# Linear Regression

# Linear Regression

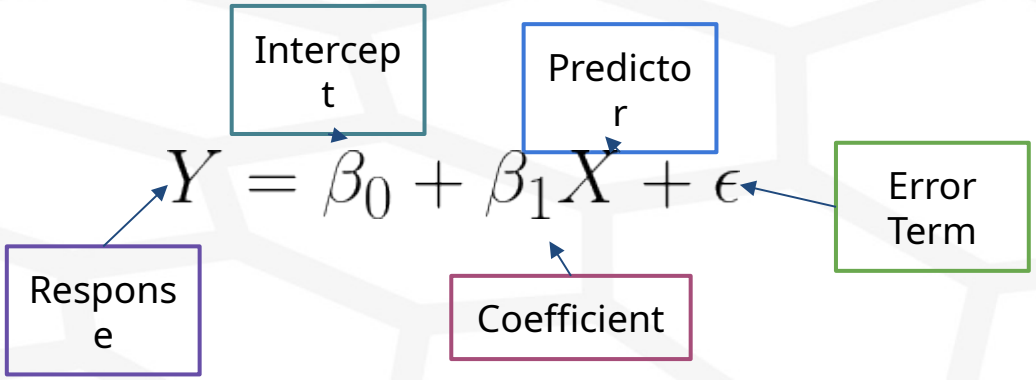
**Linear regression** is a simple form of supervised learning that involves estimating a linear fit to observed data.

Consider a case where we have data on homes in our area such as the price of the home, the square footage, the number of bedrooms, and the year it was built. Linear regression can be used to answer the following types of questions:

- Is the relationship between square footage and home price linear?
- How strong is the relationship between number of bedrooms and home price?
- How accurately can we predict future home prices?
- Which has a larger effect on home price - year it was built or number of bedrooms?

Linear regression is a simple method that is *interpretable* meaning that we can understand why the model is making its predictions.

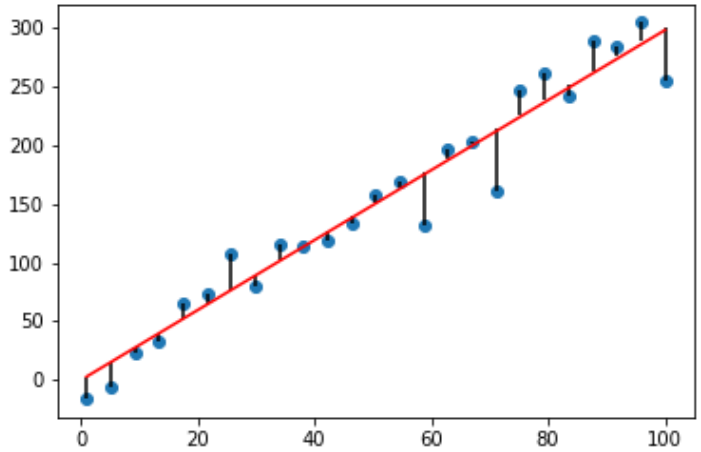
# Simple Linear Regression




$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

# How Do We Fit a Line?

We want to fit a line that is as close as possible to our data points. Linear regression uses the **least squares criterion** to measure "closeness".




Residual

$$e_i = y_i - \hat{y}_i$$

Residual Sum of Squares

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

Least squares chooses intercept & coefficients to minimize the RSS

# Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression when we have more than one predictor.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

$$Stock\ Price = Int + \beta_1(Unemployment) + \beta_2(Interest\ Rate)$$

# Polynomial & Nonlinear Regression

**Polynomial regression** can be used when there is a higher order relationship between the predictors & the response, but the relationship between predictors is linear:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \epsilon$$

**Nonlinear regression** can be used when there are nonlinear relationships between the predictors and between the predictors and the response:

$$Y = \beta_0 + \frac{\beta_1 X_1}{\beta_2 X_2} + \epsilon$$

In both cases, you must specify the form of the relationship between the predictors and the response (which can be difficult in practice).

# Implementation in Python

Both Sklearn vs Statsmodels have methods for performing linear regression.

## Sklearn

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
myfit = model.fit(X,y)
```

- Automatically estimates the intercept
- Can be used with other Sklearn methods such as “cross_val_score”
- Does not have a summary of model fit

## Statsmodels

```
import statsmodels.api as sm
X = sm.add_constant(X)
myfit = smOLS(y,X).fit()
```

- Need to add a constant to your data frame in order for statsmodels to estimate an intercept
- Can’t use Sklearn methods like “cross_val_score”
- Can return summary of model fit

# Understanding Regression Results from Statsmodels

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Stock_Index_Price | **R-squared:** | 0.898 |
| **Model:** | OLS | **Adj. R-squared:** | 0.888 |
| **Method:** | Least Squares | **F-statistic:** | 92.07 |
| **Date:** | Mon, 24 May 2021 | **Prob (F-statistic):** | 4.04e-11 |
| **Time:** | 15:00:56 | **Log-Likelihood:** | -134.61 |
| **No. Observations:** | 24 | **AIC:** | 275.2 |
| **Df Residuals:** | 21 | **BIC:** | 278.8 |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 1798.4040 | 899.248 | 2.000 | 0.059 | -71.685 | 3668.493 |
| **Interest_Rate** | 345.5401 | 111.367 | 3.103 | 0.005 | 113.940 | 577.140 |
| **Unemployment_Rate** | -250.1466 | 117.950 | -2.121 | 0.046 | -495.437 | -4.856 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 2.691 | **Durbin-Watson:** | 0.530 |
| **Prob(Omnibus):** | 0.260 | **Jarque-Bera (JB):** | 1.551 |
| **Skew:** | -0.612 | **Prob(JB):** | 0.461 |
| **Kurtosis:** | 3.226 | **Cond. No.** | 394. |

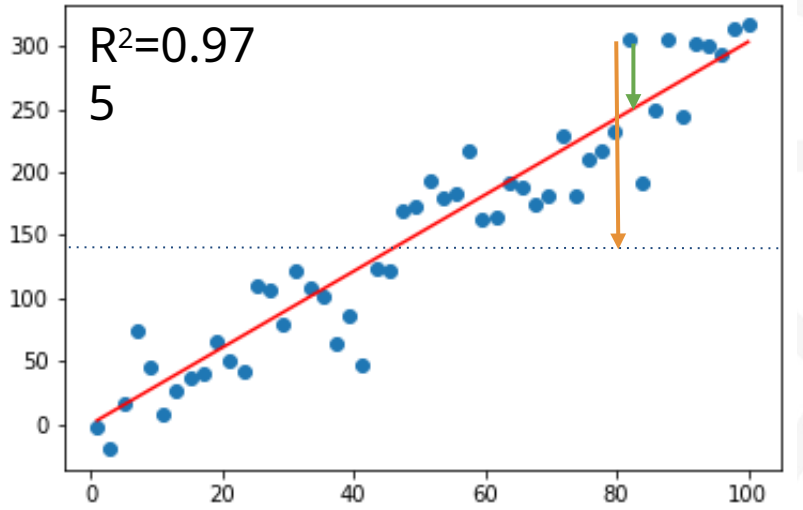
https://www.datarobot.com/blog/ordinary-least-squares-in-python/

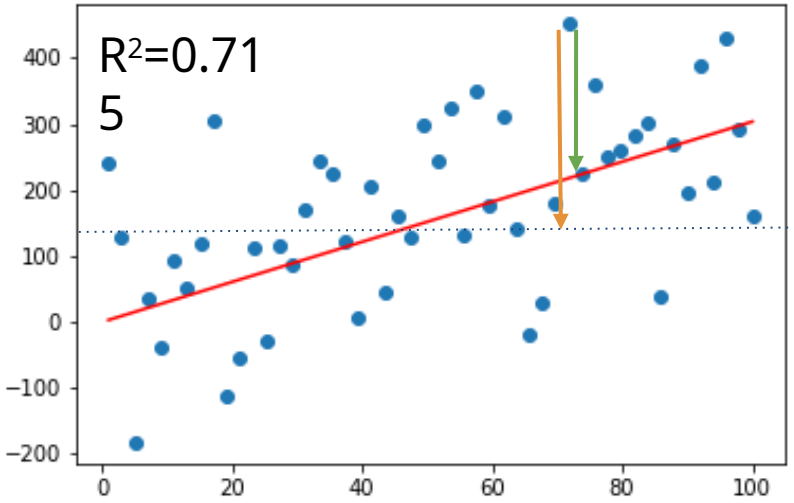
# Assessing Our Model

A common way to assess the fit of our model is using the $R^2$ statistic (called the coefficient of determination).

$R^2$ measures the proportion of variability in Y that can be explained using X.

$$R^2 = 1 - \frac{RSS}{TSS} \qquad TSS = \sum (y_i - \bar{y})^2$$



# Regression Assumptions

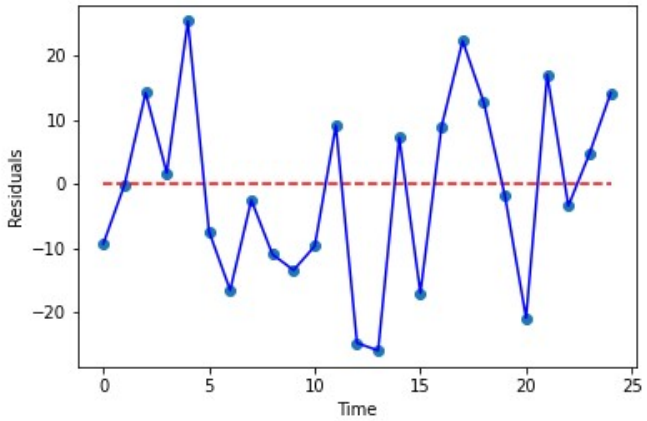
| Assumption | Description | Solution if Assumption is Not Met |
|---|---|---|
| **Linear relationship** | The relationship between the predictors and the response is (approximately) linear. | Use polynomial regression, nonlinear regression or a more flexible ML method. |
| **Independent residuals** | Error terms are independent (no correlation). Error terms are often dependent in time series data. | Plot residuals across time. If there is a pattern, more advanced methods may be required to adjust for this. |
| **Constant variance of residuals** | Variance of the residuals should not change over time (“cone shaped residuals”) | We may need to transform the data. |
| **No Collinearity** | Two or more predictors are closely related. | Look at a correlation matrix. You may decide to drop one of the predictors or combine both predictors into one. Note: there may be higher order correlations not visible on the correlation matrix. |
| **Residuals are normally distributed** | Residuals should be normally distributed. | We may need to transform data or check for outliers. |

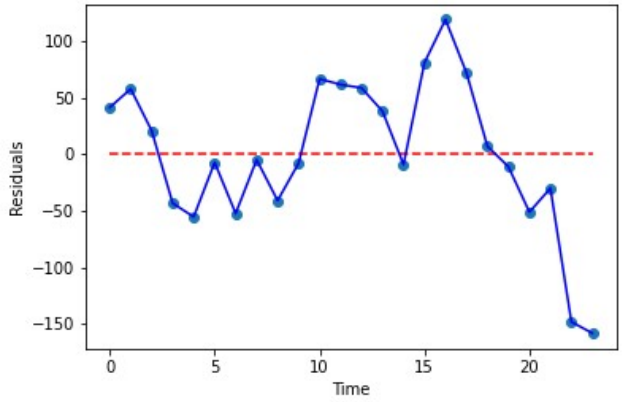
# Residual Analysis - Residuals vs Time

Residual analysis should be performed to determine if assumptions are met.

*"The problem is that checking the quality of the model is often a less prioritized aspect of a data science task flow where other priorities dominate — prediction, scaling, deployment, and model tuning."*

**Residuals vs time plot** allows us to see if the residuals are independent.

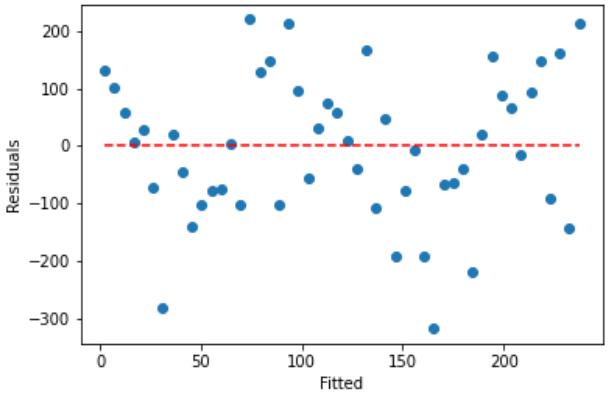


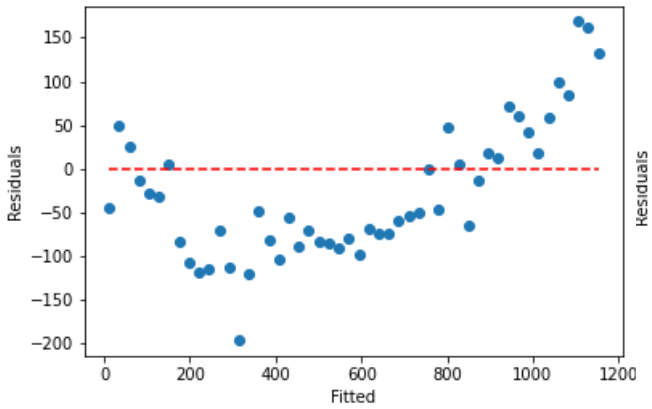
# Residual Analysis - Residuals vs Fitted

**Residuals vs fitted values plot** allows us to see if there are is non-linearity and if the error terms have constant variance.

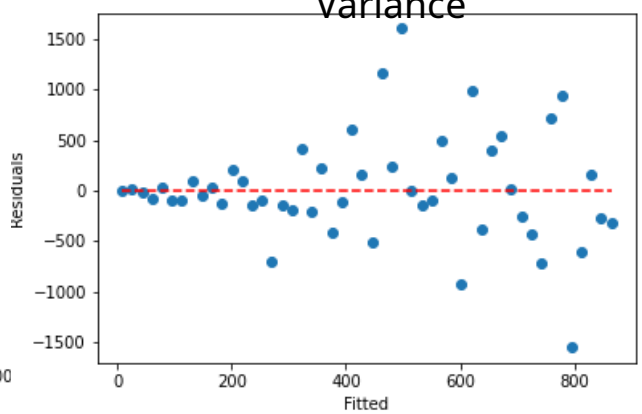
Good




Nonlinear

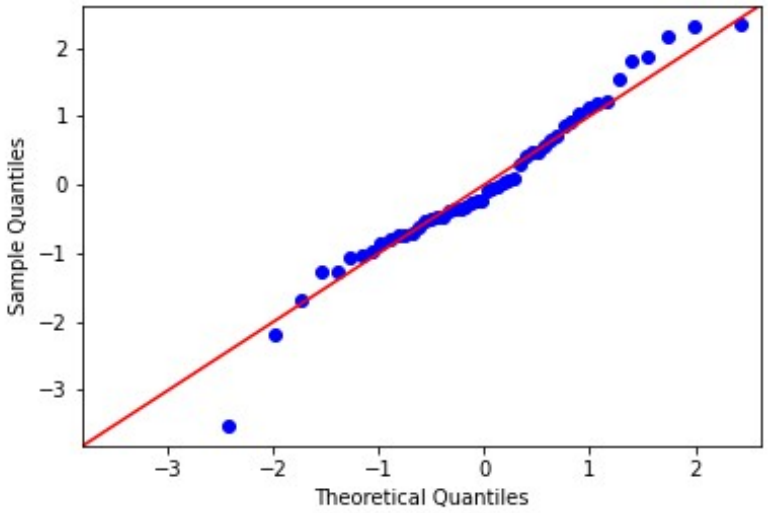


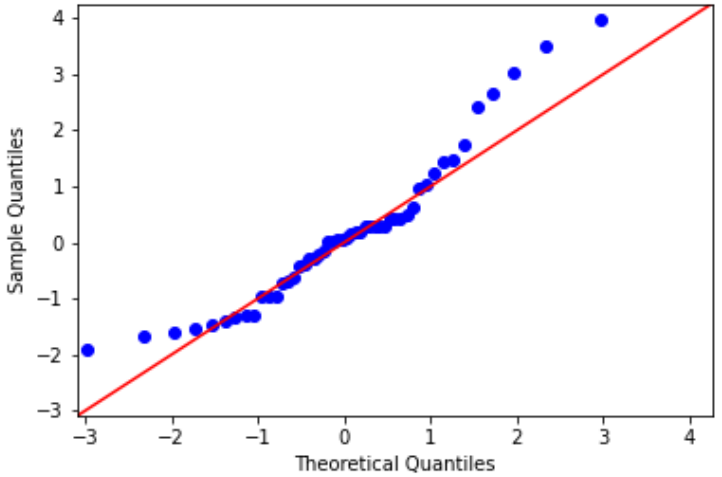

Nonconstant Variance



# Residual Analysis - QQ Plots

**QQ plots** plot sample quantiles vs theoretical quantiles of a normal distribution to determine if the residuals are (approximately) normally distributed.



# Other Things to Consider

- **Outliers** - linear regression can be very sensitive to outliers. Scatter and distribution plots of the data can help identify outliers. Remember, only remove outliers if you have a valid reason!

- **Multicollinearity** - pair plots and correlation maps are a good way to identify highly correlated predictors. If two predictors are correlated, you can remove one predictor or combine both predictors into one.

# Pros & Cons

## Pros

- Simple and easy to interpret
- Performs well when assumption of linearity holds
- There are methods to prevent overfitting (we will discuss these later on)

## Cons

- Assumes a linear relationship between predictors and response
- Sensitive to outliers and multicollinearity