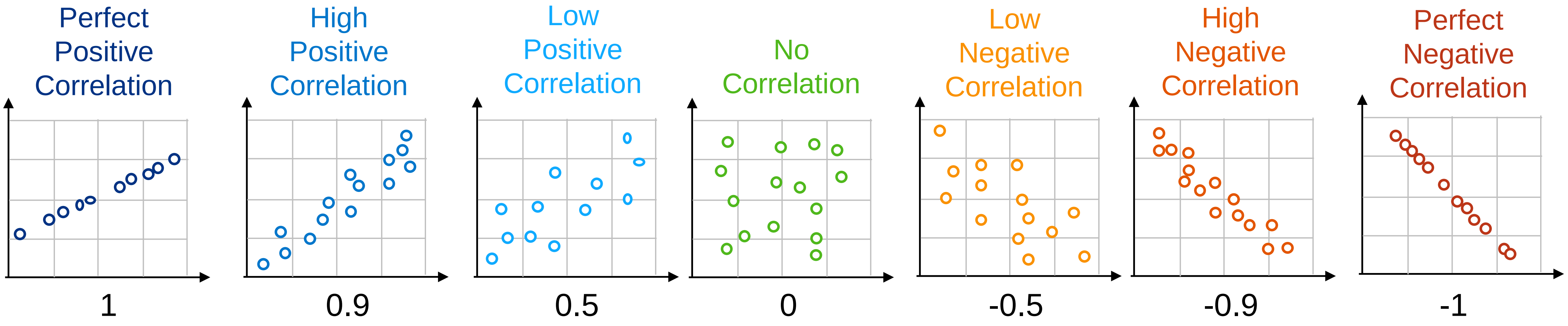
Week 3: Correlation and Linear Regression

# Unit 1: Correlation as a Statistical Measure

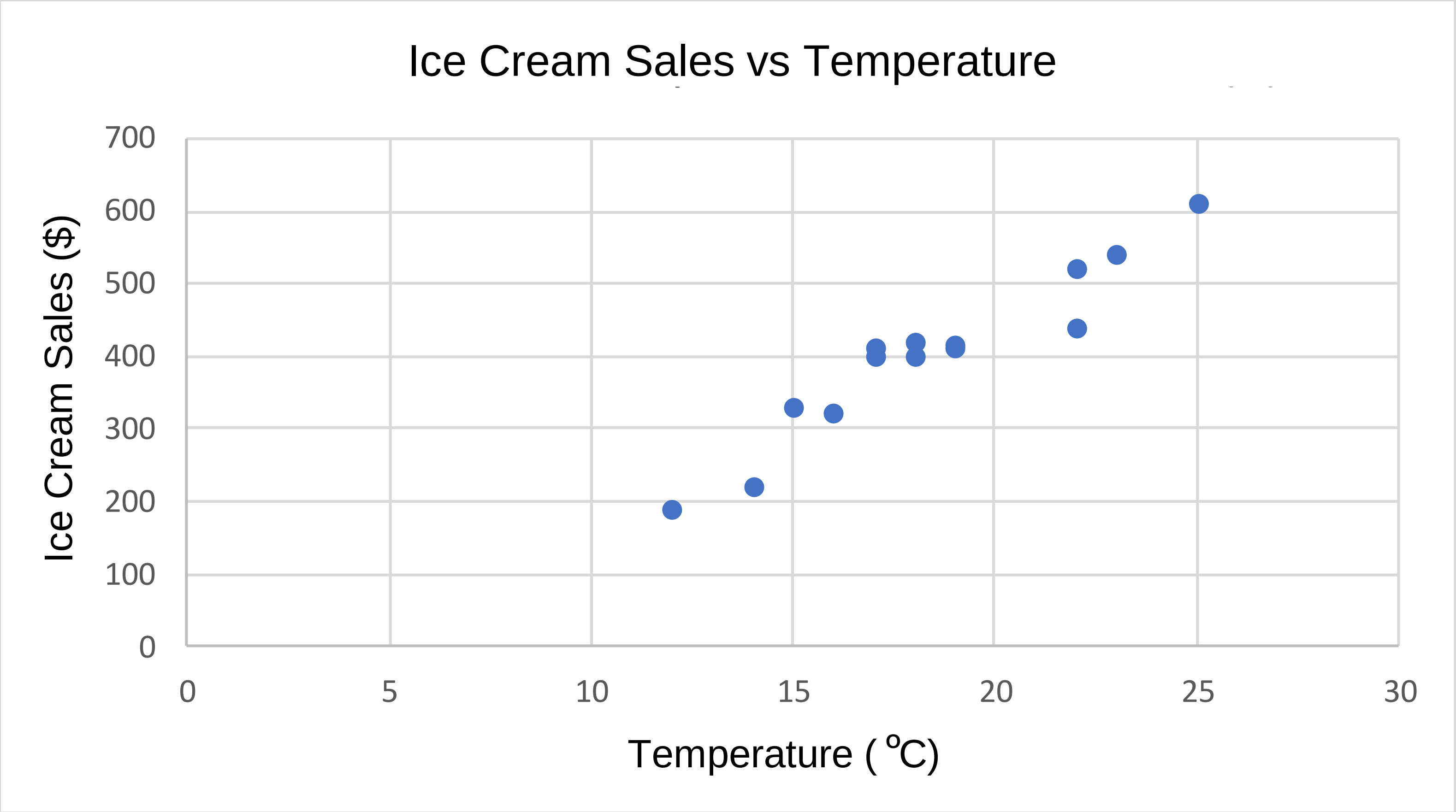
# Introduction

- When two sets of data are strongly linked together, we say they have a "**high correlation**".
- Correlation is **positive** when the values **increase** together.
- Correlation is **negative** when one value **decreases** as the other increases.



# Example

| Ice Cream Sales vs Temperature | |
|---|---|
| Temperature ( $^{o}$C) | Ice Cream Sales ($) |
| 25 | 610 |
| 18 | 400 |
| 16 | 320 |
| 22 | 440 |
| 22 | 520 |
| 19 | 410 |
| 18 | 420 |
| 17 | 410 |
| 23 | 540 |
| 14 | 220 |
| 12 | 190 |
| 15 | 330 |
| 17 | 400 |
| 19 | 415 |



# Pearson’s correlation coefficient

- The **Pearson product-moment correlation coefficient** is a measure of the strength and direction of the linear relationship between two variables.

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

Where:

*x* and *y* are the two variables

Σ is Sigma, the symbol for "sum up"

$(x_i - \bar{x})$ is each x-value minus the mean of x

$(y_i - \bar{y})$ is each y-value minus the mean of y

# Calculation of correlation

**Step 2** – Subtract mean

**Step 3** – Calculate $(x - \bar{x})$ x $(y - \bar{y})$, $(x - \bar{x})^2$ and $(y - \bar{y})^2$ for each value

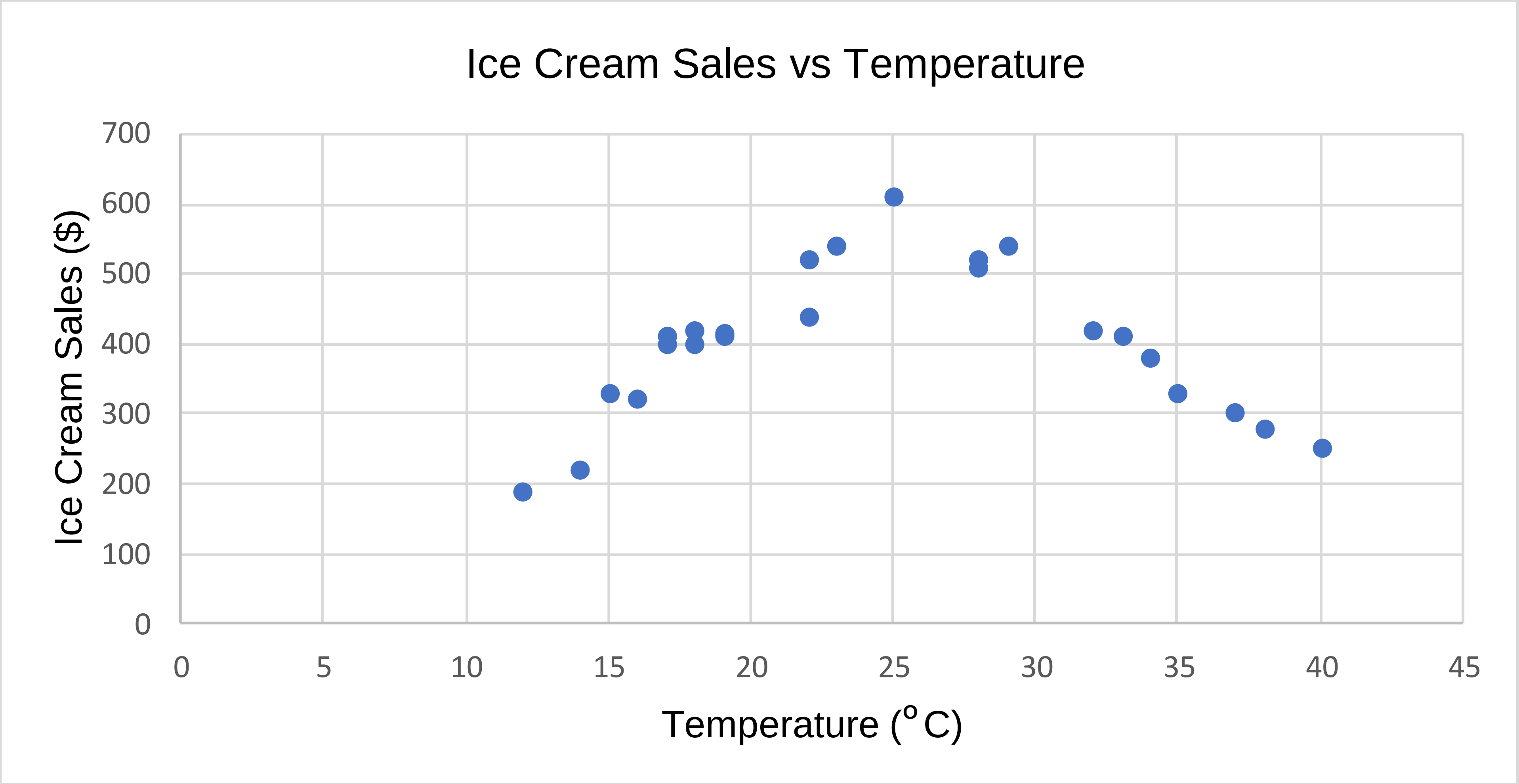
| Sales Ice Cream vs Temperature | | | | | | |
|---|---|---|---|---|---|---|
| Temperature (°C) | Ice Cream Sales ($) | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})$ x $(y - \bar{y})$ | $(x - \bar{x})^2$ | $(y - \bar{y})^2$ |
| 25 | 610 | 6.64 | 208.21 | 1383.14 | 44.13 | 43353.19 |
| 18 | 400 | -0.36 | -1.79 | 0.64 | 0.13 | 3.19 |
| 16 | 320 | -2.36 | -81.79 | 192.78 | 5.56 | 6688.90 |
| 22 | 440 | 3.64 | 38.21 | 139.21 | 13.27 | 1460.33 |
| 22 | 520 | 3.64 | 118.21 | 430.64 | 13.27 | 13974.62 |
| 19 | 410 | 0.64 | 8.21 | 5.28 | 0.41 | 67.47 |
| 18 | 420 | -0.36 | 18.21 | -6.51 | 0.13 | 331.76 |
| 17 | 410 | -1.36 | 8.21 | -11.15 | 1.84 | 67.47 |
| 23 | 540 | 4.64 | 138.21 | 641.71 | 21.56 | 19103.19 |
| 14 | 220 | -4.36 | -181.79 | 792.07 | 18.98 | 33046.05 |
| 12 | 190 | -6.36 | -211.79 | 1346.35 | 40.41 | 44853.19 |
| 15 | 330 | -3.36 | -71.79 | 240.99 | 11.27 | 5153.19 |
| 17 | 400 | -1.36 | -1.79 | 2.42 | 1.84 | 3.19 |
| 19 | 415 | 0.64 | 13.21 | 8.49 | 0.41 | 174.62 |
| 18.36 | 401.79 | | | 5166.07 | 173.21 | 168280.36 |

**Step 4** - Sum up

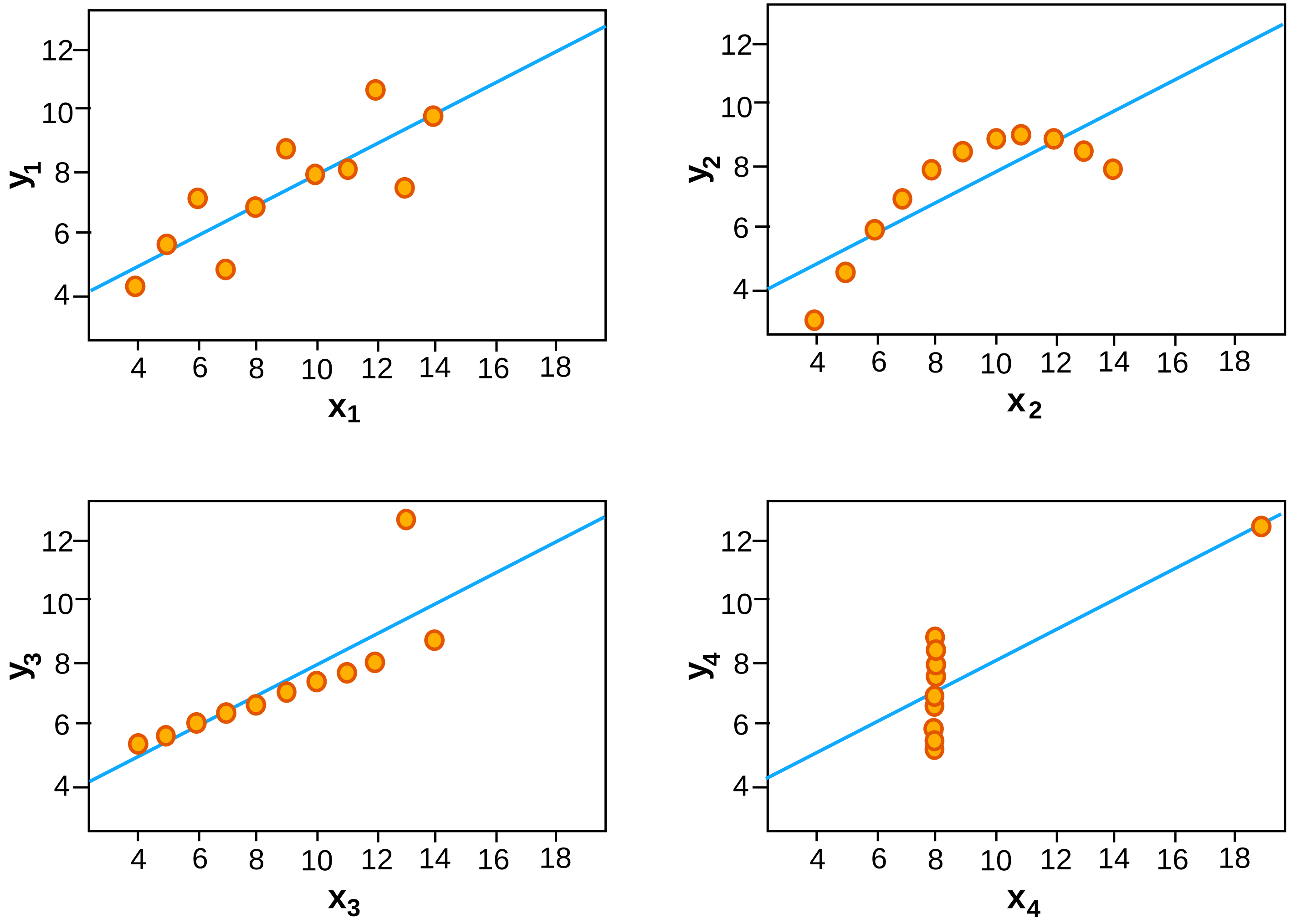
**Step 1** – mean values

Pearson's Correlation Coefficient $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$ $= \frac{5166.07}{\sqrt{173.21 \times 168280.36}} = 0.96$

# Limitations of correlation – Linear relationships only



# Limitations of correlation – Data visualization should not be ignored




For all 4 data sets:

| Property | Value |
|---|---|
| Mean of *x* | 9 |
| Sample variance of *x* | 11 |
| Mean of *y* | 7.50 |
| Sample variance of *y* | 4.125 |
| Correlation between *x* and *y* | 0.816 |
| Linear regression line | $y = 3.00 + 0.500x$ |
| Coefficient of determination of the linear regression | 0.67 |

# Summary

- Correlation measures the strength of association between two variables.
- The most common correlation coefficient is called the Pearson product-moment correlation coefficient.
- The sign and the absolute value of a Pearson correlation coefficient describe the direction and the magnitude of the relationship between two variables.
  – The value of a correlation coefficient ranges between -1 and +1.
  – The greater the absolute value of a correlation coefficient, the stronger the linear relationship.
  – The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
  – The weakest linear relationship is indicated by a correlation coefficient of 0.
  – A positive correlation indicates that as one variable increases in value, the other variable tends to increase as well.
  – A negative correlation indicates that as one variable increases in value, the other variable tends to reduce in value.
- Remember to use scatter plots to visualize the relationship as well as calculating the correlation coefficient.

# Thank you.

**Contact information:**

**open@sap.com**

Follow all of SAP

www.sap.com/contactsap

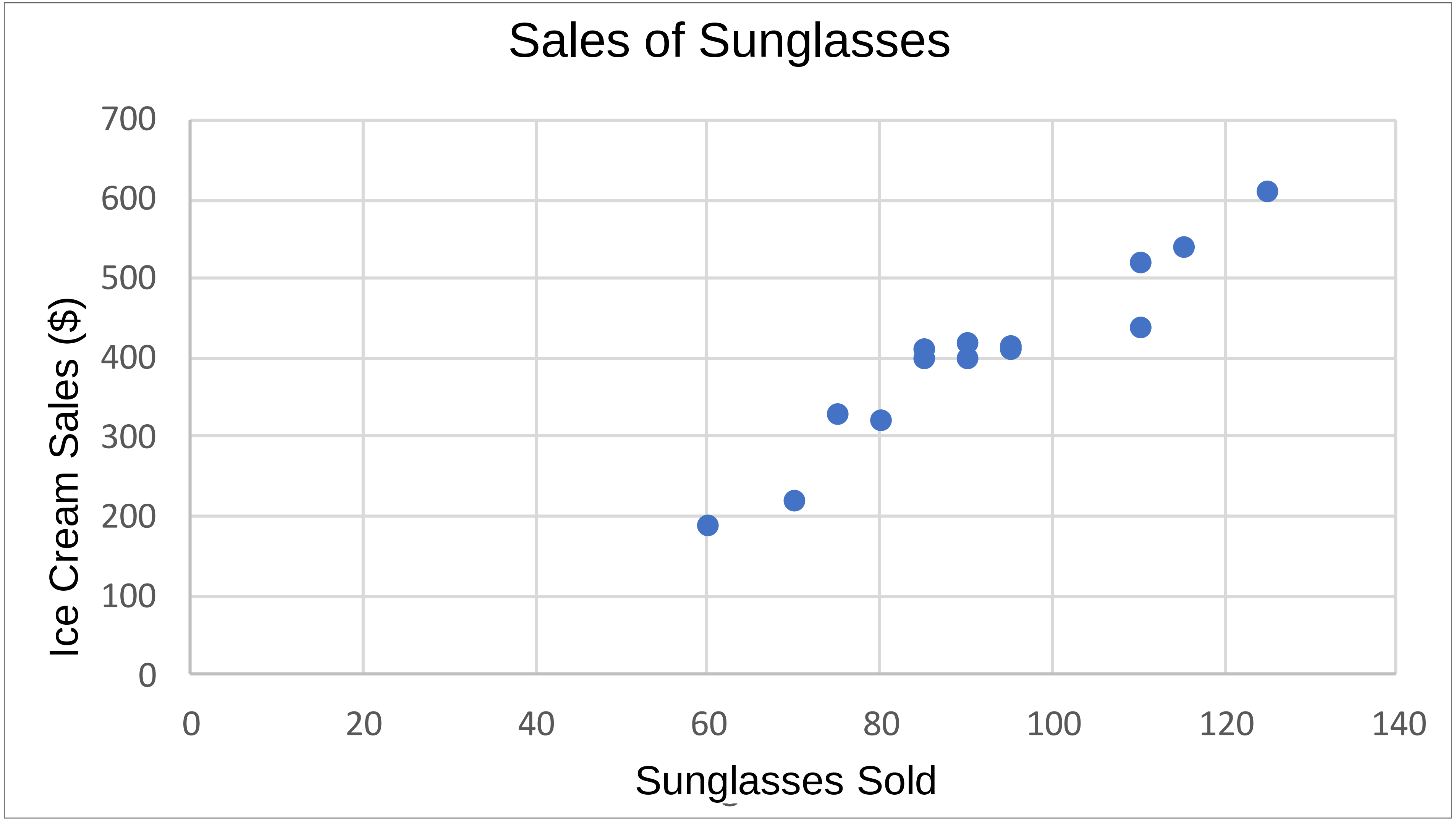


THE BEST RUN SAP

Week 3: Correlation and Linear Regression

# Unit 2: Correlation Versus Causation

# Introduction

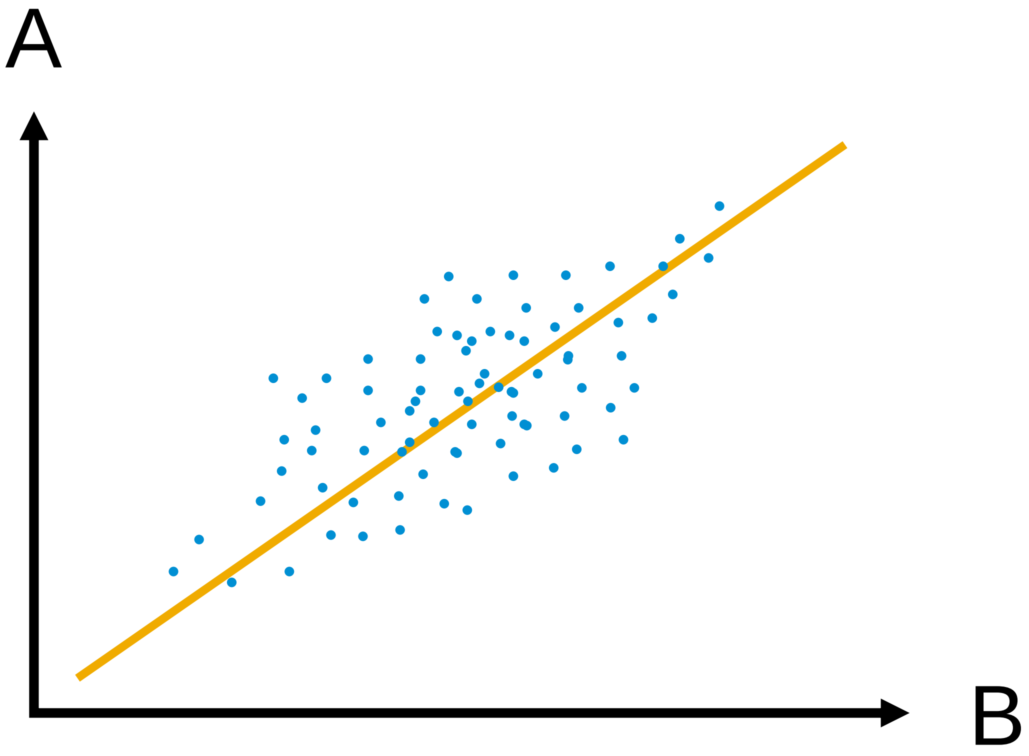


# HRT example

Numerous epidemiological studies showed that women taking combined hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD), leading doctors to propose that HRT was protective against CHD.

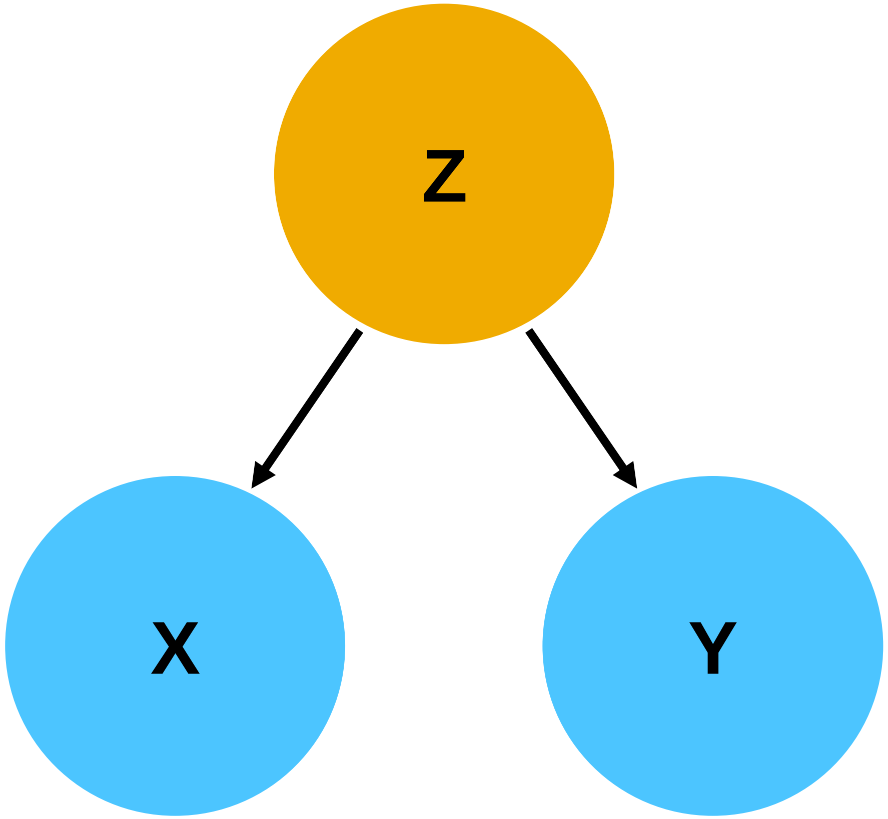
Lawlor DA, Davey Smith G, Ebrahim S (June 2004). "Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology?". Int J Epidemiol. **33** (3): 464-467.

# Cause and effect




See: https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

# Spurious relationships




- https://en.wikipedia.org/wiki/Confounding
- http://www.virmanimath.com/start-page-2012-2013/ap-stats-2012-2013/chapter-2/apstatonlineclass/confounding-and-lurking-variables
- https://docs.google.com/presentation/d/1OU4VBPWVUi0M6Mc-vdgjbkPaVsQKlHV-ONQc_T9Nlb4/edit#slide=id.p20

# Detecting spurious relationships

In experimental research, spurious relationships can often be identified by “controlling” for other factors, including those that have been theoretically identified as possible confounding factors.

# Summary

- Correlation is a statistical measure (expressed as a number) that describes the size and direction of a linear relationship between two variables.
- Causation indicates that one event is the result of the occurrence of the other event, i.e. there is a causal relationship between the two events. This is also referred to as “cause and effect.”
- For A to cause B, we tend to say that, at a minimum, A must precede B, the two must covary (vary together), and no competing explanation can better explain the covariance of A and B. Taken alone, however, these three requirements cannot prove cause – they are necessary but not sufficient.
- **Lurking** and **confounding** variables can make it difficult to conclude that it was the explanatory variables alone that affected the observed changes in the response variable.

# Thank you.

**Contact information:**

**open@sap.com**

Follow all of SAP

www.sap.com/contactsap

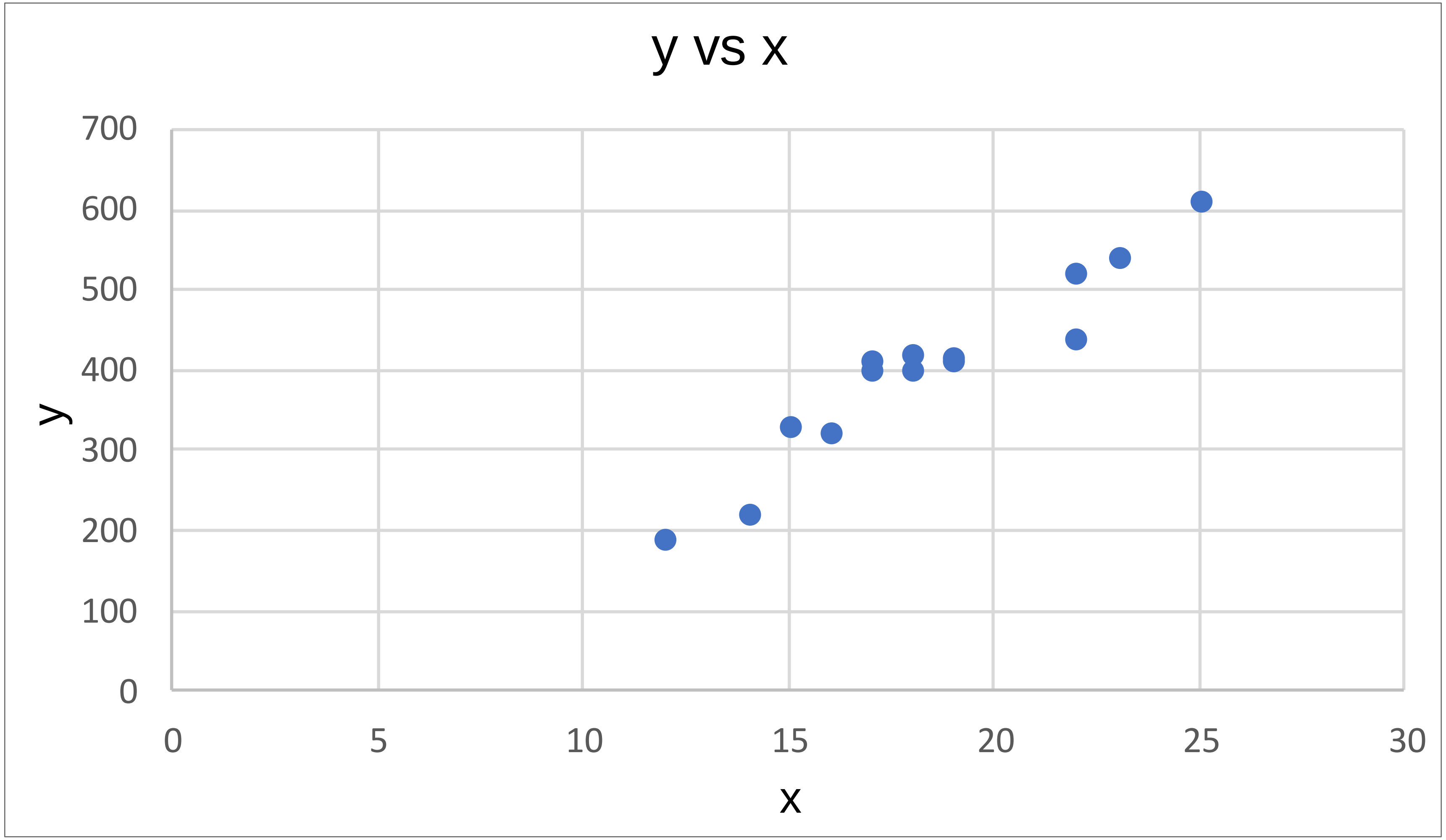


THE BEST RUN

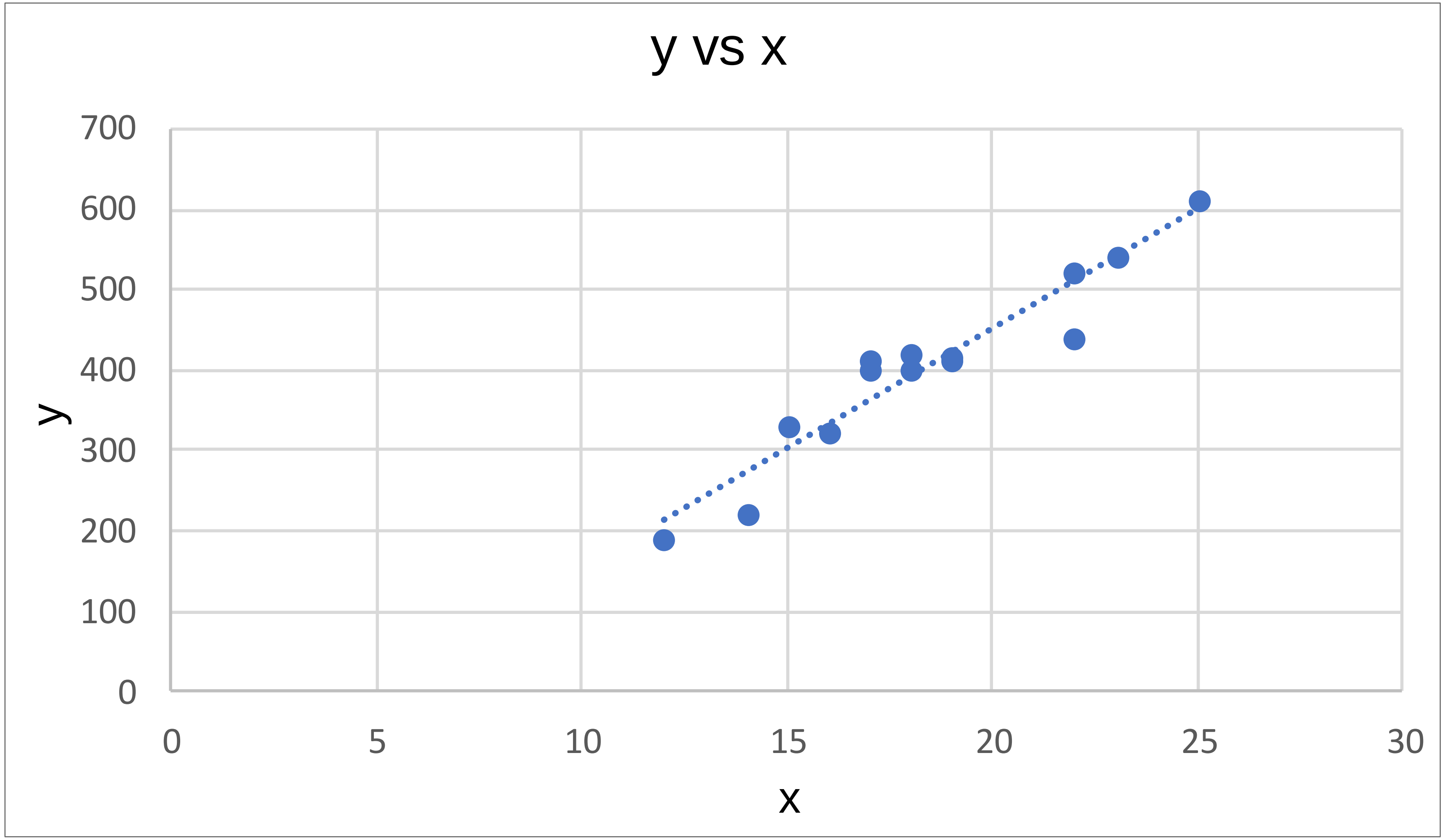
Week 3: Correlation and Linear Regression

# Unit 3: Scatter Plots and Line of Best Fit

# Scatter plots



# Line of best fit



# Line of best fit

- The equation of a **line**:

$$y = mx + b$$

- The **slope**:

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

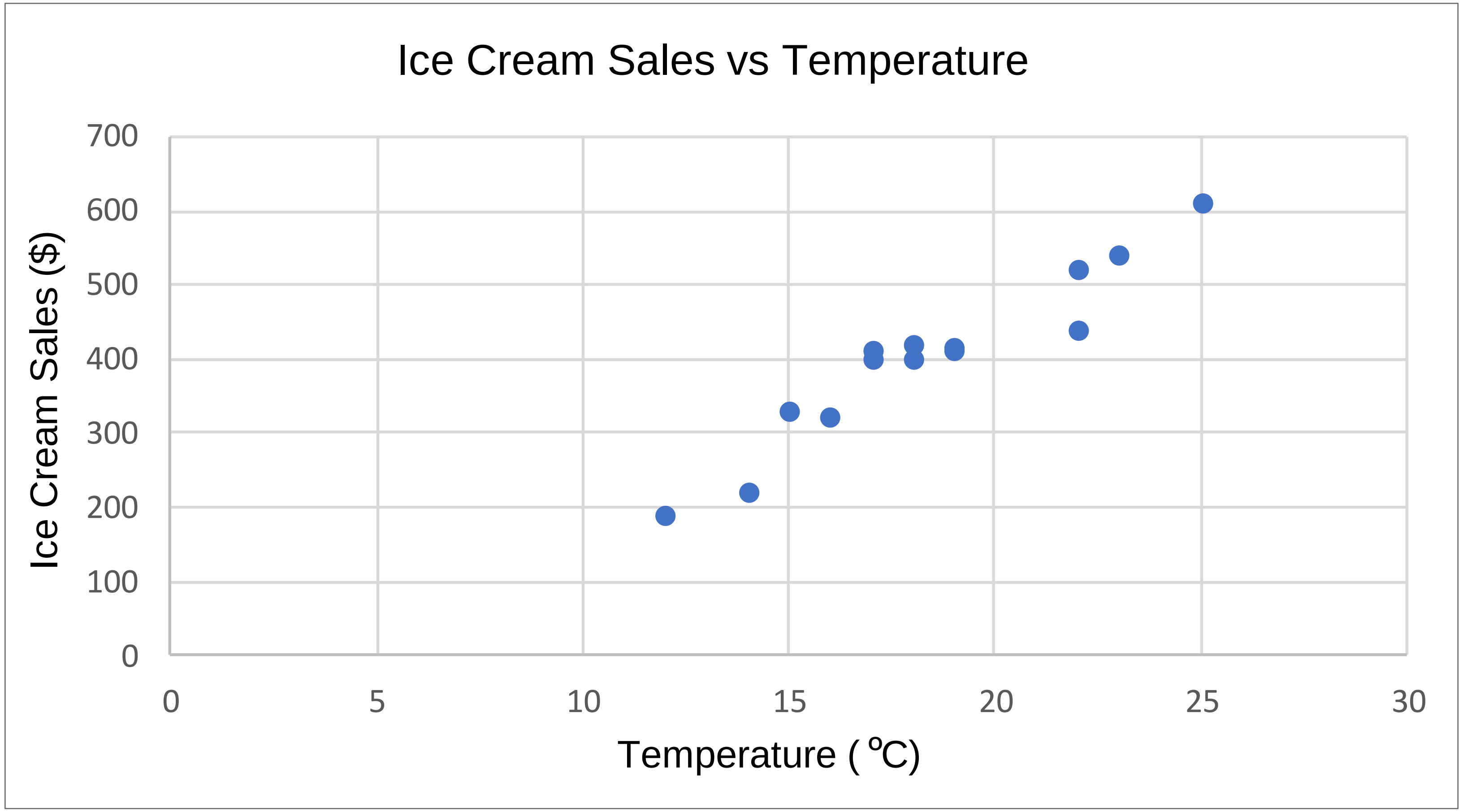
- The ***y*-intercept**:

$$b = \bar{y} - m\bar{x}$$

where:
*m* is the slope
*b* is the *y*-intercept
*n* is the number of observations
*x* is the set of *x*-values of the observations
*y* is the set of *y*-values of the observations
$\bar{x}$ is the mean of the x-values
$\bar{y}$ is the mean of the y-values

# Ice cream sales example

| Ice Cream Sales vs Temperature | |
|---|---|
| Temperature (°C) | Ice Cream Sales ($) |
| 25 | 610 |
| 18 | 400 |
| 16 | 320 |
| 22 | 440 |
| 22 | 520 |
| 19 | 410 |
| 18 | 420 |
| 17 | 410 |
| 23 | 540 |
| 14 | 220 |
| 12 | 190 |
| 15 | 330 |
| 17 | 400 |
| 19 | 415 |




See: https://www.varsitytutors.com/hotmath/hotmath_help/topics/line-of-best-fit

# Ice cream sales example

1. Calculate the mean of the x-values:

$$\bar{x} = \frac{25 + 18 + 16 + 22 + 22 + 19 + 18 + 17 + 23 + 14 + 12 + 15 + 17 + 19}{14} = 18.36$$

2. Calculate the mean of the y-values:

$$\bar{y} = \frac{610 + 400 + 320 + 440 + 520 + 410 + 420 + 410 + 540 + 220 + 190 + 330 + 400 + 415}{14} = 401.79$$

# Ice cream sales example

| | Sales Ice Cream vs Temperature | | Step 3 | | Step 4 | |
|---|---|---|---|---|---|---|
| $i$ | Temperature (°C) | Ice Cream Sales ($) | $x-\bar{x}$ | $y-\bar{y}$ | $(x-\bar{x})$ x $(y-\bar{y})$ | $(x-\bar{x})^2$ |
| 1 | 25 | 610 | 6.64 | 208.21 | 1383.14 | 44.13 |
| 2 | 18 | 400 | -0.36 | -1.79 | 0.64 | 0.13 |
| 3 | 16 | 320 | -2.36 | -81.79 | 192.78 | 5.56 |
| 4 | 22 | 440 | 3.64 | 38.21 | 139.21 | 13.27 |
| 5 | 22 | 520 | 3.64 | 118.21 | 430.64 | 13.27 |
| 6 | 19 | 410 | 0.64 | 8.21 | 5.28 | 0.41 |
| 7 | 18 | 420 | -0.36 | 18.21 | -6.51 | 0.13 |
| 8 | 17 | 410 | -1.36 | 8.21 | -11.15 | 1.84 |
| 9 | 23 | 540 | 4.64 | 138.21 | 641.71 | 21.56 |
| 10 | 14 | 220 | -4.36 | -181.79 | 792.07 | 18.98 |
| 11 | 12 | 190 | -6.36 | -211.79 | 1346.35 | 40.41 |
| 12 | 15 | 330 | -3.36 | -71.79 | 240.99 | 11.27 |
| 13 | 17 | 400 | -1.36 | -1.79 | 2.42 | 1.84 |
| 14 | 19 | 415 | 0.64 | 13.21 | 8.49 | 0.41 |
| | 18.36 | 401.79 | | | 5166.07 | 173.21 |
| | Step 1 $\bar{x}$ | Step 2 $\bar{y}$ | | | | |

# Ice cream sales example

5. Calculate the slope:

$$m = \frac{\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$m = \frac{5166.07}{173.21} = 29.82$$

6. Calculate the y-intercept:

$$b = \bar{y} - m\bar{x}$$

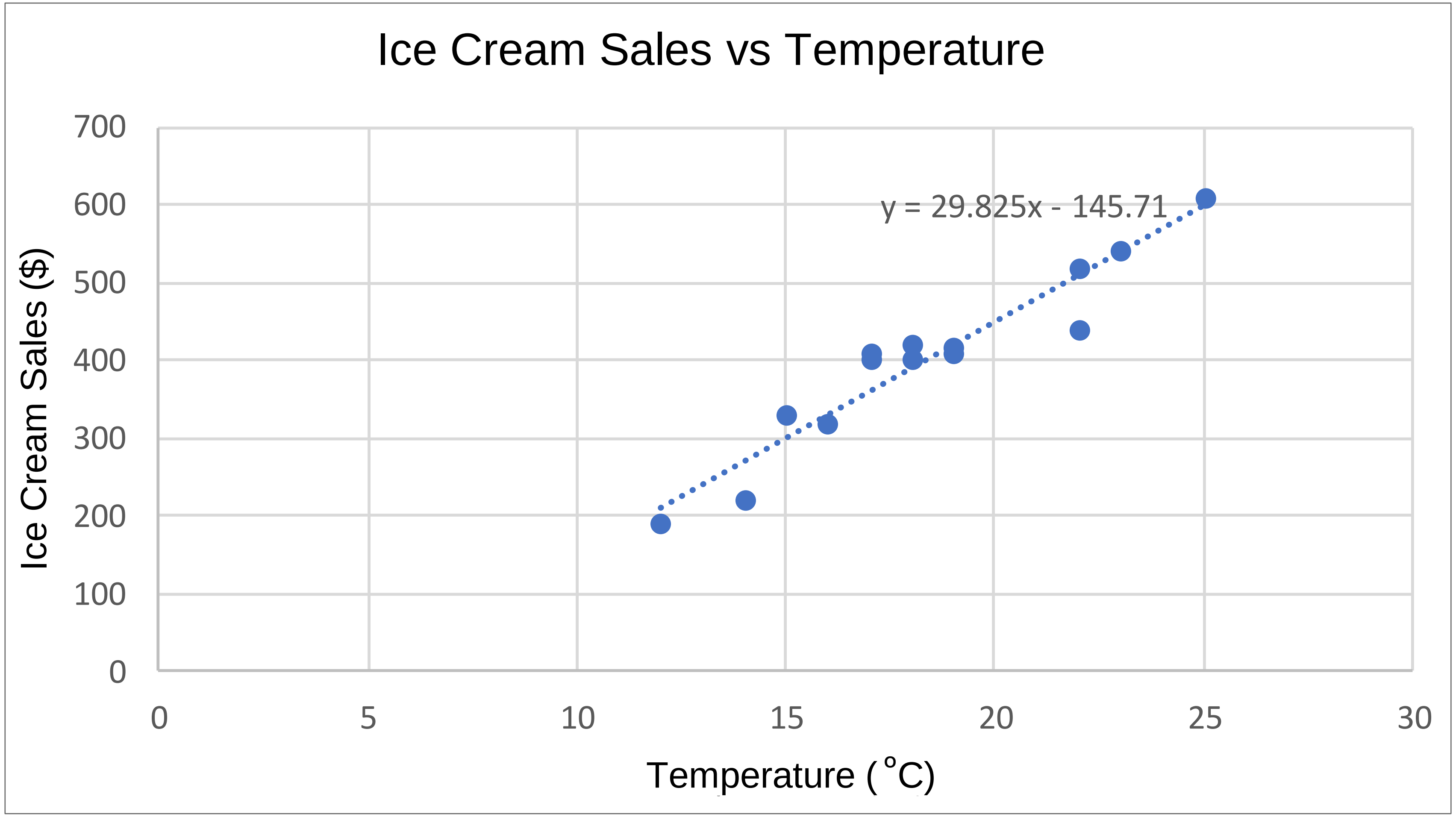
$b$ =401.79 – (29.82 x 18.36) = -145.71

Therefore the line of best fit is:

***y* = 29.82*x* -145.71**

# Ice cream sales example




For MS Excel instructions, see:
https://support.office.com/en-ie/article/add-a-trend-or-moving-average-line-to-a-chart-fa59f86c-5852-4b68-a6d4-901a745842ad

## Scatter Plots and Line of Best Fit

# Summary

- A scatter plot can be used to show the relationship between two variables.
- The **line of best fit** is the line that describes the relationship between the two variables, where the sum of the squares of the residual errors between the individual data values and the line is at its minimum.
- Therefore, it is the best possible straight line that fits the data.
- Slope (*m*) and *y*-intercept (*b*) are the two values needed to define the equation of a straight line $y = mx + b$.

# Thank you.

**Contact information:**

**open@sap.com**

Follow all of SAP

www.sap.com/contactsap

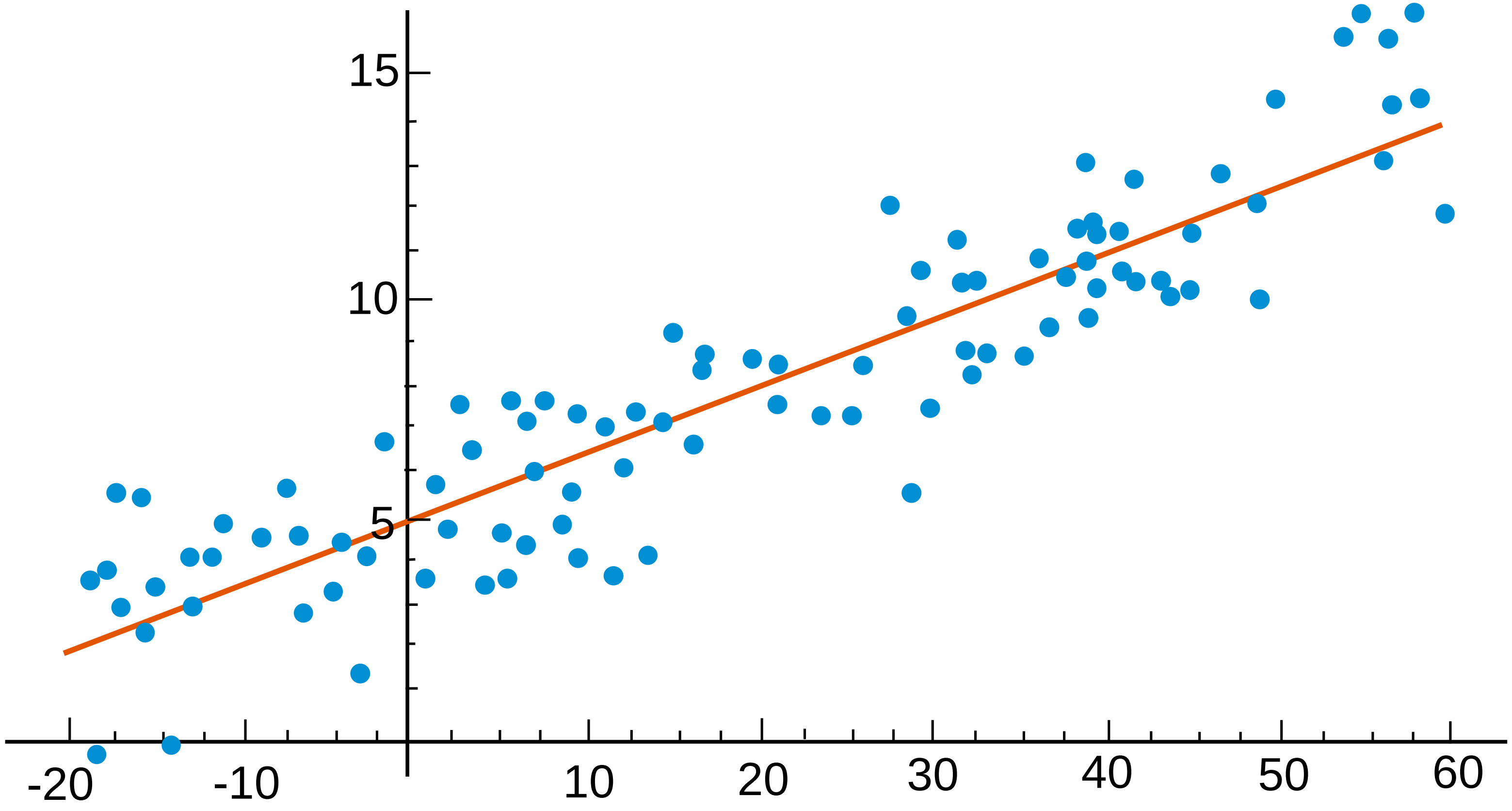


THE BEST RUN SAP

Week 3: Correlation and Linear Regression

# Unit 4: Linear Regression

# Introduction



# Uses

## **Prediction** or **forecasting**

- Fit a predictive model to an observed data set of target and explanatory variables
- Make predictions of the target values

## **Explain variation** in target

- Quantify the strength of the relationship between the target and explanatory variables
- Determine if some explanatory variables have no linear relationship with the target

Linear Regression

# Least squares

Residual = Fitted Value – Observed Value

| | Sales Ice Cream vs Temperature | | | |
|---|---|---|---|---|
| *i* | Temperature ($^{o}$C) x | Ice Cream Sales ($) y | Regression y = 29.82x - 145.71 | Residual |
| 1 | 25 | 610 | 600.0 | -10.0 |
| 2 | 18 | 400 | 391.2 | -8.8 |
| 3 | 16 | 320 | 331.5 | 11.5 |
| 4 | 22 | 440 | 510.5 | 70.5 |
| 5 | 22 | 520 | 510.5 | -9.5 |
| 6 | 19 | 410 | 421.0 | 11.0 |
| 7 | 18 | 420 | 391.2 | -28.8 |
| 8 | 17 | 410 | 361.4 | -48.6 |
| 9 | 23 | 540 | 540.3 | 0.3 |
| 10 | 14 | 220 | 271.9 | 51.9 |
| 11 | 12 | 190 | 212.2 | 22.2 |
| 12 | 15 | 330 | 301.7 | -28.3 |
| 13 | 17 | 400 | 361.4 | -38.6 |
| 14 | 19 | 415 | 421.0 | 6.0 |
| | | | Observed Value | Fitted Value |

https://en.wikipedia.org/wiki/Ordinary_least_squares

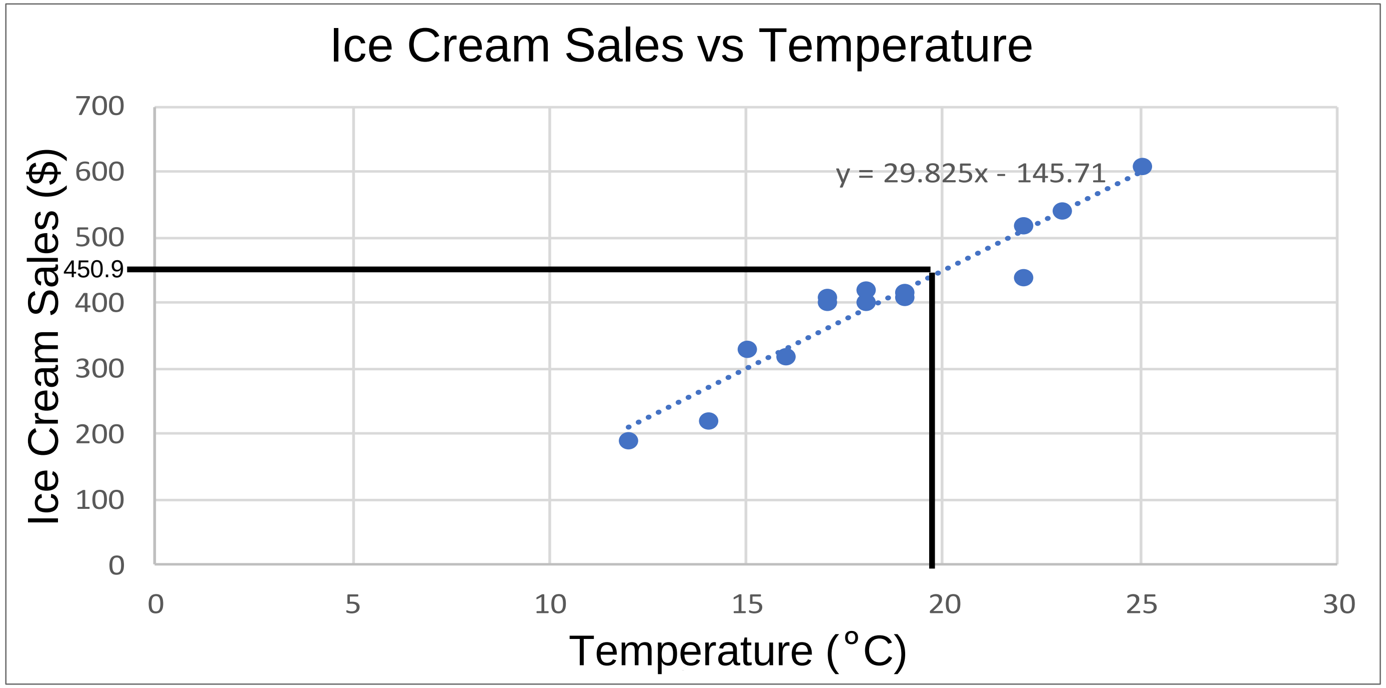
# Forecast future values

- In the example, the linear regression model is:

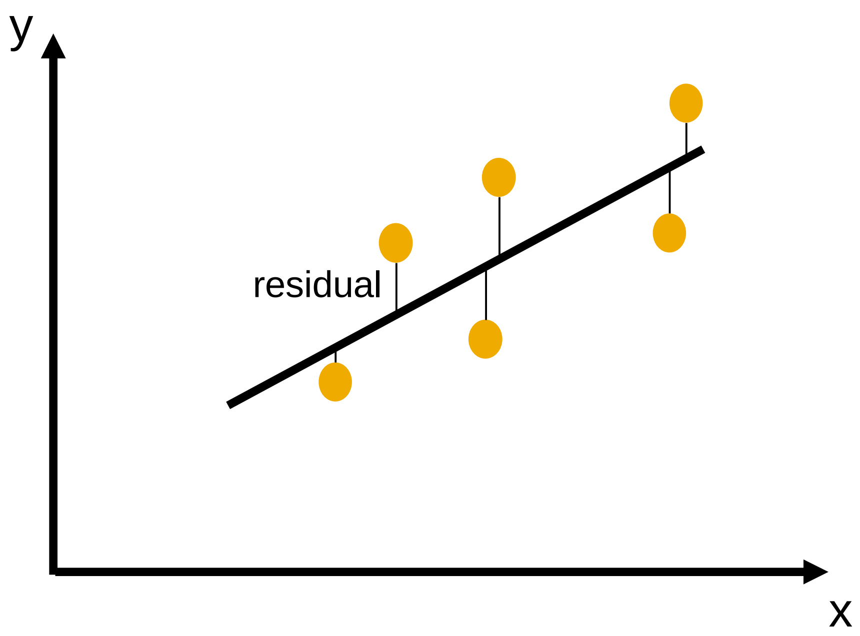
$$\text{Sales (\$)} = 29.82 \times \text{Temperature } (^{o}\text{C}) - 145.71$$

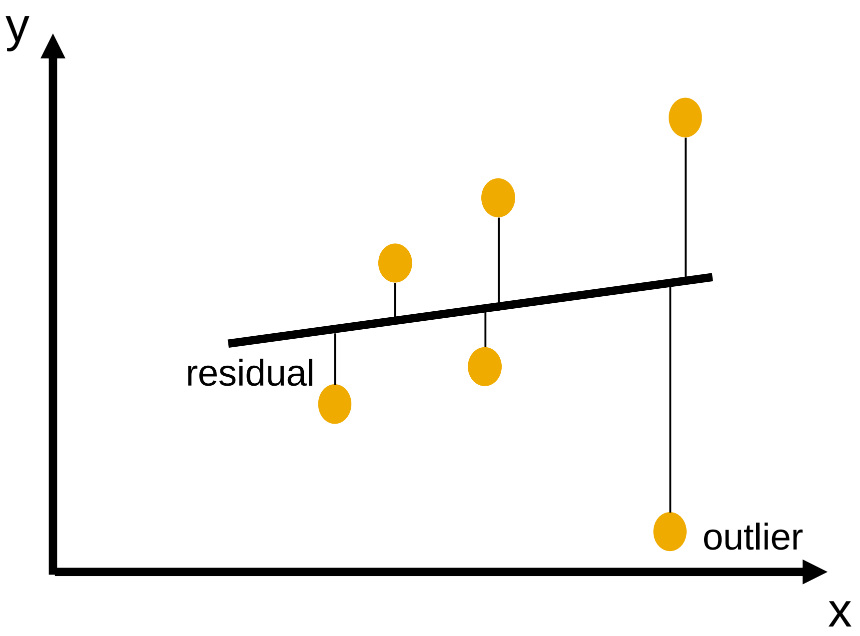
- Therefore, if the temperature tomorrow is forecasted to be 20°C, then the store can expect to sell:

$$(29.82 \times 20) - 145.71 = \$450.9 \text{ worth of ice cream tomorrow}$$



# Least squares and outliers



# Assumptions

**Linear relationship** between the explanatory and target variables

No or low "**multicollinearity**"

No "**auto-correlation**"

**Homoscedastic**

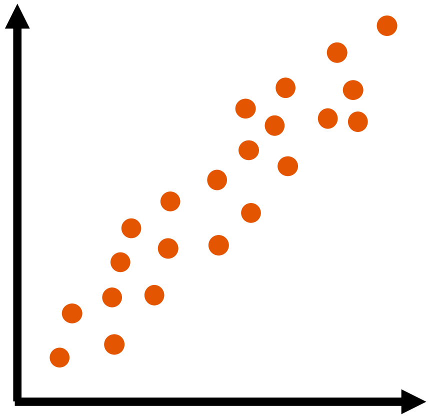
See: http://r-statistics.co/Assumptions-of-Linear-Regression.html and http://statisticsbyjim.com/regression/ols-linear-regression-assumptions/

# Assumption 1 – Linear relationship

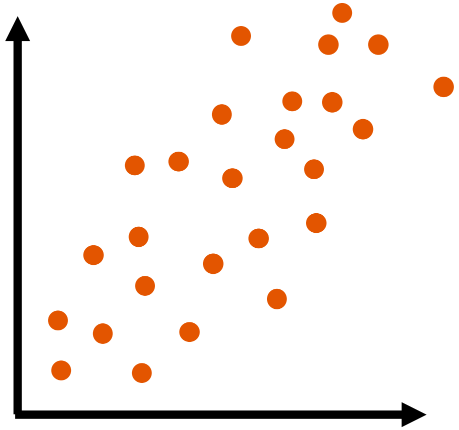
- This rule constrains the model to one type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

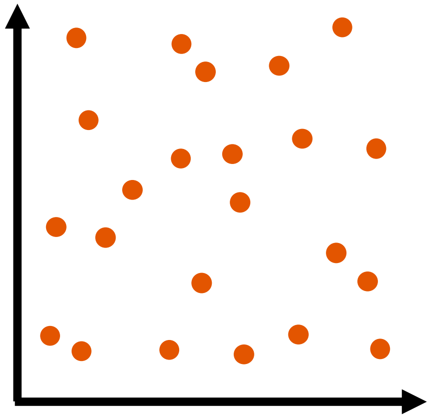
- The linearity assumption can be tested with scatter plots:



Strong linear relationship



Weaker linear relationship

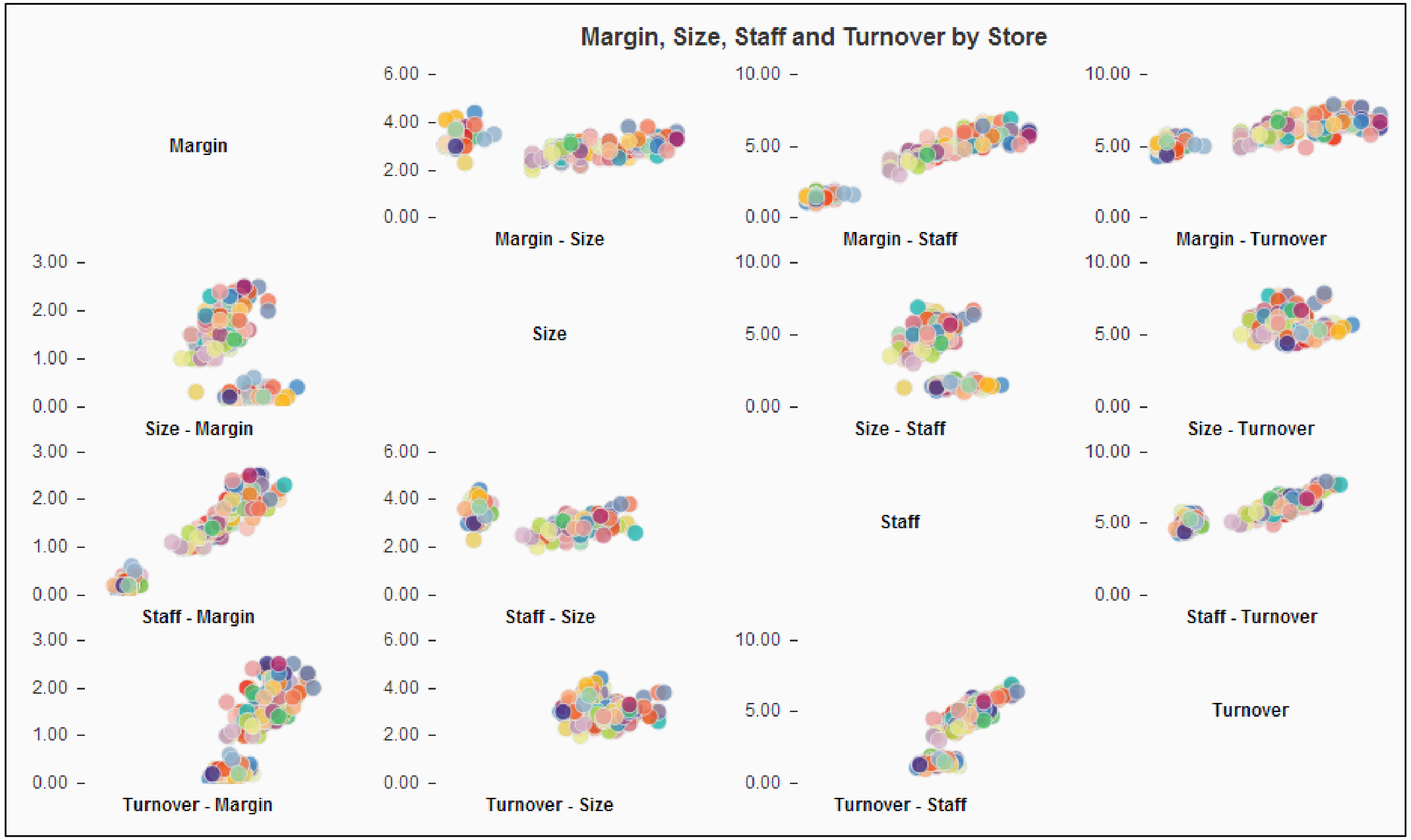


No linear relationship

See Appendix

# Assumption 2 – No or low multicollinearity

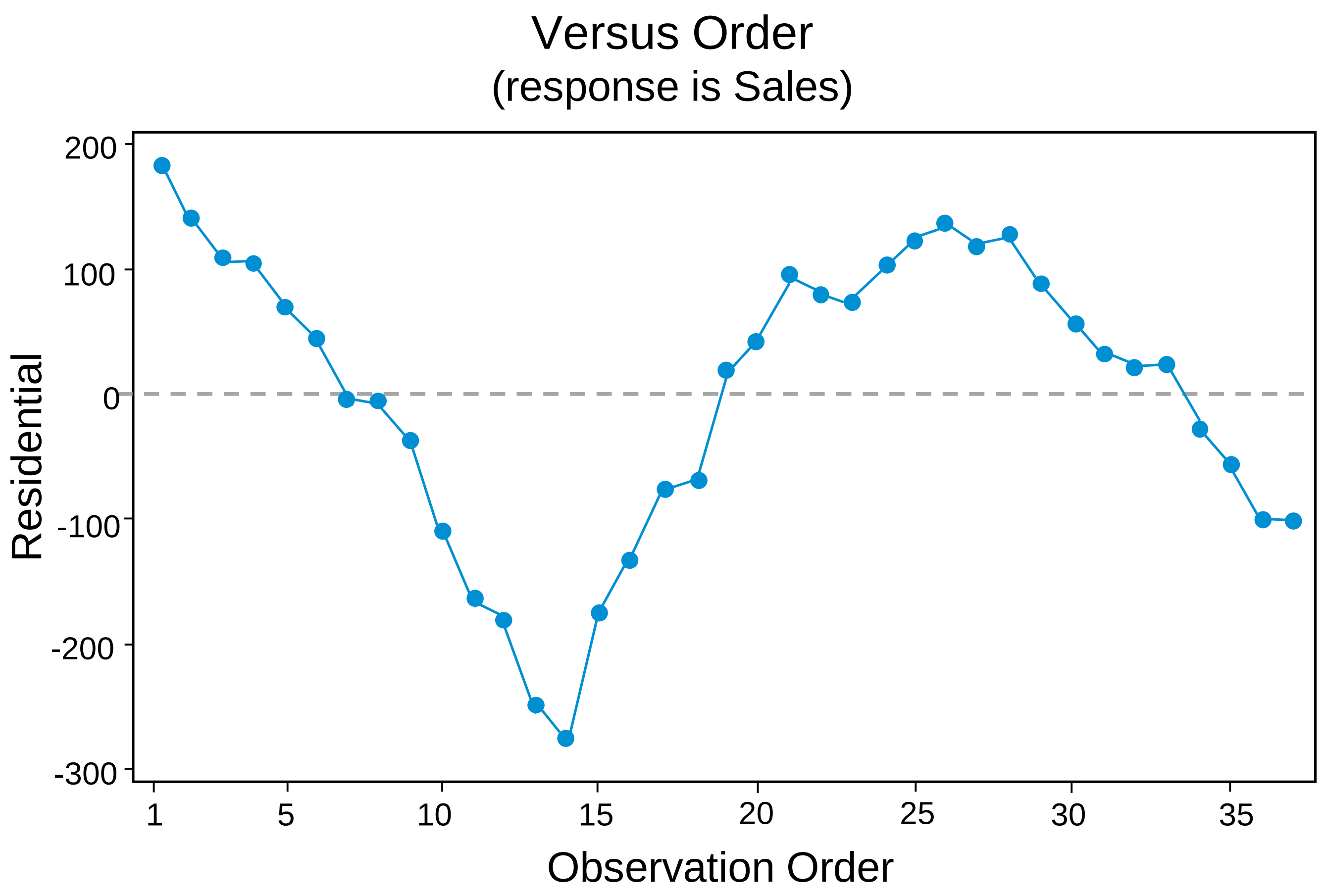
- This visualization shows the scatter plot matrix in SAP Lumira.
- In this example, there are four variables that are plotted against each other.
- It is a powerful way of visualizing the correlation between variables and identifying patterns and groups in the data.




See: https://en.wikipedia.org/wiki/Multicollinearity

See Appendix

# Assumption 3 – No autocorrelation

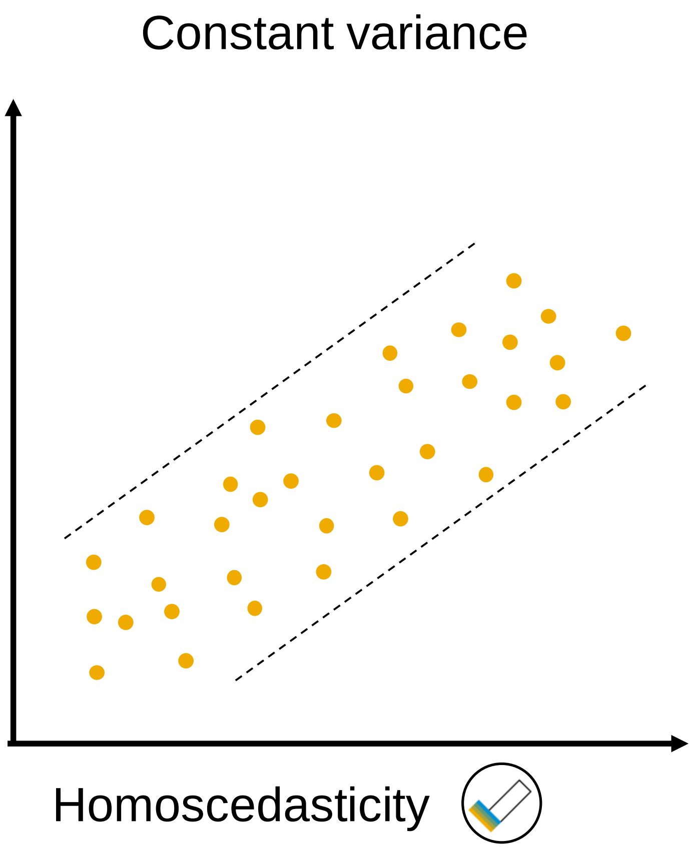


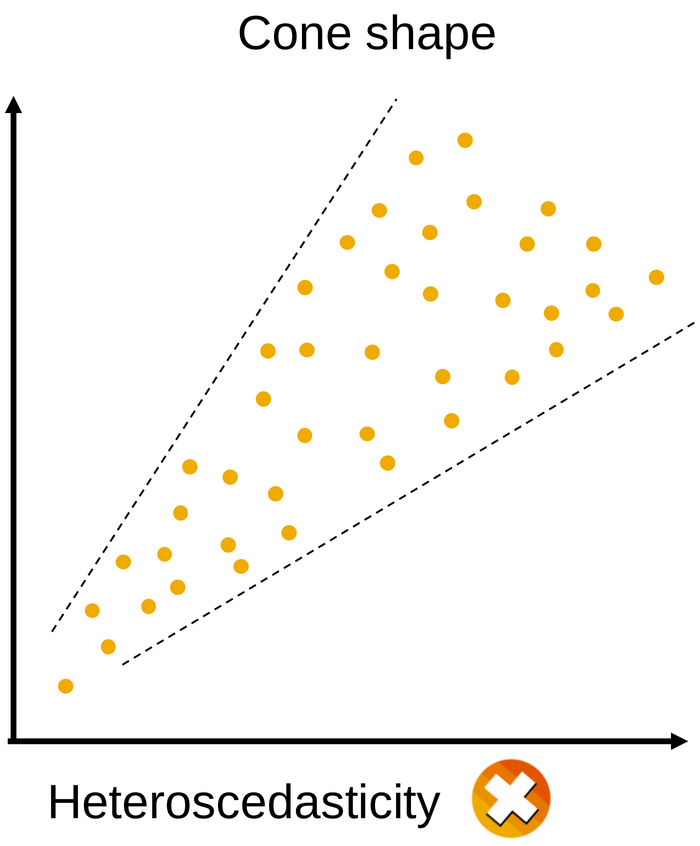

In this example, there appears to be a cyclical pattern with a positive correlation.

http://statisticsbyjim.com/regression/ols-linear-regression-assumptions/

See Appendix

# Assumption 4 – Homoscedasticity

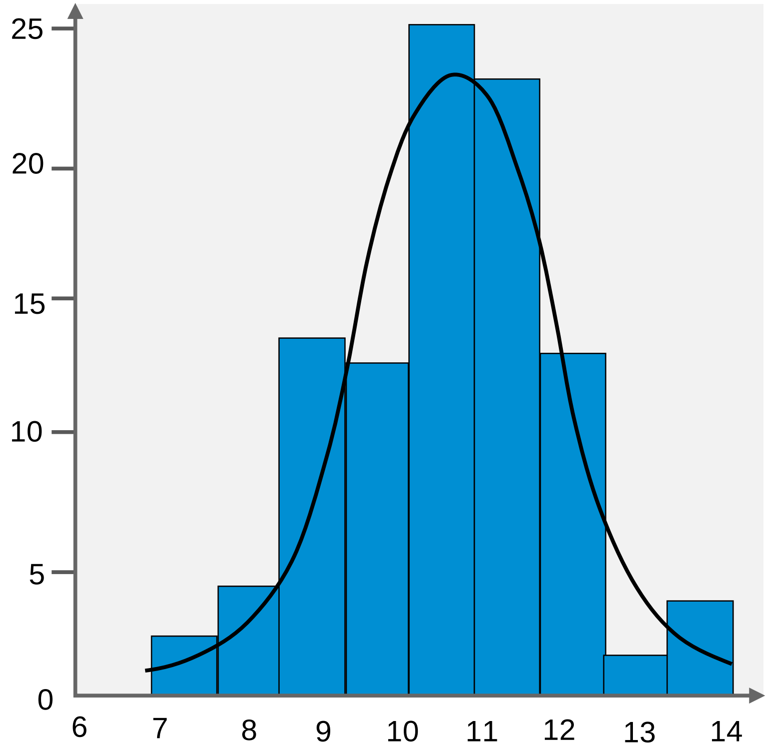


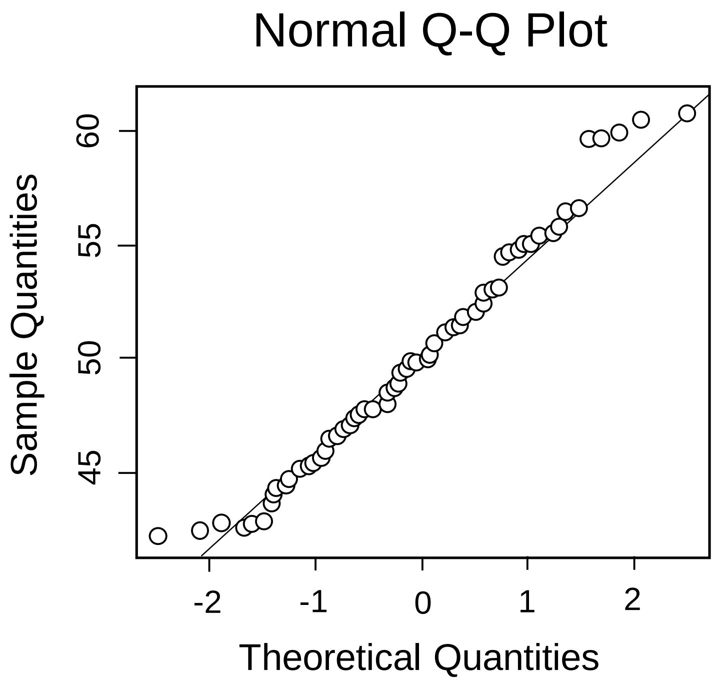

See Appendix

https://en.wikipedia.org/wiki/Heteroscedasticity

# Understanding the distribution of the explanatory variables




Histogram




Q-Q Plot

See: https://www.theanalysisfactor.com/the-distribution-of-independent-variables-in-regression-models/ and https://www.researchgate.net/post/Should_I_transform_non-normal_independent_variables_in_logistic_regression

Linear Regression

# Summary

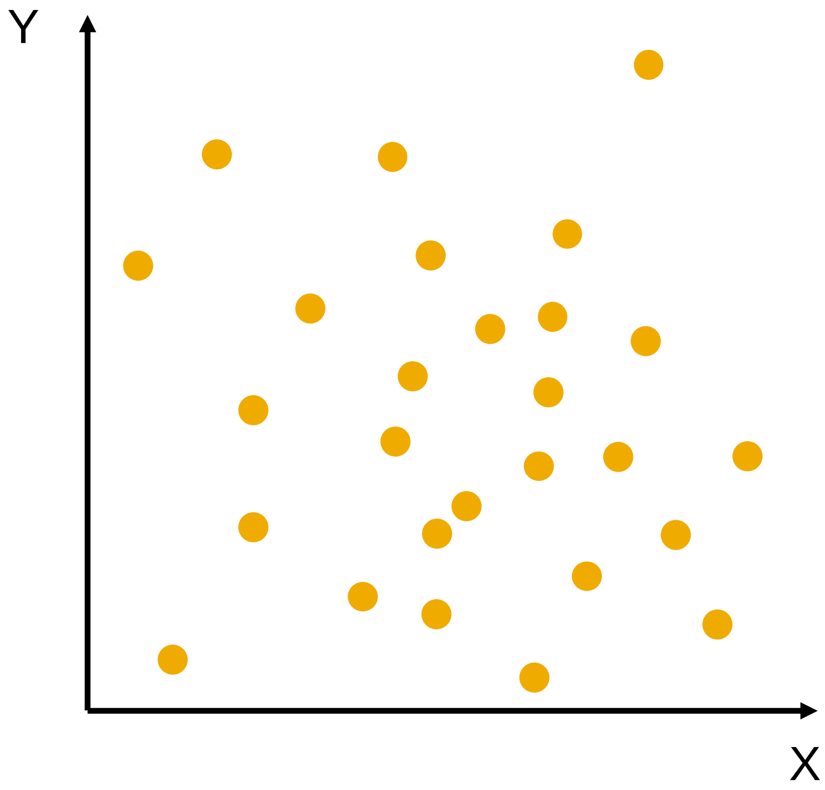
- Linear regression is an approach to modeling the linear relationship between a target variable and one or more explanatory variables.
- **Simple linear regression** has one explanatory variable and **multiple linear regression** has more than one explanatory variable.
- In summary, your linear regression model should produce residuals that have a mean of zero, have a constant variance, and are not correlated with themselves or other variables. If these assumptions are true, then the ordinary least squares regression procedure will create the best possible estimates.

# Appendix – Assumption 1 – Linear relationship

- Linear regression requires the relationship between the explanatory and target variables to be linear.
- This assumption addresses the functional form of the model. The regression model is linear when all terms in the model are either the constant or a parameter multiplied by an explanatory variable.
- You build the model equation only by adding the terms together. This rule constrains the model to one type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

- In this equation, the betas (βs) are the parameters that the ordinary least squares process estimates, and epsilon (ε) is the random error.
- The linearity assumption can be tested with scatter plots.




This scatter plot example shows there is **no** linear relationship between the target (Y) variable and the explanatory (X) variable.

# Appendix – Assumption 2 – No or low multicollinearity

- Multicollinearity occurs when the explanatory variables are highly correlated with each other.
- Multicollinearity may be analyzed in a variety of ways. For example:
  1. Correlation matrix – compute the matrix of Pearson's correlation coefficients for each explanatory variable.
  2. Tolerance – the tolerance (T) measures the influence of one explanatory variable on all the other independent variables. It is calculated by regressing the explanatory variable of interest onto the remaining explanatory variables included in the regression analysis. Then the tolerance is used to calculate the "variance inflation factor".

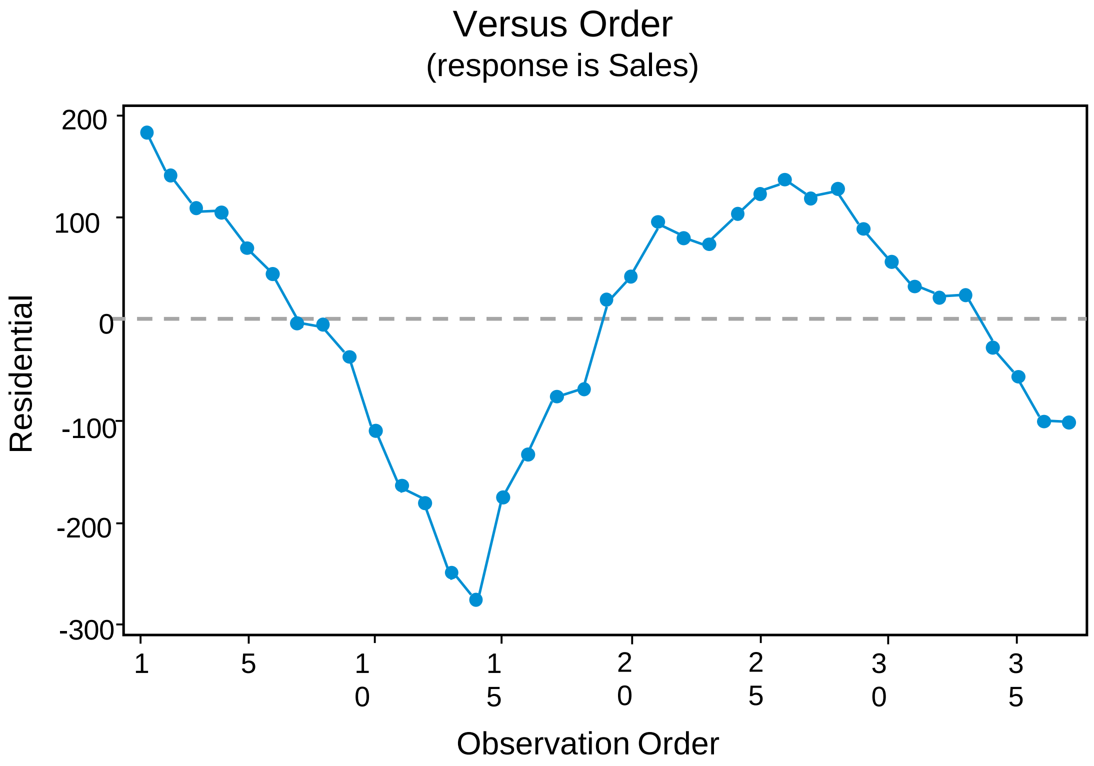
Variance Inflation Factor (VIF) = 1/Tolerance

- With VIF > 10 there is an indication that multicollinearity may be present. With VIF > 100 there is definitely multicollinearity among the variables.
- If multicollinearity is found in the data the simplest way to address the problem is to remove one of the correlated variables! However, there are a range of more sophisticated techniques available.


This visualization shows the scatter plot matrix in SAP Lumira. In this example, there are four variables that are plotted against each other. It is a powerful way of visualizing the correlation between variables and identifying patterns and groups in the data.

# Appendix – Assumption 3 – No autocorrelation

- Linear regression analysis requires that there is little or no "**autocorrelation**" in the residuals. This means that the error terms must be uncorrelated so that one observation of the error term should not predict the next observation.
- Autocorrelation occurs when the residuals are not independent of each other, in other words, when the value of y(x+1) is not independent of the value of y(x).
- For instance, if the error for one observation is positive and that increases the probability that the following error is positive, then there is a positive correlation.
- You can assess if this assumption is violated by graphing the residuals in the order that the data was collected. You hope to see a randomness in the plot.

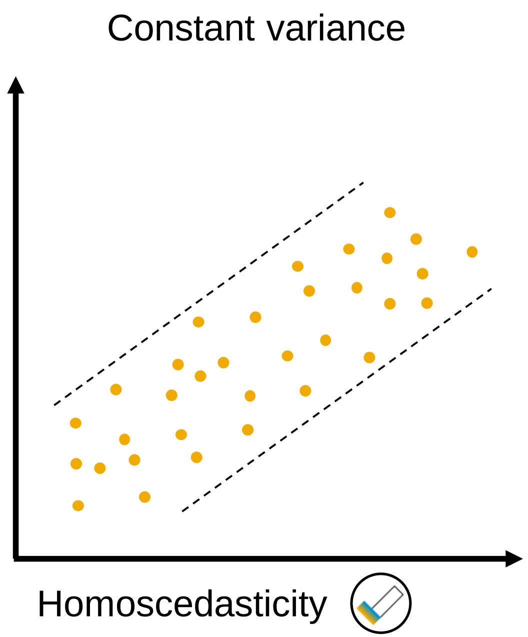


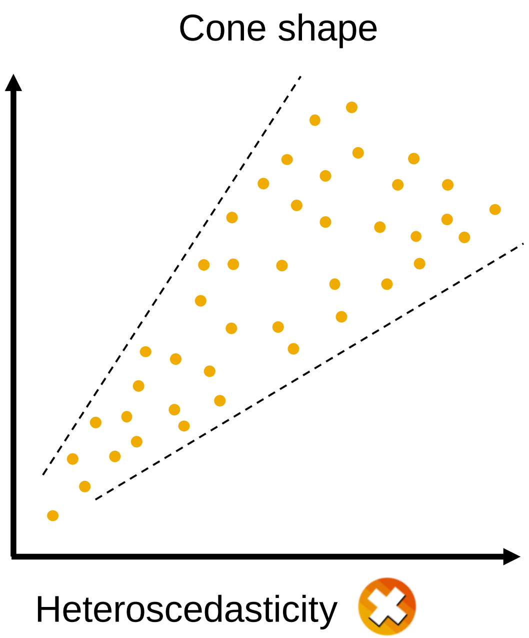

In this example there appears to be a cyclical pattern with a positive correlation.

[http://statisticsbyjim.com/regression/ols-linear-regression-assumptions/](http://statisticsbyjim.com/regression/ols-linear-regression-assumptions/)

# Appendix – Assumption 4 – Homoscedasticity

- The variance of the errors should be consistent for all observations. This means that the variance does not change for each observation or for a range of observations.
- The scatter plot is good way to check whether the data is **homoscedastic** (which simply means that the residuals are equal across the regression line).
- You can check this assumption by plotting the residuals against the fitted values. Heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction.
- The following scatter plots show examples of data that is not homoscedastic (i.e., it is **heteroscedastic**).




https://en.wikipedia.org/wiki/Heteroscedasticity

# Thank you.

**Contact information:**

**open@sap.com**

Follow all of SAP

www.sap.com/contactsap

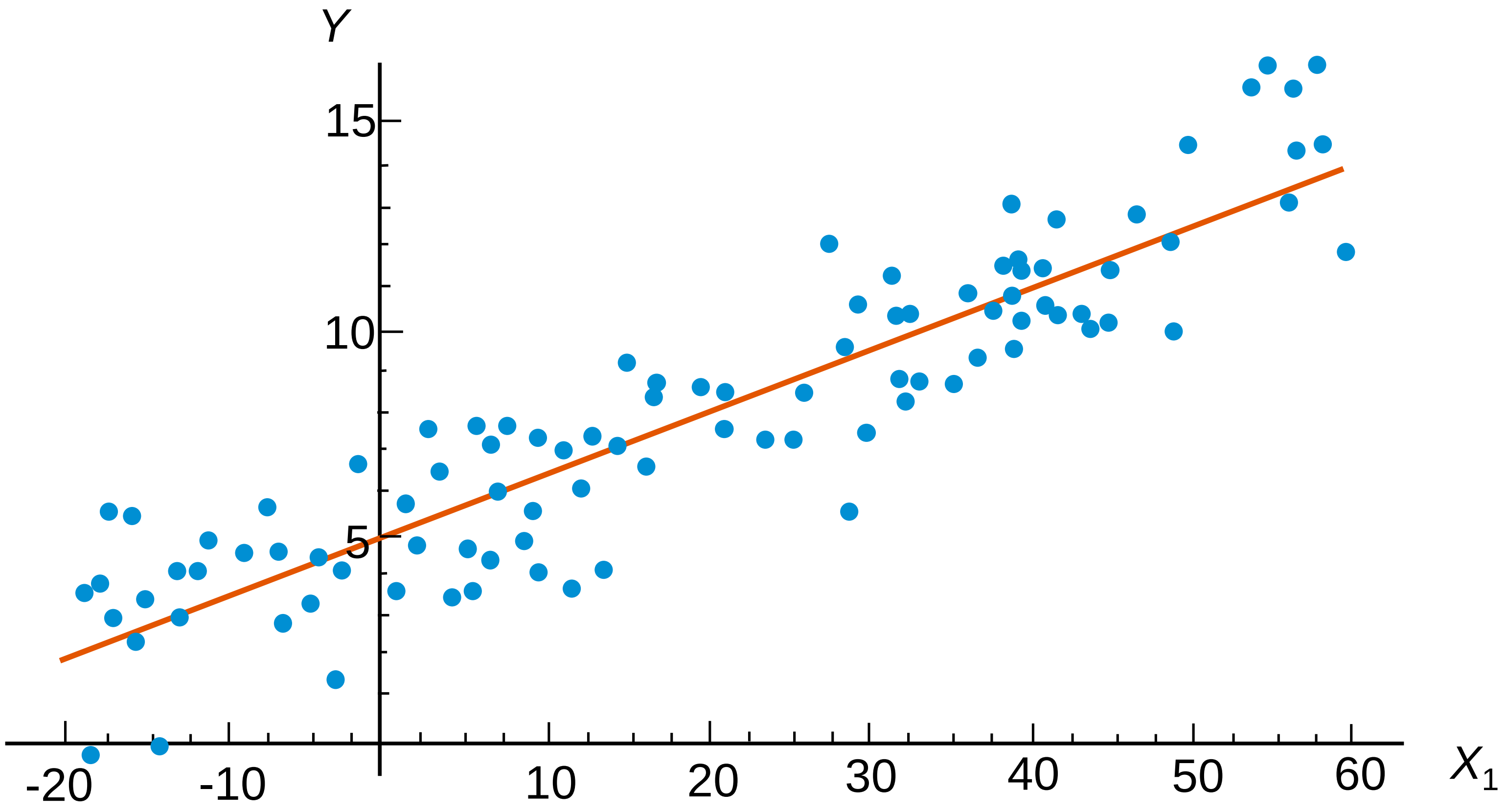


THE BEST RUN SAP®

Week 3: Correlation and Linear Regression

# Unit 5: Interpreting Results

## Reminder




$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

# Example using R

| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors |
|---|---|---|---|---|---|---|
| 1 | 221900 | 3 | 3 | 1180 | 5650 | 1 |
| 2 | 538000 | 3 | 3 | 2570 | 7242 | 2 |
| 3 | 180000 | 2 | 2 | 770 | 10000 | 1 |
| 4 | 604000 | 4 | 4 | 1960 | 5000 | 1 |
| 5 | 510000 | 3 | 3 | 1680 | 8080 | 1 |
| 6 | 1230000 | 4 | 4 | 5420 | 101930 | 1 |
| 7 | 257500 | 3 | 3 | 1715 | 6819 | 2 |
| 8 | 291850 | 3 | 3 | 1060 | 9711 | 1 |
| 9 | 229500 | 3 | 3 | 1780 | 7470 | 1 |
| 10 | 323000 | 3 | 3 | 1890 | 6560 | 2 |
| 11 | 662500 | 3 | 3 | 3560 | 9796 | 1 |
| 12 | 468000 | 2 | 2 | 1160 | 6000 | 1 |
| 13 | 400000 | 3 | 3 | 1370 | 9680 | 1 |

```
     price            bedrooms         bathrooms        sqft_living
 Min.   :  75000   Min.   : 0.000   Min.   : 0.000   Min.   :  290
 1st Qu.: 320000   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 1420
 Median : 450000   Median : 3.000   Median : 3.000   Median : 1910
 Mean   : 539458   Mean   : 3.369   Mean   : 3.369   Mean   : 2080
 3rd Qu.: 640000   3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 2550
 Max.   :7700000   Max.   :10.000   Max.   :10.000   Max.   :13540
    sqft_lot            floors
 Min.   :    520   Min.   :1.000
 1st Qu.:   5050   1st Qu.:1.000
 Median :   7616   Median :1.000
 Mean   :  15092   Mean   :1.447
 3rd Qu.:  10665   3rd Qu.:2.000
 Max.   :1651359   Max.   :3.000
```

There are 17384 rows and 6 columns in the data set.

"house_train_data" is available from openSAP

Import data into RStudio

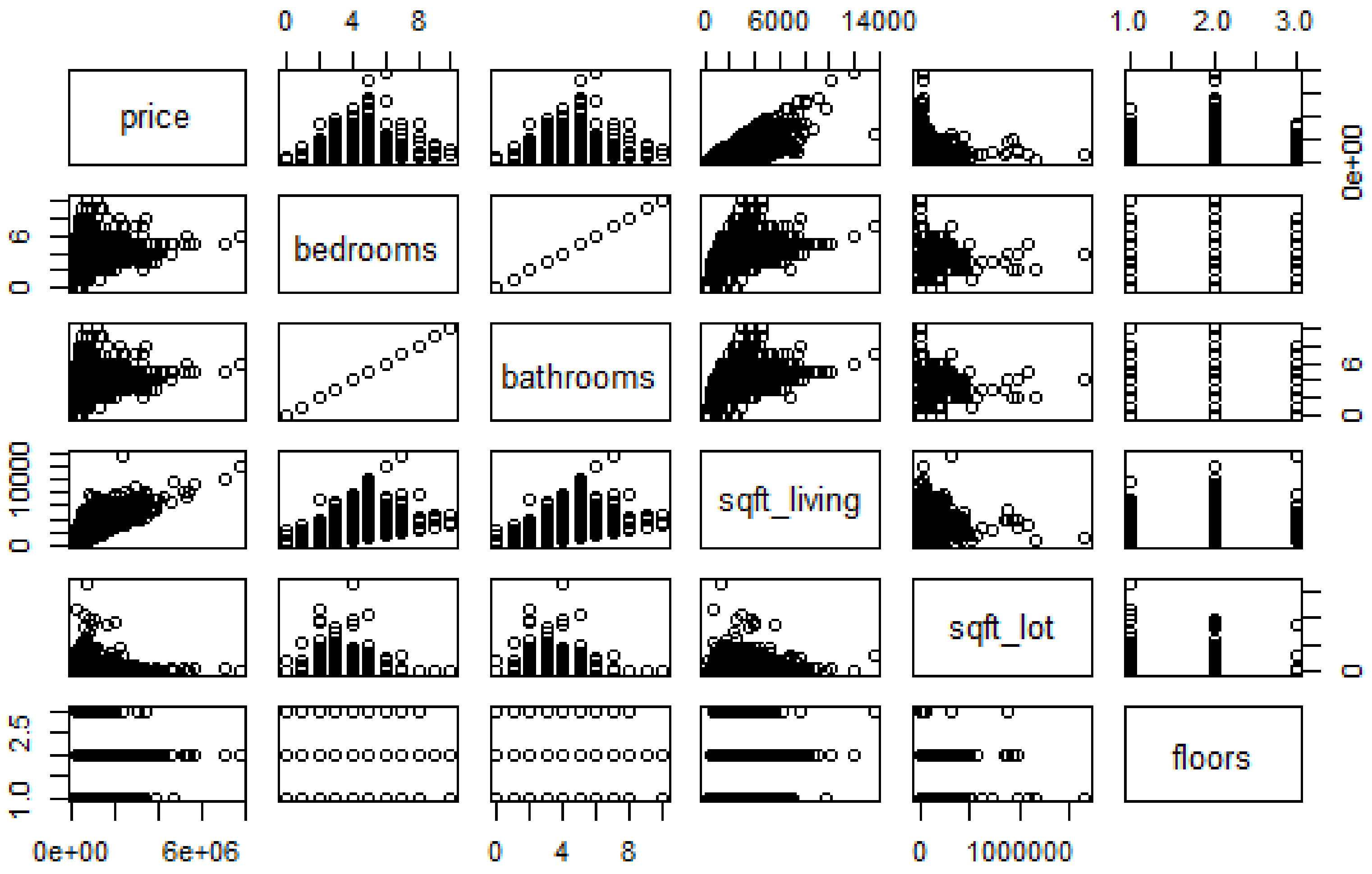
Data Summary results

using the R command
>summary(house_train_data)

For RStudio download see: https://www.rstudio.com/

For an introduction to RStudio see: https://datascienceplus.com/introduction-to-rstudio/

# Visualize the data




This is a simple visualization using the R command: >plot(house_train_data)

# Building the linear regression model

- Fit a multiple linear regression model with price as the target variable and the other variables as the explanatory variables.
- The R command is:

```
>results=lm(price~bedrooms+bathrooms+sqft_living+sqft_lot+floors,data=(house_train_data))

>results
```

# Results of the linear regression model

- The output gives the parameters for the linear regression model that is built:

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors, data = (house_train_data))

Coefficients:
(Intercept)     bedrooms    bathrooms  sqft_living     sqft_lot       floors
  1.169e+05   -6.710e+04           NA    3.281e+02   -3.883e-01   -1.934e+04
```

- This output indicates that the fitted value is given by:

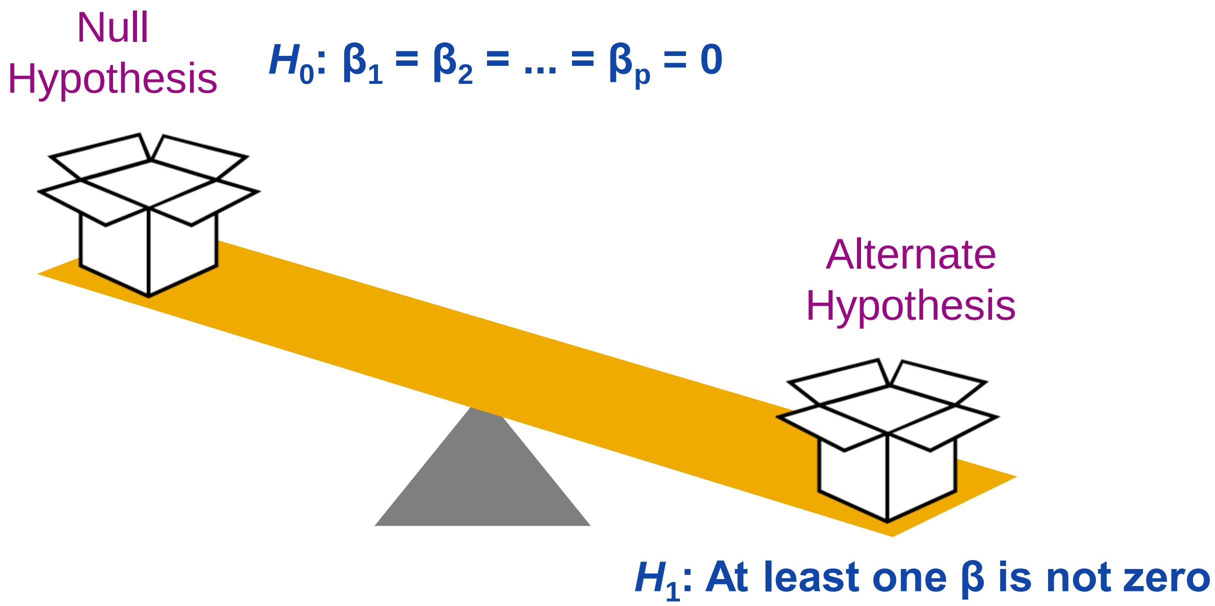
$$\hat{y} = 1.169e^{5} - 6.710e^{4}x_1 + 3.281e^{2}x_3 - 3.883e^{-1}x_4 - 1.934e^{4}x_5$$

# Rejecting the null hypothesis

- The "**null hypothesis**" is represented by $H_0$:

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$$

- When you "test the null hypothesis" it means you are assessing the probability that there is **no** relationship between the explanatory variables and the target variable.
- You then either accept or reject this null hypothesis.



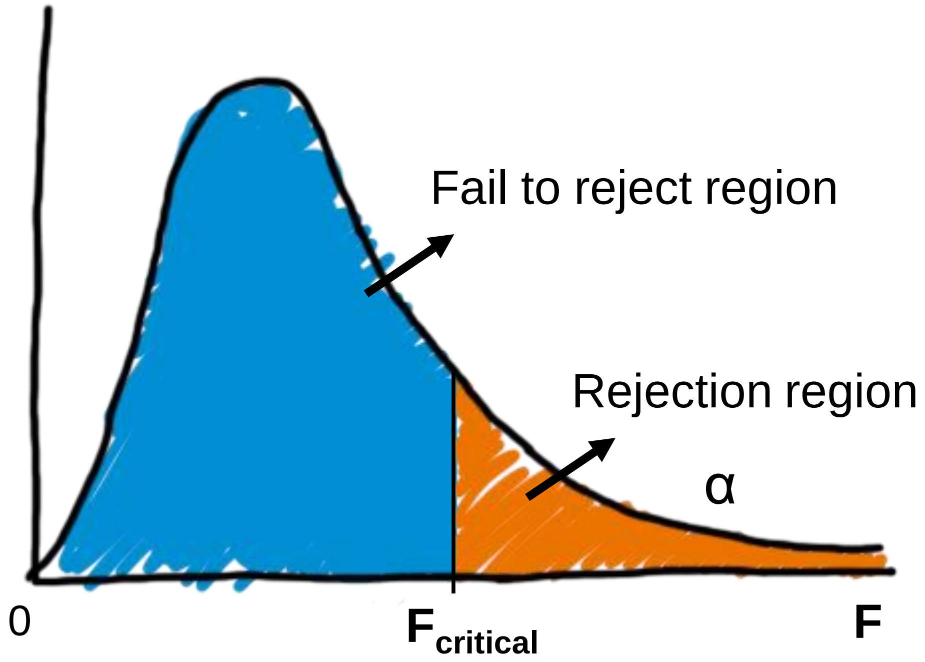

See: https://stattrek.com/regression/slope-test.aspx

# Rejecting the null hypothesis

- The **F-Test** and **F-statistic** refers to the test statistic used to decide whether the model **as a whole** has statistically significant predictive capability.
- The R command is:

  > summary(results)

```
Residual standard error: 258300 on 17379 degrees of freedom
Multiple R-squared:  0.5127,    Adjusted R-squared:  0.5126
F-statistic:  4571 on 4 and 17379 DF,  p-value: < 2.2e-16
```




Learn more about the F-statistic here: http://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/ and http://facweb.cs.depaul.edu/sjost/csc423/documents/f-test-reg.htm and https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/f-statistic-value-test/

Example F values here: http://www.stat.purdue.edu/~jtroisi/STAT350Spring2015/tables/FTable.pdf and explained here: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm

# Significance of individual variables

The results show that the variable “bedrooms” is significant (controlling for the other explanatory variables) with a p < 2.2e-16 (which is less than 0.05).

```
Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.169e+05  8.670e+03  13.478  < 2e-16 ***
bedrooms    -6.710e+04  2.696e+03 -24.894  < 2e-16 ***
bathrooms           NA         NA      NA       NA    
sqft_living  3.281e+02  2.841e+00 115.484  < 2e-16 ***
sqft_lot    -3.882e-01  4.824e-02  -8.048 8.98e-16 ***
floors      -1.934e+04  3.816e+03  -5.068 4.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 258300 on 17379 degrees of freedom
Multiple R-squared:  0.5127,	Adjusted R-squared:  0.5126 
F-statistic:  4571 on 4 and 17379 DF,  p-value: < 2.2e-16
```

Interpreting Results

## Significance of individual variables

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.169e+05 | 8.670e+03 | 13.478 | < 2e-16 |
| bedrooms | -6.710e+04 | 2.696e+03 | -24.894 | < 2e-16 |
| bathrooms | NA | NA | NA | NA |
| sqft_living | 3.281e+02 | 2.841e+00 | 115.484 | < 2e-16 |
| sqft_lot | -3.882e-01 | 4.824e-02 | -8.048 | 8.98e-16 |
| floors | -1.934e+04 | 3.816e+03 | -5.068 | 4.06e-07 |

## Interpreting Results
# R-squared

- R-squared is the percentage of the target variable variation that is explained by a linear model.
- R-squared is always between 0 and 100% (if there is an intercept value):
  – 0% indicates that the model explains none of the variability of the target data around its mean.
  – 100% indicates that the model explains all the variability of the target data around its mean.

```
Multiple R-squared:  0.5127,    Adjusted R-squared:  0.5126
```

See: https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
https://en.wikipedia.org/wiki/Coefficient_of_determination

# Interpreting regression coefficients

```
Coefficients: (1 not defined because of singulariti
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.169e+05  8.670e+03  13.478  < 2e-16
bedrooms    -6.710e+04  2.696e+03 -24.894  < 2e-16
bathrooms           NA         NA      NA       NA
sqft_living  3.281e+02  2.841e+00 115.484  < 2e-16
sqft_lot    -3.882e-01  4.824e-02  -8.048 8.98e-16
floors      -1.934e+04  3.816e+03  -5.068 4.06e-07
---
```

Interpreting Results

# Multicollinearity

- There is an error identified in the output:

```
Coefficients: (1 not defined because of singularities)
```

- The R command to produce the correlation matrix  is:

```
> cor(house_train_data)
```

```
                 price   bedrooms  bathrooms sqft_living     sqft_lot
price       1.00000000 0.31283767 0.31283767   0.7029343  0.088236721
bedrooms    0.31283767 1.00000000 1.00000000   0.5910598  0.030179053
bathrooms   0.31283767 1.00000000 1.00000000   0.5910598  0.030179053
sqft_living 0.70293434 0.59105983 0.59105983   1.0000000  0.166967283
sqft_lot    0.08823672 0.03017905 0.03017905   0.1669673  1.000000000
floors      0.23296365 0.16065981 0.16065981   0.3521290 -0.007346747
                  floors
price        0.232963645
bedrooms     0.160659811
bathrooms    0.160659811
sqft_living  0.352129048
sqft_lot    -0.007346747
floors       1.000000000
```

Perfect positive
correlation

Interpreting Results

# Summary

- Multiple linear regression is used to describe data and to explain the relationship between one target variable and two or more explanatory variables
- The analysis requires you to analyze the correlation and directionality of the data, train (fit) the model, and then evaluate the validity and usefulness of the model.
- This unit introduced you to some of the results that are generated that will help you evaluate your model. You must assess these very carefully so that you can be sure that your model is valid.
- Remember that it is also very important that you check and confirm that the basic assumptions for linear regression that were discussed in the previous unit hold true.

# Thank you.

**Contact information:**

**open@sap.com**

Follow all of SAP

www.sap.com/contactsap



THE BEST RUN SAP