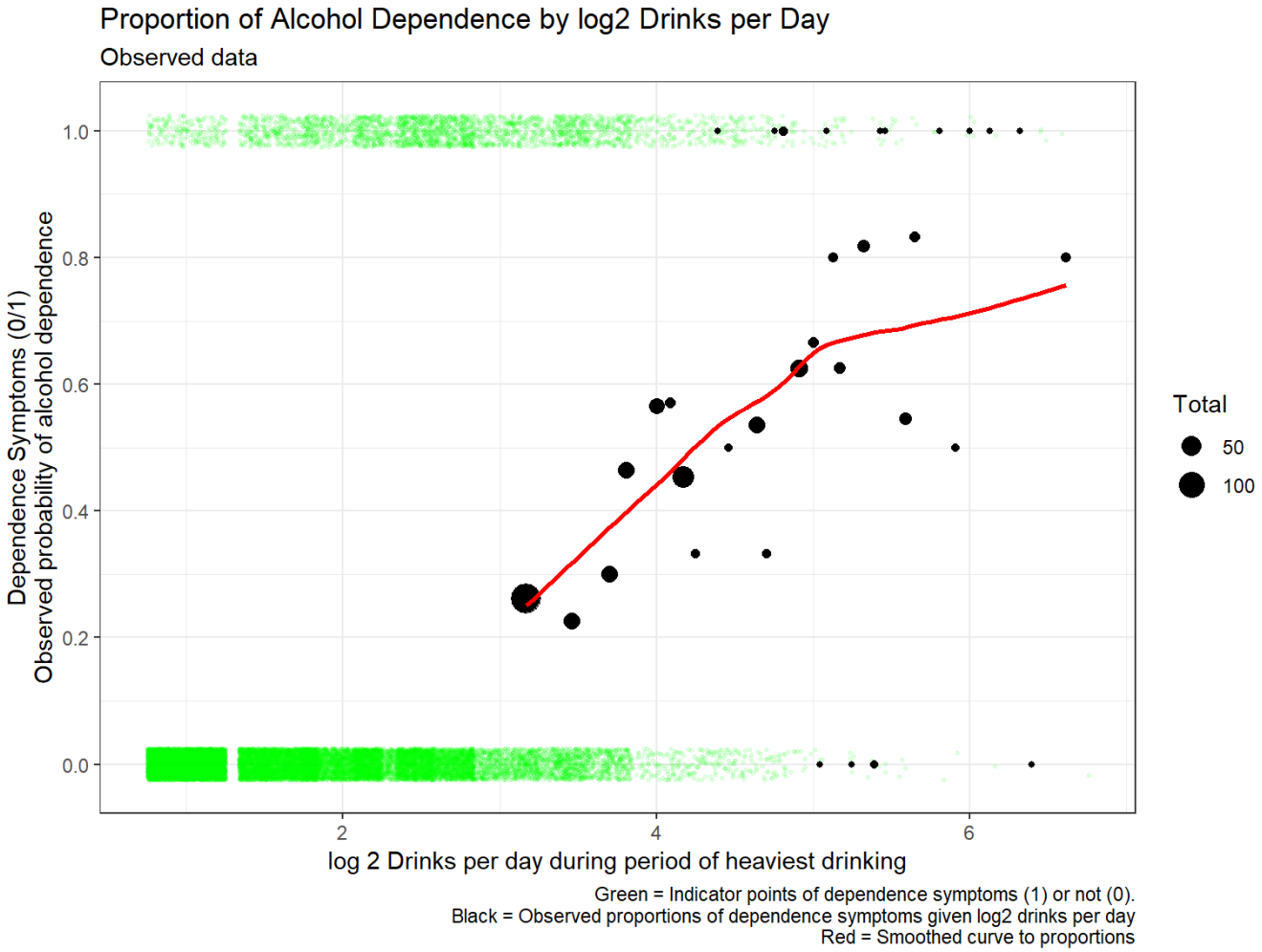
6.21.4 (1 p) Plot the $\hat{p}$ values vs the x-variable and plot the empirical logits vs the x-variable.

6.21.4.1 Probability/proportion scale

```
library(ggplot2)
p1 <- ggplot(nesarc_sub, aes(x = DrinkQuantity_Drinkers_log2, y =
        DrinkDependence01))
p1 <- p1 + theme_bw()
# data
p1 <- p1 + geom_jitter(width = 0.25, height = 0.025, size = 0.5, alpha = 1/10,
        colour = "green")
# summaries
p1 <- p1 + geom_point(data = dat_drinkexp_depend_sum, aes(x =
        DrinkQuantity_Drinkers_log2, y = p_hat, size = Total))
p1 <- p1 + geom_smooth(data = dat_drinkexp_depend_sum, aes(x =
        DrinkQuantity_Drinkers_log2, y = p_hat,weight = Total), se = FALSE, colour
        = "red")  # just for reference
# axes
p1 <- p1 + scale_y_continuous(breaks = seq(0, 1, by = 0.2))
p1 <- p1 + expand_limits(y = c(0, 1))
#p1 <- p1 + scale_x_continuous(breaks = seq(0, 100, by = 2))
# labels
p1 <- p1 + labs(
    title   = "Proportion of Alcohol Dependence by log2 Drinks per Day"
  , subtitle= "Observed data"
  , x       = "log 2 Drinks per day during period of heaviest drinking"
  , y       = "Dependence Symptoms (0/1)\nObserved probability of alcohol
        dependence"
  , caption = paste(
                "Green = Indicator points of dependence symptoms (1) or not (0)."
              , "Black = Observed proportions of dependence symptoms given log2
        drinks per day"
              , "Red = Smoothed curve to proportions"
              , sep = "\n" # separate by new lines
              )
  )
```

```
  )
print(p1)
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'

Warning: Removed 20894 rows containing missing values (geom_point).
```

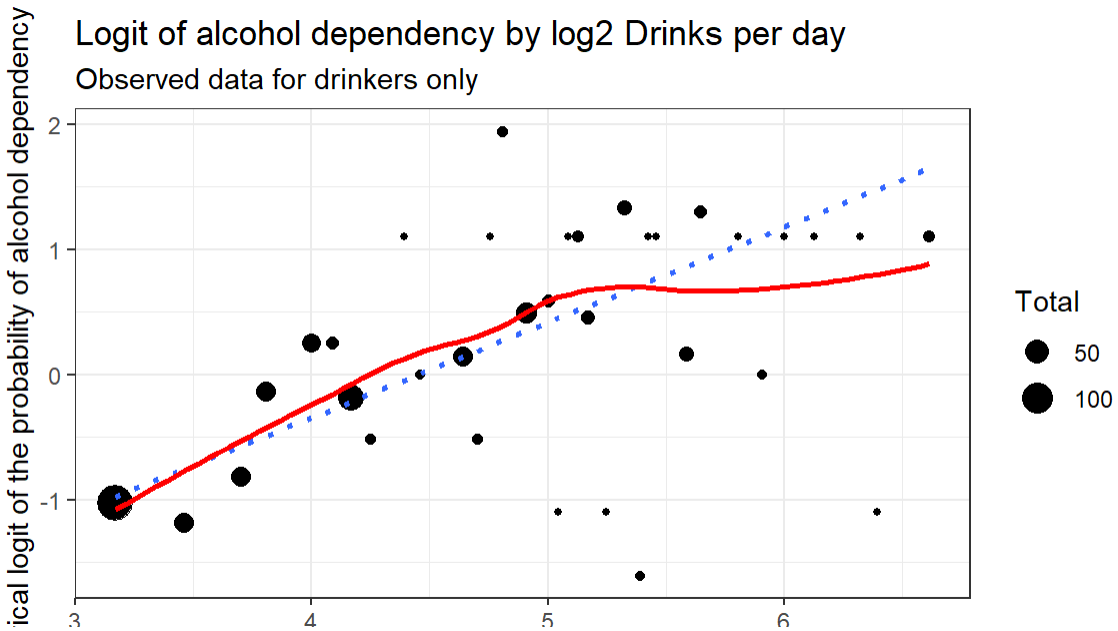


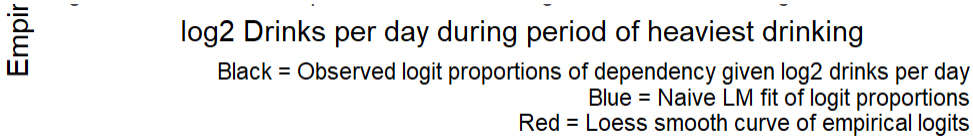
6.21.4.2 Logit scale

```
# emperical logits
dat_drinkexp_depend_sum <-
  dat_drinkexp_depend_sum %>%
  mutate(
    emp_logit = log((p_hat + 0.5/Total) / (1 - p_hat + 0.5/Total))
  )

# plot on logit scale
library(ggplot2)
p1 <- ggplot(dat_drinkexp_depend_sum, aes(x = DrinkQuantity_Drinkers_log2, y =
         emp_logit))
p1 <- p1 + theme_bw()
```

```
p1 <- p1 + theme_bw()

# summaries
p1 <- p1 + geom_point(aes(size = Total))
p1 <- p1 + stat_smooth(aes(weight = Total), method = "lm", se = FALSE, linetype =
        3)  # just for reference
p1 <- p1 + geom_smooth(aes(weight = Total), se = FALSE, colour = "red")  # just for
        reference

# axes
#p1 <- p1 + scale_y_continuous(breaks = seq(-10, 10, by = 0.5))
#p1 <- p1 + scale_x_continuous(breaks = seq(0, 100, by = 2))
# labels
p1 <- p1 + labs(
    title    = "Logit of alcohol dependency by log2 Drinks per day"
  , subtitle= "Observed data for drinkers only"
  , x        = "log2 Drinks per day during period of heaviest drinking"
  , y        = "Empirical logit of the probability of alcohol dependency"
  , caption = paste(
              "Black = Observed logit proportions of dependency given log2 drinks
        per day"
            , "Blue = Naive LM fit of logit proportions"
            , "Red = Loess smooth curve of empirical logits"
            , sep = "\n" # separate by new lines
            )
  )
print(p1)
```

```
`geom_smooth()` using formula 'y ~ x'

`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```





6.21.5 (1 p) Describe the logit-vs-x plot. Is it linear? If not, consider a transformation of x to improve linearity; describe the transformation you chose if you needed one.

- The original logit-vs-x plot was not linear. However, taking the log2 of drinks per day increased linearity. A straight line describes the data well from 3 to 5.5, but does not fit the data well when x is 6 or greater.

6.21.6 (1 p) Fit the `glm()` model and assess the deviance lack-of-fit test.

6.21.6.1 Fit the model.

```
# For our summarized data (with frequencies and totals for each age)
# The left-hand side of our formula binds two columns together with cbind():
#   the columns are the number of "successes" and "failures".
# For logistic regression with logit link we specify family = binomial,
#   where logit is the default link function for the binomial family.

glm_drinkexp_depend <-
  glm(
    cbind(Success, Total - Success) ~ DrinkQuantity_Drinkers_log2
  , family = binomial
  , data   = dat_drinkexp_depend_sum
  )
```

6.21.6.2 Deviance statistic for lack-of-fit

```
# Test residual deviance for lack-of-fit (if > 0.10, little-to-no lack-of-fit)
glm_drinkexp_depend$deviance
```

```
[1] 32.96572
```

```
glm_drinkexp_depend$df.residual
```

```
[1] 32
```

```
dev_p_val <- 1 - pchisq(glm_drinkexp_depend$deviance,
```

```
          glm_drinkexp_depend$df.residual)
dev_p_val
```

```
[1] 0.4196573
```

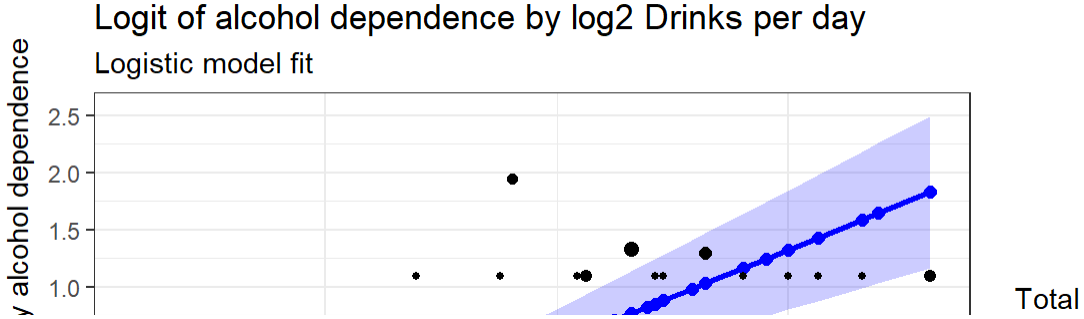
- H_0 : the model fits the data versus H_A : the model does not fit the data.
- D = 33 with
- 32 df, giving
- p-value = 0.42
- Because the p-value is greater than 0.05, we fail to reject H_0, concluding that the model does fit the data.

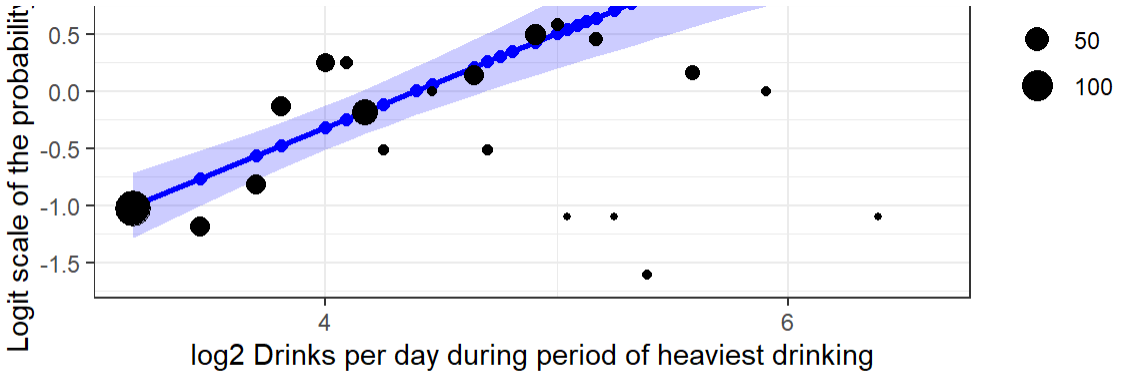
6.21.7 (1 p) Calculate the confidence bands around the model fit/predictions. Plot on both the logit and $\hat{p}$ scales.

```
# predict() uses all the Load values in dataset, including appended values
fit_logit_pred <-
  predict(
    glm_drinkexp_depend
  , data.frame(DrinkQuantity_Drinkers_log2 =
         dat_drinkexp_depend_sum$DrinkQuantity_Drinkers_log2)
  , type   = "link"
  , se.fit = TRUE
  ) %>%
  as_tibble()

# put the fitted values in the data.frame
dat_drinkexp_depend_sum <-
  dat_drinkexp_depend_sum %>%
  mutate(
    # logit scale values
    fit_logit       = fit_logit_pred$fit
  , fit_logit_se    = fit_logit_pred$se.fit
  , fit_logit_lower = fit_logit - 1.96 * fit_logit_se
  , fit_logit_upper = fit_logit + 1.96 * fit_logit_se
  # proportion scale values
  , fit_p           = exp(fit_logit) / (1 + exp(fit_logit))
  , fit_p_lower     = exp(fit_logit_lower) / (1 + exp(fit_logit_lower))
  , fit_p_upper     = exp(fit_logit_upper) / (1 + exp(fit_logit_upper))
  )
```

6.21.7.1 Logit scale

```
# plot on logit scale
library(ggplot2)
p1 <- ggplot(dat_drinkexp_depend_sum, aes(x = DrinkQuantity_Drinkers_log2, y =
        emp_logit))
p1 <- p1 + theme_bw()

# MODEL FIT
# predicted curve and point-wise 95% CI
p1 <- p1 + geom_ribbon(aes(x = DrinkQuantity_Drinkers_log2, ymin = fit_logit_lower,
        ymax = fit_logit_upper), fill = "blue", linetype = 1, alpha = 0.2)
p1 <- p1 + geom_line(aes(x = DrinkQuantity_Drinkers_log2, y = fit_logit), colour =
        "blue", size = 1)
# fitted values
p1 <- p1 + geom_point(aes(y = fit_logit), colour = "blue", size = 2)

# summaries
p1 <- p1 + geom_point(aes(size = Total))

# axes
p1 <- p1 + scale_y_continuous(breaks = seq(-10, 10, by = 0.5))
p1 <- p1 + scale_x_continuous(breaks = seq(0, 100, by = 2))
# labels
p1 <- p1 + labs(
    title   = "Logit of alcohol dependence by log2 Drinks per day"
  , subtitle= "Logistic model fit"
  , x       = "log2 Drinks per day during period of heaviest drinking"
  , y       = "Logit scale of the probability alcohol dependence"
  , caption = paste(
                "Black = Observed logit proportions of alcohol dependence given
        log2 drinks per day"
             , "Blue = Logistic model fitted logit proportions"
             , sep = "\n" # separate by new lines
             )
  )
print(p1)
```





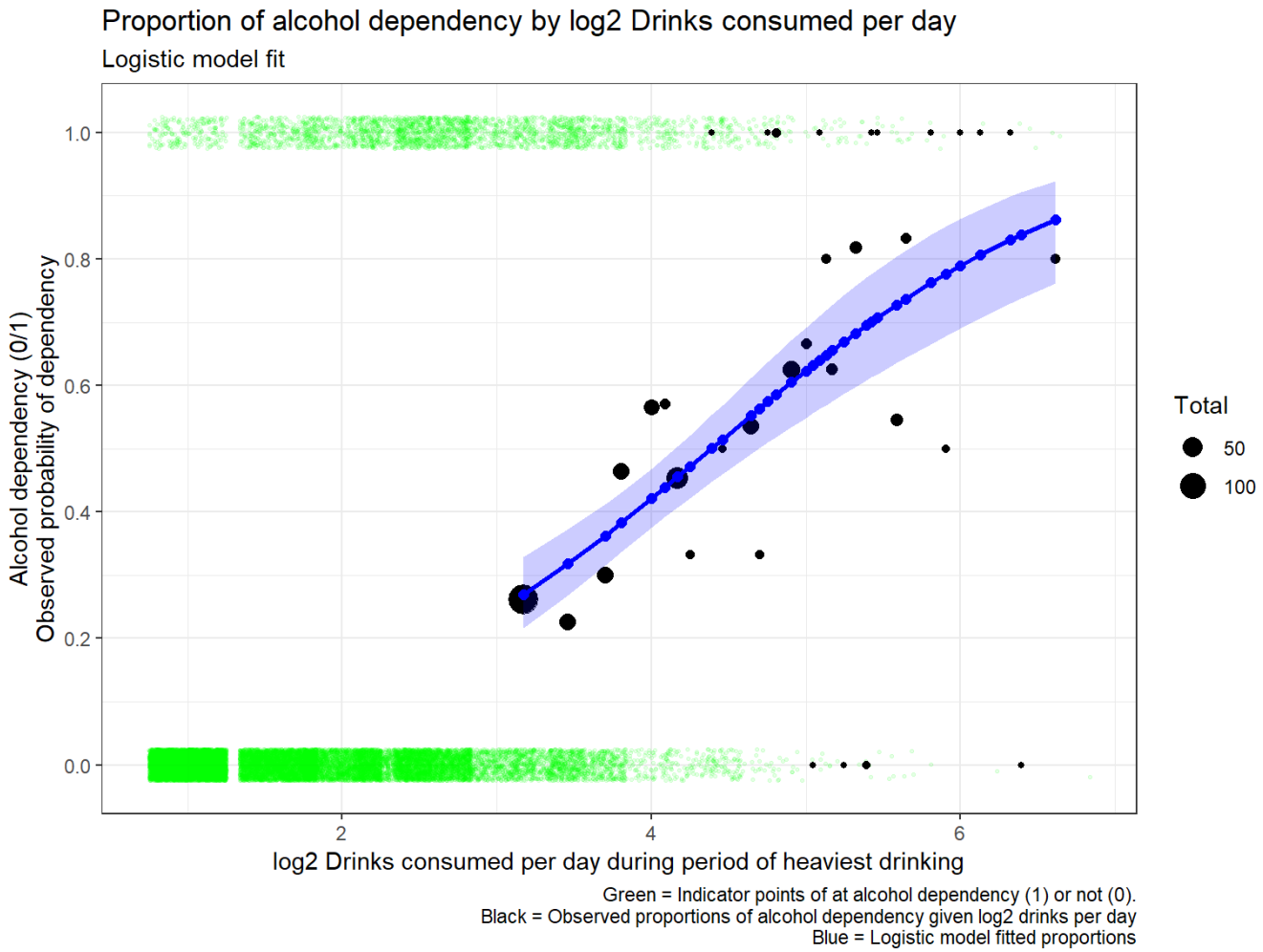
Black = Observed logit proportions of alcohol dependence given log2 drinks per day
Blue = Logistic model fitted logit proportions

6.21.7.2 Probability/proportion scale

```
# plot on probability scale
library(ggplot2)
p1 <- ggplot(nesarc_sub, aes(x = DrinkQuantity_Drinkers_log2, y =
        DrinkDependence01))
p1 <- p1 + theme_bw()
# data
p1 <- p1 + geom_jitter(width = 0.25, height = 0.025, size = 0.5, alpha = 1/10,
        colour = "green")
# summaries
p1 <- p1 + geom_point(data = dat_drinkexp_depend_sum, aes(x =
        DrinkQuantity_Drinkers_log2, y = p_hat, size = Total))

# MODEL FIT
# predicted curve and point-wise 95% CI
p1 <- p1 + geom_ribbon(data = dat_drinkexp_depend_sum, aes(x =
        DrinkQuantity_Drinkers_log2, y = fit_p, ymin = fit_p_lower, ymax =
        fit_p_upper), fill = "blue", linetype = 1, alpha = 0.2)
p1 <- p1 + geom_line(data = dat_drinkexp_depend_sum, aes(x =
        DrinkQuantity_Drinkers_log2, y = fit_p), colour = "blue", size = 1)
# fitted values
p1 <- p1 + geom_point(data = dat_drinkexp_depend_sum, aes(y = fit_p), colour =
        "blue", size = 2)

# axes
p1 <- p1 + scale_y_continuous(breaks = seq(0, 1, by = 0.2))
p1 <- p1 + expand_limits(y = c(0, 1))
p1 <- p1 + scale_x_continuous(breaks = seq(0, 100, by = 2))
# labels
p1 <- p1 + labs(
    title    = "Proportion of alcohol dependency by log2 Drinks consumed per day"
  , subtitle= "Logistic model fit"
  , x        = "log2 Drinks consumed per day during period of heaviest drinking"
  , y        = "Alcohol dependency (0/1)\nObserved probability of dependency"
```

```
  , caption = paste(
                "Green = Indicator points of at alcohol dependency (1) or not (0)."
              , "Black = Observed proportions of alcohol dependency given log2
        drinks per day"
              , "Blue = Logistic model fitted proportions"
              , sep = "\n" # separate by new lines
              )
  )
print(p1)
```

```
Warning: Removed 20894 rows containing missing values (geom_point).
```



6.21.8 (1 p) Interpret the sign ($+$ or $-$) of the slope parameter and test whether the slope is different from zero, $H_A : \beta_1 \neq 0$.

The summary table gives MLEs and standard errors for the regression parameters. The z-value column is the parameter estimate divided by its standard error. The p-values are used to test whether the corresponding parameters of the logistic model are zero.

```
summary(glm_drinkexp_depend)
```

```
Call:
glm(formula = cbind(Success, Total - Success) ~ DrinkQuantity_Drinkers_log2,
    family = binomial, data = dat_drinkexp_depend_sum)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1801  -0.6628   0.2276   0.8115   1.7929

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.6030     0.5173  -6.965 3.29e-12 ***
DrinkQuantity_Drinkers_log2  0.8211     0.1258   6.527 6.72e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80.989  on 33  degrees of freedom
Residual deviance: 32.966  on 32  degrees of freedom
AIC: 93.005

Number of Fisher Scoring iterations: 3
```

```
# see names(summary(glm_drinkexp_depend)) to find the object that has the
      coefficients.
# can also use coef(glm_drinkexp_depend)
```

6.21.8.1 Interpret the sign ($+$ or $-$) of the slope

- The positive slope indicates that as drinks per day increases, the probability of alcohol dependence increases.

6.21.8.2 Hypothesis test

- $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$
- $\beta_1 = 0.82$
- Since the p-value = $6.72 \times 10^{-11} < 0.05$, we must reject H_0 and conclude that the slope is not equal to 0.

7 Poster

7 Poster

7.1 Classes 24, 25, and 26: Poster Preparation

7.1.1 Class 24, Poster Preparation, research questions, data sources, analyses

See items under Class 26.

From the list in Class 26, complete Items 2, 3, 4, and 5.

7.1.2 Class 25, Poster Preparation, literature review, references, discussion, future work

See items under Class 26.

From the list in Class 26, complete Items 1, 6, 7, and 8.

Citation help is available at this page: https://quarto.org/docs/authoring/footnotes-and-citations.html

7.1.3 Class 26, Poster Preparation, complete content

Rubric

Organize the content of your poster.

Complete the content for each of these sections:

Title: A short title that reveals the main result of the poster.

1. **(Class 25) (3 p) Introduction**
    - (Lit Review) 2-4 bullets describing the study, previous research.
2. **Research Questions**

- Is the population mean of heavy drink consumption less for people with manic symptoms than for those without manic symptoms?
- Is there an association between manic symptoms and ethnicity?

3. **Methods**

- Data source
    - An extensive battery of questions addressing present and past alcohol consumption,

alcohol use disorders (AUDs), and questions that operationalized the criteria set forth in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM–IV) for five mood disorders, including hypomania.

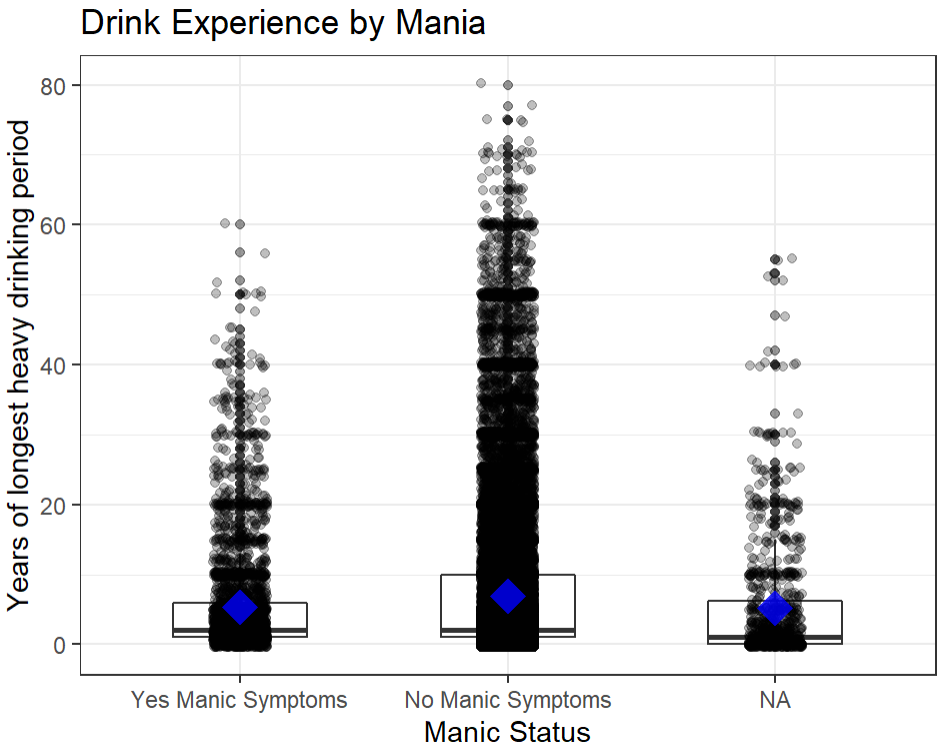
  - We focus on drinkers.
  - The study's oversampling of Blacks and Hispanics as well as the inclusion of Hawaii and Alaska in its sampling frame yielded enough minority respondents to make NESARC an ideal vehicle for addressing the critical issue of race and/or ethnic disparities in comorbidity and access to health care services.

- Variables
  - Duration (years) of period of heaviest drinking (`DrinkExperience` = consecutive years of drinking during heaviest period)
  - Had 1+ week period irritable/easily annoyed that caused you to shout/break things/start fights or arguments (`Mania` , 0=No, 1=Yes)
  - Imputed race/ethnicity (`Ethnicity` , 1=White, Not Hispanic or Latino, 2=Black, Not Hispanic or Latino, 3=American Indian/Alaska Native, Not Hispanic or Latino, 4=Asian/Native Hawaiian/Pacific Islander, Not Hispanic or Latino, 5=Hispanic or Latino)
- Statistical methods used to answer the research questions
  - A two-sample t-test comparing duration of period of heaviest drinking by manic symptoms.
  - A χ^2 analysis of a two-way categorical analysis of manic symptoms by ethnicity.

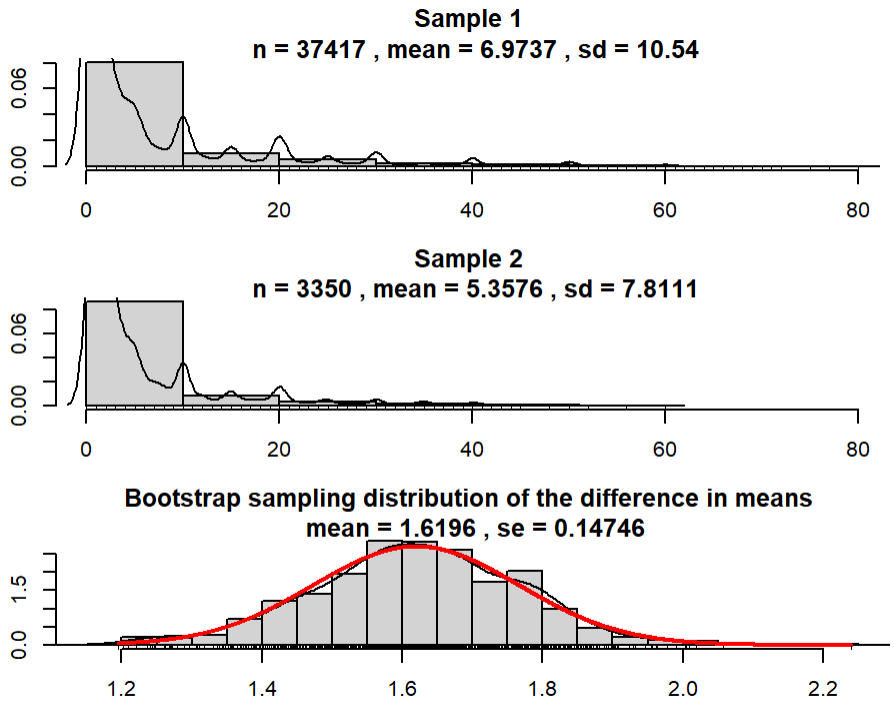
4. **Research Question 1**

- Is the population mean of heavy drink consumption (`DrinkExperience`) less for people with manic symptoms than for those without manic symptoms (`Mania`)? That is, $H_0 : \mu_{YM} = \mu_{NM}$ versus $H_A : \mu_{YM} < \mu_{NM}$
- In Figure Figure 3 (a), the distribution of years of longest heavy drinking period is heavily right skewed, and the mean of those without mania is higher (7) than that of those with mania (5.4).
- In Figure Figure 3 (b), the bootstrap sampling distribution of the difference in means is close to normal (bottom panel), so the model assumptions are met and we can rely on the results.

- Our one-sided two-sample t-test resulted in a large test statistic ($t_s = -11.1$). Thus, based on our p-value ($p = 1.39 \times 10^{-28}$) < 0.05, we reject H_0 in favor of H_A and conclude that the population mean total of drink consumption is higher among those without manic symptoms than those with manic symptoms.



(a) Samples are right skewed



(b) Model assumptions: sampling distribution of the mean is normal

Figure 3: Total drink consumption is higher among those without mania than those with mania

5. **Research Question 2**

- Is there is an association between manic symptoms (`Mania`) and `Ethnicity` ? That is, $H_0 : p(i \text{ and } j) = p(i)p(j)$ for all row categories i and column categories j versus $H_A : p(i \text{ and } j) \neq p(i)p(j)$, for at least one row category i and column category j.
- In Figure 4, along the Yes Manic symptoms row, we see that the Caucasian group is slightly higher than expected, while the African American, Hispanic, and Asian groups are all slightly lower than expected. The Native American group has a very high positive Pearson residual, however, and will likely be the primary cause for rejecting H_0.
- The expected count for each cell (minimum 55) is greater than 5, so model assumptions have been met.
- Our two-way categorical analysis yielded a large test statistic ($X^2 = 97.57$). Thus, based on our p-value (3.24×10^{-20}) < 0.05, we must reject H_0 in favor of H_A and conclude that there is an association between Ethnicity and manic symptoms.

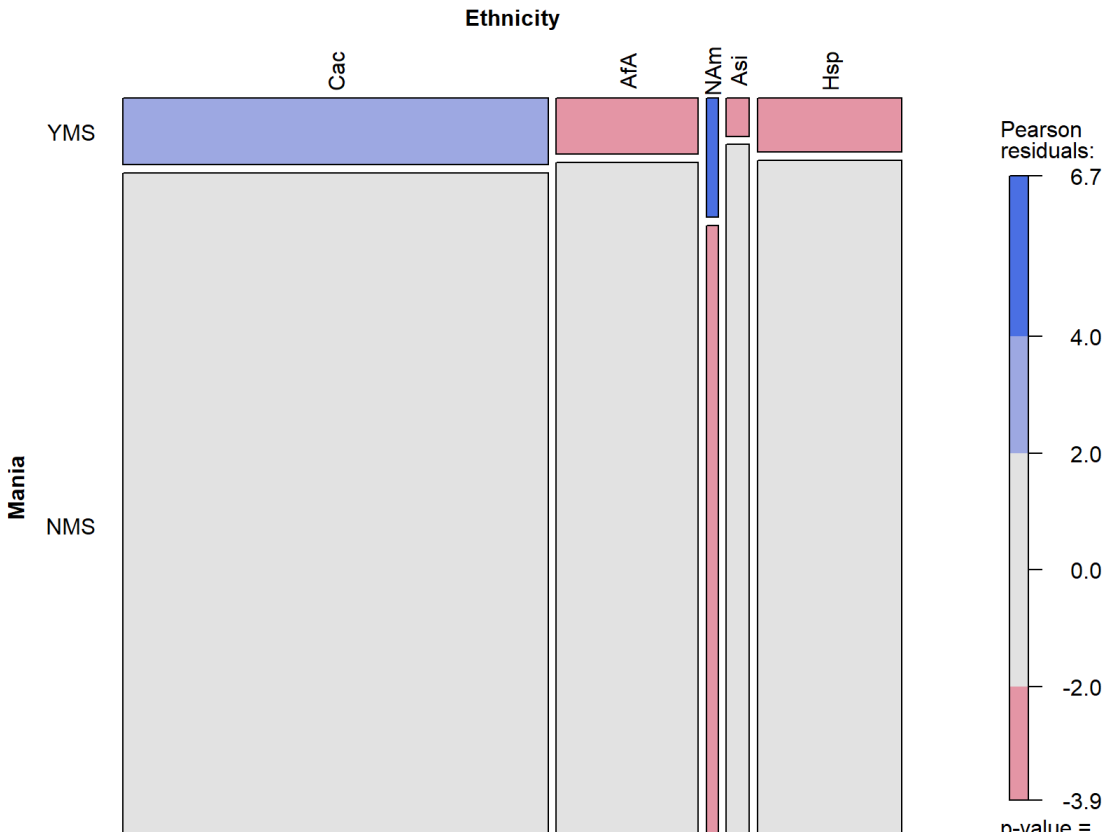


Figure 4: Association between Ethnicity and Mania

6. **(Class 25) (4 p) Discussion**
    - Put the results you found for each research question in the context you provided in your introduction.
7. **(Class 25) (1 p) Further directions or Future work or Next steps or something else that indicates there more to do and you've thought about it.**
    - What do these results lead you to want to investigate?
8. **(Class 25) (2 p) References**
    - By citing sources in your introduction, this section will automatically have your bibliography.

References

References are supposed to appear here, but they may appear at the end of the document.

7.2 Class 27, Poster Preparation, into poster template

- https://github.com/brentthorne/posterdown

7.3 Class 28, Poster Preparation, reviewed by instructor

7.4 Class 29, Poster Presentations

- Graduate students.

7.5 Class 30, Poster Presentations

- Undergraduate students.

[End]

References