This project looks at correlation. Before beginning, download the following data set:

```
> source("http://wiener.math.csi.cuny.edu/st/R/fat.R")
```

This loads a data set with 19 measurements for 252 subjects. The purpose of the data set is to explore relationships that can be used to predict a person's body fat. The body fat is an excellent measure of health – more informative that than the BMI – but isn't used so much, as it is relatively difficult to calculate. The variables, `body.fat` and `body.fat.siri` are measurements of bady fat taken two different ways.

Scatterplots

The idea here is to find variables that are easy to measure that are related with the body fat. These variables are **not** independent, and should be treated in groups, such as pairs. The standard graphic to look at the relationship between two numeric variables with paired values (from the same subject) is the scatterplot. That is, we plot (x, y) values as points. In R we use the `plot` command to make the scatterplot. In the command below, we use the formula notation to produce Figure 1.

```
> plot(weight ~ height, data = fat)
```

The `y` variable (`weight`) is often referred to as the *response* variable in statistics and the `x` variable a *predictor* variable. The basic idea we will want to pursue is to see how well the predictor variable *predicts* the *mean* value of the y values associated to a given x value.

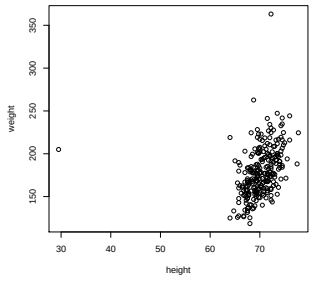


Figure 1: A scatterplot of weight and height

Question 0.1. Look at the scatterplot of weight versus height. As height increases, in general does weight increase or decrease, or not depend on height?

Question 0.2. Is it true that if person A is taller than person B that person A will weigh more than person B?

Question 0.3. Is it true that *on average* if person A is taller than person B then person A will weigh more than person B?

Question 0.4. The plot in Figure 1 shows an outlier at (30,200). It can be removed many ways, but a simple one is to subset by looking at heights 50 inches or greater with:

```
> plot(weight ~ height, data = fat, subset = height > 50)
```

Make this plot, describe any differences.

Question 0.5. Make a plot of `BMI` versus `body.fat`. Does one increase (decrease) as the other increases?

Question 0.6. For your graph of `BMI` and `body.fat` can you describe the pattern as a *linear trend*? Or does the *trend* seem to curve? (A *trend* is basically how one describes the average value of the y variable as the x variable increases.)

Question 0.7. Make a plot of `body.fat` and `age`. Does one increase (decrease) as the other increases?

Question 0.8. Remake the plot with `weight` as the response and `height` as the predictor variable. We identified the point (29.5,205) – case 42 – as an outlier. But why? A *rough* idea of an outlier is

> An outlier is any value far from the trend of the data

However, this gives many different ideas of an outlier:

1. An outlier in the bivariate sense is a value that does not fit the overall pattern of the data. In this example, the overall pattern is an increase in weight as height gets larger. In this case, we'd expect from eyeballing a much lower height for a 29.5 inch person (clearly this is a typo in the height, I'd guess it was supposed to be 69.5.)

2. An outlier in the univariate sense. Here "trend" means the center of the data. If we were to make a boxplot of the height variable, we'd see that this value of 29.5 stands out. Check:

   ```
   > boxplot(fat$height)
   ```

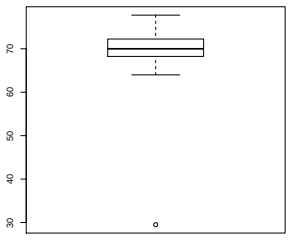
   

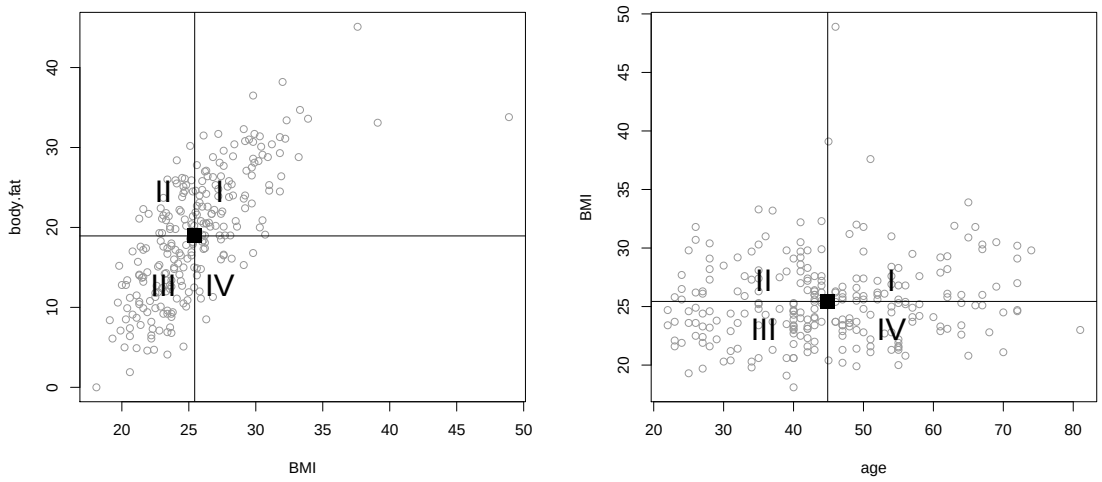
   Figure 2: Boxplot of height showing outlier.

   However, the value 205 is not an outlier in the univariate sense for weight.

Make a plot of `BMI` as a response variable with `weight` as a predictor variable. Is case 39 with a value of (363.15,48.9) an outlier in the bivariate sense? In the univariate sense for `weight`? In the univariate sense for the variable `BMI`?

Correlation

The correlation of two variables numerically measures the idea that as one value gets big the other gets big (or small). The two figures show two scatterplots. For each the mean of the response and predictor is drawn with a line. For the left



scatterplot, there are few values in II and IV – points with large x values (larger than the mean) *typically* have large y values (larger than the mean). This is not so with the right scatterplot

The correlation quantifies this. First we look at this

$$C = \sum (x_i - \bar{x})(y_i - \bar{y})$$

Question 0.9. For a figure, like that in the left side (`BMI` versus `body.fat`), will C be positive or negative – why? Can you say the same for the right side? Why or why not?

The correlation is similar to C above, only it first forms z-scores, and then averages

$$\text{correlation} = \frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

The correlation is close to 1 when the values more or less fall on a line with positive slope, close to -1 when the values fall on a line with a negative slope and close to 0 when scattered.

To compute the correlation, we have the `cor` function:

```
> with(fat, cor(body.fat, BMI))

[1] 0.7279942

> with(fat, cor(age, BMI))

[1] 0.1188513
```

Question 0.10. Find the correlation for `wrist` and `neck`; for `height` and `weight`; and for `wrist` and `ankle`. Write down the values.

Question 0.11. Which of the the scatterplots in Figure 3 shows the largest correlation? The smallest. (Points are more correlated if they fall closest to a line.) After guessing you can check your answer with one (long) command:

```
> vars <- c("wrist", "neck", "abdomen", "weight", "body.fat")
> cor(subset(fat, select = vars))
```

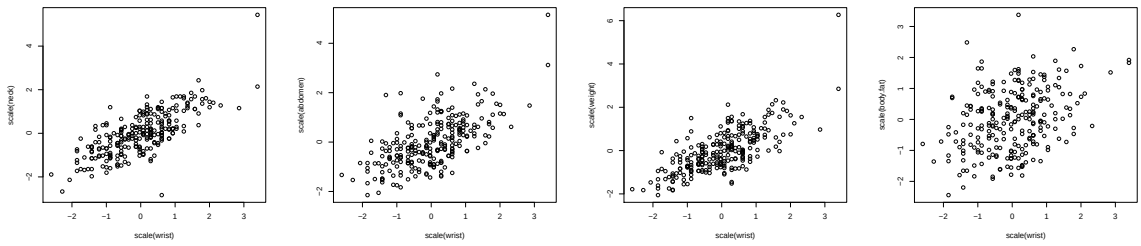


Figure 3: Which shows the greatest correlation?

Trend lines

For paired data that shows a trend, it is natural to try to draw a line that captures that trend. The simple regression line, is a straight line that attempts to capture the mean y value for a given x value. A line needs two numbers – a slope and intercept, or point and slope to be drawn.

The least squares regression line is one for which the intercept and slope are found by minimizing the squared residuals from a line. (A residual is just observed minus expected, or the difference in the y value between the line and the point for a given x value.)

These values are found with the `lm` function:

```
> lm(body.fat ~ BMI, data = fat)

Call:
lm(formula = body.fat ~ BMI, data = fat)

Coefficients:
(Intercept)          BMI
    -20.405        1.547
```

The slope is 1.547, the intercept is -20.405. To add this line to our plot, we have the `abline` command. Here are the steps.

```
> plot(body.fat ~ BMI, data = fat)
> res <- lm(body.fat ~ BMI, data = fat)
> abline(res)
```

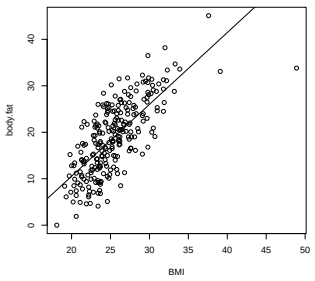


Figure 4: scatter plot with regression line

Question 0.12. Find the slope of the regression line for `wrist` and `neck`; for `height` and `weight`; and for `wrist` and `ankle`. Write down the values and compare to the correlations you found. Are there any similarities?

Question 0.13. Verify that it does not matter if you interchange the response and predictor for correlation, but that it does for the regression line slope by looking the values for `BMI` and `body.fat`.

Question 0.14. The regression line can be influenced by outliers – but it need not be. When a point has an influence, we call it an influential point. To see if the outlier in the height/weight data is influential we will fit two models – one with and one without the point:

```
> res <- lm(weight ~ height, data = fat)
> res.1 <- lm(weight ~ height, data = fat, subset = height > 30)
> plot(weight ~ height, data = fat)
> abline(res, col = "blue")
> abline(res.1, col = "red")
```

Run the above commands. Is the regression line seriously affected by the presence of the one case? Comment