Scatterplots, Correlation, and Regression

Colby Community College

Definition

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variables.

Definition

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variables.

Definition

A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Definition

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variables.

Definition

A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Definition

A **scatterplot** is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis for the second variable (y).

Definition

A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variables.

Definition

A **linear correlation** exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Definition

A **scatterplot** is a plot of paired (x, y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first variable (x), and the vertical axis for the second variable (y).

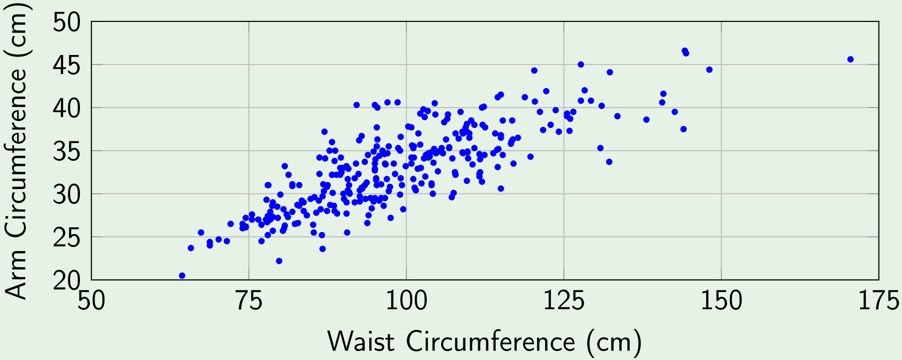
Warning

The presence of a correlation between two variables is not evidence that one of the variables causes the other.

Correlation does not imply causality!

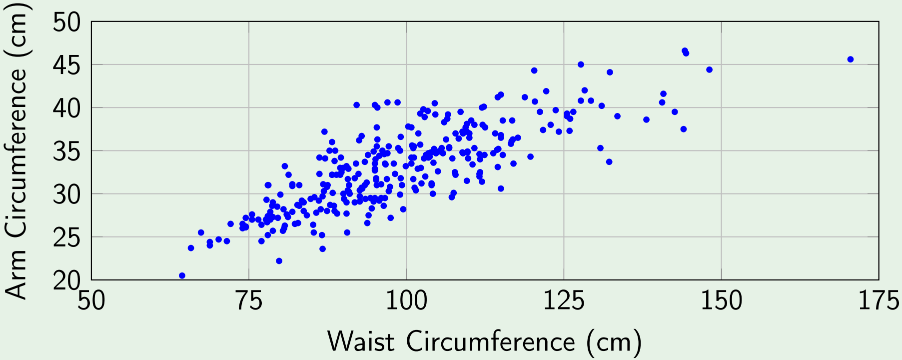
Example 1

Data Set 1 "Body Data" in Appendix B includes waist circumference and arm circumference (cm) of randomly selected adult subjects.



Example 1

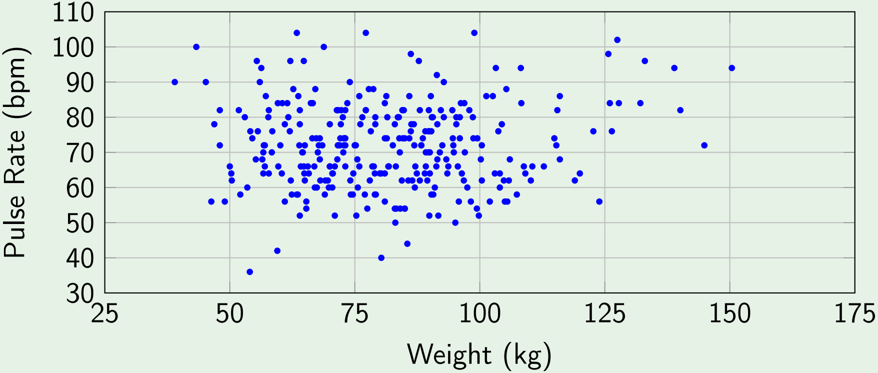
Data Set 1 "Body Data" in Appendix B includes waist circumference and arm circumference (cm) of randomly selected adult subjects.



The points show a pattern of increasing values from left to right. This pattern suggests that there is a relationship between waist circumferences and arm circumferences.

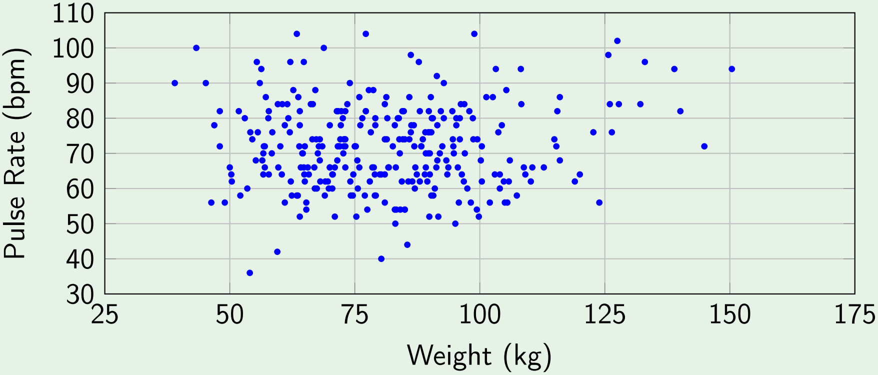
Example 2

Data Set 1 "Body Data" in Appendix B includes weights (kg) and pulse rates (bpm) of randomly selected adult subjects.



Example 2

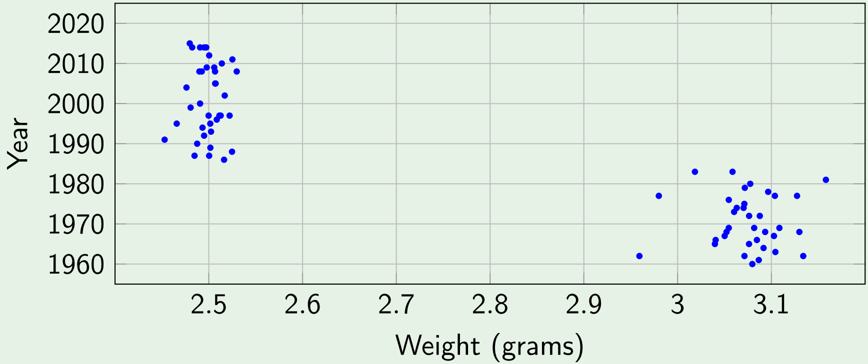
Data Set 1 "Body Data" in Appendix B includes weights (kg) and pulse rates (bpm) of randomly selected adult subjects.



The points do not show any obvious pattern, and this lack of a pattern suggests that there is no relationship between weights and pulse rates.

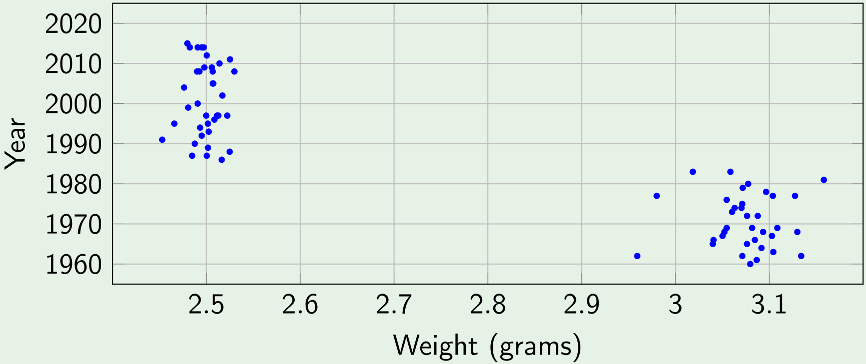
Example 3

Consider the scatterplot that depicts data consisting of the weight (grams) and year of production for 72 pennies.



Example 3

Consider the scatterplot that depicts data consisting of the weight (grams) and year of production for 72 pennies.



While it may look like there is a relationship, looking at the individual clusters we see that there is not a relationship between the weight of a penny and year is was produced.

Definition

The **linear correlation coefficient** is denoted by r, and it measures the strength of the linear association between two variables.

Definition

The **linear correlation coefficient** is denoted by r, and it measures the strength of the linear association between two variables.

Note

The correlation coefficient is always between -1 and 1. The closer r is to zero, the weaker the linear correlation. The closer r is to either -1 or 1, the stronger the correlation.

Definition

The **linear correlation coefficient** is denoted by r, and it measures the strength of the linear association between two variables.

Note

The correlation coefficient is always between -1 and 1. The closer r is to zero, the weaker the linear correlation. The closer r is to either -1 or 1, the stronger the correlation.

Note

In chapter 10 we will talk in detail about how correlation is calculated. For now, we will use software to calculate correlation and focus on how to interpret the results.

Example 4

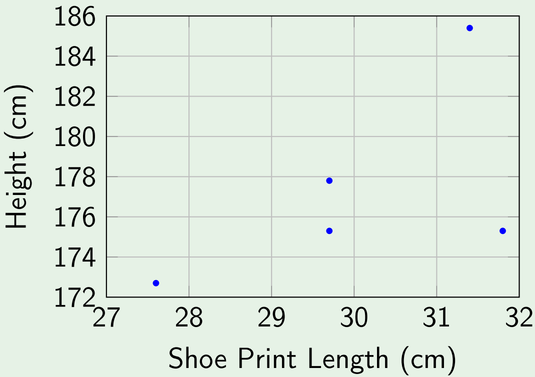
The table contains the shoe size and height of five random people. (Data Set 2 Appendix B)

Shoe Print Length (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7

Example 4

The table contains the shoe size and height of five random people. (Data Set 2 Appendix B)

Shoe Print Length (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7



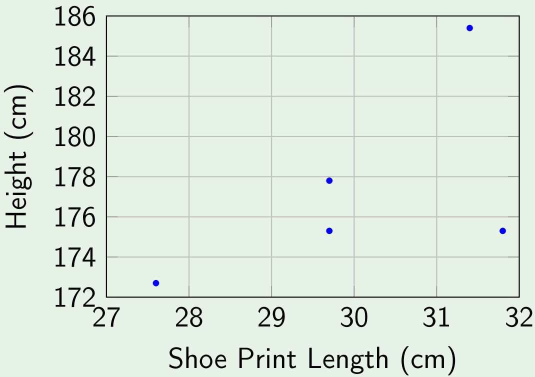
Statdisk Output

```
Sample Size, n:   5
Degrees of Freedom: 3

Correlation Results:
Correlation Coeff, r:  0.59127
Critical r:            ±0.87834
P-Value (two-tailed):  0.29369

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      125.40733
Slope, b1:            1.72745

Total Variation:       95.02
Explained Variation:   33.21891
Unexplained Variation: 61.80109
Standard Error:        4.53876
Coeff of Det, R^2:     0.3496
```

Example 4

The table contains the shoe size and height of five random people. (Data Set 2 Appendix B)

Shoe Print Length (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7



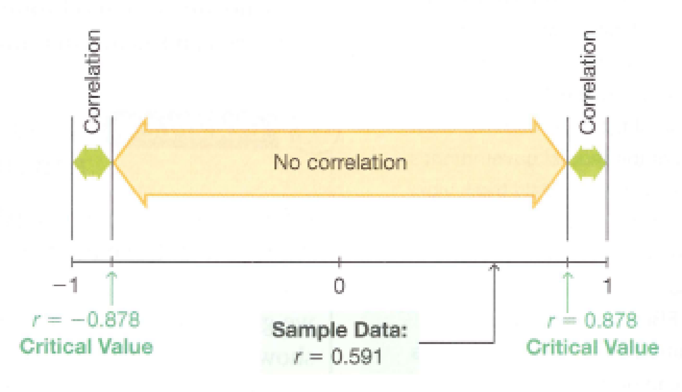
Statdisk Output

```
Sample Size, n:    5
Degrees of Freedom: 3

Correlation Results:
Correlation Coeff, r:  0.59127
Critical r:            ±0.87834
P-Value (two-tailed):  0.29369

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      125.40733
Slope, b1:            1.72745

Total Variation:       95.02
Explained Variation:   33.21891
Unexplained Variation: 61.80109
Standard Error:        4.53876
Coeff of Det, R^2:     0.3496
```

Is $r = 0.59127$ "close" to 1? Or is r "close" to 0?

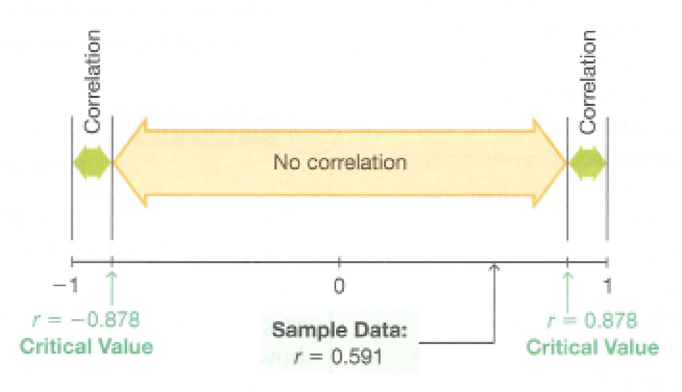
Interpreting r

The critical value tells us the separation between "close to -1" or "close to 1" and "close to 0."



Interpreting r

The critical value tells us the separation between "close to -1" or "close to 1" and "close to 0."



We see that there isn't sufficient evidence in Example 4 for correlation between shoe print length and height.

Definition

If there really is no linear correlation between two variables, the **P-value** is the probability of getting paired sample data with linear correlation coefficient r that is at least as extreme as the one obtained from the paired data.

Definition

If there really is no linear correlation between two variables, the **P-value** is the probability of getting paired sample data with linear correlation coefficient r that is at least as extreme as the one obtained from the paired data.

Note

Using tables of critical values is becoming obsolete, the more common approach is to use a P-value.
(We will talk more about P-values in chapter 8.)

Definition

If there really is no linear correlation between two variables, the **P-value** is the probability of getting paired sample data with linear correlation coefficient r that is at least as extreme as the one obtained from the paired data.

Note

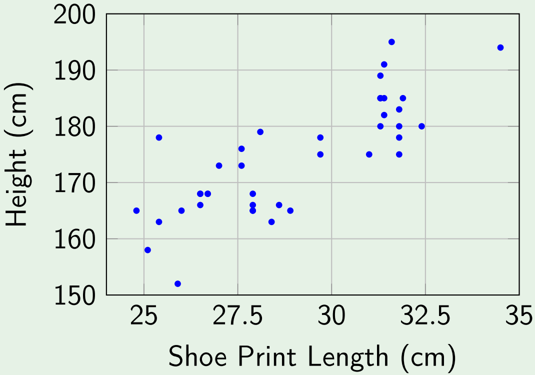
Using tables of critical values is becoming obsolete, the more common approach is to use a P-value.
(We will talk more about P-values in chapter 8.)

Note

Only a small P-value, such as 0.05 or less, suggests that the sample results are not likely to occur by chance when there is no linear correlation. The smaller the P-value the stronger the evidence that there is a linear correlation between the two variables.

Example 5

Using the full data set on shoe print length and height. (Data Set 2)

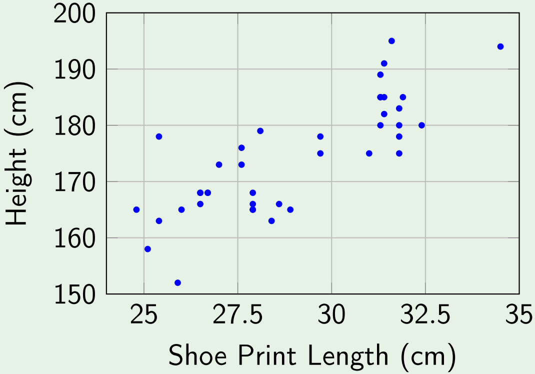


Statdisk Output

```
Sample Size, n:   40
Degrees of Freedom: 38

Correlation Results:
Correlation Coeff, r:  0.81295
Critical r:            ±0.31201
P-Value (two-tailed):  0

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      80.93041
Slope, b1:            3.21856

Total Variation:       3958.755
Explained Variation:   2616.27965
Unexplained Variation: 1342.47535
Standard Error:        5.94376
Coeff of Det, R^2:     0.66088
```

Example 5

Using the full data set on shoe print length and height. (Data Set 2)



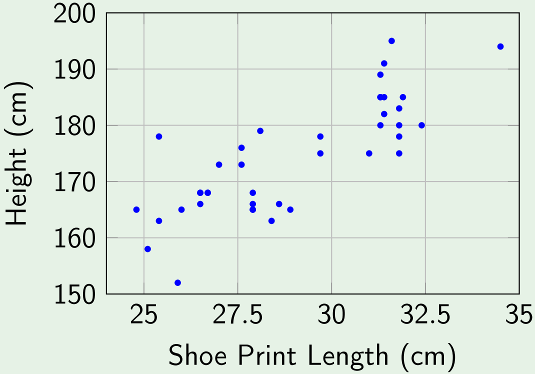
Statdisk Output

```
Sample Size, n:   40
Degrees of Freedom: 38

Correlation Results:
Correlation Coeff, r:  0.81295
Critical r:           ±0.31201
P-Value (two-tailed):  0

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      80.93041
Slope, b1:            3.21856

Total Variation:       3958.755
Explained Variation:   2616.27965
Unexplained Variation: 1342.47535
Standard Error:        5.94376
Coeff of Det, R^2:     0.66088
```

In Example 4 the Statdisk output gives a P-value of 0.29369. This is much larger than 0.05, which suggests with only 5 pairs of data, there isn't evidence of a correlation between shoe print length and height.

Example 5

Using the full data set on shoe print length and height. (Data Set 2)



Statdisk Output

```
Sample Size, n:   40
Degrees of Freedom: 38

Correlation Results:
Correlation Coeff, r:  0.81295
Critical r:            ±0.31201
P-Value (two-tailed):  0

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      80.93041
Slope, b1:            3.21856

Total Variation:       3958.755
Explained Variation:   2616.27965
Unexplained Variation: 1342.47535
Standard Error:        5.94376
Coeff of Det, R^2:     0.66088
```

In Example 4 the Statdisk output gives a P-value of 0.29369. This is much larger than 0.05, which suggests with only 5 pairs of data, there isn't evidence of a correlation between shoe print length and height.

But, with the full 40 data pairs, we get a P-value of 0. This is strong evidence of a correlation between shoe print length and height.

Definition

Given a collection of paired sample data, the **regression line** (or **line of best fit**) is the straight line that "best" fits the scatterplot of the data.

Definition

Given a collection of paired sample data, the **regression line** (or **line of best fit**) is the straight line that "best" fits the scatterplot of the data.

Note

In chapter 10 we will talk about to calculate a "best" fit.

Definition

Given a collection of paired sample data, the **regression line** (or **line of best fit**) is the straight line that "best" fits the scatterplot of the data.

Note

In chapter 10 we will talk about to calculate a "best" fit.

Regression Equation

The regression equation

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the regression line.

Definition

Given a collection of paired sample data, the **regression line** (or **line of best fit**) is the straight line that "best" fits the scatterplot of the data.

Note

In chapter 10 we will talk about to calculate a "best" fit.

Regression Equation

The regression equation

$$\hat{y} = b_0 + b_1 x$$

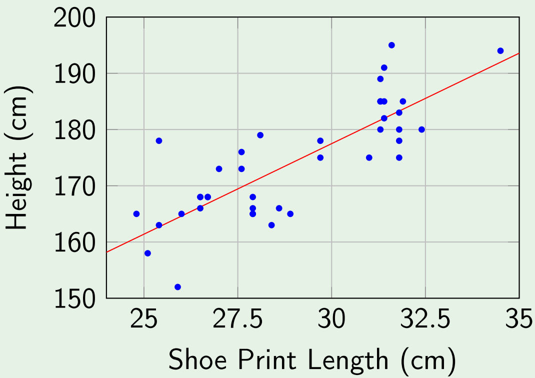
algebraically describes the regression line.

Note

A regression line is used to make predictions about a population using the sample data.

Example 6

Using the full data set on shoe print length and height. (Data Set 2)

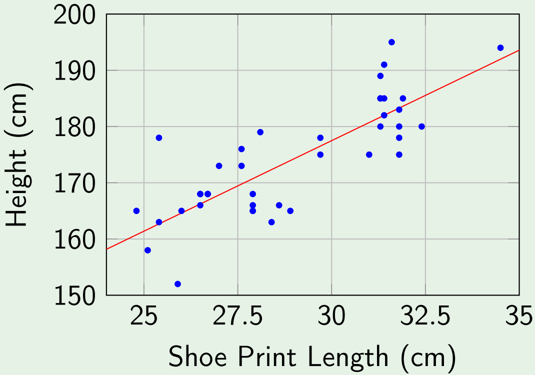


Statdisk Output

```
Sample Size, n:   40
Degrees of Freedom: 38

Correlation Results:
Correlation Coeff, r:  0.81295
Critical r:            ±0.31201
P-Value (two-tailed):  0

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      80.93041
Slope, b1:            3.21856

Total Variation:       3958.755
Explained Variation:   2616.27965
Unexplained Variation: 1342.47535
Standard Error:        5.94376
Coeff of Det, R^2:     0.66088
```

Example 6

Using the full data set on shoe print length and height. (Data Set 2)



Statdisk Output

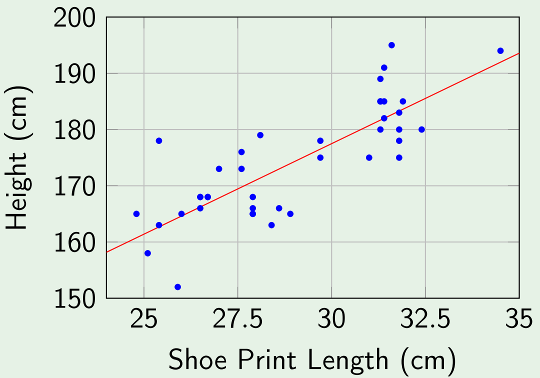
```
Sample Size, n:   40
Degrees of Freedom: 38

Correlation Results:
Correlation Coeff, r:  0.81295
Critical r:            ±0.31201
P-Value (two-tailed):  0

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      80.93041
Slope, b1:            3.21856

Total Variation:       3958.755
Explained Variation:   2616.27965
Unexplained Variation: 1342.47535
Standard Error:        5.94376
Coeff of Det, R^2:     0.66088
```

The regression line can also be expressed as, rounding to one decimal place,

$$\text{Height} = 80.9 + 3.2(\text{Shoe Print Length})$$

Example 6

Using the full data set on shoe print length and height. (Data Set 2)



Statdisk Output

```
Sample Size, n:   40
Degrees of Freedom: 38

Correlation Results:
Correlation Coeff, r:  0.81295
Critical r:           ±0.31201
P-Value (two-tailed):  0

Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      80.93041
Slope, b1:            3.21856

Total Variation:       3958.755
Explained Variation:   2616.27965
Unexplained Variation: 1342.47535
Standard Error:        5.94376
Coeff of Det, R^2:     0.66088
```

The regression line can also be expressed as, rounding to one decimal place,

$$\text{Height} = 80.9 + 3.2(\text{Shoe Print Length})$$

We can expect a person with a shoe length of 30 cm to be $80.9 + 3.2(30) = 176.9$ cm tall.