Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

November 10, 2019

Topics

Objectives

Objectives:

- Use the equation of a fitted regression line to predict y values from x values.
- Use the equation of a fitted regression line to describe the change in y associated with a given change in x.
- Compute the errors in predictions associated with a fitted regression line.
- State the principal of least squares for determining the equation of a regression line.

Objectives (Cont'd):

- Determine if a prediction is an extrapolation, and identify influential outliers in a regression analysis.

Linear Regression (14.1, 14.2)

Introduction to Linear Regression

- The goal of a ***linear regression analysis*** is to determine the line that "best fits" the points in a scatterplot.

Linear Regression (14.1, 14.2)

Introduction to Linear Regression

- The goal of a ***linear regression analysis*** is to determine the line that "best fits" the points in a scatterplot.
- Recall that the equation for a straight line is

$$y = b_0 + b_1 x$$

where b_0 is the **y-intercept** and b_1 is the **slope.**

- **Fitting a line to a scatterplot is useful for**:

  1. **Predicting** the value of y from a given value x (by plugging the x value into the equation of the line).

- **Fitting a line to a scatterplot is useful for**:

  1. **Predicting** the value of y from a given value x (by plugging the x value into the equation of the line).
  2. **Describing** a typical **change** in y associated with a given **change** in x (by looking at the **slope** of the line).

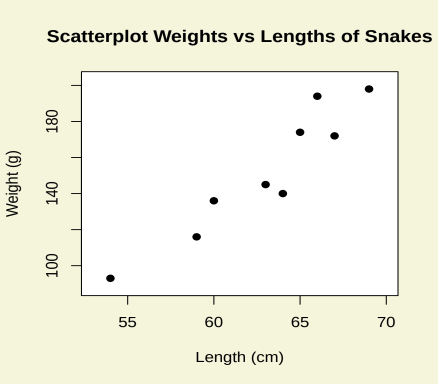
- **Fitting a line to a scatterplot is useful for**:

  1. **Predicting** the value of y from a given value x (by plugging the x value into the equation of the line).
  2. **Describing** a typical **change** in y associated with a given **change** in x (by looking at the **slope** of the line).
  3. Adding the line to the scatterplot to enhance its **appearance**.

Exercise

Here are the data on **lengths** and **weights** of snakes and the scatterplot, to which we add the **fitted regression line**.

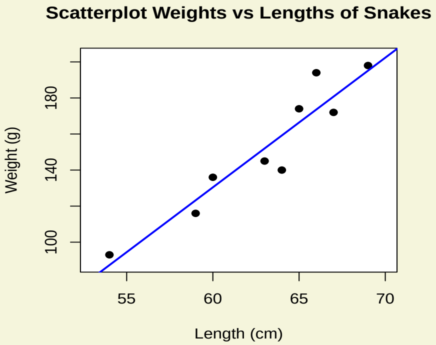
Snake	Length	Weight
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



Exercise

Here are the data on **lengths** and **weights** of snakes and the scatterplot, to which we add the **fitted regression line.**

Snake	Length	Weight
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



The **equation** of the **fitted regression line** is

$$\hat{y} \ = \ -301.09 + 7.19x$$

(where y is **weight** and x is **length**).

The **equation** of the **fitted regression line** is

$$\hat{y} \;=\; -301.09 + 7.19x$$

(where y is **weight** and x is **length**).

The **"hat"** over the y above indicates that its a **fitted regression line** (not just any line).

The **equation** of the **fitted regression line** is

$$\hat{y} \;=\; -301.09 + 7.19x$$

(where y is **weight** and x is **length**).

The **"hat"** over the y above indicates that its a **fitted regression line** (not just any line).

a) What **weight** would we **predict** for a snake whose **length** is **62** cm? What **weight** would we **predict** for a **66** cm long snake?

The **equation** of the **fitted regression line** is

$$\hat{y} \; = \; -301.09 + 7.19x$$

(where y is **weight** and x is **length**).

The **"hat"** over the y above indicates that its a **fitted regression line** (not just any line).

a) What **weight** would we **predict** for a snake whose **length** is **62** cm? What **weight** would we **predict** for a **66** cm long snake?

b) What's a typical **change** in **weight** for each **1** cm **elongation**? What would we expect the **change** in **weight** to be for a **5** cm **elongation**?

Prediction Errors

- In the last exercise a **66** cm long snake was **predicted** to weigh **173.6** g.

Prediction Errors

- In the last exercise a **66** cm long snake was **predicted** to weigh **173.6** g.

  But in the data set, an **observed 66** cm long snake weighs **194** g (snake number 3).

Prediction Errors

- In the last exercise a **66** cm long snake was **predicted** to weigh **173.6** g.

  But in the data set, an **observed 66** cm long snake weighs **194** g (snake number 3).

  In other words, there was a slight **error** in the prediction (because the observed snake's weight didn't lie on the line).

- In a regression analysis, an ***error*** (also called a ***residual***) is the vertical deviation of a point away from the fitted line in a scatterplot:

Error (or Residual):

$$\begin{aligned} \text{Error} &= \text{Observed } y - \text{Predicted } y \\ &= y - \hat{y} \end{aligned}$$

Example

The **error** when we tried to predict the weight of the **66** cm long snake was:

$$\begin{aligned} \text{Error} &= 194 - 173.6 \\ &= \mathbf{20.4} \end{aligned}$$

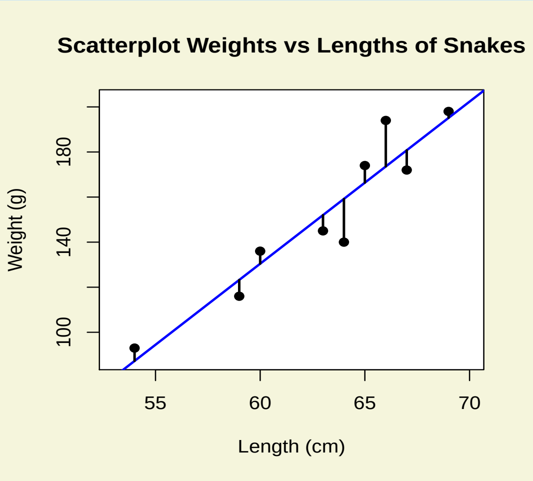
Example

The **error** when we tried to predict the weight of the **66** cm long snake was:

$$\begin{aligned} \text{Error} &= 194 - 173.6 \\ &= \mathbf{20.4} \end{aligned}$$

It's **positive** because the snake's actual weight lies **above** the regression line.

The scatterplot (next slide) uses vertical lines to represent **all** of the **errors**.


Scatterplot Weights vs Lengths of Snakes
Weight (g)
100
140
180
55
60
65
70
Length (cm)

- In general, an **error** will be **positive** or **negative** depending on whether the observed y lies **above** or **below** the line.

Exercise

Refer to the regression analysis of snakes' **weights** and **lengths**.

a) Use the regression line to **predict** the weight of a **64** cm long snake.

Exercise

Refer to the regression analysis of snakes' **weights** and **lengths**.

a) Use the regression line to **predict** the weight of a **64** cm long snake.

b) One of the snakes in the data set actually was **64** cm long (snake number 4). Calculate the **error** in the prediction from Part *a*.

(**Optional**)

Fitting a Regression Line to Data: The Principle of Least Squares

- When fitting a line to a scatterplot, we want the **errors** to be **small.**

(**Optional**)

Fitting a Regression Line to Data: The Principle of Least Squares

- When fitting a line to a scatterplot, we want the **errors** to be **small.**
- They'll be small whenever their **squares** are small, and it turns out to be easier to find a line that makes the **squared errors** small.

(**Optional**)

- The ***principle of least squares*** says that the line that ”best fits” the data is the one that makes the ***sum of squared errors***

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

as small as possible.

(**Optional**)

- The ***principle of least squares*** says that the line that "best fits" the data is the one that makes the ***sum of squared errors***

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

  as small as possible.

  The line that results in the *smallest possible* sum of squared errors is called the ***least squares regression line*** (or just the ***fitted regression line***).

(**Optional**)

- The ***principle of least squares*** says that the line that ”best fits” the data is the one that makes the ***sum of squared errors***

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

  as small as possible.

  The line that results in the *smallest possible* sum of squared errors is called the ***least squares regression line*** (or just the ***fitted regression line***).

  (This is the line shown in the scatterplots of the snakes data in previous examples.)

(**Optional**)

Computing the Slope and Intercept of the Regression Line

- It can be shown that the **slope** of the fitted regression line, which is denoted $\boldsymbol{b_1}$, is computed from the data by the formula:

Slope of the Regression Line:

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

where $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \sum(x_i - \bar{x})^2$.

(**Optional**)

- Once the slope has been computed, the $\boldsymbol{y}$**-intercept**, denote $\boldsymbol{b_0}$, is computed using the formula:

$\boldsymbol{Y}$**-Intercept of the Regression Line**:

$$b_0 = \bar{y} - b_1\bar{x}$$

(**Optional**)

- The resulting fitted **least squares regression line** is

Regression Line Equation:

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

where the values of b_1 and b_0 are obtained using the formulas on the previous two slides.

(**Optional**)

Properties of the Regression Line

- The regression line has the following **properties**:

    1. The line **always** passes through the **”point of averages”**, $(\bar{x}, \bar{y})$.

(**Optional**)

Properties of the Regression Line

- The regression line has the following **properties**:

    1. The line **always** passes through the **"point of averages"**, $(\bar{x}, \bar{y})$.

       Thus an individual who's **average** in the x **variable** is predicted to be **average** in the y **variable** too.

(**Optional**)

- (cont'd)

  2. An alternative (but equivalent) formula for computing the slope b_1 of the regression line is

> **Slope of the Regression Line (Alternative Formula)**:
>
> $$b_1 = r \times \frac{s_y}{s_x}$$
>
> where $\boldsymbol{r}$ is the correlation and $\boldsymbol{s_x}$ and $\boldsymbol{s_y}$ are the $\boldsymbol{x}$ and $\boldsymbol{y}$ standard deviations.

(**Optional**)

- (cont'd)

  3. From the alternative formula for the slope b_1 (previous slide), since s_x and s_y are always positive the **slope** will **always** have the **same sign** as the **correlation** r.

(**Optional**)

- **Cautions** about the **regression line**:

  1. Beware of ***extrapolation*** (using the regression line to make predictions far outside the range of the x values in the original data set).

     Extrapolation can lead to **faulty predictions.**

(**Optional**)

- **Cautions** about the **regression line**:

  1. Beware of ***extrapolation*** (using the regression line to make predictions far outside the range of the x values in the original data set).

     Extrapolation can lead to **faulty predictions**.

  2. Beware of outliers that are ***influential*** (i.e. that have a strong influence on the slope of the regression line).

     Outliers in the **horizontal** (x) direction can be particularly **influential**.

(**Optional**)

- **Cautions** about the **regression line**:

  1. Beware of ***extrapolation*** (using the regression line to make predictions far outside the range of the x values in the original data set).

     Extrapolation can lead to **faulty predictions.**

  2. Beware of outliers that are ***influential*** (i.e. that have a strong influence on the slope of the regression line).

     Outliers in the **horizontal** (x) direction can be particularly **influential.**

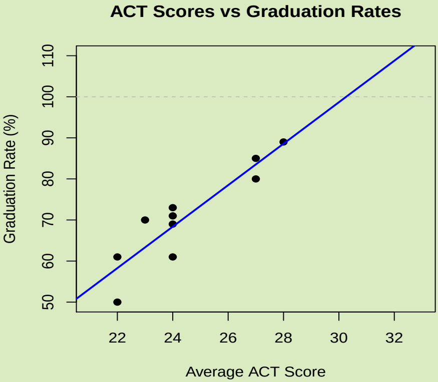
- The next example illustrates the danger of **extrapolation.**

(**Optional**)

Exercise

ACT exam scores are often used to predict **graduation rates** at universities. The average **ACT score** and **percentage** of freshmen who **graduate** are presented below for ten large universities.

University	ACT Average	Graduation Rate (%)
Illinois	27	80
Indiana	24	69
Iowa	24	61
Michigan	27	85
Michigan State	23	70
Minnesota	22	50
Northwestern	28	89
Ohio State	22	61
Purdue	24	71
Wisconsin	24	73



(**Optional**)

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

(**Optional**)

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

Another university's average **ACT score** is **32.**

(**Optional**)

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

Another university's average **ACT score** is **32.**

a) Would a **prediction** of its **graduation rate** based on the regression line be an **extrapolation**?

(**Optional**)

The equation of the **fitted regression line** is

$$\hat{y} = -52.8 + 5.05x.$$

(where y is **graduation rate** and x is average **ACT score**).

Another university's average **ACT score** is **32.**

a) Would a **prediction** of its **graduation rate** based on the regression line be an **extrapolation**?

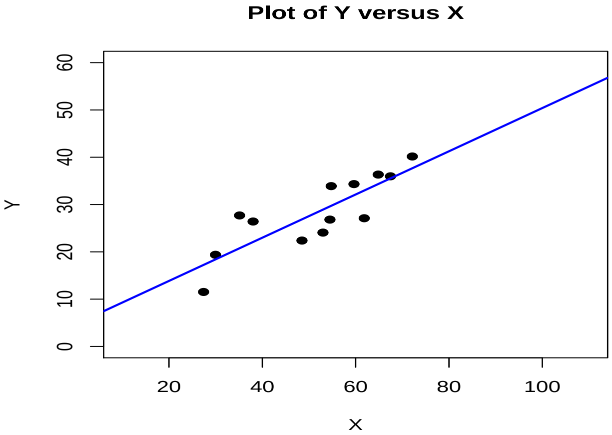
b) Would the **prediction** be trustworthy?

(**Optional**)

- This next several slides show that an **outlier** can be **influential** (on the regression line), but not all outliers are.

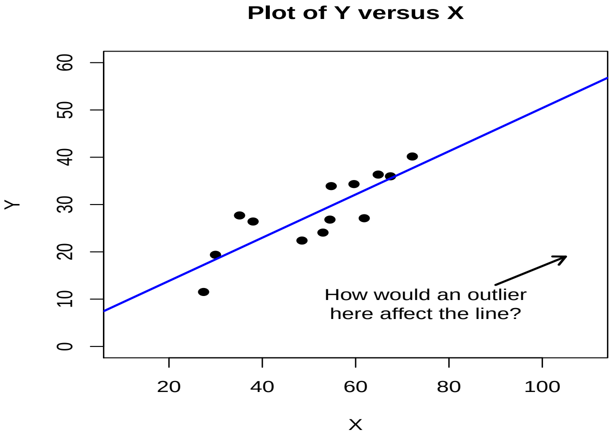
(**Optional**)

- Some outliers are **influential**.



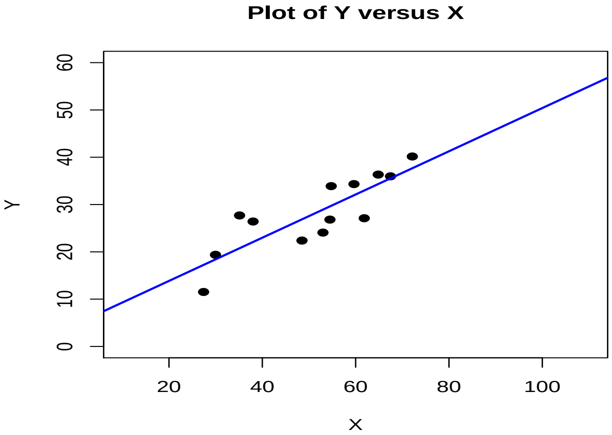
(**Optional**)

- Some outliers are **influential**.



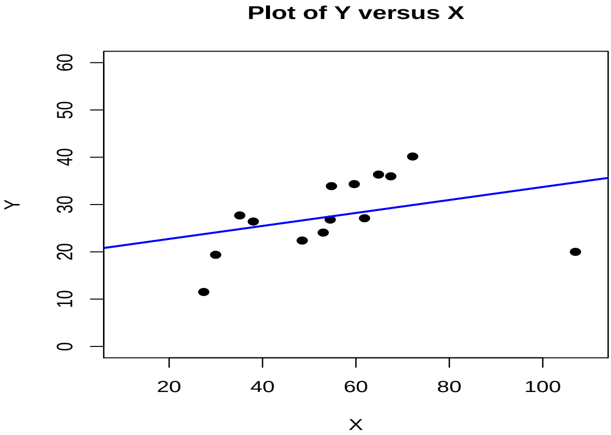
(**Optional**)

- Some outliers are **influential.**



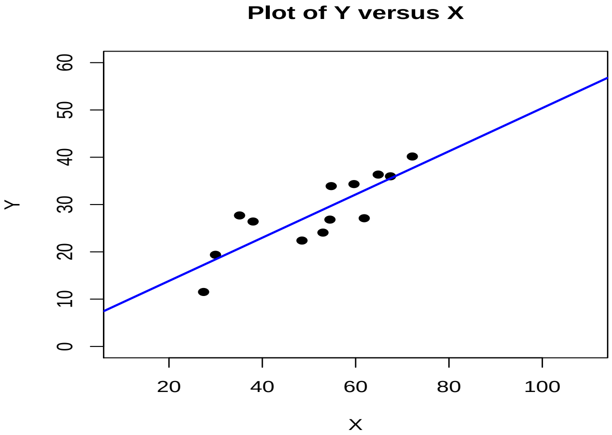
(**Optional**)

- Some outliers are **influential.**



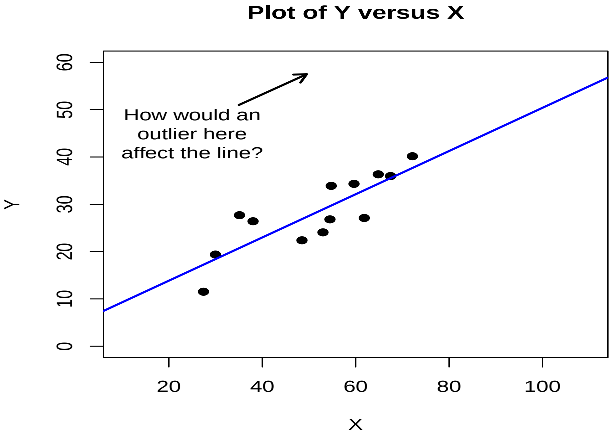
(**Optional**)

- Other outliers are **not** influential.



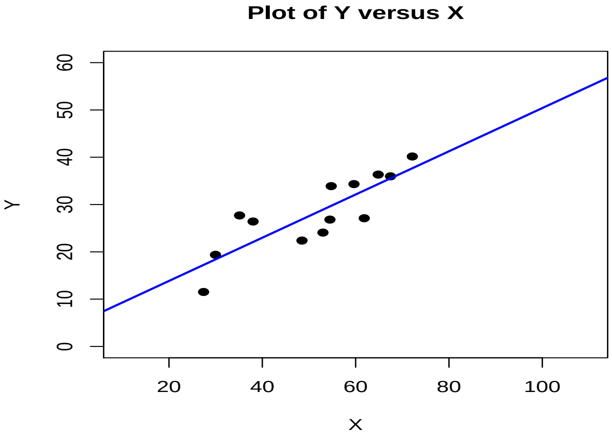
(**Optional**)

- Other outliers are **not** influential.



(**Optional**)

- Other outliers are **not** influential.



(**Optional**)

- Other outliers are **not** influential.