We will use pmg in this project, please load it.

Let's look at data set stored in the body `fat` data set available online through:

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

This data set has several measurements of body parts for many different subjects. The idea is the body fat index is hard to compute – it requires a trained user of calipers or a tank of water. Yet it is an important measurement of health. As such, one looks into variables which are naturally correlated with body fat to see if one can predict the body fat from these easily measured variables. The available values (mostly circumference measurements) are:

```
> names(fat)
```

```
 [1] "case"          "body.fat"      "body.fat.siri" "density"
 [5] "age"           "weight"        "height"        "BMI"
 [9] "ffweight"      "neck"          "chest"         "abdomen"
[13] "hip"           "thigh"         "knee"          "ankle"
[17] "bicep"         "forearm"       "wrist"
```

1 Plotting scatterplots, adding a regression line

How do we use pmg to make a scatterplot? The `lattice explorer` is an easy way. Open this under the plots menu. To make a scatterplot, select `xyplot` and then drag first a y variable and then an x variable. A graphic should be produced. A least squares regression line can be added by the drop-down box in the upper right. Select `lmline`.

If you drag a third variable, plots for each level of that variable are made. To make a new plot, use the `clear` button in the upper left.

1. Make plots with `BMI` as the *response* variable (y) for different predictors: `Age`, `wrist`, `height`. Which pairs of variables appear to be strongly correlated? Is this positive or negative correlation?

2. The BMI measures overall fitness. Values more than 30 are considered overweight. The waist size is a simpler measure of overall fitness. Values over 100cm (the units used) are considered overweight. Make the scatterplot of both and see if these two values tend to correlate with each other. Comment.

3. Make a scatterplot of `neck` and `wrist` size. Do the two variables appear to be strongly correlated?

4. Under the `Plots::Multivariate` menu is the `pairs` dialog. Open this up, then drag the data set `fat` into the data area. Click OK. What kind of graphic is produced? You may need to make it larger to see the detail. Which pairs of variables seem highly correlated?

2 Finding regression lines

To find the regression coefficients is done many different ways. Open the `Models::Dynamic models` dialog. Drag `BMI` to the response and drag `age` to the predictor area. A summary of the model fit is presented. You should see this about half way down:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.89261    0.84739  28.196   <2e-16 ***
fat$age      0.03441    0.01818   1.893   0.0596 .
```

This says that the estimate for the intercept is 23.8 and the estimate for the slope is 0.03441, so the regression line becomes

$$\hat{y} = 23.89261 + 0.03441 \cdot \text{age}$$

To change the predictor variable can be done two ways. First, one can clear the formul using the action "clear formula." This requires you to drag over the response and the new predictor again. Otherwise, one can edit the predictor: click on it and after the dollar sign type the variable name; when done hit enter.

1. Fit the model with `wrist` predicting `neck`. What are the coefficients?

2. Use your linear model to predict the neck size of a person with an 18cm wrist circumference.

3. Fit the model with `neck` predicting `abdoment`. What are the coefficients?

4. Use your linear model to predict the neck size of a person with an 40cm neck circumference.

3 The correlation

The output of the linear model also includes the correlation, in this example between BMI and wrist circumference, in the line

```
Multiple R-squared: 0.3918,Adjusted R-squared: 0.3893
```

We can ignore the Adjusted value, as it applies to the case with more than one predictor variable (multiple regression) and look at the value for R^2 given. The correlation is R. If you want to get this from R^2 you need to a) take the square root and b) make sure it has the same sign as the slope.

1. What is the correlation between `BMI` and `abdomen`?

4 Linear model

The simple linear regression model for a paired data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

If we assume the ε_i are a *random sample* from a mean 0 normal distribution with variance σ^2, then there are 3 paramters to estiamte from the data: β_0, β_1, and σ. I'll use hats to indicate the estimates based on the sample, e.g. $\hat{\beta}_1$ is the estimate for the unknown value of β_1.

For the β's we have the following fact about the *sampling distribution*

$$T = \frac{\text{obs} - \text{exp}}{\text{exp}} = \frac{\hat{\beta} - \beta}{\mathsf{SE}}$$

has a t distribution with $n-2$ degrees of freedom. This fact allows us to do

1. Confidence intervals for either β_0 or β_1:

$$\hat{\beta}_i \pm t^* \mathsf{SE}(\hat{\beta}_i)$$

2. Significance tests about either use T as a test statistic.

Now, how to find the estimates and the standard errors using the computer?

The full output of a linear model fit contains more than previously described. For instance for a model of `neck` size predicted by `wrist` size, we have:

1. The "call" which just echoes your command
2. A summary of the Residuals. These have a median of -0.01920. The mean is 0, why?
3. The important summary of the coefficients. This is where the action is. More below
4. A summary of the standard error. The value 1.625 is found from

$$\sqrt{\frac{1}{n-2}\sum e_i^2}$$

   Or found within R (at the command line)

```
sqrt(sum(resid(res)^2)/(252-2))
```

5. That value of `R-squared` might sound familiar, especially if you were to type `cor(neck, wrist)^2`

The Coefficients have not only the estimates, but also the standard errors computed. Additionally a two-sided significance test for whether β is 0 is performed and summarized with an observed value (`t value`) and a p-value, `Pr(>|t|)`.

From the output we see that the estimate $\hat{\beta}_0 = 2.6370$ is *not* statistically significant from 0. (Why?) Yet the smaller value $\hat{\beta}_1 = 1.9364$ *is* statistically significant from 0.

What about the natural question prompted by the following passage from *Gulliver's Travels* by Jonathan Swift (in the giant's voice)

> Then they measured my right Thumb, and desired no more; for by a mathematical Computation, that twice round the Thumb is once round the Wrist, and so on to the Neck and the Waist, and by the help of my old Shirt, which I displayed on the Ground before them for a Pattern, they fitted me exactly.

That is, in the language of MTH 214, is the data consistent with

$$H_0 : \beta_1 = 2, \quad H_A : \beta_1 \neq 2?$$

Here we can use T as the test statistic. The observed value is

```
> Tobs = (1.9364 - 2)/0.1099
> Tobs
```

```
[1] -0.5787079
```

The p-value is computed with the t-distribution and $n-2$ degrees of freedom:

```
> pt(Tobs, df = 250)

[1] 0.2816536
```

That is the difference between the 1.9364 and the hypothesized 2 is not statistically significant.

The `pt` function is accessed through the `Data::Random data::Cumulative probabilities` menu item, or can be typed in directly at the command line.

Okay, your turn.

1. Use T to find a 95% confidence interval for β_0 based on $\hat{\beta}_0$. You can use $t^* = 1.96$ (why?) or see what it is yourself with `qt(.975, df = 250)`.

2. Is twice around the neck once around the waist? Let's model by the `hip` values using the `waist` values as a predictor.

   Do two-sided tests of the following

   $$H_0: \quad \beta_0 = 0$$
   $$H_0: \quad \beta_1 = 2$$

   What are your p-values?

3. The BMI is a simple measurement of fitness: weight divided by height squared. Larger BMIs mean a person is heavier than taller. The body fat is a much more precise measurement of fitness, but is much harder to measure. Does the BMI predict the body fat well? and if so, how does one convert?

   Model the variable `body.fat` using `BMI` as a predictor.

   Find a 95% CI for β_0.

   Perform a two-sided significance test of

   $$H_0: \quad \beta_1 = 1.5$$

   Comment on this proposed relationship between BMI and body fat:

   > To compute body fat from BMI do the following triple the BMI, divide by 2 and subtract 20.

4. Does body fat depend on height? Model body fat using height as a predictor and perform a one-sided significance test of $H_0 : \beta_1 = 0$ against $H_A : \beta_1 < 0$. Is the difference significant at the $\alpha = 0.05$ level?

5. For a model of ankle size modeled by wrist size find 95% CIs for β_0 and β_1. Is 0 in the first one? Is 1 in the latter? If so, comment on the summary that wrist and ankle sizes are the same *on average*.