Week 3: Basic regression

2. Introduction to linear model

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Main ideas
    1. Line of best fit
    2. Finding the least square line in R
    3. Prediction
    4. Conditions for linear regression

2. Summary

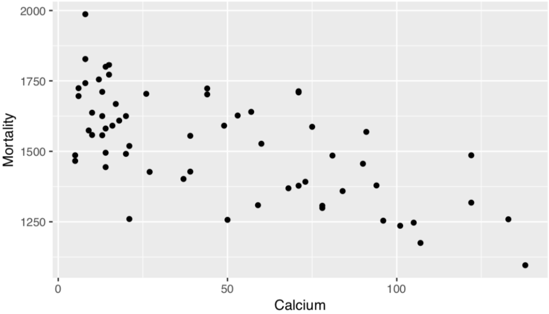
Outline

1. Main ideas
    1. Line of best fit
    2. Finding the least square line in R
    3. Prediction
    4. Conditions for linear regression
2. Summary

Mortality and potion

Scientists believe that water with high concentrations of calcium and magnesium is beneficial for health.

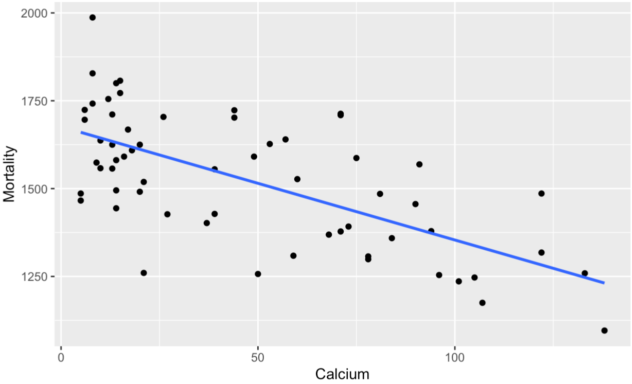
We have recordings of the mortality rate (deaths per 100,000 population) and concentration of calcium in drinking water (parts per million) in 61 large towns in England and Wales.



Review of terms

- **Response variable**: variable whose behavior or variation you are trying to understand, on the y-axis (dependent variable)
- **Explanatory variables**: other variables that you want to use to explain the variation in the response, on the x-axis (independent variables); these are also referred to as predictors or features.

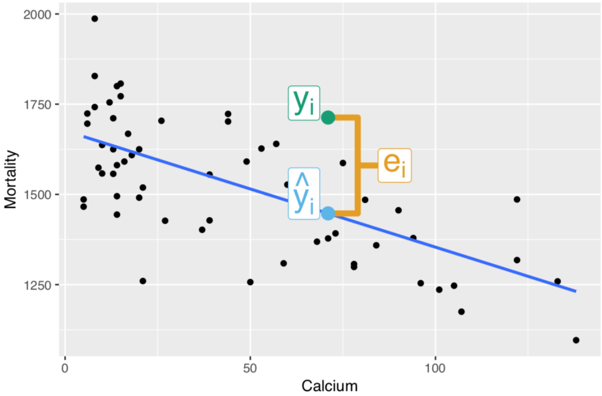
Add a line



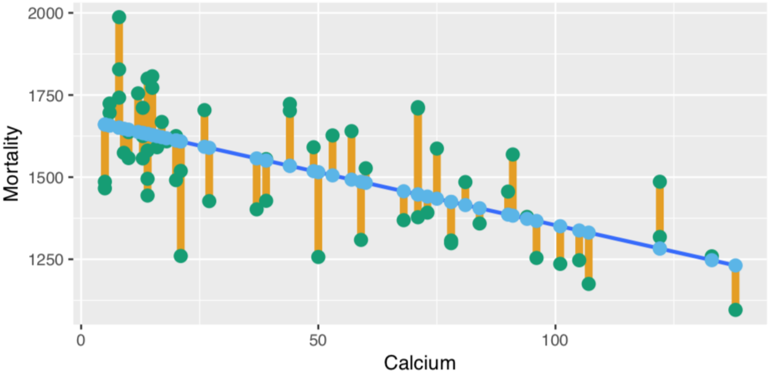
What is the best line of fit?
What qualities are important here?

Residual = Observed - Predicted

1. **Predicted value** $\hat{y}$: estimate made from a model
2. **Observed value** y: value in the dataset



The line of best fit is the line for which the sum of the squared residuals is smallest, the **least squares line**.



Outline

1. Main ideas
    1. Line of best fit
    2. Finding the least square line in R
    3. Prediction
    4. Conditions for linear regression

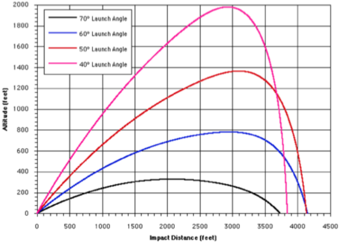
2. Summary

Lines

The algebraic equation for a line is

$$Y = b_0 + b_1 X$$

The use of coordinate axes to show functional relationships was invented by Rene Descartes (1596-1650). He was an artillery officer, and probably got the idea from pictures that showed the trajectories of cannonballs.



A new function

General form of ‘lm‘ command:

```
lm(y_variable ~ x_variable, data = data_frame)
```

Use this to estimate the intercept and the slope of line in the Mortality/Health data.

```
linear_fit ← lm(Mortality ~ Calcium, data = mortality_water)
linear_fit
```

```
Coefficients:
(Intercept)      Calcium
   1676.356       -3.226
```

From the coefficients, we can write the regression line in the Mortality/Health data as

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

Abstractly,

$$\hat{y} = 1676 - 3x$$

Outline

1. Main ideas
    1. Line of best fit
    2. Finding the least square line in R
    3. Prediction
    4. Conditions for linear regression
2. Summary

One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

Predict

One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

By hand:

$$\widehat{\text{Mortality}} = 1676 - 3 \times 71 = 1463$$

The predicted value for the mortality rate in that town is 1463 deaths per 100,000 population.

Predict

By R:

The general form of ‘predict‘ command:

predict(linear_model, newdata = data_frame)

Code:

```
linear_fit <- lm(Mortality ~ Calcium, data = mortality_water)
predict_data <- data.frame( Calcium = 71)
predict(linear_fit, newdata = predict_data)
```

Output:

```
       1
1447.303
```

Predict

By R:

The general form of ‘predict‘ command:

predict(linear_model, newdata = data_frame)

Code:

```
linear_fit <- lm(Mortality ~ Calcium, data = mortality_water)
predict_data <- data.frame( Calcium = 71)
predict(linear_fit, newdata = predict_data)
```

Output:

```
       1
1447.303
```

The outputs by hand and by R are different because of rounding errors.

Outline

1. Main ideas
    1. Line of best fit
    2. Finding the least square line in R
    3. Prediction
    4. Conditions for linear regression

2. Summary

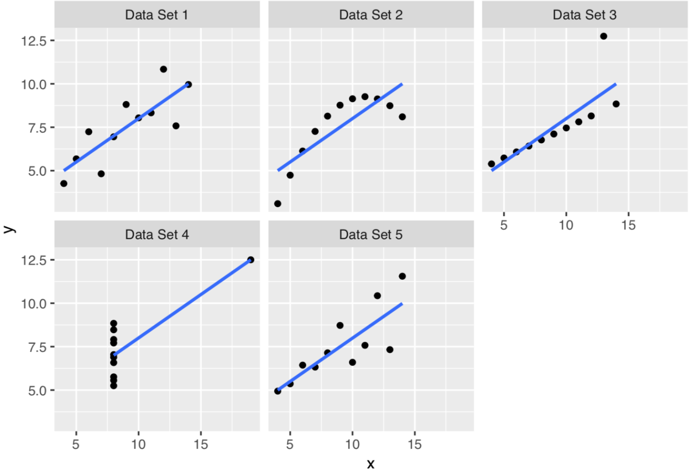
Anscombe's data

```
x1, y1     x2, y2     x3, y3     x4, y4     x5, y5
b0 = 3     b0 = 3     b0 = 3     b0 = 3     b0 = 3
b1 = 0.5   b1 = 0.5   b1 = 0.5   b1 = 0.5   b1 = 0.5
R2 = 67%   R2 = 67%   R2 = 67%   R2 = 67%   R2 = 67%
```

All 5 have essentially the same estimated intercept, slope, R^2!

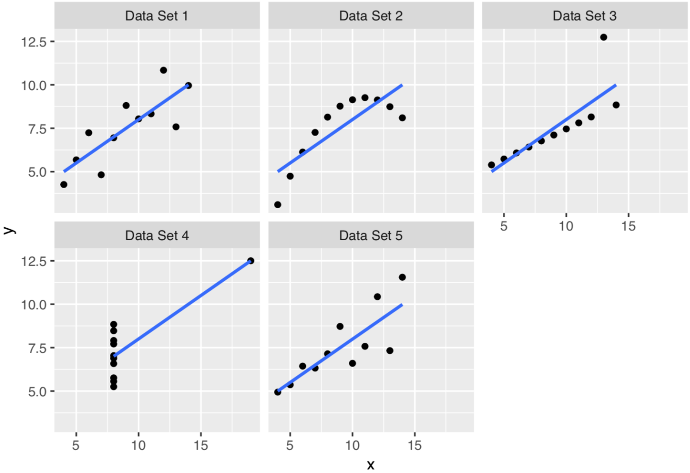
That means the five data sets should be pretty much the same right?

The scatterplots tell a different story.



Words of caution: always plot your data!

Is a linear model useful here?



Checking conditions

Be sure to check the conditions for linear regression before reporting or interpreting a linear model.

- From the scatterplot of y against x, check the
  - **Straight Enough Condition** Is the relationship between y and x straight enough to proceed with a linear regression model?
  - **Outlier Condition** Are there any outliers that might dramatically influence the fit of the least squares line?
  - **Does the Plot Thicken? Condition** Does the spread of the data around the generally straight relationship seem to be consistent for all values of x?

Outline

1. Main ideas
    1. Line of best fit
    2. Finding the least square line in R
    3. Prediction
    4. Conditions for linear regression

2. Summary

Summary of main ideas

1. Line of best fit
2. Finding the least square line in R
3. Prediction
4. Conditions for linear regression