Proposal

Sakura Hu, Ruikang Wang, Peizeng Yuan

2024-10-10

Contributions

Sakura Hu: Ruikang Wang: Peizeng Yuan:

Introduction

Data Description

Preliminary Results

Bibliography

```
# Install the NHANES package
install.packages("NHANES")
```

```
##
## The downloaded binary packages are in
##  /var/folders/km/sjr_hj0n7lj6hf7j8l7nrfpm0000gn/T//Rtmp5V8o5W/downloaded_packages
```

```
# Load the package
library(NHANES)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

# View available datasets
data(package = "NHANES")

# Load the main dataset
data(NHANES)

# Select only the columns of the predictor variables and filter out rows with missing values
```

```r
nhanes_data <- NHANES %>%
  select(BMI, Poverty, PhysActiveDays, Age, SleepHrsNight, Gender) %>%
  filter(
    !is.na(BMI),!is.na(Poverty),!is.na(PhysActiveDays),!is.na(Age),!is.na(SleepHrsNight),!is.na(Gender)
  )

# Check the first few rows to make sure everything looks correct
head(nhanes_data)
```

```
## # A tibble: 6 x 6
##     BMI Poverty PhysActiveDays   Age SleepHrsNight Gender
##   <dbl>   <dbl>          <int> <int>         <int> <fct>
## 1  27.2     5                5    45             8 female
## 2  27.2     5                5    45             8 female
## 3  27.2     5                5    45             8 female
## 4  23.7     2.2              7    66             7 male
## 5  23.7     5                5    58             5 male
## 6  26.0     2.2              1    54             4 male
```

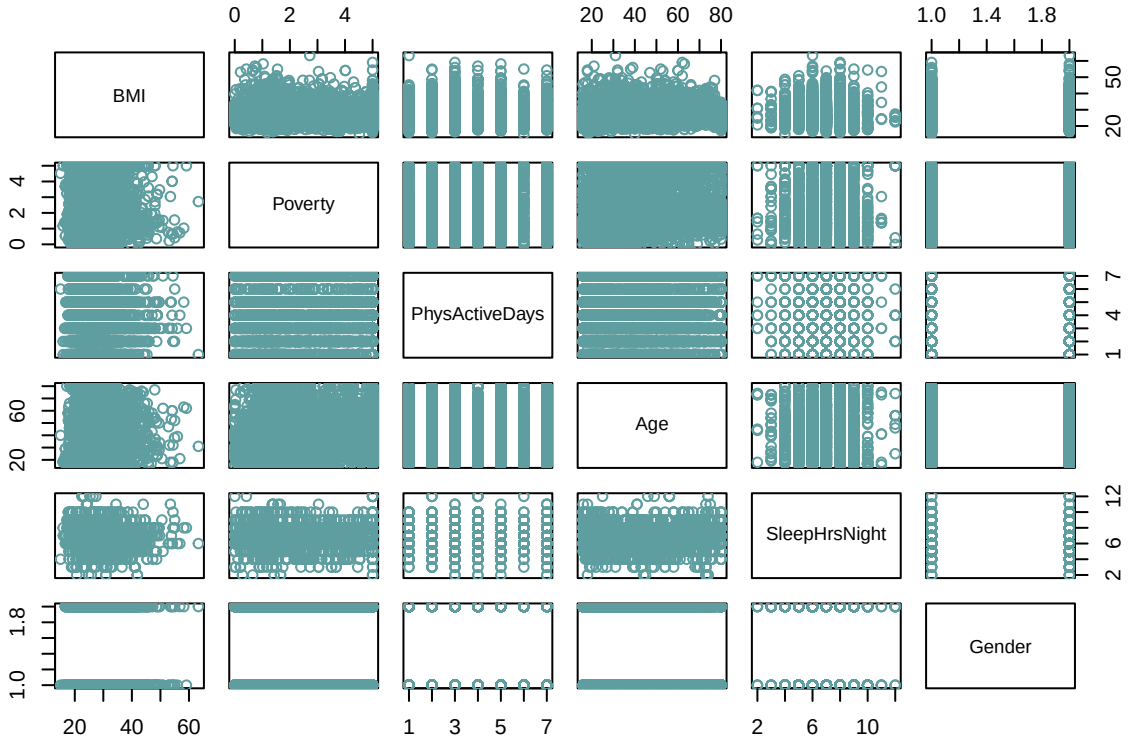
```r
# Convert Gender to a factor
nhanes_data$Gender <- as.factor(nhanes_data$Gender)

# Fit a linear model
lm_model <- lm(log(BMI) ~ Poverty + PhysActiveDays + Age + SleepHrsNight + Gender,
               data = nhanes_data)

# Summary of the model
summary(lm_model)
```

```
##
## Call:
## lm(formula = log(BMI) ~ Poverty + PhysActiveDays + Age + SleepHrsNight +
##     Gender, data = nhanes_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58922 -0.14816 -0.01533  0.13461  0.82036
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.3596886  0.0229080 146.660  < 2e-16 ***
## Poverty        -0.0101187  0.0021355  -4.738 2.24e-06 ***
## PhysActiveDays -0.0043391  0.0019351  -2.242 0.025004 *
## Age             0.0012950  0.0002051   6.313 3.07e-10 ***
## SleepHrsNight  -0.0102776  0.0027065  -3.797 0.000149 ***
## Gendermale      0.0212251  0.0070538   3.009 0.002639 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2099 on 3567 degrees of freedom
## Multiple R-squared:  0.02335,    Adjusted R-squared:  0.02198
## F-statistic: 17.06 on 5 and 3567 DF,  p-value: < 2.2e-16
```

```r
plot(nhanes_data[, c(1, 2, 3, 4, 5, 6)], col="cadetblue")
```




```r
png("histogram_plots.png", width = 1200, height = 1200) # Adjust the size
par(
  mfrow = c(3, 3),
  # 2x4 grid layout
  mar = c(5, 5, 4, 2),
  # Plot margins
  oma = c(5, 5, 5, 5)
)     # adds blank space around edges

# Customize text sizes
par(cex.main = 2,
    cex.lab = 1.5,
    cex.axis = 1.5)

# Plot histograms for each variable

hist(nhanes_data$BMI, main="figure1: BMI Distribution", xlab="BMI", col="lightblue")
hist(log(nhanes_data$BMI), main="figure2: log(BMI) Distribution", xlab="log(BMI)", col="lightblue")
hist(nhanes_data$Poverty, main="figure3: Poverty Index Distribution", xlab="Poverty Index", col="lightg
hist(nhanes_data$PhysActiveDays, main="figure4: Physical Activity Days Distribution", xlab="Phys Active
hist(nhanes_data$Age, main="figure5: Age Distribution", xlab="Age", col="lightyellow")
hist(nhanes_data$SleepHrsNight, main="figure6: Sleep Hours Distribution", xlab="Sleep Hours", col="ligh
hist(as.numeric(nhanes_data$Gender), main="figure7: Gender Distribution", xlab="Gender (0=Female, 1=Mal
```

```r
dev.off() # Save the file
```

```
## pdf
##   2
```

```r
png("NHANES_plots.png", width = 3000, height = 3000) # Adjust the size
par(
  mfrow = c(4, 4),
  # 4x4 grid layout
  mar = c(5, 5, 4, 2),
  # Plot margins
  oma = c(5, 5, 5, 5)
)     # adds blank space around edges

# Customize text sizes
par(cex.main = 4,
    cex.lab = 3,
    cex.axis = 2)

# Extract the fitted values from the linear model
fitted_values = fitted(lm_model)

# Extract the residuals
residual_values = resid(lm_model)

# Plot the fitted values vs. the residual values
plot(
  fitted_values,
  residual_values,
  main = "Figure8: fitted versus residual values",
  xlab = "Fitted",
  ylab = "Residuals"
)

# Extract the standardized residuals
sresidual_values = rstandard(lm_model)

# Plot fitted values vs. standardized residuals
plot(
  fitted_values,
  sresidual_values,
  main = "Figure9: fitted vs. standardized residuals values",
  xlab = "Fitted",
  ylab = "Standardized Residuals"
)

# Plot a histogram of the standardized residuals
hist(sresidual_values, main = "Figure10: Standardized residuals histogram", xlab = "Standardized residu

# Plot residuals against Poverty
plot(
  nhanes_data$Poverty,
  residual_values,
```

```
  main = "Figure11: Poverty vs. Residuals",
  xlab = "Poverty",
  ylab = "Residuals"
)

# Plot residuals against Age
plot(
  nhanes_data$Age,
  residual_values,
  main = "Figure12: Age vs. Residuals",
  xlab = "Age",
  ylab = "Residuals"
)

# Plot residuals against Gender
plot(
  nhanes_data$Gender,
  residual_values,
  main = "Figure13: Gender vs. Residuals",
  xlab = "Gender",
  ylab = "Residuals"
)

#Log-transform the BMI values from the dataset
BMI_values = log(nhanes_data$BMI)
fitted_values = fitted(lm_model)

# Plot the fitted values (predicted log(BMI)) against the actual log(BMI) values
plot(
  fitted_values,
  BMI_values,
  main = "Figure14: regression fit vs. BMI",
  xlab = "Fitted values",
  ylab = "log(BMI)"
)
abline(0, 1, col = c("blue"), lty = 1)
legend(
  "bottomright",
  legend = c("y = x line"),
  col = c("blue"),
  lty = 1
)


# Normal Q-Q plot
plot(lm_model, which = 2)
title(main = "Figure15: Normal Q-Q Plot")

# Scatter plot for Age vs. log(BMI)
plot(
  nhanes_data$Age,
  log(nhanes_data$BMI),
  main = "Figure16: Age vs. BMI",
```

```r
  xlab = "Age",
  ylab = "log(BMI)"
)

# Add the fitted regression line
fit <- lm(log(nhanes_data$BMI) ~ nhanes_data$Age)
abline(fit, col = "red", lwd = 2)


# Scatter plot for Gender vs. log(BMI)
plot(
  nhanes_data$Gender,
  log(nhanes_data$BMI),
  main = "Figure17: Gender vs. BMI",
  xlab = "Gender",
  ylab = "log(BMI)"
)

# Add the fitted regression line
fit <- lm(log(nhanes_data$BMI) ~ nhanes_data$Gender)
abline(fit, col = "red", lwd = 2)

# Scatter plot for Poverty vs. log(BMI)
plot(
  nhanes_data$Poverty,
  log(nhanes_data$BMI),
  main = "Figure18: Poverty vs. BMI",
  xlab = "Poverty",
  ylab = "log(BMI)"
)

# Add the fitted regression line
fit <- lm(log(nhanes_data$BMI) ~ nhanes_data$Poverty)
abline(fit, col = "red", lwd = 2)

# Scatter plot for SleepHrsNight vs. log(BMI)

plot(
  nhanes_data$SleepHrsNight,
  log(nhanes_data$BMI),
  main = "Figure19: SleepHrsNight vs. BMI",
  xlab = "SleepHrsNight",
  ylab = "log(BMI)"
)

# Add the fitted regression line
fit <- lm(log(nhanes_data$BMI) ~ nhanes_data$SleepHrsNight)
abline(fit, col = "red", lwd = 2)

# Scatter plot for PhysActiveDays vs. log(BMI)
plot(
  nhanes_data$PhysActiveDays,
  log(nhanes_data$BMI),
```

```
  main = "Figure20: Physical Active Days vs. BMI",
  xlab = "PhysActiveDays",
  ylab = "log(BMI)"
)

# Add the fitted regression line
fit <- lm(log(nhanes_data$BMI) ~ nhanes_data$PhysActiveDays)
abline(fit, col = "red", lwd = 2)

dev.off() # Save the file
```

```
## pdf
##   2
```