# Data analysis tools

# Week 3 assignment

# Pearson correlation

I chose the **addhealth** as my data set. Two correlations were tested:

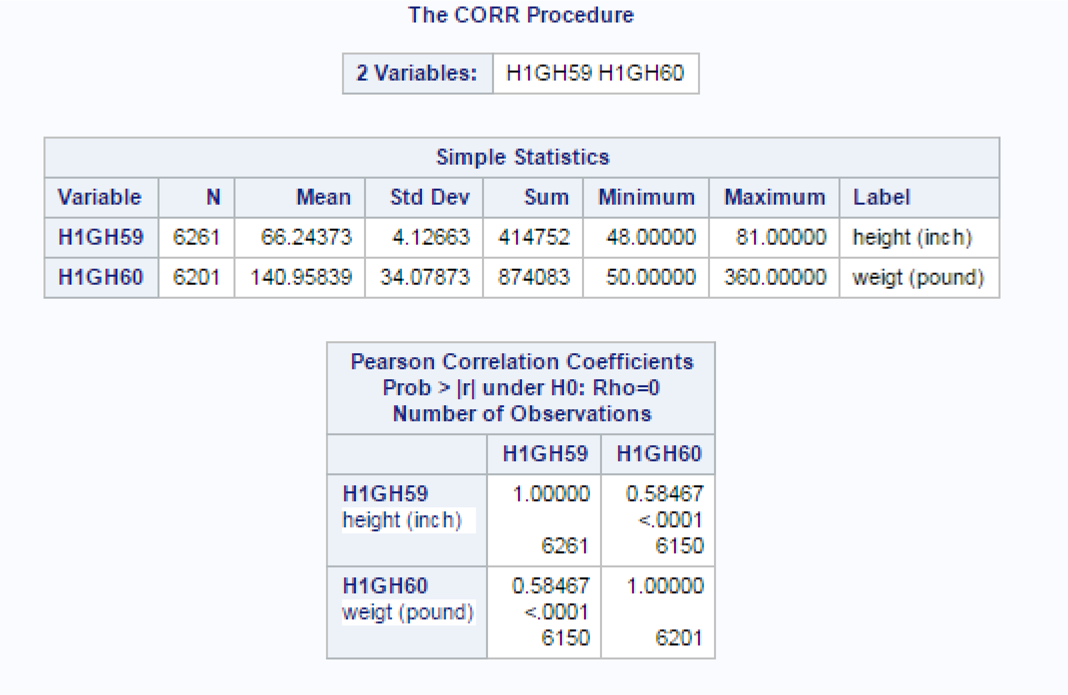
1. height and weight;

2. BMI, general health and frequency of headache.

Because both the general health and frequency of headache have more than 3 levels and the means of them are meaningful, they can be used to generate correlation coefficient.

Here are my results.

**1. the correlation between height and weight.**



The CORR Procedure

| 2 Variables: | H1GH59 H1GH60 |
|---|---|

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum | Label |
|---|---|---|---|---|---|---|---|
| H1GH59 | 6261 | 66.24373 | 4.12663 | 414752 | 48.00000 | 81.00000 | height (inch) |
| H1GH60 | 6201 | 140.95839 | 34.07873 | 874083 | 50.00000 | 360.00000 | weigt (pound) |

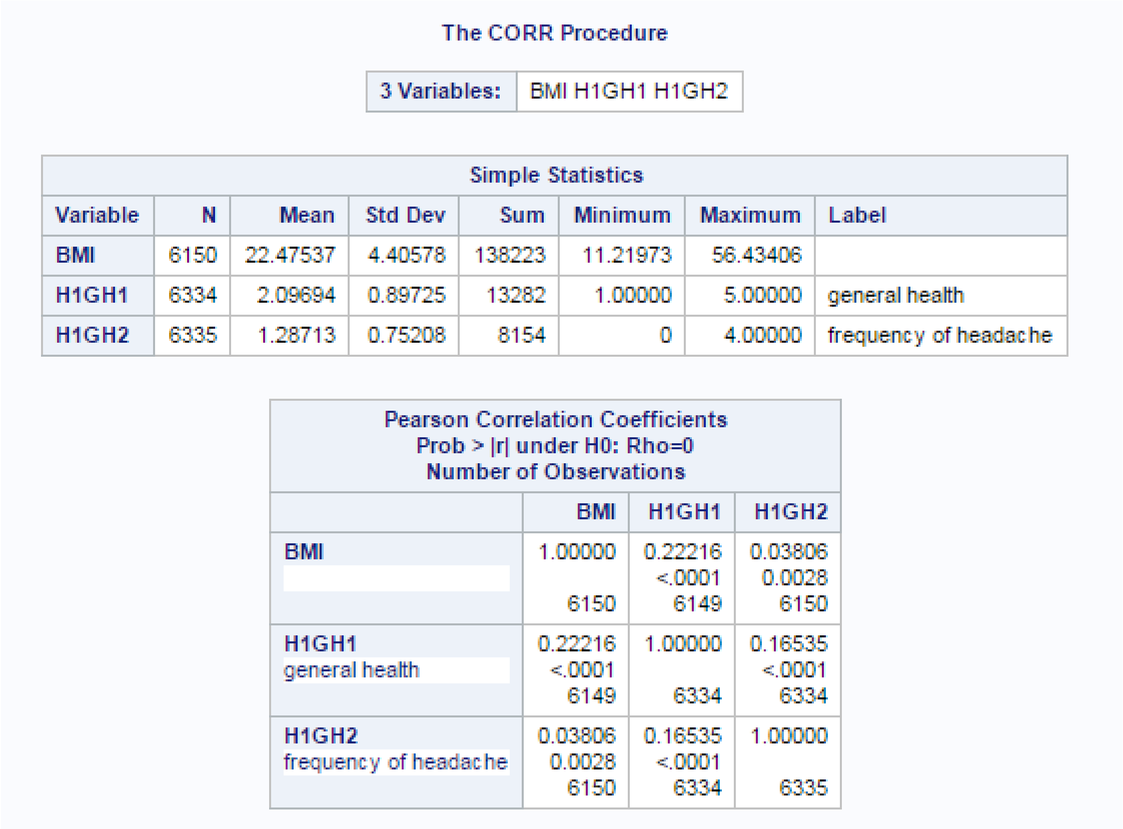
Pearson Correlation Coefficients
Prob > \|r\| under H0: Rho=0
Number of Observations

| | H1GH59 | H1GH60 |
|---|---|---|
| H1GH59<br>height (inch) | 1.00000<br><br>6261 | 0.58467<br><.0001<br>6150 |
| H1GH60<br>weigt (pound) | 0.58467<br><.0001<br>6150 | 1.00000<br><br>6201 |

P-value is less than 0.05, so at 95% confidence level, weight and height has **significant correlation**. **Correlation coefficient is 0.58467.**

R square equals to 0.3418, which means 34.18% of the variability in height is described by variability in weight.

**2. correlation between BMI, general health and frequency of headache.**



The CORR Procedure

| 3 Variables: | BMI H1GH1 H1GH2 |
|---|---|

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum | Label |
|---|---|---|---|---|---|---|---|
| BMI | 6150 | 22.47537 | 4.40578 | 138223 | 11.21973 | 56.43406 | |
| H1GH1 | 6334 | 2.09694 | 0.89725 | 13282 | 1.00000 | 5.00000 | general health |
| H1GH2 | 6335 | 1.28713 | 0.75208 | 8154 | 0 | 4.00000 | frequency of headache |

Pearson Correlation Coefficients
Prob > \|r\| under H0: Rho=0
Number of Observations

| | BMI | H1GH1 | H1GH2 |
|---|---|---|---|
| BMI | 1.00000<br><br>6150 | 0.22216<br><.0001<br>6149 | 0.03806<br>0.0028<br>6150 |
| H1GH1<br>general health | 0.22216<br><.0001<br>6149 | 1.00000<br><br>6334 | 0.16535<br><.0001<br>6334 |
| H1GH2<br>frequency of headache | 0.03806<br>0.0028<br>6150 | 0.16535<br><.0001<br>6334 | 1.00000<br><br>6335 |

At 95% confidence level, all the three variables are **significantly correlated**, because p-value of each pair is less than 0.05.

For BMI and general health **correlation coefficient is 0.22216.** R square is 0.0494, which means 4.94% of variability in BMI can be described by variability in general health.

For BMI and frequency of headache, **correlation coefficient is 0.03806**. R square is 0.0014, which means 0.14% of variability in BMI can be described by variability in frequency of headache.

For general health and frequency of headache, **correlation coefficient is 0.16335**. R square is 0.0273, which means 2.73% of variability in general health can be described by variability in frequency of headache.