Homework0

Takuto Yoshida

2023-01-24

1. Read in the nhefs.xlsx file from the EPI 289 course website. Show your log to demonstrate that the file was successfully assigned.

```
# Question1
install.packages("readxl", repos = "http://cran.us.r-project.org") # read excel file
```

```
##
## The downloaded binary packages are in
##  /var/folders/01/yxgsk4rn7r9gh2ch_khv9cp40000gn/T//RtmpimRUJD/downloaded_packages
```

```
library("readxl")
df <- read_excel("/Users/yoshidatakuto/Dropbox/HSPH/MPH-CLE/Spring/EPI289/Homework0/nhefs.xlsx")
```

2. Sort the data set by the variable seqn. Print out the ID number, age, and sex for the first 10 observations.

```
# Question2
## Sorting the data
df[1:10, c("seqn", "age", "sex")]
```

```
## # A tibble: 10 x 3
##     seqn   age   sex
##    <dbl> <dbl> <dbl>
##  1   233    42     0
##  2   235    36     0
##  3   244    56     1
##  4   245    68     0
##  5   252    40     0
##  6   257    43     1
##  7   262    56     1
##  8   266    29     1
##  9   419    51     0
## 10   420    43     0
```

3. Find the mean systolic blood pressure and standard error for men and for women.

```
# Question3
summary(df$sbp) # 77 missing values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    87.0   116.0   126.0   128.7   140.0   229.0      77
```

```
# Separate the dataset by gender
df_men <- subset(df, sex==0)
df_women <- subset(df, sex==1)
summary(df_men$sbp) # 45 missing values
```

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

```
##    90.0   118.0   129.0   131.2   141.0   229.0      45
```

```r
summary(df_women$sbp) # 32 missing values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    87.0   113.0   124.0   126.3   137.0   212.0      32
```

```r
# Find the mean and standard error of systolic blood pressure in men
mean_sbp_men <- mean(df_men$sbp, na.rm=T)
se_sbp_men <- sd(df_men$sbp, na.rm=T)/sqrt(length(df_men$sbp)-45)

# Find the mean and standard error of systolic blood pressure in women
mean_sbp_women <- mean(df_women$sbp, na.rm=T)
se_sbp_women <- sd(df_women$sbp, na.rm=T)/sqrt(length(df_women$sbp)-32)

# Print the results
cbind(Mean = mean_sbp_men, SE = se_sbp_men)
```

```
##          Mean        SE
## [1,] 131.2467 0.6866385
```

```r
cbind(Mean = mean_sbp_women, SE = se_sbp_women)
```

```
##         Mean        SE
## [1,] 126.312 0.6703827
```

4. What is the mean, 25th percentile, 50th percentile, 75th percentile, and interquartile range of weight in 1971 (in kilograms).

```r
# Question 4
summary(df$wt71) # No missing value
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   36.17   59.65   69.40   71.05   79.95  169.19
```

```r
mean(df$wt71)
```

```
## [1] 71.05213
```

```r
quantile(df$wt71, probs = c(0.25, 0.50, 0.75))
```

```
##   25%   50%   75% 
## 59.65 69.40 79.95
```

```r
IQR(df$wt71)
```

```
## [1] 20.3
```

5a. Using ifelse statements, create a new categorical variable corresponding to quartiles of weight in 1971 as based on the cut-points from Question (4). Give a tabulation of your results.

```r
# Question 5a
## Create the new variables
df$wt71_cat <- ifelse(df$wt71<=59.65, 1, ifelse(df$wt71<=69.40, 2, ifelse(df$wt71<=79.95, 3, 4)))

## Tabulate the categorical variable
table(df$wt71_cat)
```

```
## 
##   1   2   3   4 
## 414 402 406 407
```

5b. Create quartiles for weight in 1971 using cut in R. Give a tabulation of your results. Do your results match those of Question (5a)? Why or why not?

```
# Question 5b
df_quart <- cut(df$wt71, breaks = c(-Inf, quantile(df$wt71, probs = c(0.25)), quantile(df$wt71, probs =

table(df_quart)
```

```
## df_quart
## (-Inf,59.6] (59.6,69.4]   (69.4,80]   (80, Inf]
##         414         402         406         407
```

6. Using lm in R, fit a univariate linear regression model for the outcome weight in 1971 with number of cigarettes smoked per day in 1971 as the predictor. Report the parameter estimate for cigarettes smoked per day.

```
# Question 6
linear_model <- lm(df$wt71 ~ df$smokeintensity, data = df)
summary(linear_model)
```

```
##
## Call:
## lm(formula = df$wt71 ~ df$smokeintensity, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.345 -11.452  -1.718   8.840  99.891
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       68.73497    0.78011  88.109  < 2e-16 ***
## df$smokeintensity  0.11275    0.03292   3.425  0.00063 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.68 on 1627 degrees of freedom
## Multiple R-squared:  0.007159,   Adjusted R-squared:  0.006549
## F-statistic: 11.73 on 1 and 1627 DF,  p-value: 0.0006298
```

```
coef(linear_model)
```

```
##       (Intercept) df$smokeintensity
##        68.7349725         0.1127502
```

```
confint(linear_model)
```

```
##                         2.5 %     97.5 %
## (Intercept)       67.20484656 70.2650984
## df$smokeintensity  0.04818256  0.1773178
```

7. Create a cross-tabulation between sex and race.

```
# Question 7
table(df$sex, df$race, dnn = c("sex", "race"))
```

```
##    race
## sex   0   1
##   0 705  94
##   1 709 121
```

8. Using lm in R, fit a multivariate linear regression model for the outcome weight in 1971 with age, sex, and race as the predictors. From this model, print the observed and predicted values of weight in 1971 for the first 5 observations. What is the predicted value of weight in 1971 for an individual of age 40, female, and of Black or other race/ethnicity?

```
# Question8
## Fit the multivariable linear regression model
df$sex <- as.factor(df$sex)
df$race <- as.factor(df$race)
adj_linear_model <- lm(wt71 ~ age + sex + race, data = df)

## Get the predicted values for the first 5 observations
predicted_values <- predict(adj_linear_model, newdata = df[1:5,])

## Get the observed value for the first 5 observations
observed_values <- df$wt71[1:5]

## Print the observed and predicted values
cbind(Observed = observed_values, Predicted = predicted_values)
```

```
##   Observed Predicted
## 1    79.04  82.33576
## 2    58.63  76.87137
## 3    56.81  69.49903
## 4    59.42  82.17154
## 5    87.09  76.84610
```

```
# Predicted value
covariate_values <- data.frame(age = 40, sex = "1", race = "1")
predicted_values_q8 <- predict(adj_linear_model, newdata = covariate_values)
print(predicted_values_q8)
```

```
##        1
## 69.60009
```

9. Fit the same model from Question (8) using glm in R and compare your results.

```
# Question9
adj_linear_model_2 <- glm(wt71 ~ age + sex + race, data = df)

## Get the predicted values for the first 5 observations
predicted_values_2 <- predict(adj_linear_model_2, newdata = df[1:5,])

## Get the observed value for the first 5 observations
observed_values <- df$wt71[1:5]

## Print the observed and predicted values
cbind(Observed = observed_values, Predicted = predicted_values_2)
```

```
##   Observed Predicted
## 1    79.04  82.33576
## 2    58.63  76.87137
## 3    56.81  69.49903
## 4    59.42  82.17154
## 5    87.09  76.84610
```

10. Using glm with family specified as binomial in R, fit a multivariate logistic regression model for the

outcome asthma diagnosis in 1971 with age, sex, race, and usual physical activity status (var active) as the predictors. Print the predicted probabilities of asthma diagnosis for the individuals with the first 5 ID numbers.

```
# Question10
## Fit the multivariable logistic regression model
df$asthma <- as.factor(df$asthma)
df$active <- as.factor(df$active)
adj_linear_model_3 <- glm(asthma ~ age + sex + race + active, data = df, family = binomial(link = 'logi

## Get the predicted values for the first 5 observations
predicted_values_3 <- predict(adj_linear_model_3, newdata = df[1:5,], type = "response")
predicted_values_3
```

```
##          1          2          3          4          5
## 0.02287751 0.03464734 0.04262889 0.02952948 0.03570667
```

11. (Optional) Create a graph that plots systolic blood pressure on the Y-axis and usual physical activity status (var active) on the X-axis.

```
# Question 11
boxplot(sbp ~ active, data = df, xlab="Usual physical activity status", ylab="Systolic blood pressure",
```