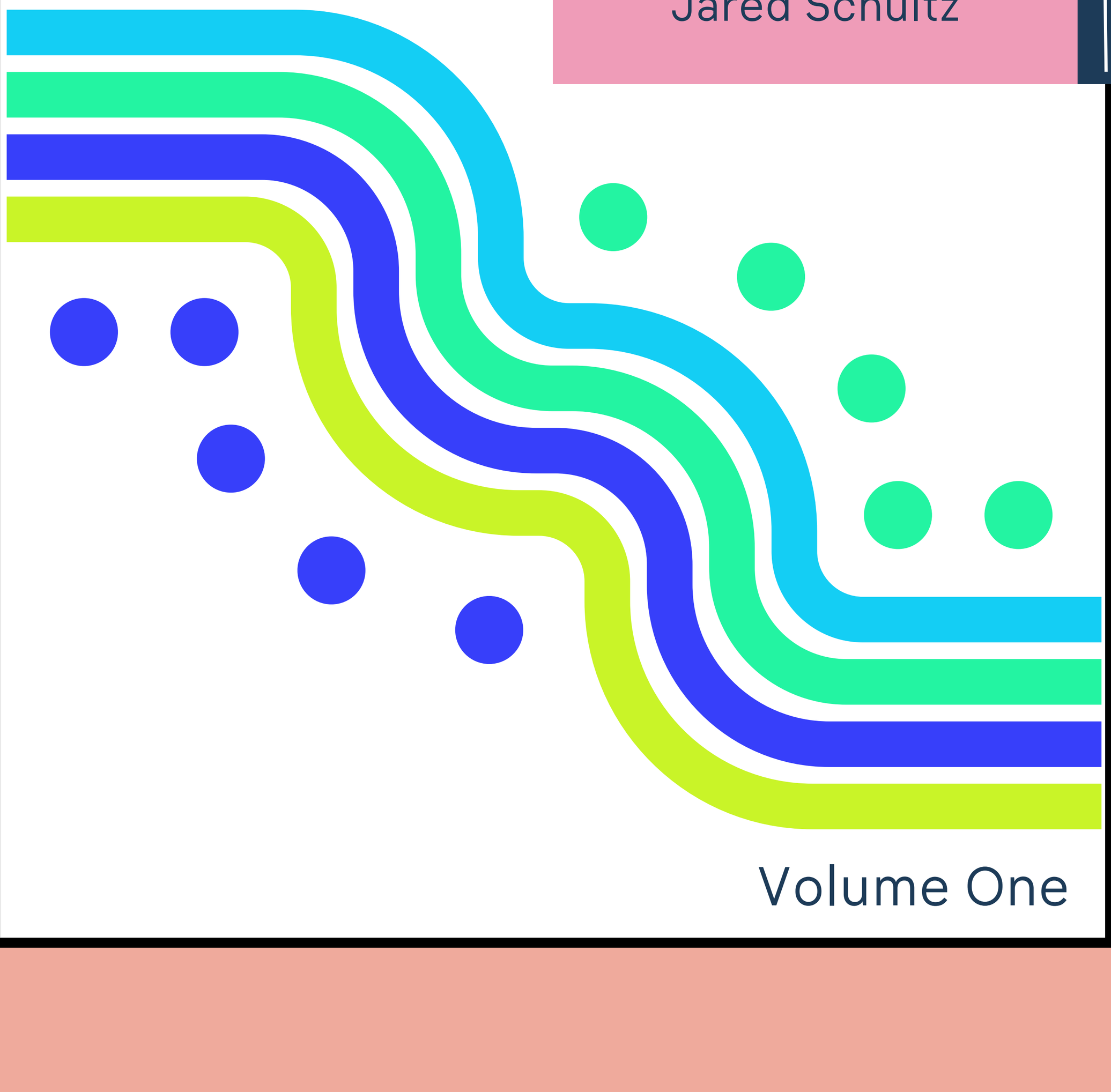
REGRESSION FOR EVERYONE

A Simple Guide To Simple Linear Regression

Jared Schultz



Volume One

Table of Contents

REGRESSION FOR EVERYONE #1

What is Regression? *

Regression is a form of machine learning that examines the relationship between variables and allows us insight on explaining patterns in data.

*This is a very casual definition.

Housing Market Example Guide

- *In this example we are interested in finding what variables affect housing price in a specific region.*
- *We also want our model to be good fit.*

Example

In this example the only data we have about housing in our region is the **house value** and the **age of the house**.

Simple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + error_i$$

$$i = 1, \ldots, n$$

* ***error*** shorthand will be e

Simple Regression Model Legend

Y_i Value of the house for the *ith* house (Dependent or Response variable)

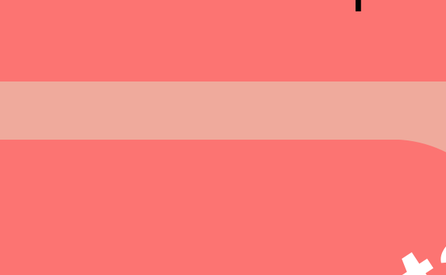
X_i Age of the house for the *ith* house (Independent or Explanatory variable)

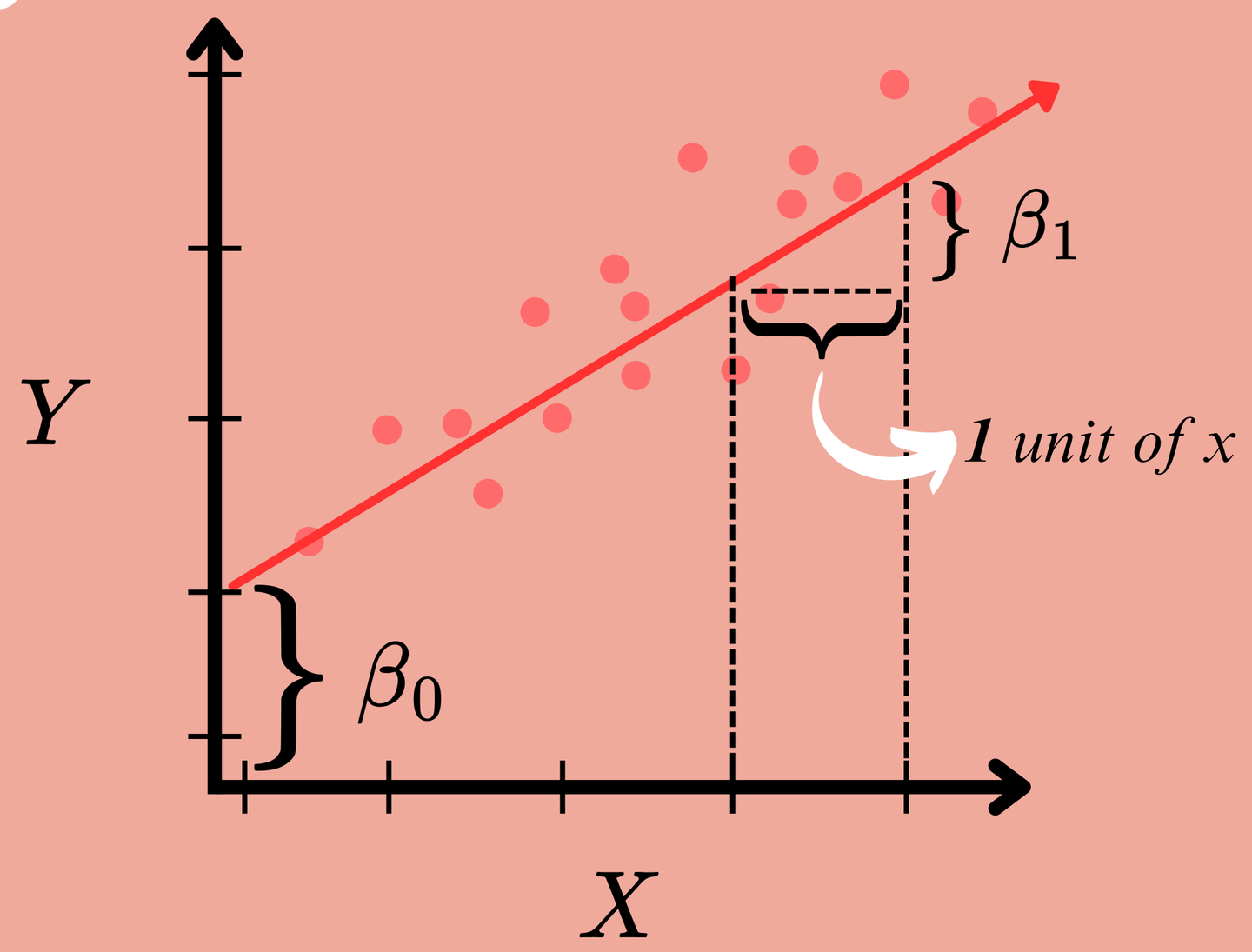
β_0 Regression Intercept (unknown parameter)

β_1 Regression Slope (unknown parameter)

e_i Random variables that have a mean of 0. All have equal variance and are uncorrelated.

Graphical Interpretation





How do we pick our unknown parameters to get the best line that fits our data?

We can use the **Ordinary Least Squares (OLS)** method. The goal of OLS is to fit a line that minimizes the sum of squared difference from the observed and fitted values.

OLS Formula and Graphical Interpretation

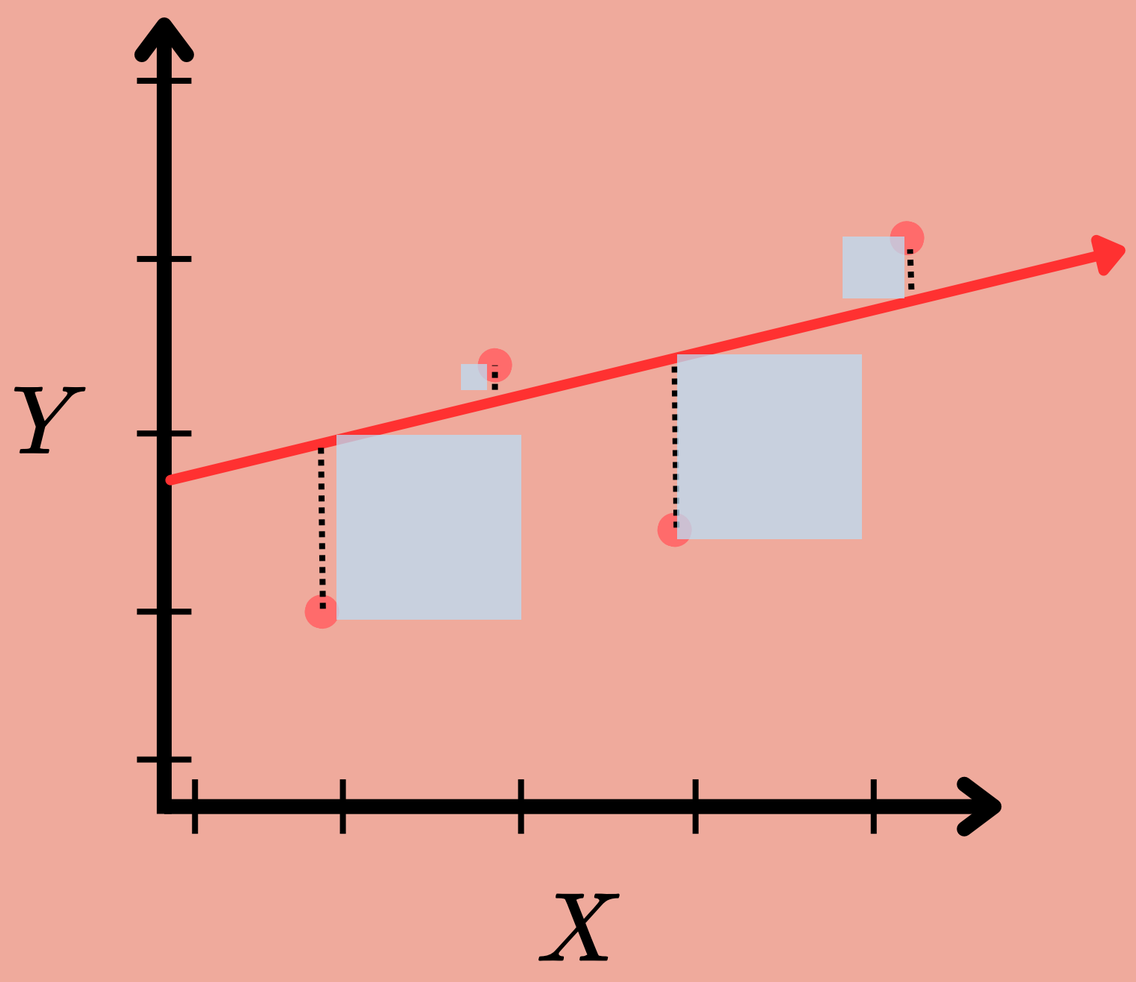
$$Q\left(b_{0,}b_1\right) = arg(min)\sum_{i=1}^{n}\left(Y_i - \left(b_0 + b_1X_i\right)\right)^2$$

This formula looks complicated but it boils down to saying that our parameters we will pick will be the values that minimizes the right-hand side of the equation.

Lets take a graphical approach to understanding this:

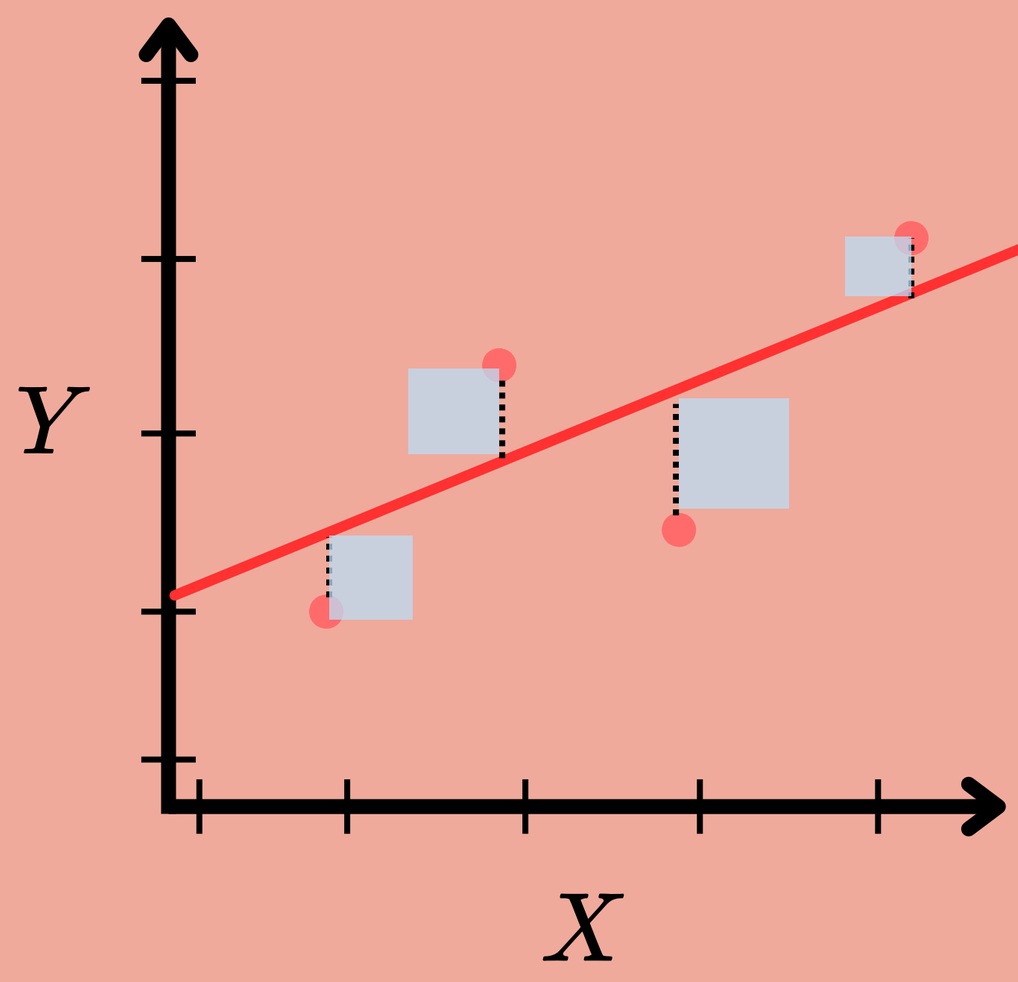
Suppose on the left I create a line with an intercept of 3 and a slope of 0.5.

$$y = 3 + 0.5x$$



Suppose on the right I create a line with an intercept of 2 and a slope of 0.7.

$$y = 2 + 0.7x$$



The OLS formula in a graphical context is the sum of the squares that are shown in blue. In this picture Q(3,0.5) > Q(2,0.7) thus the line that best fits the data is Q(2,0.7) since it contains a smaller sum of error.

OLS Estimators

Obviously, it would be very time consuming find the pair of parameters by hand. Luckly, using calculus we can find the normal equations that allow us to solve for our estimators.

$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widetilde{\beta}_0 = \bar{Y} - \widetilde{\beta}_1 \bar{X}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

*Typically a hat notation is used instead of ~.

What does our fitted line look like using our data?

$$\widetilde{Y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 X_i$$

Fitted values which are predictions by the OLS line. Which means that they are the respective points of the fitted line.

Residuals are the differences between the observed values and the respective fitted values.

$$Y_i - \widetilde{Y}_i$$

REGRESSION FOR EVERYONE #2

RECAP

- Simple linear regression is uses OLS to determine the coefficients of our regression relation.

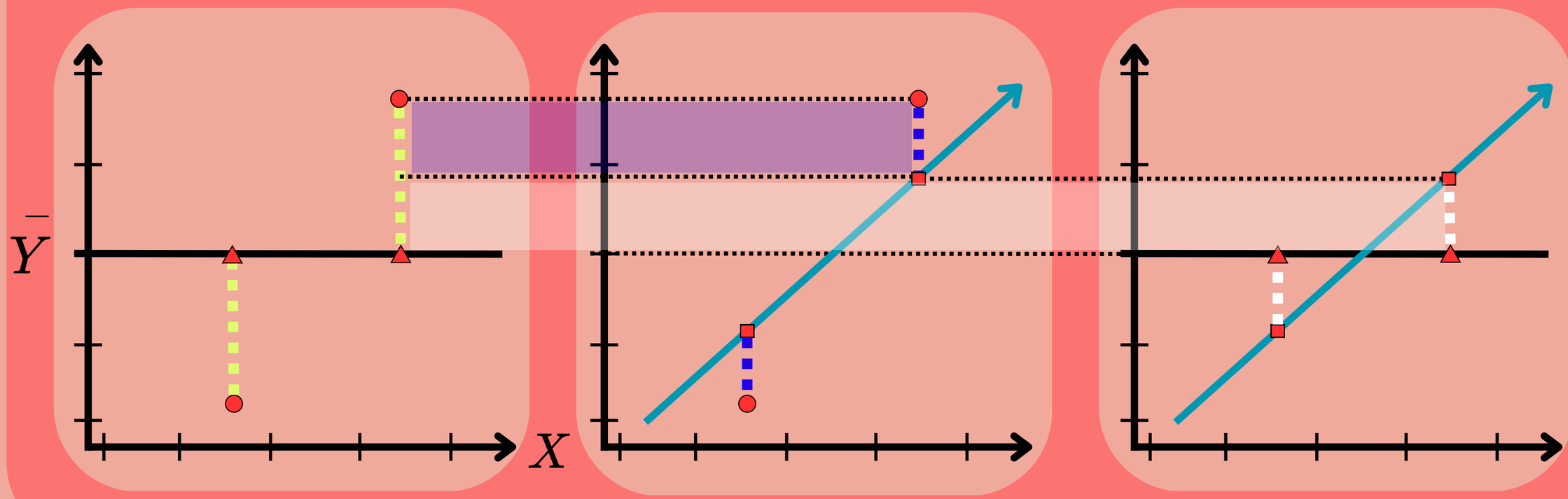
- But how can we quantify how much error is in the model

WHAT IS ANALYSIS OF VARIANCE (ANOVA)?

Basic Idea: Attributing variation in the data to different sources through decomposition of total variation.

Graphical Representation of the Partition of Total Deviation

Total Deviations $Y_i - \bar{Y}$ = **Residuals** $Y_i - \tilde{Y}_i$ + **Deviations** $\tilde{Y}_i - \bar{Y}$



$i = 1, \ldots, n$

Decomposition of Total Variation

Now let's take a look at the sum of squares of the total variation.

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\tilde{Y}_i\right)^2+\sum_{i=1}^{n}\left(\tilde{Y}_i-\bar{Y}\right)^2$$

SSTO = SSE + SSR

Total Sum of Squares (SSTO)

Variation of the observations around the sample mean.

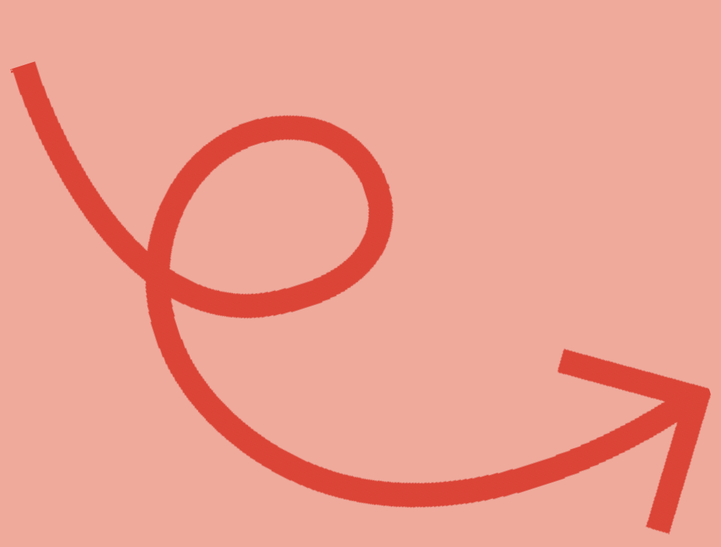
Error Sum of Squares (SSE)

Variation of the observations around the fitted regression line.

Regression Sum of Squares (SSR)

Variation of the fitted values around the sample mean.

Should I use SSE to measure the error of the model?



NO!

Explanation and Solution

SSE is not a good representation of error in the model. Let's take this example to see why:

Suppose that we have a model with 100 points of data and an SSE of 250. We get more data which should give us a better model since we have more data, but our SSE is higher!? This is due to the fact that SSE increases with more data.

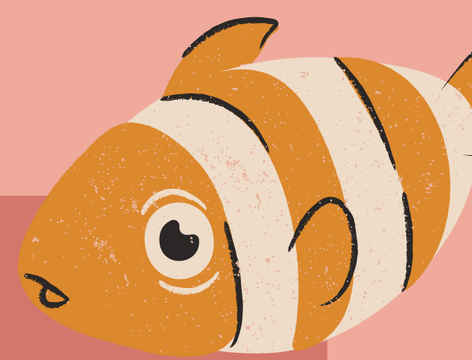
Mean Squared Error (MSE)

Error sum of squares divided by its *degrees of freedom (df)* gives an ***unbiased estimate of the true error variance***.

$$MSE = \frac{SSE}{df\,(SSE)}$$

Degrees of Freedom (df)

The number of components that are allowed to vary. For simple linear regression the df(SSE) is n-2.



Regression Mean Square (MSR)

Regression sum of squares divided by its *degrees of freedom (df). In simple linear regression the df(SSR) is 1.*

$$MSR = \frac{SSR}{df\,(SSR)}$$

REGRESSION FOR EVERYONE #3

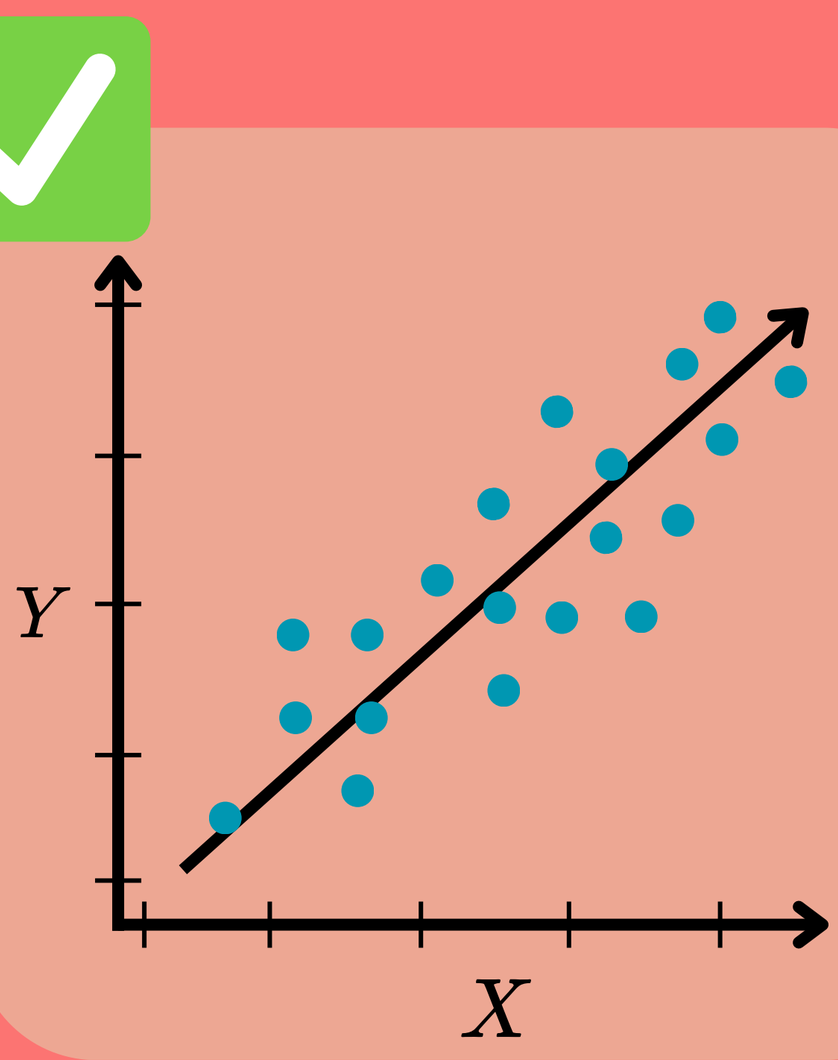
MODEL DIAGNOSTICS

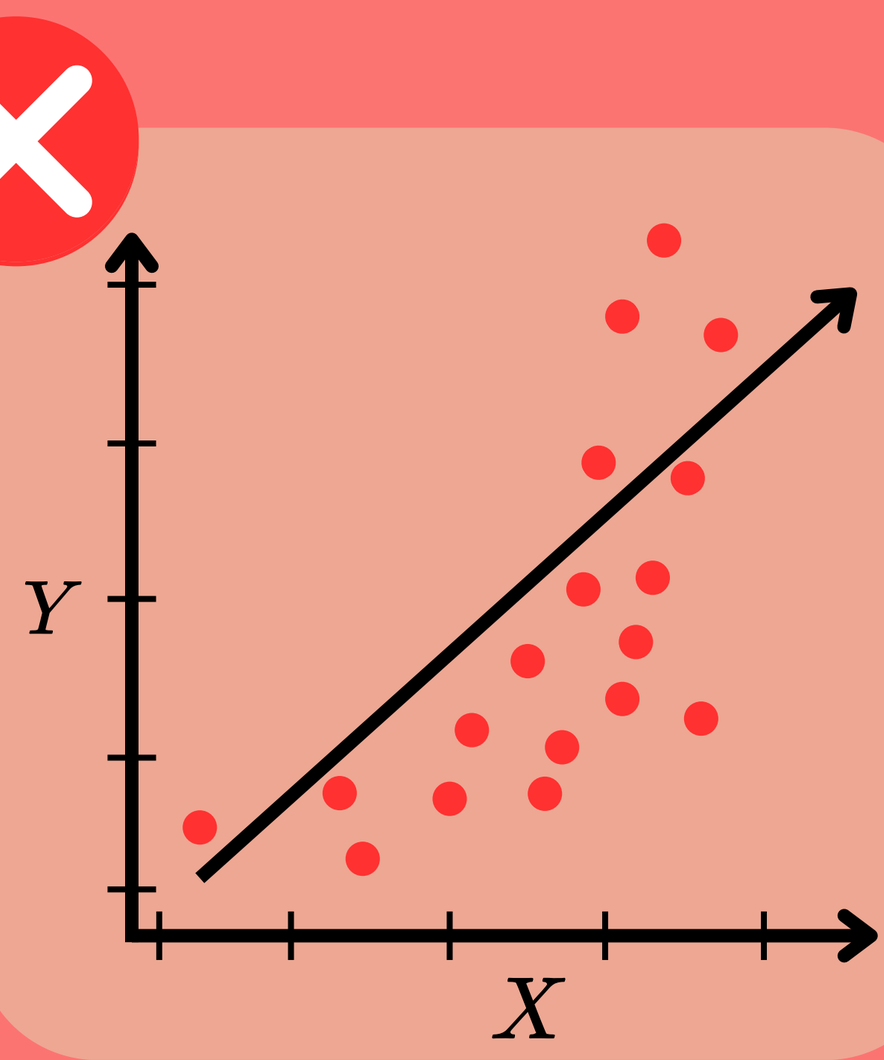
Basic Idea: Our model has assumptions. We have to make sure that our assumptions of the normal error model in simple linear regression are correct. Otherwise we do not have a valid model.

Simple Regression Model Assumptions

1) Linearity

There is a linear relationship between Y and the coefficients of the model.





How to Check Our Assumptions

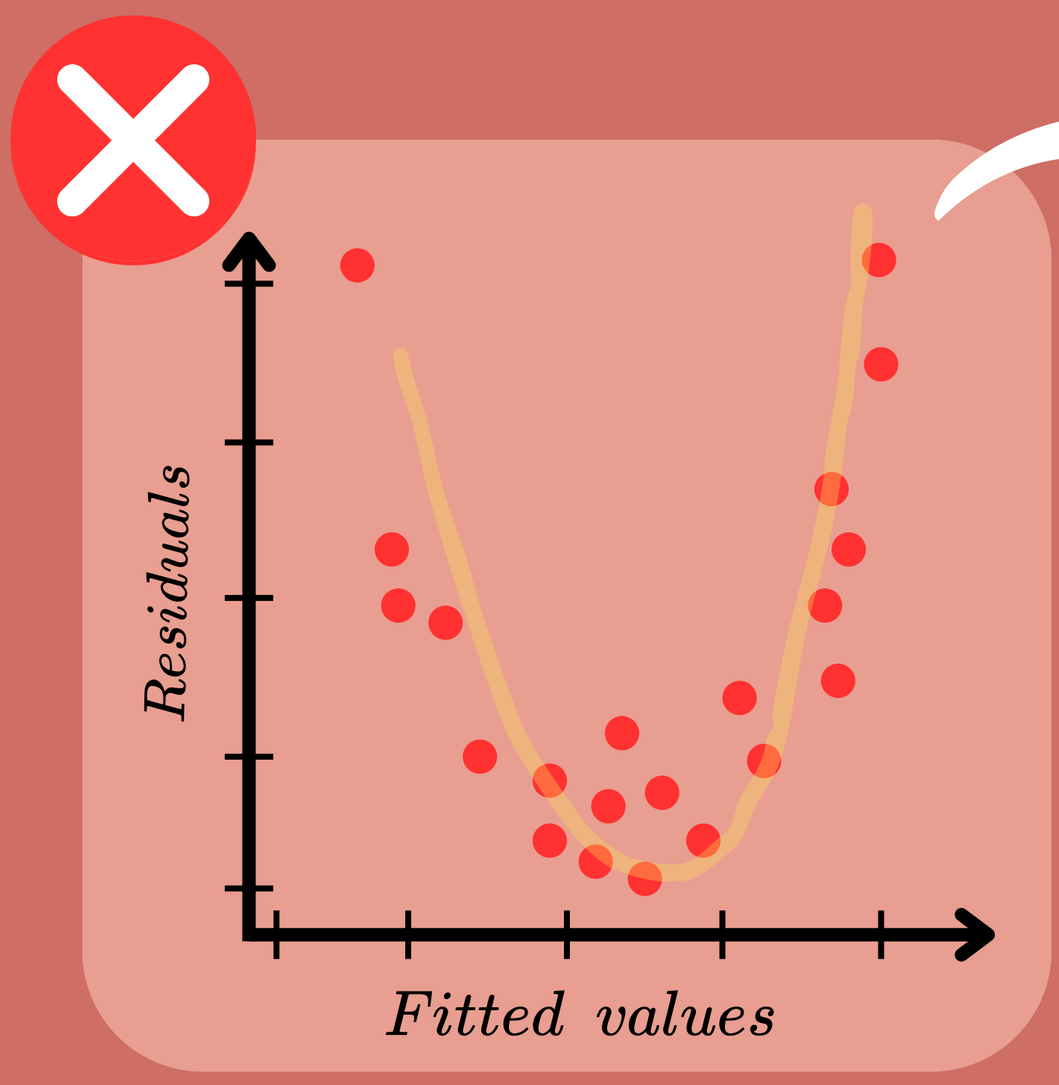
- Unless it is clearly obvious in the scatterplot, we will often need to do a **residual analysis** to check our model assumptions.
- Residuals contain the leftover variation in the data after accounting for our model fit.

Detection of Non-Linearity

- Residuals vs. fitted values plot.
- Residuals vs. X variable plot.

- If either of these **show a clear non-linear pattern**, then there is a possible indication of non-linearity.
- Non-linearity unaccounted for by the model will be left in the residuals.

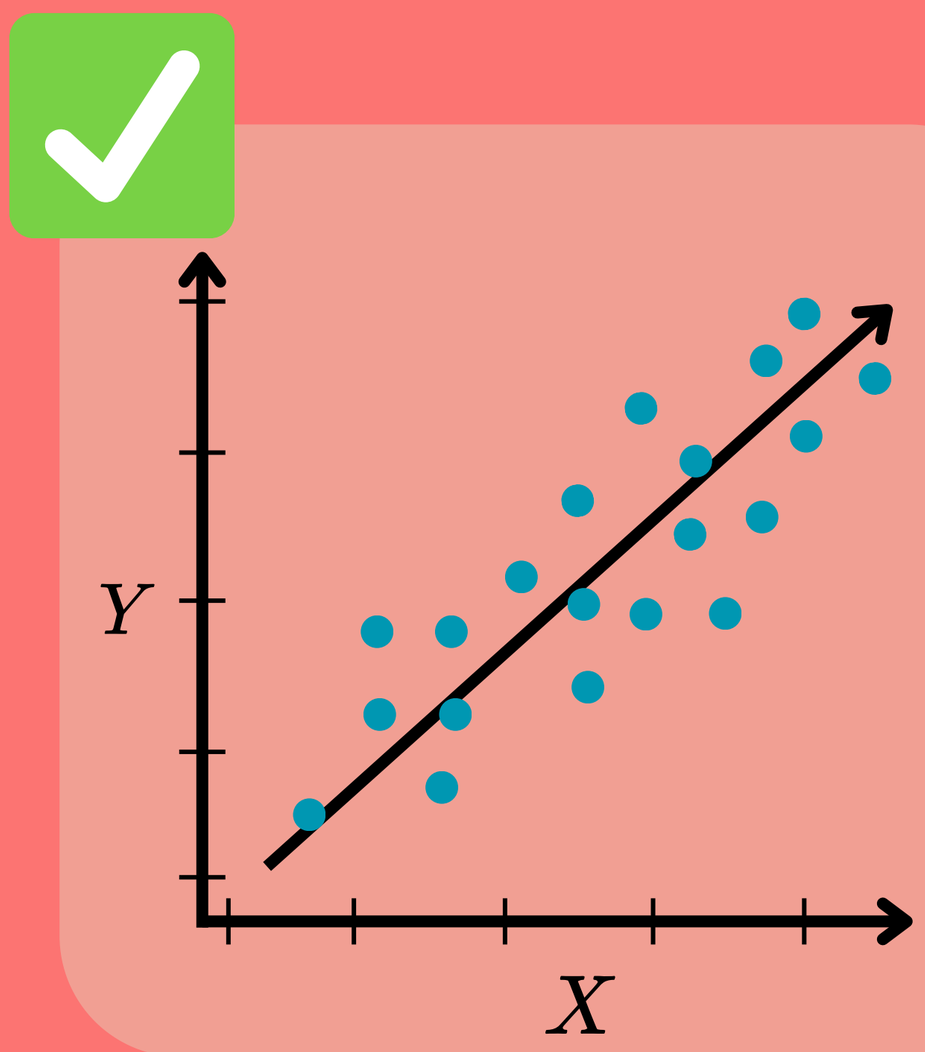
Residual Analysis



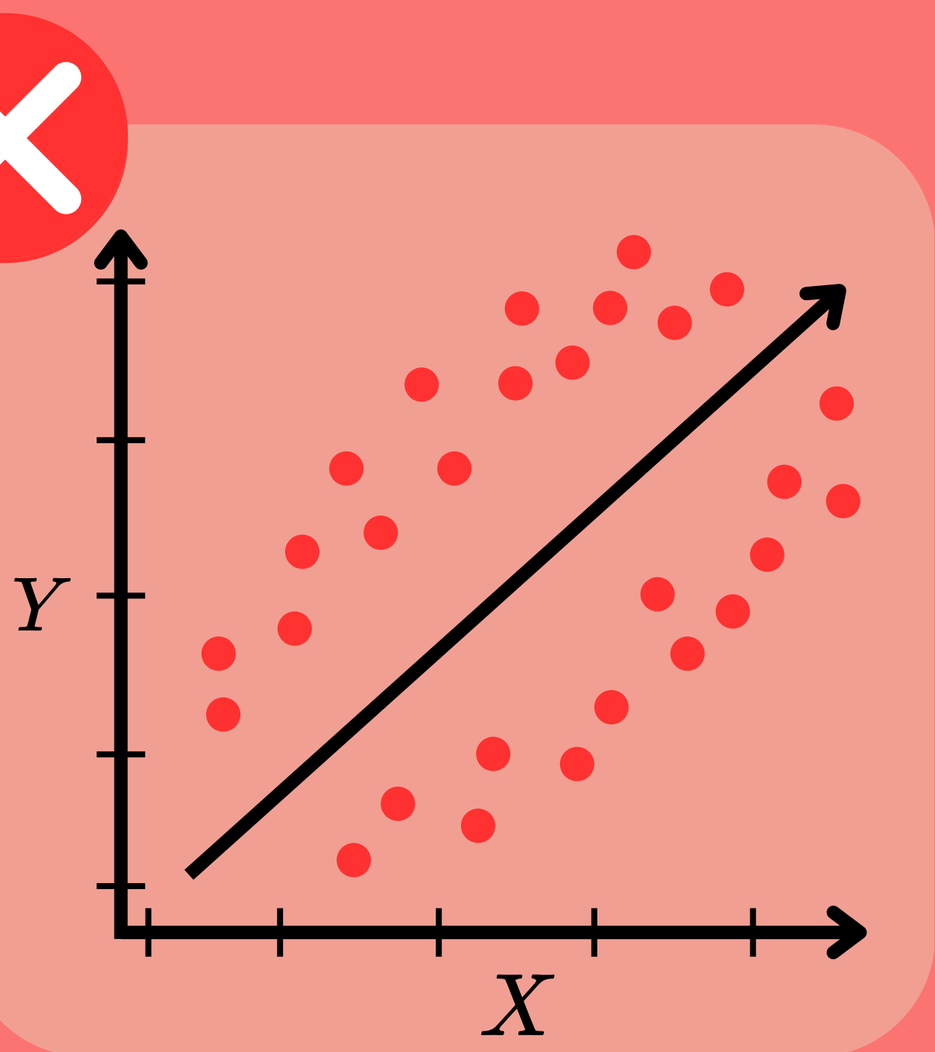
Assumption Violation:
There is a clear quadratic pattern indicating that our regression relation is non-linear.

2) Normality

We assume that our error distribution is normally distributed.



Error Distribution ~

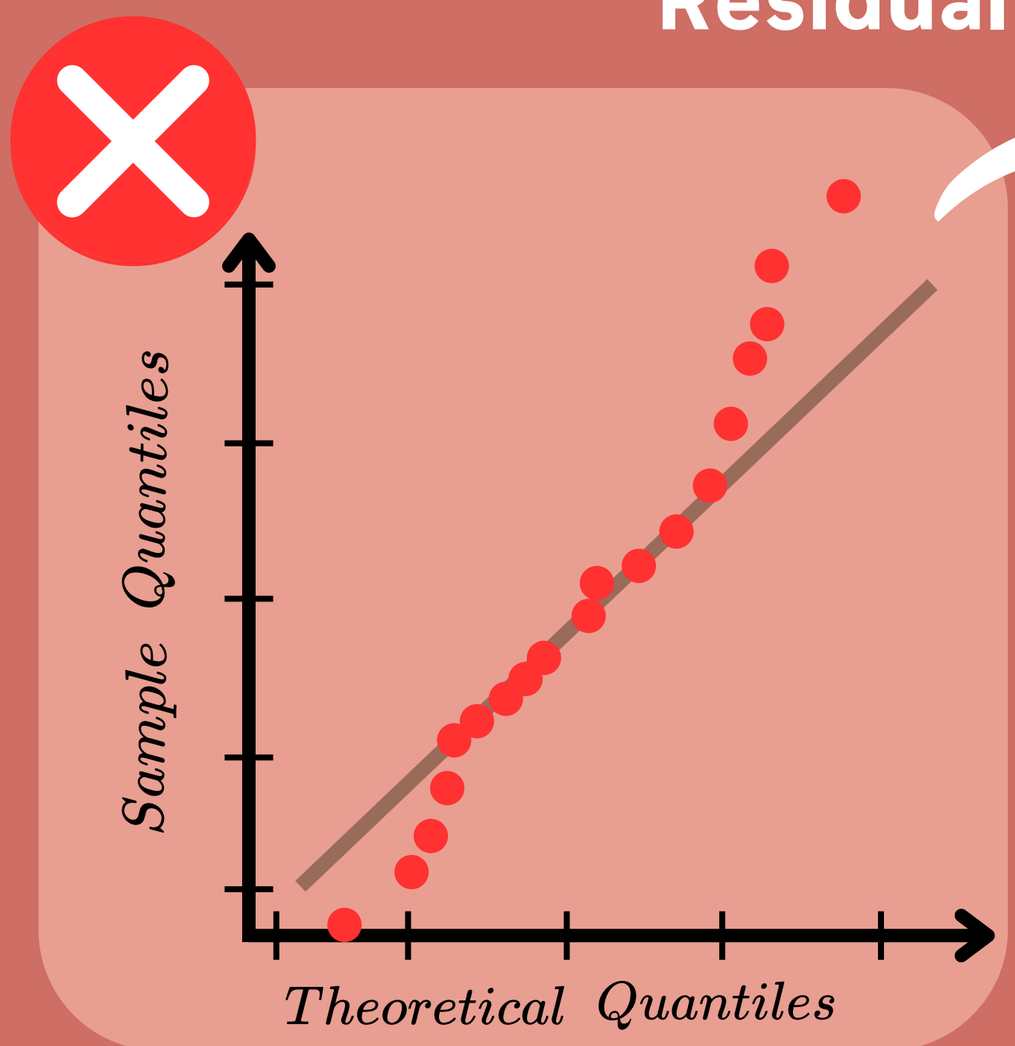


Error Distribution ~

Detection of Non-Normality

- Normal Q-Q plot of the residuals.
- If the residuals are normally distributed, then the points of the Q-Q plot **should be nearly straight** on a line.

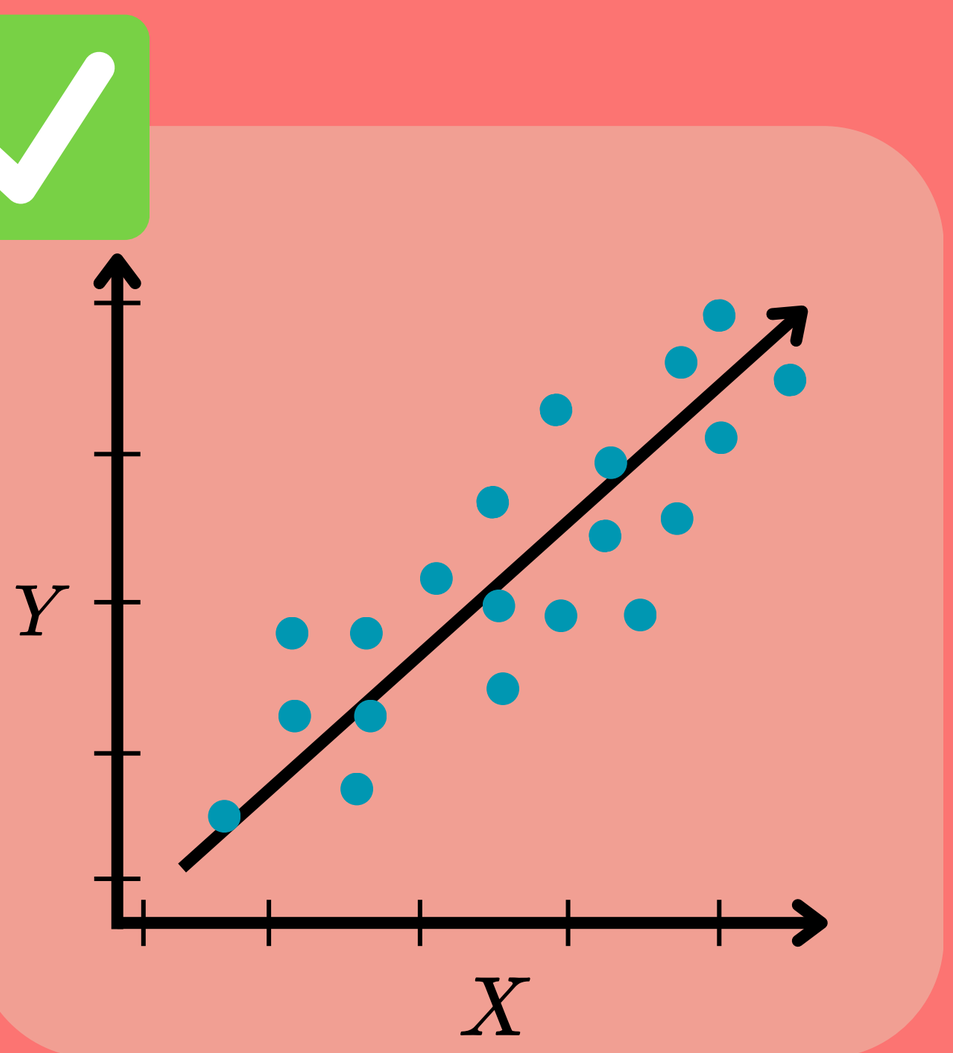
Residual Analysis

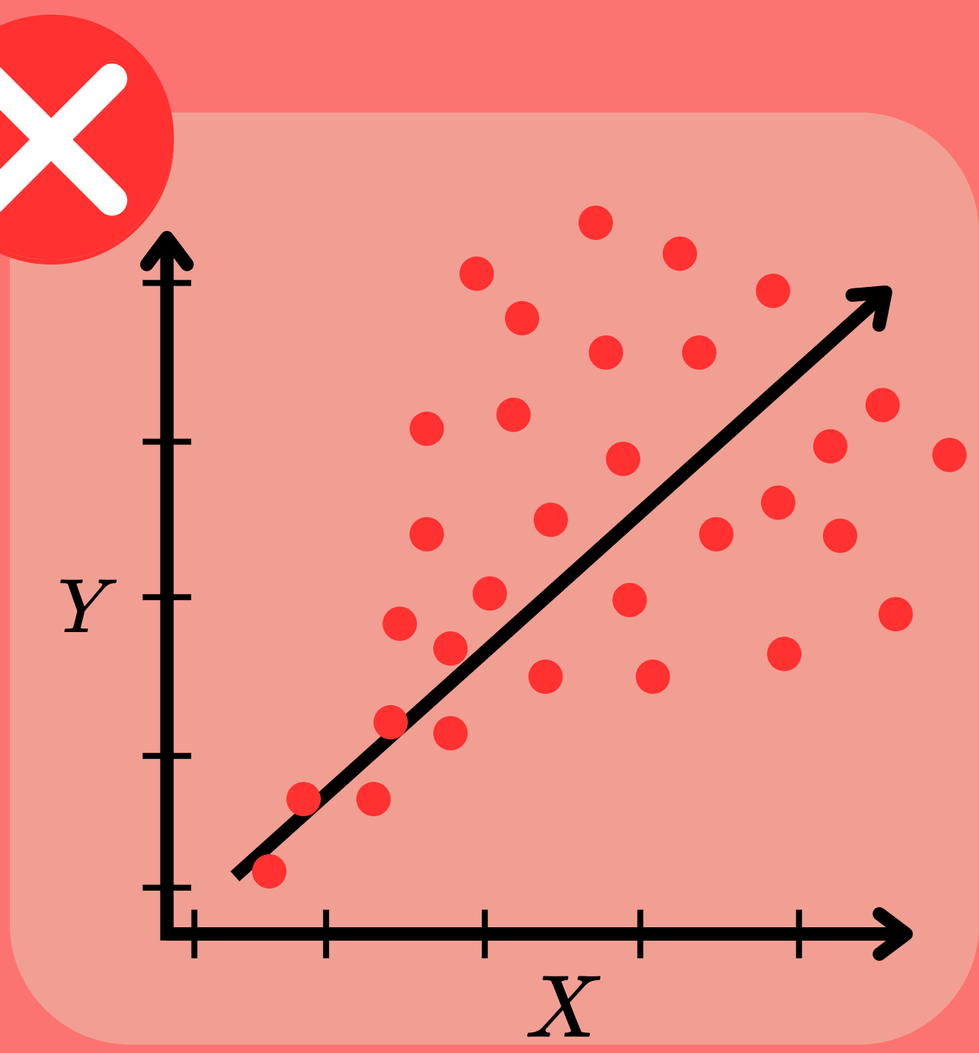


Assumption Violation:
This plot shows more probability mass on both tails. Distribution has heavy tails and is not normal.

3) Constant Variance

We assume that all of the errors have equal variance.

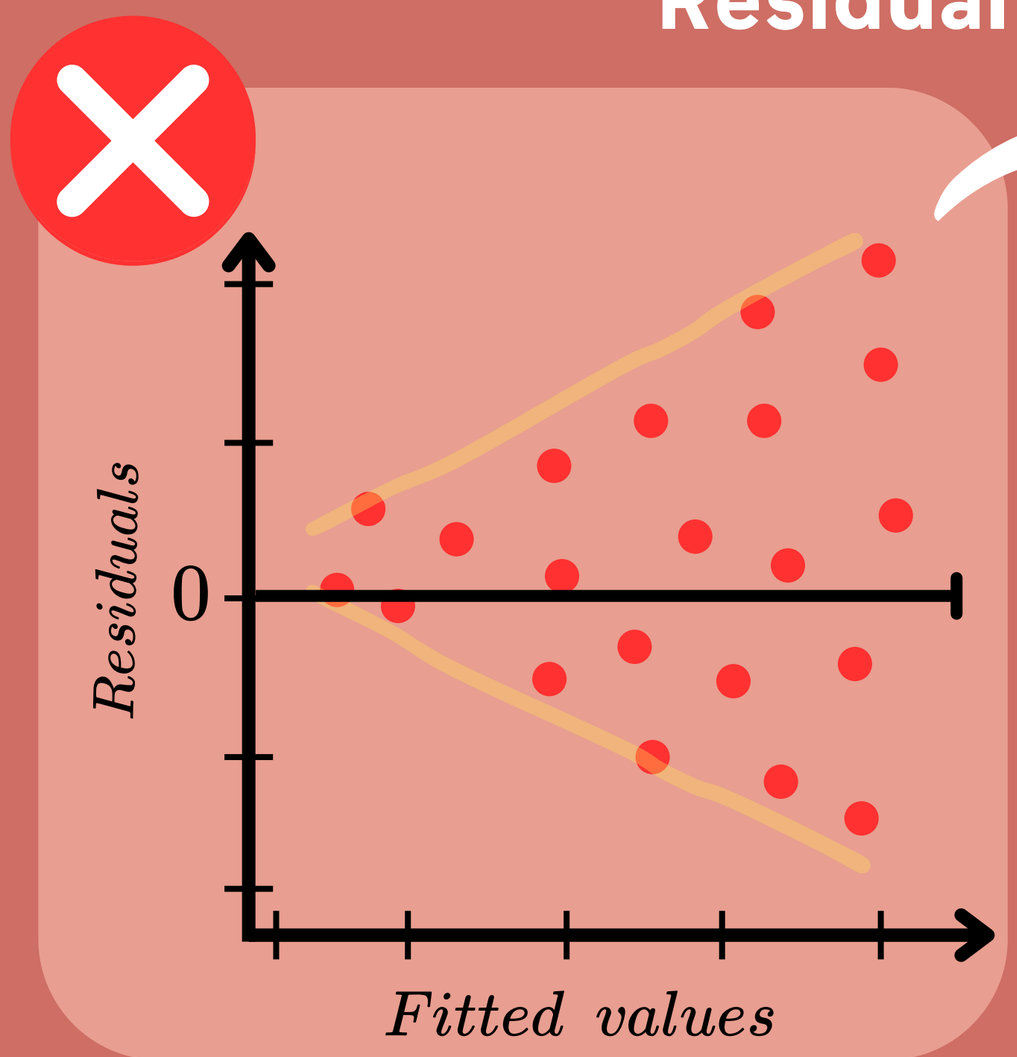




Detection of Non-Constant Variance

- Residuals vs. fitted values plot.
- If this plot **shows a clear increasing or decreasing spread**, then there is a possible indication of non-constant variance.

Residual Analysis



Assumption Violation:
There is a clear increasing pattern indicating that our variance is not constant.

4) Independence

We assume that all of the errors are independent.

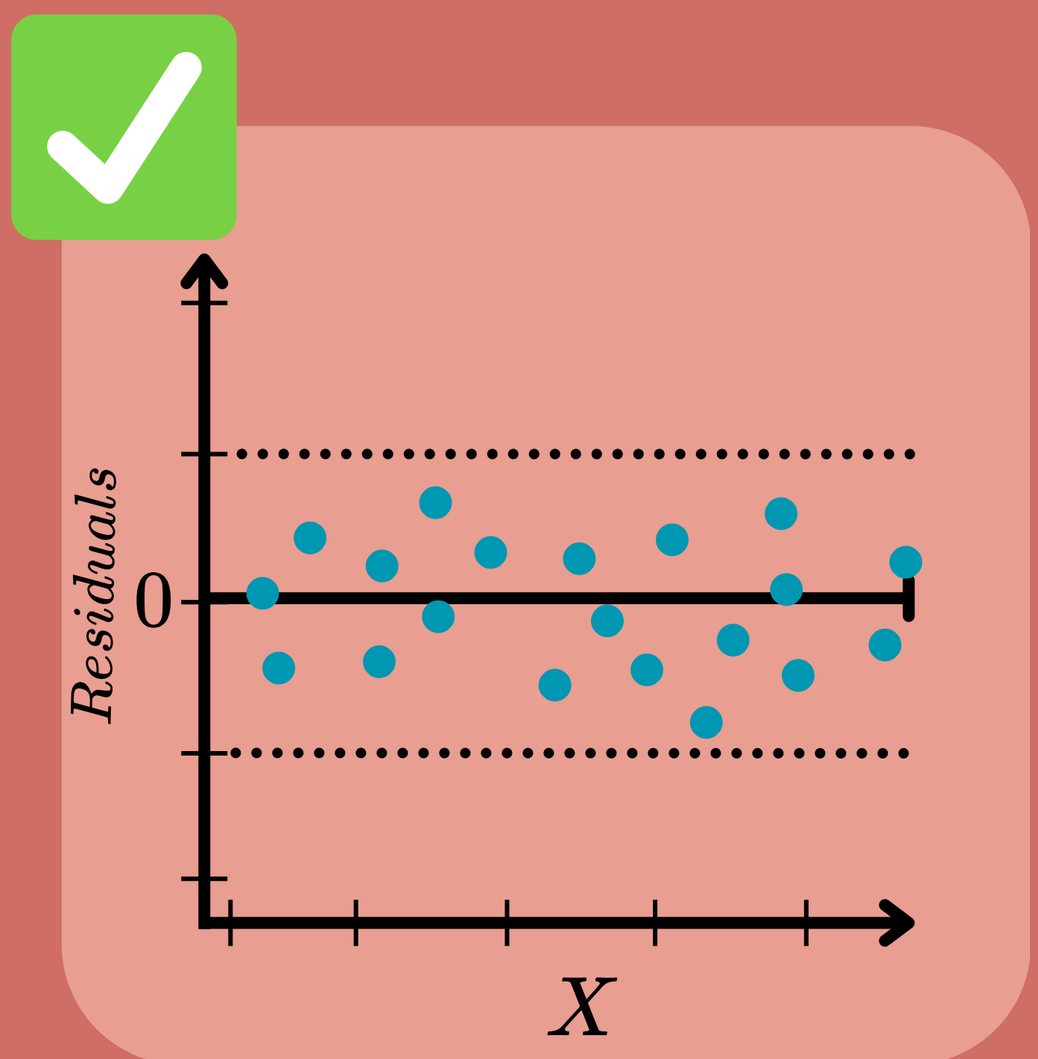
- This assumption is often overlooked since it can be done during the data collection stage. They make observations collected independent.

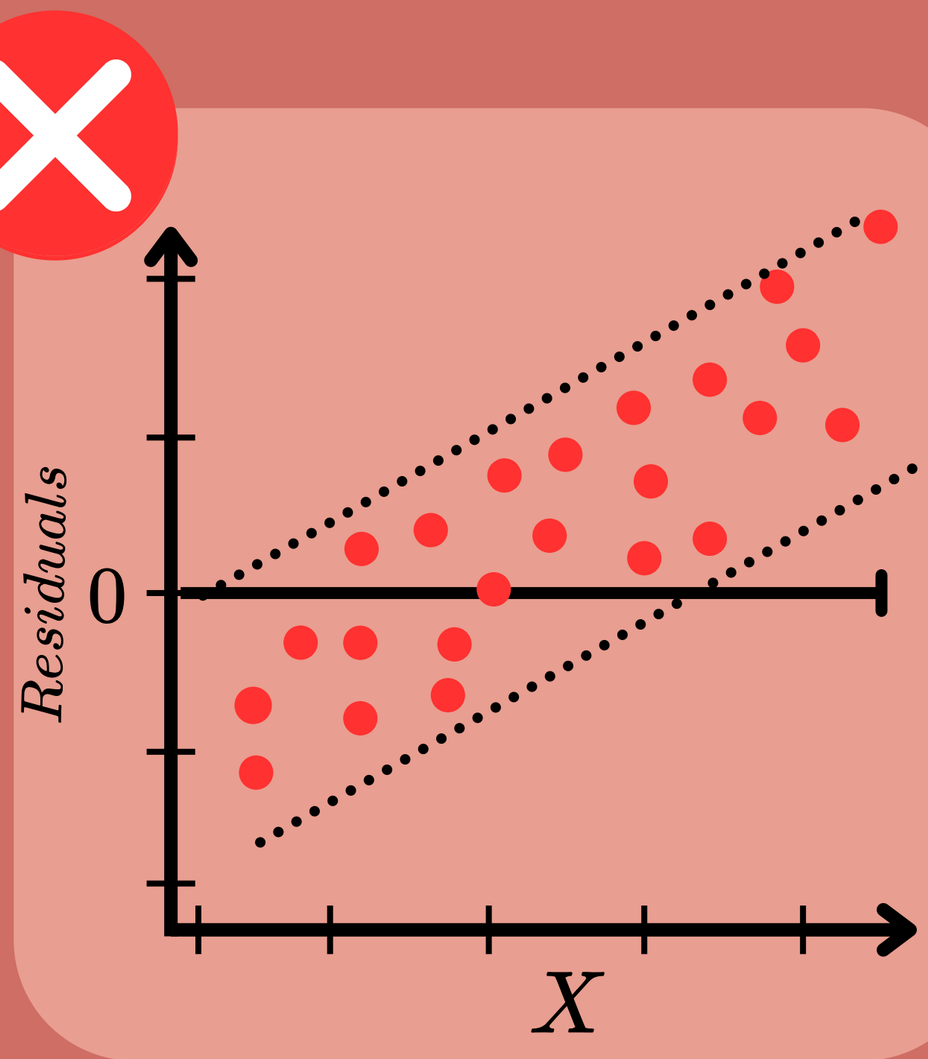
- The assumption might be broken when working with longitudinal data, time series data or cluster collected data.

Detection of Dependence

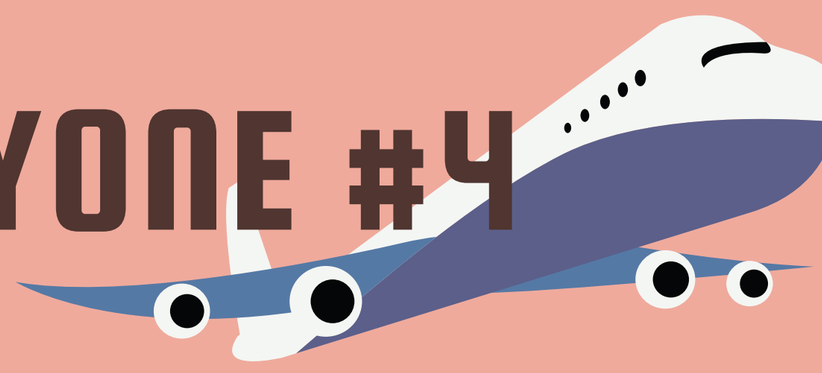
- Residuals vs. Time (X)
- If this plot **shows residuals deviating outside of a 95% CI around 0**, then there is a possible indication of dependence.

Residual Analysis





REGRESSION FOR EVERYONE #4



FIXING MODEL DEPARTURES

Basic Idea: Our model might have departed from the assumptions. Thus, we need to fix our model in such a way that our assumptions still hold true.

WHAT SHOULD I DO ?

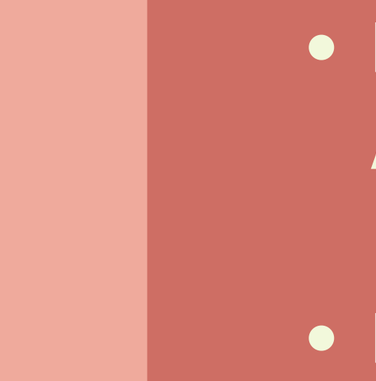
First, mild departures of our model do not need to be fixed. Serious departures in our model include:

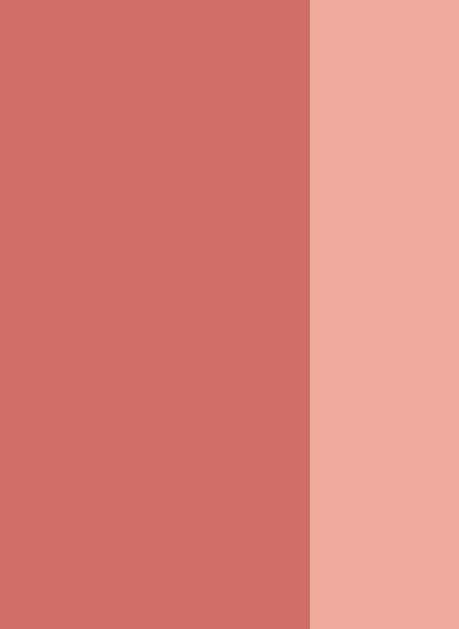
- **Fix Regression Relation** (Linearity Assumption)**:** Transformation of the Y and/or X variable may be needed.

- **Fix Error Distribution** (Normality and Equal Variance Assumption)**:** Transformation of the Y variable.

- **Fix Outliers** (Influential Cases)**:** Exclusion or robust regression.

NOTE: Fixing departures can take some time and exploration. Below are common methods of fixing them. Remember, when applying transformations, it affects the interpretation of the model results and can affect model interpretability.





Transformations of X

We may want to linearize a non-linear relationship:

- Data is *increasing* and *concave downward*:

$$X^{\times} = log\,(X) \qquad X^{\times} = \sqrt{X}$$

- Data is *increasing* and *concave upward*:

$$X^{\times} = exp\,(X) \qquad X^{\times} = X^2$$

- Data is *decreasing* and *concave upward*:

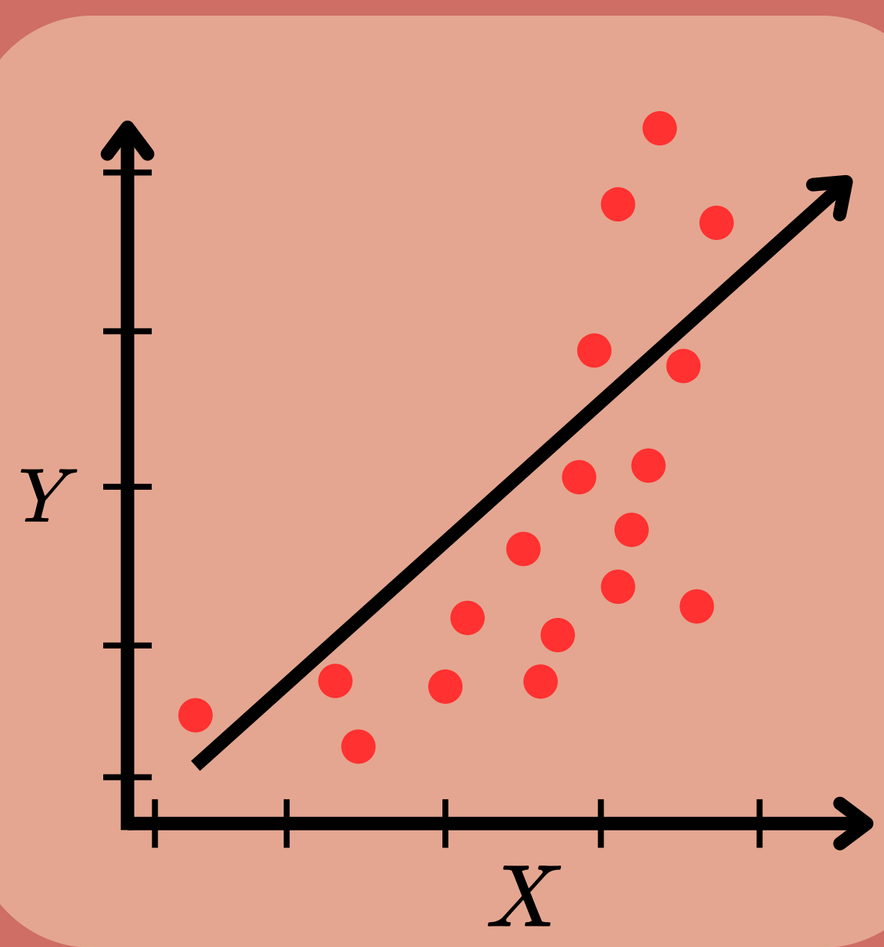
$$X^{\times} = exp\,(-X) \qquad X^{\times} = \frac{1}{X}$$

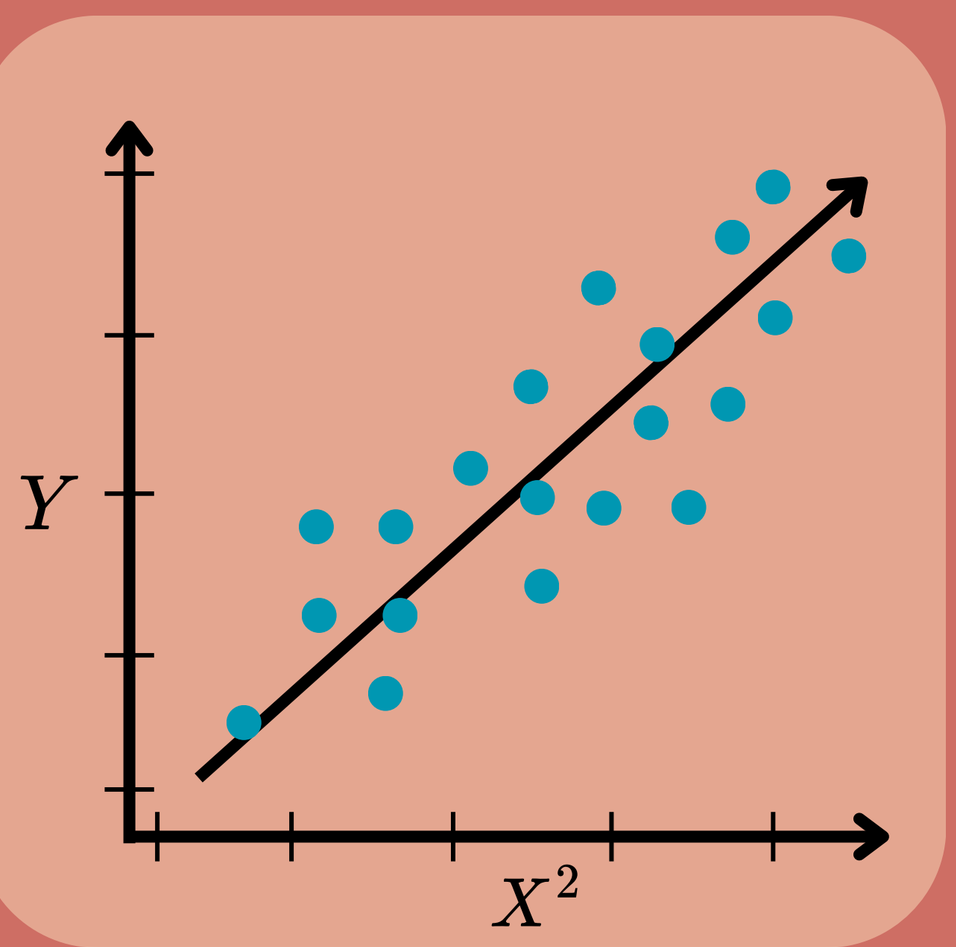
- Add constants to the transformation to avoid negatives or zeros.

$$X^{\times} = \frac{1}{c + X}$$

Example Application

$$Y \sim X \qquad\qquad Y \sim X^2$$





Transformations of Y

Fixing error distribution such as **unequal variance** and/or **non-normality**:

$$Y^{\times} = log\,(Y) \qquad Y^{\times} = \sqrt{Y} \qquad Y^{\times} = \frac{1}{Y}$$

Box-Cox Procedure

A method for picking **a power transformation** on the Y variable to make the distribution normal. (Use R library: MASS)

The procedure is as follows:

- For each λ, fit a regression model on the transformed data $Y^{\times}$ and record the SSE for each choice of λ.
- Find the λ that minimizes SSE and apply the corresponding power transformation to Y.

$$Y_i^{\times} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda \eta^{\lambda - 1}} & if, \quad \lambda \neq 0 \\ \eta log\,(Y_i) & if, \quad \lambda = 0 \end{cases}$$

$$\eta = exp\left(\frac{1}{n}\sum \ln\,(Y_i)\right)$$

Rather than using the entire transformation above, a simpler one you can try after getting λ is:

λ	-3	-2	-1	-0.5	0	0.5	1	2	3
$Y^{\times}$	$\frac{1}{Y^3}$	$\frac{1}{Y^2}$	$\frac{1}{Y}$	$\frac{1}{\sqrt{Y}}$	$log\,(Y)$	$\sqrt{Y}$	Y	Y^2	Y^3

REGRESSION FOR EVERYONE #5

CONFIDENCE INTERVALS

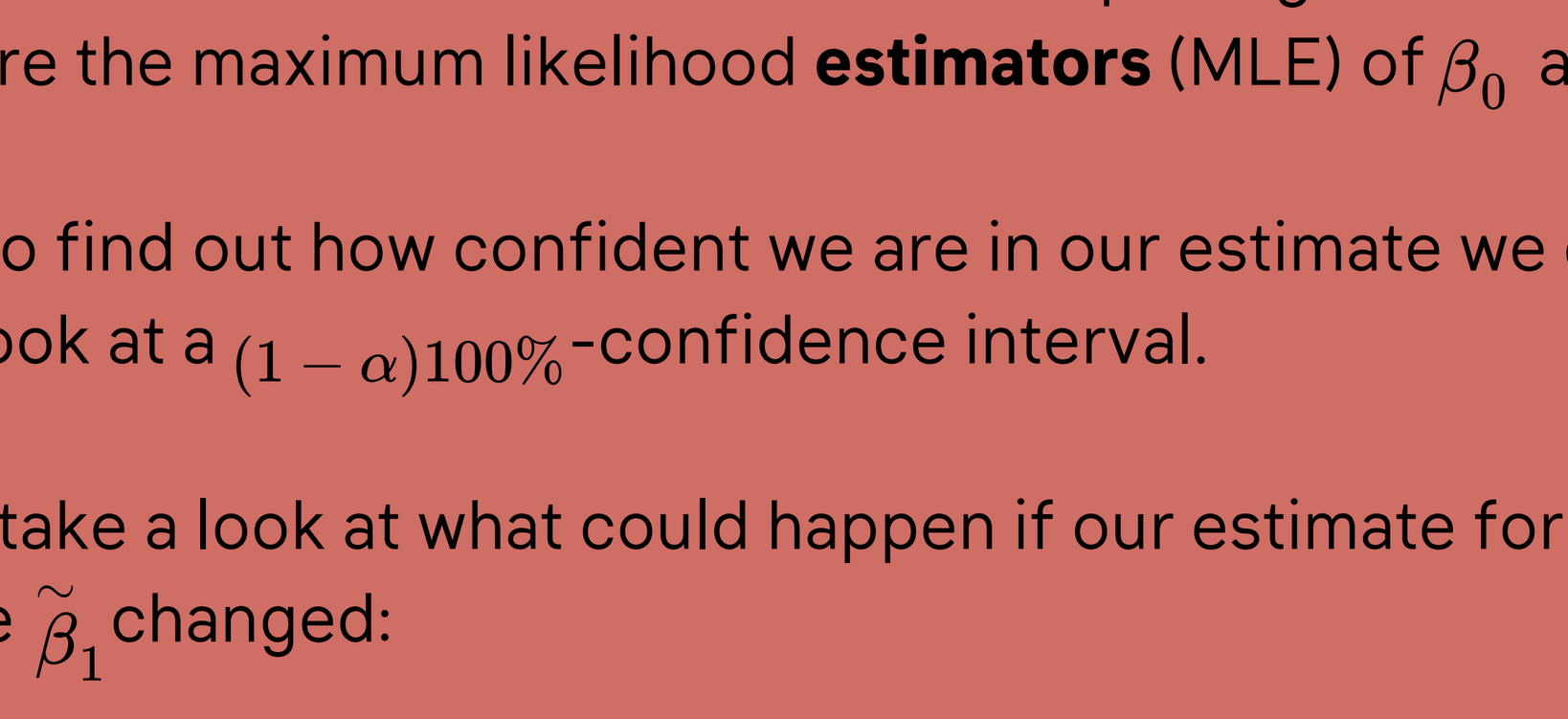
Basic Idea: We want to have a measure of confidence regarding our estimates that come from our simple linear regression model.

Confidence Intervals

Recall: Our regression coefficients that we find are $\widetilde{\beta}_0$ and $\widetilde{\beta}_1$.

- Under the normal error model for simple regression these are the maximum likelihood **estimators** (MLE) of β_0 and β_1.
- To find out how confident we are in our estimate we can look at a $(1-\alpha)100\%$-confidence interval.

Lets take a look at what could happen if our estimate for the slope $\widetilde{\beta}_1$ changed:

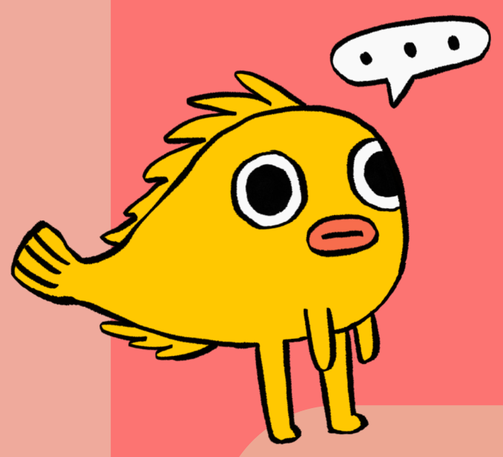


Confidence Interval for $\widetilde{\beta}_1$

The $(1-\alpha)100\%$-confidence interval form:

$$\widetilde{\beta}_1 \pm t\left(1-\frac{\alpha}{2}; n-2\right) s\{\widetilde{\beta}_1\}$$

Accuracy *Precision*



KEY

α Amount of type one error allowed.

$t\,(\sim)$ Critical t-value related to confidence level and degrees of freedom.

$s\{\sim\}$ Standard error of the estimated coefficient.

n Sample size.

Accuracy and Precision

$(1-\alpha)100\%$ is called the **confidence level** and represents the *accuracy* of the confidence interval.

- The higher the confidence level, the more accurate the confidence interval.

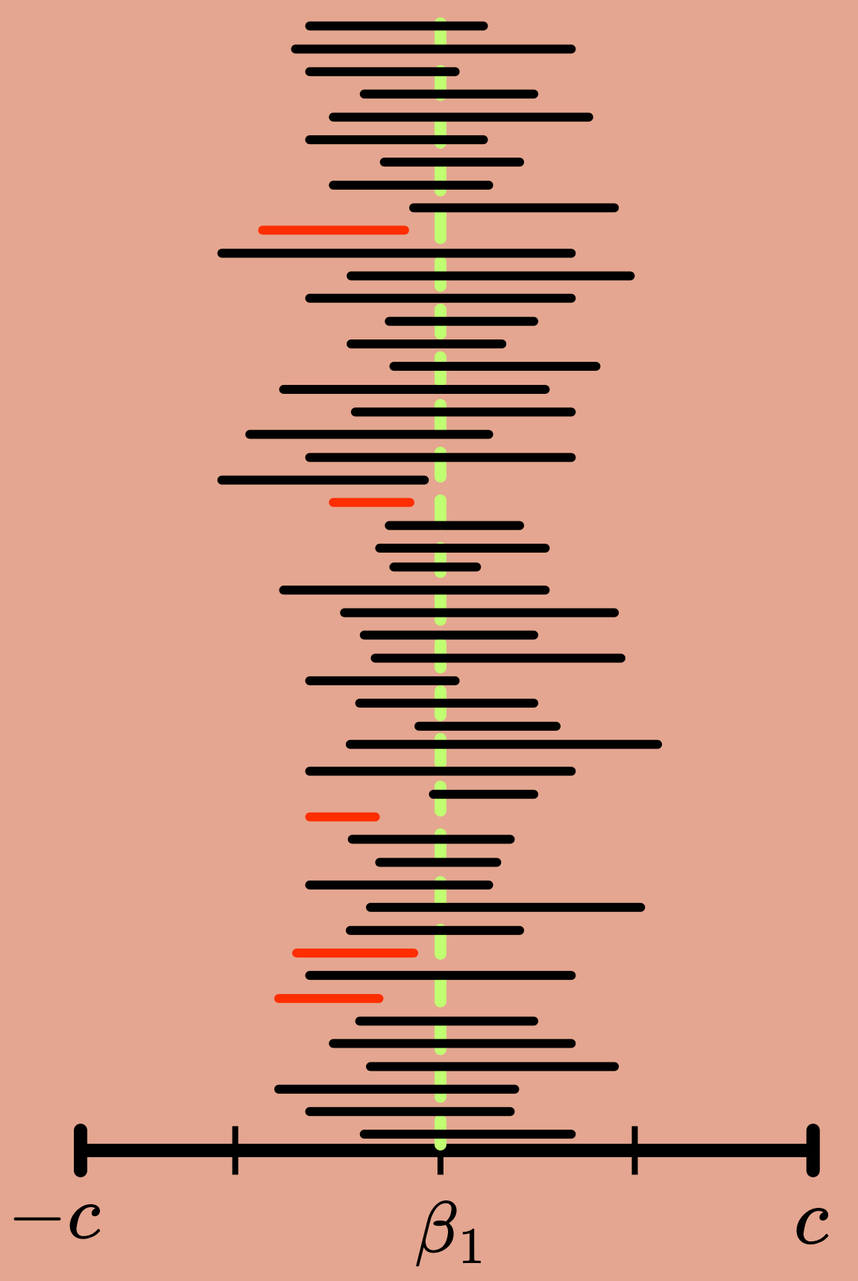
$t\,(\sim)s\{\sim\}$ is called the **half-width** and represents the *precision* of the confidence interval.

- The larger the sample size n, the narrower the confidence interval.
- The larger the standard error, the wider the confidence interval.

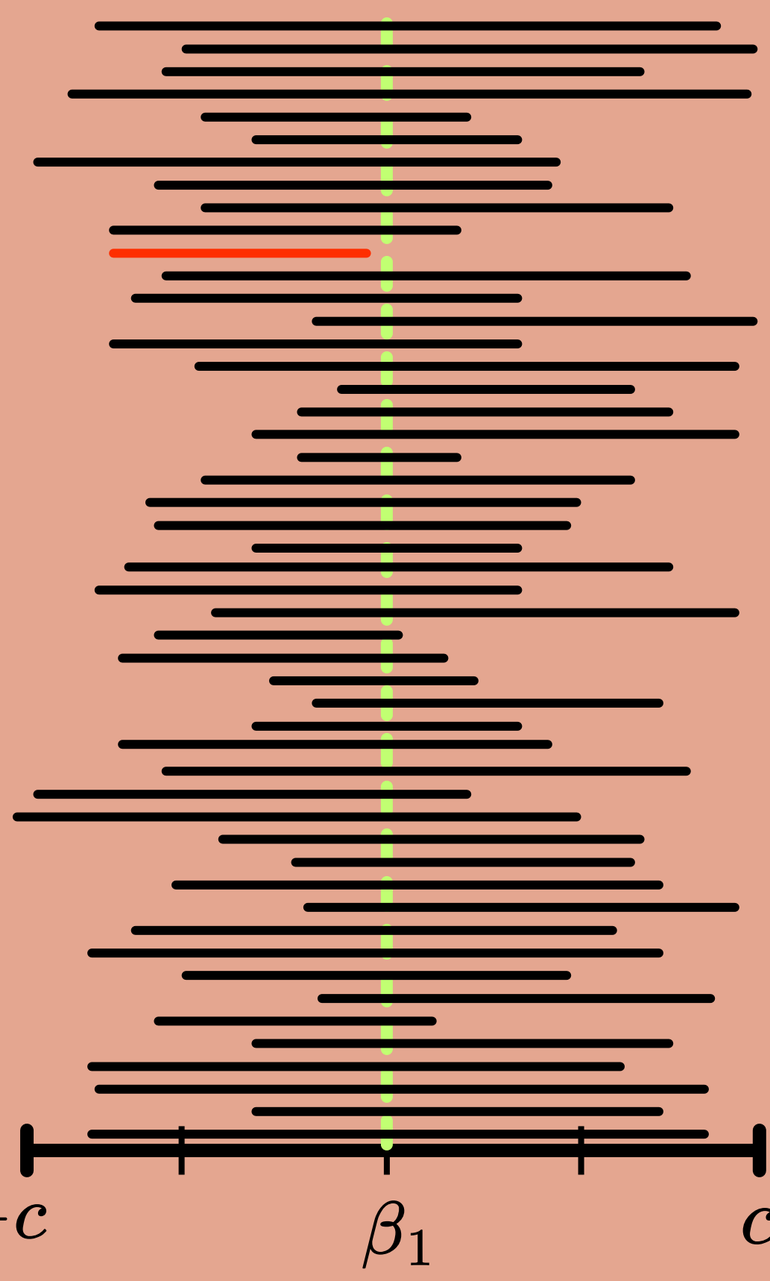
Tradeoff: To add *more accuracy* to a confidence interval it must become *less precise* with all other things equal.

Visual Representation

90% C.I of $\widetilde{\beta}_1$



98% C.I of $\widetilde{\beta}_1$



Confidence Interval Interpretation

A (1- α)100% confidence interval can be interpreted in the following way:

- **(1- α)100% of the time we are confident that the confidence interval captures the true parameter.**

To get a better understanding, look at the visual example again and notice how not all of the confidence intervals capture the true parameter β_1 and that is reflected in the confidence level.

Confidence Interval of the Mean Response

Suppose that we want to create a confidence interval for a specific point in our data. The formula:

$$\widetilde{Y}_h \pm t\left(1-\frac{\alpha}{2}; n-2\right) s\{\widetilde{Y}_h\}$$

Example (Not real data)

Suppose we go back to the housing example where we regressed housing price on the age of the house. Then we want a 95% confidence interval of the housing price for a house aged at $X_h = 30$ years old in the dataset.

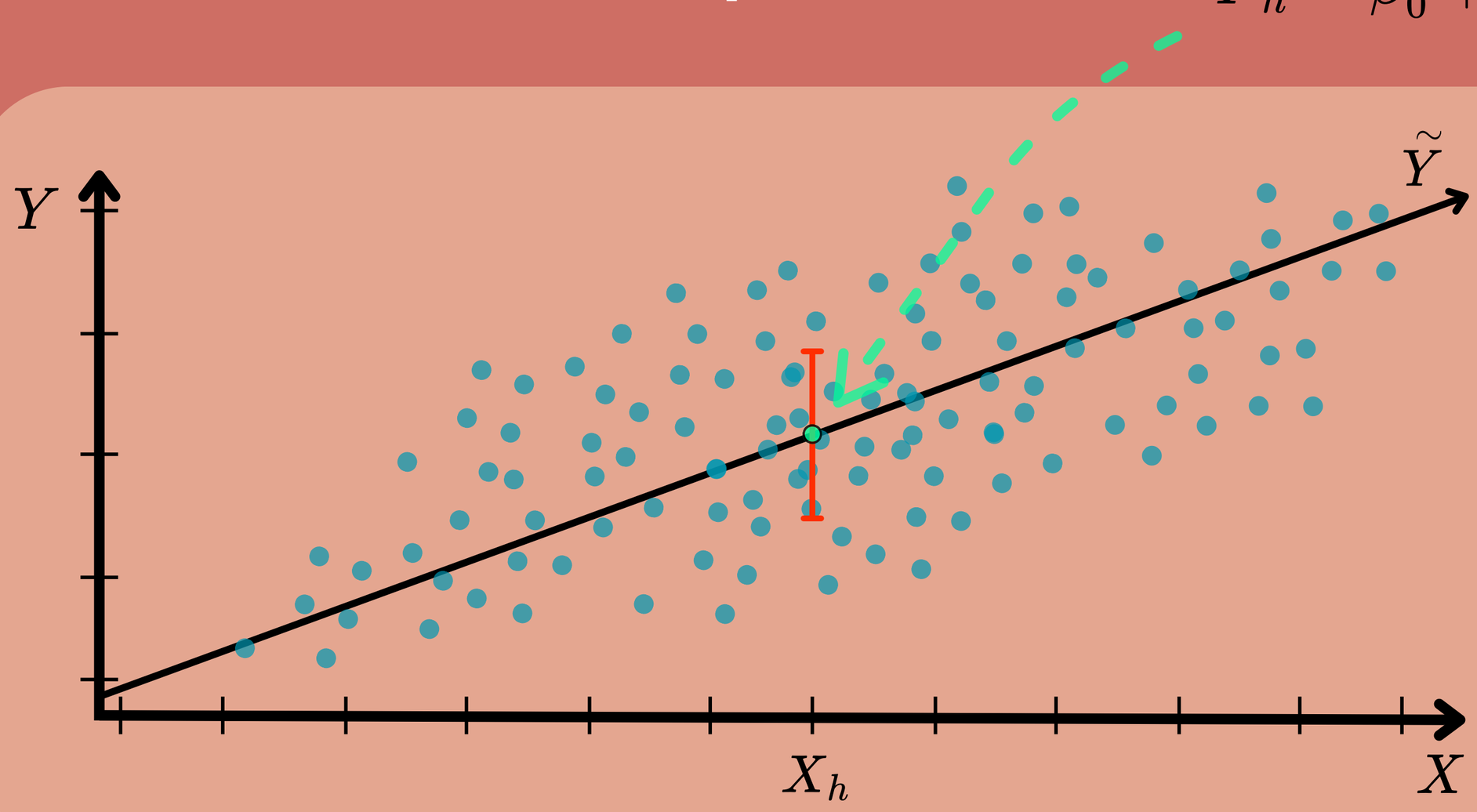
- We could say: We are 95% confident that the average home value of homes 30 years old is between [320,000,350,000].

NOTE:

- This confidence interval is for making inferences *within the scope of our dataset*.
- The farther away our choice of X_h is from the mean X, then the larger the standard error for our confidence interval.

Visual Representation

$$\widetilde{Y}_h = \widetilde{\beta}_0 + \widetilde{\beta}_1 X_h$$



REGRESSION FOR EVERYONE #6

Model Evaluation & Validation

Basic Idea: We want to understand metrics that can indicate if we have a good model with significant predictors and validate our model to make sure we are not over or underfitting.

Coefficient of Determination

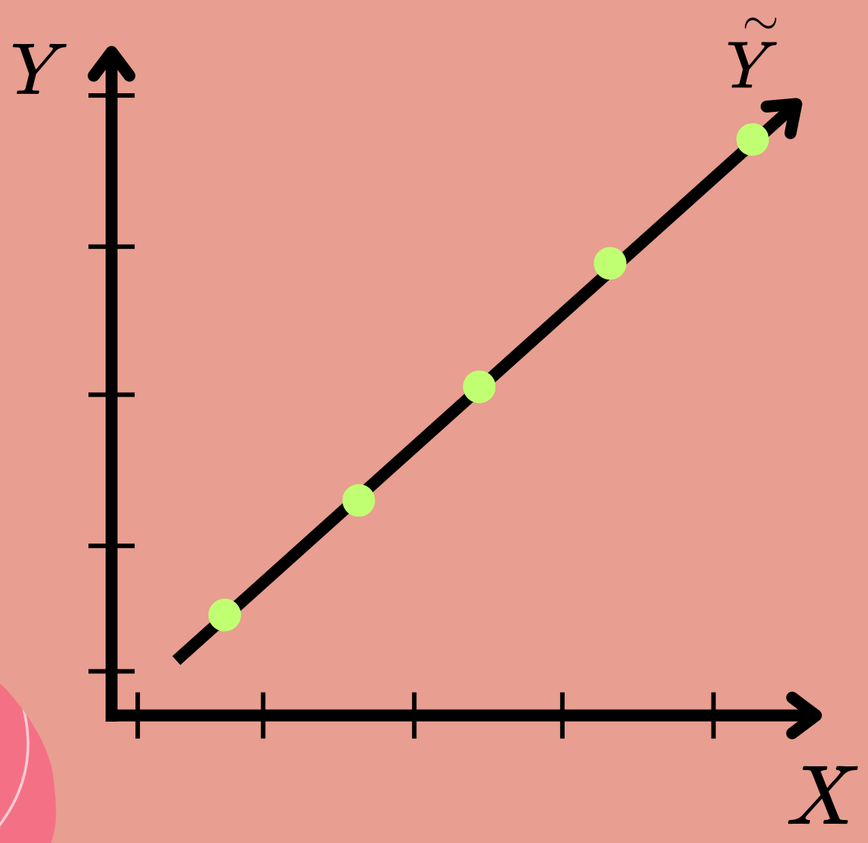
Coefficient of Determination: A descriptive measure for the linear association between X and Y.

Interpretation: Tells us the amount of variation of Y that is explained by X.

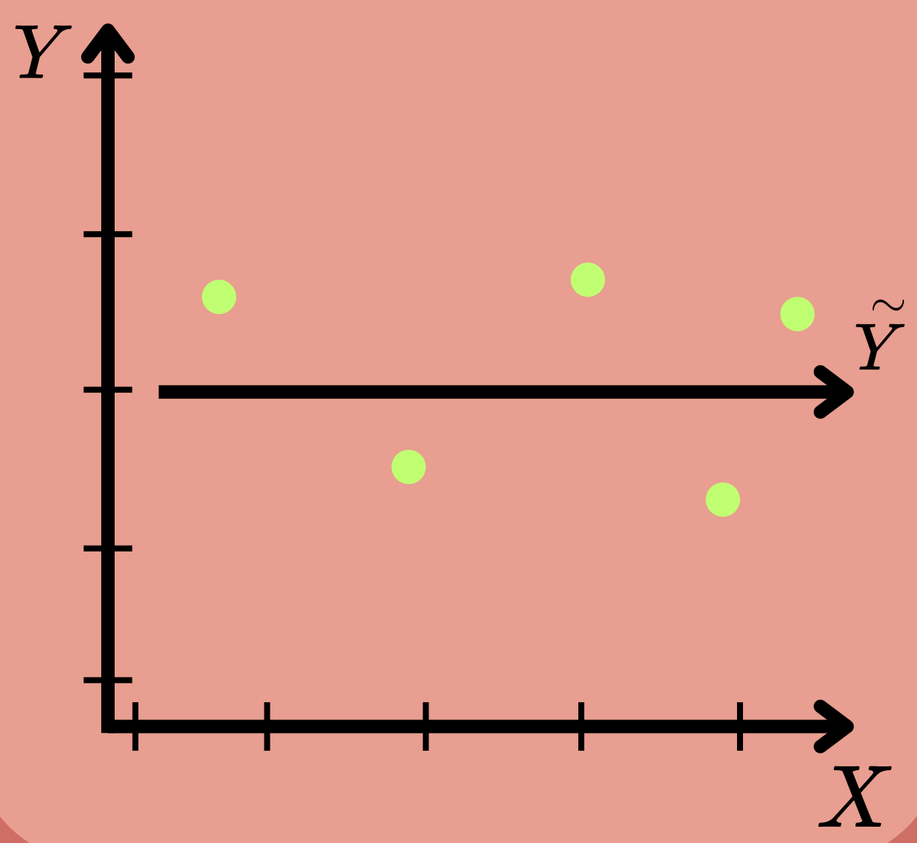
$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad 0 \leqslant R^2 \leqslant 1$$

Visual Representation

$R^2 = 1$



$R^2 = 0$



Warnings:

- When the relationship between X and Y is non-linear, R^2 is not a meaningful measure.
- A large R^2 does not necessarily mean the estimated regression line it a good fit.
- A near zero R^2 does not necessary mean that X and Y are not related.

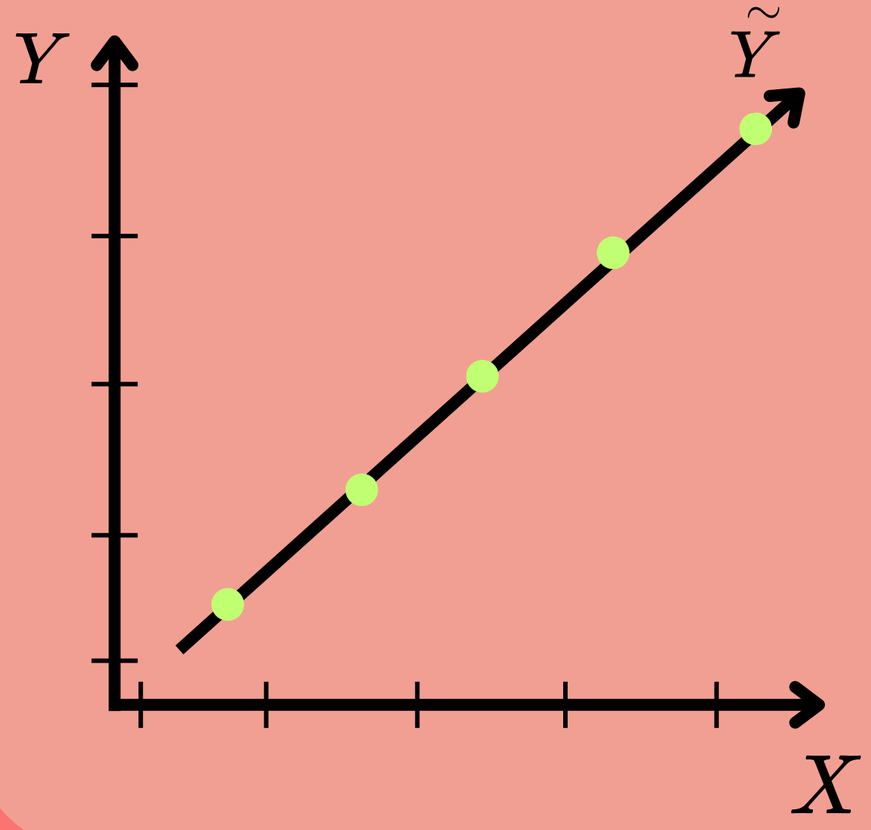
Pearsons Correlation Coefficient

A statistical measure that evaluates the strength and direction of the relationship between two variables.

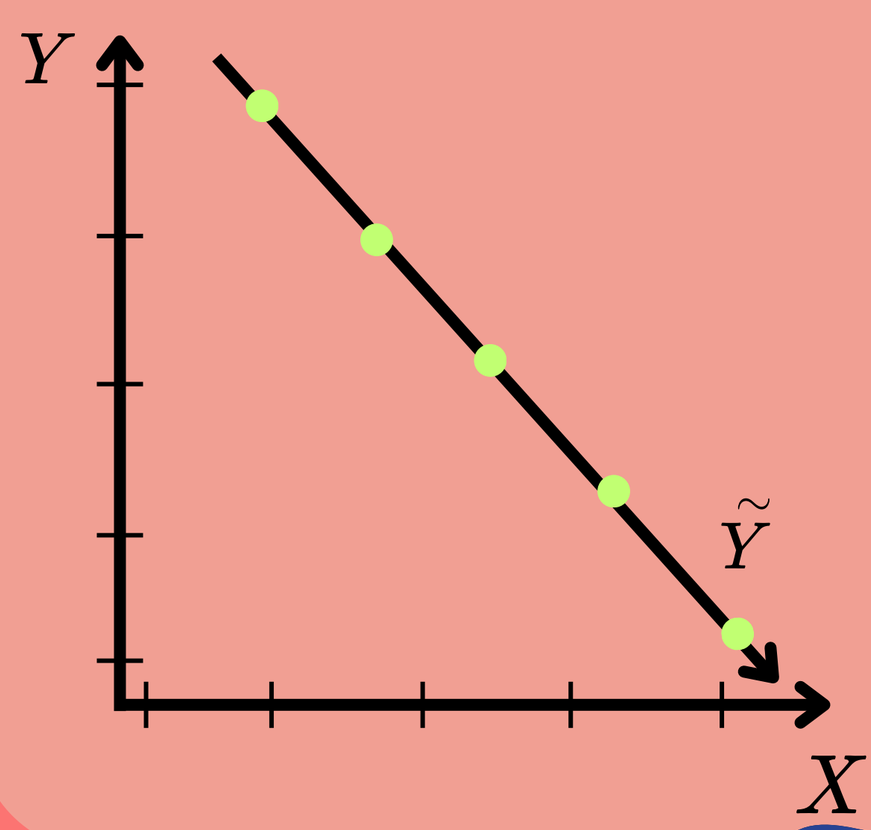
$$r = sign\{\widetilde{\beta}_1\}\sqrt{R^2} \qquad -1 \leqslant r \leqslant 1$$

Visual Representation

$r = 1$



$r = -1$



Mean Squared Error

Recall: MSE is an unbiased estimate of the true error variance.

- MSE is the averaged squared distance of from the observed to the predicted values.
- Since MSE deals with squared distances it is often harder to interpret.
- A model with a good fit should have a low MSE.

Root Mean Squared Error (RMSE)

RMSE: The square root of MSE. Measures the average distance between the predicted and actual values.

- It represents the standard deviation of residuals which is a good quantifier of how dispersed the residuals are in the model.
- A low RMSE is an indication that the model is a good fit and has precise predictions

Tests for Linear Association

These tests are testing the following hypothesis:

$$H_o : \beta_1 = 0 \qquad VS \qquad H_a : \beta_1 \neq 0$$

Testing Methods:

- T-test
- F-test

Note:

- For simple linear regression these tests are equivalent.

Interpretation: In these tests we are testing to see if the slope is significantly different from zero.

Model Validation

Model validation is a form of quality checking your model to make sure that it preforms as expected.

- **Internal validation**: Checking the validity of the model using the same data when fitted
- **External validation**: Checking the validity of the model using new or holdout data.

For a data set with a sufficiently large sample size, one option for internal validation uses *training* and *validation* (testing) data to check the model validity.

- **Training data**: Must be large enough so that a reliable model can be built. Model is trained on this data.
- **Validation data**: Often smaller in size. Use the fitted model and the new data to see how model preforms.

Note:

- The distribution of both datasets should be the same when comparing variables.

Common Methods

- **Leave one out cross validation**
- **K-fold Cross Validation**

Authors Note: Thanks for reading this far! This volume covered a simple overview of simple linear regression. The next volume will cover multiple regression and will build off of what has already been covered.