Lab Assignment 6 BIS 505b

Wenxin Xu

4/25/2021

Contents

Instructions

This Lab Assignment continues to analyze data from the observational study presented in lecture that studies factors related to fractures in women with osteoporosis. You may keep the sections on **Data Background** and the **Data Key** in your submission if you wish. Perform your work in the **Assignment** section below. In this assignment, report any p-values that are less than 0.001 as **<0.001** and round values reported in your narrative text to **3** decimal places. **Be sure to clearly state the reference category when interpreting the effects of categorical variables in any regression model.** Perform all hypothesis testing at the $\alpha = 0.05$-level.

Data Background

In this observational study, female patients were recruited by their primary care physician after receiving a diagnosis of osteoporosis. These women were given the opportunity to enroll in a strength training program [`strength`]. After consent was obtained, baseline data were collected. Data elements collected at the first visit (at diagnosis) included quality of life (scale 0-100) [`qol`], pain assessment (10 point scale) [`pain`], a measure of physical activity [`act`], current calcium use [`cal`], age [`age`], and race [`race`]. Data on the number of healthcare utilizations (HCUs) [`hcu`] (emergency room, urgent treatment center, and hospital visits) were collected by telephone interview every 6 months. Medical records were accessed to verify information collected in the telephone interviews. Follow-up time for each participant is recorded [`period`]. A CSV file [`hcu.csv`] is provided which contains data from the women in the study.

Data Key – `hcu.csv`

Variable Name	Definition
`qol`	Quality of life (QoL) index (higher: better QoL)
`cal`	Calcium use at initial visit 0 = No (reference) 1 = Yes
`race`	Race

Variable Name	Definition
	1 = White (reference)
	2 = Black
	3 = Other
`strength`	Participation in strength training program
	0 = No (reference)
	1 = Yes
`act`	Activity level at initial visit
	1 = None (reference)
	2 = Limited/Moderate
	3 = Rigorous
`age`	Age at initial visit (years)
`pain`	Pain score at initial visit (higher: greater pain)
`period`	Number of years participating in study
`hcu`	Number of new health care utilizations reported

Assignment

1. [5 points] Import the CSV file `hcu.csv` in the third code chunk above. Name your data frame `hcu` and create the factor variables `cal_factor` (reference = "No"), `race_factor` (reference = "White"), `strength_factor` (reference = "No") and `act_factor` (reference = "None"). After these steps, `hcu` should contain 13 variables. [**Note:** When creating factor variables, **do not** use the `ordered=TRUE` option to create ordinal variables. No written response is required for this question. Display the code chunk(s) that perform the requested data management steps.]

```r
# create factor variables
hcu <- mutate(hcu,
              cal_factor = factor(cal,
                                  levels = c(0, 1),
                                  labels = c("No", "Yes")),
              race_factor = factor(race,
                                   levels = c(1, 2, 3),
                                   labels = c("White", "Black", "Other")),
              strength_factor = factor(strength,
                                       levels = c(0, 1),
                                       labels = c("No", "Yes")),
              act_factor = factor(act,
                                  levels = c(1, 2, 3),
                                  labels = c("None", "Limited", "Rigorous")))
```

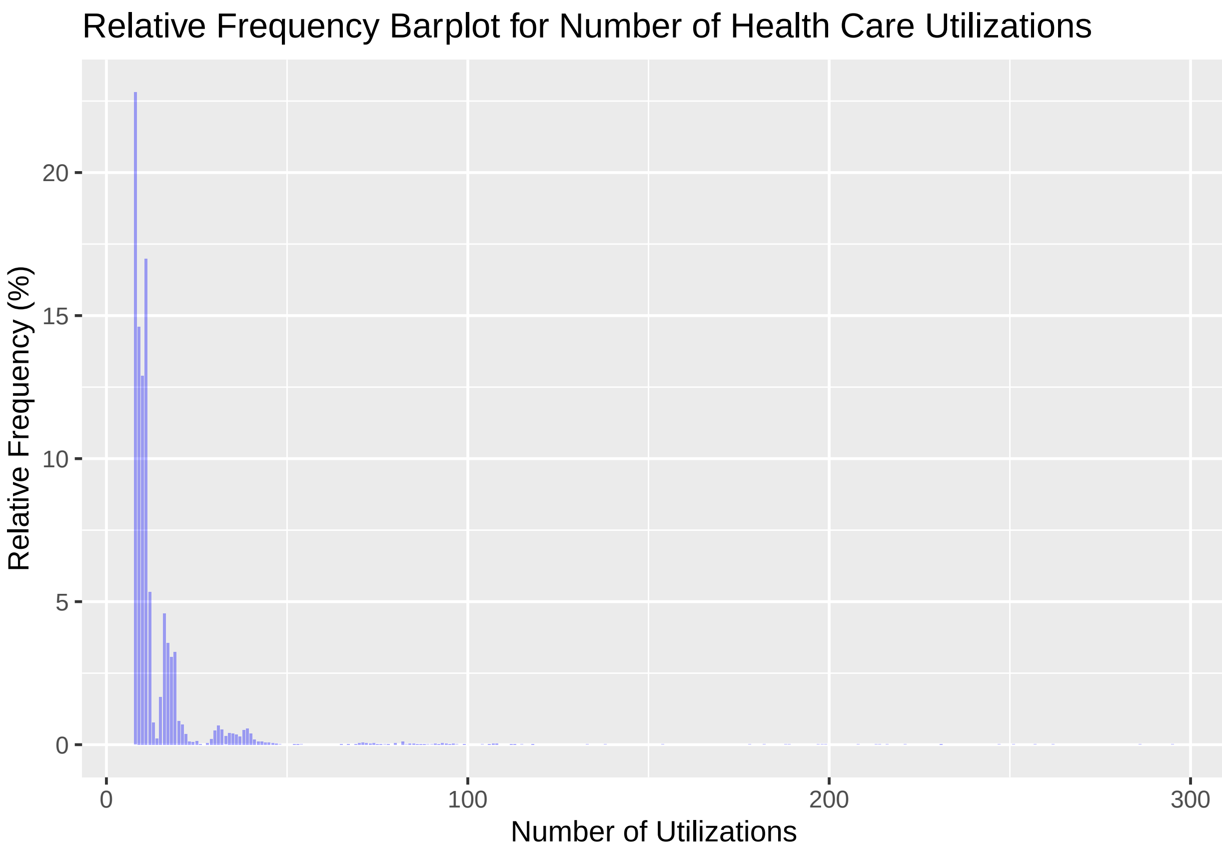
```r
# check # of variables
ncol(hcu)
```

```
## [1] 13
```

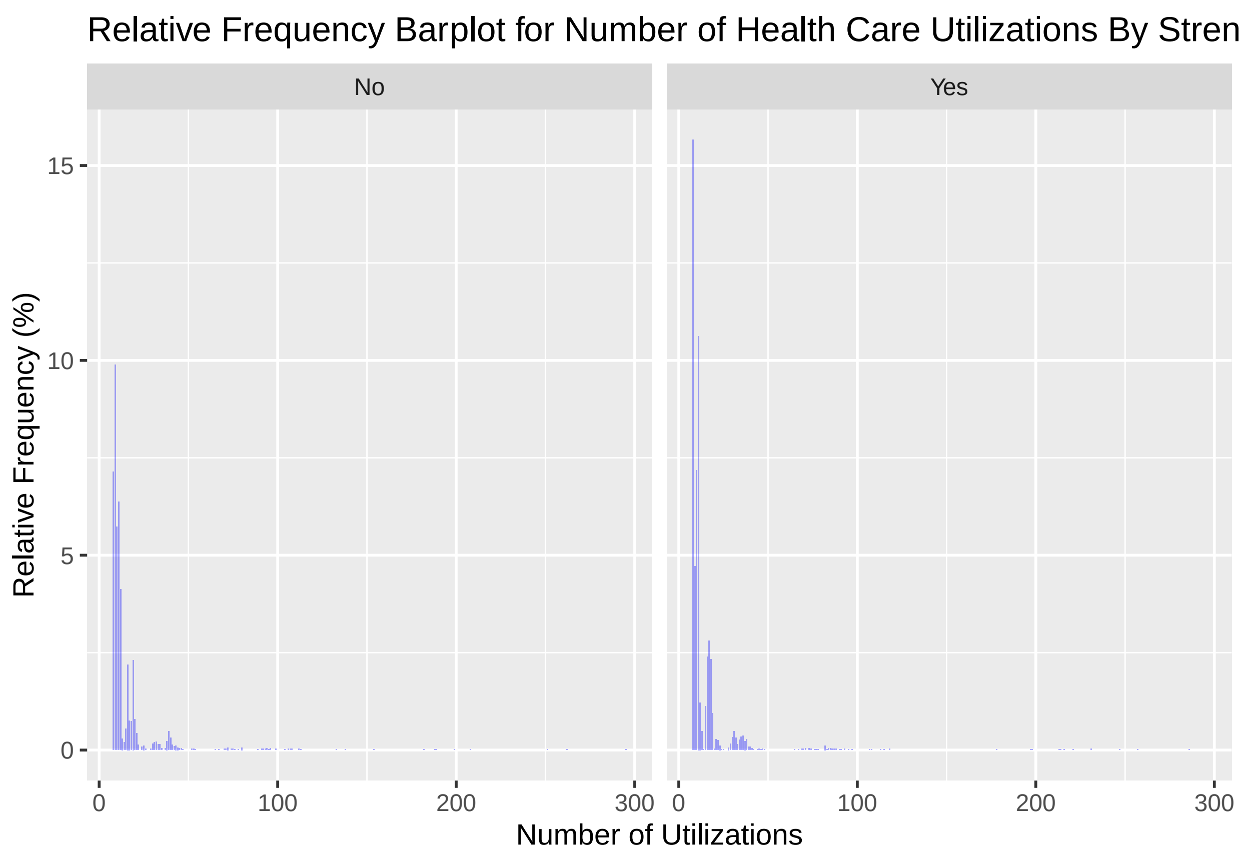
2. The **research question** of this study is to determine if *healthcare utilization* is related to participation in the *strength training program*. We will begin our analysis with some descriptive statistics and graphical summaries.

2a. [5 points] Provide a graphical summary of the number of healthcare utilizations observed per patient in this study (`hcu`) and the number of years individuals participated in the study (`period`) for both the full sample and by levels of the strength training variable. Use a relative frequency barplot for `hcu` and a relative frequency histogram for `period`. Comment on what you see in the plots (overall and comparing the two groups).

```r
# relative frequency barplot for hcu for the full sample
ggplot(data = hcu,
       aes(x = hcu, y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Barplot for Number of Health Care Utilizations",
       x = "Number of Utilizations",
       y = "Relative Frequency (%)")
```

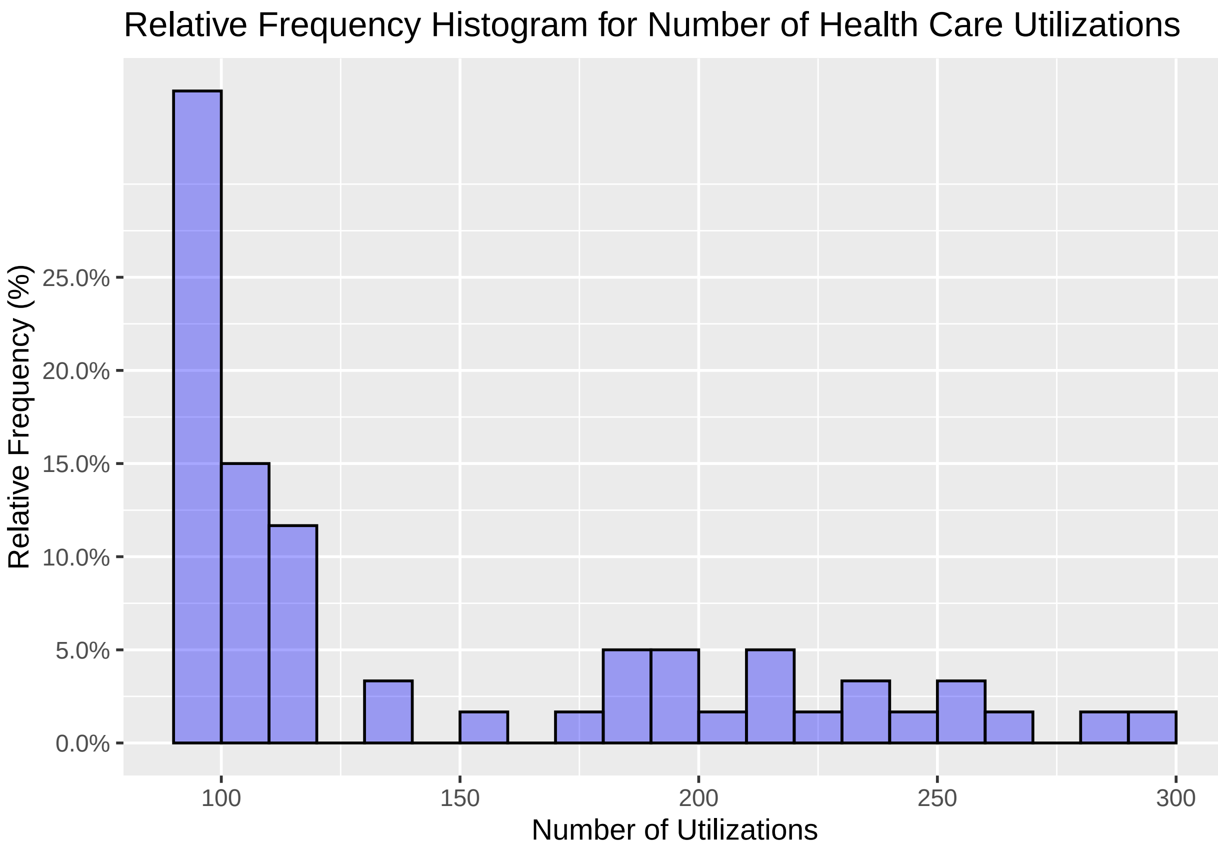



```r
# Vertical relative frequency barplot for hcu by levels of the strength training variable
ggplot(data = hcu,
       aes(x = hcu, y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Barplot for Number of Health Care Utilizations By Strength Training"
       x = "Number of Utilizations",
       y = "Relative Frequency (%)")+
  facet_wrap(~ strength_factor, nrow = 1)
```



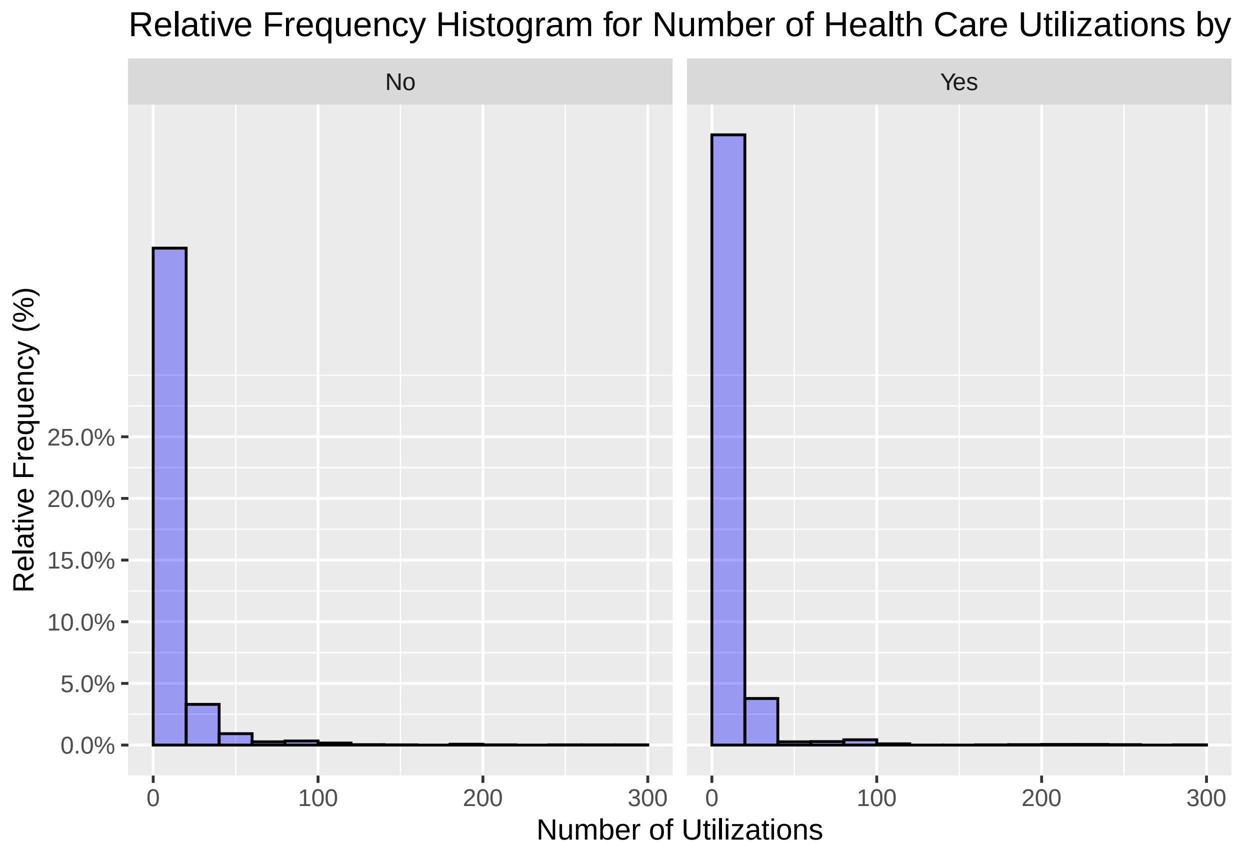

```
# relative frequency histogram for hcu for the full sample
ggplot(data = hcu, aes(x = hcu)) +
  geom_histogram(aes(y = stat(count)/sum(stat(count))),
                 breaks = seq(90,300,by=10),
                 col = "black", fill = "blue", alpha = 0.35,
                 closed = "left", na.rm = TRUE) +

  scale_y_continuous(name = "Relative Frequency (%)",
                     labels = scales::percent_format(),
                     breaks = seq(0, 0.25, by=0.05)) +
  ggtitle("Relative Frequency Histogram for Number of Health Care Utilizations") +
  xlab("Number of Utilizations")
```




```r
# relative frequency histogram for hcu by levels of the strength training variable
ggplot(data = hcu, aes(x = hcu)) +
  geom_histogram(aes(y = stat(count)/sum(stat(count))),
                 breaks = seq(0,300,by=20),
                 col = "black", fill = "blue", alpha = 0.35,
                 closed = "left", na.rm = TRUE) +

  scale_y_continuous(name = "Relative Frequency (%)",
                     labels = scales::percent_format(),
                     breaks = seq(0, 0.25, by=0.05)) +
  ggtitle("Relative Frequency Histogram for Number of Health Care Utilizations by Strength Training") +
  xlab("Number of Utilizations")+
  facet_grid(~ strength_factor)
```



Comment on what you see in the plots (overall and comparing the two groups). The overall pattern of counts of health care utilizations are heavily right skewed. There are a few individuals with a large observed number of utilizations. Between strength training groups, those who participated in strength training program had a **lower** counts of health care utilizations than those who not.

2b. [10 points] Report the percentage of participants in each group that had 8, 9, 10-14, 15-19, 20-49, and 50+ HCUs. Round your percentages to 1 decimal place. Also create a relative frequency barplot that graphically displays this information. Describe any differences that you see.

```
# create factor variable for hcu
hcu$hcu2[hcu$hcu == 8] <- 0
hcu$hcu2[hcu$hcu == 9] <- 1
hcu$hcu2[(hcu$hcu >= 10) && (hcu$hcu <=14)] <- 2
hcu$hcu2[(hcu$hcu >= 15) && (hcu$hcu <=19)] <- 3
hcu$hcu2[(hcu$hcu >= 20) && (hcu$hcu <=49)] <- 4
hcu$hcu2[(hcu$hcu >= 50)] <- 5

hcu <- mutate(hcu,
              hcu_factor = factor(hcu2,
                                  levels = c(0,1,2,3,4,5),
                                  labels = c("8", "9", "10-14","15-19","20-49","50+")))
```

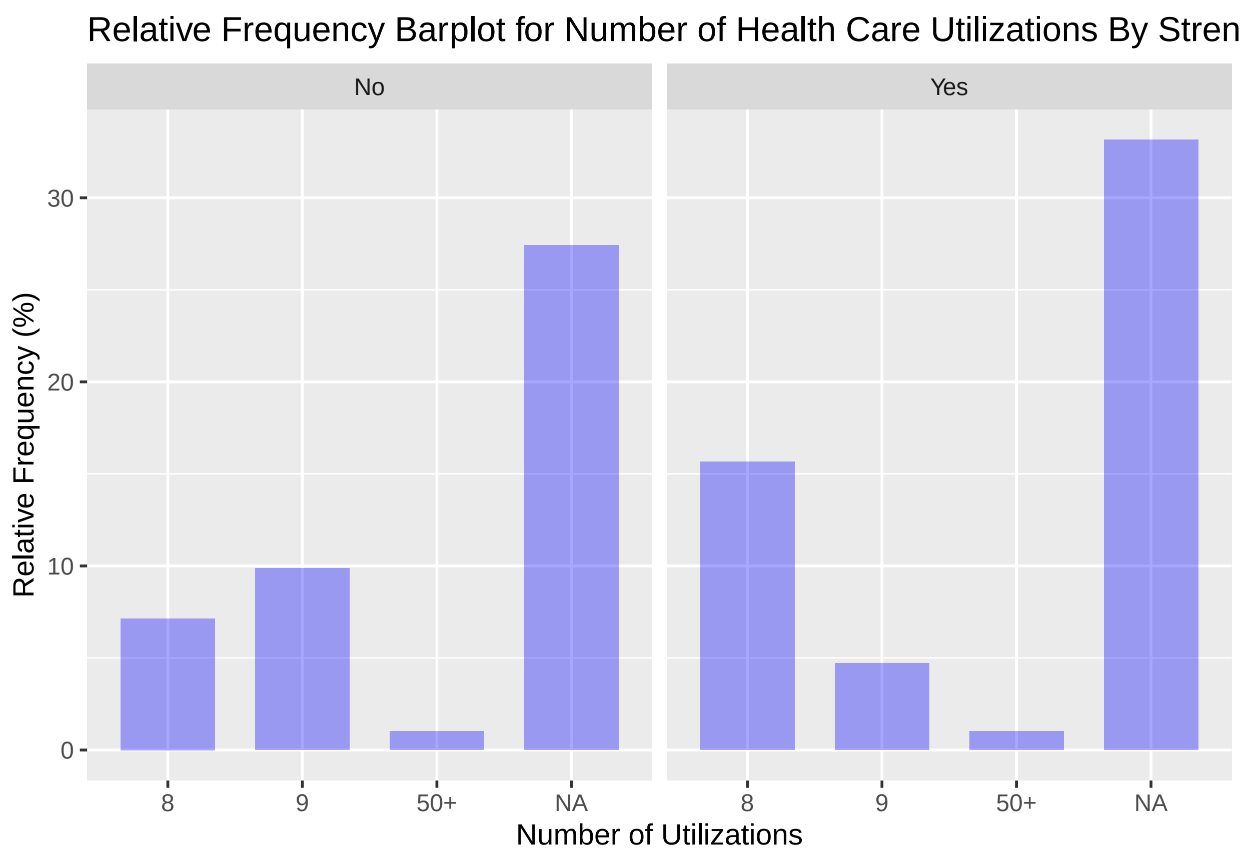
```
tab <- table(hcu$hcu_factor, hcu$strength_factor, useNA = "ifany")

relfreq <- round(100*prop.table(tab, margin=2), 1)

relfreq
```

```
##
##          No  Yes
##   8      15.7 28.7
##   9      21.7  8.7
##   10-14   0.0  0.0
##   15-19   0.0  0.0
##   20-49   0.0  0.0
##   50+     2.3  1.8
##   <NA>   60.3 60.8
```

```r
# Vertical relative frequency barplot for hcu by levels of the strength training variable
# remove NA: data = remove_missing(hcu, na.rm = TRUE)
ggplot(data = hcu,
       aes(x = hcu_factor, y = 100*(stat(count))/sum(stat(count)))) +
  geom_bar(fill = "blue", width = 0.7, alpha = 0.35) +
  labs(title = "Relative Frequency Barplot for Number of Health Care Utilizations By Strength Training"
       x = "Number of Utilizations",
       y = "Relative Frequency (%)")+
  facet_wrap(~ strength_factor, nrow = 1)
```



For participants in training program, 28.7 had 8 HCUs, 8.7 had 9 HCUs, 1.8 had 50+ HCUs. For those who don't participant in training program, 15.7 had 8 HCUs, 21.7 had 9 HCUs, 2.3 had 50+ HCUs. For both group, none of them had 10-49 HCUs. There are also many NA for both groups, 60.8 for participants in training program, 60.3 for those who don't parcitipate.

2c. [10 points] Use the `tableby()` function in the `arsenal` package (syntax in **Lab 1**) to create a single summary table of the number of HCUs observed per patient in this study and the number of years individuals participated in the study for the full group (overall) and by levels of the strength training variable. Report the mean (SD) and the median (range) in your table to 1 decimal place. Based on the results in the table, comment on any differences in the two groups. Next, compute the mean ratio of HCUs in those who participated in strength training vs. those who did not participate and interpret the mean ratio.

```r
# specify statistics: mean, sd, median, range
my_controls <- tableby.control(
  test = F,
  total = T,
  numeric.stats = c("meansd","medianrange"),
  stats.labels = list(
    meansd = "Mean (SD)",
    medianrange = "Median (Range)"
  ),
  digits = 1
)

# label variables
my_labels <- list(
  hcu = "Number of Utilizations",
  period = "Period (years)",
  strength_factor = "Participate or Not"
)

table <- tableby(strength_factor ~ hcu + period,
                 data = hcu,
                 control = my_controls)

kable(summary(table,
        labelTranslations = my_labels,
        title = "Summary Statistics of HCUs and Period",
        term.name = TRUE))
```

Participate or Not	No (N=2755)	Yes (N=3305)	Total (N=6060)
Number of Utilizations			
Mean (SD)	14.9 (17.2)	14.3 (16.9)	14.6 (17.0)
Median (Range)	10.0 (8.0, 295.0)	10.0 (8.0, 286.0)	10.0 (8.0, 295.0)
Period (years)			
Mean (SD)	7.0 (1.2)	7.1 (1.2)	7.0 (1.2)
Median (Range)	7.0 (3.0, 11.2)	7.1 (2.7, 11.1)	7.0 (2.7, 11.2)

Those who not participate in the strength training program have a slightly larger average number of health care utilizations compared to those who participate (14.9 vs. 14.3) while the mean are the same (10). The mean and median of participating period in study is similar (mean: 7 vs. 7.1; median: mean: 7 vs. 7.1).

```r
# compute mean ratio of HCUs
mean(hcu$hcu[which(hcu$strength_factor=="Yes")], na.rm=TRUE) / mean(hcu$hcu[which(hcu$strength_factor==
```

```
## [1] 0.960306
```

Mean ratio of HCUs in those who participated in strength training vs. those who did not participate is 0.96.

2d. [6 points] Compute the healthcare utilization rate in the overall sample and by levels of the strength training variable. Also compute the HCU rate ratio in those who participated in strength training vs. those who did not participate and interpret the rate ratio.

```r
# compute healthcare utilization rate
bygroup.sum <- aggregate(x = list(hcu = hcu$hcu, period = hcu$period),
                         by = list(group = hcu$strength_factor),
                         FUN = sum,
                         na.rm = TRUE)

bygroup.rate <- cbind(bygroup.sum, rate = bygroup.sum[,2]/bygroup.sum[,3])

bygroup.rate
```

```
##   group   hcu   period     rate
## 1    No 41166 19231.00 2.140606
## 2   Yes 47424 23304.63 2.034961
```

Participators in the strength training program have a health care utilization rate of $\hat{\lambda}_1 = 2.035$ times/year, while those who don't participate have a have a health care utilization rate of $\hat{\lambda}_0 = 2.141$ times/year.

```r
# HCU rate ratio (participate vs. not)
lamhat1 = bygroup.rate[2,4]

lamhat0 = bygroup.rate[1,4]

rateratio = lamhat1 / lamhat0

rateratio
```

```
## [1] 0.9506471
```

The HCU rate ratio of in those who participated in strength training vs. those who did not participate is 0.951, indicating that those participate in strength training program have 0.951 times the rate of healthcare utilizations compared to those who not.

3. [10 points] **Model 1:** Fit a simple Poisson regression model of the healthcare utilization rate using participation in the strength training program. Assume the reference level specified in Question **1**. Report the equation of the fitted Poisson regression model. Interpret the estimated intercept. Report and interpret the unadjusted rate ratio associated with the strength training variable, report its 95% confidence interval and perform a hypothesis test to determine if there is a significant association between the healthcare utilization rate and participation in the strength training program. (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

Fit a simple Poisson regression model of the healthcare utilization rate using participation in the strength training program.

```r
mod.rate1 <- glm(hcu ~ strength_factor + offset(log(period)),
                 data = hcu,
                 family = poisson(link = "log"))

summary(mod.rate1)
```

```
## 
## Call:
## glm(formula = hcu ~ strength_factor + offset(log(period)), family = poisson(link = "log"),
##     data = hcu)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
```

```
## -3.488  -1.712  -1.062   0.084  36.172
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.761089   0.004929 154.420  < 2e-16
## strength_factorYes -0.050612   0.006736  -7.513 5.76e-14
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 54429  on 6059  degrees of freedom
## Residual deviance: 54372  on 6058  degrees of freedom
## AIC: 80690
##
## Number of Fisher Scoring iterations: 5
```

- Report the equation of the fitted Poisson regression model.

The fitted model is $\log(\hat{\lambda}) = 0.761 - 0.051$ Strength.

- Interpret the estimated intercept.

The estimated intercept $a = 0.761$ is equal to log-rate of healthcare utilizations in the reference group (non-participators). The exponentiated intercept $e^a = 2.141$ times/year is equal to the yearly rate of healthcare utilizations in those don't participate the strength training program.

- Report and interpret the unadjusted rate ratio associated with the strength training variable, report its 95% confidence interval.

```r
# rate ratio and 95% CI
cbind(bj = coef(mod.rate1),
      RR = exp(coef(mod.rate1)),
      exp(confint.default(mod.rate1)))
```

```
##                              bj        RR     2.5 %    97.5 %
## (Intercept)         0.76108888 2.1406058 2.1200270 2.1613843
## strength_factorYes -0.05061238 0.9506471 0.9381782 0.9632817
```

The unadjusted rate ratio is given by the exponentiated slope e^b, $\hat{RR} = e^b = 0.951$ [95% CI (0.938, 0.963)], indicating that participators in strength training program had 0.951 times the rate of healthcare utilizations of those who don't participate.

- Perform a hypothesis test to determine if there is a significant association between the healthcare utilization rate and participation in the strength training program.

(i) State the null and alternative hypotheses;

$H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic is -7.513, p-value $<.001$.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that the rate of healthcare utilizations is significantly different in those who participate in the strength training program and those who not.

4. [7 points] The **research question** would like to determine if there is an association between the *healthcare utilization rate* and participation in the *strength training program*. Given the description of the study in the introduction of this assignment, do you believe it is important to control for the other variables that were collected in these subjects (e.g., quality of life index, calcium use, race, activity level, age, and pain score) when assessing the impact of strength training program on the HCU rate? Explain. Using the `tableby()`

function, create a table that summarizes these baseline variables (quality of life index, calcium use, race, activity level, age, and pain score) by participation in the strength training program. Report mean (SD) and median (range) for quantitative variables and count (%) for categorical variables to 1 decimal place. Comment on any differences that you observe.

I believe it is important to control for the other variables when assessing the impact of strength training program on the HCU rate because these variables (e.g., quality of life index, calcium use, race, activity level, age, and pain score) are the potential confounders in the study so that the results may not reflect the actual association without controlling for it.

```r
# specify statistics: mean, sd, median, range
my_controls <- tableby.control(
  test = F,
  total = T,
  numeric.stats = c("meansd","medianrange"),
  cat.stats = c("countrowpct"),
  stats.labels = list(
    meansd = "Mean (SD)",
    medianrange = "Median (Range)",
    countrowpct = "Count (%)"
  ),
  digits = 1
)

# label variables
my_labels <- list(
  qol = "Quality of Life Index",
  cal_factor = "Calcium Use",
  race_factor = "Race",
  act_factor = "Activity Level",
  age = "Age (years)",
  pain = "Pain Score"
)



table <- tableby(strength_factor ~ qol + cal_factor + race_factor + act_factor + age + pain,
                 data = hcu,
                 control = my_controls)

kable(summary(table,
        labelTranslations = my_labels,
        title = "Summary Statistics of Baseline Variables",
        term.name = TRUE))
```

strength_factor	No (N=2755)	Yes (N=3305)	Total (N=6060)
Quality of Life Index			
Mean (SD)	44.5 (15.4)	44.8 (14.9)	44.7 (15.2)
Median (Range)	45.0 (0.0, 97.0)	45.0 (0.0, 96.0)	45.0 (0.0, 97.0)
Calcium Use			
No	691 (44.9%)	849 (55.1%)	1540 (100.0%)
Yes	2064 (45.7%)	2456 (54.3%)	4520 (100.0%)
Race			
White	1804 (45.2%)	2184 (54.8%)	3988 (100.0%)
Black	542 (45.2%)	657 (54.8%)	1199 (100.0%)

strength_factor	No (N=2755)	Yes (N=3305)	Total (N=6060)
Other	409 (46.8%)	464 (53.2%)	873 (100.0%)
Activity Level			
None	1112 (51.5%)	1047 (48.5%)	2159 (100.0%)
Limited	1527 (46.0%)	1794 (54.0%)	3321 (100.0%)
Rigorous	116 (20.0%)	464 (80.0%)	580 (100.0%)
Age (years)			
Mean (SD)	60.7 (9.2)	56.9 (9.7)	58.6 (9.7)
Median (Range)	61.0 (39.0, 85.0)	56.0 (31.0, 85.0)	58.0 (31.0, 85.0)
Pain Score			
Mean (SD)	5.9 (2.2)	5.5 (2.7)	5.7 (2.5)
Median (Range)	6.0 (0.0, 10.0)	6.0 (0.0, 10.0)	6.0 (0.0, 10.0)

Except from average quality of life index is similar between participation group, average or percentage value of other baseline variables is different between participation group.

5. [5 points] **Model 2:** Extend Model 1 to control for an individual's quality of life index, calcium use, race, activity level, age, and pain score at baseline. Categorical variables should use the reference levels specified in Question **1**. Using Model 2's residual deviance and residual degrees of freedom, assess if overdispersion is a problem. [Note: No interpretation of the fitted model is required]

```r
mod.rate2 <- glm(hcu ~ strength_factor + qol + cal_factor + race_factor + act_factor + age + pain + off
                 data = hcu,
                 family = poisson(link = "log"))

summary(mod.rate2)
```

```
##
## Call:
## glm(formula = hcu ~ strength_factor + qol + cal_factor + race_factor +
##     act_factor + age + pain + offset(log(period)), family = poisson(link = "log"),
##     data = hcu)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.805  -1.709  -1.060   0.101  35.625
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.7606323  0.0267962  28.386  < 2e-16
## strength_factorYes  -0.0332172  0.0069899  -4.752 2.01e-06
## qol                 -0.0007861  0.0002214  -3.551 0.000383
## cal_factorYes       -0.0888436  0.0075636 -11.746  < 2e-16
## race_factorBlack     0.0019338  0.0086087   0.225 0.822263
## race_factorOther    -0.0232858  0.0098534  -2.363 0.018117
## act_factorLimited   -0.0382573  0.0071882  -5.322 1.03e-07
## act_factorRigorous  -0.1059297  0.0127554  -8.305  < 2e-16
## age                  0.0022578  0.0003548   6.364 1.96e-10
## pain                -0.0012560  0.0013400  -0.937 0.348588
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 54423  on 6056  degrees of freedom
## Residual deviance: 54093  on 6047  degrees of freedom
```

```
##   (3 observations deleted due to missingness)
## AIC: 80415
##
## Number of Fisher Scoring iterations: 5
```

- The **fitted model** is $\log(\hat{\lambda}) = 0.761 - 0.033$ Strength $- 0.001$ Quality of Life Index $- 0.089$ Calcium Use $+ 0.002$ Black $- 0.023$ Other $- 0.038$ Limited $- 0.106$ Rigorous $+ 0.002$ Age $- 0.001$ Pain.
- Check overdispersion

```r
# check overdispersion
deviance(mod.rate2)/mod.rate2$df.residual
```

```
## [1] 8.945488
```

In the multiple Poisson regression model, the residual deviance is equal to 54093 and the residual degrees of freedom is equal to 6047. Their ratio, 8.945 is much larger than 1, indicating that overdispersion is a problem in these data.

6. Model 3: Re-fit Model 2 using a negative binomial regression model.

```r
mod.NBrate <- glm.nb(hcu ~ strength_factor + qol + cal_factor + race_factor + act_factor + age + pain +
                 data = hcu)

summary(mod.NBrate)
```

```
##
## Call:
## glm.nb(formula = hcu ~ strength_factor + qol + cal_factor + race_factor +
##     act_factor + age + pain + offset(log(period)), data = hcu,
##     init.theta = 2.769510399, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4945  -0.7549  -0.4851  -0.0032   9.7875
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.7559215  0.0671791  11.252  < 2e-16
## strength_factorYes -0.0255047  0.0175596  -1.452  0.14637
## qol                -0.0006272  0.0005555  -1.129  0.25885
## cal_factorYes      -0.1024330  0.0192780  -5.313 1.08e-07
## race_factorBlack   -0.0038565  0.0216004  -0.179  0.85830
## race_factorOther   -0.0123338  0.0245267  -0.503  0.61505
## act_factorLimited  -0.0430480  0.0181342  -2.374  0.01760
## act_factorRigorous -0.1023122  0.0312466  -3.274  0.00106
## age                 0.0028192  0.0008891   3.171  0.00152
## pain               -0.0014838  0.0033541  -0.442  0.65820
##
## (Dispersion parameter for Negative Binomial(2.7695) family taken to be 1)
##
##     Null deviance: 6087.4  on 6056  degrees of freedom
## Residual deviance: 6027.9  on 6047  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 42895
##
## Number of Fisher Scoring iterations: 1
##
```

```
##
##              Theta:  2.7695
##          Std. Err.:  0.0548
##
##  2 x log-likelihood:  -42872.5610
```

```r
# Rate ratio and 95% CI
cbind(bj = coef(mod.NBrate),
      RR = exp(coef(mod.NBrate)),
      exp(confint.default(mod.NBrate)))
```

```
##                                bj        RR     2.5 %    97.5 %
## (Intercept)          0.7559215448 2.1295731 1.8668506 2.4292686
## strength_factorYes -0.0255047304 0.9748178 0.9418390 1.0089513
## qol                -0.0006271854 0.9993730 0.9982856 1.0004616
## cal_factorYes      -0.1024329815 0.9026386 0.8691695 0.9373966
## race_factorBlack   -0.0038565306 0.9961509 0.9548581 1.0392294
## race_factorOther   -0.0123338061 0.9877419 0.9413829 1.0363839
## act_factorLimited  -0.0430480205 0.9578654 0.9244184 0.9925225
## act_factorRigorous -0.1023122312 0.9027476 0.8491202 0.9597620
## age                 0.0028192411 1.0028232 1.0010773 1.0045722
## pain               -0.0014838202 0.9985173 0.9919747 1.0051030
```

6a. [7 points] Using Model 3, report and interpret the rate ratio associated with the strength training variable, report its 95% confidence interval and perform a hypothesis test to determine if there is a significant association between the healthcare utilization rate and participation in the strength training program. (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

- Report and interpret the rate ratio associated with the strength training variable, report its 95% confidence interval.

The adjusted rate ratio associated with the strength training variable is $\hat{RR} = e^{b_1}$ = -0.026 [95% CI (0.942, 1.009)], indicating the rate of healthcare utilizations in those who participate in the strength training program is -0.026 times compared to those who don't.

- Perform a hypothesis test to determine if there is a significant association between the healthcare utilization rate and participation in the strength training program.

(i) State the null and alternative hypotheses;

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic is -0.012, p-value <.001.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We fail to reject H_0 and conclude that the rate of healthcare utilizations is not significantly different in those who participate in the strength training program and those who not.

6b. [5 points] Notice that there are some parameters (slopes) that were found to be statistically significant in Model 2, but are no longer statistically significant in Model 3. For which parameters does this occur? What is the reason for this loss of statistical significance?

These params are: quality of life index, other race. Because standard errors of params from the negative binomial model are **larger** than those in the Poisson model, as a result, the individual Wald test p-values are **larger** in the negative binomial model, we are more harder to reject H_0.

7. Our goal is to now refine Model 3 to give a parsimonious model that will be used to identify the factors that are independently associated with the healthcare utilization rate in this population of women.

7a. [10 points] Begin by removing the variable from Model 3 with the largest p-value and re-fit the model. You may use either a Wald test or likelihood ratio test to assess statistical significance of binary and quantitative predictors but should use a likelihood ratio test to assess overall statistical significance of categorical predictors made up of >2 levels. [Note: Statistical decisions involving categorical variables with >2 levels should be based on the result of the likelihood ratio test.] Repeat this process, removing one variable at a time, until there are only statistically significant predictors (at the $\alpha = 0.05$-level) remaining in the model. At each stage, clearly state which variable is being dropped and why. Report the equation of the final fitted negative binomial regression model.

- Step 1

```
Anova(mod.NBrate)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: hcu
##                 LR Chisq Df Pr(>Chisq)
## strength_factor   2.0913  1   0.148137
## qol               1.2583  1   0.261974
## cal_factor       28.3092  1  1.034e-07
## race_factor       0.2565  2   0.879640
## act_factor       12.1830  2   0.002262
## age               9.8985  1   0.001654
## pain              0.1950  1   0.658808
```

The variable race has largest p-value (0.88), which means the overall effect of race is not statistically significant in the presence of the other variables in this negative binomial model, so we can remove it from the full model.

- Step 2

```
# negative binomial regression model after romving race factor
mod.NBrate2 <- glm.nb(hcu ~ strength_factor + qol + cal_factor + act_factor + age + pain + offset(log(pe
                 data = hcu)

Anova(mod.NBrate2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: hcu
##                 LR Chisq Df Pr(>Chisq)
## strength_factor   2.1018  1   0.147122
## qol               1.2646  1   0.260775
## cal_factor       28.5396  1   9.18e-08
## act_factor       12.1078  2   0.002349
## age               9.9684  1   0.001593
## pain              0.1829  1   0.668887
```

The variable pain has largest p-value (0.669), which means the overall effect of pain score is not statistically significant in the presence of the other variables in this negative binomial model, so we can remove it from the second model.

- Step 3

```
# negative binomial regression model after romving race factor and pain
mod.NBrate3 <- glm.nb(hcu ~ strength_factor + qol + cal_factor + act_factor + age + offset(log(period))
```

```r
                 data = hcu)

Anova(mod.NBrate3)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: hcu
##                 LR Chisq Df Pr(>Chisq)
## strength_factor   2.0212  1   0.155111
## qol               1.2748  1   0.258874
## cal_factor       28.4108  1  9.812e-08
## act_factor       12.0368  2   0.002434
## age              10.0516  1   0.001522
```

The variable quality of life index has largest p-value (0.259), which means the overall effect of quality of life index is not statistically significant in the presence of the other variables in this negative binomial model, so we can remove it from the third model.

- Step 4

```r
# negative binomial regression model after romving race factor, pain and quality of life index
mod.NBrate4 <- glm.nb(hcu ~ strength_factor + cal_factor + act_factor + age + offset(log(period)),
                 data = hcu)

Anova(mod.NBrate4)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: hcu
##                 LR Chisq Df Pr(>Chisq)
## strength_factor   2.0730  1   0.149925
## cal_factor       28.2240  1  1.081e-07
## act_factor       11.9889  2   0.002493
## age               9.9001  1   0.001653
```

The variable strength has largest p-value (0.15), which means the overall effect of strength training program is not statistically significant in the presence of the other variables in this negative binomial model, so we can remove it from the fourth model.

- Step 5

```r
# negative binomial regression model after romving race factor, painm quality of life index and strengt
mod.NBrate5 <- glm.nb(hcu ~ cal_factor + act_factor + age + offset(log(period)),
                 data = hcu)

Anova(mod.NBrate5)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: hcu
##            LR Chisq Df Pr(>Chisq)
## cal_factor   28.132  1  1.133e-07
## act_factor   14.170  2  0.0008377
## age          12.330  1  0.0004457
```

Now all the predictors in this model are statistically significant.

```r
summary(mod.NBrate5)
```

```
##
## Call:
## glm.nb(formula = hcu ~ cal_factor + act_factor + age + offset(log(period)),
##     data = hcu, init.theta = 2.76848855, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5155  -0.7533  -0.4846  -0.0023   9.8011
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         0.6888686  0.0551139  12.499  < 2e-16
## cal_factorYes      -0.1019526  0.0192696  -5.291 1.22e-07
## act_factorLimited  -0.0443812  0.0181008  -2.452 0.014211
## act_factorRigorous -0.1097793  0.0307851  -3.566 0.000362
## age                 0.0030786  0.0008717   3.532 0.000413
##
## (Dispersion parameter for Negative Binomial(2.7685) family taken to be 1)
##
##     Null deviance: 6086.5  on 6059  degrees of freedom
## Residual deviance: 6030.9  on 6055  degrees of freedom
## AIC: 42907
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  2.7685
##          Std. Err.:  0.0547
##
##  2 x log-likelihood:  -42894.5800
```

The equation of the final fitted negative binomial regression model is $\log(\hat{\lambda}) = 0.689 - 0.102$ Calcium Use $-$ 0.044 Limited Activity $-$ 0.11 Rigorous Activity $+$ 0.003 Age.

7b. [10 points] Using your final model from Question **7a**, interpret each rate ratio and report the 95% confidence interval for each rate ratio. Perform a hypothesis test of each slope parameter. (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

- Controlling for all the other variables in the model, the rate of healthcare utilizations in those who **use calcium** is -9.7% lower than those who don't (ref); adjusted $\hat{RR} = e^{b_1} = 0.903$ [95% CI (0.87, 0.938)].

Perform a hypothesis test.

(i) State the null and alternative hypotheses;

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic is -5.291, p-value <.001.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that the rate of healthcare utilizations is significantly different in those who use calcium in the strength training program and those who not.

- Controlling for all the other variables in the model, the rate of healthcare utilizations in those who with **limited** level of activity is -4.3% lower than those who with None level of activity (ref); adjusted $\hat{RR} = e^{b_2} = 0.957$ [95% CI (0.923, 0.991)].

Perform a hypothesis test.

(i) State the null and alternative hypotheses;

$H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic is -2.452, p-value = 0.014.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that the rate of healthcare utilizations is significantly different in those who with limited level of activity and who with limited level of activity.

- Controlling for all the other variables in the model, the rate of healthcare utilizations in those who with **rigorous** level of activity is -10.4% lower than those who with None level of activity (ref); adjusted $\hat{RR} = e^{b_3} = 0.896$ [95% CI (0.844, 0.952)].

Perform a hypothesis test.

(i) State the null and alternative hypotheses;

$H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic is -3.566, p-value <.001.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that the rate of healthcare utilizations is significantly different in those who with rigorous level of activity and who with limited level of activity.

- Controlling for all the other variables in the model, as **age** increases, the rate of healthcare utilizations increases. A 1-year increase in age increases the rate of healthcare utilizations by 0.3%; adjusted $\hat{RR} = e^{b_4} = 1.003$ [95% CI (1.001, 1.005)].

Perform a hypothesis test.

(i) State the null and alternative hypotheses;

$H_0 : \beta_4 = 0$ vs. $H_1 : \beta_4 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic is 3.532, p-value <.001.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that there is a siginificant linear relationship between the rate of healthcare utilizations and age.

7c. [7 points] Using your final model from Question **7a**, estimate the yearly healthcare utilization rate for all combinations of factor levels included in your final model. When specifying your `newdata` data frame for use in the `predict()` function, set the value of any quantitative variables included in your final model at their mean value. For example, if your final model includes *quality of life*, *calcium use* and *race*, predict the HCU rate when (1) *quality of life* = 44.679, *race* = White, and *calcium use* = No; (2) *quality of life* = 44.679, *race* = Black, and *calcium use* = No; (3) *quality of life* = 44.679, *race* = Other, and *calcium use* = No; (4) *quality of life* = 44.679, *race* = White, and *calcium use* = Yes; (5) *quality of life* = 44.679, *race* = Black, and *calcium use* = Yes; (6) *quality of life* = 44.679, *race* = Other, and *calcium use* = Yes. As your answer to this question, create a simple table that reports the fitted annual rates and their corresponding x values. What are the values of x in your table from that have the lowest estimated annual healthcare utilization rate? Do the trends that you observe in the annual rates agree with the direction of the rate ratios associated with the categorical predictors in your model? Explain. For example, holding *quality of life* and *calcium use* constant, do the fitted HCU rates in whites, blacks, and others follow the trends that you observed in the rate ratios?

```
# new data frame includes all possible combinations of x for prediction
pred.x <- data.frame(cal_factor = c("No", "No","No","Yes", "Yes", "Yes"), act_factor = c("None", "Limite

# fitted value
lambdahat <- predict(mod.NBrate5, newdata = pred.x, type = "response")

table = cbind(fitted = lambdahat, pred.x)

table
```

```
##     fitted cal_factor act_factor  age period
## 1 16.69628         No       None 58.6      7
## 2 15.97148         No    Limited 58.6      7
## 3 14.96040         No   Rigorous 58.6      7
## 4 15.07795        Yes       None 58.6      7
## 5 14.42340        Yes    Limited 58.6      7
## 6 13.51032        Yes   Rigorous 58.6      7
```

Calcium use = Yes, activity level = Rigorous, age = 58.6 has the lowest healthcare utilization rate (13.51 times/year). The trends that I observed in the annual rates agree with the direction of rate ratios associated with categorical predictors (calcium use and activity level) in my model. Holding activity level constant, fitted HCU rates in those who use calcium is lower than those who not (ref). Holding calcium use constant, fitted HCU rates in activity level group is: rigorous < limited < none (ref).

7d. [3 points] If you were to increase the values of any quantitative predictors in your model while holding the categorical variables fixed/constant, would you expect the fitted rates to increase or decrease? Why?

The quantitative predictor in my model is **age**, I expect the fitted rates to increase because the coefficient of age is positive (0.003), which means as **age** increases, the rate of healthcare utilizations increases.