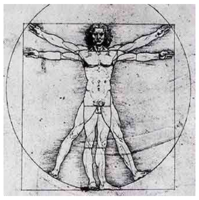

Human Proportions

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)
```

1.1 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

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

[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`.

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 1: 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 2: 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 3: 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.

Confidence intervals

Confidence intervals for β_0 and β_1 found with, for example,

$$\hat{\beta}_1 \pm t^* \mathbf{SE}(\hat{\beta}_1)$$

where t^* is related to the value of α by $P(-t^* < T_{n-2} < t^*) = 1 - \alpha$.

For instance, a confidence interval for the value of β_1 in the father-son model is produced with

```
> alpha = 0.05
> tstar = qt(1 - alpha/2, df = length(fheight) - 2)
> 0.514 + c(-1, 1) * tstar * 0.027

[1] 0.4610214 0.5669786
```

Question 4: Find a 90% confidence interval for the value of β_1 in the model of neck size modeled by wrist size using the data in the `fat` data set.

Question 5: Find a 90% confidence interval for the value of β_1 in the model of abdomen size modeled by neck size using the data in the `fat` data set.

Question 6: Find a 90% confidence interval for the value of β_1 in the model of body fat percentage modeled by BMI using the data in the `fat` data set.

1.2 Assessing the linear model

The simple regression model makes distributional assumptions on the error terms ϵ_i. The residuals, $e_i = y_i - \hat{y}_i$ should reflect these, although e_i is not an estimate for ϵ_i. By looking at the residuals we can assess whether the linear model is an appropriate one for the data. Graphs of the residuals are produced by applying `plot()` to the output of `lm()`.

For instance, the following commands produce four diagnostic plots:

```
> res = lm(sheight ~ fheight, father.son)
> par(mfrow = c(2, 2))
> summary(res)


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

Residuals:
      Min        1Q    Median        3Q       Max
-8.877151 -1.514415 -0.007896  1.628512  8.968479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.88660    1.83235   18.49   <2e-16 ***
fheight      0.51409    0.02705   19.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-Squared: 0.2513,       Adjusted R-squared: 0.2506
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

(The command `par(mfrow=c(2,2))` forces all four graphs to apear in same figure.)

The graphs are

Residuals vs. fitted This plots the fitted values $\hat{y}_i$ versus e_i. An appropriate model should show no trend.

Normal Q-Q plot The residuals are roughly speaking normally distributed sample. If this is so, then this graph should appear linear.

Scale-Location plot An assumption on the error terms is that the variance, σ^2, is constant for all values of x, the predictor variable. This will be the case if this graph shows no tendency to have larger points at the left or right.

Cook's distance plot This shows points which are influential in the regression model. Large values may indicated a more robust method for fitting the data is warranted.

Question 7: Produce the diagnostic plots for `fheight` modeling `sheight`. Outside of a point or two, these graphs indicate the linear model seems appropriate. Which point is most unusual for these graphs?

Question 8: Make diagnostic plots for the model of `wrist` circumference predicting `body.fat`. Does the linear model seem to apply. Discuss.

Question 9: Make diagnostic plots for the model of `BMI` predicting `body.fat`. Does the linear model seem to apply. Discuss.