HaoMing Chen

Professor Zhen Ma

ECON 382

23 May 2020

The data files that I'm using are the National Health and Nutrition Examination Survey from 2013-2014 Data Documentation [Body Measures (BMX_H)] and NHANES 2013-2014 Demographics Data. The variables that I am interested in are **RIDAGEYR** - Age in years at screening (0 to 79 are the range of values), (80 stands for 80 years of age and over). **BMXWT** - Weight in kg (range of values are from 3.1 to 222.6). **BMXHT**-standing height in cm (range of values are from 79.7 to 202.6). **BMXWAIST** - Waist Circumference in cm (range of values are from 40.2 to 177.9**). BMXARML** - Upper Arm Length in cm (range of values are from 9.9 to 47.9). **BMXLEG** - Upper Leg Length in cm (range of values are from 24.4 to 51.9). **RIAGENDR** – Gender (1 stands for male), (2 stands for female).

Before forming the regression line, we need to combine the 2 data, blood measures and demographic into 1 by using the command merge in R. Then, remove the missing data (N/A) using the command data=subset(demo_merge_blood, RIDAGEYR!=""&BMXWT!=""&BMXHT!=""&BMXWAIST!=""&BMXARML!=""&BMXLEG!=""&female!=""). After that, we need to be aware that RIAGENDR (gender) is a dummy variable and outcome can either be male or female. In this case, we need to leave out one variable as the benchmark or base. I choose male as the benchmark but choosing whichever one does not really matter that much. However, it is recommended to always leave out the variable that you want to study more of because you interpret the relationship between every other variable to the variable that you left out. Since I have chosen male as the benchmark, it means that I will use female as the variable in my regression, and in my interpretation, this female variable will be comparing against male because I left out male. Male and female both belong to the variable, RIAGENDR, but if you put male and female in the same regression, one of them will be dropped automatically or the regression will not perform. Therefore, I only choose to put Female in the regression instead of putting in both Male and Female.

Regression line:

BMXWT= RIDAGEYR+ BMXHT+ BMXWAIST+ BMXARML+ BMXLEG+ RIAGENDR(female/male)

BMXWT= RIDAGEYR+ BMXHT+ BMXWAIST+ BMXARML+ BMXLEG+ Female

BMXWT=-108.9-0.1601 RIDAGEYR+0.3546 BMXHT+1.095 BMXWAIST+0.4464 BMXARML+0.3400 BMXLEG+0.009947 Female

```
> summary(reg1)

Call:
lm(formula = BMXWT ~ RIDAGEYR + BMXHT + BMXWAIST + BMXARML +
    BMXLEG + female, data = demo_merge_blood2)

Residuals:
    Min      1Q  Median      3Q     Max
-32.166  -3.807  -0.188   3.478  41.981

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -1.089e+02  1.065e+00 -102.279   <2e-16 ***
RIDAGEYR    -1.601e-01  4.022e-03  -39.812   <2e-16 ***
BMXHT        3.546e-01  1.392e-02   25.464   <2e-16 ***
BMXWAIST     1.095e+00  5.111e-03  214.347   <2e-16 ***
BMXARML      4.464e-01  4.813e-02    9.274   <2e-16 ***
BMXLEG       3.400e-01  3.447e-02    9.864   <2e-16 ***
female       9.947e-03  1.596e-01    0.062     0.95
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.024 on 7325 degrees of freedom
Multiple R-squared:  0.9384,    Adjusted R-squared:  0.9384
F-statistic: 1.86e+04 on 6 and 7325 DF,  p-value: < 2.2e-16
```

Summary Statistics (the data of the variables of interest are the following):

Minimal value: BMXWT=16.9, RIDAGEYR=8, BMXHT=113, BMXWAIST=47.8, BMXARML=22.7, BMXLEG=24.4, FEMALE=0

Maximum value: BMXWT= 195.4, RIDAGEYR= 80, BMXHT= 202.6, BMXWAIST= 172.5, BMXARML= 47.9, BMXLEG= 51.9, FEMALE=1

Mean: BMXWT= 73.91092, RIDAGEYR= 38.26664, BMXHT= 163.9536, BMXWAIST= 92.64025, BMXARML= 36.26701, BMXLEG= 38.58299, FEMALE= 0.5083047

Medium: BMXWT= 72.1, RIDAGEYR= 37, BMXHT= 164.6, BMXWAIST= 91.95, BMXARML= 36.5, BMXLEG= 38.6, FEMALE= 1

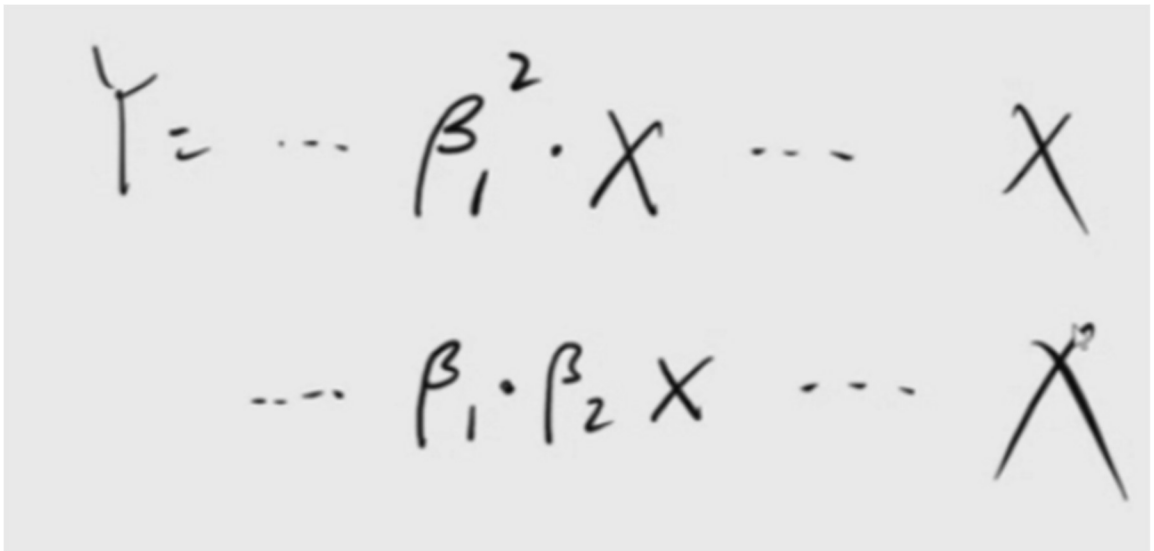
Variance: BMXWT= 588.774, RIDAGEYR= 471.8578, BMXHT= 166.2415, BMXWAIST= 361.4121, BMXARML= 12.905, BMXLEG= 16.33076, FEMALE= 0.2499556

Standard Deviation: BMXWT= 24.26467, RIDAGEYR= 21.72229, BMXHT= 12.89347, BMXWAIST= 19.01084, BMXARML= 3.592353, BMXLEG= 4.041134, FEMALE= 0.4999556

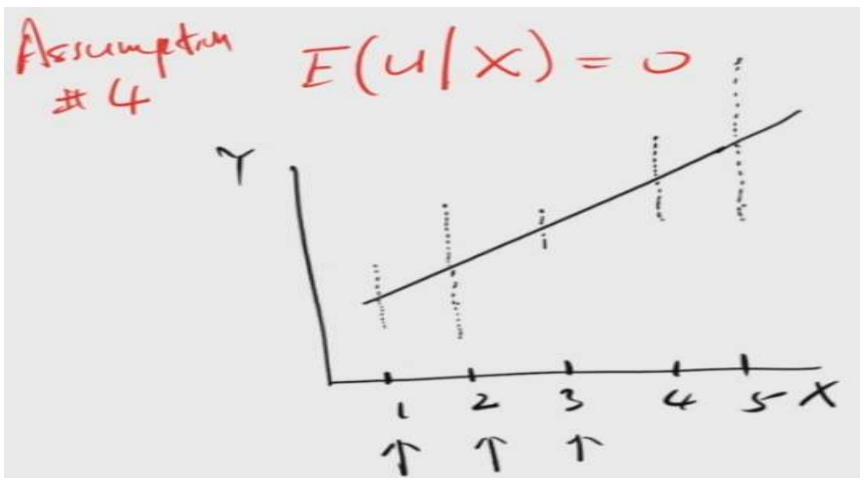
R-squared=0.9384, R-squared measures the goodness of the fit or how well it fits, when it fits well, it indicates there is a strong relationship between X and Y. The R-squared value in this regression is quite large, indicating a strong relationship between X and Y. This also translates to 93.84% of the total variation in Y is explained by this regression. Therefore, I think there are not any key variables that I am missing. For the x variable FEMALE, 0 stands for female, and 1 stands for female.

Assumptions of linear regression:

1. Linear in parameters: this means that there must be a linear relation between x and y. My regression qualified for this assumption because it is a linear graph and there is nothing that changes the coefficient like the following examples:



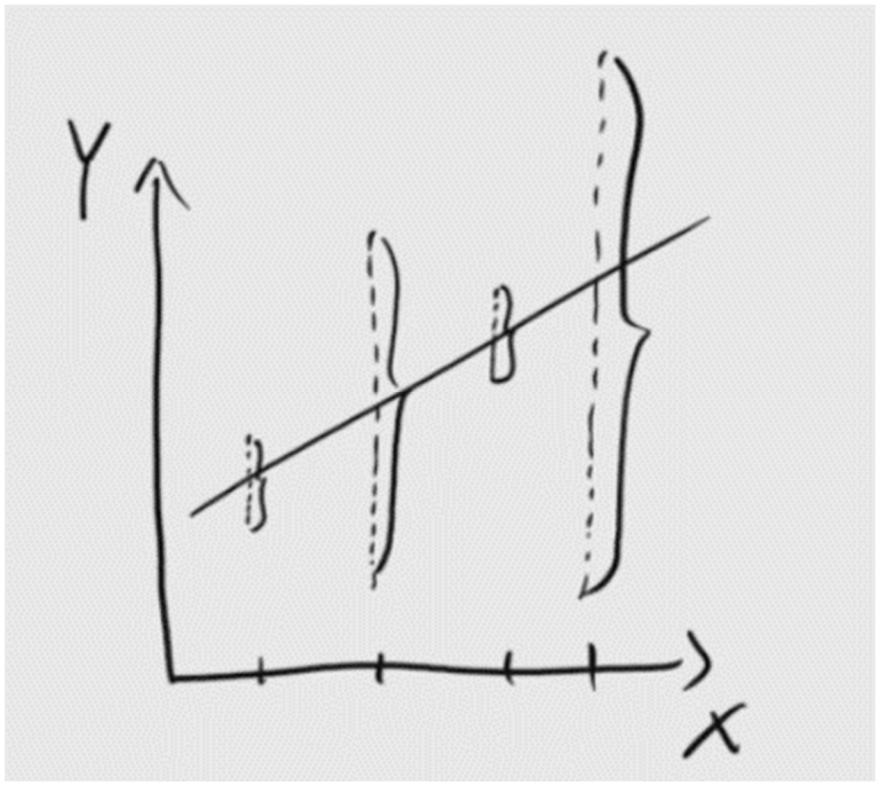

2. Random sampling: Sample must be a random representing the population. According to the website for codebook, I have not seen anywhere does it stated that the samples are randomly selected. Therefore, it is not clear if this assumption is met.
3. Sample Variation in the Explanatory Variable: You need some variation in X. If no variation, no regression line.
4. Zero Conditional Mean: the average of U given all the X values. The positive and negative of each residual(u) should cancel each other out, making sure that the regression line is in the middle. If not violated, picture looks as follow:




For assumption #4, even if it is violated, it does not violate the property. Therefore, I can still run a regression because if I add up all the residuals, they might still be 0. However, I can't say the average of all my estimate will be equal to the true population because assumption 4 is violated.

5. Homoskedasticity: equal variance of residuals (u), if violated, it is called heteroskedasticity. My regression is confirmed to be heteroskedasticity because I did a Breusch Pagan Method to test for heteroskedasticity with the F test.

I will be using Breusch Pagan Method to test for heteroskedasticity with the F test. First, I need to use the R software to run another regression with residual squared from the regression that I have formed instead of BMXWT, I will name this new regression reg2. The null hypothesis will be BMXWT= 0, RIDAGEYR= 0, BMXHT= 0, BMXWAIST= 0, BMXARML= 0, BMXLEG= 0, FEMALE=0. The alternative hypothesis will be BMXWT≠ 0, RIDAGEYR≠ 0, BMXHT≠ 0, BMXWAIST≠ 0, BMXARML≠0, BMXLEG≠ 0, FEMALE≠0. Then, I need to calculate the F statistics. According to the summary of reg2, R-squared=0.09004, k (number of X variables) =6,Denominator degree of freedom (n-k-1)= 7325. Therefore, F=[(0.09004)/6]/[(1-0.09004)/7325] ≈120.8007. Assuming 5% critical values of the F distribution, numerator degrees of freedom=6, Denominator degree of freedom=7325. Therefore, Fc=2.10. Since F statistic is greater than F critical value, we reject our null hypothesis, indicating it is heteroskedasticity. This is a problem because when the regression is confirmed to be heteroskedasticity, a hypothesis test for significance cannot be conducted. A heteroskedasticity looks like the following graph:




Heteroskedasticity means that I have a problem because the outliers have a heavier pull. In order to fix it, I need to do the weighted least squares estimation and what it does is basically removes the outliers. I will do that in the R studio with the command: coeftest(reg1,vcov=vcovHC(reg1,type="HC0")). After I fixed the heteroskedasticity with weighted least squares estimation, I get the following result:

```
t test of coefficients:

              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -1.0889e+02  1.1048e+00 -98.5633 < 2.2e-16 ***
RIDAGEYR    -1.6013e-01  4.5133e-03 -35.4796 < 2.2e-16 ***
BMXHT        3.5459e-01  1.4821e-02  23.9252 < 2.2e-16 ***
BMXWAIST     1.0954e+00  7.1636e-03 152.9169 < 2.2e-16 ***
BMXARML      4.4640e-01  5.4837e-02   8.1404 4.592e-16 ***
BMXLEG       3.4005e-01  3.8667e-02   8.7941 < 2.2e-16 ***
female       9.9473e-03  1.5744e-01   0.0632    0.9496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the result, there are changes in T value, standard deviation, and P value. We can see that the P value of FEMALE changed from 0.95 to 0.9496 and BMXARML changed from 2e-16 to 4.592e-16, every other x variables just changed from 2e-16 to 2.2e-16. Since the P value for RIDAGEYR, BMXHT, BMXWAIST, and BMXLEG are all 2.2e in the corrected regression summary, assuming alpha =0.05, since P value is less than alpha, we reject the null hypothesis, indicating these variables are statistically significant. As of BMXARML, assuming alpha=0.05, the P value is also less than alpha, so we reject the null hypothesis, indicating these variables are statistically significant. Lastly, the P value for FEMALE is 0.9496. Assuming alpha=0.05, P value is clearly larger than alpha, so we do not reject null hypothesis, indicating it is not significant. Since FEMALE is not statistically significant, it is recommended to remove this variable from the regression because keeping this variable might reduce the probability for a precise prediction. T value for RIDAGEYR changed from -1.601e-01 to -1.6013e-01, BMXHT changed from 3.546e-01 to 3.5459e-01, BMXWAIST changed from 1.095 to1.0954, BMXARML changed from 4.464e-01 to 4.4640e-01, BMXLEG changed from 3.400e-01 to 3.4005e-01, FEMALE changed from 9.947e-03 to 9.9473e-03. Standard error for RIDAGEYR changed from 4.022e-03 to 4.5133e-03, BMXHT changed from 1.392e-02 to 1.4821e-02, BMXWAIST changed from 5.111e-03 to 7.1636e-03, BMXARML changed from 4.813e-02 to 5.4837e-02, BMXLEG changed from 3.447e-02 to 3.8667e-02, FEMALE changed from 1.596e-01 to 1.5744e-01.

NHANES body measures data are taken in order to monitor trends in infant and their growth. With this data, it is possible to estimate the possibility for the children, teenagers, and adults in United States to be overweight. In the report, I will be analyzing how a person's weight are under the influence of gender, age, standing height, waist circumference, upper arm length, and upper leg length. The interpretation of this regression line for y-intercept will be even if all the other variables are 0, the weight in kg will be negative 108.9. This does not make sense because no one can have a negative weight. Therefore, the y intercept is not valuable. Holding all the other variables constant, the interpretation of this regression line for RIDAGEYR (Age in years at screening) will be when a person' age increase by 1 year old, his weight will decrease by 0.1601kg. This makes sense because the older a person gets, the less health that person becomes. Therefore, you get sick more often and get skinnier because of that. Holding all the other variables constant, the interpretation of this regression line for BMXHT (standing height in cm) will be when a person's standing height goes up by 1 cm, his weight will increase by 0.3546kg. This interpretation makes sense because the taller you are, the heavier you become. Holding all the other variables constant, the interpretation of this regression line for BMXWAIST (waist circumference in cm) will be when a person's waist circumference increases by 1 cm, his weight will increase by 1.095kg. This makes sense because the larger the waist circumference, the heavier that person becomes. Holding all the other variables constant, the interpretation of this regression line for BMXARML (upper arm length in cm) will be when a person's upper arm length in cm increases by 1, that person's weight will increase by 0.4464kg. This interpretation makes sense because the longer your arm length is, the heavier you become. Holding all the other variables constant, the interpretation of this regression line for BMXLEG (upper leg length in cm) will be when a person's upper leg length in cm increases by 1, the weight of that person will increase by 0.3400kg. This interpretation makes sense because the longer your leg length is, the heavier you become. Holding all the other variables constant, the interpretation of this regression line for FEMALE will be if a person is female, she will be 0.009947 or 0.9947 percent point higher weight than a male. This makes sense because some women take birth control pills, and this pill does have a temporary side effect of weight gaining due to fluid retention. On the other hand, there are not any birth control pills for men. However, you can also argue that not all women take birth control pills. Overall, out of all the variables listed in the regression line, BMXWAIST (Waist Circumference in cm) seem to have the largest influence on weight because it has the largest coefficient compare against other variables. Therefore, people with lower waist circumference in cm are less likely to be overweight.

Appendix

```
> library(haven)
> mydata=read_xpt(file.choose())
> demographic=mydata
> mydata=read_xpt(file.choose())
> blood_measures=mydata
> View(blood_measures)
> View(demographic)
> View(demographic)
> attach(demographic)
> female=ifelse(RIAGENDR==1,0,1)
> demographic2=cbind(demographic,female)
> demo_merge_blood=merge(demographic2,Blood_measures)
>demo_merge_blood2=subset(demo_merge_blood,RIDAGEYR!=""&BMXWT!=""&BMXHT!=""&BMXWAIST!=""&BMXARML!=""&BMXLEG!=""&female!="")
>reg1=lm(BMXWT~RIDAGEYR+BMXHT+BMXWAIST+BMXARML+BMXLEG+female,data=demo_merge_blood2)
> summary(reg1)

Call:
lm(formula = BMXWT ~ RIDAGEYR + BMXHT + BMXWAIST + BMXARML +
    BMXLEG + female, data = demo_merge_blood2)

Residuals:
    Min      1Q  Median      3Q     Max
-32.166  -3.807  -0.188   3.478  41.981

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -1.089e+02  1.065e+00 -102.279   <2e-16 ***
RIDAGEYR    -1.601e-01  4.022e-03  -39.812   <2e-16 ***
BMXHT        3.546e-01  1.392e-02   25.464   <2e-16 ***
BMXWAIST     1.095e+00  5.111e-03  214.347   <2e-16 ***
BMXARML      4.464e-01  4.813e-02    9.274   <2e-16 ***
BMXLEG       3.400e-01  3.447e-02    9.864   <2e-16 ***
female       9.947e-03  1.596e-01    0.062     0.95
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.024 on 7325 degrees of freedom
Multiple R-squared:  0.9384,    Adjusted R-squared:  0.9384
F-statistic: 1.86e+04 on 6 and 7325 DF,  p-value: < 2.2e-16


> u2=residuals(reg1)^2
>reg2=lm(u2~RIDAGEYR+BMXHT+BMXWAIST+BMXARML+BMXLEG+female,data=demo_merge_blood2)
> summary(reg2)
```

```
Call:
lm(formula = u2 ~ RIDAGEYR + BMXHT + BMXWAIST + BMXARML + BMXLEG +
    female, data = demo_merge_blood2)

Residuals:
    Min      1Q  Median      3Q     Max
-155.67  -29.72  -12.87    9.42 1674.94

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.05036   12.86377  -1.714   0.0865 .
RIDAGEYR     -0.36690    0.04860  -7.550 4.89e-14 ***
BMXHT        -0.74652    0.16825  -4.437 9.26e-06 ***
BMXWAIST      1.44042    0.06175  23.328  < 2e-16 ***
BMXARML      -0.11560    0.58157  -0.199   0.8425
BMXLEG        1.66491    0.41654   3.997 6.48e-05 ***
female        2.45441    1.92838   1.273   0.2031
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72.79 on 7325 degrees of freedom
Multiple R-squared:  0.09004,   Adjusted R-squared:  0.0893
F-statistic: 120.8 on 6 and 7325 DF,  p-value: < 2.2e-16
```

> library(lmtest)

> library(sandwich)

>attach(reg1)

>coeftest(reg1,vcov=vcovHC(reg1,type="HC0"))

```
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -1.0889e+02  1.1048e+00 -98.5633 < 2.2e-16 ***
RIDAGEYR    -1.6013e-01  4.5133e-03 -35.4796 < 2.2e-16 ***
BMXHT        3.5459e-01  1.4821e-02  23.9252 < 2.2e-16 ***
BMXWAIST     1.0954e+00  7.1636e-03 152.9169 < 2.2e-16 ***
BMXARML      4.4640e-01  5.4837e-02   8.1404 4.592e-16 ***
BMXLEG       3.4005e-01  3.8667e-02   8.7941 < 2.2e-16 ***
female       9.9473e-03  1.5744e-01   0.0632    0.9496
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> attach(demo_merge_blood2)

> min(BMXWT)

[1] 16.9

> min(RIDAGEYR)

[1] 8

> min(BMXHT)

[1] 113

> min(BMXWAIST)

[1] 47.8

> min(BMXARML)

[1] 22.7

> min(BMXLEG)

[1] 24.4

> min(female)

[1] 0

> max(BMXWT)

[1] 195.4

> max(RIDAGEYR)

[1] 80

```
> max(BMXHT)
[1] 202.6
> max(BMXWAIST)
[1] 172.5
> max(BMXARML)
[1] 47.9
> max(BMXLEG)
[1] 51.9
> max(female)
[1] 1
> mean(BMXWT)
[1] 73.91092
> mean(RIDAGEYR)
[1] 38.26664
> mean(BMXHT)
[1] 163.9536
> mean(BMXWAIST)
[1] 92.64025
> mean(BMXARML)
[1] 36.26701
> mean(BMXLEG)
[1] 38.58299
> mean(female)
[1] 0.5083047
> median(BMXWT)
[1] 72.1
> median(RIDAGEYR)
[1] 37
> median(BMXHT)
[1] 164.6
> median(BMXWAIST)
[1] 91.95
> median(BMXARML)
[1] 36.5
> median(BMXLEG)
[1] 38.6
> median(female)
[1] 1
> var(BMXWT)
```

```
[1] 588.774

> var(RIDAGEYR)

[1] 471.8578

> var(BMXHT)

[1] 166.2415

> var(BMXWAIST)

[1] 361.4121

> var(BMXARML)

[1] 12.905

> var(BMXLEG)

[1] 16.33076

> var(female)

[1] 0.2499556

> sd(BMXWT)

[1] 24.26467

> sd(RIDAGEYR)

[1] 21.72229

> sd(BMXHT)

[1] 12.89347

> sd(BMXWAIST)

[1] 19.01084

> sd(BMXARML)

[1] 3.592353

> sd(BMXLEG)

[1] 4.041134

> sd(female)

[1] 0.4999556
```