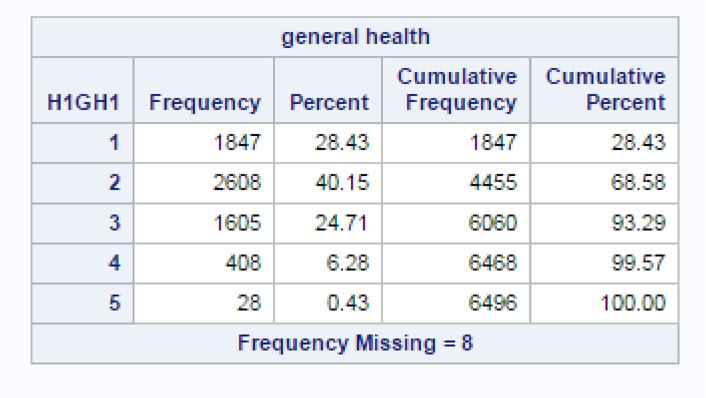
# Regression models in practice

## Week 2 assignment

## Basic linear regression model

I chose addhealth as my data set. The relationship between general health and body mass index (BMI) was analyzed.

Response variable is general health, which is categorical. There are five values of general health. 1 means excellent and 5 mean poor.



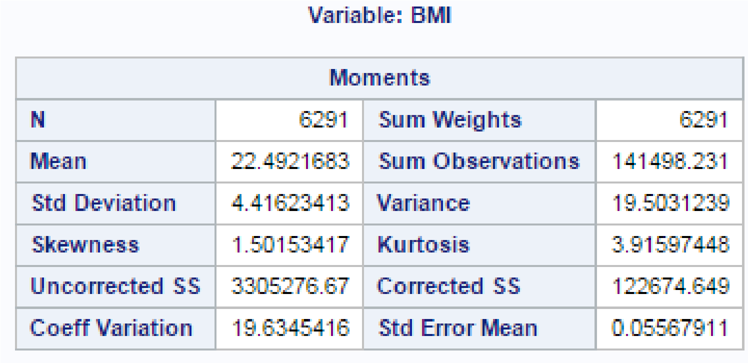
general health

| H1GH1 | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 1847 | 28.43 | 1847 | 28.43 |
| 2 | 2608 | 40.15 | 4455 | 68.58 |
| 3 | 1605 | 24.71 | 6060 | 93.29 |
| 4 | 408 | 6.28 | 6468 | 99.57 |
| 5 | 28 | 0.43 | 6496 | 100.00 |

Frequency Missing = 8

Frequency distribution of general health

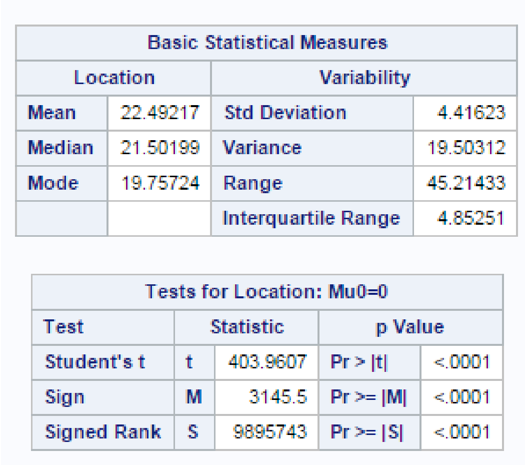
Explanatory variable is BMI, which is quantitative. The mean is 22.49. P-value is less than 0.0001, which means we can reject that it is equal to 0.



Variable: BMI

Moments

| | | | |
|---|---|---|---|
| N | 6291 | Sum Weights | 6291 |
| Mean | 22.4921683 | Sum Observations | 141498.231 |
| Std Deviation | 4.41623413 | Variance | 19.5031239 |
| Skewness | 1.50153417 | Kurtosis | 3.91597448 |
| Uncorrected SS | 3305276.67 | Corrected SS | 122674.649 |
| Coeff Variation | 19.6345416 | Std Error Mean | 0.05567911 |



Basic Statistical Measures

| Location | | Variability | |
|---|---|---|---|
| Mean | 22.49217 | Std Deviation | 4.41623 |
| Median | 21.50199 | Variance | 19.50312 |
| Mode | 19.75724 | Range | 45.21433 |
| | | Interquartile Range | 4.85251 |

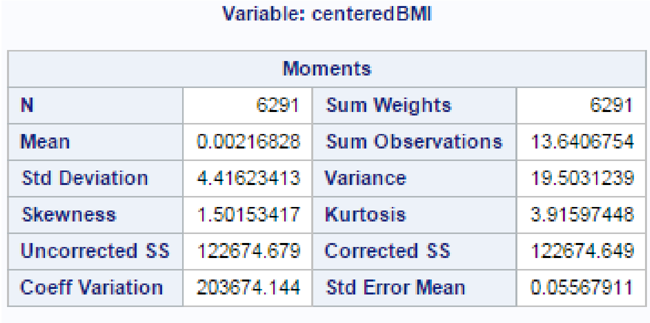
Tests for Location: Mu0=0

| Test | Statistic | | p Value | |
|---|---|---|---|---|
| Student's t | t | 403.9607 | Pr > \|t\| | <.0001 |
| Sign | M | 3145.5 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | 9895743 | Pr >= \|S\| | <.0001 |

Descriptive statistics of BMI

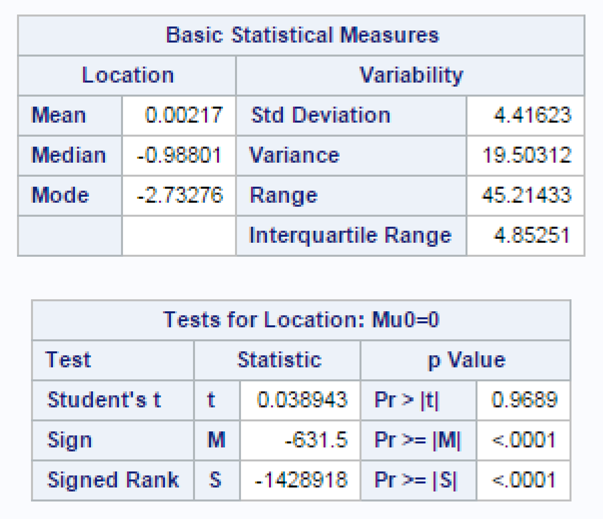
Therefore, I subtract the mean from each value of BMI and get the centered BMI, saved in the new variable named centeredBMI.

The mean of centered BMI is 0.00217. P-value is 0.9689. At 95% confidence level, we can assume the mean is equal to 0.



Variable: centeredBMI

| Moments | | | |
|---|---|---|---|
| N | 6291 | Sum Weights | 6291 |
| Mean | 0.00216828 | Sum Observations | 13.6406754 |
| Std Deviation | 4.41623413 | Variance | 19.5031239 |
| Skewness | 1.50153417 | Kurtosis | 3.91597448 |
| Uncorrected SS | 122674.679 | Corrected SS | 122674.649 |
| Coeff Variation | 203674.144 | Std Error Mean | 0.05567911 |



| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 0.00217 | Std Deviation | 4.41623 |
| Median | -0.98801 | Variance | 19.50312 |
| Mode | -2.73276 | Range | 45.21433 |
| | | Interquartile Range | 4.85251 |

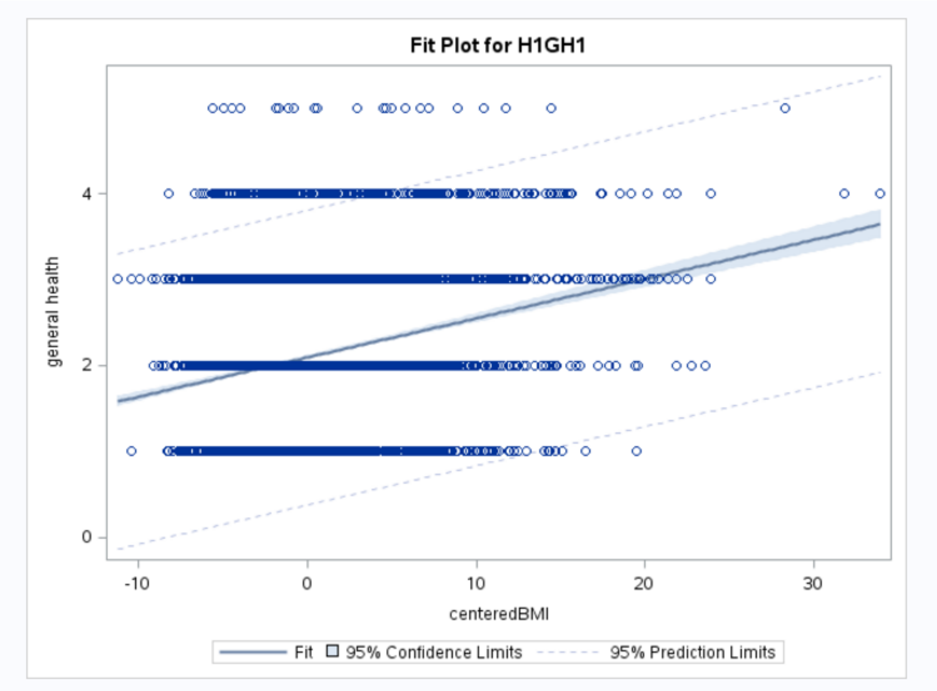
| Tests for Location: Mu0=0 | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Student's t | t | 0.038943 | Pr > \|t\| | 0.9689 |
| Sign | M | -631.5 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | -1428918 | Pr >= \|S\| | <.0001 |

Descriptive statistics of centered BMI

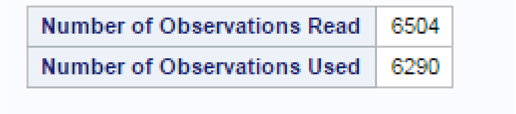
At last, the linear regression model was processed.

General health = 0.04 * BMI + 2.09

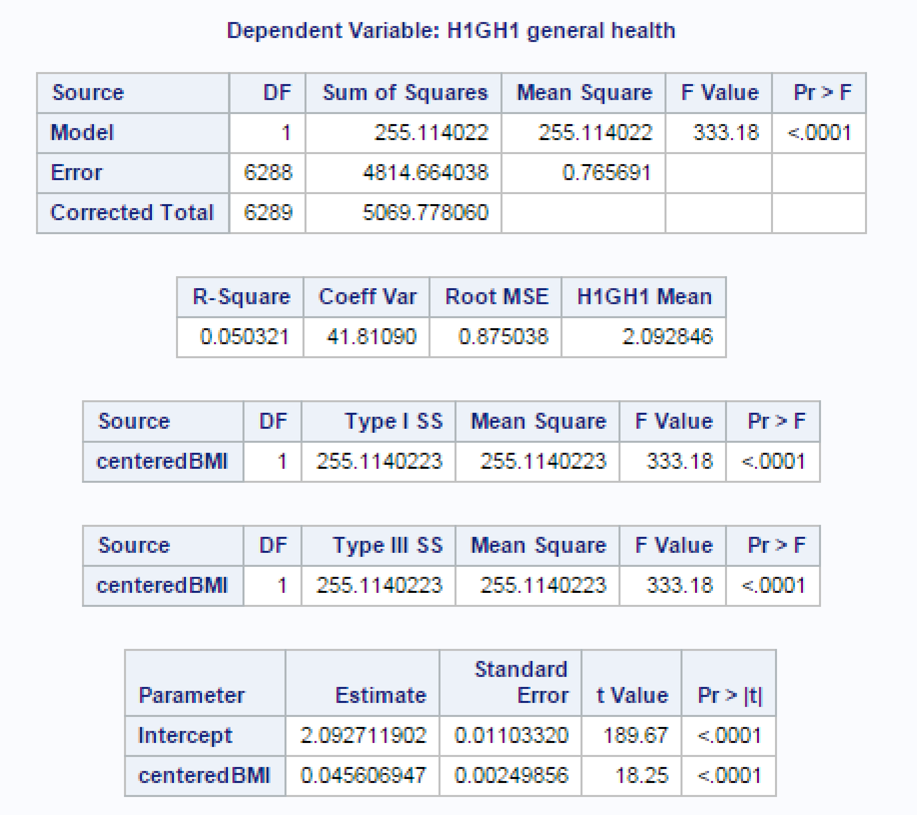
P-values are less than 0.0001.



Linear regression model



| Number of Observations Read | 6504 |
|---|---|
| Number of Observations Used | 6290 |



Dependent Variable: H1GH1 general health

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 255.114022 | 255.114022 | 333.18 | <.0001 |
| Error | 6288 | 4814.664038 | 0.765691 | | |
| Corrected Total | 6289 | 5069.778060 | | | |

| R-Square | Coeff Var | Root MSE | H1GH1 Mean |
|---|---|---|---|
| 0.050321 | 41.81090 | 0.875038 | 2.092846 |

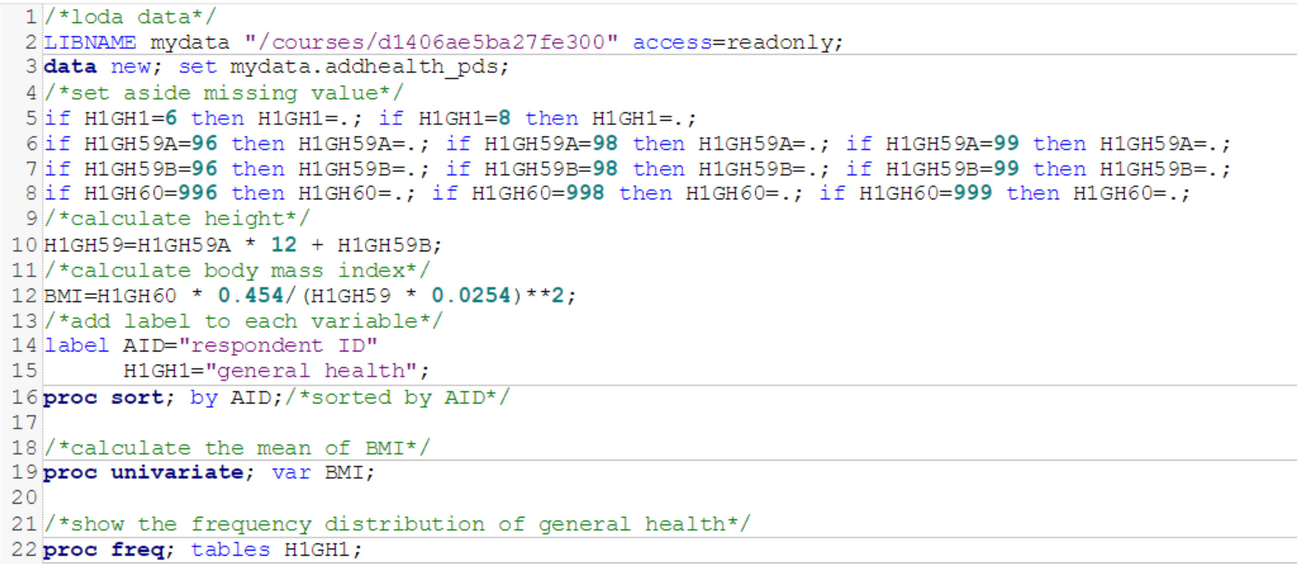
| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| centeredBMI | 1 | 255.1140223 | 255.1140223 | 333.18 | <.0001 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| centeredBMI | 1 | 255.1140223 | 255.1140223 | 333.18 | <.0001 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 2.092711902 | 0.01103320 | 189.67 | <.0001 |
| centeredBMI | 0.045606947 | 0.00249856 | 18.25 | <.0001 |

Statistic information of the linear regression model

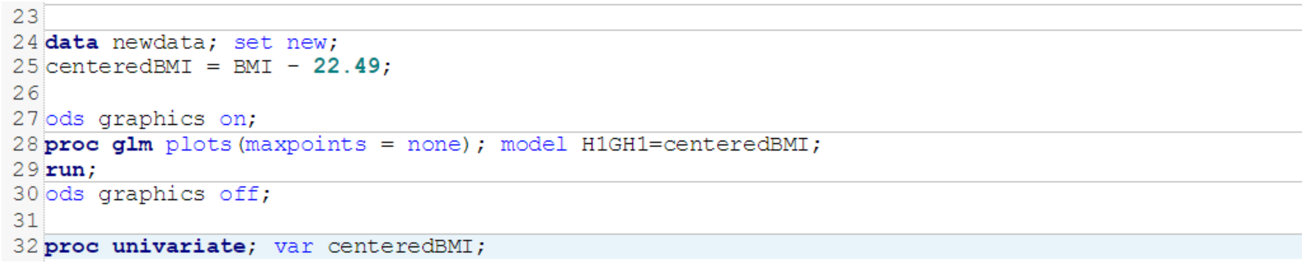
## My code



```
/*loda data*/
LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
data new; set mydata.addhealth_pds;
/*set aside missing value*/
if H1GH1=6 then H1GH1=.; if H1GH1=8 then H1GH1=.;
if H1GH59A=96 then H1GH59A=.; if H1GH59A=98 then H1GH59A=.; if H1GH59A=99 then H1GH59A=.;
if H1GH59B=96 then H1GH59B=.; if H1GH59B=98 then H1GH59B=.; if H1GH59B=99 then H1GH59B=.;
if H1GH60=996 then H1GH60=.; if H1GH60=998 then H1GH60=.; if H1GH60=999 then H1GH60=.;
/*calculate height*/
H1GH59=H1GH59A * 12 + H1GH59B;
/*calculate body mass index*/
BMI=H1GH60 * 0.454/(H1GH59 * 0.0254)**2;
/*add label to each variable*/
label AID="respondent ID"
      H1GH1="general health";
proc sort; by AID;/*sorted by AID*/

/*calculate the mean of BMI*/
proc univariate; var BMI;

/*show the frequency distribution of general health*/
proc freq; tables H1GH1;
```



```
data newdata; set new;
centeredBMI = BMI - 22.49;

ods graphics on;
proc glm plots(maxpoints = none); model H1GH1=centeredBMI;
run;
ods graphics off;

proc univariate; var centeredBMI;
```