Impact of Alcohol Frequency on Headache: An Exploration of NHEFS Data

Wenjuan Bian

2023-06-20

Introduction:

Headaches, often underestimated as minor discomforts, are in fact a pressing global health issue. The World Health Organization asserts that nearly half of all adults worldwide have experienced a headache at least once in the past year (Stovner et al., 2020). The ramifications of these pervasive conditions extend beyond transient discomfort. Chronic headaches pose severe disruptions to daily life, leading to productivity losses and a decline in overall quality of life (Buse et al., 2019).

From an economic standpoint, headaches impose a considerable burden, encompassing not only the direct medical costs tied to physician consultations, medications, and potential hospital stays (Linde et al., 2012), but also the indirect costs arising from lost work days (Munakata et al., 2009). Additionally, the psychological toll inflicted on chronic headache sufferers is significant, with many individuals grappling with associated mental health challenges such as depression and anxiety, thereby escalating the overall health impact of this prevalent ailment (Minen et al., 2016).

Given these considerable impacts, investigations into potential correlations between headache frequency and lifestyle factors, such as alcohol consumption, hold great promise. Research in this area, as suggested in this study, could illuminate new preventative measures and contribute valuable insights to future public health strategies and individual lifestyle recommendations for those prone to headaches (Chen et al., 2023)."

Despite the broad impacts, the correlation between headaches and lifestyle factors, including alcohol consumption, remains partially understood. The project aims to explore the potential relationship between the frequency of alcohol consumption and the occurrence of headaches. A subset of the data from the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS) is used, and statistical methods such as standardization with a parametric outcome model and potential outcomes with single and multiple predictors are employed.

Description of Data:

The data used for my final project is from a subset of the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS). This comprehensive health and lifestyle survey was initiated by the National Center for Health Statistics and the National Institute on Aging, and it has collected invaluable information from adults and children in the U.S.

For my project, several variables from this subset were used. The subset was originally used in the book "Causal Inference: What If" (Hernán and Robins, 2020) to investigate the impact of smoking cessation on weight gain. Although this subset was initially used for studying smoking cessation and weight gain, it also contains crucial data for my project. I am specifically focusing on the variables related to the self-reported frequency of alcohol consumption ('alcoholfreq'), the incidence of headaches ('headache'), other pain symptoms reported by the subjects ('otherpain'), their income levels ('income'), age ('age'), race ('race'), and education levels ('education').

By employing a generalized linear model with a binomial family, my objective is to estimate the causal effect of the frequency of alcohol consumption on the occurrence of headaches, while controlling for the effects of other pain, income, age, race, and education. This project serves as an opportunity to apply the principles and methodologies of causal inference that I have learned throughout the course, allowing me to explore these relationships in a multivariate context.

The NHEFS dataset is publicly accessible, which encourages additional research to uncover the intricate relationships between lifestyle habits, socioeconomic factors, demographic characteristics, and health outcomes.

Data Analysis:

(a) *Variables Chosen from The Dataset*

The key variables chosen for this study include the occurrence of headaches (headache), frequency of alcohol consumption (alcoholfreq), other pain conditions (otherpain), income levels (income), age, race, and education level. Table 1 shows the first 6 records of the dataset with the selected variables from NHEFS.

Table 1 First 6 records of the selected variables from NHEFS

ID#	HEADACHE	OTHERPAIN	INCOME	AGE	ALCOHOLFREQ	RACE	EDUCATION
1	1	0	19	42	1	1	1
2	1	0	18	36	0	0	2
3	1	1	15	56	3	1	2
4	0	1	15	68	2	1	1
5	1	0	18	40	2	0	2
6	1	0	11	43	3	1	2

Table 2 shows the transformed dataset by the original dataset. The alcohol frequency, age, education, and income variables have been recoded to binary representation for further analysis.

Table 2 The dataset after transformations

ID#	HEADACHE	OTHERPAIN	INCOME	AGE	ALCOHOLFREQ	RACE	EDUCATION
1	1	0	0	1	0	1	0
2	1	0	0	0	0	0	0
3	1	1	1	1	1	1	0
4	0	1	1	1	1	1	0
5	1	0	0	1	1	0	0
6	1	0	1	1	1	1	0

Headache: This binary variable indicates whether the subject used headache medication (1: Yes, 0: No).

Alcoholfreq: This variable represented the frequency of alcohol consumption in 1971. It was transformed into a binary variable, coded as 0 for those who drank less than 2 times a week, and 1 for those who drank 2 or more times a week.

Otherpain: This binary variable indicates whether the subject used other pain medication (1: Yes, 0: No).

Income: For this study, the income variable was transformed into a binary variable. Incomes less than $5000 (original categories 11 to 15) were coded as 1, while incomes of $5000 and above (original categories 16 to 22) were coded as 0.

Age: The age variable represented the subject's age in 1971. This was recoded into a binary variable, with 1 representing ages 40 and above, and 0 representing ages below 40.

Race: This binary variable indicates the subject's race (0: White, 1: Black or Other).

Education: The education variable represented the amount of education achieved by 1971. It was transformed into a binary variable, coded as 1 for those who attended college or had more education, and 0 for those who had less than a college education.

(b) *Methods From the Book*

The methods used in this project are based on those presented in Chapters 3 and 6. The analysis primarily employs logistic regression modeling, which was chosen because of the binary nature of our outcome variable (headache: yes/no). To accommodate this, all other variables were transformed into binary form. The unadjusted estimates and outcome-model standardization were utilized to estimate potential outcomes and effect measures through Bootstrapping. The prop.r was used to check whether there is any missing arrow by propensity and prognostic score.

(c) *Analysis Results*

1. Visualization of the selected data

Figure 1 shows the distribution of the variable 'alcoholfreq'. The chart reveals that approximately 21% of participants reported consuming alcohol "Almost every day", and 14% claimed to do so "2-3 times per week". Around 31% reported consumption "1-4 times per month", while 34% claimed to consume alcohol fewer than 12 times per year or not at all. A few participants were recorded as 'unknown', but the number in this category is negligible, close to zero percent. Therefore, in the modeling process, those reported as 'unknown' were excluded.

The proportion of participants reporting headaches in each category of 'alcoholfreq' is shown in Figure 2. The chart indicates that the percentage of individuals reporting headaches increases as the frequency of alcohol consumption decreases, except for those who did not consume alcohol in the last year. Specifically, 54% of those drinking almost daily reported headaches, rising to 57% for those drinking 2-3 times per week, 64% for those drinking 1-4 times per month, and peaking at 72% for those drinking less than 12 times per year. Interestingly, the percentage drops slightly to 69% for those who did not consume alcohol in the last year.

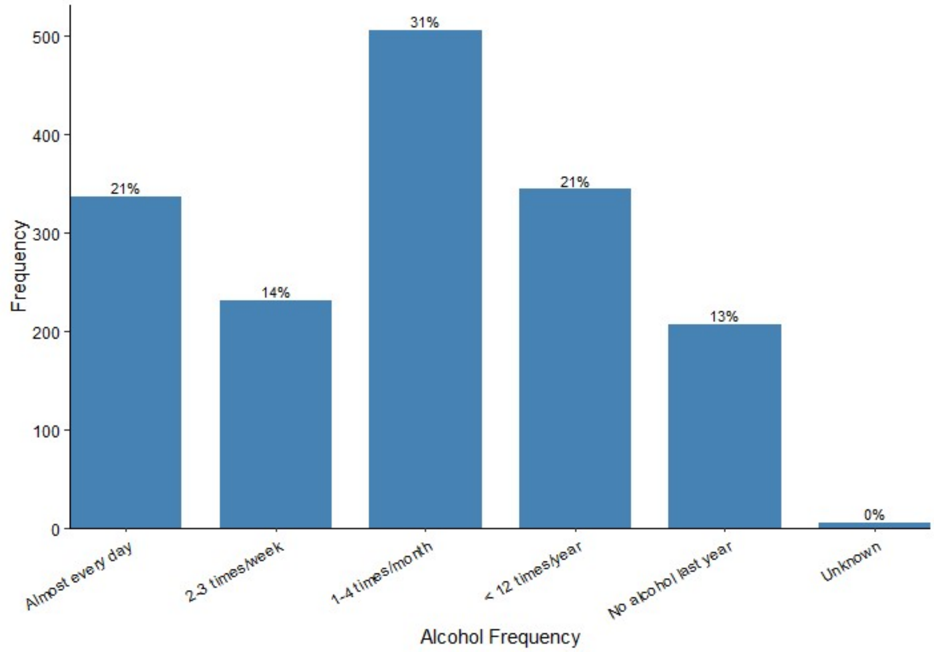

Figure 1 The distribution of the variable Alcohol Frequency

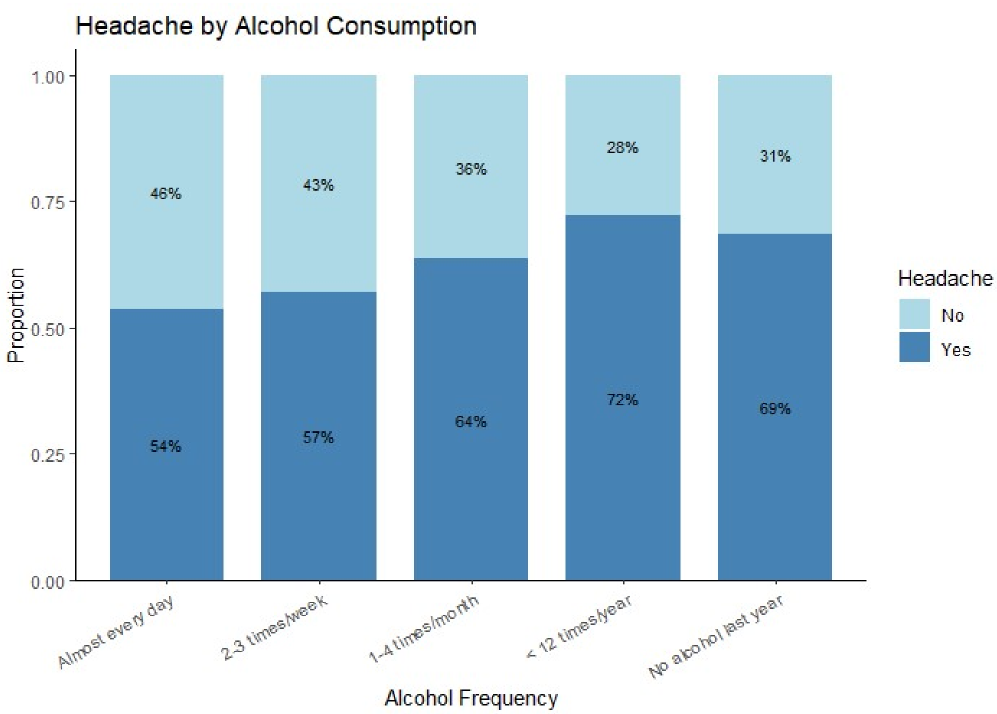

Figure 2 Proportion of reporting headaches in each category of 'alcoholfreq'

2. Potential outcomes and effect measures by unadjusted estimates and outcome-model standardization

A bootstrap method was used to estimate the potential effect of alcohol frequency (alcoholfreq) on the occurrence of headaches, with alcoholfreq as the only variable. The output is shown in Table 3. The estimated probability of experiencing a headache for individuals who frequently consume alcohol is 0.550 (95% CI: 0.509 to 0.591). In contrast, those who do not frequently consume alcohol have a higher probability of experiencing a headache, estimated to be 0.674 (95% CI: 0.644 to 0.703). The confidence intervals of four association measures exclude the null association, which is 0 for the RD and 1 for the other three measures. These results, demonstrating a potential inverse relationship between frequent alcohol consumption and the likelihood of experiencing headaches, align with previous findings reported by Yokoyama et al. (2009), which also indicated a similar association.

Table 3 The effect meansures relating more frequent consumption of alcohol

MEASURE	ESTIMATE	95% CONFIDENCE INTERVAL
P1HAT	0.55	(0.509, 0.591)
P0HAT	0.674	(0.644, 0.703)
RDHAT	-0.123	(-0.175, -0.072)
RRHAT	0.817	(0.748, 0.892)
RRSTARHAT	0.726	(0.637, 0.827)
ORHAT	0.593	(0.477, 0.737)

Table 4 shows the R output for logistic regression analysis. By the estimated standard errors and p-values returned by glm, all of the coefficients are statistically significant. The coefficient for frequent alcohol consumption (alcoholfreq) was statistically significant ($p < 0.001$), indicating a significant impact on the likelihood of experiencing a headache. The presence of other pain (otherpain), age, race, and education also had statistically significant coefficients ($p < 0.05$), indicating their significant contributions to the likelihood of experiencing a headache. The coefficient for income had a marginally significant p-value ($p = 0.0568$), suggesting a potential association that warrants further investigation. According to this model, individuals who drink alcohol less frequently, use other pain medications, have lower income, are younger, are of "White" race, and have some college education are more likely to use headache medication.

A more comprehensive approach of the bootstrap method was taken to estimate the sampling distributions for $p1 = E(Headache| alcoholfreq =1, H=1)$, and $p0 = E(Headache| alcoholfreq =0, H=1)$ by including additional covariates such as other types of pain (otherpain), income, age, race, and education, which means H=(otherpain, income, age, race, education). The analysis results are shown in Table 5.

With the inclusion of other covariates, the estimated probability of experiencing a headache for individuals who frequently consume alcohol is 0.634 (95% CI: 0.524 to 0.745), and for those who do not frequently consume alcohol is 0.745 (95% CI: 0.655 to 0.836).

Comparing Table 3 and Table 5, we can see that the risk difference, relative risk, and odds ratio suggest a decreased likelihood of headaches with higher alcohol frequency. However, when accounting for other variables as shown in Table 5, the estimated probabilities for headaches, relative risk, and odds ratio all changed slightly, suggesting potential confounding effects from these covariates. This demonstrates the importance of considering other covariates when assessing the relationship between alcohol frequency and the occurrence of headaches.

Table 4 The effect meansures relating more frequent consumption of alcohol

```
logistic_model <- glm(headache ~ alcoholfreq +otherpain + income + age + rac
e + education, data = f_nhefs, family = "binomial")
# Print the summary of the model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = headache ~ alcoholfreq + otherpain + income + age +
##     race + education, family = "binomial", data = f_nhefs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9967  -1.3277   0.8043   0.9990   1.4858
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8738     0.1103   7.923 2.32e-15 ***
## alcoholfreq  -0.5242     0.1136  -4.615 3.93e-06 ***
## otherpain     0.6130     0.1333   4.599 4.25e-06 ***
## income        0.2782     0.1461   1.905  0.05680 .
## age          -0.5272     0.1135  -4.644 3.41e-06 ***
## race         -0.5233     0.1610  -3.249  0.00116 **
## education     0.3602     0.1456   2.475  0.01333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Table 5 Estimates of Sampling distributions and four effect measures

MEASURE	ESTIMATE	95% CONFIDENCE INTERVAL
P1HAT	0.634	(0.524, 0.745)
P0HAT	0.745	(0.655, 0.836)
RDHAT	-0.111	(-0.163, -0.060)
RRHAT	0.851	(0.778, 0.931)
RRSTARHAT	0.696	(0.591, 0.819)
ORHAT	0.592	(0.472, 0.742)

The method of outcome-model standardization was employed to adjust for confounding in the causal analysis. This method involves identifying a set of covariates that blocks all backdoor paths from the exposure to the outcome (in this case, the headache). Through standardization, these covariates were adjusted, and a logistic model was used to simulate potential outcomes under different scenarios. The simulated outcomes were then compared to estimate the causal effect. The standardized estimates are presented in Table 6. These estimates are almost identical to the unadjusted estimates, as reported in Table 3. These similarities suggest that either the arrow from H (otherpain, income, age, race, education) to T (alcoholgreq) or the arrow from H to Y (Headache) is missing. To investigate this, the statistical significance of the coefficients of H in the parametric exposure model was checked by propensity score, via prop.r. The results the statistical significance of the variables 'income' and 'education' in predicting T ('alcoholfreq'), with p-values less than 0.05. Other variables ('otherpain', 'age', 'race') are not statistically significant as their p-values exceed 0.05. Therefore, the arrow from H to T is missing. However, for the prognostic score, via prog.r, the coefficients of the model for E(Y|H) are statistically significant. The variables 'otherpain', 'income', 'age', and 'race' are statistically significant with p-values less than 0.05, while 'education' shows a trend towards significance with a p-value slightly over 0.05. Thus, the arrow from H to Y is not missing.

Table 6 Outcome-model Standardization of effect of alcolholfreq on Headache

MEASURE	ESTIMATE	95% CONFIDENCE INTERVAL
EY0	0.675	(0.646, 0.703)
EY1	0.556	(0.513, 0.599)
RD	-0.118	(-0.170, -0.067)
RR	0.824	(0.752, 0.897)
OR	0.604	(0.470, 0.738)
EY0	0.675	(0.646, 0.703)

Conclusion:

In conclusion, the exploration of the subset of National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study (NHEFS) data yielded intriguing insights into the potential relationship between alcohol frequency and headache occurrence. Findings indicate an inverse relationship between alcohol frequency and headaches, with more frequent alcohol consumers less likely to experience headaches, as evidenced across all methods.

Through standardization with a parametric outcome model (Method 1, Table 6), and potential outcomes with a single predictor (Method 3, Table 3), similar estimates were obtained, suggesting the robustness of this inverse relationship. The average probability of headache occurrence for those who frequently consume alcohol was estimated at 0.556 and 0.55, respectively, compared to 0.675 in both cases for less frequent consumers. When more covariates were incorporated in the potential outcomes estimates and effect measures, we again found evidence of this inverse relationship, but with a slightly higher headache occurrence probability (0.634) for frequent alcohol consumers, compared to infrequent consumers (0.745).

However, correlation doesn't mean causation, and there may be other influencing factors not examined in this project. More comprehensive research is needed to fully understand this relationship and reveal potential causal mechanisms. Further investigations could consider variables such as alcohol type, quantity, or individuals' genetic predispositions.

References:

Buse, D. C., Manack, A., Serrano, D., Turkel, C., & Lipton, R. B. (2019). Sociodemographic and comorbidity profiles of chronic migraine and episodic migraine sufferers. Journal of Neurology, Neurosurgery & Psychiatry, 90(6), 644-651.

Linde, M., Gustavsson, A., Stovner, L. J., Steiner, T. J., Barre, J., Katsarava, Z., ... Andrée, C. (2012). The cost of headache disorders in Europe: the Eurolight project. European Journal of Neurology, 19(5), 703-711.

Minen, M. T., Begasse De Dhaem, O., Kroon Van Diest, A., Powers, S., Schwedt, T. J., Lipton, R., & Silbersweig, D. (2016). Migraine and its psychiatric comorbidities. Journal of Neurology, Neurosurgery, and Psychiatry, 87(7), 741-749.

Munakata, J., Hazard, E., Serrano, D., Klingman, D., Rupnow, M. F., Tierce, J., Reed, M., & Lipton, R. B. (2009). Economic burden of transformed migraine: results from the American Migraine Prevalence and Prevention (AMPP) Study. Headache: The Journal of Head and Face Pain, 49(4), 498-508.

Stovner, L. J., Nichols, E., Steiner, T. J., Abd-Allah, F., Abdelalim, A., Al-Raddadi, R. M., ... Collaborators, G. B. D. H. (2020). Global, regional, and national burden of migraine and tension-type headache, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology, 19(4), 391-401.

Yokoyama, M., Yokoyama, T., Funazu, K., et al. (2009). Associations between headache and stress, alcohol drinking, exercise, sleep, and comorbid health conditions in a Japanese population. Journal of Headache Pain, 10(3), 177-185.

Hernán, M. A., & Robins, J. M. (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

R code for the project:

```
library(boot)
library(AER)
library(car)
library(ggplot2)
library(dplyr)
library(tidyverse)

nhefs <- read.csv("C:/Users/fwjbi/OneDrive -
Bowling Green State University/Summer
2023/6820-Causal/Final Project/nhefs.csv")

f_nhefs <- nhefs[,
c("headache","otherpain" ,"income", "age",
"alcoholfreq", "race", "education")]
head(f_nhefs)
f_nhefs <- f_nhefs[f_nhefs$alcoholfreq != 5, ]
f_nhefs$alcoholfreq <- ifelse(f_nhefs$alcoholfreq <
2, 1, 0)
f_nhefs$age <- ifelse(f_nhefs$age < 40, 0, 1)
f_nhefs$education <- ifelse(f_nhefs$education >=4,
1, 0)
f_nhefs$income <- ifelse(f_nhefs$income < 16, 1, 0)
head(f_nhefs)

logistic_model <- glm(headache ~ alcoholfreq
+otherpain + income + age + race + education, data
= f_nhefs, family = "binomial")
summary(logistic_model)

lmodboot.r <- function() {
  estimator <- function(data, ids) {
    dat <- data[ids, ]
    coef <- glm (headache ~ alcoholfreq, family =
binomial, data = dat)$coef

    xbeta1 <- sum(coef)
    xbeta0 <- sum(coef)-(coef[2])

    p1 <- exp (xbeta1) / (1+exp (xbeta1) )
    p0 <- exp (xbeta0) / (1+exp (xbeta0) )
    rd <- p1 - p0
    logrr <- log(p1) - log(p0)
    logrrstar <- log(1-p0) - log(1-p1)
    logor <- log(p1/(1-p1)) - log(p0/(1-p0))

    c(p1, p0, rd, logrr, logor, logrrstar)
  }

  boot.out <- boot(data = f_nhefs, statistic =
estimator, R = 1000)

  beta0hat <- summary(boot.out)$original[7]
  beta1hat <- summary(boot.out)$original[8]

  p0hat <- summary(boot.out)$original[2]
  p0lci <- p0hat - 1.96 *
summary(boot.out)$bootSE[2]
  p0uci <- p0hat + 1.96 *
summary(boot.out)$bootSE[2]

  p1hat <- summary(boot.out)$original[1]
  p1lci <- p1hat - 1.96 *
summary(boot.out)$bootSE[1]
  p1uci <- p1hat + 1.96 *
summary(boot.out)$bootSE[1]

  rdhat <- summary(boot.out)$original[3]
  logrrhat <- summary(boot.out)$original[4]
  logorhat <- summary(boot.out)$original[5]
  logrrstarhat <- summary(boot.out)$original[6]

  rdhatlci <- rdhat - 1.96 *
summary(boot.out)$bootSE[3]
  rdhatuci <- rdhat + 1.96 *
summary(boot.out)$bootSE[3]
  logrrhatlci <- logrrhat - 1.96 *
summary(boot.out)$bootSE[4]
  logrrhatuci <- logrrhat + 1.96 *
summary(boot.out)$bootSE[4]

  logrrstarlci <- logrrstarhat - 1.96 *
summary(boot.out)$bootSE[6]
  logrrstaruci <- logrrstarhat + 1.96 *
summary(boot.out)$bootSE[6]

  logorhatlci <- logorhat - 1.96 *
summary(boot.out)$bootSE[5]
  logorhatuci <- logorhat + 1.96 *
summary(boot.out)$bootSE[5]

  rrhat <- exp(logrrhat)
  rrstarhat <- exp(logrrstarhat)
```

```
  orhat <- exp(logorhat)

  p0ci <- c(p0lci, p0uci)
  p1ci <- c(p1lci, p1uci)
  rdci <- c(rdhatlci, rdhatuci)
  rrci <- exp(c(logrrhatlci, logrrhatuci))
  rrstarci <- exp(c(logrrstarlci, logrrstaruci))
  orci <- exp(c(logorhatlci, logorhatuci))

  list(
    p1hat = p1hat,
    p1ci = p1ci,
    p0hat = p0hat,
    p0ci = p0ci,
    rdhat = rdhat,
    rdci = rdci,
    rrhat = rrhat,
    rrci = rrci,
    rrstarhat = rrstarhat,
    rrstarci = rrstarci,
    orhat = orhat,
    orci = orci
  )
}

lmodboot.out <- lmodboot.r()
lmodboot.out

lmodboot.r1 <- function() {
  estimator <- function(data, ids) {
    dat <- data[ids, ]
    coef <- glm (headache ~ alcoholfreq +otherpain +
income + age + race + education, family = binomial,
data = dat)$coef

    xbeta1 <- sum(coef)
    xbeta0 <- sum(coef)-(coef[2])

    p1 <- exp (xbeta1) / (1+exp (xbeta1) )
    p0 <- exp (xbeta0) / (1+exp (xbeta0) )
    rd <- p1 - p0
    logrr <- log(p1) - log(p0)
    logrrstar <- log(1-p0) - log(1-p1)
    logor <- log(p1/(1-p1)) - log(p0/(1-p0))

    c(p1, p0, rd, logrr, logor, logrrstar)
  }

  boot.out <- boot(data = f_nhefs, statistic =
estimator, R = 1000)

  beta0hat <- summary(boot.out)$original[7]
  beta1hat <- summary(boot.out)$original[8]

  p0hat <- summary(boot.out)$original[2]
  p0lci <- p0hat - 1.96 *
summary(boot.out)$bootSE[2]
  p0uci <- p0hat + 1.96 *
summary(boot.out)$bootSE[2]

  p1hat <- summary(boot.out)$original[1]
  p1lci <- p1hat - 1.96 *
summary(boot.out)$bootSE[1]
  p1uci <- p1hat + 1.96 *
summary(boot.out)$bootSE[1]

  rdhat <- summary(boot.out)$original[3]
  logrrhat <- summary(boot.out)$original[4]
  logorhat <- summary(boot.out)$original[5]
  logrrstarhat <- summary(boot.out)$original[6]

  rdhatlci <- rdhat - 1.96 *
summary(boot.out)$bootSE[3]
  rdhatuci <- rdhat + 1.96 *
summary(boot.out)$bootSE[3]
  logrrhatlci <- logrrhat - 1.96 *
summary(boot.out)$bootSE[4]
  logrrhatuci <- logrrhat + 1.96 *
summary(boot.out)$bootSE[4]

  logrrstarlci <- logrrstarhat - 1.96 *
summary(boot.out)$bootSE[6]
  logrrstaruci <- logrrstarhat + 1.96 *
summary(boot.out)$bootSE[6]

  logorhatlci <- logorhat - 1.96 *
summary(boot.out)$bootSE[5]
  logorhatuci <- logorhat + 1.96 *
summary(boot.out)$bootSE[5]

  rrhat <- exp(logrrhat)
  rrstarhat <- exp(logrrstarhat)
  orhat <- exp(logorhat)

  p0ci <- c(p0lci, p0uci)
  p1ci <- c(p1lci, p1uci)
```

```
  rdci <- c(rdhatlci, rdhatuci)
  rrci <- exp(c(logrrhatlci, logrrhatuci))
  rrstarci <- exp(c(logrrstarlci, logrrstaruci))
  orci <- exp(c(logorhatlci, logorhatuci))

  list(
    p1hat = p1hat,
    p1ci = p1ci,
    p0hat = p0hat,
    p0ci = p0ci,
    rdhat = rdhat,
    rdci = rdci,
    rrhat = rrhat,
    rrci = rrci,
    rrstarhat = rrstarhat,
    rrstarci = rrstarci,
    orhat = orhat,
    orci = orci
  )
}

lmodboot.out1 <- lmodboot.r1()
lmodboot.out1

f_nhefs <- f_nhefs %>% drop_na(alcoholfreq,
otherpain, income, age, race, education, headache)


boot_estimator <- function(dat, indices) {
  dat <- dat[indices, ]

  lmod <- glm(headache ~ alcoholfreq + otherpain +
income + age + race + education, family = binomial,
data = dat)

  dat0 <- dat1 <- dat
  dat0$alcoholfreq <- 0
  dat1$alcoholfreq <- 1

  EYhat0 <- predict(lmod, newdata = dat0, type =
"response")
  EYhat1 <- predict(lmod, newdata = dat1, type =
"response")

  EY0 <- mean(EYhat0)
  EY1 <- mean(EYhat1)

  rd <- EY1 - EY0
  rr <- EY1 / EY0
  or <- (EY1 / (1 - EY1)) / (EY0 / (1 - EY0))

  c(EY0, EY1, rd, rr, or)
}

stand.r <- function(data = f_nhefs, ids =
c(1:nrow(f_nhefs))) {
  stand.out <- boot(data = data[ids, ], statistic =
boot_estimator, R = 1000)

  stand.est <- summary(stand.out)$original
  stand.SE <- summary(stand.out)$bootSE

  stand.lci <- stand.est - 1.96 * stand.SE
  stand.uci <- stand.est + 1.96 * stand.SE

  list(stand.est = stand.est, stand.SE = stand.SE,
stand.lci = stand.lci, stand.uci = stand.uci)
}

result2 <- stand.r(f_nhefs)
print(result2)


prop.r <- function (data = f_nhefs)
{
  out <- glm (alcoholfreq ~ otherpain + income + age
+ race + education,
            family = binomial, data = data)
  list (e = fitted(out), emod = summary(out))
}
prop.r()$emod
prop.r <- function (data = f_nhefs)
{
  out <- glm (headache ~ otherpain + income + age +
race + education,
            family = binomial, data = data)
  list (d = fitted(out), dmod = summary(out))
}
prop.r()$dmod


nhefs1 <- nhefs %>%
  group_by(alcoholfreq, headache) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count)) %>%
  ungroup()
```

```
nhefs1$alcoholfreq <- factor(nhefs1$alcoholfreq,
levels = c("0", "1", "2", "3", "4", "5"), labels =
c("Almost every day", "2-3 times/week", "1-4
times/month", "< 12 times/year", "No alcohol last
year", "Unknown"))

levels(nhefs1$alcoholfreq)

ggplot(data = nhefs1, aes(x = alcoholfreq, y = prop,
fill = factor(headache))) +
  geom_bar(position = "fill", stat = "identity", width =
0.75) +
  geom_text(aes(label = scales::percent(prop,
accuracy = 1)), position = position_fill(vjust = 0.5),
size = 3) +
  labs(x = "Alcohol Frequency", y = "Proportion", fill
= "Headache") +
  scale_fill_manual(name = "Headache", values =
c("0" = "lightblue", "1" = "steelblue"), labels =
c("No", "Yes")) +
  ggtitle("Headache by Alcohol Consumption") +
  scale_y_continuous(expand = expansion(mult =
c(0, 0))) +
  coord_cartesian(ylim = c(0, 1.05)) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 30, hjust
= 1))


nhefs$alcoholfreq <- factor(nhefs$alcoholfreq,
levels = c(0, 1, 2, 3, 4, 5), labels = c("Almost every
day", "2-3 times/week", "1-4 times/month", "< 12
times/year", "No alcohol last year", "Unknown"))

nhefs2 <- nhefs %>%
  group_by(alcoholfreq) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))

ggplot(data = nhefs2, aes(x = alcoholfreq, y = count))
+
  geom_bar(stat = "identity", fill = "steelblue", width
= 0.75) +
  geom_text(aes(label = scales::percent(prop,
accuracy = 1)), vjust = -0.3, size = 3) +
  labs(x = "Alcohol Frequency", y = "Frequency") +
  theme_classic() +
  theme(axis.line = element_line(colour = 'black'),
      axis.text = element_text(colour = 'black'),
      axis.title = element_text(colour = 'black'),
      axis.text.x = element_text(angle = 30, hjust = 1))
+
  coord_cartesian(ylim = c(0, max(nhefs2$count)),
expand = FALSE, clip = 'off') +
  scale_y_continuous(expand = expansion(mult =
c(0, 0.1)))


f_nhefs <- f_nhefs %>% drop_na(alcoholfreq,
otherpain, income, age, race, education, headache)
vars <- c("alcoholfreq", "otherpain", "income",
"age", "race", "education", "headache")
f_nhefs[vars] <- lapply(f_nhefs[vars], factor)

nhefs_long <- f_nhefs %>%
  tidyr::pivot_longer(cols = c(alcoholfreq, otherpain,
income, age, race, education), names_to =
"variable", values_to = "value")

nhefs_long <- nhefs_long %>%
  group_by(variable, value, headache) %>%
  summarise(count = n(), .groups = 'drop')

total_counts <- nhefs_long %>%
  group_by(variable) %>%
  summarise(total_count = sum(count), .groups =
'drop')

nhefs_long <- nhefs_long %>%
  left_join(total_counts, by = "variable")

nhefs_long <- nhefs_long %>%
  mutate(prop = count / total_count)

ggplot(nhefs_long, aes(x = variable, y = prop, fill =
interaction(headache, value, sep = "_"))) +
  geom_bar(stat = "identity", position = "stack",
width = 0.75) +
  labs(x = "Variable", y = "Proportion", fill =
"Headache & Value") +
  scale_fill_manual(values = c("0_0" = "lightgray",
"1_0" = "lightblue", "0_1" = "gray", "1_1" =
"steelblue")) +
  scale_y_continuous(expand = expansion(mult =
c(0, 0.1))) +
```

```
  coord_cartesian(ylim = c(0, NA)) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 20, hjust
= 1)) +
  ggtitle("Headache by Various Factors")
```