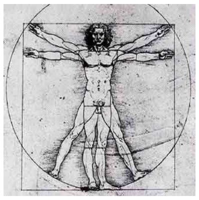

Human Proportions

1 The shape of body part measurements

The human body comes in various shapes and sizes. However, as daVinci knew, there are certain proportions that are consistent throughout. For this project two data sets are used which contain various measurements of human bodies.

To download the data sets issue these commands:

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

The `normtemp` data set[1] contains measurements of normal body temperature for 300 healthy adults in the variable `temperature`. The variable `gender` records the gender of the subject, and `hr` the heart rate in beats per minute.

The `fat` data set[2] contains many measurements of human bodies that can be done with a tape measure (circumference measurements), for instance the variable `wrist` contains measurements of wrist size in centimeters. Additionally, the variable `body.fat` contains body fat measurements.

After downloading the data sets, they may be attached so that the variable names are visible from the command line.

```
> attach(normtemp)
> attach(fat)
```

2 Paired data

In *Gulliver's Travels*, Jonathan Swift wrote (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.

This implies that data on wrist size and neck size for the same person should be jointly related in some manner. Are such relationships actually the case for the human body? The `fat` data set allows us to investigate to some degree.

[1] This data set was contributed to the *Journal of Statistical Education* by Allen L. Shoemaker, `http://www.amstat.org/publications/jse/v4n2/datasets.shoemaker.html`

[2] This data set was contributed to the *Journal of Statistical Education* by Roger W. Johnson, `http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html`.

Question 1: Use some measuring device (a sheet of paper may work) and see if twice around your thumb is roughly once around your wrist. Then use one hand to measure once around your wrist. Compare to using both hands to measure once around the neck.

2.1 Viewing paired data

Paired numerical data is often viewed with a scatterplot. These are produced using the `plot()` function. This function has a few different ways it can be used.

For instance, in the `fat` data set, we can plot corresponding `wrist` and `neck` measurements with any of these

seperated by a comma As `fat` is attached, we can refer to these variables by their names.

```
plot(wrist, neck)
```

Using the model formula If we think of the neck size being determined by the wrist size, then we might want to think in terms of R's model formula notation. This puts the dependent variable on the left of a tilde, ~, and the independent variable(s) on the right. Eg.

```
plot(neck ~ wrist)
```

Attaching the data set temporarily When the data set is not attached, the model formula notation allows one to briefly attach the data using the `data=` argument:

```
plot(neck ~ wrist, data=fat)
```

As well, the argument `subset=` can be used with a logical expression to reduce the number of points plotted. This example uses only subjects with wrist size more than 19cm.

```
plot(neck ~ wrist, subset=wrist > 19)
```

The first two examples will produce Figure 1

Question 2: For the data in `normtemp`, make a scatterplot of heart rater `hr` versus `temperature`. Does there appear to be a relationship?

Question 3: The following produces a plot for the just the males in the `normtemp` data set:

```
> plot(temperature ~ hr, subset = gender == "Male")
```

Make this plot. Then add the points for females. This can be done with

```
> points(temperature ~ hr, subset=gender=="female", pch=2)
```

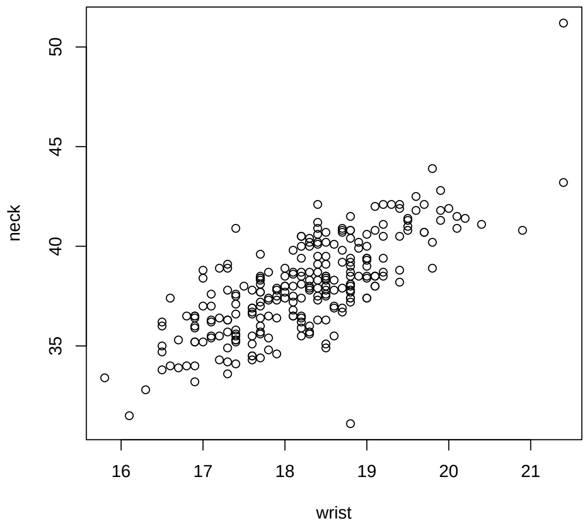

Figure 1: Scatterplot of neck size versus wrist size

Do the two scatterplots tell the same story? A different story? Explain.

The term *regression* was used by Galton in papers written in the 1880s. One of the data sets he was interested in contained data on the joint distribution of stature for the parents and adult children. The data set `father.son` [3] contains similar data. Download and attach the dataset with the command

```
> source("http://www.math.csi.cuny.edu/st/R/father.son.R")
> attach(father.son)
```

The variable `sheight` records the son's height, and `fheight` the father's height. You can use `attach()` so that the variable names are readily accessible.

Question 4: Make a scatterplot of the `father.son` data using the `fheight` variable to predict the `sheight` variable. Does there appear to be a relationship? Would you say it was a strong relationship?

2.2 Linear models

Consider the plot of wrist versus neck size in Figure 1. Although there is a bit of scatter, one could effectively summarize the trend in the data by a straight line

When two variables are related and it appears that their relationship can be summarized by a line running from a lower left point of, say, $(16, 35)$ and an upper right point of $(21, 45)$. That is a slope of $10/5 = 2$ with equation $(y - 16) = 2(x - 35)$.

[3] from `http://stat-www.berkeley.edu/users/juliab/141C/pearson.dat`

When such a *linear relationship* appears, we can summarize the strength of the relationship using a number called the *Pearson correlation coefficient*. The definition can be written as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{1}{n-1} \sum_i \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}.$$

The latter expression writes this, basically as the average product of the z-scores for x and y. These center and scale the data around $(\bar{x}, \bar{y})$.

This number is returned by the function `cor()`. For example, the correlation between wrist circumference and neck circumference is given by

```
> cor(wrist, neck)

[1] 0.7448264
```

Values of r closer to 1 or -1 indicate that the data more closely track as straight line.

Question 5: What is the correlation between `hr` and `temperature` for the `normtemp` data? Is the value positive or negative?

Question 6: What is the correlation between `sheight` and `fheight` for the `father.son` data set. Is this value close to 1 or -1?

Question 7: What is the correlation between `wrist` and `bicep` circumference measurements in the `fat` data set?

Question 8: Compare the values of r found in the last three exercises with their scatterplots. Summarize what is different for the plots with larger values of r.

2.3 Simple linear regression models

A statistical model to describe a linear relationship between wrist circumference (`wrist`) and neck circumference (`neck`) would be

$$\texttt{neck} = \beta_0 + \beta_1 \texttt{wrist} + \epsilon,$$

where β_0 is the y-intercept, β_1 the slope, and ϵ indicates an error term. Generically, we might write the model, using an i to keep track of which data point, as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

The parameters, β_0 and β_1 are estimated using the *method of least squares* by the `lm()` function. The estimates are usually denoted using a "hat," as in $\hat{\beta}_0$ and $\hat{\beta}_1$.

The `lm()` function uses the model formula syntax. In this case, the model is specified by `neck ~ wrist`:

```
> lm(neck ~ wrist)
```

```
Call:
lm(formula = neck ~ wrist)

Coefficients:
(Intercept)        wrist
      2.637        1.939
```

The estimated relationship for the mean neck size for a given wrist size is

$$\hat{y} = 2.64 + 1.94 \cdot \texttt{wrist}.$$

The value of 1.94 is "close to" but not exactly 2. Is the difference statistically significant?

Question 9: Find the estimated regression line for the relationship between `hr` and `temperature` in the `normtemp` data set. Use `hr` as the predictor (or independent) variable.

Question 10: Find the estimated regression line for the relationship between `neck` and `abdomen` in the `fat` data set. Use `neck` as the predictor variable. How close is $\hat{\beta}$ to 2?

Question 11: Find the equation for the regression line for the model of father's height, `fheight`, predicting son's height, `sheight`.

2.4 Plotting the regression line

The regression line is added to the scatterplot using `abline()` and the output of `lm()`. For instance to add the regression line to Figure 1 would be done with

```
> plot(neck ~ wrist, data = fat)
> res = lm(neck ~ wrist, data = fat)
> abline(res)
```

Question 12: Add a regression line to your scatterplot of `hr` predicting `temperature` for the `normtemp` data set.

Question 13: Add a regression line to your scatterplot of `fheight` and `sheight` for the `father.son` data set.

3 Prediction

We can use the regression line to make predictions. For the model with normally distributed and independent error terms, the prediction line at a given value of x can be used to predict either the mean value of many samples (really the population mena) for this value of x, denoted $\mu_{y|x}$. Or, the value of a single observation of y for a given value of x.

Predictions can done directly from the formula for the regression line, or using the `predict()` function. For instance, the formula for the regression line of the `neck` circumference modeled by `wrist` circumference was found to be

$$\hat{y} = 2.64 + 1.94\texttt{wrist}.$$

So a person with a 19-centimeter wrist would have a predicted neck size of

```
> 2.64 + 1.94 * 19
[1] 39.5
```

Or 39.5 centimeters.

Question 14: What size neck would be predicted for a person with a 20-centimeter wrist size?

Question 15: For the model of son's height versus father's height, what is the predicted mean heights of the sons whose father are 70 inches tall.

3.1 Predicting body fat

The `fat` data set is intended to illustrate that a person's body fat percentage can be measured fairly well with simple measurements. To actually find a person's body fat percentage, the person must be weighed in water and have this compared to a weight in air. This makes the calculation difficult. Another less precise measurement involves the use of a caliper, requiring training on the part of the measurer. Wouldn't it be better if the measurement of body fat could be made using simple, unambiguous measurements of the body using a scale and a tape measure?

For instance, the BMI, or body-mass index, is a ratio of weight to height squared in metric units. It is widely used to assess obesity, although many argue that the cut offs used are not appropriate. (A November 30, 2004 letter to the *New York Times* mentioned that Alex Rodriquez, with a BMI of 26.5, would be considered overweight.)

Question 16: The variable `BMI` records the body-mass index for each subject in the `fat` data set. The variable `body.fat` variable the body fat percentage. Fit a linear model using `BMI` to predict `body.fat`, then make a prediction for a person with a BMI of 30.

Question 17: Another December 4, 2004 letter writer to the *New York Times*, mentions that an easy way to measure body fat is simply to measure the waist. (Although this does not account for the relationship of waist size to height.) The variable `abdomen` records waist size in centimeters.

The writer proposes that a waist size of greater than 40 inches for a male is high (35 inches for a female). Use a linear model to predict the body-fat percentage of a person with a 40-inch waist.

Question 18: As wrist size is related in some way to many other variables, is it possible to predict the body fat percentage from a wrist measurement? Make a scatterplot of `wrist` and `body.fat`. If a linear model seems appropriate, find the predicted body fat percentage for a person with an 18.6 centimeter wrist size. (18 cms = 18/2.54 ins)

3.2 Statistical inferences

The linear model

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

uses the term ϵ_i to incorporate error into the data. When assumptions are placed on the distribution of the error terms statistical inference can be made. We will assume the error terms are independent of each other (and the x variable) and normally distributed with mean 0 and common variance σ^2.

With these assumptions, the following have t-distributions

$$\frac{\hat{\beta}_0 - \beta_0}{\mathbf{SE}(\hat{\beta}_0)}, \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\mathbf{SE}(\hat{\beta}_1)}$$

The standard errors are computed in the output of the `summary()` of `lm()`.

For instance, the linear model

$$\texttt{sheight} = \beta_0 + \beta_1 \texttt{fheight} + \epsilon_i$$

has the following summary:

```
> res = lm(sheight ~ fheight, father.son)
> summary(res)

Call:
lm(formula = sheight ~ fheight, data = father.son)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8772 -1.5144 -0.0079  1.6285  8.9685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   33.887      1.832    18.5   <2e-16 ***
fheight        0.514      0.027    19.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.44 on 1076 degrees of freedom
Multiple R-Squared: 0.251,      Adjusted R-squared: 0.251
F-statistic:  361 on 1 and 1076 DF,  p-value: <2e-16
```

The value of $\mathbf{SE}(\hat{\beta}_0)$ is 1.832, and $\mathbf{SE}(\hat{\beta}_1) = 0.027$.

Significance tests

Standard errors can be used to perform significance tests. For the father-son model, it might seem intuitive that $\beta_1 = 1$. A test of the hypotheses

$$H_0 : \beta_1 = 1, \qquad H_A : \beta_1 \neq 1$$

can be carried out as follows.

```
> t.obs = (0.514 - 1)/0.027
> 2 * pt(t.obs, df = length(fheight) - 2)

[1] 1.586776e-63
```

The small p-value puts much doubt on the intuitive assumption that $\beta_1 = 1$.

Question 19: For the model of wrist size predicting neck size, test the null hypothesis

$$H_0 : \beta_1 = 2, \qquad H_A : \beta_1 \neq 2$$

What is the p-value? Do you reject at the $\alpha = 0.05$ level?

Question 20: For the model of neck size predicting abdomen size, test the null hypothesis

$$H_0 : \beta_1 = 2, \qquad H_A : \beta_1 \neq 2$$

What is the p-value? Do you reject at the $\alpha = 0.05$ level?

Question 21: For the model of `hr` predicting `temperature` test the null hypothesis

$$H_0 : \beta_1 = 0, \qquad H_A : \beta_1 \neq 0$$

What is the p-value? Do you reject at the $\alpha = 0.05$ level? Then look at the full output of `summary()` to see if you can find your p-value.