

Unit 9

Correlation & Regression

Examining Relationships

In statistics, we often want to compare **two (or more) different populations** with respect to the **same variable**.

We use tools such as side-by-side boxplots and two-sample t tests to make comparisons between the samples.

Examining Relationships

Often, however, we wish to examine relationships between **several variables** for the **same population**.

When we are interested in examining the relationship between two variables, we may find ourselves in one of two situations:

- We may simply be interested in the **nature of the relationship**.
- One of the variables may be thought to **explain** or **predict** the other.

Explanatory and Response Variables

In this second case, one of the variables is an **explanatory variable** (which we denote by X) and the other is a **response variable** (denoted by Y).

A response variable takes values representing the outcome of a study, while an explanatory variable helps explain this outcome.

Example

Does the caffeine in coffee really help keep you awake? Researchers interviewed 300 adults and asked them how many cups of coffee they drink on an average day, as well as how many hours of sleep they get at night.

The response variable Y is **hours of sleep**, while the explanatory variable X is **number of cups of coffee per day**.

Practice Question

If the study finds that adults who drink more coffee tend to sleep less at night, can we conclude that drinking more coffee is the **cause**?

Practice Question

A calculus professor plans to conduct a study to determine whether a student's final exam score can be predicted by his or her midterm score. What is the explanatory variable in this study?

- (A) midterm score
- (B) the calculus course
- (C) the professor
- (D) final exam score
- (E) There is no logical explanatory or response variable here.

Example

Are students who excel in English also good at math, or are most people strictly left- or right-brained? A psychology professor locates 450 students at a large university who have taken the same introductory English course and the same Math course and compares their percentage grades in the two courses at the end of the semester.

In this case, there is no explanatory or response variable; we are simply interested in the nature of the relationship.

Scatterplots

The best way to display the relationship between two quantitative variables is with a **scatterplot**.

A scatterplot displays the values of two different **quantitative** variables measured on the **same individuals**. The data for each individual (for both variables) appears as a single point on the scatterplot. If there is an explanatory and a response variable, they should be plotted on the x - and y -axes, respectively. Otherwise, the choice of axes is arbitrary.

Example

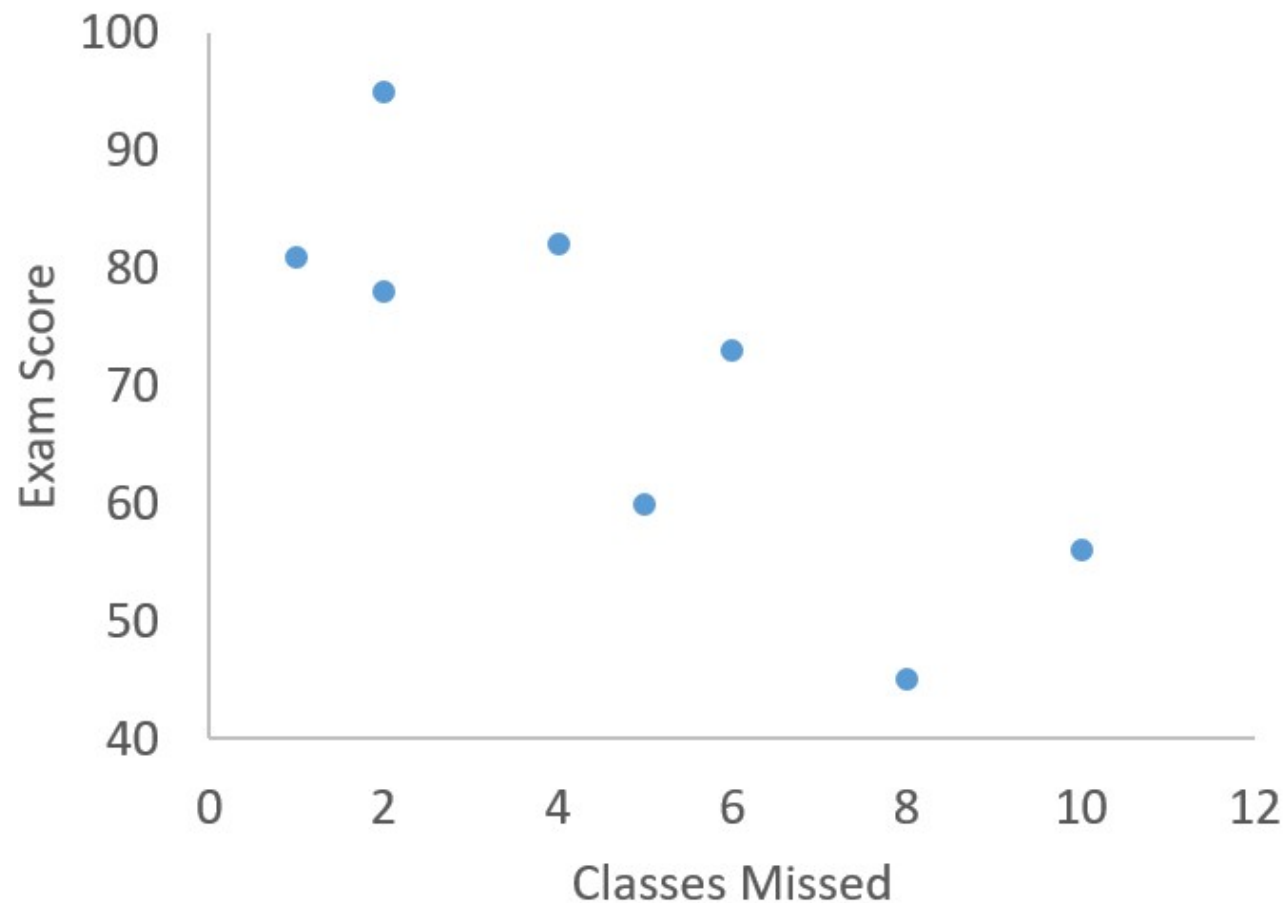
Consider the relationship between the number of classes a student misses during the term and his or her final exam score. The table on the following page gives the values for both variables for a sample of eight students.

Example

| Student | Classes Missed | Exam Score |
|---------|----------------|------------|
| 1 | 5 | 60 |
| 2 | 2 | 95 |
| 3 | 6 | 73 |
| 4 | 10 | 56 |
| 5 | 1 | 81 |
| 6 | 8 | 45 |
| 7 | 4 | 82 |
| 8 | 2 | 78 |

Example

The scatterplot for these data is shown below:



Scatterplots

We look for four things when examining a scatterplot:

1) Direction

- In this case, there is a **negative association** between the two variables. An above-average number of classes missed tends to be accompanied by a below-average exam score, and vice-versa. If the pattern of points slopes upward from left to right, we say there is a **positive association**.

Scatterplots

2) Form

- A straight line would do a fairly good job approximating the relationship between the two variables. It is therefore reasonable to assume that these two variables share a **linear relationship**.

Scatterplots

3) Strength

- The strength of the relationship is determined by how close the points lie to a simple form such as a straight line. In our example, if we draw a line which roughly approximates the relationship between the two variables, all points will fall quite close to the line. As such, the linear relationship is quite **strong**.

Scatterplots

3) Strength (cont'd)

- Not all relationships are linear in form. They can be quadratic, logarithmic or exponential, to name a few. Sometimes the points appear to be “randomly scattered”, in which case many of them will fall far from a line used to approximate the relationship. In this case, we say the linear relationship between the two variables is **weak**.

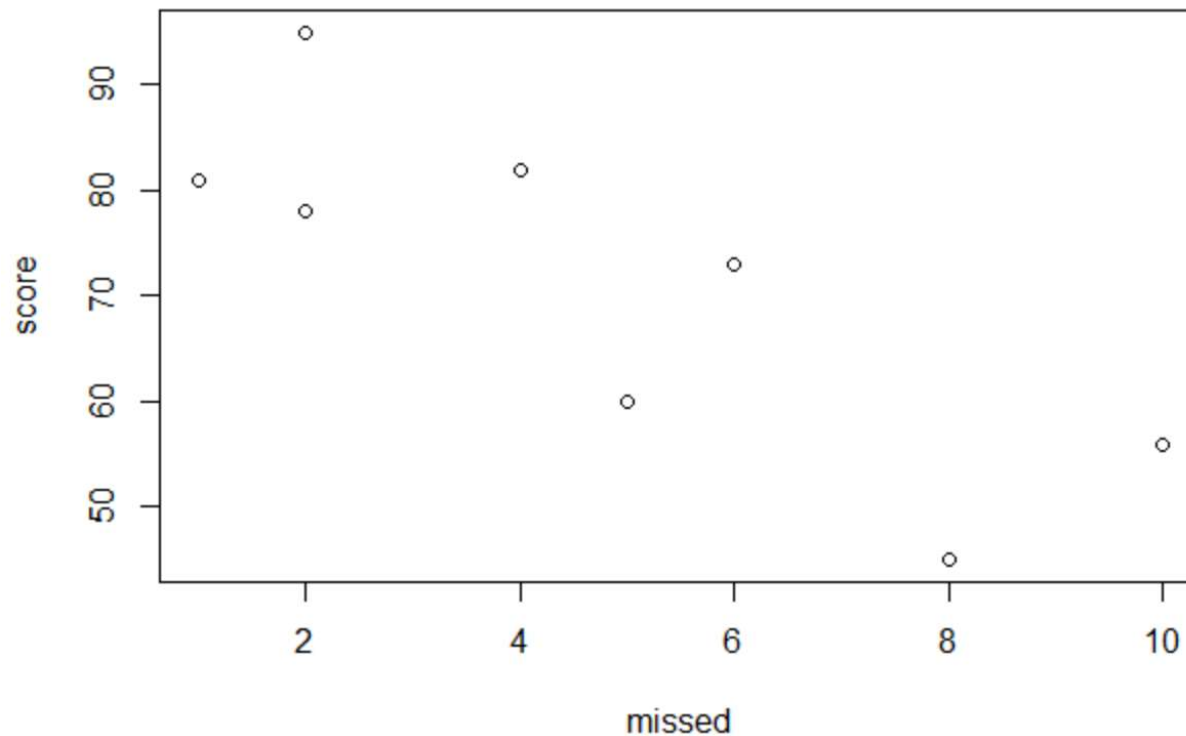
Scatterplots

4) Outliers

- There are several types of outliers for bivariate data. An observation may be outlying in either the x - or y -directions (or both). Another type of outlier occurs when an observation simply falls outside the general pattern of points, even if it is not extreme in either the x - or y -directions. Some types of outliers have more of an impact on our analysis than others, as we will discuss shortly.

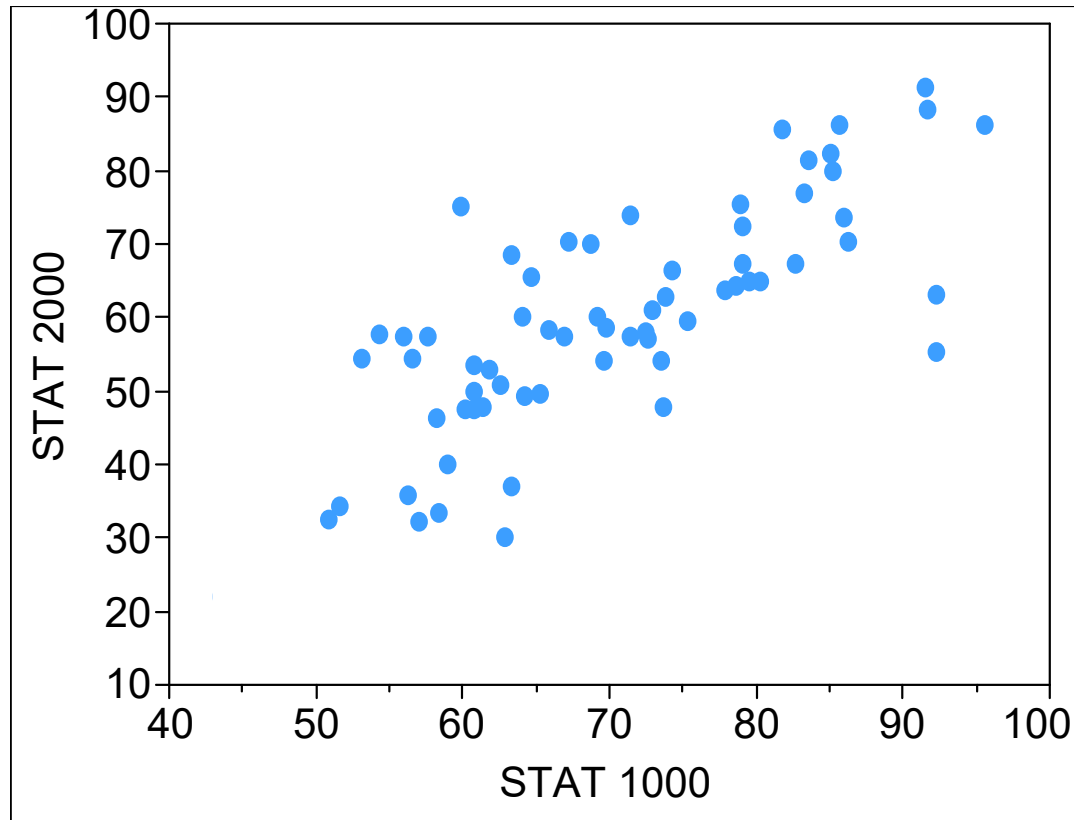
R Code

```
> missed <- c(5, 2, 6, 10, 1, 8, 4, 2)  
> score <- c(60, 95, 73, 56, 81, 45, 82, 78)  
> plot(missed, score)
```



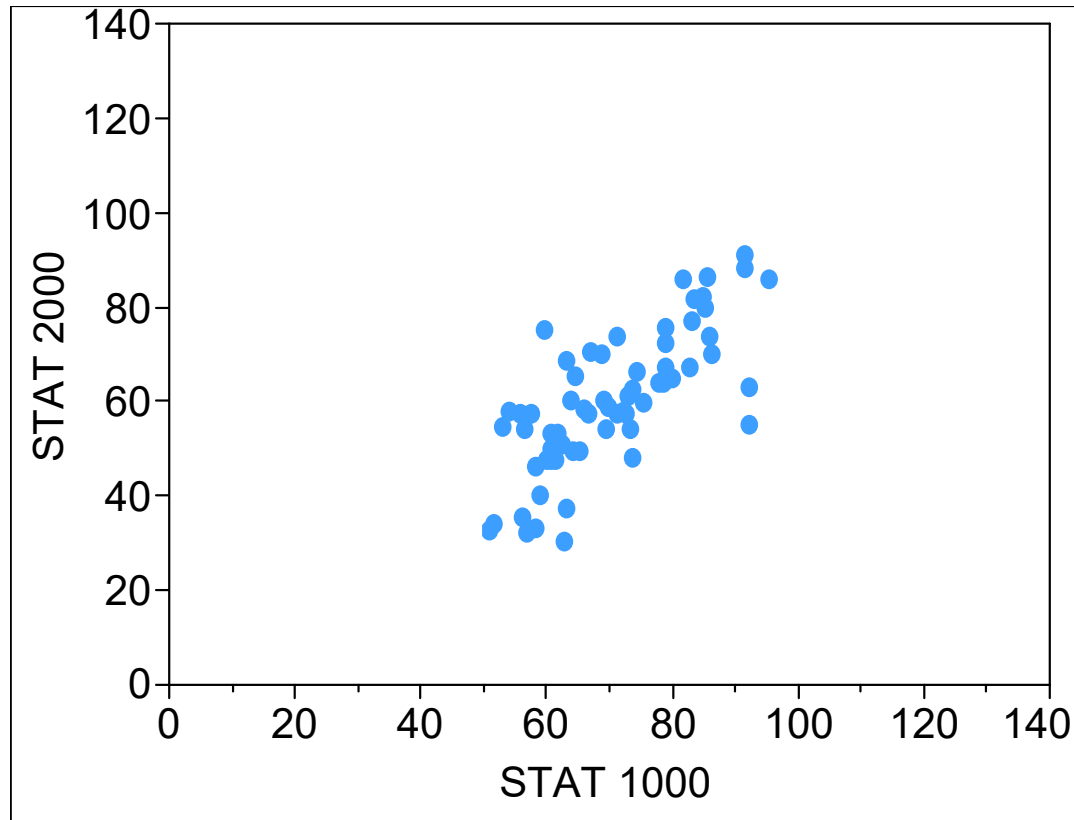
Strength of Linear Relationship

The STAT 1150 and STAT 2150 percentage grades for a sample of students who have taken both courses are displayed in the scatterplot below:



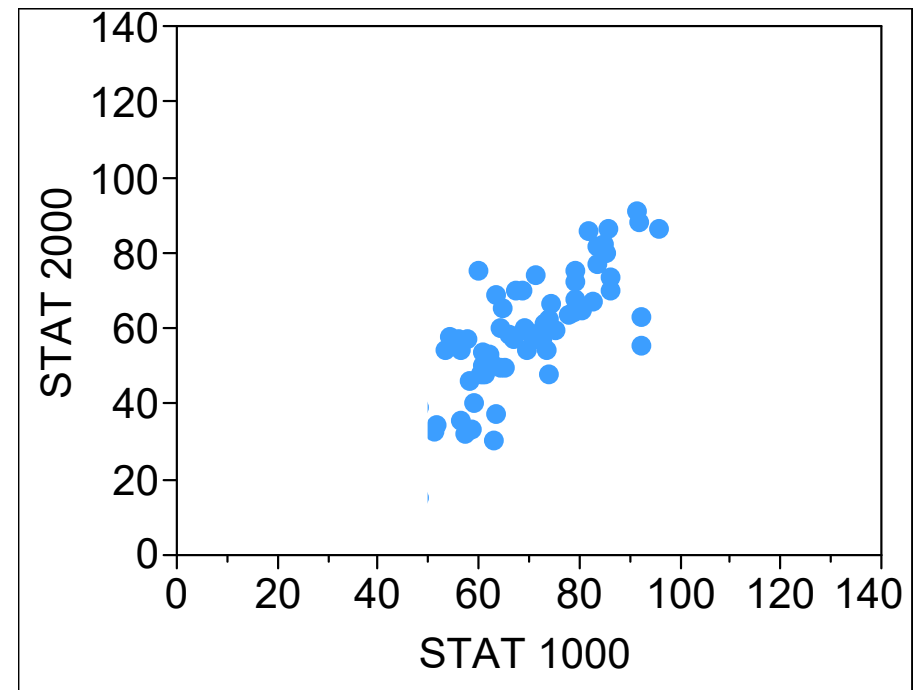
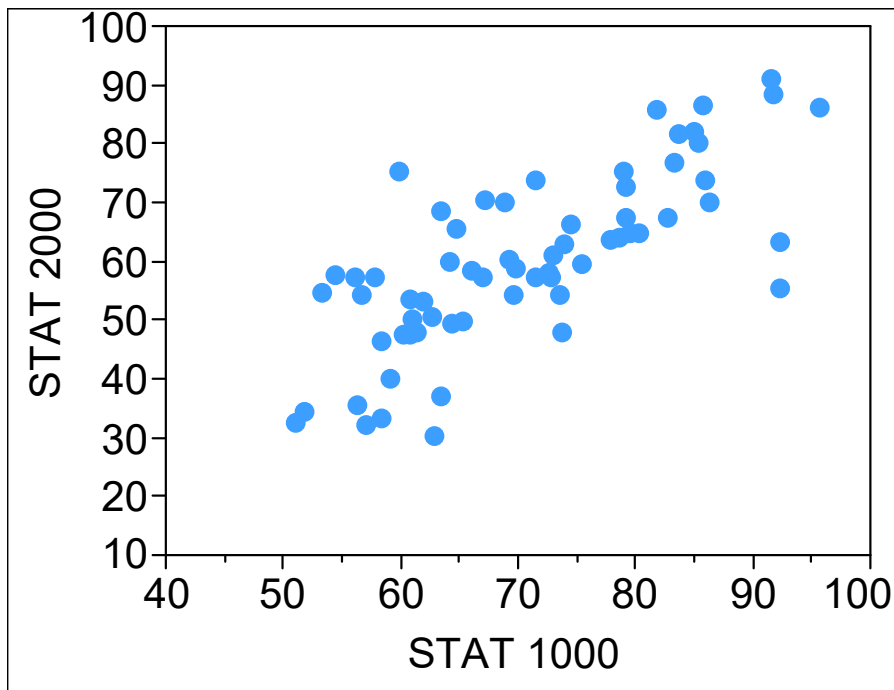
Strength of Linear Relationship

The scatterplot shows a moderately strong positive linear relationship. Does the relationship for the data in the following scatterplot appear stronger?



Strength of Linear Relationship

It might, but these are the **same data**; the scatterplots are just constructed with **different scales**!



Strength of Linear Relationship

This example shows that our eyes are not the best tools to assess the strength of relationship between two quantitative variables.

Can we find a numerical measure that will give us a concrete description of the strength of a linear relationship between two quantitative variables?

The measure we use is called **correlation**.

Correlation Coefficient

The **correlation coefficient** r measures the direction and strength of a linear relationship between two quantitative variables.

Suppose the values of two quantitative variables X and Y have been measured for n individuals. Then

$$\begin{aligned} r &= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{1}{(n-1)s_x s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

Correlation

We will use the second version of the formula, as it is computationally simpler. To calculate the correlation r :

- (i) Calculate \bar{x} , \bar{y} , s_x and s_y
- (ii) Calculate the deviations $x_i - \bar{x}$ and $y_i - \bar{y}$
- (iii) Multiply the corresponding deviations for x and y
 $(x_i - \bar{x})(y_i - \bar{y})$
- (iv) Add the n products $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- (v) Divide by $(n - 1)s_x s_y$

$$r = \frac{1}{(n - 1)s_x s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Correlation

For the Classes Missed and Exam Score example,

(i) $\bar{x} = 4.75, \bar{y} = 71.25, s_x = 3.15, s_y = 16.35$

| x_i | y_i | (ii) $x_i - \bar{x}$ | $y_i - \bar{y}$ | (iii) $(x_i - \bar{x})(y_i - \bar{y})$ |
|-------|-------|----------------------|-----------------|--|
| 5 | 60 | 0.25 | -11.25 | -2.8125 |
| 2 | 95 | -2.75 | 23.75 | -65.3125 |
| 6 | 73 | 1.25 | 1.75 | 2.1875 |
| 10 | 56 | 5.25 | -15.25 | -80.0625 |
| 1 | 81 | -3.75 | 9.75 | -36.5625 |
| 8 | 45 | 3.25 | -26.25 | -85.3125 |
| 4 | 82 | -0.75 | 10.75 | -8.0625 |
| 2 | 78 | -2.75 | 6.75 | -18.5625 |
| | | sum = 0 | sum = 0 | (iv) sum = -294.5 |

Correlation

$$\begin{aligned} \text{(v)} \quad r &= \frac{1}{(n-1)s_x s_y} \sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{7(3.15)(16.35)} (-294.5) = -0.8169 \end{aligned}$$

Note that some software programs display only the value of r^2 . If there is a positive association, r is the positive square root of r^2 , and if there is a negative association, r is the negative square root of r^2 .

R Code

```
> cor(missed, score)  
[1] -0.8165786
```

Calculations in R will often differ slightly from our calculations, as R carries more decimal places.

Association vs. Causation

As we will see shortly, a correlation of $r = -0.8169$ is representative of a fairly strong linear relationship. However, we must be careful when interpreting correlation. Despite the strong negative correlation, we **cannot** conclude that missing more classes **causes** a student's grade to decrease.

There are many other variables that could help explain the strong relationship between Classes Missed and Exam Score. One such variable is the **effort** of a student.

Lurking Variable

Students who put more effort into the course generally miss fewer classes. We also know that exam scores tend to be higher for more dedicated students.

The effort of a student in this example is known as a **lurking variable**. A lurking variable is one that helps explain the relationship between variables in a study, but which is not itself included in the study.

Association vs. Causation

Regardless of the existence of identifiable lurking variables, we must remember that correlation measures **only** the linear association between two quantitative variables. It gives us **no** information about the causal nature of the relationship.

Association does not imply causation!

Practice Question

We would like to examine how the age X of a certain model of car affects its selling price Y . The age (in years) and price (in \$) for a sample of 15 cars of the same make and model are recorded from the classified ads in the newspaper one weekend. The correlation is calculated to be $r = -0.847$.

Practice Question

Despite the strong correlation between age and price, we cannot say that a car getting older **causes** its value to decrease. This is because of the potential effect of one or more lurking variables. Which of the following is a likely lurking variable in this case?

- (A) colour
- (B) mileage
- (C) age
- (D) make and model
- (E) all of the above

Correlation

Some properties of correlation:

- Positive values of r indicate a positive association and negative values indicate a negative association.
- r falls between -1 and 1 , inclusive. Values of r close to -1 or 1 indicate a strong linear association (negative or positive, respectively). A correlation of -1 or 1 is obtained only in the case of a perfect linear relationship, i.e., when all points fall exactly on a line. Values of r close to zero indicate a weak linear relationship.

Correlation

Some properties of correlation (cont'd):

- r has no units (i.e., it is just a number).
- The correlation makes no distinction between X and Y . As such, an explanatory and response variable are not necessary.
- Changing the units of X and Y has **no effect** on the correlation, i.e., it doesn't matter if we measure a variable in pounds or kilograms, feet or metres, dollars or cents, etc.

Correlation

Some properties of correlation (cont'd):

- r measures only the strength of a **linear** relationship. In other cases, it is a useless measure.
- Because the correlation is a function of several measures that are affected by outliers, r is itself strongly affected by outliers.

Practice Question

Which of the following pairs of variables X and Y will likely have a positive correlation?

- (A) X = outdoor temperature
 Y = amount of hot chocolate sold at Tim Hortons
- (B) X = time it takes a runner to finish a race
 Y = number of competitors who finish after the runner
- (C) X = amount of alcohol consumed
 Y = reaction time of a driver
- (D) X = price of T-shirts at a clothing store
 Y = number of T-shirts sold
- (E) None of the above – all of these correlations are negative.

Practice Question

Which of the following pairs of variables would likely have a negative correlation?

- (I) We examine the sales at a Winnipeg Walmart store for a sample of days one year.

X = # of bottles of sunscreen sold

Y = # of jackets sold

- (II) We take a sample of drivers in a city.

X = # of speeding tickets received

Y = cost to renew driver's license

- (III) We take a sample of drivers in a large city.

X = distance between home and work

Y = daily time spent listening to the radio

(A) none (B) I only (C) II only (D) III only (E) I and II

Practice Question

The ages of a large sample of married men (over age 30) are recorded, along with several other variables. Consider the following pairs of variables:

- (i) $X = \text{Age}$, $Y = \text{Year of Birth}$
- (ii) $X = \text{Age}$, $Y = \text{Spouse's Age}$
- (iii) $X = \text{Age}$, $Y = \text{Height}$

Which of the following is the most likely set of correlations for the three pairs of variables?

- (A) (i) $r = -1$, (ii) $r = 0.75$, (iii) $r = 0$
- (B) (i) $r = 1$, (ii) $r = 0$, (iii) $r = 0.5$
- (C) (i) $r = -1$, (ii) $r = 0.2$, (iii) $r = 0.5$
- (D) (i) $r = 1$, (ii) $r = 0.75$, (iii) $r = 0.5$
- (E) (i) $r = -1$, (ii) $r = 0$, (iii) $r = 0$

Practice Question

The lengths (in centimetres) and the weights (in grams) of a sample of rainbow trout are measured and the correlation between length and weight is calculated to be 0.62. What would be the value of the correlation if the lengths had instead been measured in feet (1 foot = 30.48 cm) and the weights had been measured in ounces (1 ounce = 28.35 grams)?

- (A) 0.67 (B) 0.71 (C) 0.62 (D) 0.58 (E) 0.38

Practice Question

A study found a correlation of $r = -0.27$ between the gender of a worker and his or her income. You conclude that:

- (A) women earn more than men on average.
- (B) men earn more than women on average.
- (C) a calculation error was made; -0.27 is not a possible value of r .
- (D) this is nonsense; correlation makes no sense here.
- (E) the correlation is not meaningful because the relationship between Gender and Income is likely nonlinear.

Regression

When a relationship appears to be linear in nature, we often wish to estimate this relationship between variables with a single straight line.

A **regression line** is a straight line that describes how a response variable Y changes as an explanatory variable X changes. This line is often used to **predict** values of Y for given values of X .

Regression

Note that with correlation, we didn't require a response variable and an explanatory variable.

In regression, we always have an explanatory variable X and a response variable Y .

Regression

The theory behind regression analysis relies on the assumption that there is a **true line** describing the relationship between X and Y . Not all points will fall exactly on the line because of the inherent variation of the response variable. Instead, the line describes an equation for μ_Y , defined as the average value of Y for any given value of X . The equation of our theoretical line is

$$\mu_Y = \beta_0 + \beta_1 X$$

where β_0 and β_1 are the “true intercept” and “true slope”, respectively.

Regression

We also assume that:

- For any given value of X , the response variable Y follows a normal distribution.
- Observed responses y_i are independent of one another.
- The standard deviation σ of Y is the **same** for all values of X . The value of the parameter σ is unknown.

Simple Linear Regression Model

Together, these assumptions are known as the **simple linear regression model**, which can be written as follows:

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

in which:

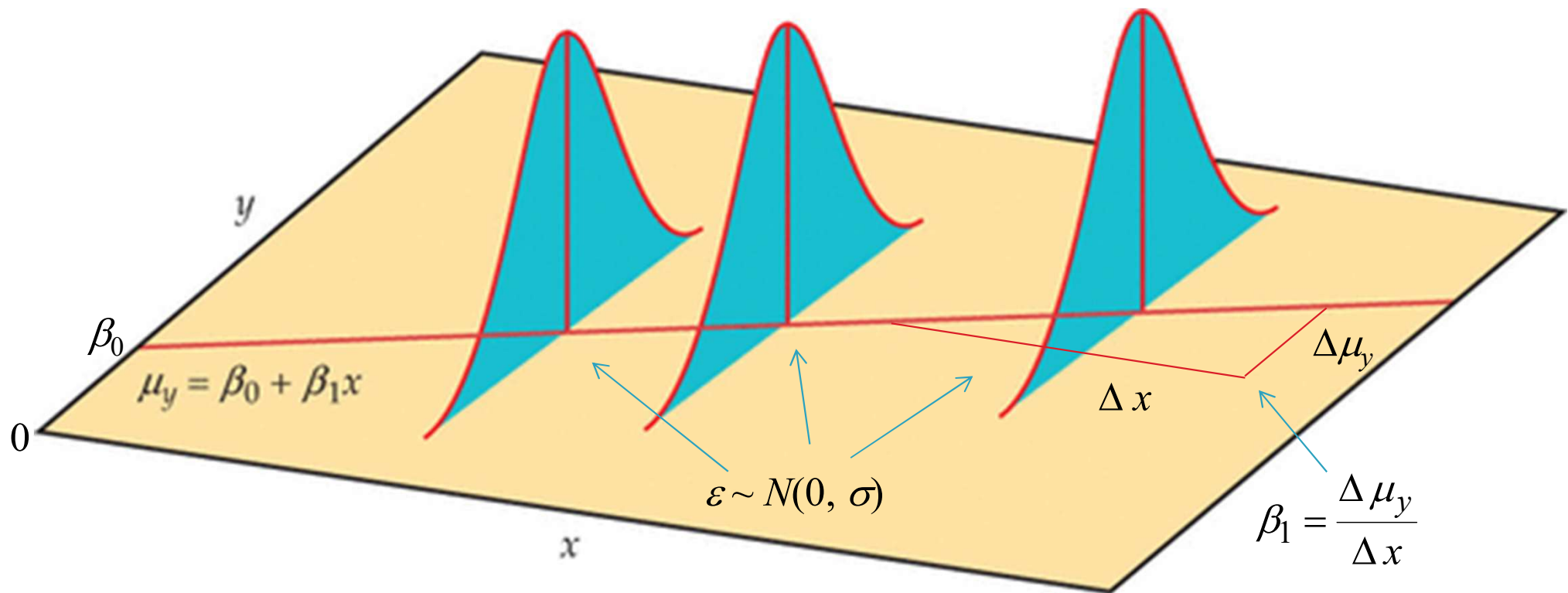
y_i is the value of the response variable for the i^{th} individual in the population,

β_0 and β_1 are parameters

x_i is the value of the explanatory variable for the i^{th} individual in the population,

ϵ_i is a random error term such that $\epsilon_i \sim N(0, \sigma)$.

Simple Linear Regression Model



Regression

Given a value of X , we would like to **predict** the corresponding value of Y . Unless there is a perfect relationship, we won't know the exact value of Y , because Y is still a variable.

Regression Line

We will instead use a sample to estimate the equation of this true line. Our estimate of the true line is

$$\hat{y} = b_0 + b_1x$$

\hat{y} is the estimated value of μ_Y for a given value of X . The values b_0 (the intercept) and b_1 (the slope) are estimates for β_0 and β_1 , respectively.

We will use this regression line to make our predictions.

Regression Line

We would like to find the line that fits our data the best. That is, we need to find the appropriate values of b_0 and b_1 .

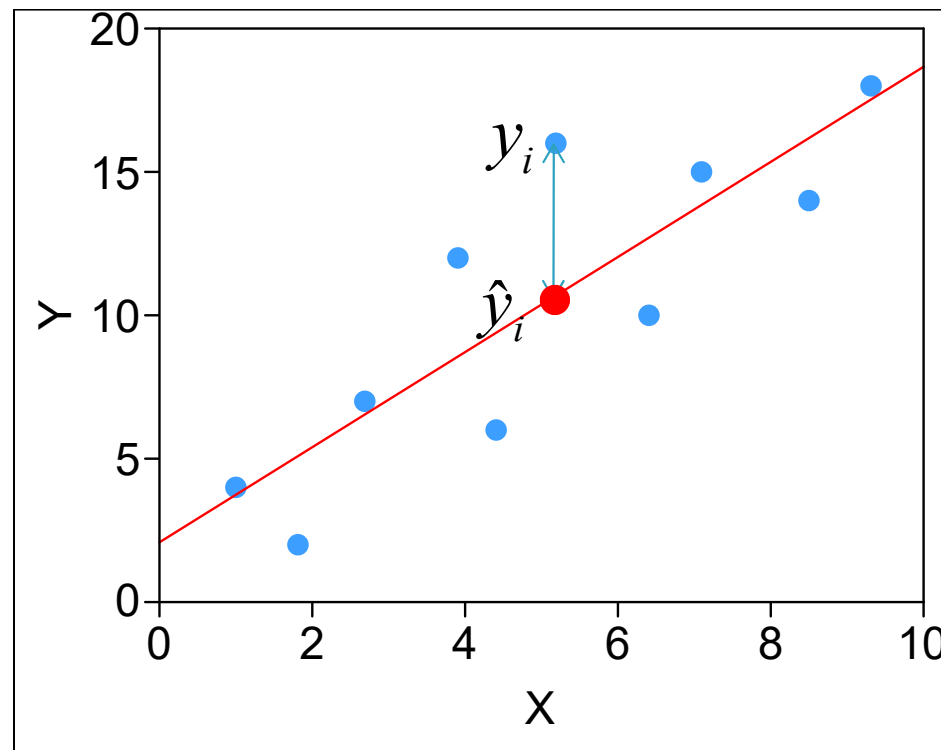
But there are infinitely many possible lines. Which one is the “best” line?

Since we are using X to predict Y , we would like the line to lie as close to the points as possible in the **vertical** direction.

Regression Line

The line we will use is the line that **minimizes the sum of squared deviations in the vertical direction**:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Least Squares Regression

The values of b_0 and b_1 that give us the line that minimizes this sum of squared deviations are:

$$b_1 = r \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

The line $\hat{y} = b_0 + b_1 x$ is called the **least squares regression line**, for obvious reasons.

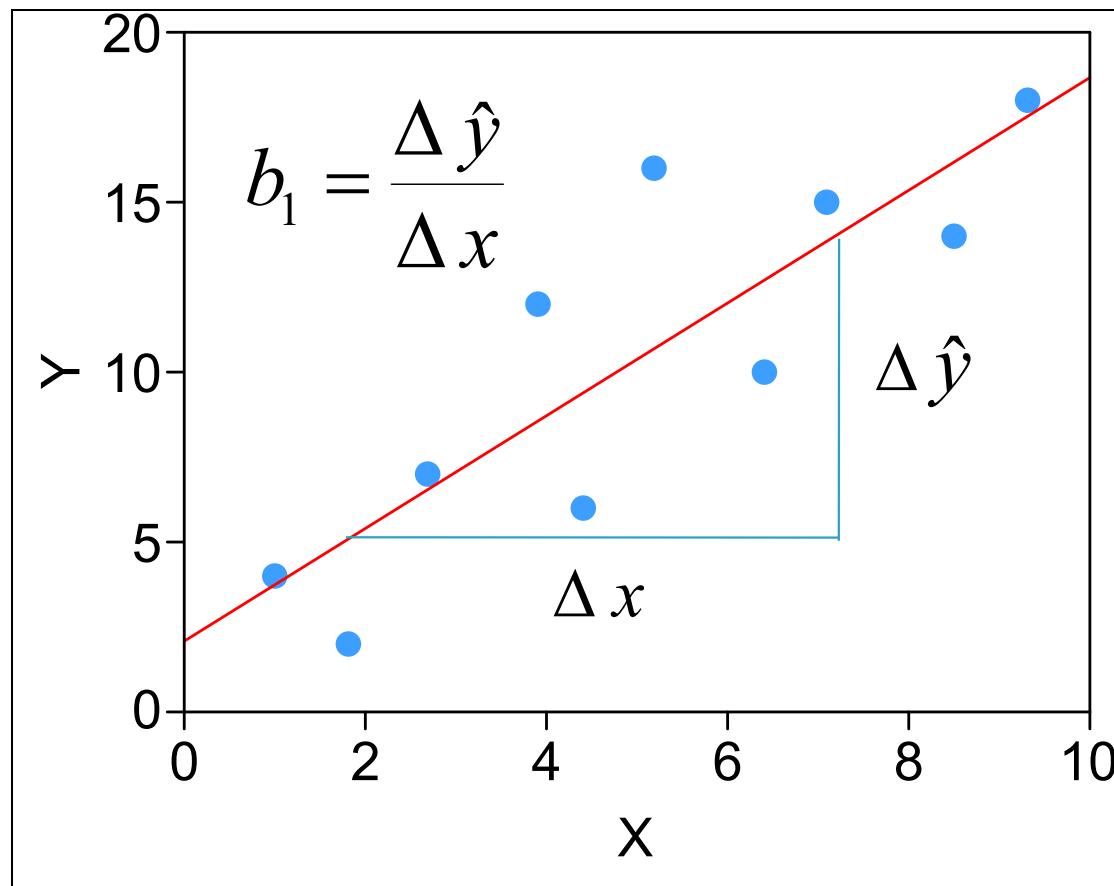
Least Squares Regression

We like to make a distinction between correlation and regression, and so we also have an equivalent formula for the least squares regression slope b_1 which does not directly involve the correlation coefficient:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

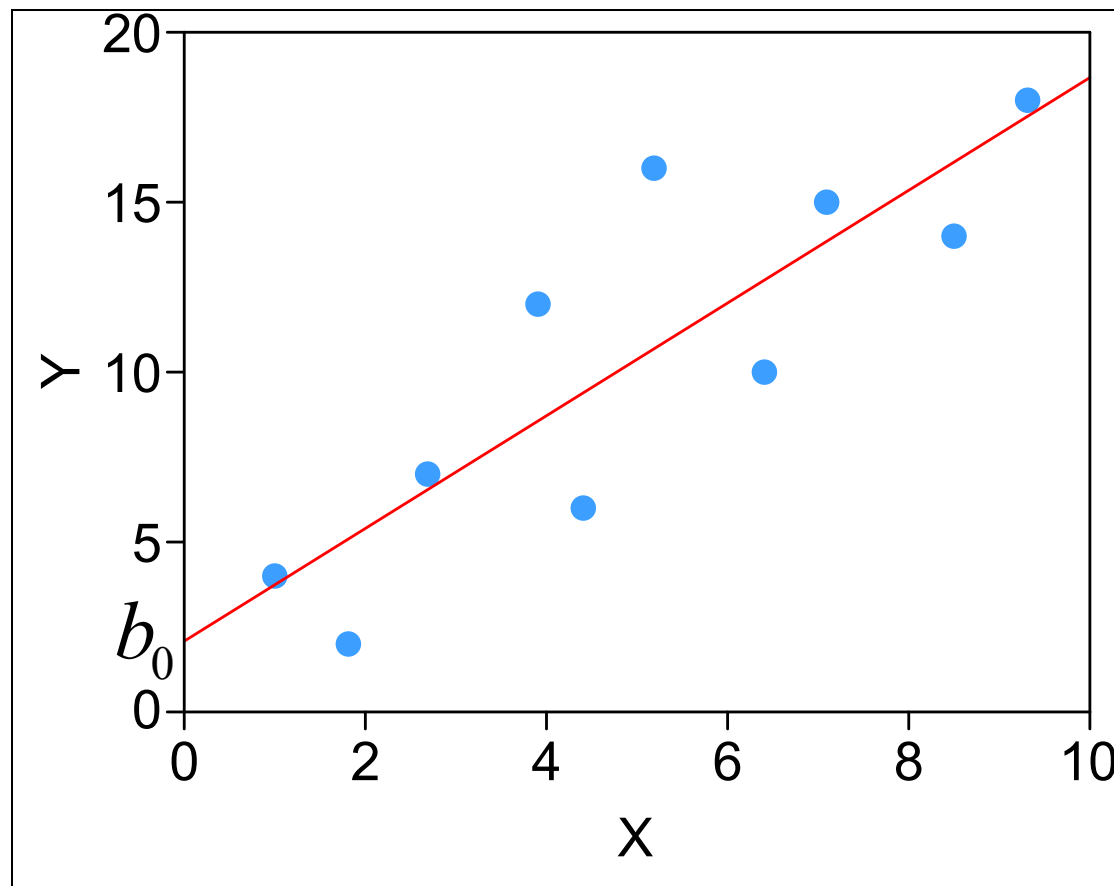
Least Squares Regression Slope

The **slope** of the regression line, b_1 , is defined as **the predicted increase in y when x increases by one unit.**



Least Squares Regression Intercept

The **intercept** of the regression line, b_0 , is defined as the **predicted value of y when $x = 0$** .



Coefficient of Determination r^2

Some variability in Y is accounted for by the fact that, as X changes, it pulls Y along with it. The remaining variation is accounted for by other factors (which we usually don't know).

The quantity r^2 is called the **coefficient of determination** and has a special meaning in least squares regression. It is **the fraction of variation in Y that is accounted for by its regression on X .**

Coefficient of Determination r^2

If $r = -1$ or 1 , then $r^2 = 1$. That is, we can predict Y exactly for any value of X , as regression on X accounts for all of the variation in Y .

If $r = 0$, then $r^2 = 0$, and so regression on X tells us nothing about the value of Y .

Otherwise, r^2 is between 0 and 1.

Example

Can the monthly rent for an apartment be predicted by the size of the apartment? The size X (in square feet) and the monthly rent Y (in \$) are recorded for a sample of ten apartments in a large city. The data are shown below:

| | | | | | | | | | | |
|-----|------|-----|------|------|-----|------|------|------|------|------|
| X | 770 | 650 | 925 | 850 | 575 | 860 | 800 | 1000 | 730 | 900 |
| Y | 1270 | 990 | 2230 | 1295 | 860 | 1925 | 1575 | 1790 | 1580 | 1550 |

Example

Our simple linear regression model is as follows:

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

in which:

Y_i is the monthly rent for the i^{th} apartment in the population,

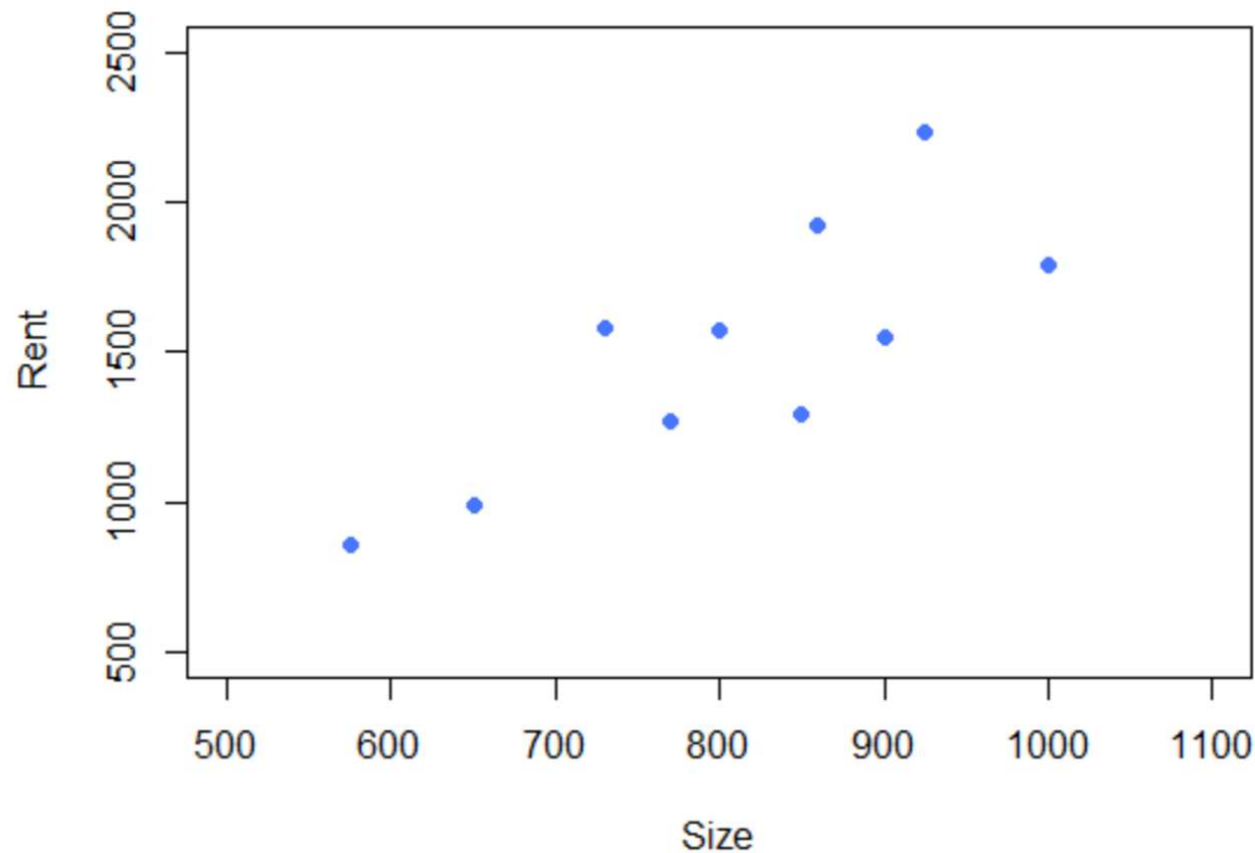
β_0 , and β_1 are parameters,

x_i is the size of the i^{th} apartment in the population,

ϵ_i is a random error term such that $\epsilon_i \sim N(0, \sigma)$.

Example

The first step in any analysis is to examine a picture of the data. A scatterplot of Rent vs. Size is shown below:



Example

Notice that Size is the explanatory variable X and Rent is the response variable Y , because the size of an apartment helps to **explain** its monthly rent.

We notice that there seems to be a **strong positive linear relationship** between the two variables.

We do some preliminary calculations and find that

$$\bar{x} = 806.0, \quad s_x = 129.22, \quad \bar{y} = 1506.5, \quad s_y = 418.51, \quad r = 0.8031$$

Other necessary calculations are shown in the table on the following pages:

Example

| Apartment | Size X | Rent Y | $(x_i - \bar{x})$ | $(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|-----------|----------|----------|-------------------|-------------------|---------------------|----------------------------------|
| 1 | 770 | 1270 | -36 | -236.5 | 1296 | 8514.0 |
| 2 | 650 | 990 | -156 | -516.5 | 24336 | 80574.0 |
| 3 | 925 | 2230 | 119 | 723.5 | 14161 | 86096.5 |
| 4 | 850 | 1295 | 44 | -211.5 | 1936 | -9306.0 |
| 5 | 575 | 860 | -231 | -646.5 | 53361 | 149341.5 |
| 6 | 860 | 1925 | 54 | 418.5 | 2916 | 22599.0 |
| 7 | 800 | 1575 | -6 | 68.5 | 36 | -411.0 |
| 8 | 1000 | 1790 | 194 | 283.5 | 37636 | 54999.0 |
| 9 | 730 | 1580 | -76 | 73.5 | 5776 | -5586.0 |
| 10 | 900 | 1550 | 94 | 43.5 | 8836 | 4089.0 |

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 150290$$

(sum of second-last column)

$$\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 390910$$

(sum of last column)

Example

We have two equivalent formulas we could use to calculate the slope:

$$b_1 = r \frac{s_y}{s_x} = 0.8031 \left(\frac{418.5}{129.2} \right) = 2.60$$

Alternatively,

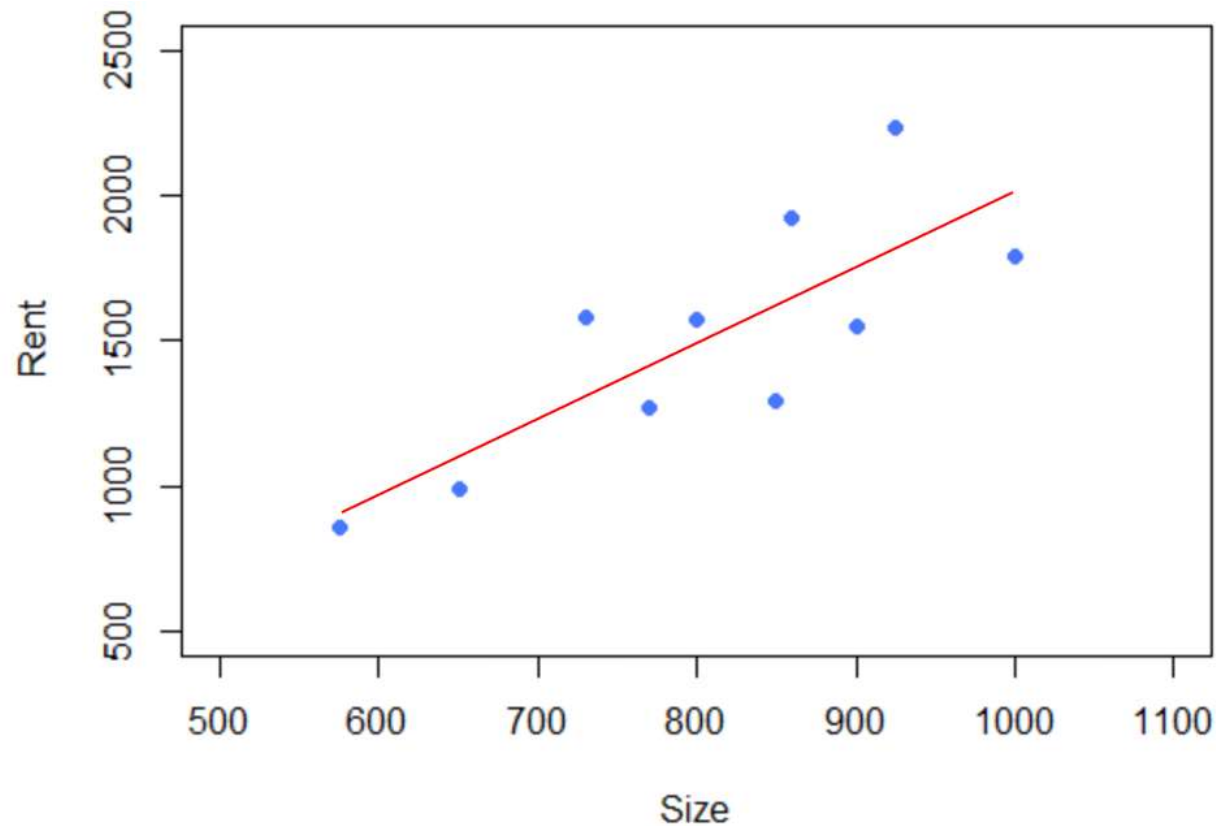
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{390910}{150290} = 2.60$$

The intercept is calculated as follows:

$$b_0 = \bar{y} - b_1 \bar{x} = 1506.5 - 2.60(806.0) = -589.10$$

Example

The equation of the least squares regression line is therefore $\hat{y} = -589.10 + 2.60x$. The line is shown on the scatterplot below:



Example

Note that, since calculations can be quite long in linear regression, we will usually rely on software for this purpose. We should, however, still be familiar with the proper use of the formulas.

R Code

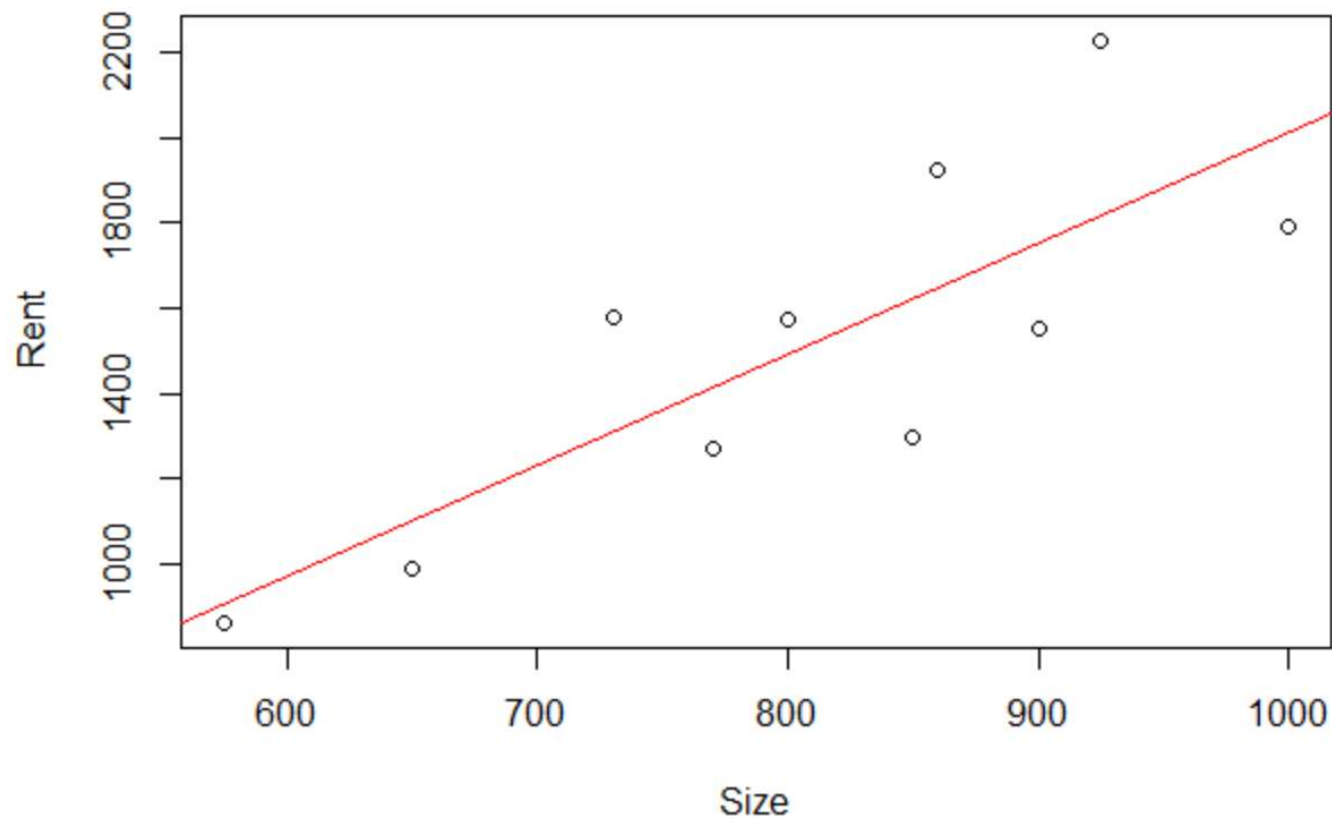
```
> Size <- c(770, 650, 925, 850, 575, 860, 800,  
            1000, 730, 900)  
> Rent <- c(1270, 990, 2230, 1295, 860, 1925,  
            1575, 1790, 1580, 1550)  
> lm(Rent ~ Size)
```

```
Call:  
lm(formula = Rent ~ Size)
```

```
Coefficients:  
(Intercept)      Size  
   -589.937      2.601
```

R Code

```
> plot(Size, Rent)
> abline(lm(Rent ~ Size), col = "red")
```



Interpretations

The slope $b_1 = 2.60$ tells us that, when the size of an apartment increases by one square foot, we predict the monthly rent to increase by \$2.60.

The intercept $b_0 = -589.10$ is statistically meaningless in this case. An apartment cannot have a size of 0 square feet, and a negative rent is impossible.

We also see that $r^2 = (0.8031)^2 = 0.645$, which tells us that 64.5% of the variation in an apartment's monthly rent is accounted for by its regression on size.

Association vs. Causation

Recall our discussion of association vs. causation. The former does not imply the latter. In the apartment example, there was a strong positive relationship between the size of an apartment and its monthly rent. However, this doesn't mean an apartment being larger **causes** its monthly rent to be higher. The observed relationship may be due to one or more **lurking variables**. For example, perhaps the apartments in our sample in nicer, more expensive parts of the city are larger. Then the neighbourhood where an apartment is located might be a lurking variable.

Prediction

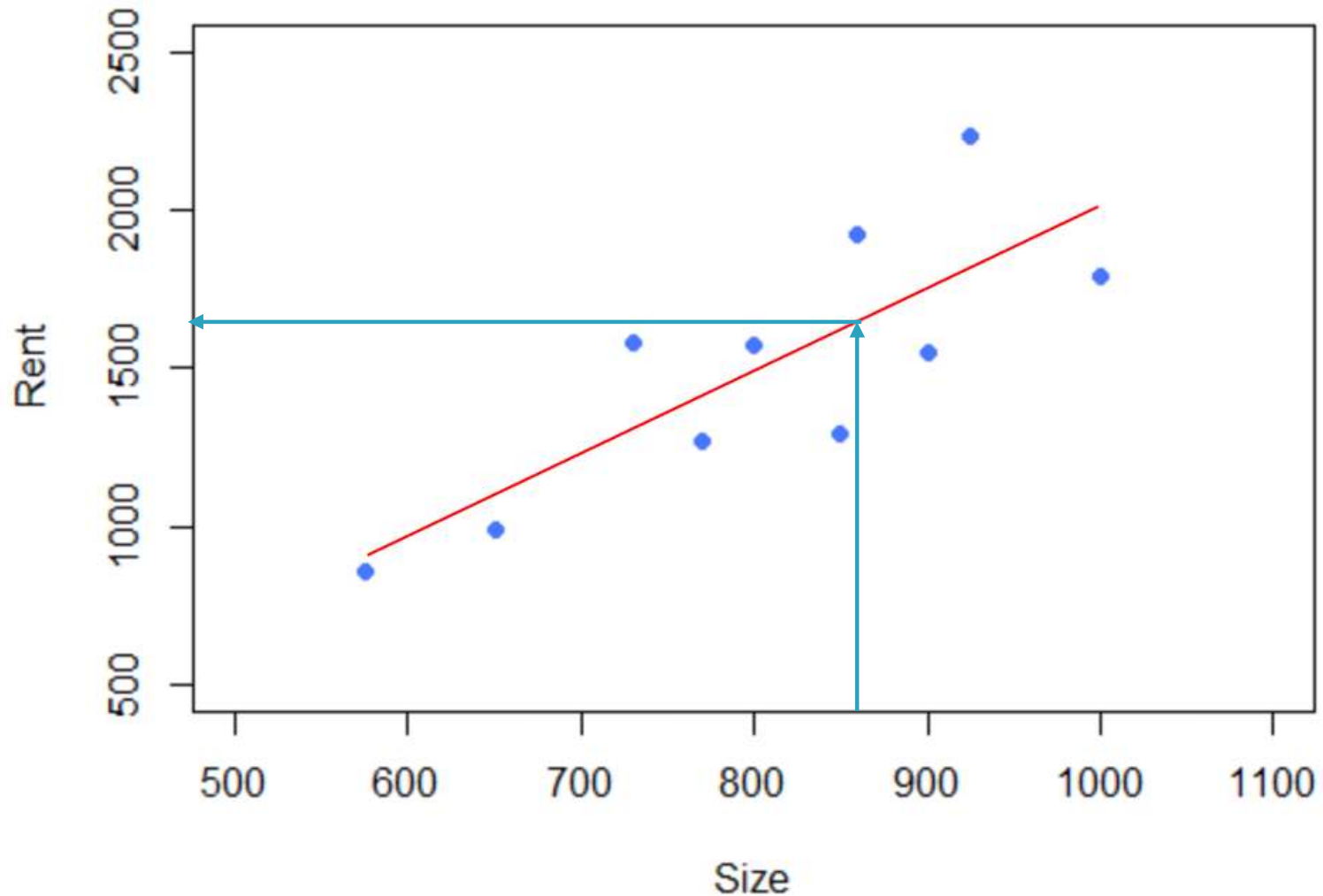
We can now use this line to predict the monthly rent for an apartment of a given size.

To do this, we simply plug the size of an apartment into the equation of the least squares regression line. For example, the predicted monthly rent for an 860 square foot apartment is

$$\hat{y} = -589.10 + 2.60(860) = 1646.90$$

Predicted Value of Y

We call this the **predicted value of Y** when $X = 860$.



Residuals

Note that there is an 860 square foot apartment in the sample. How does the actual monthly rent for this apartment compare with the predicted rent?

$$e_6 = y_6 - \hat{y}_6 = 1925 - 1646.90 = 278.10$$

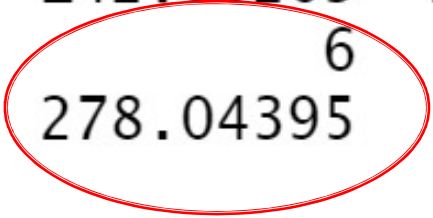
The monthly rent for this apartment is \$278.10 higher than we would have predicted it to be from our regression line.

The value $y_i - \hat{y}_i$ is called the **residual** for the i^{th} observation. The residual for any value of X reflects the **error** of our prediction.

R Code

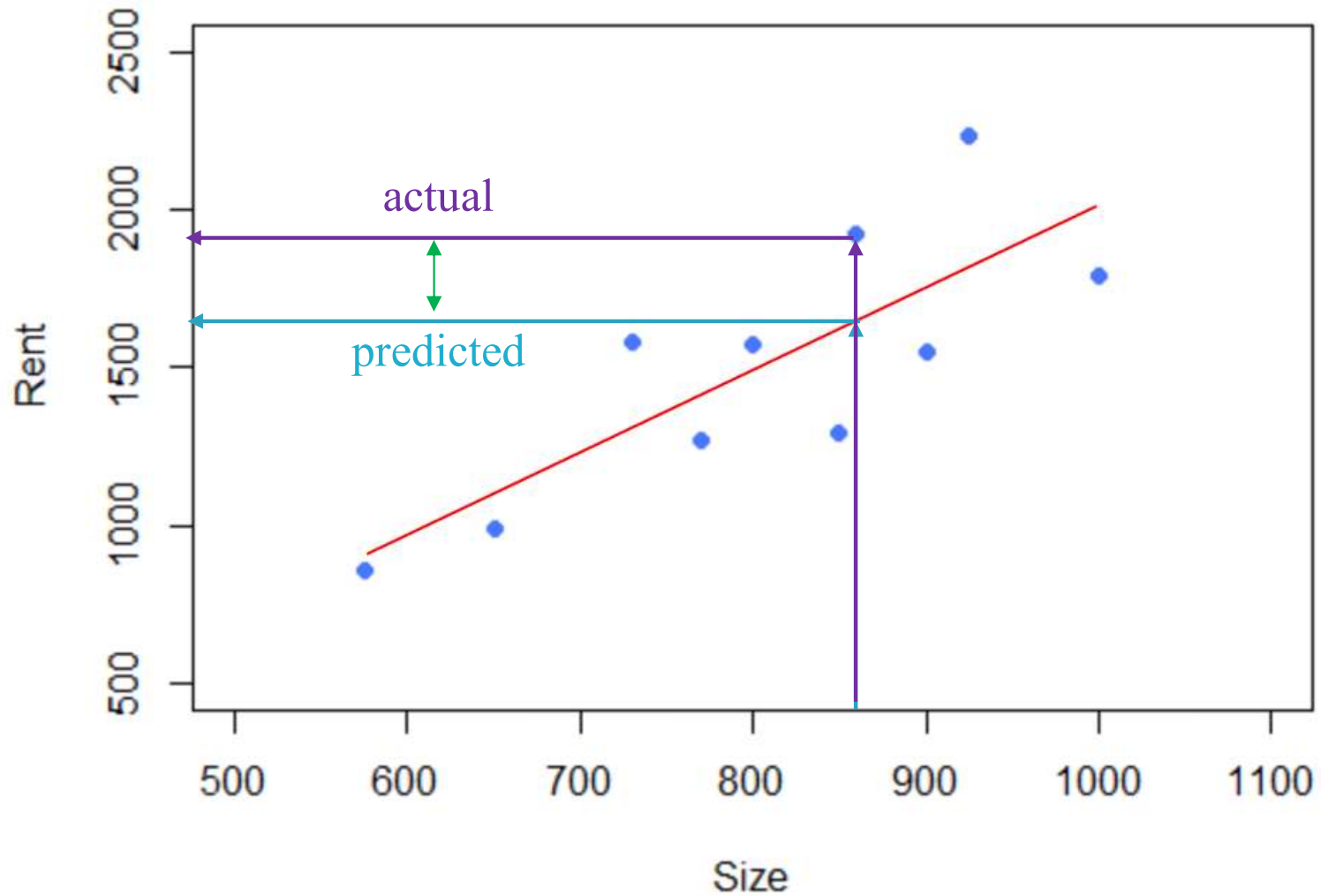
```
> lm(Rent ~ Size)$residuals
```

| | | | | |
|------------|------------|------------|------------|------------|
| 1 | 2 | 3 | 4 | 5 |
| -142.86263 | -110.73807 | 413.97648 | -325.94567 | -45.66022 |
| 6 | 7 | 8 | 9 | 10 |
| 278.04395 | 84.10623 | -221.10137 | 271.17889 | -200.99757 |



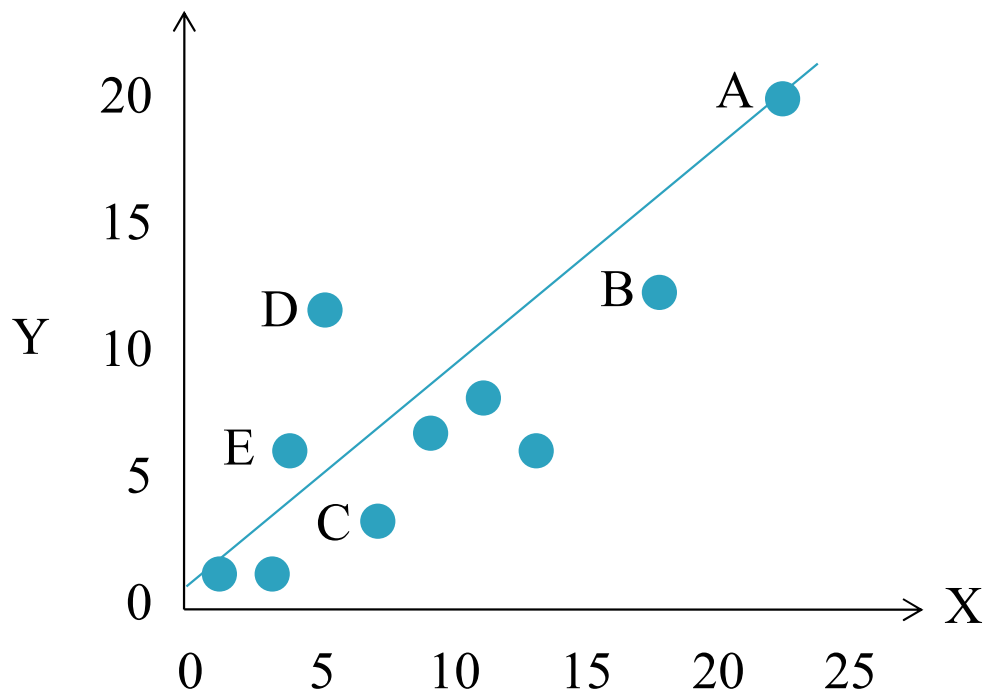
Residuals

residual = actual value of y – predicted value of y



Practice Question

Data for some explanatory variable X and some response variable Y are shown on the scatterplot below, as well as the least squares regression line:



Which of the labeled points has the largest residual?

- (A) A (B) B (C) C
(D) D (E) E

Residuals

A positive residual indicates that a point falls **above** the regression line and a negative residual indicates that it falls **below** the line. As an example, check that the residual for the 770 square foot apartment in the sample is equal to -142.90 .

Note that **it is in fact the sum of squared residuals that is minimized in calculating the least squares regression line.**

What if we want to predict the monthly rent for a 1250 square foot apartment? Our predicted value is

$$\hat{y} = -589.10 + 2.60(1250) = 2660.90$$

Extrapolation

Mathematically, there is no problem with making this prediction. However, there is a statistical problem.

Our range of values for X is from 575 to 1000 square feet. We have good evidence of a linear relationship **within this range of values**. However, we have no apartments in our sample as large as 1250 square feet, so we have no idea whether this relationship continues to hold outside our range of data.

The process of predicting a value of Y for a value of X outside our range of data is known as **extrapolation**, and should be avoided if possible.

Practice Question

Can a student's IQ score be used to predict his or her GPA? An IQ test was administered to a sample of students and their IQ score and GPA were recorded. The correlation is calculated to be 0.63 and the equation of the least squares regression line is

$$\hat{y} = 0.248 + 0.035x$$

Practice Question

This least squares regression line minimizes:

- (A) the sum of differences between actual IQs and predicted IQs.
- (B) the sum of squared differences between actual IQs and predicted IQs.
- (C) the sum of differences between actual GPAs and predicted GPAs.
- (D) the sum of squared differences between actual GPAs and predicted GPAs.
- (E) the sum of squared differences between actual IQs and predicted GPAs.

Practice Question

What is the correct interpretation of the slope of the least squares regression line?

- (A) When IQ increases by 1, we predict GPA to increase by 0.248.
- (B) When GPA increases by 1, we predict IQ to increase by 0.035.
- (C) When IQ increases by 0.035, we predict GPA to increase by 1.
- (D) When GPA increases by 0.035, we predict IQ to increase by 1.
- (E) When IQ increases by 1, we predict GPA to increase by 0.035.

Practice Question

Can a student's IQ score be used to predict his or her GPA? An IQ test was administered to a sample of students and their IQ score and GPA were recorded. The correlation is calculated to be 0.63 and the equation of the least squares regression line is

$$\hat{y} = 0.248 + 0.035x$$

What percentage of the variation in a student's GPA can be accounted for by its regression on his or her IQ?

- (A) 63.0% (B) 3.5% (C) 24.8% (D) 39.7% (E) 79.4%

Practice Question

Can a student's IQ score be used to predict his or her GPA? An IQ test was administered to a sample of students and their IQ score and GPA were recorded. The correlation is calculated to be 0.63 and the equation of the least squares regression line is

$$\hat{y} = 0.248 + 0.035x$$

One student in the sample had an IQ of 100 and a GPA of 2.4. What is the value of the residual for this student?

- (A) -3.75 (B) -1.35 (C) 0.33 (D) 3.75 (E) 1.35

Practice Question

Which of the following statements about the slope of the least squares regression line is true?

- (A) It lies between -1 and 1 , inclusive.
- (B) The larger the value of the slope, the stronger the linear relationship between the variables.
- (C) It always has the same sign as the correlation.
- (D) The square of the slope is equal to the fraction of variation in Y that is explained by regression on X .
- (E) All of the above are true.

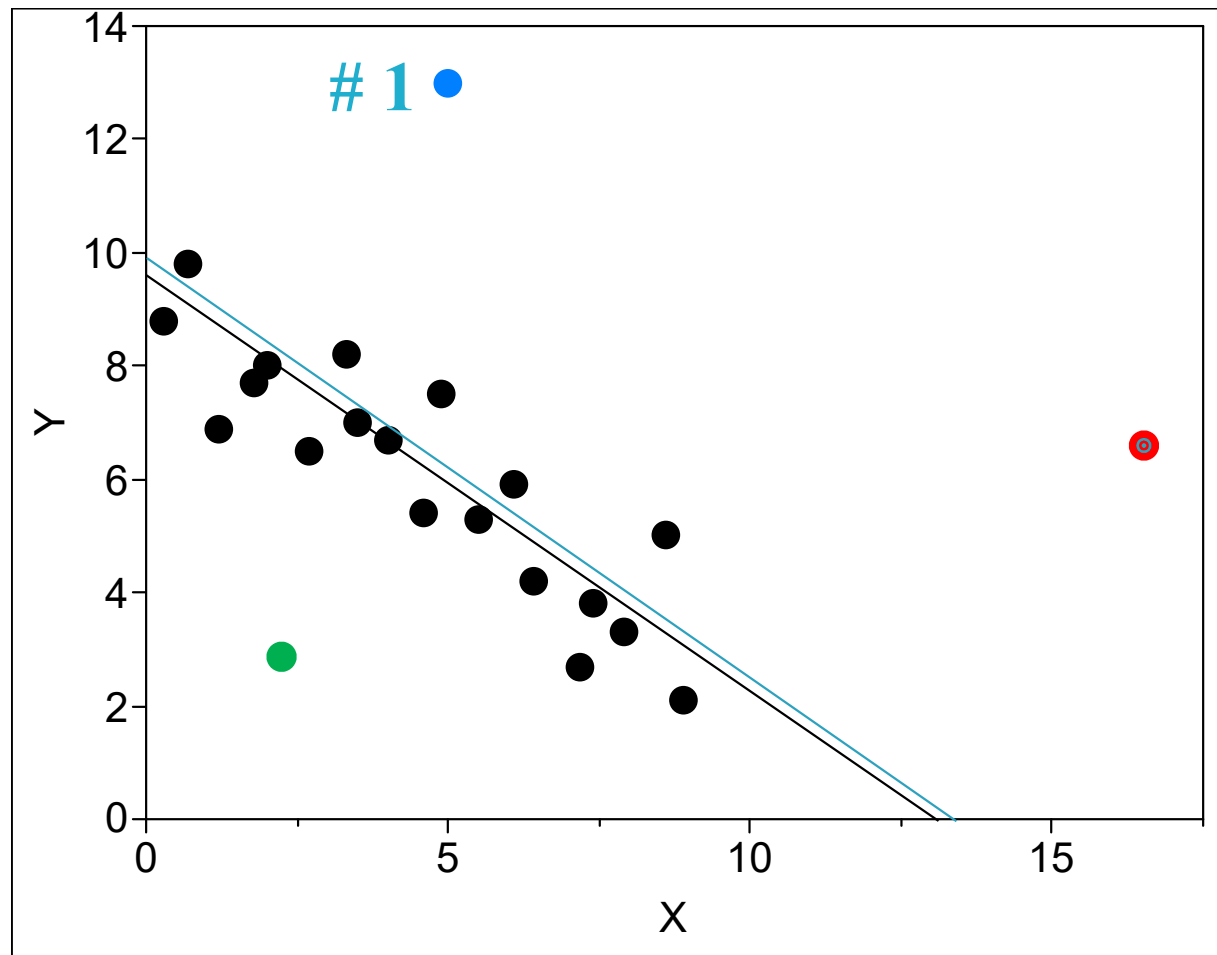
Outliers

We have seen that an outlier can be defined as a point that is far from the other data points in the x -direction or the y -direction, or if it falls outside the general pattern of points.

We now examine the effect of each of these three types of outliers.

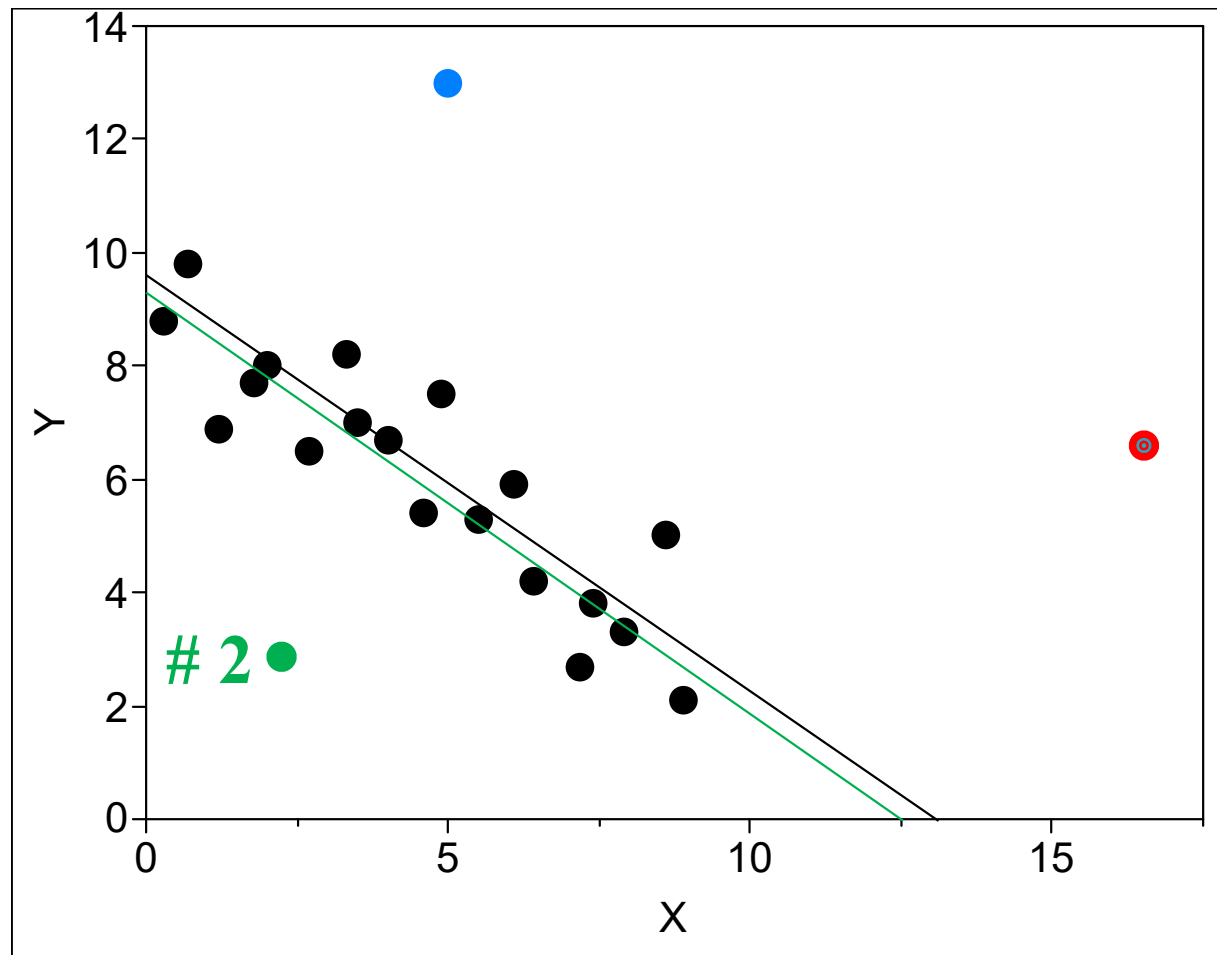
Outliers

Point # 1 is an outlier in the y -direction. It generally has little effect on the regression line.



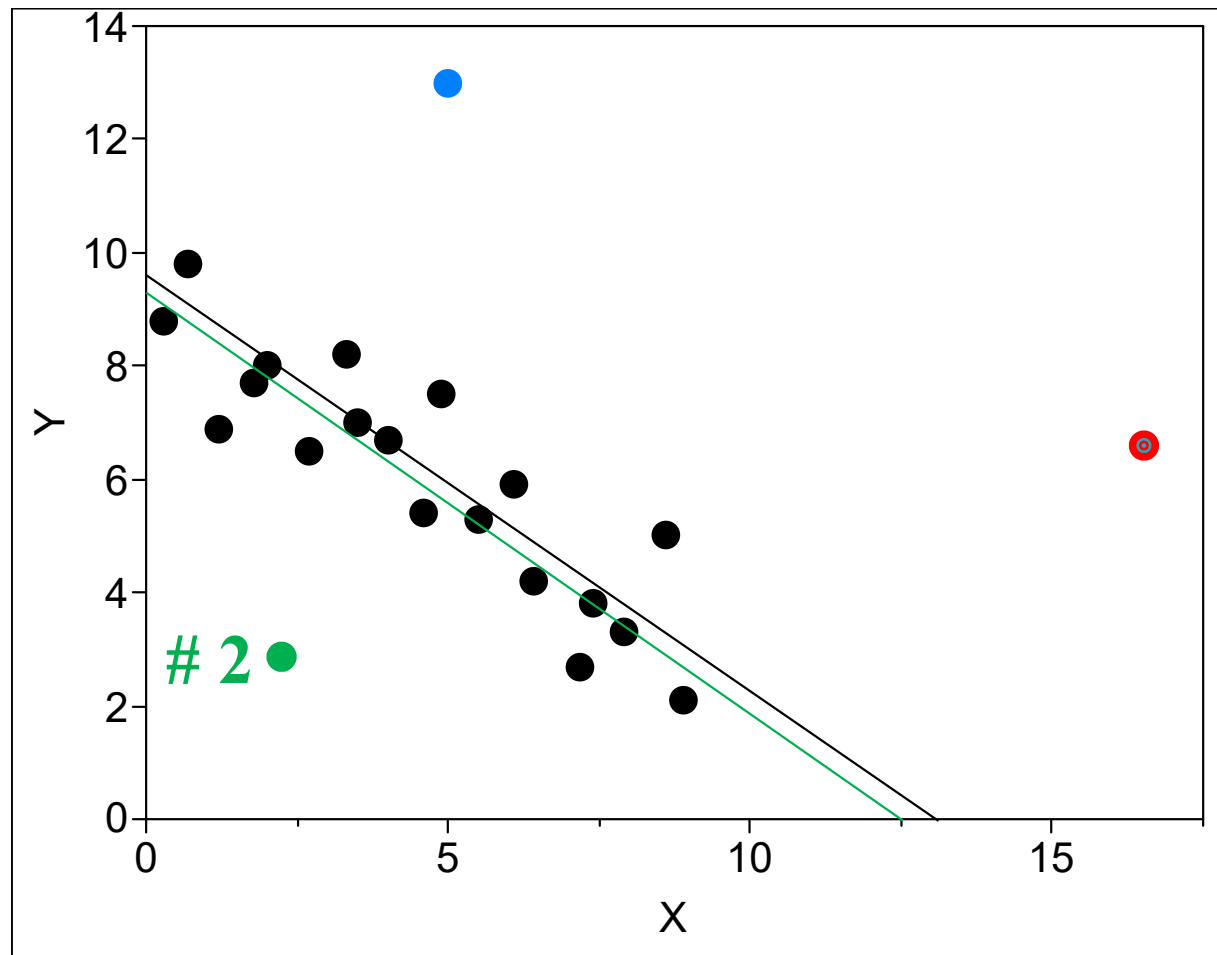
Outliers

Point # 2 is not an outlier in either the x - or y -directions, but falls outside the pattern of points.



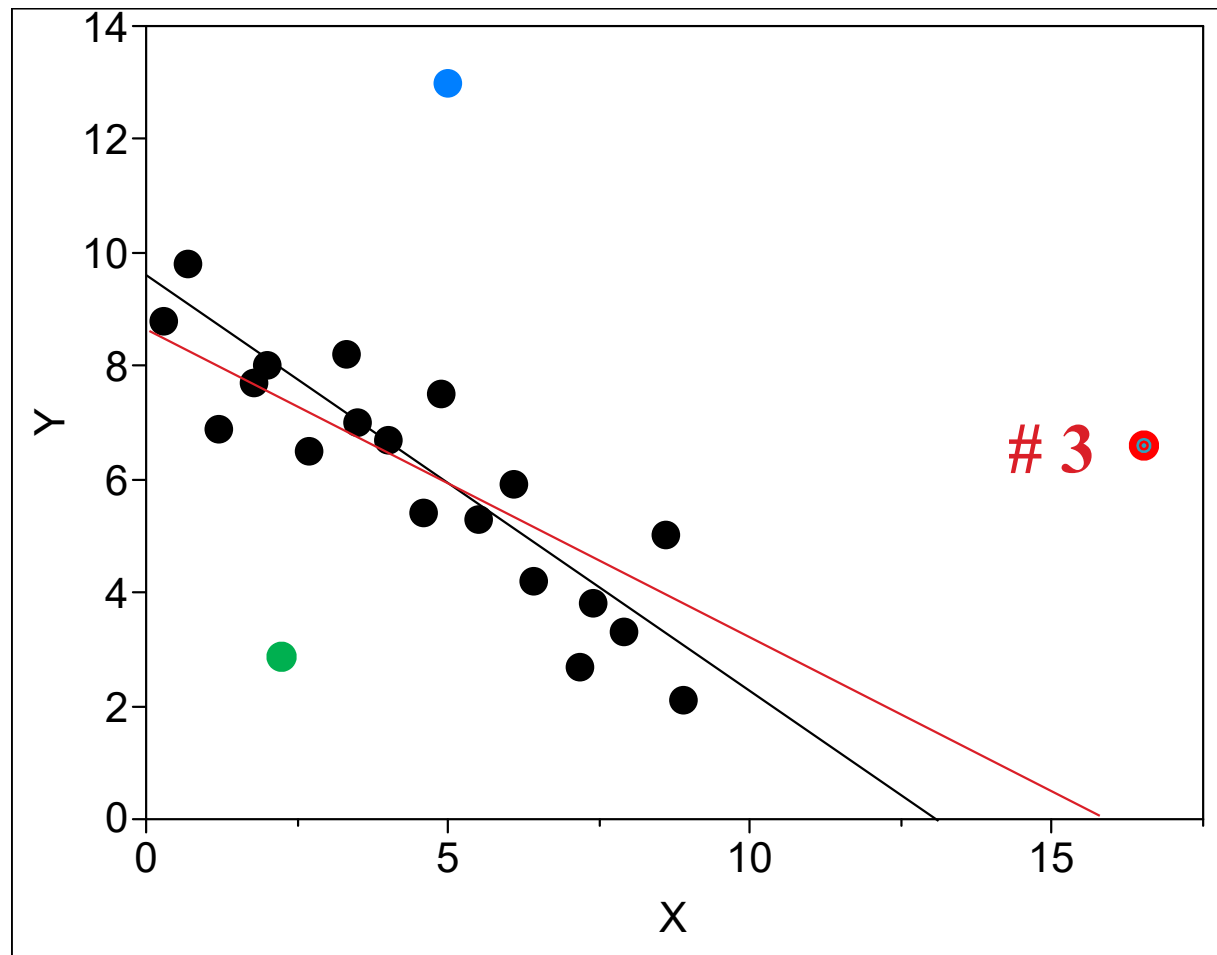
Outliers

A bivariate outlier such as this generally has little effect on the regression line.



Outliers

Point # 3 is an outlier in the x -direction. It has a strong effect on the regression line.



Influential Observations

An observation is called **influential** if removing it from the data set would dramatically alter the position of the regression line (and the value of r^2).

In the above illustration, Point # 3 is an influential observation, which is often the case for outliers in the x -direction.

Influential Observations

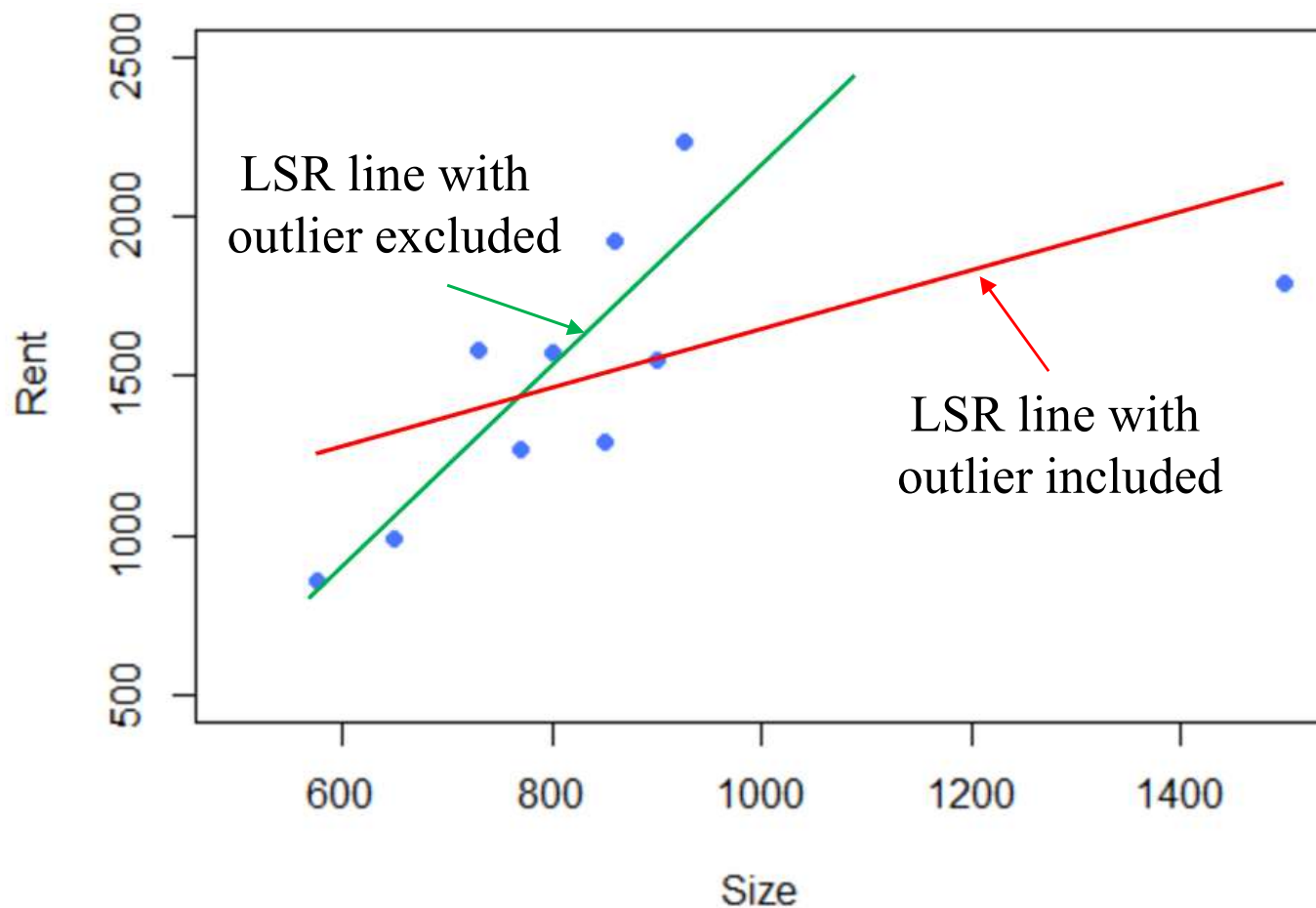
In our example, suppose the size of the largest apartment was 1500 square feet instead of 1000 square feet, and the monthly rent was still \$1790. The equation of the regression line changes to

$$\hat{y} = 705.58 + 0.94x$$

In addition, the value of r^2 reduces to 0.316. The outlying value has had a strong effect on the equation of the line and the value of r^2 .

Influential Observations

We see that, with the outlier included, the regression line is a far less accurate description of the relationship.



Estimating σ

We have seen how to estimate the parameters β_0 and β_1 in our regression model, but how do we estimate the third parameter σ ?

The standard deviation σ measures the variability of the response variable about the **true** unknown line for any given value of X . The residuals e_i measure the variability of the response variable about the **regression line** for a given value of X .

It seems natural then that our estimator for the unknown parameter σ is a function of the residuals.

Estimating σ

For simple linear regression, the estimate of σ is the **standard error**

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Notice that s_e^2 is just the “average” of the squared residuals, where we divide by $n-2$, which is the degrees of freedom for s_e .

On the next page, we see the calculations necessary in our example to find the estimated value of σ .

Example

| Apartment | Size X | Rent Y | Predicted Rent | Residual | Squared Residual |
|------------------|----------------------------|----------------------------|-----------------------|-----------------|-------------------------|
| 1 | 770 | 1270 | 1412.9 | -142.9 | 20420.41 |
| 2 | 650 | 990 | 1100.9 | -110.9 | 12298.81 |
| 3 | 925 | 2230 | 1815.9 | 414.1 | 171478.81 |
| 4 | 850 | 1295 | 1620.9 | -325.9 | 106210.81 |
| 5 | 575 | 860 | 905.9 | -45.9 | 2106.81 |
| 6 | 860 | 1925 | 1646.9 | 278.1 | 77339.61 |
| 7 | 800 | 1575 | 1490.9 | 84.1 | 7072.81 |
| 8 | 1000 | 1790 | 2010.9 | -220.9 | 48796.81 |
| 9 | 730 | 1580 | 1308.9 | 271.1 | 73495.21 |
| 10 | 900 | 1550 | 1750.9 | -200.9 | 40360.81 |

Example

Adding down the last column, we calculate

$$\begin{aligned} s_e &= \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \\ &= \sqrt{\frac{559580.9}{8}} = \sqrt{69947.6125} = 264.476 \end{aligned}$$

R Output

We save the regression line as an object with the command `regline<-lm(Rent~Size)`, and then we use the command `summary(regline)` to get the value of the standard error, as well as several other important values:

```
Call:
lm(formula = Rent ~ Size)

Residuals:
    Min       1Q   Median       3Q      Max
-325.9 -186.5  -78.2   224.4   414.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -589.9366    556.1895  -1.061   0.31981
Size          2.6010     0.6822   3.813   0.00514 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 264.5 on 8 degrees of freedom
Multiple R-squared: 0.645,    Adjusted R-squared: 0.6006
F-statistic: 14.54 on 1 and 8 DF,  p-value: 0.005143
```

intercept

slope

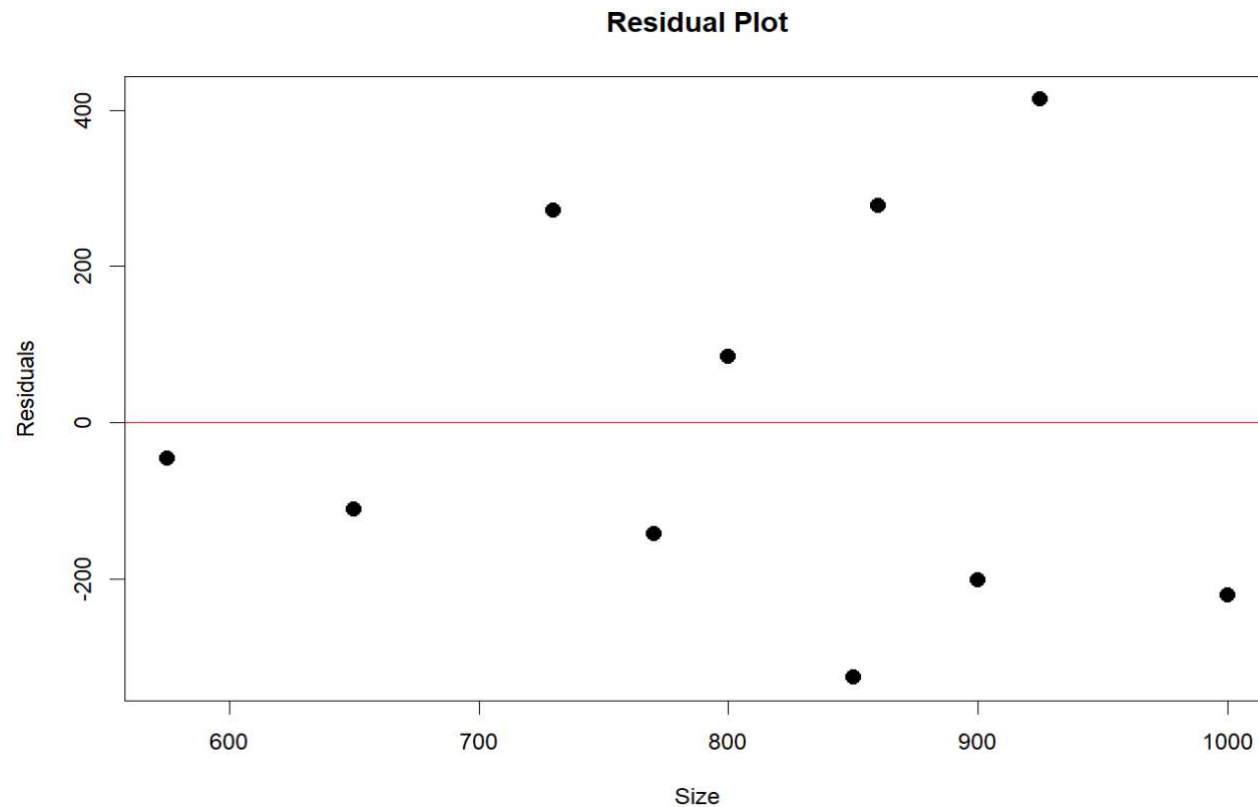
standard error

r^2

Residual Plots

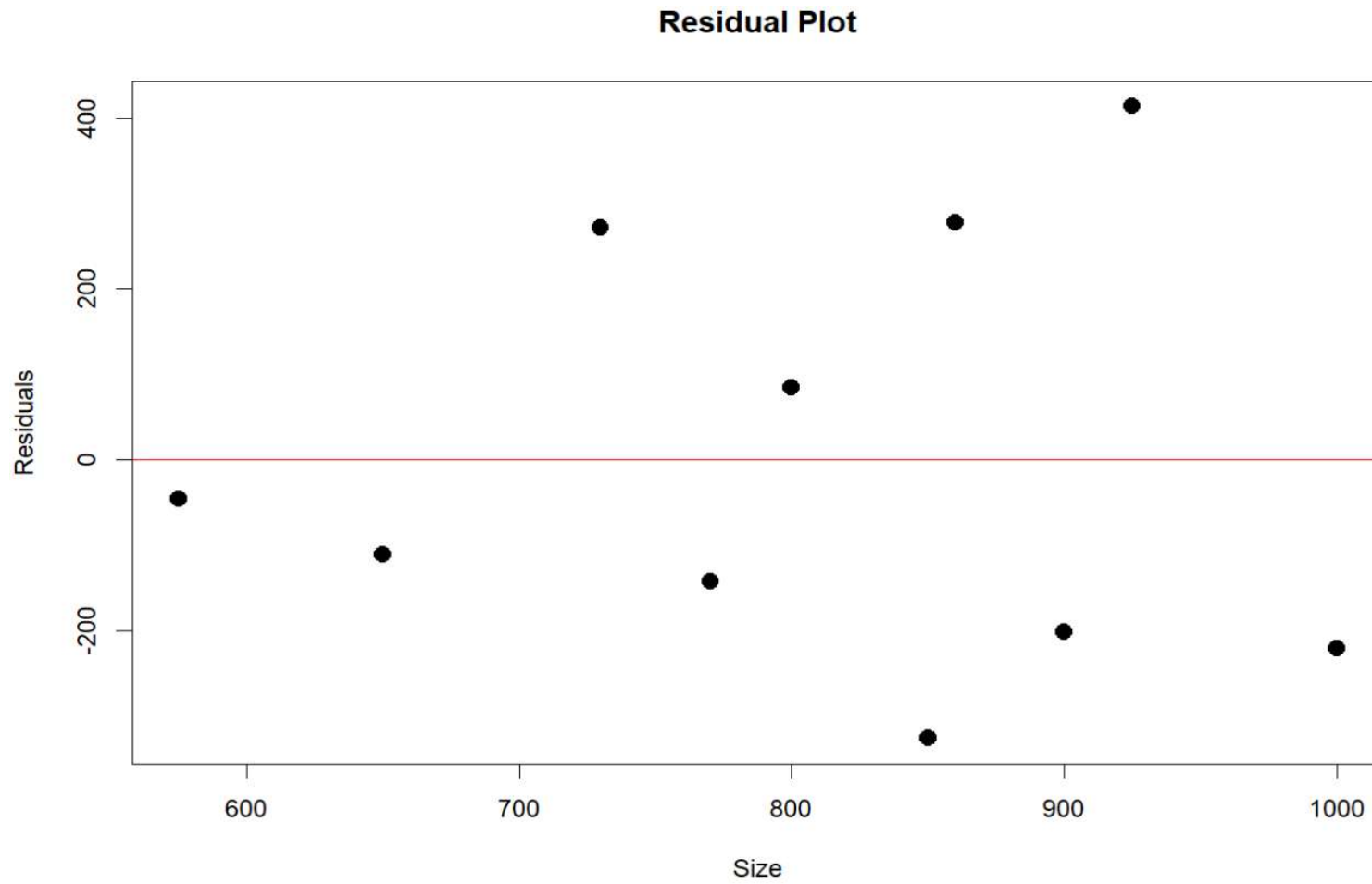
Residual analysis can be very helpful in assessing whether the model assumptions have been satisfied.

Consider the following **residual plot** of our data:



Residual Plots

A residual plot is just a plot of the residuals vs. the explanatory variable.



R Code

```
> regline<-lm(Rent~Size)
> residuals<-resid(regline)
> plot(Size, residuals, pch=19, cex=1.5,
      xlab="Size", ylab="Residuals",
      main="Residual Plot", abline(h=0,col="red"))
```

Residual Plots

Looking at the residuals this way helps us see the deviations from the line better, and helps us assess whether our model assumptions are valid.

The residuals appear to be **randomly scattered** about the line, **with no obvious patterns or outliers**.

In fact, this is exactly the type of residual plot we look for when trying to verify our regression assumptions.

Residual Plots

We assumed that the mean of Y has a linear relationship with X and that the standard deviation of Y is constant over all values of X . There is no pattern in the residuals and the variability seems to remain fairly constant over all values of X . This plot indicates that we have no reason to question our assumptions.

The **sum** of the residuals is equal to **zero**, so the **mean** of the residuals is also zero. It can be shown that **this is always the case in least squares regression**. A horizontal line is drawn at the value zero for reference.

Normality of Residuals

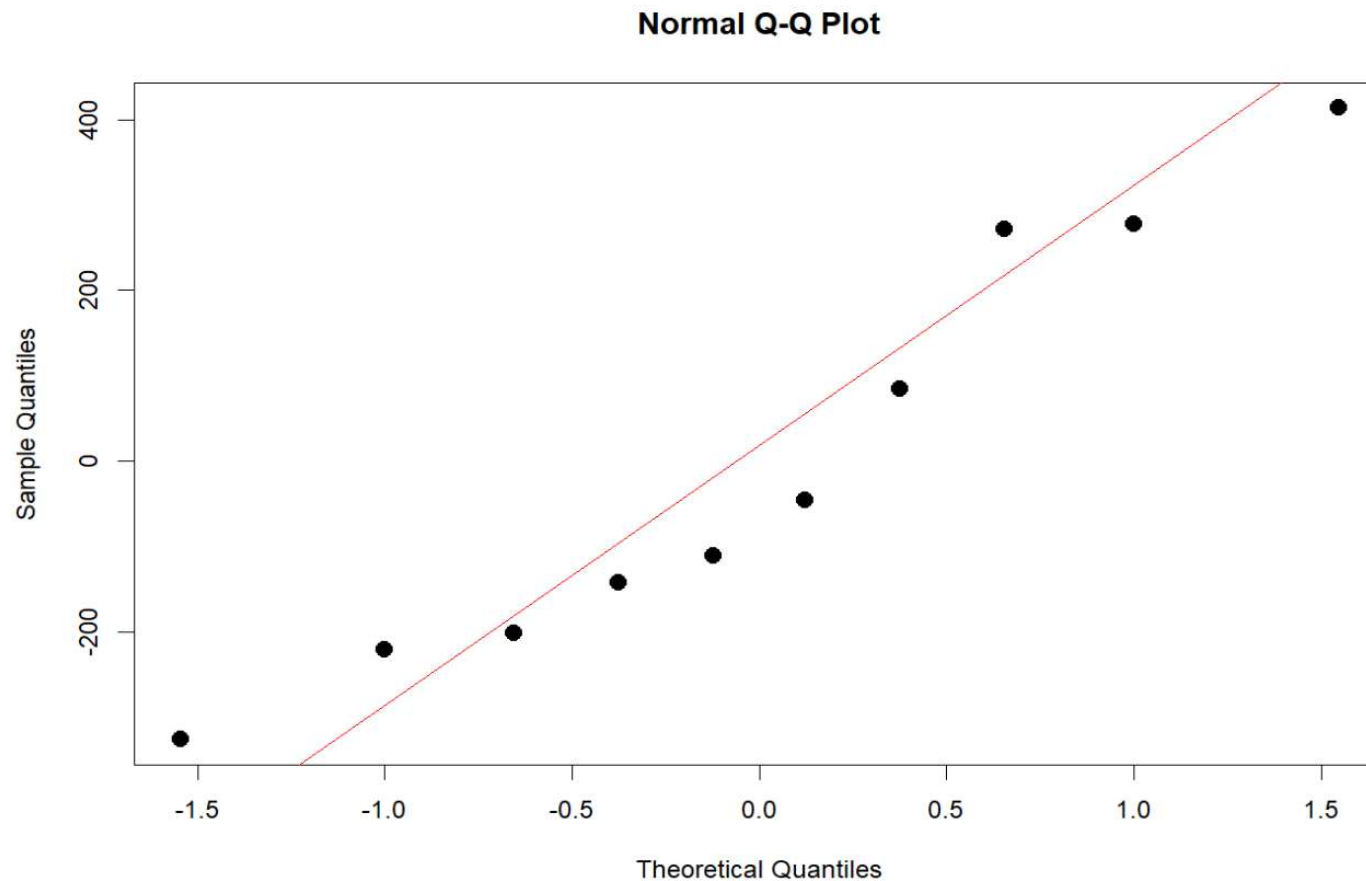
Recall that we also assumed that the error term in the simple linear regression model follows a normal distribution.

If this is the case, we expect the residuals to have an approximate normal distribution.

We create a normal quantile plot of the residuals to verify the assumption.

Normality of Residuals

```
> qqnorm(residuals, cex=1.5, pch=19)  
> qqline(residuals, col="red")
```



Normality of Residuals

All of the points fall fairly close to the diagonal line, so it appears that normality of the error terms is a reasonable assumption.

(Remember that with a small sample size, we can't expect to see a perfect normal distribution.)

Violation of Model Assumptions

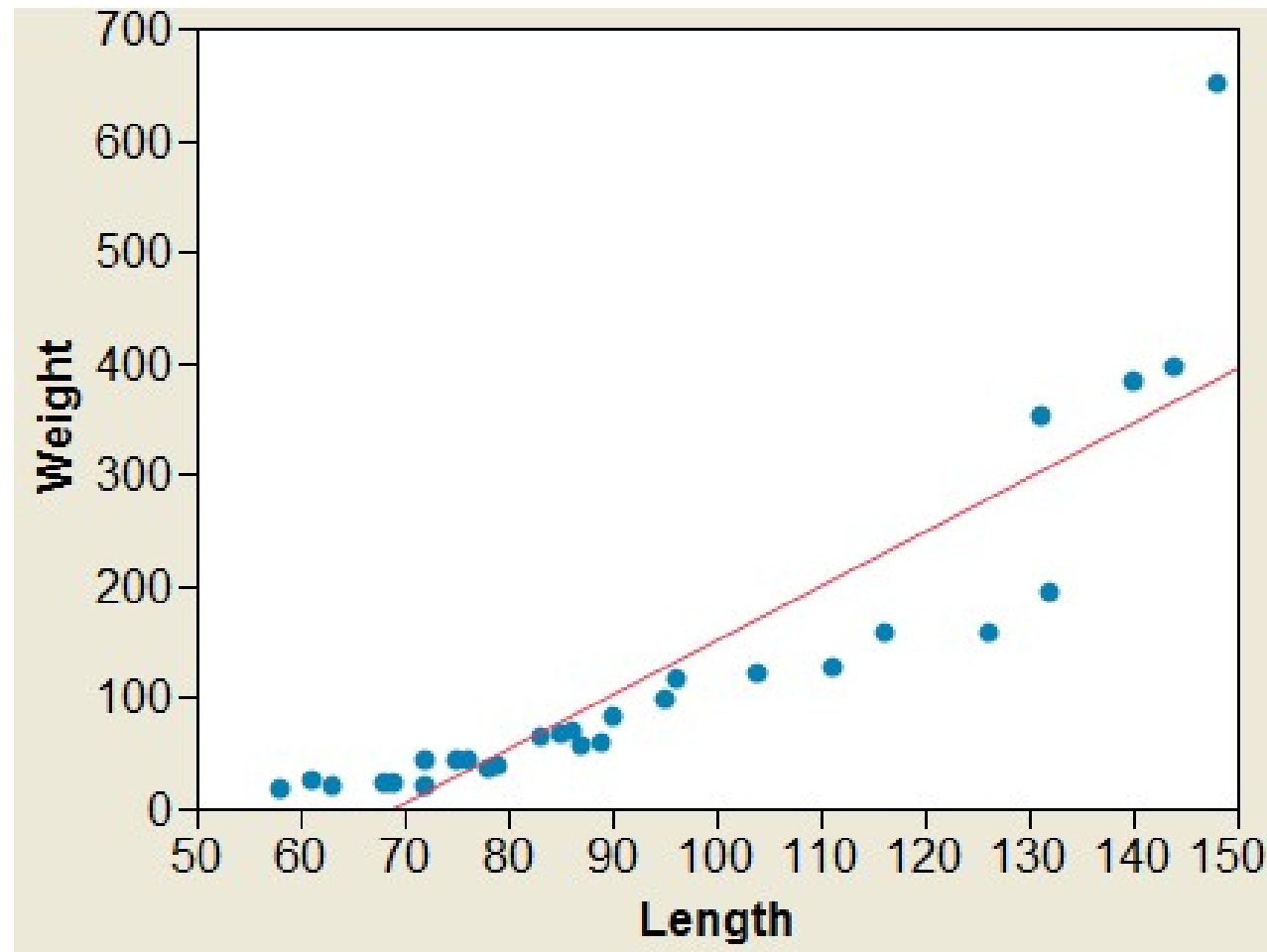
We now examine two examples where our assumptions are shown to be invalid.

Can the length of an alligator be used to predict its weight?

Scientists in Florida can measure the lengths of alligators quite precisely from an airplane flying overhead. The weights of the alligators, however, are not possible to determine in this manner. A sample of 30 alligators is caught and their lengths X (in inches) and weights Y (in pounds) are measured.

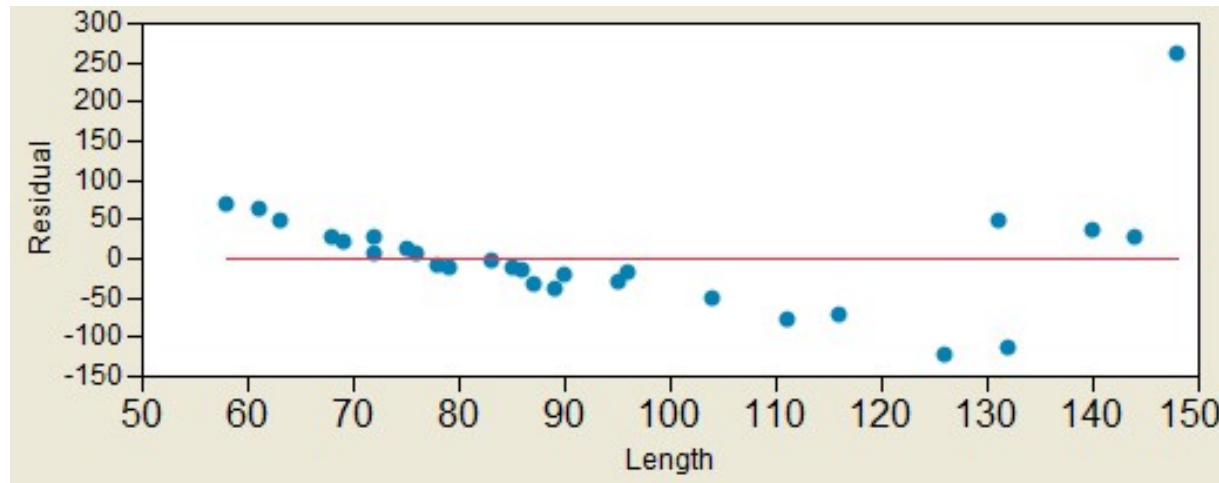
Violation of Model Assumptions

The following is a scatterplot with the least squares regression line for the data:



Violation of Model Assumptions

The residual plot is shown below:

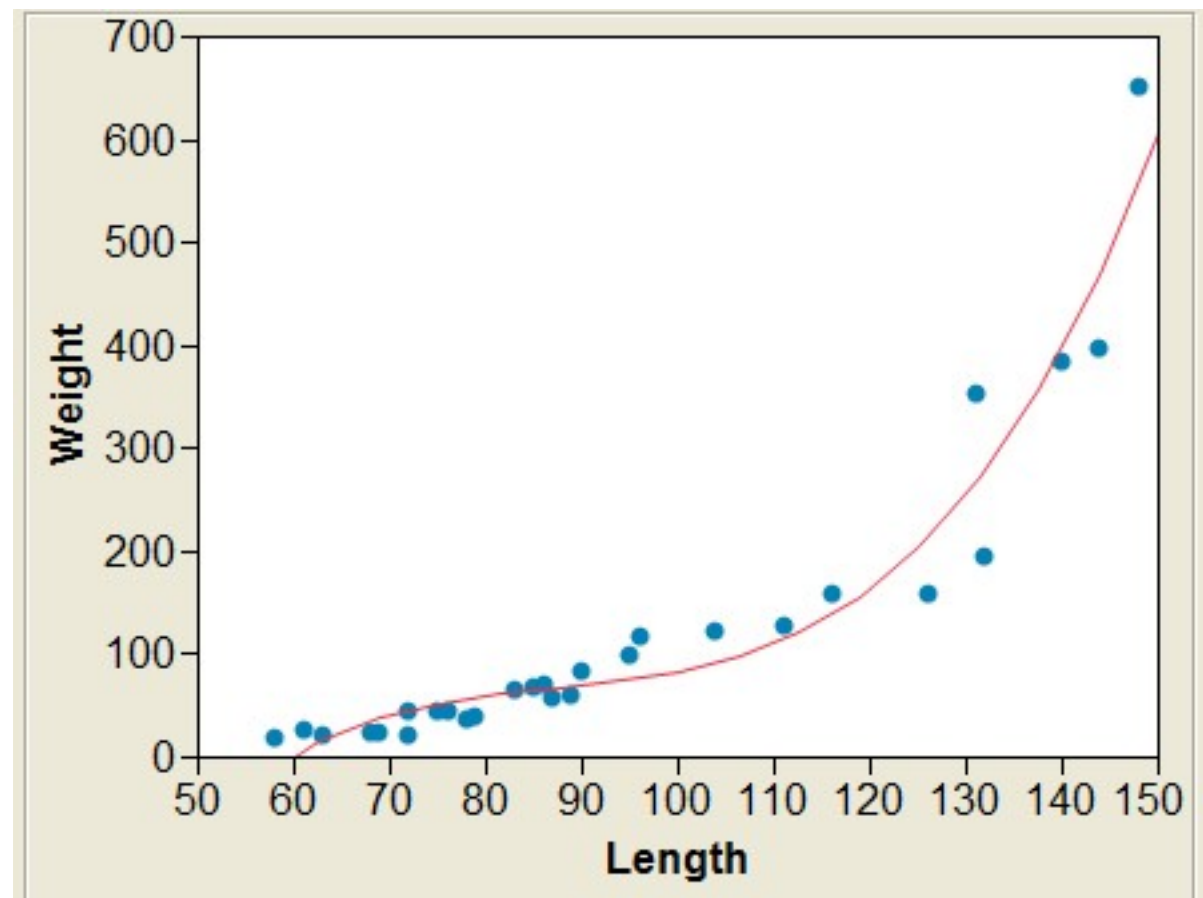


We see that the residuals have a **pattern**. Residuals for small and large lengths are positive, while residuals for medium-length alligators are negative.

This indicates that the assumption of a linear relationship between the mean of Y and X is violated. Linear regression is not appropriate in this example.

Violation of Model Assumptions

In fact, the assumption of a cubic relationship is much more appropriate in this case, as shown by the cubic polynomial fit to the data:

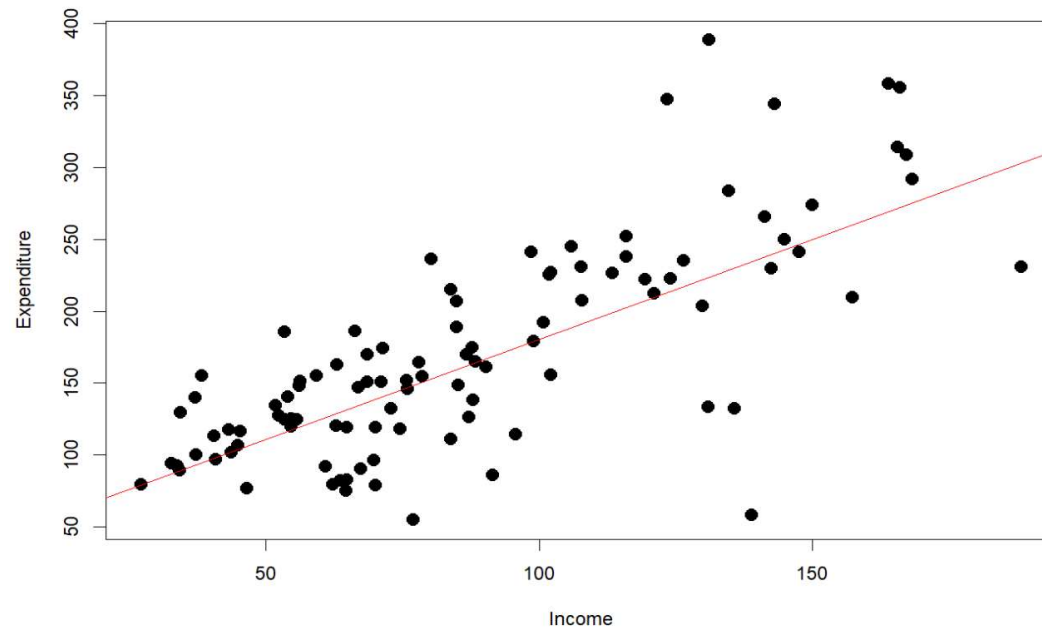


Violation of Model Assumptions

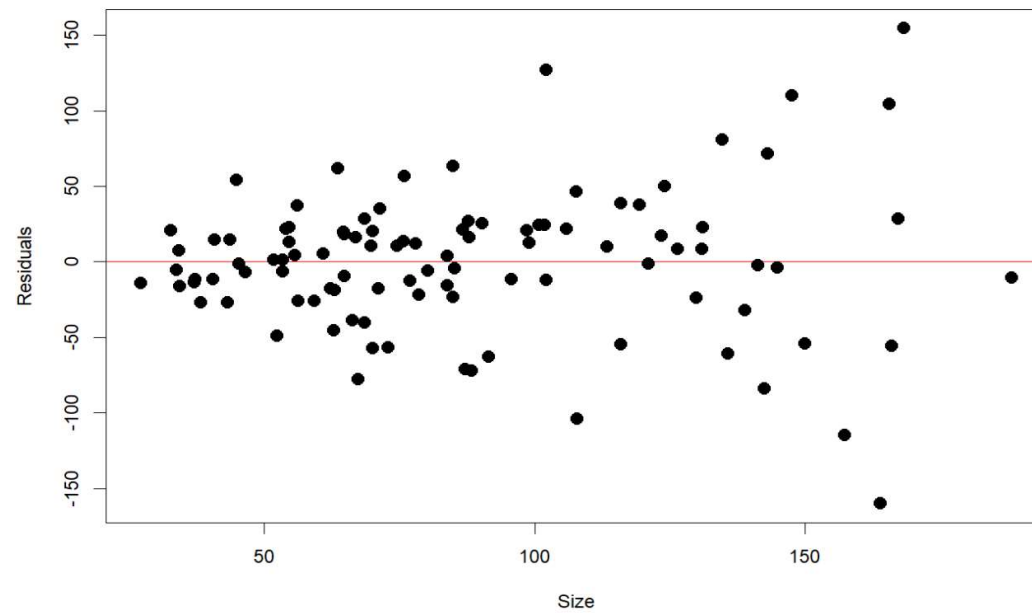
Do people who make more money spend more on groceries? The annual incomes (in thousands of dollars) and weekly grocery expenses (in \$) were recorded for a sample of 101 people.

A scatterplot of Expenses vs. Income is shown on the next page with the least squares regression line added. The residual plot is also shown.

Violation of Model Assumptions

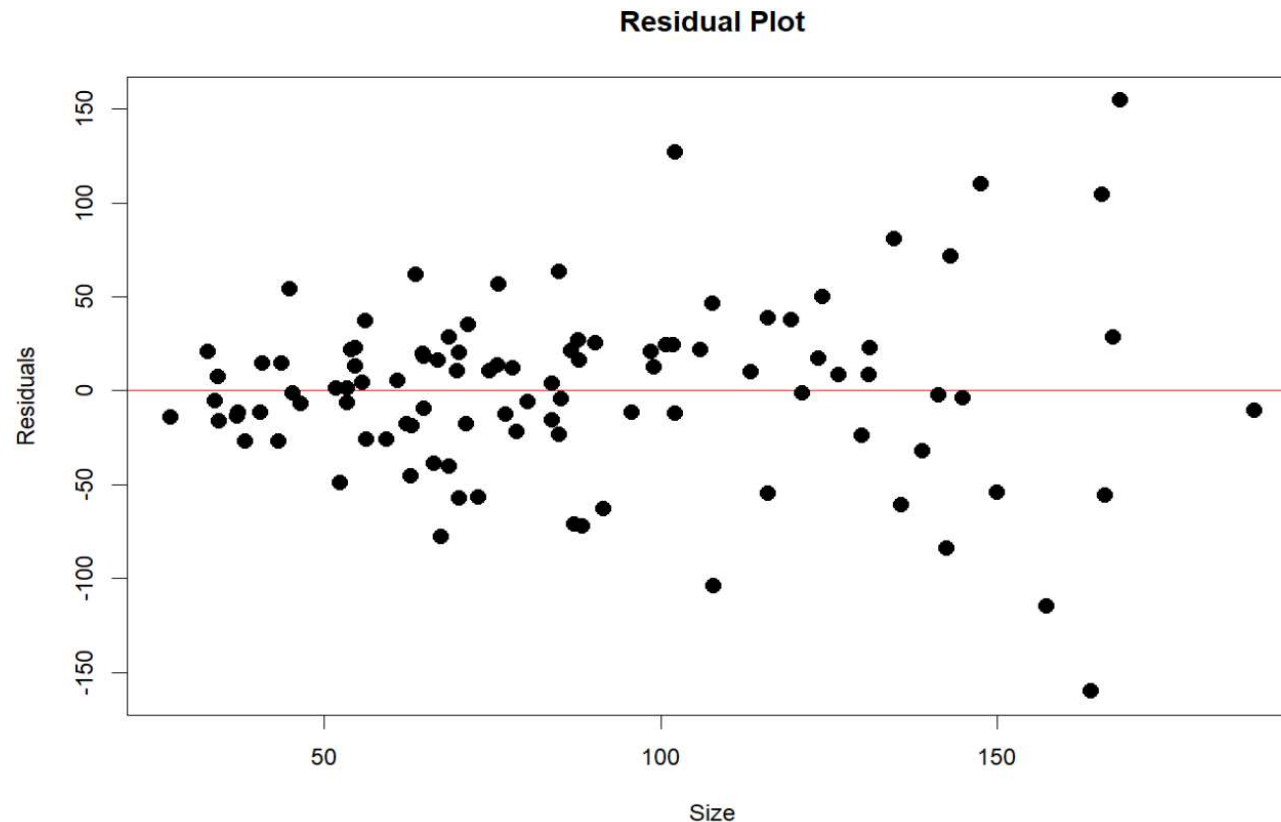


Residual Plot



Violation of Model Assumptions

In this case, the overall relationship appears to be linear. However, the magnitudes of the residuals appear to **increase** as X increases, indicating that our assumption of constant variance for Y over all values of X is not valid.



Violation of Model Assumptions

It should be noted that, in the previous two examples, the departures from our assumptions would likely have been noticed from the scatterplot before the regression was even run. Residual plots are more useful in **multiple linear regression** (when there is more than one explanatory variable). This topic, however, is not covered in this course.

There are remedies in situations such as these to “correct” for nonlinearity or non-homogeneity of variance. Data values can be “transformed” so the assumptions are met, and linear methods can be used on the transformed data. In this course, however, we will deal only with situations in which the relationship is already linear in nature.

Confidence Intervals for μ_Y

The main purpose of fitting our regression line was to enable us to **predict** the value of Y for a given value of X .

We are often not satisfied with just the predicted value, however. As is the case when we use the sample mean to estimate the population mean μ , we would like to be able to construct **confidence intervals** in the regression setting.

Confidence Intervals for μ_Y

We would like to estimate the **mean** value of the response variable Y for any given value of the explanatory variable X .

The predicted (estimated) value of y is

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

where x^* is the value of X in which we are interested.

We will construct a confidence interval for

$$\mu_Y = \beta_0 + \beta_1 x^*$$

where μ_Y is the mean value of the response variable Y when $X = x^*$.

Confidence Intervals for μ_Y

A **level C** confidence interval for the mean response μ_Y when $X = x^*$ is

$$\hat{y} \pm t^* SE_{\hat{\mu}}$$

where the standard error $SE_{\hat{\mu}}$ is

$$SE_{\hat{\mu}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Confidence Intervals for μ_Y

The value t^* is the upper $\alpha/2$ critical value from the t distribution with $n - 2$ degrees of freedom.

In practice, calculating these intervals is quite tedious and we will usually use computer software to help us. However, for this example, we will go through the calculations to see how they are done.

Example

Suppose we want a 95% confidence interval for the true mean monthly rent for all 900 square foot apartments.

We have already calculated the value of the standard error s_e (the estimate of the standard deviation σ) to be $s_e = 264.476$.

Example

The **predicted value** of Rent when $X = 900$ is

$$\hat{y} = -589.10 + 2.60(900) = 1750.90$$

We also calculate

$$(x^* - \bar{x})^2 = (900 - 806)^2 = 8836$$

We previously calculated

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 150290$$

Example

The standard error $SE_{\hat{\mu}}$ for estimating the mean value of Y when $X = 900$ is therefore:

$$SE_{\hat{\mu}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$
$$= 264.476 \sqrt{\frac{1}{10} + \frac{8836}{150290}} = 105.39$$

Example

Our 95% confidence interval for μ_y when $X = 900$ is therefore

$$\begin{aligned}\hat{y} \pm t^* SE_{\hat{\mu}} &= 1750.90 \pm 2.306(105.39) \\ &= 1750.90 \pm 243.03 = (1507.87, 1993.93)\end{aligned}$$

where $t^* = 2.306$ is the upper 0.025 critical value of the t distribution with $n - 2 = 10 - 2 = 8$ degrees of freedom.

Example

We interpret the 95% confidence interval as follows:

If we took repeated samples of 10 apartments, fit the least squares regression line and calculated the interval in a similar manner, 95% of such intervals would contain the true mean monthly rent of all 900 square foot apartments in the city.

Confidence Interval for the Slope

We may also want to construct a confidence interval for the slope β_1 of the true unknown line. It turns out that the standard error for the estimator b_1 is also a function of the standard error s_e .

A level C confidence interval for the slope β_1 of the true unknown line is

$$b_1 \pm t^* SE_{b_1}$$

where the standard error SE_{b_1} of the least squares regression slope is

$$SE_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Confidence Interval for the Slope

The value t^* in the above formula is again the upper $\alpha/2$ critical value of the t distribution with $n - 2$ degrees of freedom.

Let us calculate a 95% confidence interval for the slope of the true unknown line for Rent vs. Size.

The standard error of the least squares regression slope is

$$SE_{b_1} = \frac{s_e}{\sqrt{(x_i - \bar{x})^2}} = \frac{264.476}{\sqrt{150290}} = 0.6822$$

Example

Our 95% confidence interval for the slope β_1 is therefore

$$\begin{aligned} b_1 \pm t^* SE_{b_1} &= 2.60 \pm 2.306(0.6822) \\ &= 2.60 \pm 1.57 = (1.03, 4.17) \end{aligned}$$

where $t^* = 2.306$ is the upper 0.025 critical value of the t distribution with $n - 2 = 10 - 2 = 8$ degrees of freedom.

Example

This interval can be interpreted as follows: If we were to repeatedly take random samples of ten apartments and calculate the least squares regression line, and calculate an interval in a similar manner, then 95% of such confidence intervals would contain the true value of β_1 .

We can also conduct a hypothesis test to determine if there is a **linear relationship** between an explanatory variable X and a response variable Y .

We will conduct a test to determine whether a true linear relationship exists between Size and Rent.

Hypothesis Test for the Slope

Let $\alpha = 0.05$.

We are testing the hypotheses

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

Note that the null hypothesis states that the **true slope** of the **true unknown line** is zero, which implies that Y does not change when X changes. This is equivalent to saying the two variables do not have a linear relationship, whereas the alternative hypothesis claims that they do.

Test Statistic

We will reject the null hypothesis if the P-value $\leq \alpha = 0.05$.

The test statistic is

$$t = \frac{b_1}{SE_{b_1}} = \frac{2.60}{0.6822} = 3.81$$

Under the null hypothesis, this test statistic follows a t distribution with $n - 2 = 8$ degrees of freedom.

Example

The P-value is $2P(T(8) \geq 3.81)$. We see from Table 2 that

$$P(T(8) \geq 3.355) = 0.005 \quad \text{and} \quad P(T(8) \geq 3.833) = 0.0025$$

Since $3.355 < t = 3.81 < 3.833$, it follows that $P(T(8) \geq 3.81)$ is between 0.0025 and 0.005. But our P-value is double this probability, and so the P-value is between 0.005 and 0.01.

We can obtain the exact P-value from R as follows:

```
> 2*pt(3.81,8,lower.tail=FALSE)  
[1] 0.005162634
```

Example

Since we saved the regression line as an object in R, can use the `summary()` command to get the value of the standard error:

call:

```
lm(formula = Rent ~ Size)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|--------|--------|--------|-------|-------|
| -325.9 | -186.5 | -78.2 | 224.4 | 414.0 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|------------|
| (Intercept) | -589.9366 | 556.1895 | -1.061 | 0.31981 |
| Size | 2.6010 | 0.6822 | 3.813 | 0.00514 ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

test statistic

P-value

Residual standard error: 264.5 on 8 degrees of freedom

Multiple R-squared: 0.645, Adjusted R-squared: 0.6006

F-statistic: 14.54 on 1 and 8 DF, p-value: 0.005143

Example

Since the P-value $< \alpha = 0.05$, we reject H_0 . At the 5% level of significance, we have sufficient evidence to conclude that there exists a linear relationship between Size and Rent, i.e., that $\beta_1 \neq 0$.

The P-value can be interpreted as follows:

If there was no linear relationship between Size and Rent (i.e., if the slope of the true line was equal to zero), the probability of observing a value of b_1 at least as extreme as 2.60 would be between 0.005 and 0.01.

Example

Suppose we had instead used the critical value method.

Our decision rule would be to reject the null hypothesis if $|t| \geq 2.306$, where $t^* = 2.306$ is the upper 0.025 critical value of the t distribution with $n - 2 = 8$ degrees of freedom.

Our conclusion would be to reject the null hypothesis, since $t = 3.81 > t^* = 2.306$.

.

Example

Since this was a two-sided test, and since the confidence level of our interval and the level of significance of the test add up to one, we could have used the confidence interval to conduct the test.

Since the value 0 falls outside the 95% confidence interval for β_1 , we reject H_0 at the 5% level of significance.

Example

This conclusion should come as no surprise to us, because we have seen the scatterplot already and it is quite obvious that a linear relationship between Apartment Size and Rent is a reasonable model. In practice (as will be shown in the next example), checking for the existence of a linear relationship should be done before any analysis is done on the data.

Practice Question

A researcher recorded the number of bees in a sample of 13 hives, as well as the amount of honey produced in each hive.

We would like to conduct a t test of $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$ at the 1% level of significance. The critical region for the test is:

(A) $|t| \geq 2.328$

(B) $|t| \geq 2.681$

(C) $|t| \geq 2.718$

(D) $|t| \geq 3.055$

(E) $|t| \geq 3.106$

Example

Do people buy more ice cream when it's hot outside? Researchers collected data from nineteen American cities for the month of August. The Average Monthly Temperature X (in °F) was recorded for each city, as well as the estimated ice cream consumption Y (in pints per capita, estimated from data collected from several stores that sell ice cream). The data are shown on the next page.

Example

| City | Average Temperature | Ice Cream Consumption |
|-----------------|---------------------|-----------------------|
| Akron, OH | 70.3 | 0.613 |
| Casper, WY | 68.7 | 0.421 |
| Houston, TX | 82.3 | 0.287 |
| Louisville, KY | 75.8 | 0.337 |
| Anchorage, AK | 56.3 | 0.166 |
| Raleigh, NC | 77.1 | 0.650 |
| Worcester, MA | 68.0 | 0.153 |
| Peoria, IL | 73.1 | 0.453 |
| Minneapolis, MN | 70.5 | 0.490 |
| Phoenix, AZ | 91.5 | 0.625 |

| City | Average Temperature | Ice Cream Consumption |
|----------------|---------------------|-----------------------|
| Fargo, ND | 68.8 | 0.506 |
| Newark, NJ | 76.4 | 0.233 |
| Orlando, FL | 82.5 | 0.621 |
| Olympia, WA | 63.3 | 0.381 |
| Sacramento, CA | 75.1 | 0.536 |
| Tulsa, OK | 81.5 | 0.324 |
| Savannah, GA | 81.0 | 0.609 |
| Glasgow, MT | 69.4 | 0.472 |
| Madison, WI | 68.3 | 0.211 |

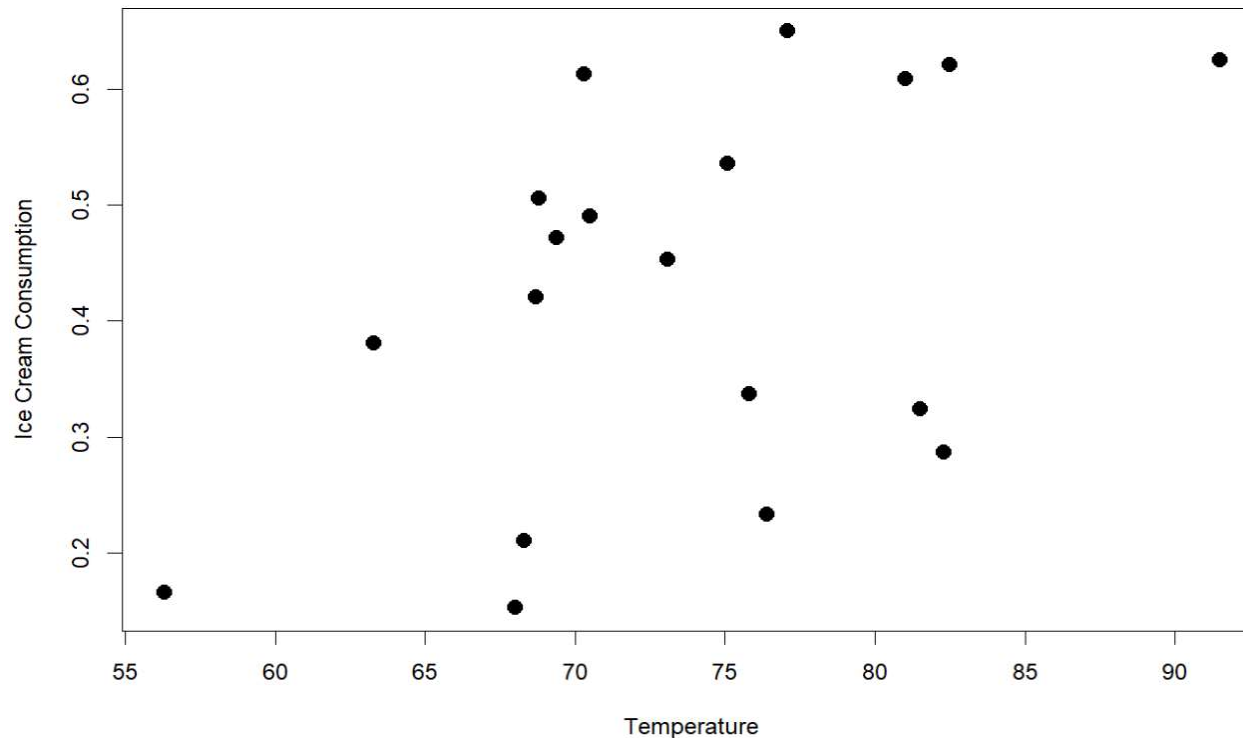
Example

In the previous example, our focus was on learning the appropriate methods of analysis.

In this example, we will also focus on the proper order of our analysis.

The first thing we do in any regression problem is to look at a scatterplot of the data.

Example



There appears to be a weak positive association. There are no obvious outliers, and although the points do not fall very close to a straight line, there is no other obvious functional relationship either. We will test for a linear relationship shortly.

Example

Our simple linear regression model is as follows:

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

in which:

y_i is the ice cream consumption for the i^{th} city in the population,

β_0 , and β_1 are parameters,

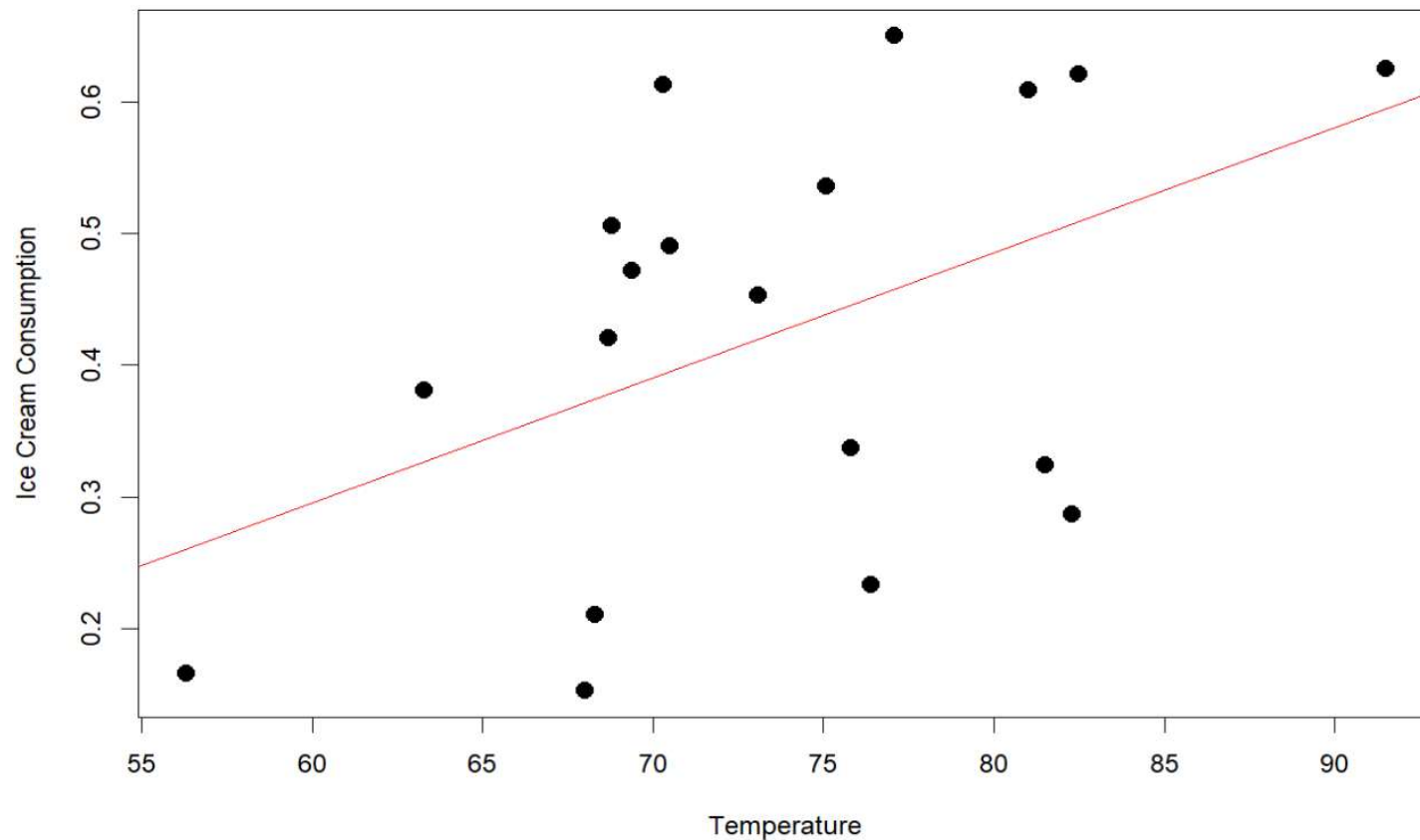
x_i is the average August temperature for the i^{th} city in the population,

ϵ_i is a random error term such that $\epsilon_i \sim N(0, \sigma)$.

Example

From R, the least squares regression line is calculated to be

$$\hat{y} = -0.273 + 0.00949x$$



Example

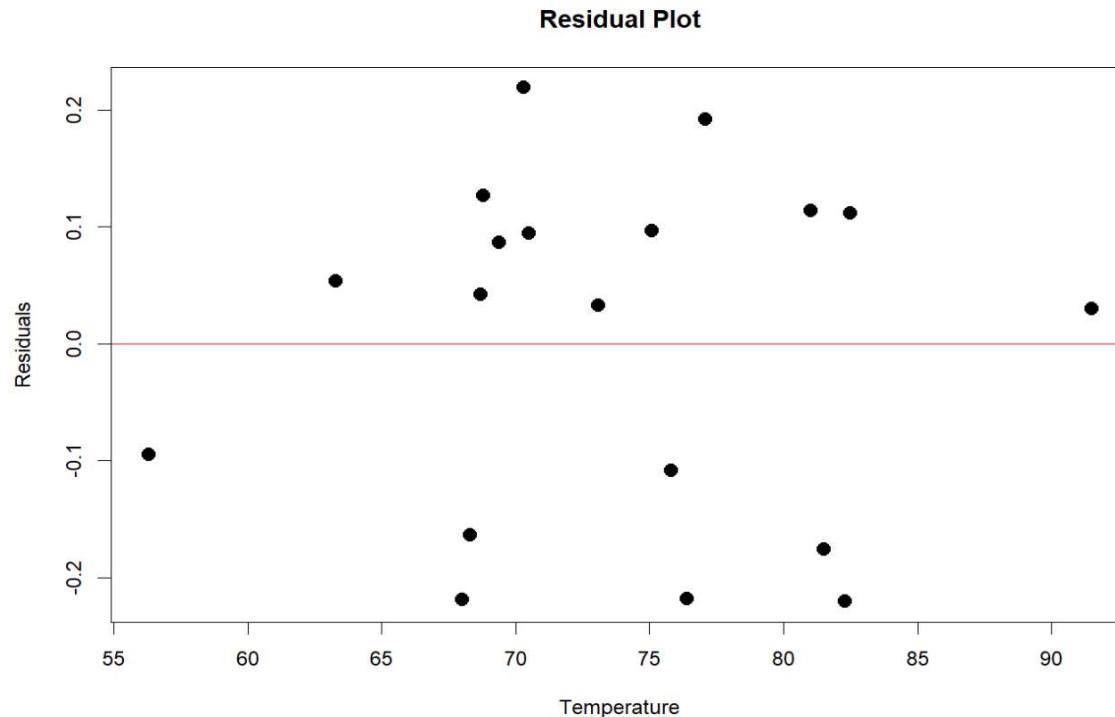
Following the same steps as the previous example (Apartment Rent vs. Size), we can calculate the following:

$$\sum (x_i - \bar{x})^2 = 1169.8116$$

$$\sum (y_i - \hat{y}_i)^2 = 0.3817$$

Example

Our next step is to examine a plot of the residuals:



There are no obvious patterns or outliers, and the points appear to be randomly scattered about the line. We have no reason to doubt the validity of our assumptions.

Example

From R, the correlation is $r = 0.465$, so $r^2 = 0.2162$, which is fairly low compared to previous examples we have seen.

This means that only about 21.62% of the variation in Ice Cream Consumption can be attributed to its regression on Temperature.

In real-life problems, it is quite common to see lower values of r^2 .

Example

When we have a response variable Y , there are usually many different explanatory variables that affect the value of Y . There are obviously many other variables that affect how much ice cream people consume, but we have only considered one – the Average Temperature of a city in August.

To be able to account for more than 20% of the variation in the response variable by examining only one explanatory variable is still quite good.

Example

Is there evidence of a linear relationship between temperature and ice cream consumption at the 10% level of significance?

Let $\alpha = 0.10$.

We are testing the hypotheses

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 \neq 0$$

Example

The estimate of the parameter σ in the regression model is

$$\begin{aligned}\hat{\sigma} = s_e &= \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \\ &= \sqrt{\frac{0.3817}{17}} = \sqrt{0.02245} = 0.1498\end{aligned}$$

The standard error of the least squares regression slope is

$$SE_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{0.1498}{\sqrt{1169.8116}} = 0.00438$$

Example

We will reject the null hypothesis if the P-value $\leq \alpha = 0.10$.

The test statistic is

$$t = \frac{b_1}{SE_{b_1}} = \frac{0.00949}{0.00438} = 2.17$$

Under the null hypothesis, this test statistic follows a t distribution with $n - 2 = 17$ degrees of freedom.

Example

The P-value is $2P(T(17) \geq 2.17)$. We see from Table 2 that

$$P(T(17) \geq 2.110) = 0.025 \quad \text{and} \quad P(T(17) \geq 2.224) = 0.02$$

Since $2.110 < t = 2.17 < 2.224$, it follows that $P(T(17) \geq 2.17)$ is between 0.02 and 0.025. But our P-value is double this probability, and so the P-value is between 0.04 and 0.05.

We can obtain the exact P-value from R as follows:

```
> 2*pt(2.17,17,lower.tail=FALSE)  
[1] 0.04446727
```

Example

Since the P-value $< \alpha = 0.10$, we reject H_0 . At the 10% level of significance, we have sufficient evidence that there exists a linear relationship between the average August temperature and ice cream consumption in a city.

If we had used the critical value method, the decision rule would be to reject the null hypothesis if $|t| \geq t^* = 1.740$, the upper 0.05 critical value of the t distribution with 17 d.f. The conclusion would be to reject the null hypothesis, since $t = 2.17 > t^* = 1.740$.

Example

We save the regression line as an object with the command `regline<-lm(IceCream~Temp)`, and then we use the command `summary(regline)` to get the value of the standard error, as well as several other values:

Call:

```
lm(formula = IceCream ~ Temp)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.22047 | -0.13623 | 0.04255 | 0.10423 | 0.21937 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | -0.273312 | 0.324621 | -0.842 | 0.4115 |
| Temp | 0.009487 | 0.004381 | 2.165 | 0.0449 * |

--- slope

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1498 on 17 degrees of freedom
 Multiple R-squared: 0.2162 Adjusted R-squared: 0.1701
 F-statistic: 4.689 on 1 and 17 DF, p-value: 0.04487

r^2

P-value

Example

It should be noted that tests for a linear relationship do not have to be two-sided. For example, in this situation, we **expect** that the two variables have a **positive** linear relationship. That is, as temperature increases, it seems logical that ice cream consumption should increase as well.

We can test for a **positive linear relationship** by making our alternative hypothesis one-sided:

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_a: \beta_1 > 0$$

The t statistic, would be the same (i.e. $t = 2.17$), but we would not double the tail area in finding the P-value, so the P-value would be half that of the two-sided test, i.e., between 0.02 and 0.025, and we would conclude that we have sufficient evidence of a positive linear relationship.

Example

Now that we have concluded that there is in fact a linear relationship between X and Y , we can use the least squares regression line for estimation and prediction purposes.

Suppose we want to find a 90% confidence interval for the value of the slope of the true regression line.

The standard error for the slope of the least squares regression line was calculated as:

$$SE_{b_1} = 0.00438$$

Example

Our 90% confidence interval for β_1 is

$$\begin{aligned} b_1 \pm t^* SE_{b_1} &= 0.00949 \pm 1.740(0.00438) \\ &= 0.00949 \pm 0.00762 = (0.00187, 0.01711) \end{aligned}$$

where $t^* = 1.740$ is the upper 0.05 critical value of the t distribution with $n - 2 = 19 - 2 = 17$ degrees of freedom.

Example

Now suppose we want to construct a 90% confidence interval for the true mean ice cream consumption for all American cities in which the average August temperature is 80°F.

Recall the form of the confidence interval:

$$\hat{y} \pm t^* SE_{\hat{\mu}}$$

where the standard error $SE_{\hat{\mu}}$ is

$$SE_{\hat{\mu}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Example

The estimated value of μ_Y when $X = 80$ is

$$\hat{y} = -0.273 + 0.00949(80) = 0.486$$

The sample mean temperature is calculated to be $\bar{x} = 73.679$, and so

$$(x^* - \bar{x})^2 = (80 - 73.679)^2 = 39.955$$

Example

We can now calculate

$$\begin{aligned} SE_{\hat{\mu}} &= s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\ &= 0.1498 \sqrt{\frac{1}{19} + \frac{39.955}{1169.8116}} = 0.0441 \end{aligned}$$

Example

Our 90% confidence interval for the true mean ice cream consumption per capita for all American cities with an average August temperature of 80°F is therefore

$$\begin{aligned}\hat{y} \pm t^* SE_{\hat{\mu}} &= 0.486 \pm 1.740(0.0441) \\ &= 0.486 \pm 0.077 = (0.409, 0.563)\end{aligned}$$

where $t^* = 1.740$ is the upper 0.05 critical value from the t distribution with $n - 2 = 19 - 2 = 17$ degrees of freedom.

Practice Question

Does a gas station sell more gas when the price is cheaper? The price of gas (in \$ per litre) and the amount of gas sold (in thousands of litres) at a gas station for a sample of nine days last year are recorded. Some summary statistics are shown below:

| | mean | std. dev. |
|----------------------|-------|-----------|
| Price Per Litre | 1.56 | 0.07 |
| 1000s of Litres Sold | 12.96 | 1.95 |

From the least squares regression line, it is calculated that 47.6% of the variation in the amount of gas sold can be accounted for by its regression on the price. We also calculate

$$\sum (x_i - \bar{x})^2 = 0.0416 \quad \text{and} \quad \sum (y_i - \hat{y}_i)^2 = 13.90$$

Practice Question

- (a) What is the value of the correlation between gas price and amount of gas sold for this sample of days?
- (b) We would like to use the price of gas to predict the amount of gas sold. Write out the simple linear regression model in the context of this example and define all terms.
- (c) Find the equation of the least squares regression line.
- (d) Provide an interpretation of the slope of the regression line in the context of this example.
- (e) What is the estimate of the parameter σ in the simple linear regression model?

Practice Question

- (f) Conduct a hypothesis test at the 5% level of significance to determine if there exists a linear relationship between the price of gas and the amount of gas sold. Use the P-value method.
- (g) Provide an interpretation of the P-value of the test in (f).
- (h) Calculate a 95% confidence interval for the parameter β_1 in the simple linear regression model.
- (i) Provide an interpretation of the interval in (h).

Practice Question

- (l) Could the interval in (h) be used to conduct the test in (f)? Why or why not? If the interval could have been used to conduct the test, what would the conclusion be, and why?
- (k) Suppose you had conducted the test in (f) using the critical value method. What would be the decision rule and the conclusion of the test?
- (l) Can we conclude that lowering the gas price causes people to buy more gas? Explain.

Practice Question

- (m) For one day in the sample, the gas price was \$1.62 and 11,250 litres of gas were sold. What is the value of the residual for this day?
- (n) Calculate a 90% confidence interval for the true mean gas sales for all days on which the gas price is \$1.70 per litre.