I. Examining Residuals:

If you find a pattern in the Residual Plot that means the residuals, (errors) are predictable. If the residuals are predictable, then a better model exists. ----If you find a pattern or bend in the residual plot it means THE LINEAR MODEL IS NOT APPROPRIATE.

Note: The sum of residual is zero.

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

A residual plot constructed with the RESIDUALS on the y-axis. On the x-axis, put the explanatory variable.

NOTE: Some software packages will put $\hat{y}$ on the x-axis. This does not change the presence of (or lack of) of a pattern.

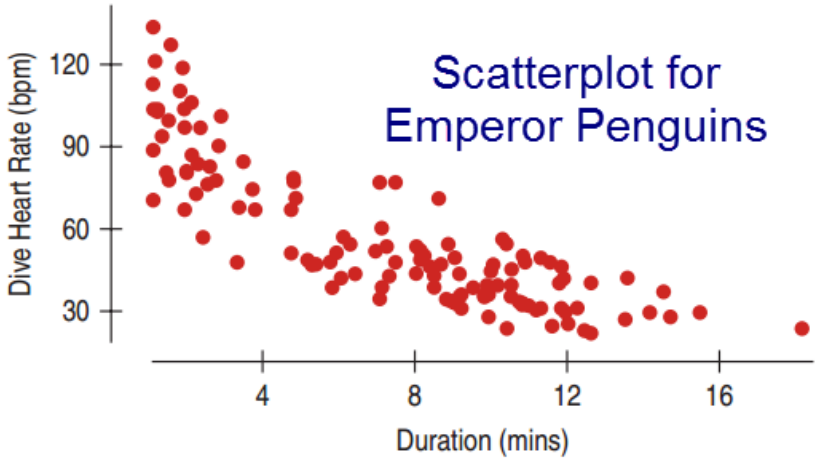
If Residual shows a no pattern or bend, it indicates that a linear model is a good fit for the data.

Predicted vs Residual

If Residual shows a pattern, it indicated a non-linear model is a good fit for the data.

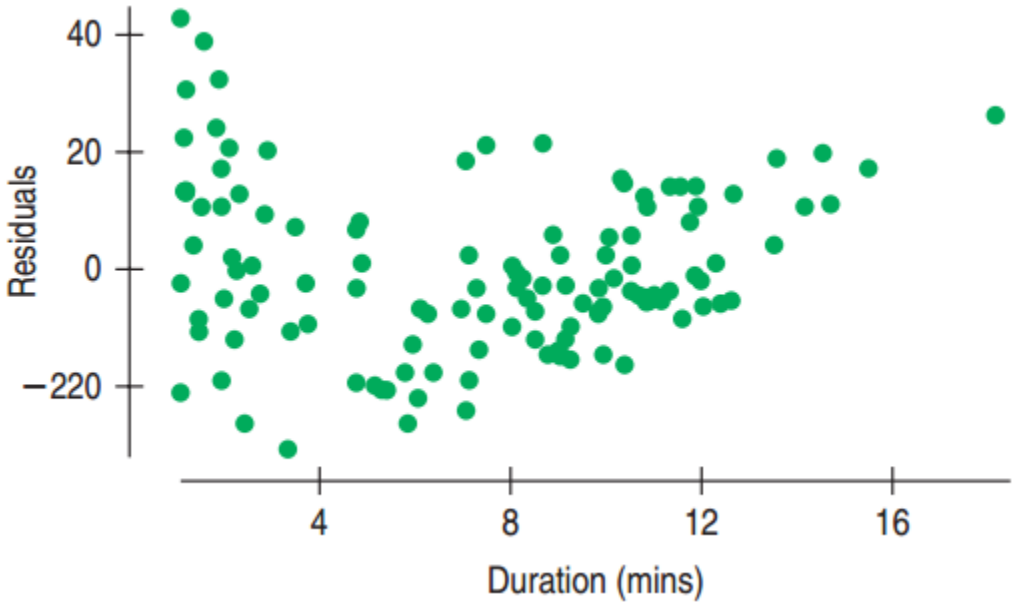
Example: We would like to analyze the relationship between the amount of time (in minutes) a penguin is under water with its heart rate (bpm).

- **What Scatter Plot tells us?**



- R^2 = 71.5%
- Moderately strong negative association
- $\hat{y} = 96.9 - 5.47\,x$
- On average, for each minute the penguin is under water, its heart rate declines by 5.47 bpm.
- The predicted heart rate before the dive is 96.9 bpm.

- **What Residual Plot tells us?**

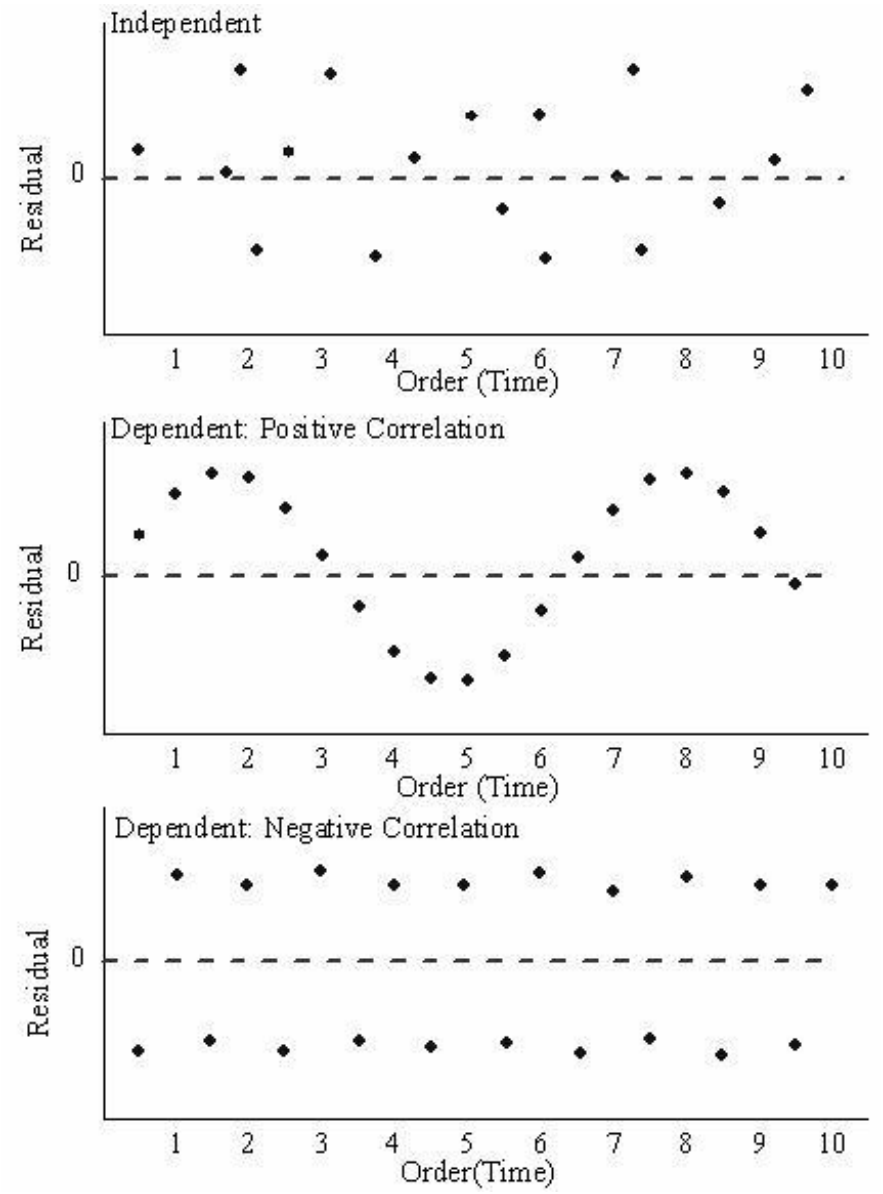


- There is a clear bend in the residual plot.
- It is more spread out for low durations and less for high durations.
- The residual plot suggests a re-expression may be needed to straighten out the data.

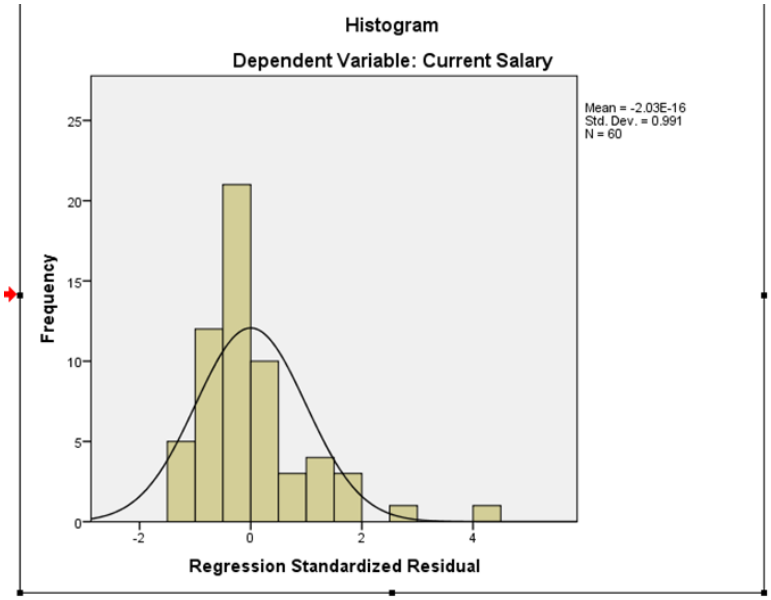
The four assumptions of linear Regression:(http://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regression/)

- Linearity: Draw scatterplot to check linearity

- Independence of Residual: Fluctuating patterns around zero will indicate that the error term is dependent.(https://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis)

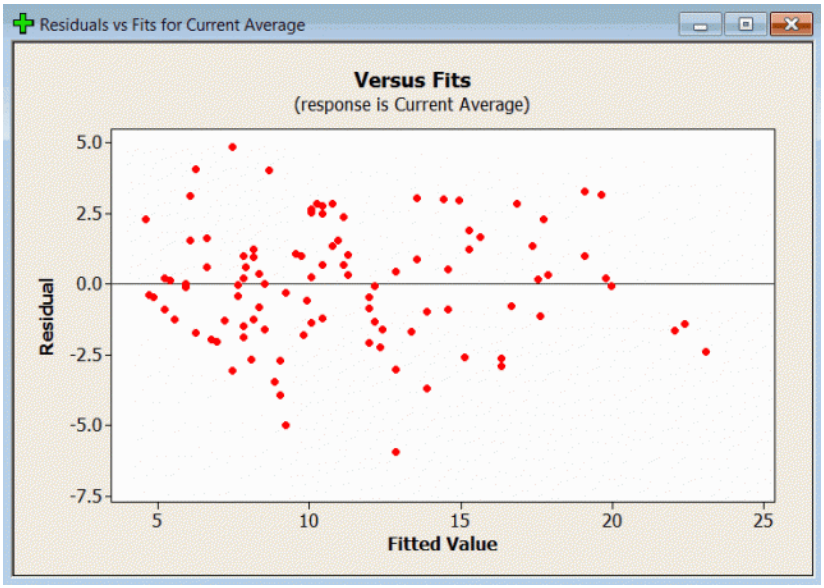


- Normality: Draw histogram of residuals to check the normality of residuals.



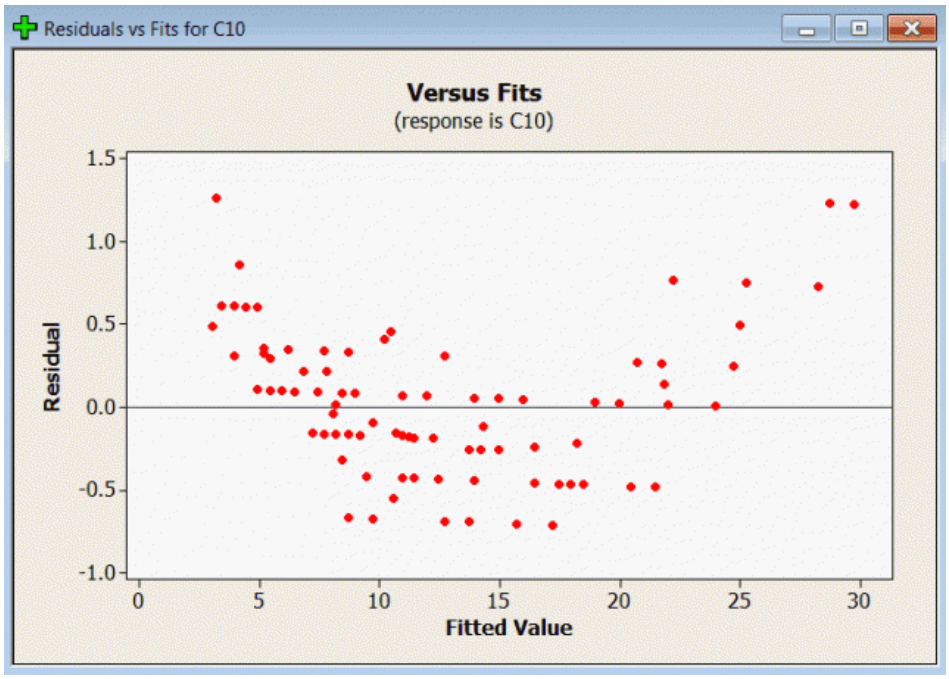
- **Homoscedasticity of residuals** or **Equality of variance:(**https://blog.minitab.com/blog/the-statistics-game/checking-the-assumption-of-constant-variance-in-regression-analyses)

In the below plot the errors have constant variance, with the residuals scattered randomly around zero. If, for example, the residuals increase or decrease with the fitted values in a pattern, the errors may not have constant variance.



Consider the next graph:

There is definitely a noticeable pattern here! The residuals (error terms) take on positive values with small or large fitted values, and negative values in the middle. The width of the scatter seems consistent, but the points are not randomly scattered around the zero line from left to right. This graph tells us we should not use the regression model that produced these results.



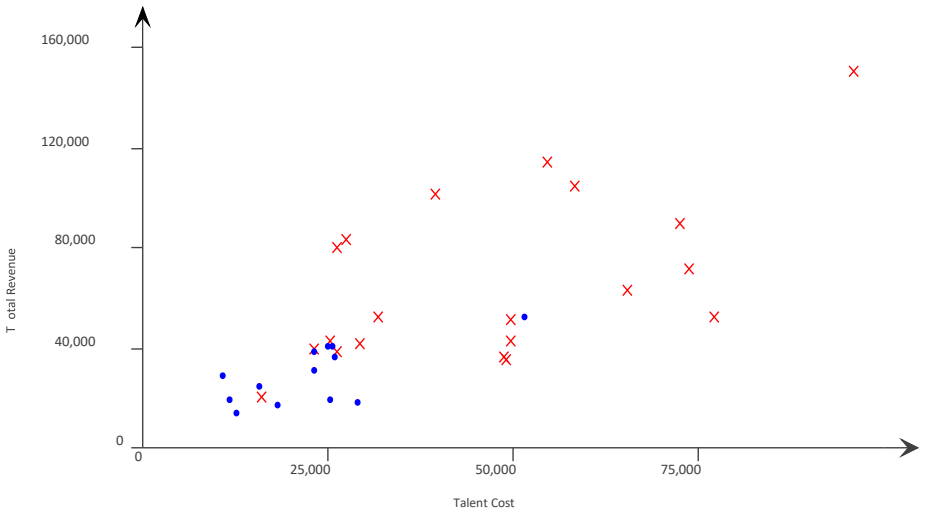
- To look for outliers, and to check the Equal Variance Assumption, a _Residual Plot_ should be created.
- Not all outliers have large residuals.

Example:

A student wants to study how gas prices in her city have changed in the last five years. She finds a website that reports the average price per gallon each month. How will the strength of this relationship differ from the relationship if she had the price of gas at each gas station each day over the same time period? The strength of the relationship will be higher for average rather than individual.

Example:

A concert production company examined its records. The manager made the following scatterplot. The company places concerts in two venues, a smaller, more intimate theater (plotted with blue circles) and a larger auditorium-style venue (red x's).



a) Describe the relationship between talent cost and total revenue. (Remember: direction, form, strength, outliers.)The scatterplot shows a strong positive linear relationship between talent cost and total revenue. There is 1 outlier that stands apart from the majority of the data.

b) How are the results for the two venues similar?

Both venues show an increase of revenue with talent cost.

c) How are they different?

The larger venue has greater variability. Revenue for that venue is more difficult to predict.

II. Interpolation and Extrapolation:

- **Interpolation:** Interpolation is the method to find the predicted value within the range of the explanatory variable.

For instance, let y represents the stock price and x represents years since 1970. Also, let assume the data collected from 1970 to 1982. Let assume the regression line for this problem is:

$$\hat{y} = 12.79 + 6.75x$$

The predicted price for year 1980 ($x = 10$) is:

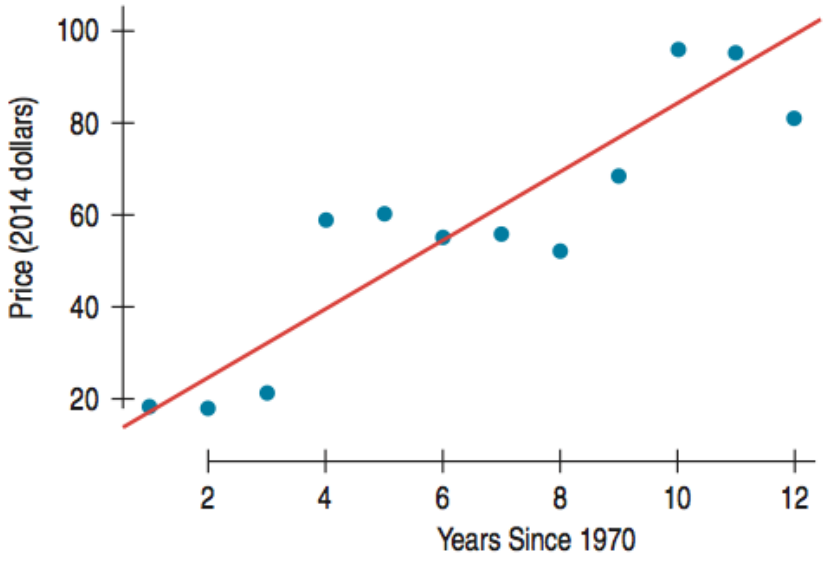
$$\hat{y} = 12.79 + 6.75(10) = \$19.29$$

Please note that year 1980, $x = 10$, is within the range of the explanatory variable (1970 to 1982)

- **Extrapolation:** Reaching Beyond the Data.

Extrapolation is the method to find the predicted value outside of the range of the explanatory variable. For instance, If we want to predict the price for year 1995 (x =25) since we are reaching beyond the data this action called Extrapolation. Please note that year 1995, $x = 25$, is outside of the range of the explanatory variable (1970 to 1982).

If we look at the model,$\hat{y} = 12.79 + 6.75x$, we see as years goes up the price goes up (a positive relation). The linear model $\hat{y} = 12.79 + 6.75x$ may not be a good model to predict the price in year 1995. Why? Because we do not know if the market has the same behavior after the year 1982. After the year 1982, the price and year may have negative relationships that mean as the year goes up the price decreases.



Using a linear model to predict a value of y for an x-value far from the ones used to find the model is called ___Extrapolating_____.

Example:

Students at a large state university system are upset over the rate at which their fees have increased in the last five years (2005-2010). A small group present before the state legislature and report a predicted fee for 2020 based on their model. What error could they be accused of?

Extrapolating

Example:

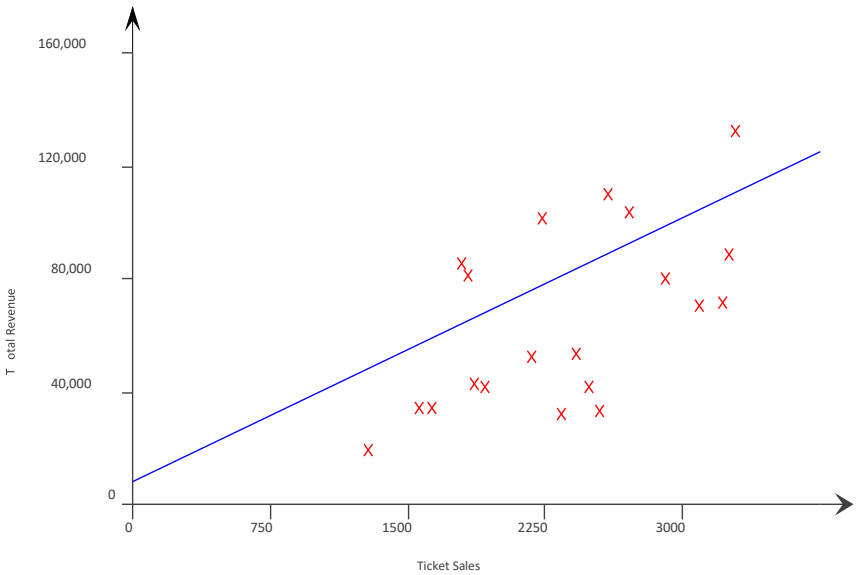
Noting a recent study predicting the increase in cell phone costs, a friend remarks that by the time he's a grandfather, no one will be able to afford a cell phone. Explain where his thinking went awry.

He is extrapolating into the future. It is impossible to know if a trend like this will continue so far into the future.

Example:

A regression of total revenue on ticket sales determined by a concert production company is given below.

$$\widehat{Revenue} = -9.225 + 32.23 \cdot Ticket\ Sales$$



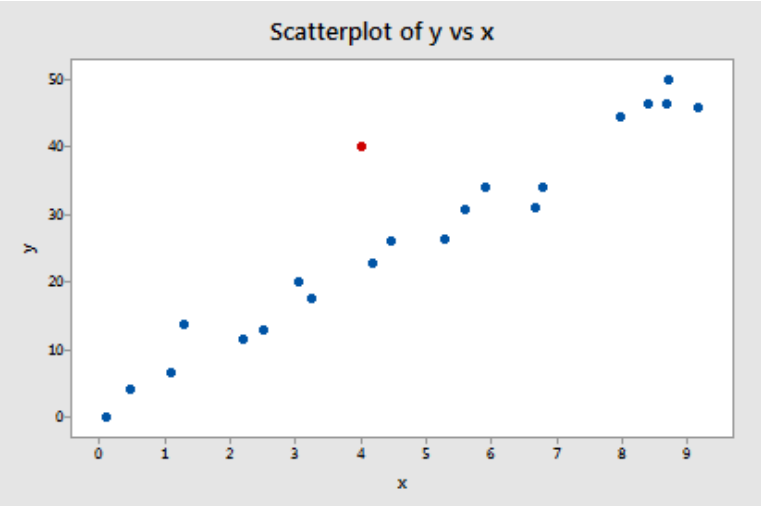
I. Management is considering adding a stadium-style venue that would seat 10,000. What does this model predict that revenue would be if the new venue were to sell out?

$$\widehat{Revenue} = -9.225 + 32.23 \cdot 10000 = \$313075$$

II. Why would it be unwise to assume that this model accurately predicts revenue for this situation? An extrapolation this far from the data is unreliable.

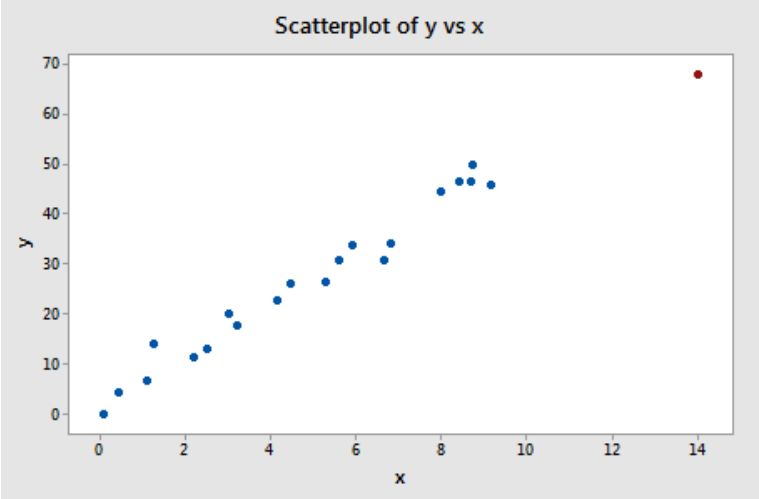
III. Outliers, Leverage and Influence: (Note from: https://newonlinecourses.science.psu.edu/stat501/node/337/)

- **Outlier:** An **outlier** is a data point that far from the overall pattern in the sample



Any Outlier need to be investigated.

- **Leverage:** A data point whose *x*-value is far from the mean of the rest of the *x*-values is said to have high leverage.
    - Leverage points have the potential to pull strongly on the regression line.

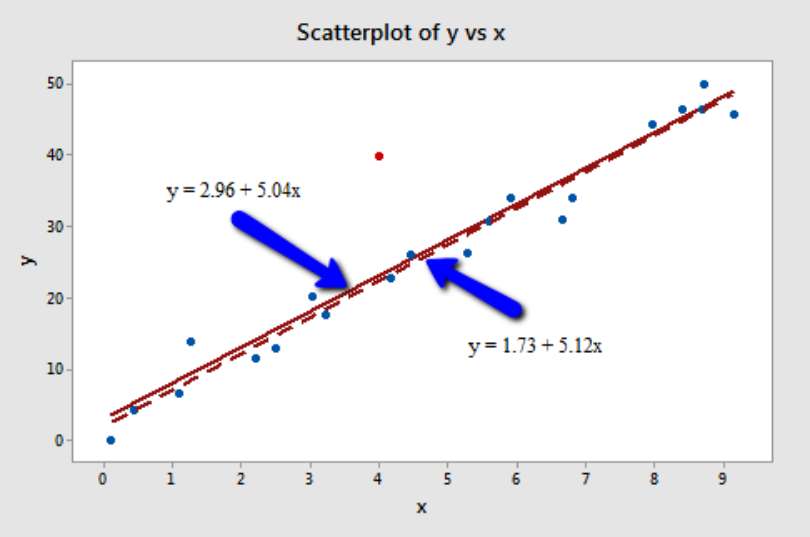


The red point has high leverage.

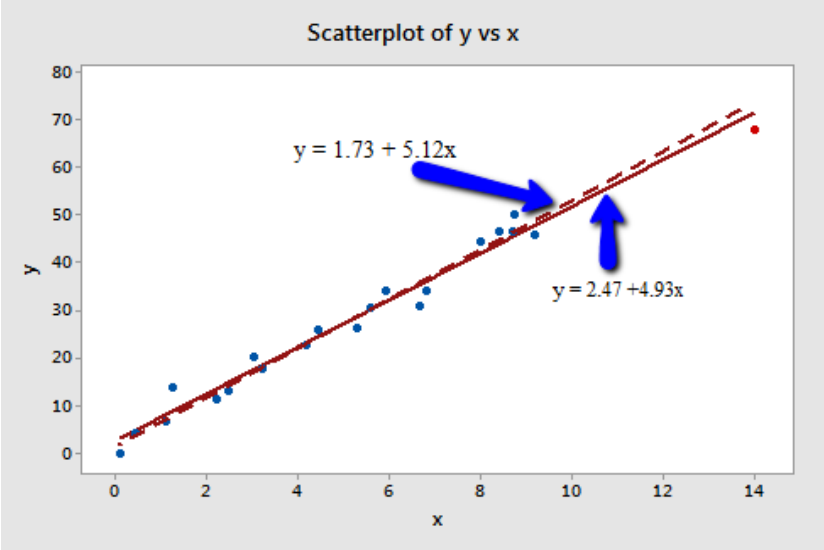
Exercise:

A data point in a scatterplot that pulls the line close to it is said to _____have high leverage__.

- **Influential point:** A point is influential if omitting it from the analysis changes the model enough to make a meaningful difference.

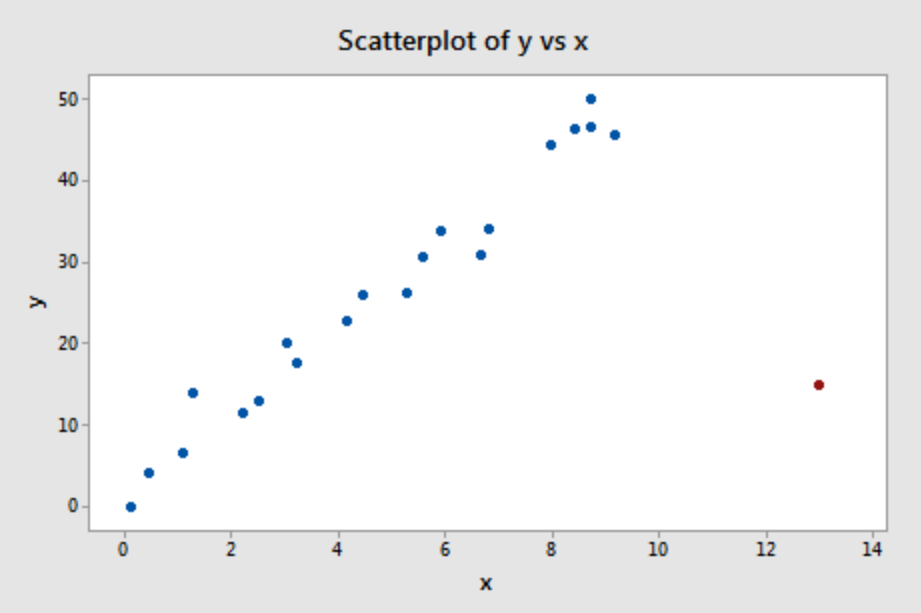


The slopes of the two lines are very similar — 5.04 and 5.12, respectively. Therefore, the red point is not an influential point.

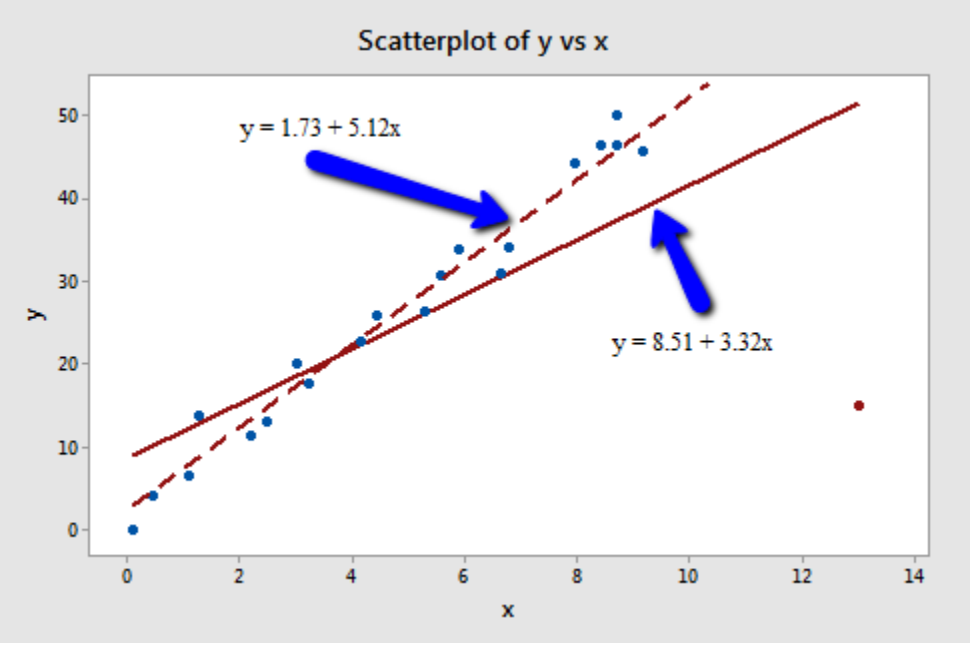


The slopes of the two lines are very similar — 5.12 and 4.93, respectively. Therefore, the red point is not an influential point.

Regression-Wisdom



The red data point does not follow the general trend of the rest of the data and it also has an extreme x value.



The red point is an influential point. Because the slopes of two points are noticeably different.

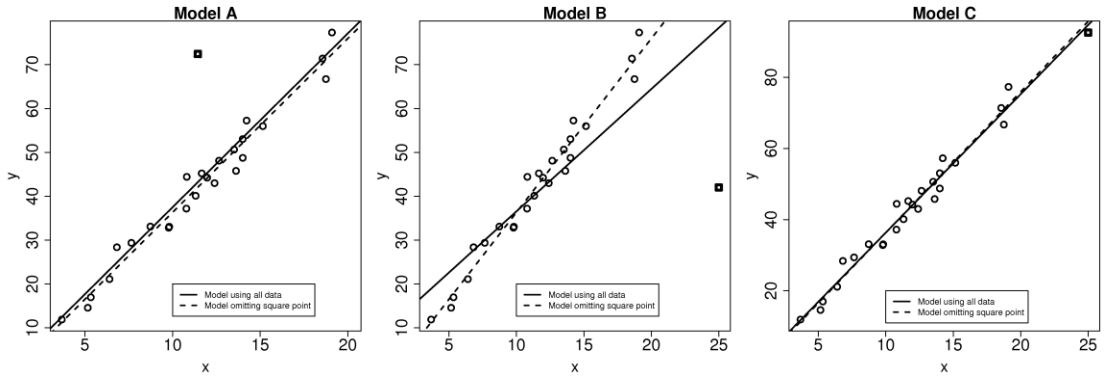
Example:

A researcher is working with a model that uses the number of rings in an Abalone's shell to predict its age. He finds an observation that he believes has been miscalculated. After deleting this outlier, he redoes the calculation. Does it appear that this outlier was exerting very much influence?

Before:		**After:**
R-squared = 59.1%		R-squared = 78.2%
Variable	**Coefficient**	**Variable** **Coefficient**
Intercept	1.791	Intercept 1.54
Rings	0.41	Rings 1.03

Yes, this observation was influential. After it was removed, the slope of the regression line changed by a large amount.

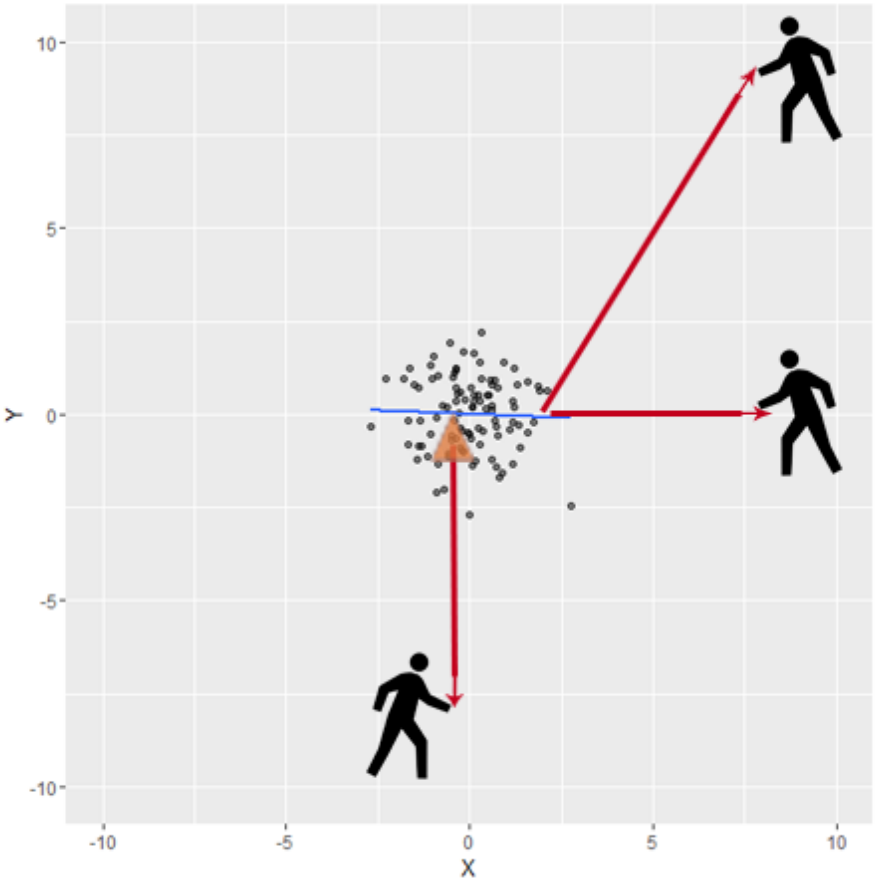
- The outlier points that are not in high leverage positions and also the points that are in high leverage position but not the outliers do not tend to be influential.



Graph from: https://learnche.org/pid/least-squares-modelling/outliers-discrepancy-leverage-and-influence-of-the-observations

Regression-Wisdom

Note from: http://omaymas.github.io/InfluenceAnalysis/



f you think of the regression line as a rod pivoted at the orange triangle, which is placed at the mean *(here [0,0] since X,Y are centered)*, you will be able to conclude the answers easily. For example:

- **At [10,0]**, the guy has high leverage since he stands far from the cloud of the X values. And although he has the potential of high influence, he will not be able to impact the regression line significantly because his force is exerted almost parallel to the line.
- **At [10,10]**, he has both high leverage and high influence, since he stands far from the rest of the observations in X and Y. So he will affect the intercept and the slope of the regression line significantly.
- **At [0,-10]**, he will not have noticeable effect on the fitted line. Since he pulls the line towards him at the pivot, the exerted force is almost perpendicular to the line, and consequently will have no influence.

IV. Lurking Variable:

Lurking Variable
A **lurking variable** is a variable which is not among the variables of a study and yet may influence the interpretation of the relationships among those variables. For example, consider the statistical relationship between ice cream sales and drowning deaths. These two variables have a positive, and potentially statistically significant, correlation with each other. One might be tempted to conclude then, that more ice cream sales *cause* more drowning deaths to occur. The real cause of a corresponding increase in both of these variables is a lurking variable – warm weather. People eat more ice cream *and* go swimming more when it is warm.

A variable that is not part of the model but affects the way variables in the model appear to be related is called a(n) ____Lurking Variable_____.

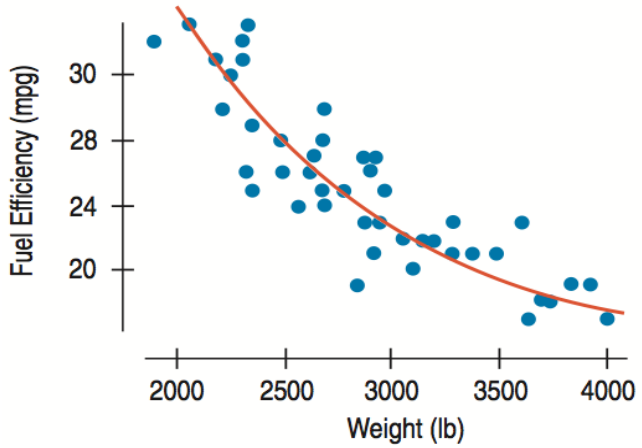
Example:

There is a strong correlation between the temperature and the number of skinned knees on playgrounds. Does this tell us that warm weather causes children to trip?

No. In warm weather, more children will go outside and play.

Please note that:

- We cannot use a linear model unless the relationship between the two variables is linear. Often re-expression can save the day, straightening bent relationships so that we can fit and use a simple linear model.
- If there's a curve in the scatterplot, why not just fit a curve to the data?



- The mathematics and calculations for "curves of best fit" are considerably more difficult than "lines of best fit."

Besides, straight lines are easy to understand. We know how to think about the slope and the *y*-intercept

Example:

team of Calculus teachers is analyzing student scores on a final exam compared to the midterm scores. One teacher proposes that they already have every teacher's class averages and they should just work with those averages. Explain why this is problematic.

Individual student scores will vary greatly. The class averages will have much less variability and may disguise important trends.

Example:

In justifying his choice of a model, a student wrote, "I know this is the correct model because R^2=99.4%."

- Is this reasoning correct? Explain.

No. The scatterplot should be examined first to see if the conditions are satisfied.

- Does this model allow the student to make accurate predictions? Explain.

No, the linear model might not fit the data everywhere

Regression-Wisdom

Regression analysis is used to understand the relationship between the response variable and explanatory variable(s).

The standard error of the estimate measures the accuracy of the prediction, predicted response ($\hat{y}$), that is made by regression line.
As we know the actual response denoted with y and predicted response denoted by $\hat{y}$
The residual is obtained by " $y - \hat{y}$"
Residual squared defined by $(y - \hat{y})^2$
We define the standard error with the Square root of the average of Residual squared. $\sigma_E = \sqrt{\frac{\sum(y-\widehat{y})^2}{N}}$

We usually do not know the population parameters. In which case we calculate the standard error estimate from a collected sample.

$$s_E = \sqrt{\frac{\sum(y-\widehat{y})^2}{N-2}}$$

Remember: $\hat{y} = b_0 + b_1 x$ and $y = b_0 + b_1 x + \varepsilon$
The error of predicting single values of y for a value of x does not converge to 0, it's

measured by Epsilon, ε
in the $y = b_0 + b_1 x + \varepsilon$
With more and more points, the s_E of the Estimate will converge to the std dev of ε.

The assumptions made in the calculation of the regression line are:

- The errors of prediction are normally distributed.
- The variance around the regression line is the same for all values of x.
- The relationship between x and y is linear.

Residual Plot:
A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.