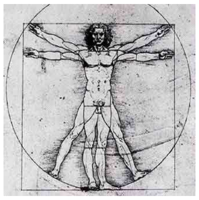

Human Proportions

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 *t*-based statistical Inferences

What is normal body temperature? Common wisdom is that it is 98.6 °F. The `normtemp` data set may be used to test this assumption using a test of significance.

2.1 Basics of significance tests

We will test the *null hypothesis* that the distribution of normal body temperatures is normal with mean $\mu = 98.6$ and unspecified variance, against the *alternative hypothesis* that the mean is less than this. That is

$$H_0 : \mu = 98.6, \qquad H_A : \mu < 98.6.$$

---

[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`.

As we assume the data is normally distributed, we will use the T-statistic as a test statistic:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\text{observed} - \text{expected}}{\textbf{SE}}.$$

Finally, we'll use $\alpha = 0.05$ for the *level of significance.*

Question 1: An assumption for using the t-statistic is that the data be normally distributed, or n be large enough. In this case, n is quite large, but nonetheless, verify graphically that the data appears to come from a normal distribution. Which graph did you use, and how did it show normality.

We can find the p-value with the following commands:

```
> xbar = mean(temperature)
> s = sd(temperature)
> n = length(temperature)
> T.obs = (xbar - 98.6)/(s/sqrt(n))
> pt(T.obs, df = n - 1)

[1] 1.205316e-07
```

We see that the p-value is tiny, about 10^{-7}, and is much less than the significance level α. The p-value is computed using the alternative hypothesis to specify what values of the test statistic are more extreme than the observed value, `T.obs`. In this case, as smaller values of $\bar{x}$ support H_A, we see that smaller values are more extreme.

Alternatively, one can find the critical value(s) for this problem using H_A and α. For this problem, the lone critical value corresponds to having $\alpha \cdot 100\%$ of the area to the left of the critical value. That is, the critical value is the 0.05 quantile of the t-distribution and is found with

```
> qt(0.05, df = n - 1)

[1] -1.656752


> T.obs

[1] -5.454823
```

One can see that `T.obs` is much less than the critical value.

For this problem, the computer can do all of the work with the `t.test()` function. To use this function, one must specify the data (`temperature`), the null hypothesis (`mu=98.6`), and the alternative hypothesis (`alt="less"`). The function returns the observed value of T, the degrees of freedom, the p-value and more:

```
> t.test(temperature, mu = 98.6, alt = "less")


        One Sample t-test

data:  temperature
t = -5.4548, df = 129, p-value = 1.205e-07
alternative hypothesis: true mean is less than 98.6
95 percent confidence interval:
     -Inf 98.35577
sample estimates:
mean of x
 98.24923
```

You are left to compare the p-value to α if so desired.

The alternative hypothesis may be specified three ways to the `t.test` function: $H_A : \mu < \mu_0$ is specified with `alt="less"`, $H_A : \mu > \mu_0$ is specified with `alt="greater"`, and $H_A : \mu \neq \mu_0$ is specified with `alt="two.sided"`.

Question 2: Some have suggested that 98.2°F is a better value for the mean body temperature. Do a test of significance of

$$H_0 : \mu = 98.2, \qquad H_A : \mu \neq 98.2.$$

Is the difference statistically significant at the $\alpha = 0.05$ level?

Question 3: Do females have a higher body temperature? We will ask this by testing if the mean body temperature for females is 98.2°F or is it more. That is

$$H_0 : \mu = 98.2, \qquad H_A : \mu > 98.2.$$

You can extract the temperatures for just the females in the study with this command:

```
> females = temperature[gender == "female"]
```

1. First check that using the t-statistic is appropriate for this data. That is, either the data appears to come from a normal population or the data set (n) is large enough.

   Is the statistic appropriate and why?

2. Now apply `t.test()` to find the p-value. Are the differences statistically significant at a level of $\alpha = 0.05$.

Question 4: A watch band maker assumes that the average wrist size is seven inches. We can use the `wrist` variable to check this assumption using a significance test of

$$H_0 : \mu = 7 \text{ inches}, \qquad H_A : \mu \neq 7\text{inches}.$$

1. Check that a t-test is appropriate for this data, by verifying normality of the data, or a large n. Why is the t-test appropriate?
2. As the measurements in `wrist` are in centimeters, convert the null and alternative hypotheses into centimeters (1 inch = 2.54 cms).
3. Test to see if the difference between the mean wrist size of the data is statistically significant at the $\alpha = 0.05$ level.

Question 5: The body mass index, `BMI`, is a measurement of weight divided by height squared in metric units. "Normal" values are in the range $[18,5,25]$. Do a significance test at the $\alpha = 0.05$ level to see if the mean BMI is 25 versus an alternative that it is more.

1. First plot the data in `BMI` and argue if the t-test is appropriate for this data and why.
2. Write down the null and alternative hypotheses

$$H_0 : \qquad\qquad H_A :$$

3. What is the p-value?
4. Is the difference statistically significant?

Question 6: A manufacturer of one-size-fits-all sweatpants wishes to know whether the mean waist size (the data is stored in the variable `abdomen`) has changed from the last time it designed a pair of sweatpants. Before, the manufacturer assumed the mean waist size was 34 inches. Do a significant test at the $\alpha = 0.05$ level, testing

$$H_0 : \mu = 34, \qquad H_A : \mu > 34.$$

1. Why is a t-test approporiate for the data in `abdomen`?
2. Rewrite the null and alternative hypotheses using cms (1 inch = 2.54 centimeters).
3. What is the p-value?
4. Is the difference statistically significant?

Question 7: The mean height of 17-year old Finnish boys is estimated to be 180cms. Suppose the data collected in the `height` variable is a random sample from adult males in the United States. Is the mean height statistically different from that of full grown 17-year olds from Finland at the $\alpha = 0.05$ level of significance?

1. Specify the null and alternative hypotheses using inches.
2. Comment on the appropriateness of the t-test?
3. What is the p-value?

2.2 two-sample tests

Is there a statistically significant difference between the body temperatures for males and females? Let μ_m be the population mean for males, and μ_f be the population mean for females. If we assume that the two populations are normally distributed, then a significance test of

$$H_0 : \mu_1 = \mu_2, \qquad H_A : \mu_1 \neq \mu_2$$

(or $H_A : \mu_1 < \mu_2$, or $H_A : \mu_1 > \mu_2$) can also be done using a t-distributed statistic.

The `t.test()` will still perform this test. In the two-sample case, the null hypothesis is not specified as it is always the same.

For example, the two-sample test of equivalence of means for the temperature data can be performed as follows

```
> males = temperature[gender == "Male"]
> females = temperature[gender == "female"]
> t.test(males, females, alt = "two.sided")
```

```

        Welch Two Sample t-test

data:  males and females
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

The small p-value indicates that there is a statistically significant difference.

Question 8: The normality of the two populations (sampled in `males` and `females`) was not verified. Comment on the validity of this assumption. (A large sample size is also sufficient to use the t-test.)

Question 9: If the population variances can be assumed to be equal, then the extra argument `var.equal=TRUE` to `t.test()` will generally give smaller p-values, as the sampling distribution of the test statistic has smaller degrees of freedom in most cases. First check if this is a reasonable assumption about our data, and then check if it makes a difference in the t-test.

Question 10: An alternate syntax (the model formula syntax) for performing a two-sample test when there is one variable (`gender`) which is used to indicate which level of some factor the subject has may be used. The t-test above, may have been carried out with

```
> t.test(temperature ~ gender)
```

Verify that this gives the same output and also that the default value for `alt=` is `"two.sided"`.

Question 11: Perform a two-sample t-test of the resting heart rate (`hr`) for males and females. Is the difference statistically significant at the $\alpha = 0.10$ level?

3 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.

Question 12: 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.

3.1 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

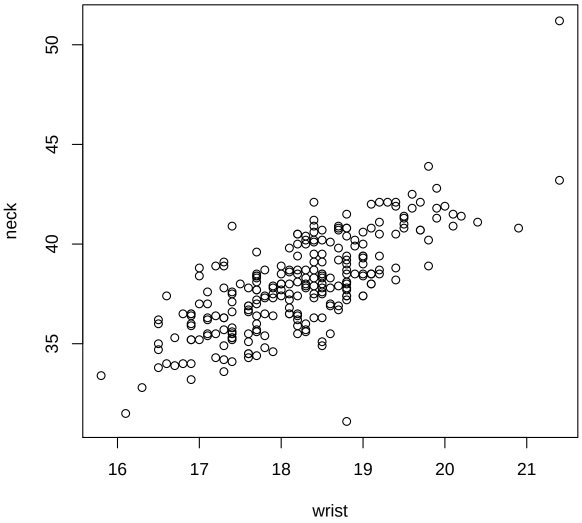

Figure 1: Scatterplot of neck size versus wrist size

Question 13: For the data in `normtemp`, make a scatterplot of heart rater `hr` versus `temperature`. Does there appear to be a relationship?

Question 14: 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)
```

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

---

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

```
> 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 15: 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?

3.2 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 $(y - 16) = 2(x - 35)$.

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

Question 17: What is the correlation between `sheight` and `fheight` for the `father.son` data set. Is this value close to 1 or -1?

Question 18: What is the correlation between `wrist` and `bicep` circumference measurements in the `fat` data set?

Question 19: 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.

3.3 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 20: 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 21: 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 22: Find the equation for the regression line for the model of father's height, `fheight`, predicting son's height, `sheight`.

3.4 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 23: Add a regression line to your scatterplot of `hr` predicting `temperature` for the `normtemp` data set.

Question 24: Add a regression line to your scatterplot of `fheight` and `sheight` for the `father.son` data set.

4 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 25: What size neck would be predicted for a person with a 20-centimeter wrist size?

Question 26: 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.

4.1 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 27: 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 28: 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 29: 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)

4.2 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 30: 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 31: 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 32: 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 33: 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 34: 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 35: 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.

4.3 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 36: 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 37: Make diagnostic plots for the model of `wrist` circumference predicting `body.fat`. Does the linear model seem to apply. Discuss.

Question 38: Make diagnostic plots for the model of `BMI` predicting `body.fat`. Does the linear model seem to apply. Discuss.