

Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

November 18, 2019

Topics

1 Assessing the Fit of a Regression Line Using R Squared

Objectives

Objectives:

- Compute and interpret the R-squared associated with a linear regression analysis.

Assessing the Fit of a Regression Line Using R Squared

(14.3)

Introduction

- The square of the correlation, r^2 , is called the **coefficient of determination**, but is more often referred to as **"R squared"**.

Assessing the Fit of a Regression Line Using R Squared

(14.3)

Introduction

- The square of the correlation, r^2 , is called the **coefficient of determination**, but is more often referred to as **"R squared"**.
- r^2 is used to **measure how well the regression line fits the data** in a scatterplot:

Assessing the Fit of a Regression Line Using R Squared

(14.3)

Introduction

- The square of the correlation, r^2 , is called the **coefficient of determination**, but is more often referred to as **"R squared"**.
- r^2 is used to **measure how well** the **regression line fits** the **data** in a scatterplot:
 - Values of r^2 **close** to **1.0** mean the regression line **fits** the data quite **well**.

Assessing the Fit of a Regression Line Using R Squared

(14.3)

Introduction

- The square of the correlation, r^2 , is called the **coefficient of determination**, but is more often referred to as **"R squared"**.
- r^2 is used to **measure how well** the **regression line fits** the **data** in a scatterplot:
 - Values of r^2 **close to 1.0** mean the regression line **fits** the data quite **well**.
 - Values **close to 0** mean it **doesn't fit** very well at all.

- More formally, r^2 represents ***the proportion of variation in y that can be explained by variation in x .***

- More formally, r^2 represents ***the proportion of variation in y that can be explained by variation in x .***

What does this mean? The value of y will **vary** from one individual to the next.

We can attribute this y **variation** to **two sources**:

- More formally, r^2 represents ***the proportion of variation in y that can be explained by variation in x .***

What does this mean? The value of y will **vary** from one individual to the next.

We can attribute this y **variation** to **two sources**:

1. **Variation** in y **due** to **variation** in x .

- More formally, r^2 represents ***the proportion of variation in y that can be explained by variation in x .***

What does this mean? The value of y will **vary** from one individual to the next.

We can attribute this y **variation** to **two sources**:

1. **Variation** in y **due** to **variation** in x .

This is what leads to an overall upward or downward **sloping pattern** in the scatterplot.

- More formally, r^2 represents ***the proportion of variation in y that can be explained by variation in x .***

What does this mean? The value of y will **vary** from one individual to the next.

We can attribute this y **variation** to **two sources**:

1. **Variation in y due to variation in x .**

This is what leads to an overall upward or downward **sloping pattern** in the scatterplot.

2. **Variation in y due to variation in all *other* factors (*besides x*).**

- More formally, r^2 represents ***the proportion of variation in y that can be explained by variation in x .***

What does this mean? The value of y will **vary** from one individual to the next.

We can attribute this y **variation** to **two sources**:

1. **Variation in y due to variation in x .**

This is what leads to an overall upward or downward **sloping pattern** in the scatterplot.

2. **Variation in y due to variation in all *other* factors (*besides x*).**

This is what leads to residual **errors** in the scatterplot.

- Now consider the two extremes:

- Now consider the two extremes:
 - If **all (100%)** of the variation in y was due to x , ...

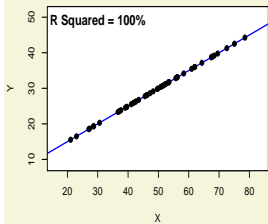
- Now consider the two extremes:
 - If **all (100%)** of the variation in y was due to x , ...
... the points in the scatterplot would lie **exactly** on a **straight line**, and r^2 would equal **1.0**.

- Now consider the two extremes:
 - If **all (100%)** of the variation in y was due to x , ...
... the points in the scatterplot would lie **exactly** on a **straight line**, and r^2 would equal **1.0**.
 - If **none (0%)** of the variation in y was due to x (i.e. *all* of it was due to other factors *besides* x), ...

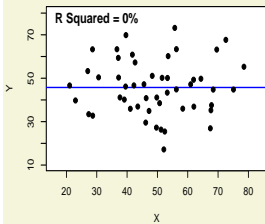
- Now consider the two extremes:
 - If **all (100%)** of the variation in y was due to x , ...
... the points in the scatterplot would lie **exactly** on a **straight line**, and r^2 would equal **1.0**.
 - If **none (0%)** of the variation in y was due to x (i.e. *all* of it was due to other factors *besides* x), ...
... the points in the scatterplot would form a **round "blob"**, and r^2 would equal **0.0**.

- Now consider the two extremes:
 - If **all (100%)** of the variation in y was due to x , ...
... the points in the scatterplot would lie **exactly** on a **straight line**, and r^2 would equal **1.0**.
 - If **none (0%)** of the variation in y was due to x (i.e. *all* of it was due to other factors *besides* x), ...
... the points in the scatterplot would form a **round "blob"**, and r^2 would equal **0.0**.
- Most data sets fall somewhere between these two extremes.

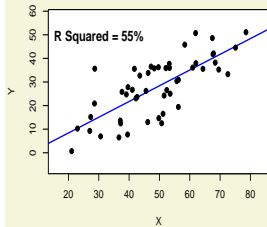
Scatterplot of Y vs X



Scatterplot of Y vs X



Scatterplot of Y vs X

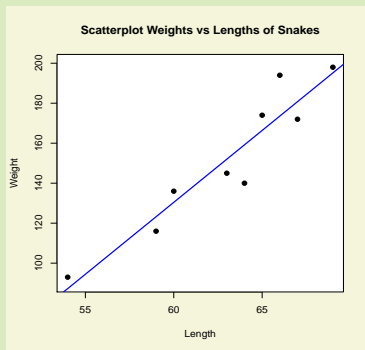


Exercise

The data and scatterplot (with fitted regression line) for the **lengths** (x) and **weights** (y) of snakes are below.

Lengths and Weights of Female Snakes

Snake	Length (cm)	Weight (g)
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



Statistical software gives the **equation** of the **regression line**:

$$\hat{y} = -300.97 + 7.19x$$

The **correlation** between **lengths** and **weights** is

$$r = 0.94.$$

- a) Note: the **weight** of a snake isn't solely determined by its **length** (otherwise the points in the scatterplot would lie exactly on the line).

Statistical software gives the **equation** of the **regression line**:

$$\hat{y} = -300.97 + 7.19x$$

The **correlation** between **lengths** and **weights** is

$$r = 0.94.$$

- a) Note: the **weight** of a snake isn't solely determined by its **length** (otherwise the points in the scatterplot would lie exactly on the line).

List a few of the other factors (*besides* length) that determine how much a snake weighs.

b) Calculate r^2 , the **coefficient of determination**.

- b) Calculate r^2 , the **coefficient of determination**.
- c) Based r^2 , what **percentage** of the **variation** in snakes' **weights** that's attributable to **variation** in their **lengths**?

- b) Calculate r^2 , the **coefficient of determination**.
- c) Based r^2 , what **percentage** of the **variation** in snakes' **weights** that's attributable to **variation** in their **lengths**?
- What **percentage** is due to **all other factors besides length**?

(Optional Section) An Explanation for Why r^2 Measures the Proportion of y Variation That's Due to

x (14.3)

- To see why r^2 is interpreted as the *proportion of variation in y that can be explained by x* :

(Optional Section) An Explanation for Why r^2 Measures the Proportion of y Variation That's Due to

x (14.3)

- To see why r^2 is interpreted as the *proportion of variation in y that can be explained by x* :
 - First consider the **total sum of squares of y** , **SST**, defined to be

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2,$$

the sum of squared deviations of the observed y_i 's away from their mean \bar{y} .

(Optional Section) An Explanation for Why r^2 Measures the Proportion of y Variation That's Due to

x (14.3)

- To see why r^2 is interpreted as the *proportion of variation in y that can be explained by x* :
 - First consider the **total sum of squares of y** , **SST**, defined to be

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2,$$

the sum of squared deviations of the observed y_i 's away from their mean \bar{y} .

SST is a measure of the **total variation** in the **observed y values**.

- (cont'd)
 - Second, consider the **error sum of squares, SSE**, defined to be

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

the "**sum of squared errors**" (deviations of the observed y_i 's away from their predicted values \hat{y}_i).

SSE is a measure of the **variation** in the **y values** that's due to **all other factors *besides* x** .

- (cont'd)

- The ratio **SSE/SST** has the following interpretation:

$$\frac{\text{SSE}}{\text{SST}} = \frac{\text{Variation in } y \text{ that's } \textit{not} \text{ due to } x}{\text{Total variation in } y}$$

= Proportion of the total variation in y that's *not* due to x .

- (cont'd)
 - It can be shown that

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

- (cont'd)
 - It can be shown that

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

The right side has the following interpretation:

$$1 - \frac{\text{SSE}}{\text{SST}} = 1 - \text{The proportion of the total variation in } y$$

that's *not* due to x

$$= \text{The proportion that } \textit{is} \text{ due to } x.$$