Relationship between two variables

- For quantitative variables, you previously studied:
  - Measures of Central Tendency (the Average)
  - Variability: Dispersion or Spread
- These explained much about a single quantitative variable like the height of professional soccer players or the number of children in a family.
- However, social scientists are also interested in how two quantitative variables move together.
- Specifically, if a value in series A is higher, would we expect a corresponding value in series B to be higher, lower, or the same.
- Examples include the following research questions:
  - Do people with more education have higher income?
  - Do people with lower socio-economic status have a higher likelihood of committing a crime?
  - Do businesses with better corporate governance have higher value?

Relationship between two variables: Correlation

- Consider two quantitative variables, x, y, with n observations each.
- A measure of how two variables move together is the **Pearson Correlation**.

$$r_{x,y} = \frac{\Sigma\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\Sigma\left(x-\bar{x}\right)^2 \Sigma\left(y-\bar{y}\right)^2}}$$

- The **Pearson Correlation** coefficient, $r_{x,y}$, is intimidating!
- Let's re-write $r_{x,y}$ in parts to better understand:
- First, multiply the numerator and denominator by $1/n$:

$$r_{x,y} = \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(y-\bar{y})^2}{n}}}$$

- The numerator of is the sum of the mean deviations in x times the mean deviatons in y called the **covariance**, a measure of how x and y move together:

$$\sigma_{x,y} = \frac{\Sigma\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{n}$$

$$\sigma_{x,y} = \frac{\Sigma\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{n}$$

- The **covariance** is a measure of how x and y move together.
- $X = \{0, 3, 6\} \qquad Y = \{2, -4, -4\}$

Relationship between two variables: Correlation

$$r_{x,y} = \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(y-\bar{y})^2}{n}}}$$

- The denominator is the square root of the product of the variance in x and the variance in y.
- Simplifying, so that $n/n = 1$, gives a simplified expression for correlation:

$$r_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{\text{Covariance}_{x,y}}{(\sigma_x)(\sigma_y)}$$

- So, **Correlation** is the **covariance** of x, y divided by the product of the standard deviation of x and standard deviation of y.

Example

$$r_{x,y} = \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(y-\bar{y})^2}{n}}}$$

- $X = \{3, 4, 5\} \quad Y = \{1, 7, 7\}$
- Calculate the means, standard deviations, covariance, and correlation of X and Y.

Relationship between two variables: Correlation

$$r_{x,y} = \frac{\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(y-\bar{y})^2}{n}}}$$

- It's no accident that our correlation was less than 1.
- To illustrate why, let $Y = X$, so $\bar{Y} = \bar{X}$ and $r_{x,y}$ becomes:

$$r_{x,y} = \frac{\frac{\Sigma(x-\bar{x})(x-\bar{x})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(x-\bar{x})^2}{n}}}$$

- A series has a correlation of 1 with itself.
- So, the maximum correlation between two series is 1.

- Now let Y=-X, so $\bar{Y} = -\bar{X}$ and $r_{x,y}$ becomes:

$$\begin{aligned} r_{x,y} &= \frac{\frac{\Sigma(x-\bar{x})(-x-(-\bar{x}))}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(x-\bar{x})^2}{n}}} \\ r_{x,y} &= \frac{\frac{-\Sigma(x-\bar{x})(x-\bar{x})}{n}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\frac{\Sigma(x-\bar{x})^2}{n}}} \\ r_{x,y} &= -\frac{\sigma_x^2}{\sqrt{\sigma_x^2\sigma_x^2}} \\ r_{x,y} &= -1 \end{aligned}$$

- Correlation can be no less than -1.
- $-1 < r_{x,y} < 1$

- Show the series $Y = \{1, 1, 7, 7\}$ has a correlation of 1 with itself.

Interpreting a Correlation Coefficient

Type of Correlation	Correlation Coefficient	Change in X	Change in Y
Positive	0.00 to 1.00	Increase	Increase
Positive	0.00 to 1.00	Decrease	Decrease
Negative	−1.00 to 0.00	Increase	Decrease
Negative	−1.00 to 0.00	Decrease	Increase

Type	$r_{x,y}$	Example
Positive	0.00 to 1.00	Child Brain Weight and Gestation Period
Positive	0.00 to 1.00	Calories Burned and Time Running
Negative	−1.00 to 0.00	Income and Crime Rate
Negative	−1.00 to 0.00	Altitude and Temperature

Interpreting a Correlation Coefficient

- The correlation coefficient reflects the strength of correlation.
  - A correlation of $+0.4$ is stronger than a correlation of $+0.2$.
  - A correlation of -0.8 is stronger than a correlation of $+0.4$.
- To calculate a correlation coefficient, there must be at least two values (data points, observations) in each series.
- $r_{x,y}$ Is the general notation for the Pearson Correlation coefficient.
- $r_{IQ,Vocabulary}$ is the notation for the correlation between IQ and vocabulary.
- Correlation measures the variability that is shared between two variables.
  - If one variables is constant, so it has no variance then the correlation coefficient will be zero.
  - If you restrict the range of a series, all else the same, the correlation coefficient will be lower.

Interpreting a Correlation Coefficient

Coefficient	*Interpretation*
0.8 to 1.0	Very Strong
0.6 to 0.8	Strong
0.4 to 0.6	Moderate
0.2 to 0.4	Weak
0.0 to 0.2	Weak or None

Source: Salkind, Neil J. Statistics for people who (think they) hate Statistics. Sage Publications, 2017.

Coefficient	*Interpretation*
-0.8 to -1.0	Very Strong
-0.6 to -0.8	Strong
-0.4 to -0.6	Moderate
-0.2 to -0.4	Weak
0.0 to -0.2	Weak or None

Auto-Thefts in the City of Toronto

Year	Reported	Cleared
2014	3619	644
2015	3546	356
2016	3519	445
2017	3730	494
2018	4946	714
2019	5461	581
2020	5820	418

$\sigma_r = 925.9$, $\sigma_c = 119.7$, $\sigma_{r,c} = 21826.4$

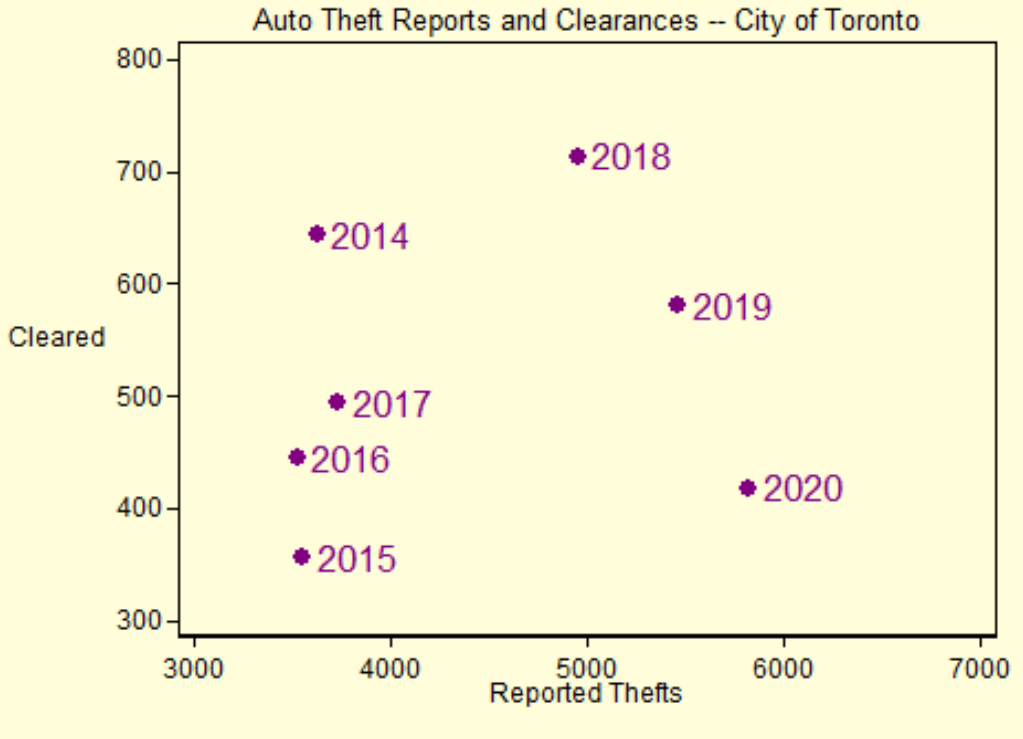
Find the Correlation between Reported and Cleared auto-thefts

Visualizing Correlation with a Scatterplot

Year	Reported	Cleared
2014	3619	644
2015	3546	356
2016	3519	445
2017	3730	494
2018	4946	714
2019	5461	581
2020	5820	418

$\sigma_r = 925.9$, $\sigma_c = 119.7$, $\sigma_{r,c} = 21826.4$

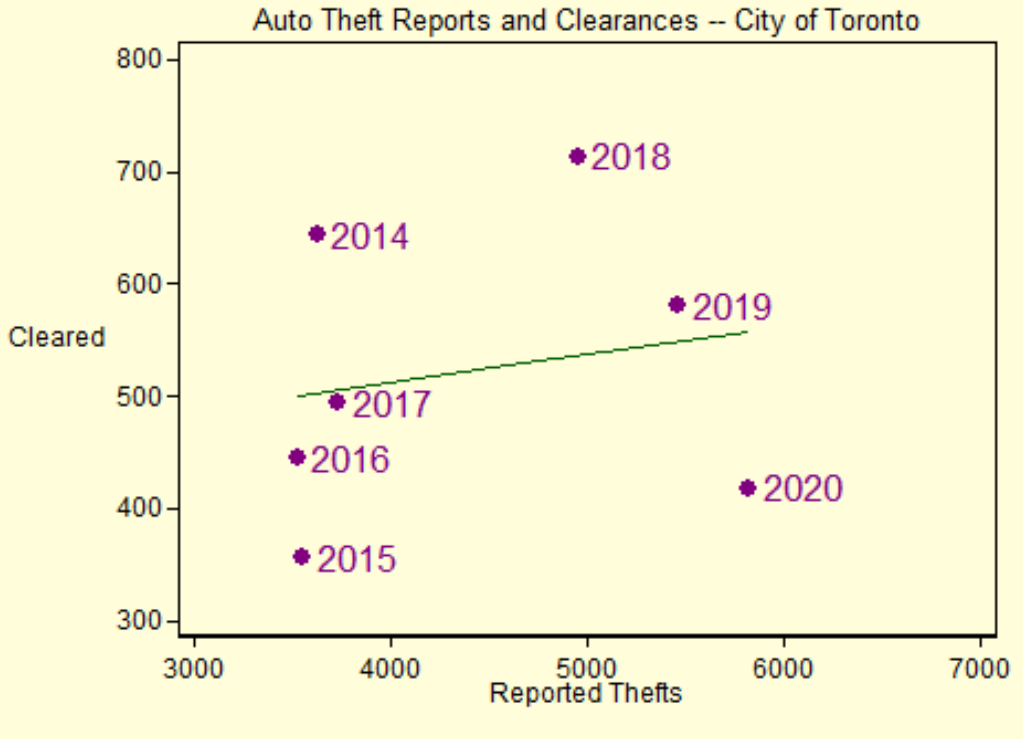
$$r_{r,c} = 0.197$$



Visualizing Correlation with a Scatterplot

Year	Reported	Cleared
2014	3619	644
2015	3546	356
2016	3519	445
2017	3730	494
2018	4946	714
2019	5461	581
2020	5820	418

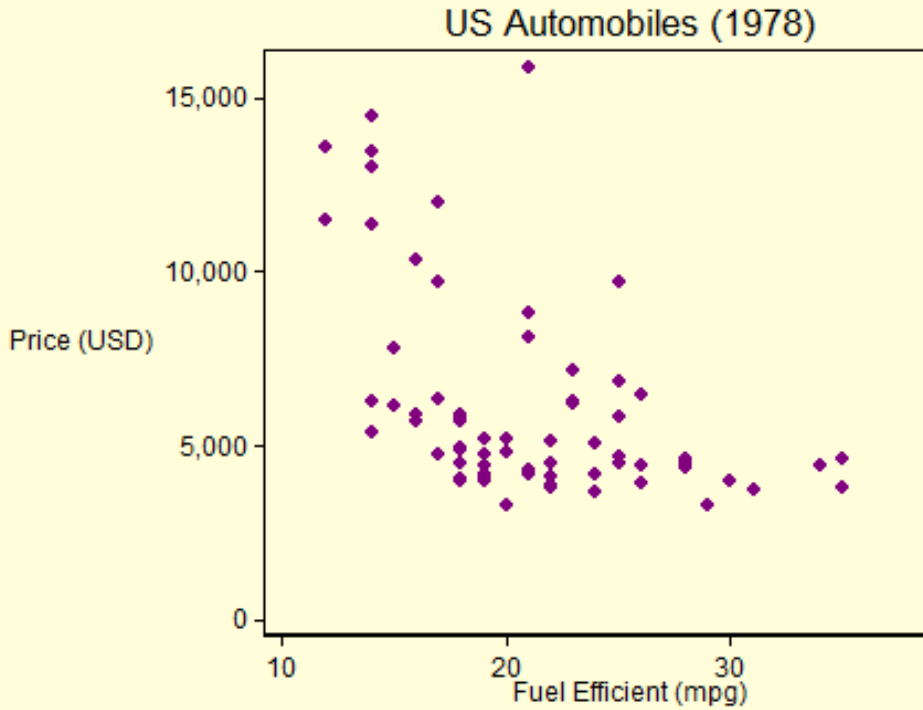
$r_{\text{Reported,Cleared}} = 0.197$



Visualizing Correlation with a Scatterplot

Make	Price	MPG	Weight
Honda Accord	5799	25	2240
Honda Civic	4499	28	1760
Mazda GLC	3995	30	1980
Peugeot 604	12990	14	3420
Renault Le Car	3895	26	1830
and more ...			

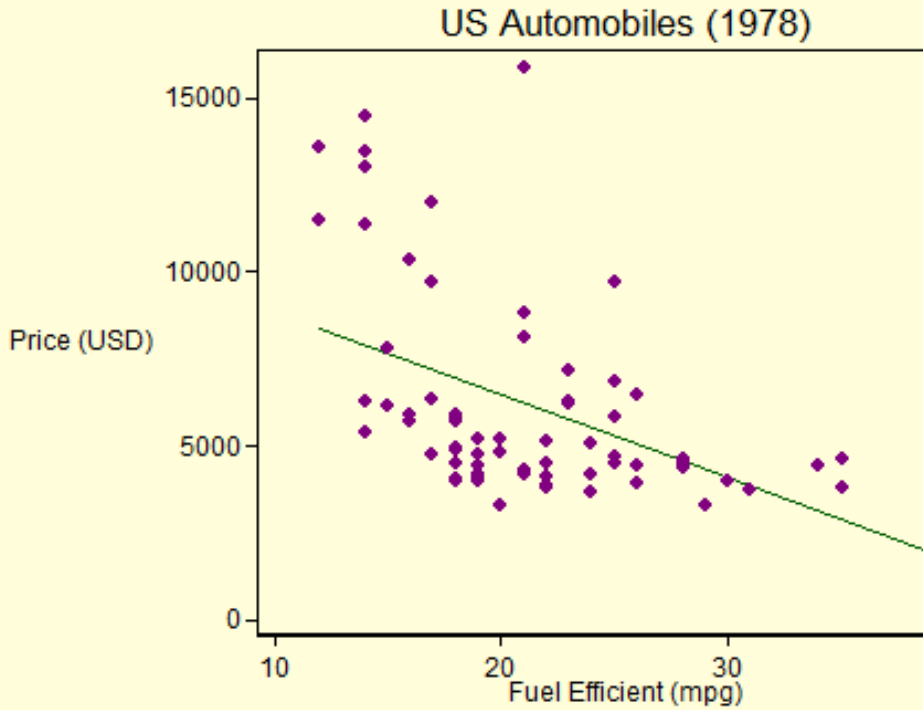
`https://www.fueleconomy.gov/feg/download.shtml`



Visualizing Correlation with a Scatterplot

Make	Price	MPG	Weight
Honda Accord	5799	25	2240
Honda Civic	4499	28	1760
Mazda GLC	3995	30	1980
Peugeot 604	12990	14	3420
Renault Le Car	3895	26	1830
and more ...			

`https://www.fueleconomy.gov/feg/download.shtml`



Correlation Matrix

	Price	MPG	Trunk	Weight
price	1			
mpg	-0.47	1		
trunk	0.31	-0.58	1	
weight	0.54	-0.81	0.67	1

- Correlation matrix is used to present correlations between more than two series.

Correlation of Determination

- The **Correlation of Determination** is the percentage of variance in one variable that is explained by the variance in another variable.
- The correlation of Determination is the square of the correlation coefficient, r^2.

Correlation Matrix

	Price	MPG	Trunk	Weight
price	1			
mpg	-0.47	1		
trunk	0.31	-0.58	1	
weight	0.54	-0.81	0.67	1

Correlation of Determination

	price	mpg	trunk	weight
price	1			
mpg	0.22	1		
trunk	0.10	0.34	1	
weight	0.29	0.65	0.45	1

When To Use Correlation?

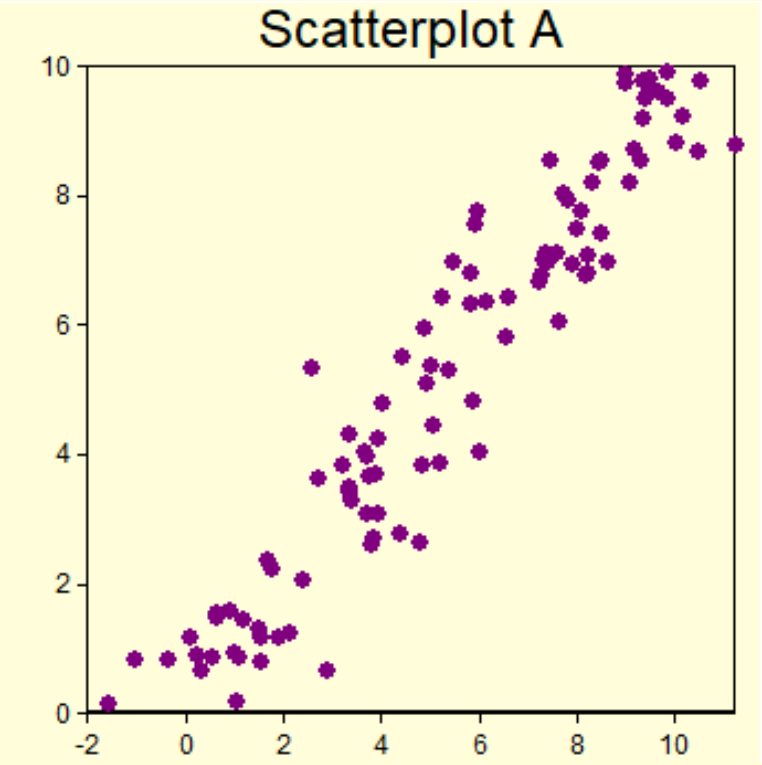
Correlation is a **linear measure** of the strength of a relationship between two **quantitative** variables.

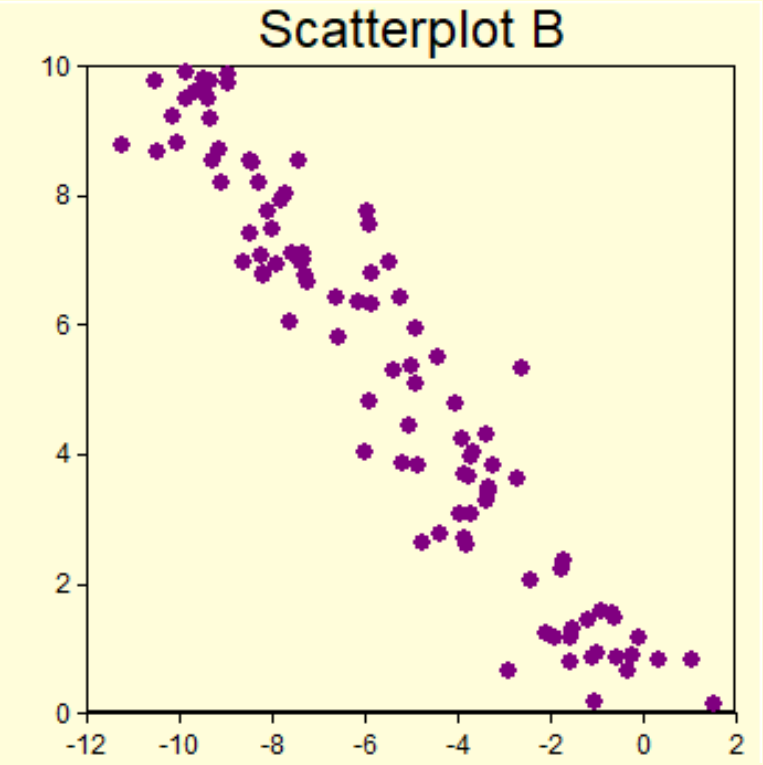
Consequently, correlation must satisfy **three conditions** to be accurate:

1. Do not use the correlation coefficient to measure the relationship between categorical variables.
2. Make sure the relationship between the quantitative variables is linear or "straight."
3. Avoid outliers – outliers can artificially increase the correlation coefficient. This can make a weak relationship look strong.

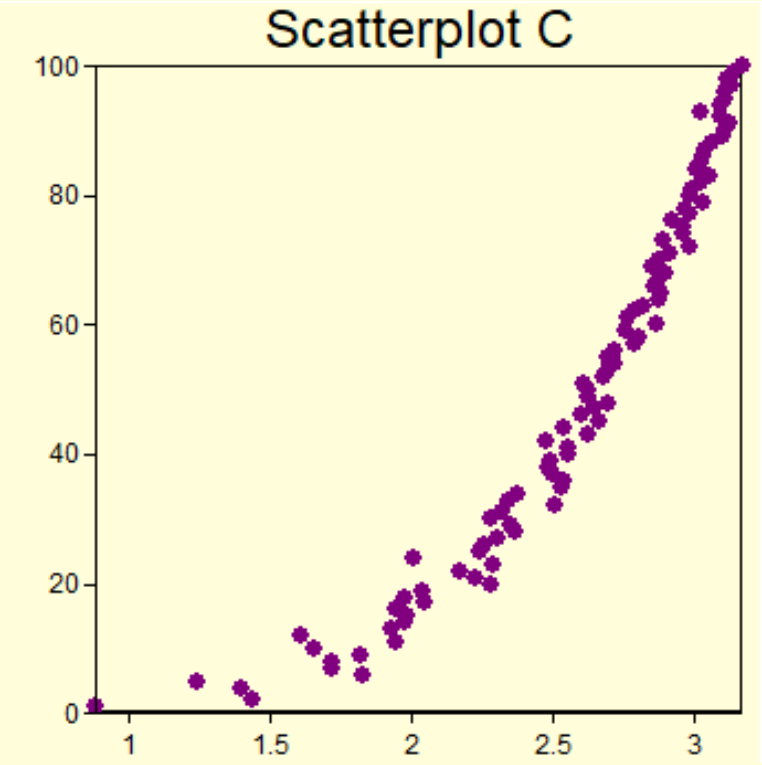
You can check the above conditions with a scatterplot diagram.

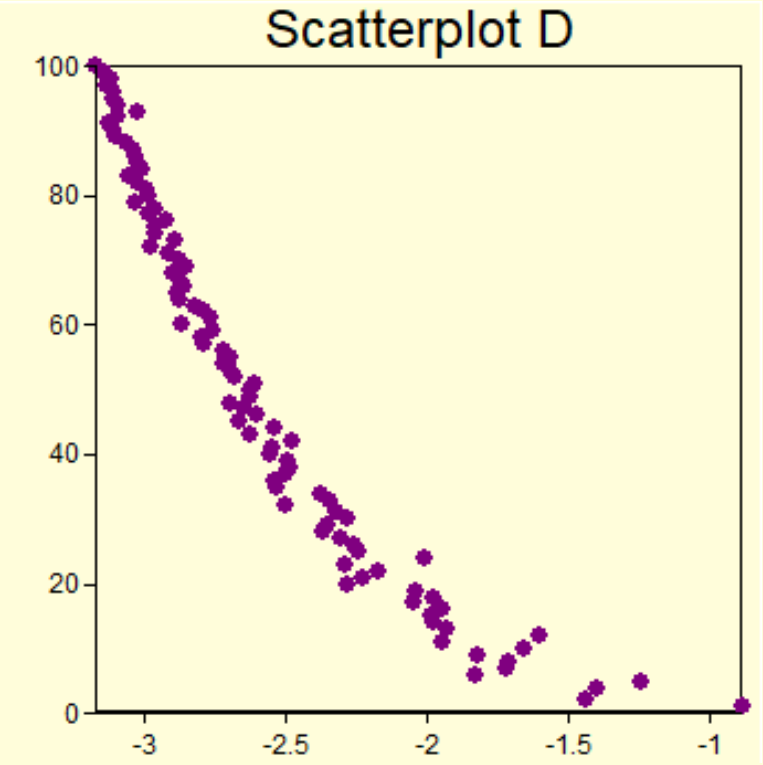
Are the Correlation Conditions Satisfied?



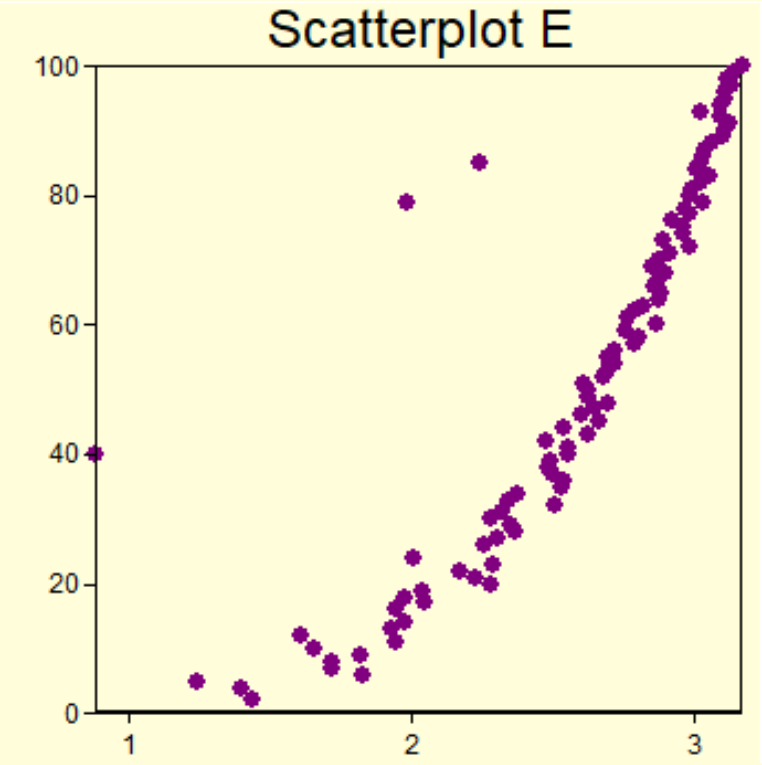


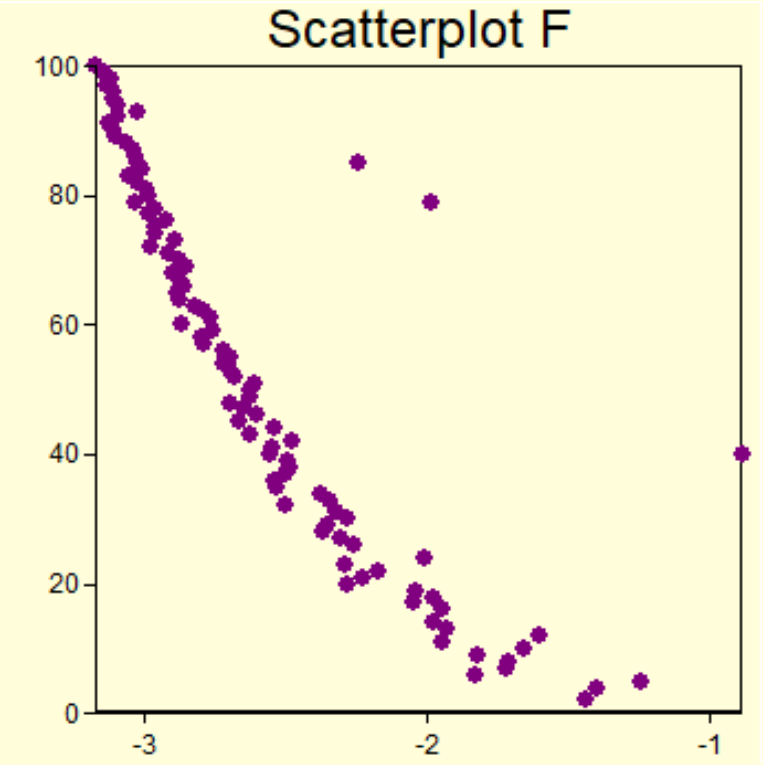
Are the Correlation Conditions Satisfied?



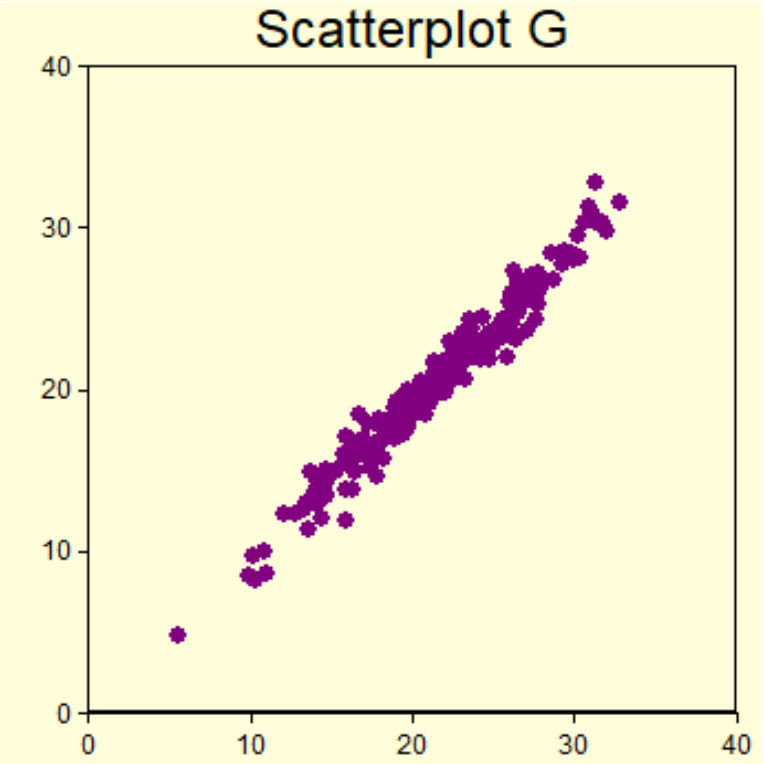


Are the Correlation Conditions Satisfied?

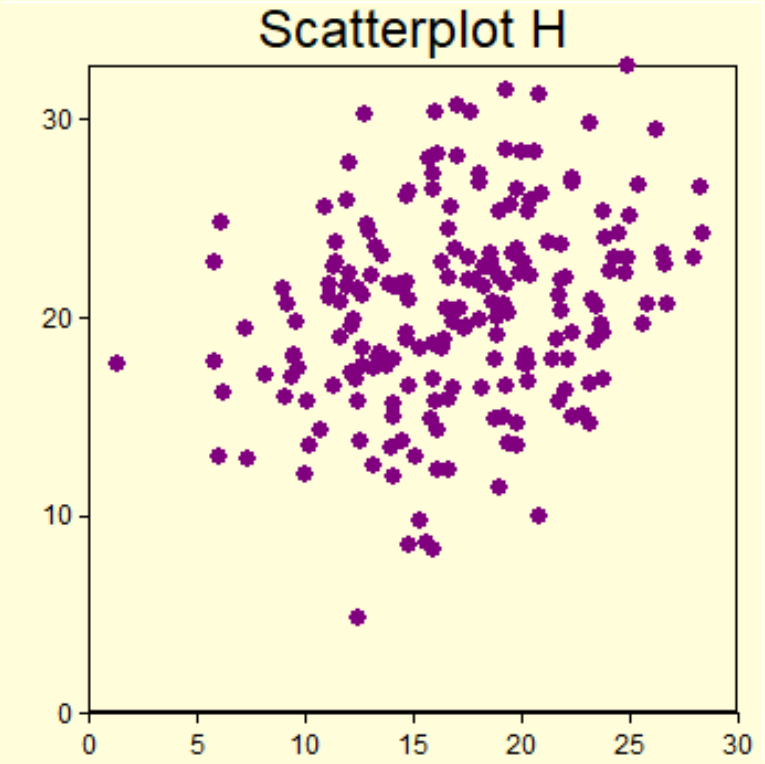




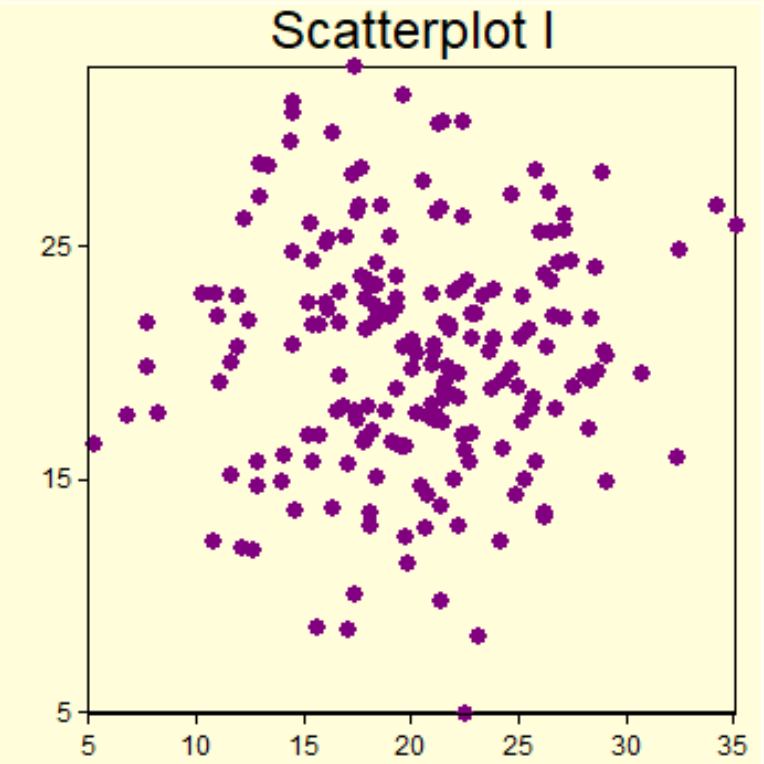
Strong v. Weak Correlation



Correlation $= 0.98$, $r^2 = 0.96$



Correlation $= 0.26$, $r^2 = 0.07$



Correlation $= 0.0421$, $r^2 = 0.00$

Properties of Correlation (r)

- Correlation is always $\{-1, +1\}$.
- The correlation of X with Y is also the correlation of Y with X.
- The sign of the correlation coefficient determines the type of relationship.
- Correlation ignores units: i.e. converting celsius to fahrenheit or meters to yards does not change the correlation coefficient, so always *present correlation coefficients without units*.
- Do not confuse correlation with causation.
- Correlation can be misleading if the two quantitative variables have a non-linear relationship.
- Correlation is sensitive to outliers – outliers could reduce the correlation coefficient even when the relationship is strong.
- Or outliers could increase the coefficient even when the relationship is weak.
- The correlation of determination, r^2, measures the variability in one variable explained by the variability of another variable.

Making Predictions with Simple Linear Regression

- Simple **Linear Regression** predicts a y-value given a x-value using a linear model: $y = mx + b$
- We can re-write the above as:

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

- where $\hat{y}$ is the predicted value of y.
- b_0 is the intercept.
- b_1 is the slope.
- b are known as the **coefficients**.
- The **Least Squares** method uses the x, y values to compute these coefficients and generate the **regression line** or **line of best fit**.
- A common terminology is to refer to the y series as the **dependent** variable and the x series as the **independent** variable.

Least Squares Method: The Slope

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

- The **Least Squares** method uses the standard deviations and correlation (or covariance) of x, y to compute the slope coefficient, b_1:

$$b_1 = r_{x,y} \frac{s_y}{s_x}$$

- where r is the correlation between x and y.
- s_y is the sample standard deviation of y.
- s_x is the sample standard deviation of x.
- b_1 is the ratio of y per unit of x.
- An alternate formula replaces covariance with correlation which simplifies to:

$$b_1 = r \frac{s_{y,x}}{s_x}$$

- The intercept is the difference.

Least Squares Method: The Intercept

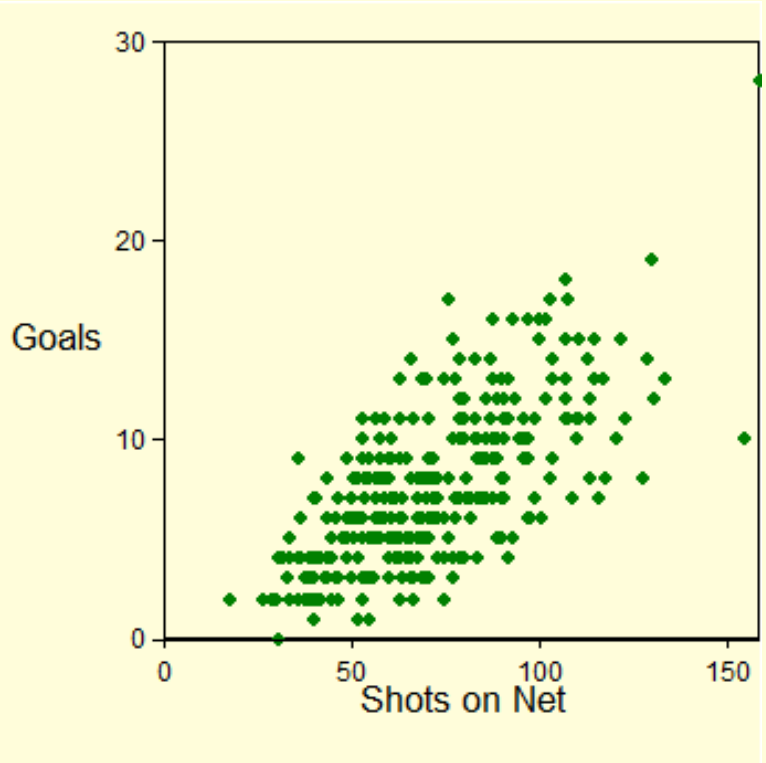
$$\hat{y} = b_0 + b_1 x$$

- The **Least Squares** method uses the means of x, y to compute the intercept coefficient: b_1:

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

- where $\bar{y}$ is mean of y, and
- $\bar{x}$ is the mean of x.

Linear Regression – NHL Shots and Goals

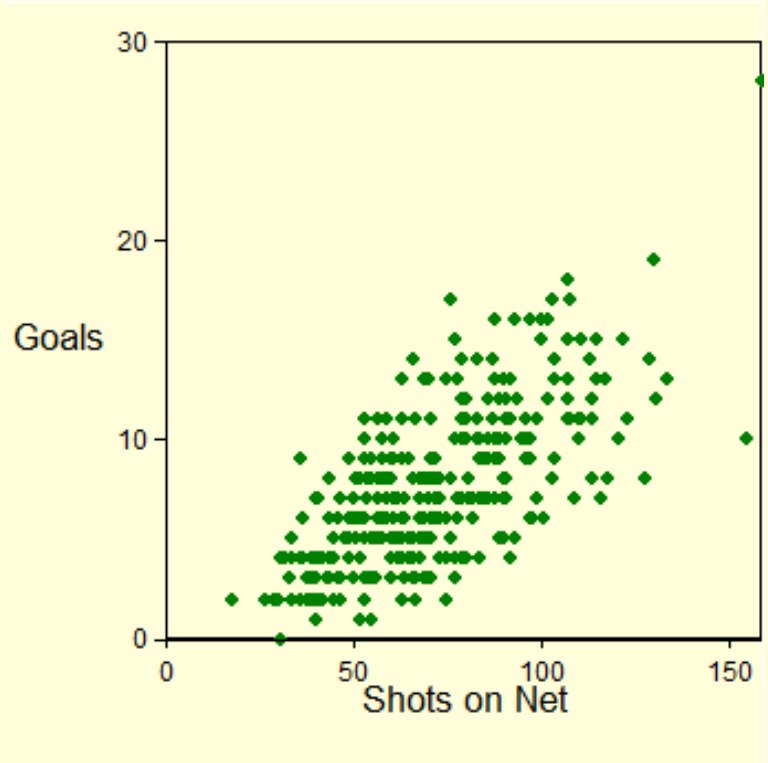


Source: MoneyPuck.com – Forwards, 40 games or more, 5-on-5 play, 2020-21

To be accurate, linear regression requires:

1. The data should be quantitative.
2. The relationship should be linear.
3. There are no outliers.

Linear Regression – NHL Shots and Goals



Source: MoneyPuck.com – Forwards, 40 games or more, 5-on-5 play, 2020-21

	Goals	Shots
Mean	7.59	72
Sample Std. Dev.	3.97	24.06
Correlation	0.675	0.675

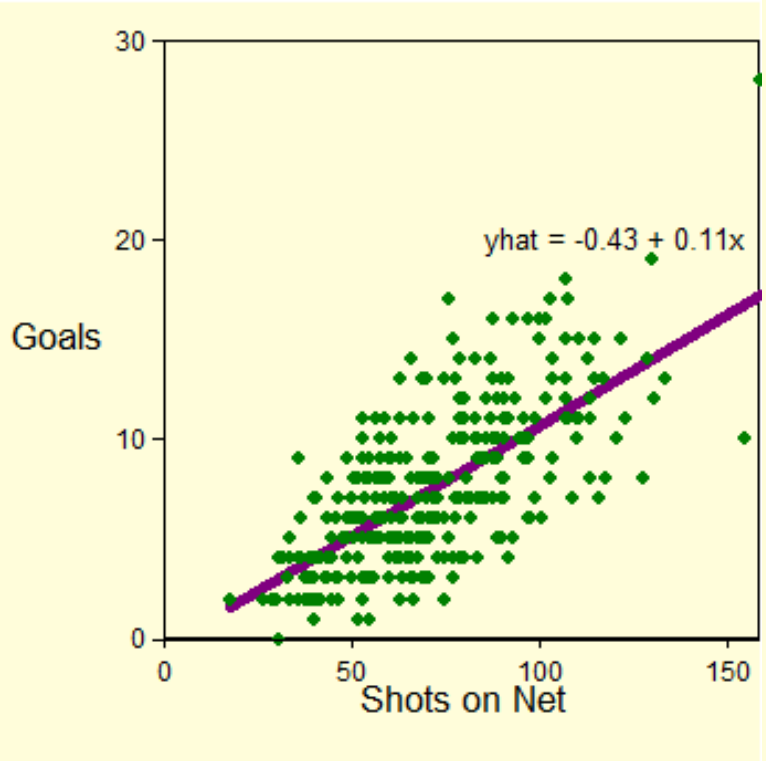
1. Use a simple linear regression to predict the number of goals a NHL forward would score with 75 shots on net.

Linear Regression – NHL Shots and Goals

	Goals	Shots
Mean	7.59	72
Sample Std. Dev.	3.97	24.06
Correlation	0.675	0.675

1. Use a simple linear regression to predict the number of goals a NHL forward would score with 75 shots on net.

NHL Shots and Goals



Source: MoneyPuck.com – Forwards, 40 games or more, 5-on-5 play, 2020-21

$$\hat{y} = -0.43 + 0.11x$$

- The model predicts that a player who shoots more also scores more.

"You miss one hundred percent of the shots you don't take"
– Wayne Gretzky

Linear Regression – NHL Shots and Goals

	Goals	Shots
Mean	7.59	72
Sample Std. Dev.	3.97	24.06
Correlation	0.675	0.675

$$\hat{y} = -0.43 + 0.11x$$

2. How much of the variability in goals scored is explained by the variability in shots taken?

Linear Regression – NHL Shots and Goals

	Goals	Shots
Mean	7.59	72
Sample Std. Dev.	3.97	24.06
Correlation	0.675	0.675

$$\hat{y} = -0.43 + 0.11x$$

3. For each additional shot taken, what is the predicted change in goals scored?

Residuals – a Measure of Error in the Model

- Every model is imperfect. We call this error a **residual**.

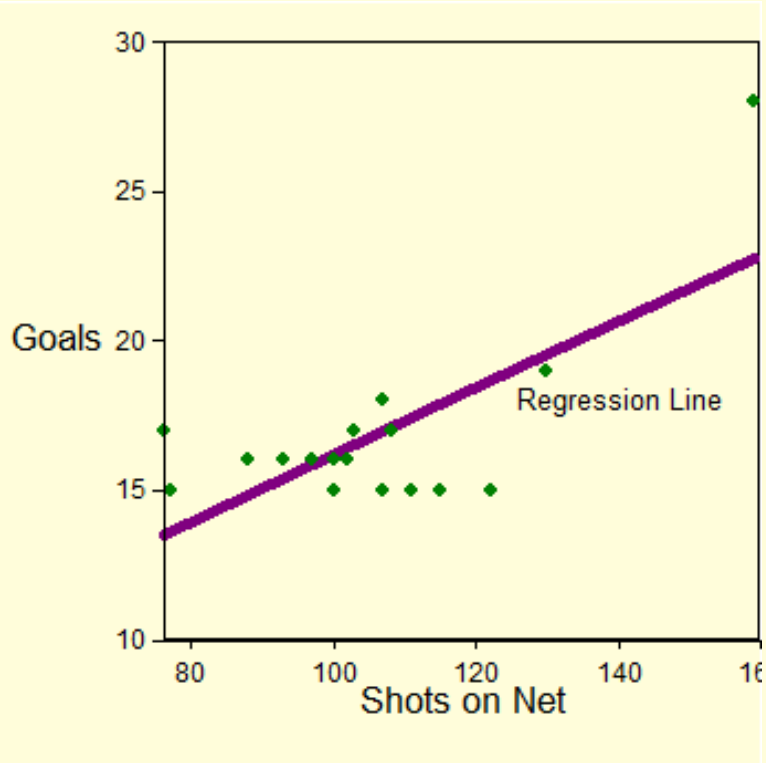
$$\text{Data} = \text{Model} + \text{Residual}$$

- The residuals, e, are the difference between the predicted values and the data

$$\begin{aligned} \text{Residual} &= \text{Data} - \text{Predicted Value} \\ e &= y - \hat{y} \\ e &= y - b_0 - b_1 x \end{aligned}$$

- We can use these residuals, e, to measure the accuracy of the model.

Residuals – a Measure of Error in the Model



Source: MoneyPuck.com – Forwards, 40-plus games, 5-on-5 play with 15 or more goals, 2020-21

$$\begin{aligned} \text{Residual} &= \text{Data} - \text{Predicted Value} \\ e &= y - \hat{y} \\ e &= y - b_0 - b_1 x \end{aligned}$$

- The **regression line** represents the model's predicted value, $\hat{y}$, for a given x.
- The dots on the diagram represent the data.

Residuals – a Measure of Error in the Model

	Goals	Shots
Mean	7.59	72
Sample Std. Dev.	3.97	24.06
Correlation	0.675	0.675

$$\hat{y} = -0.43 + 0.11x$$

A) If a player takes 120 shots, predict the number of goals.

B) Suppose one player had 120 shots and 8 goals – calculate the residual from part (A).

C) Did the model over-estimate or under-estimate the number of goals?

Residuals – a Measure of Error in the Model

- **Negative residuals** imply the actual value is less than the predicted value – an **over-estimation**.
- **Positive residuals** imply the actual value is larger than the predicted value – an **under-estimation**.
- The regression line gives us information about **average values**.
- The accuracy of the model prediction can be measured by the residuals.

- The sum of the residuals is equal to zero:

$$\sum (y - \hat{y}) = 0$$

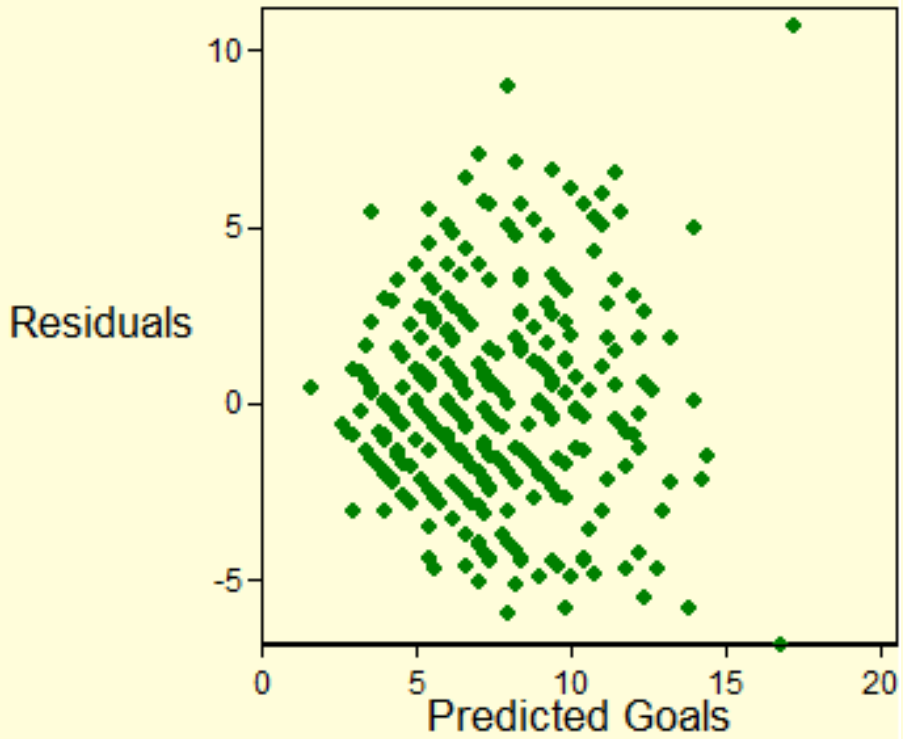
- A measure of a model's accuracy is the **sum of squared-residuals**, *SSR*.

$$\text{SSR} = \sum (y - \hat{y})^2$$

- The regression line or **line of best fit** is also called the **least squares line** because it minimizes the sum of squared-residuals.

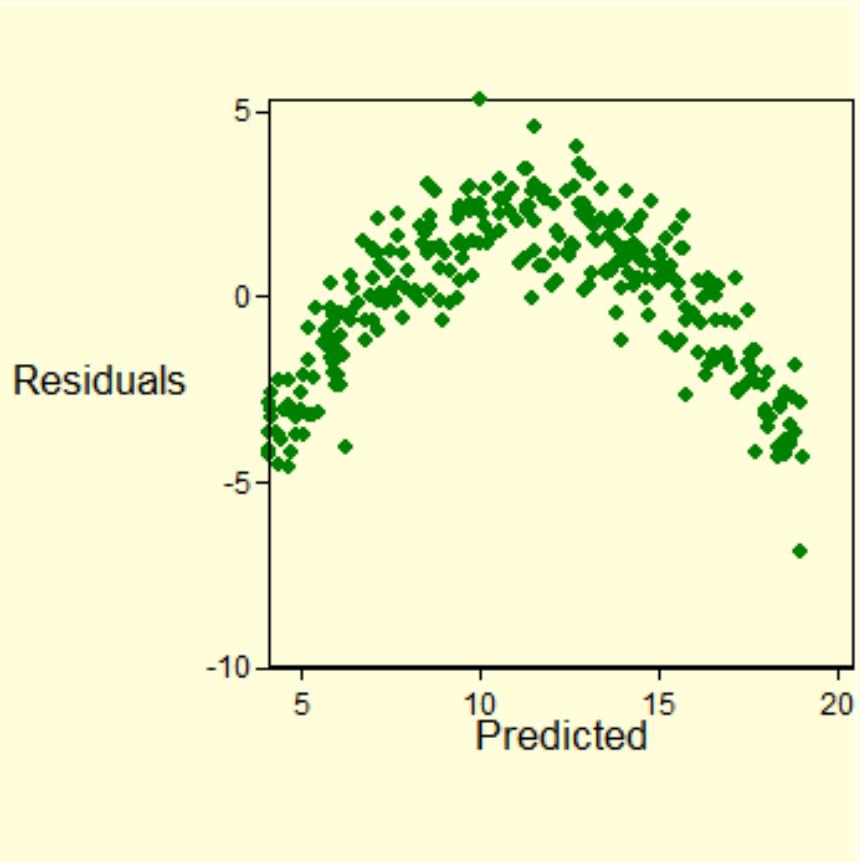
Residuals – a Measure of Error in the Model

- An accurate model should not contain any useful information in the residuals.
- A scatterplot of the *residuals*, e, and the *predicted values*, $\hat{y}$, can help verify that the residuals have no useful information.
- On a scatterplot of the residuals and the predicted values, we should not see any trends, patterns, shapes, or outliers.



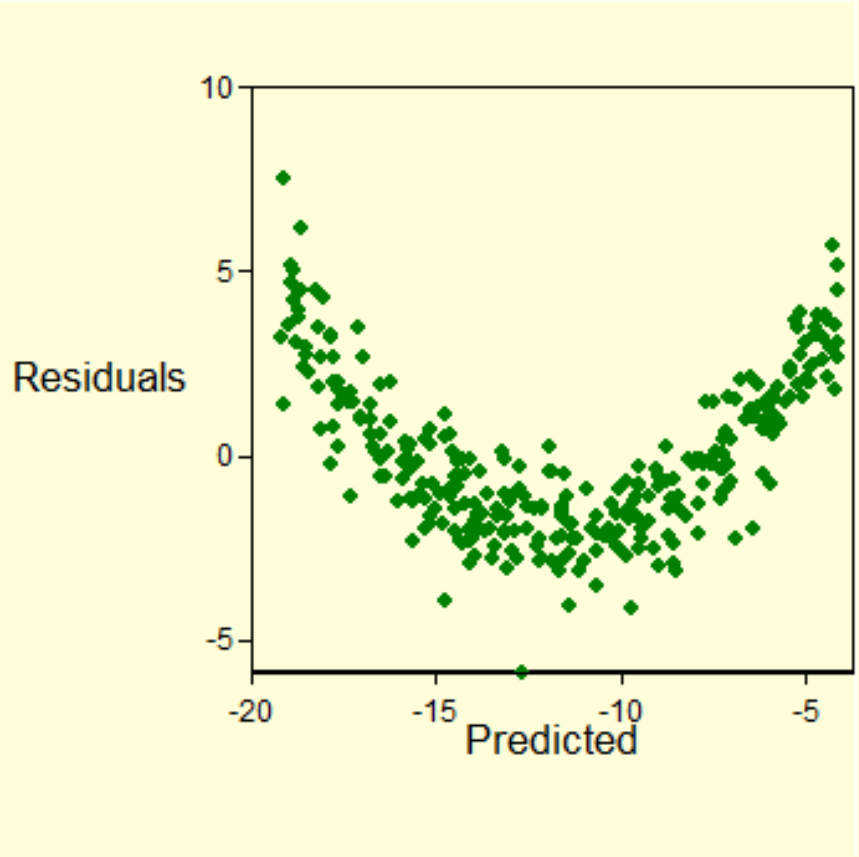
Scatterplot of the residuals from our model of NHL shots and goals

Residuals – A nonlinear relationship



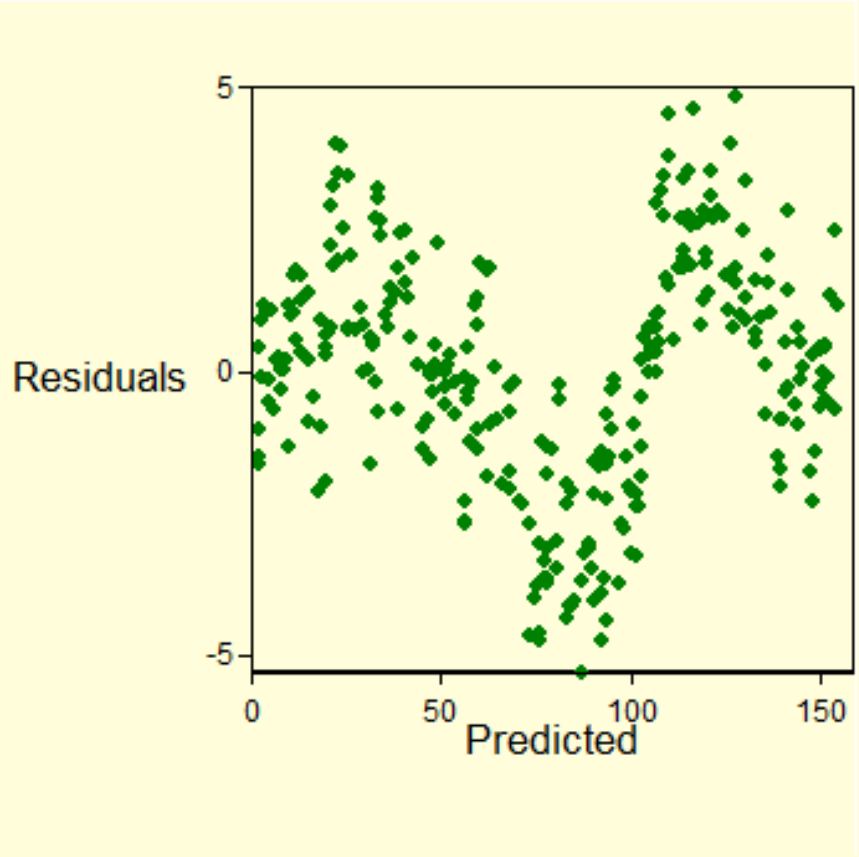
- These residuals have an interesting shape.
- The residuals do not appear in a horizontal line/cloud.
- In this case, the relationship between the two variables was not linear – hence the curve-shape in the scatterplot.

Residuals – A nonlinear relationship



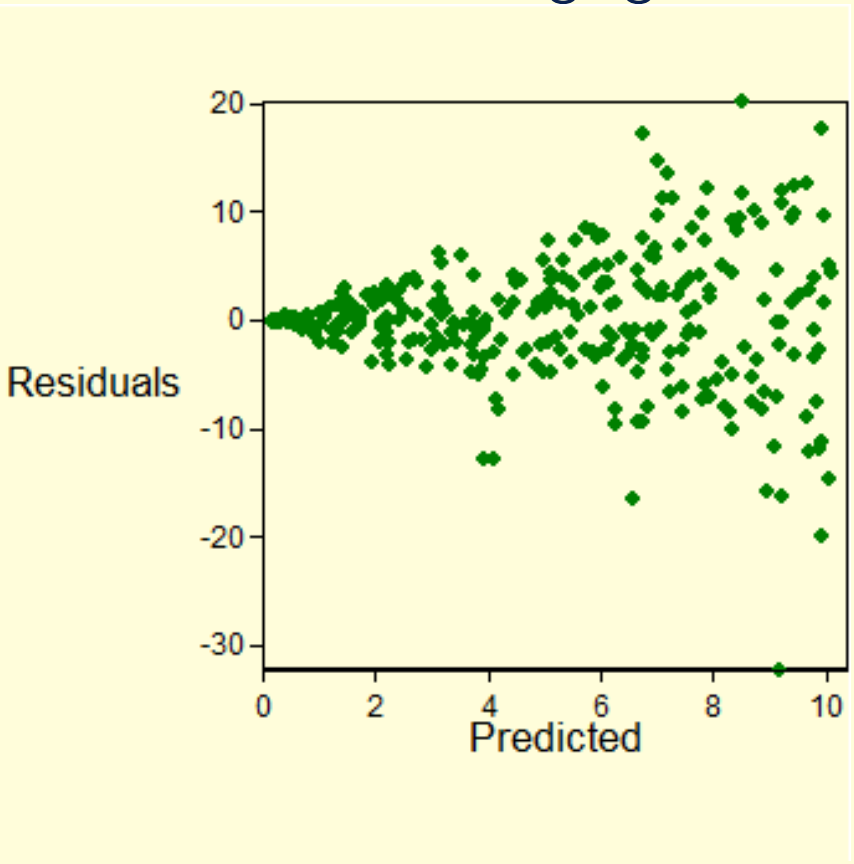
- These residuals have an interesting shape.
- The residuals do not appear in a horizontal line/cloud.
- In this case, the relationship between the two variables was not linear – hence the curve-shape in the scatterplot.

Residuals – A nonlinear relationship



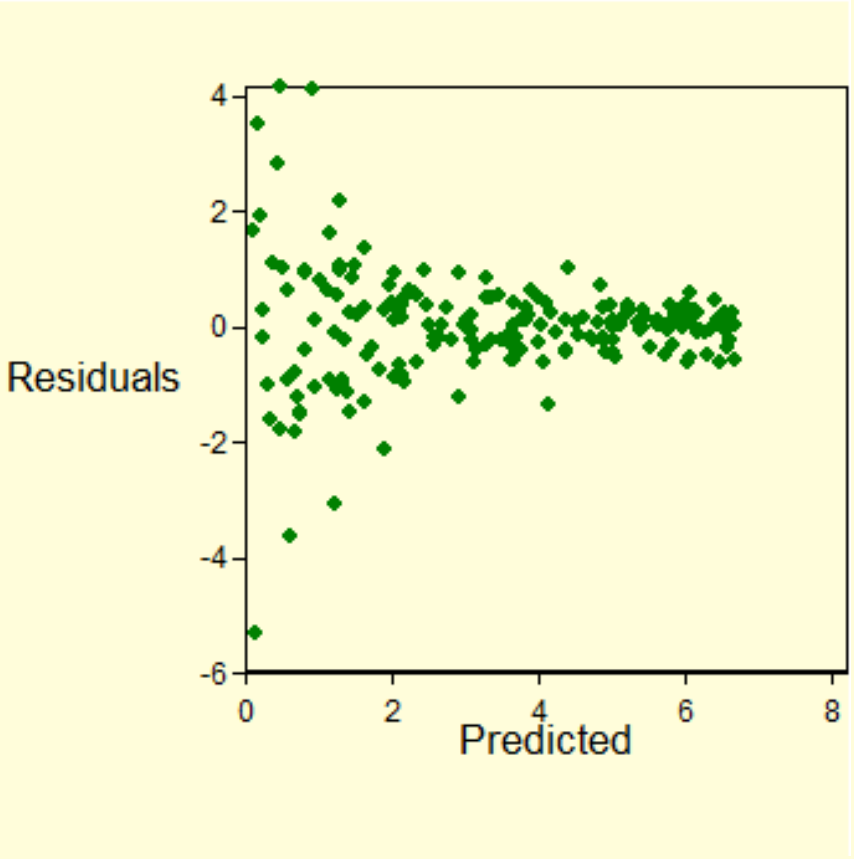
- These residuals have an interesting shape.
- The residuals do not appear in a horizontal line/cloud.
- In this case, the relationship between the two variables was not linear.

Residuals – Changing Variance



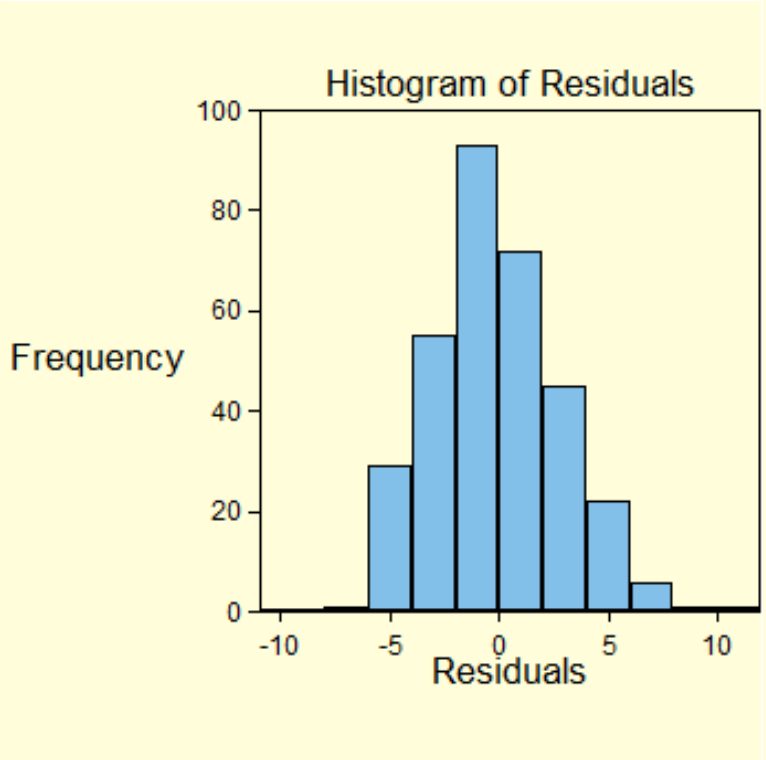
- The residuals appear to become **larger** as the predicted value increases.
- The variance (variability) of the residuals is not constant – a new problem.

Residuals – Changing Variance



- The residuals appear to become **smaller** as the predicted value increases.
- The variance (variability) of the residuals is not constant.

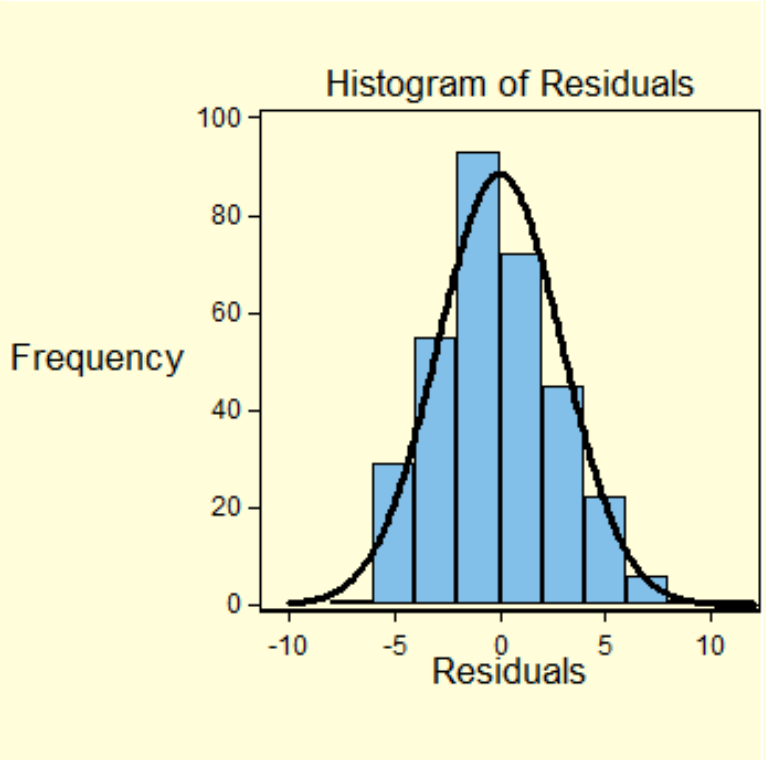
Residuals – a Measure of Error in the Model



Scatterplot of the residuals from our model of NHL shots and goals

- Similarly, a histogram of the residuals should have the appearance of a bell curve:
  - Many residuals clustered around zero.
  - A gradual decline in frequency as the value of the residuals moves further from zero.

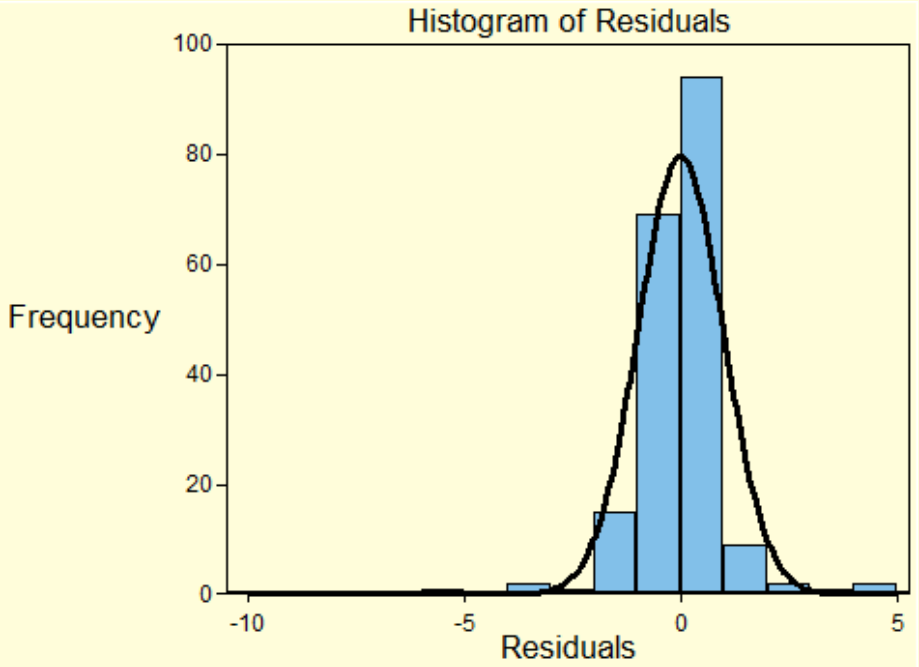
Residuals – a Measure of Error in the Model



Scatterplot of the residuals from our model of NHL shots and goals

- Similarly, a histogram of the residuals should have the appearance of a bell curve:
  - Many residuals clustered around zero.
  - A gradual decline in frequency as the value of the residuals moves further from zero.
  - Here, the black curve represents a normal distribution.

Residuals – a Measure of Error in the Model



Scatterplot of the residuals from our model of NHL shots and goals

- This histogram of residuals may have some problems.
- There seem to be more observations around 0 than expected.
- There are some large, negative values.

A Summary of Linear Regression

- Only use simple linear regression when the underlying relationship between variables is linear.
- Identify outliers and investigate why they are unusual.
- Check the residuals for:
  - unusual patterns and shapes
  - changing variability.
- A higher R^2 is not helpful unless the above conditions have been satisfied.