



Regression Analysis: Fundamentals and Practical Applications

Course Learning Objectives



Learn what linear regression is and how to use it to make real world predictions



Learn how to perform simple regression calculations in Excel & RegressIt



Create linear regression models in Python using both statsmodels and sklearn modules



Understand the implicit assumptions behind linear regression



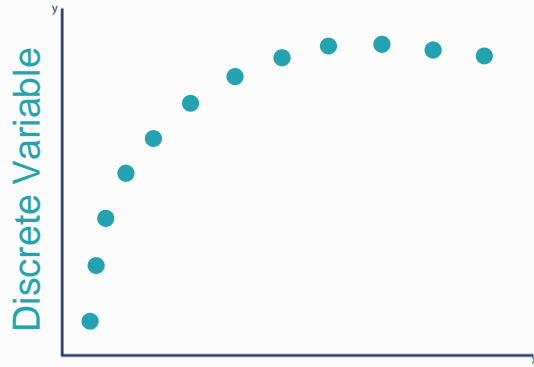
Be able to interpret regression coefficients, p-values and other metrics to evaluate a model



Become familiar with more advanced regression techniques and when to use them

What is Regression?

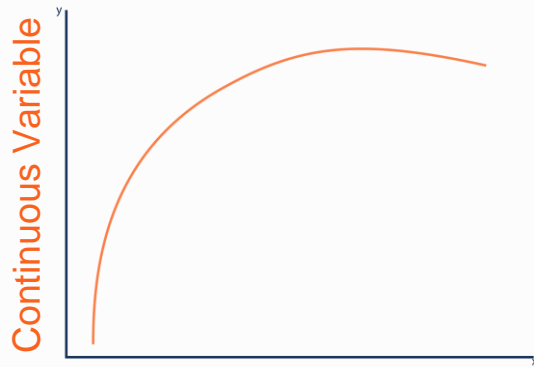
Regression refers to a specific type of Machine Learning model, used to make predictions.



Discrete variables

Can only take certain values.

For **discrete predictions**, we use **classification**.



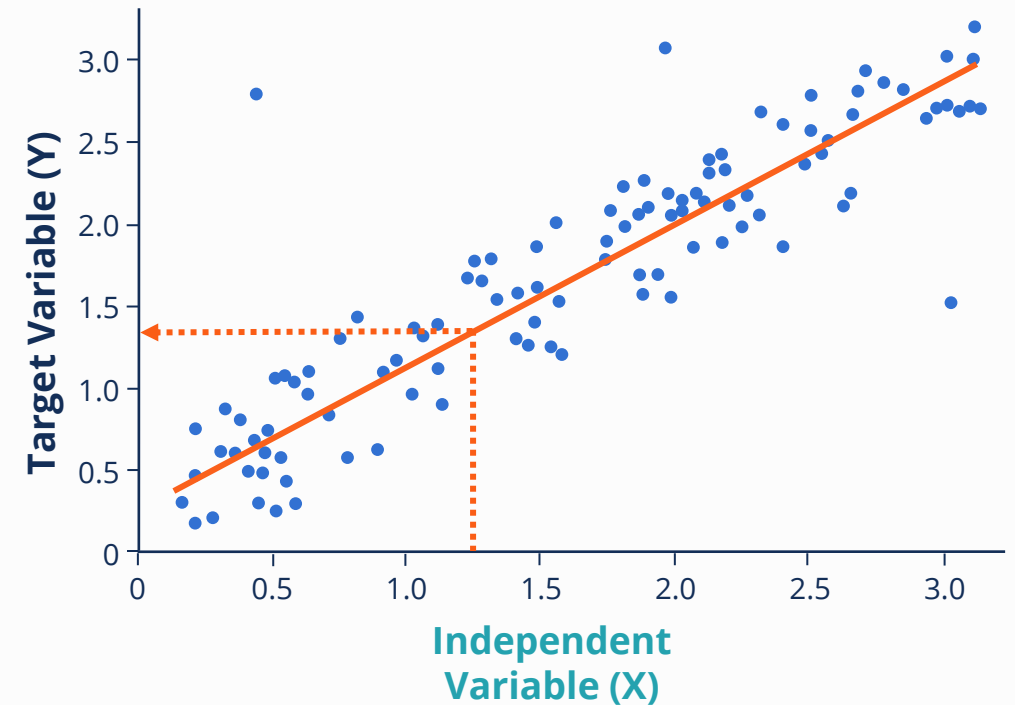
Continuous variables

Can take any value along a continuous scale.

For **continuous predictions** we use **regression**.

Regression Definitions

- The **target variable** is known as the **Y variable**.
- Input data is known as **independent or X** variable.
- We plot our **sample data points** with known X and Y.
- Then we plot a **line of best fit** to help us **make predictions** when X is known, but the target Y is unknown.

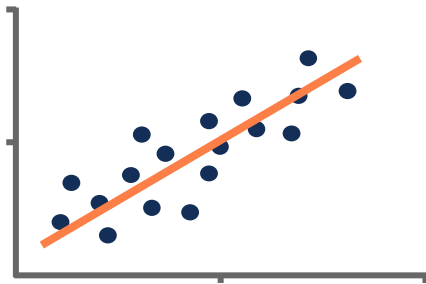


Regression can be applied to a wide range of problems where we want to predict a continuous value such as revenue, costs, life expectancy or film review scores.

Types of Linear Regression

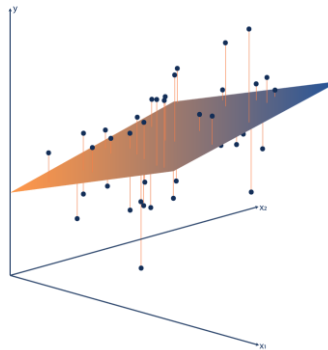
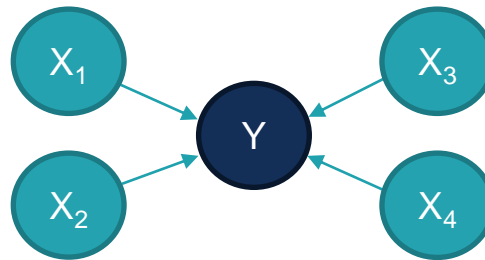
Simple Linear Regression

Target predicted using only **one** independent variable



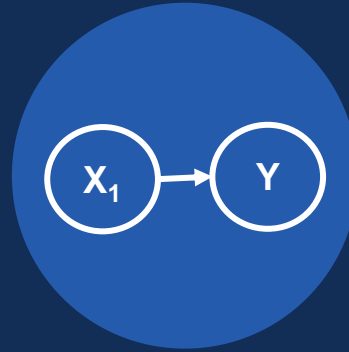
Multiple Linear Regression

Target predicted using **multiple** independent variables



Non-Linear Regression

Target predicted using **multiple** independent variables



Simple Linear Regression

Simple Linear Regression



Break linear regression down into its simplest form, including lines of best fit and errors.



Calculate the SSE (sum of squared errors) to summarize total error in the model.



Learn how R^2 can be used to explain the performance of the model.



Be comfortable using regression terminology in conversation.



Understand basic limitations of a linear regression model.

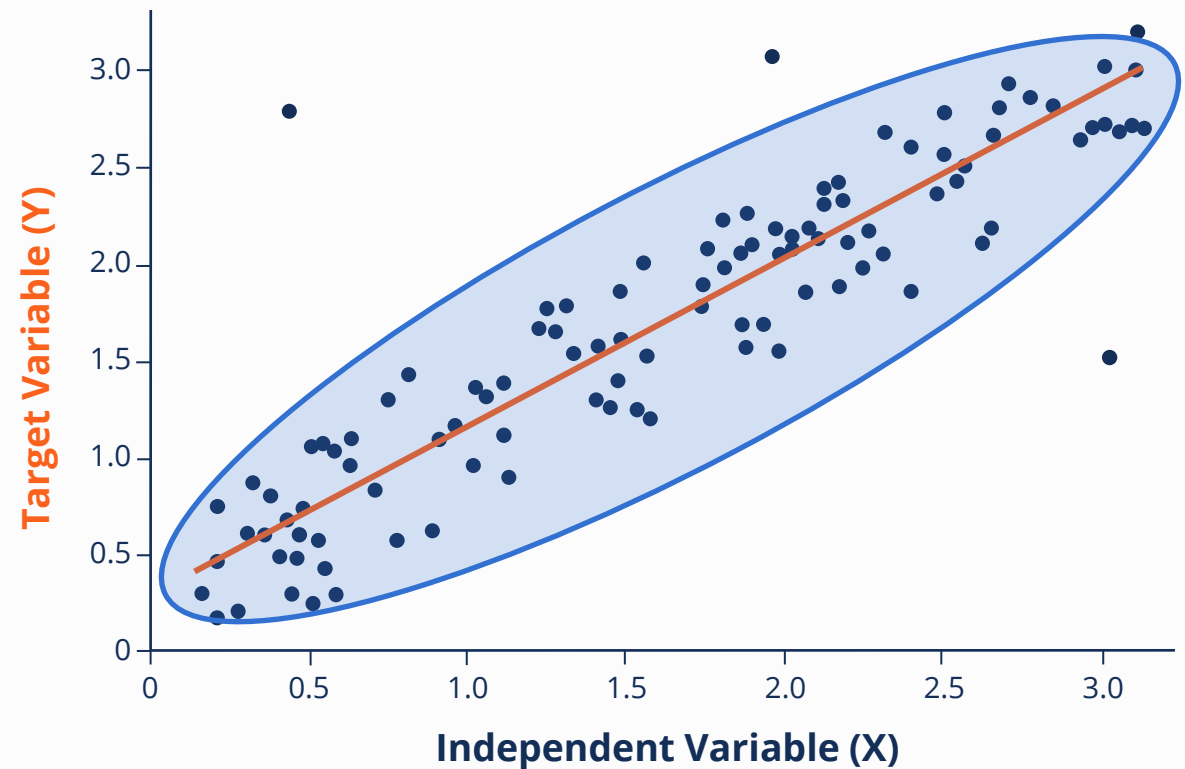


Practice a basic regression scenario in Excel and Python.

Simple Linear Regression

The aim of simple linear regression is to fit a straight-line relationship between **two variables, X and Y**.

- We summarize the data points with the **line of best fit**.
- Use the line of best fit to **predict y-values** for any x-values
- Some random variation **cannot be explained**.

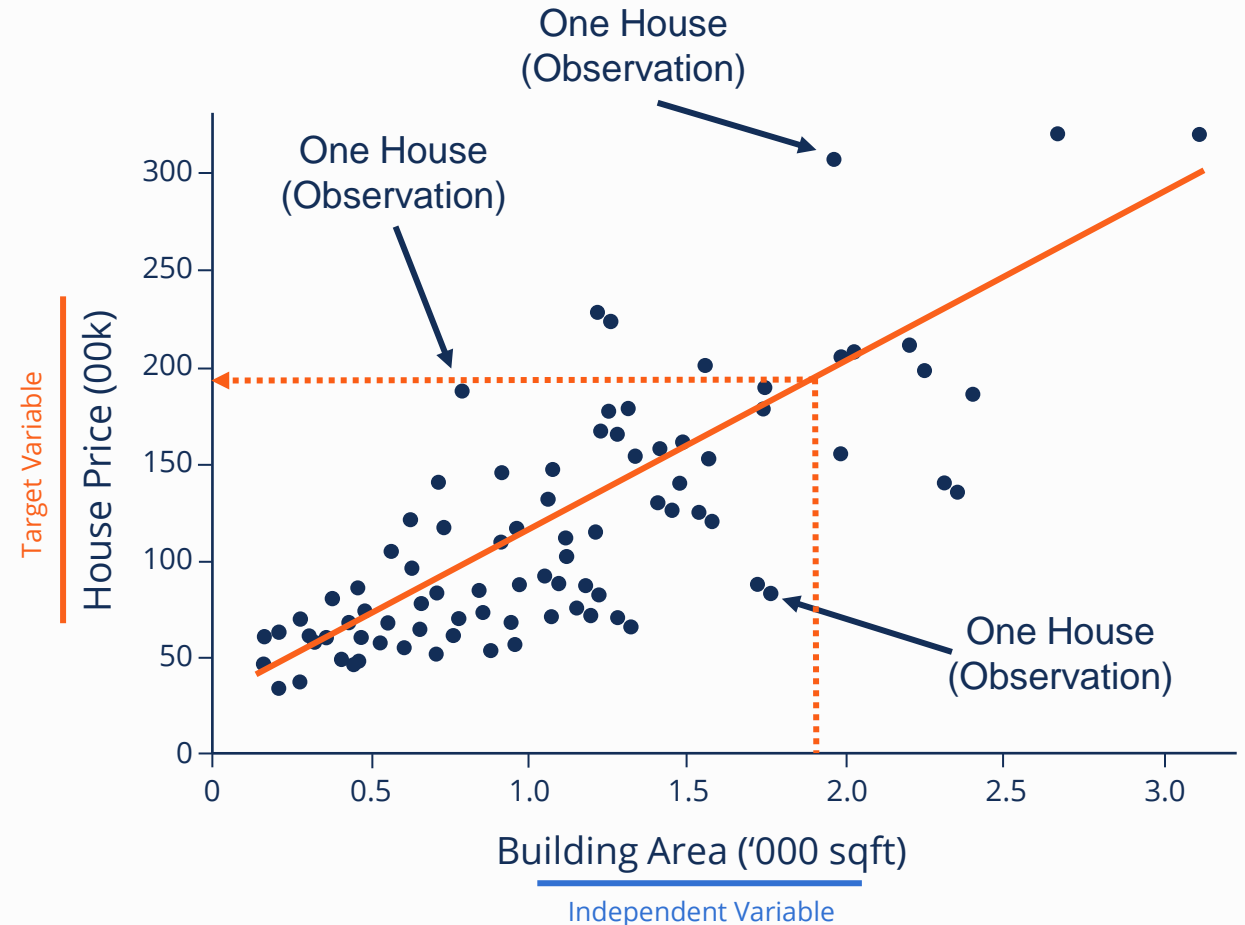


House Price Prediction

- We can use building area to predict house prices
 - **Independent Variable:** Building Area (Size)
 - **Target Variable:** House Price
- Each blue point is a single house with a known area and corresponding house price
- From the **line of best fit** we can predict house prices for any house building area

Additional Complexity

- House prices are determined by more than just the area of the house
- Multiple Linear Regression can be used to include other variables.

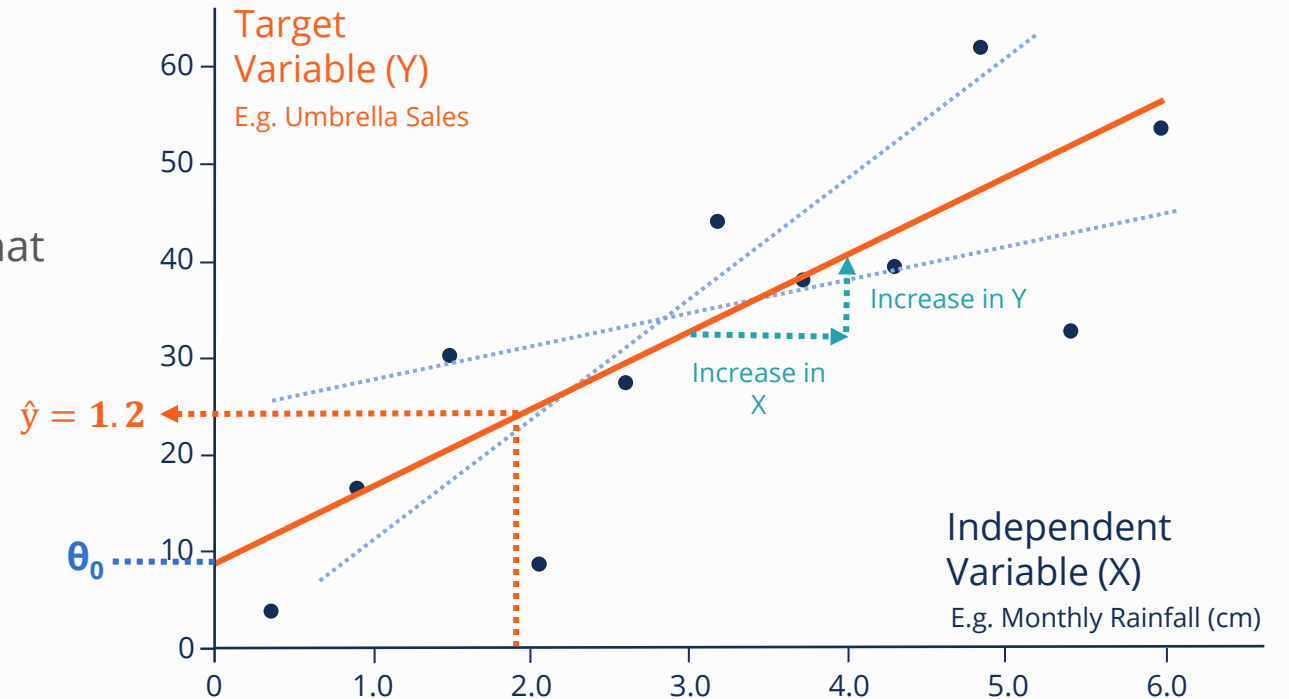


Linear Regression Algorithm

- The line of best fit can be described by:

$$\hat{y} = \theta_1 x + \theta_0$$

- \hat{y} is our predicted value
- x is the independent variable
- θ_0 and θ_1 are the **parameters** (or **coefficients**) that define the regression line.
 - θ_0 represents the intercept
 - θ_1 represents the slope



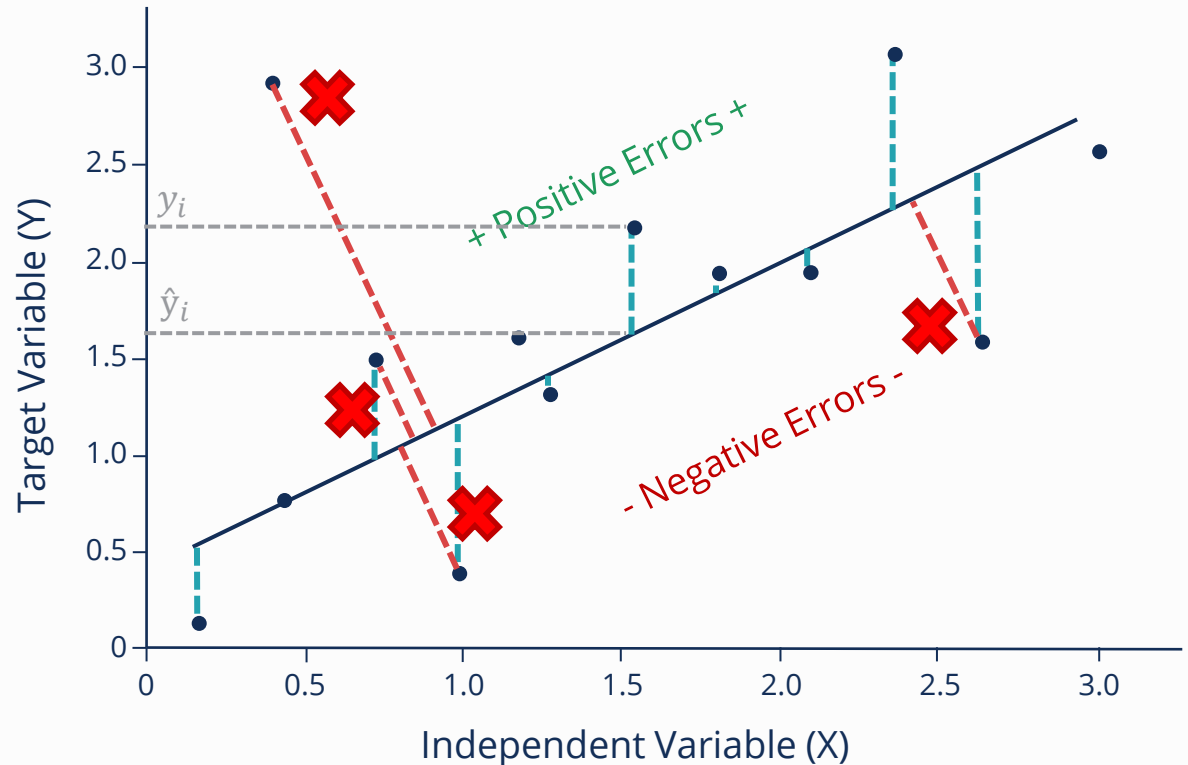
To find the optimal line, we can use **Ordinary Least Squares**

Ordinary Least Squares

- Ordinary Least Squares method finds the best fitting line by minimizing the **sum of square errors** (SSE) also known as the **sum of square residuals**
- **Errors**, or residuals are the difference between the observed and predicted values of the data (vertical lines on chart)
- In OLS, we **square each error** to remove negative values. We then add all squared errors together.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

- When the SSE is smaller, the model fits the data better overall.



Ordinary Least Squares Calculation

- If the orange line has the following equation:

$$\hat{y} = 0.7x + 0.45$$

- Then the **errors** can be calculated as:

$$y_i - \hat{y}_i = y_i - ((0.7 \times x_i) + 0.45) = \text{error}$$

$$y_1 - \hat{y}_1 = 0.10 - ((0.7 \times 0.2) + 0.45) = -0.49$$

$$y_2 - \hat{y}_2 = 0.75 - ((0.7 \times 0.4) + 0.45) = +0.02$$

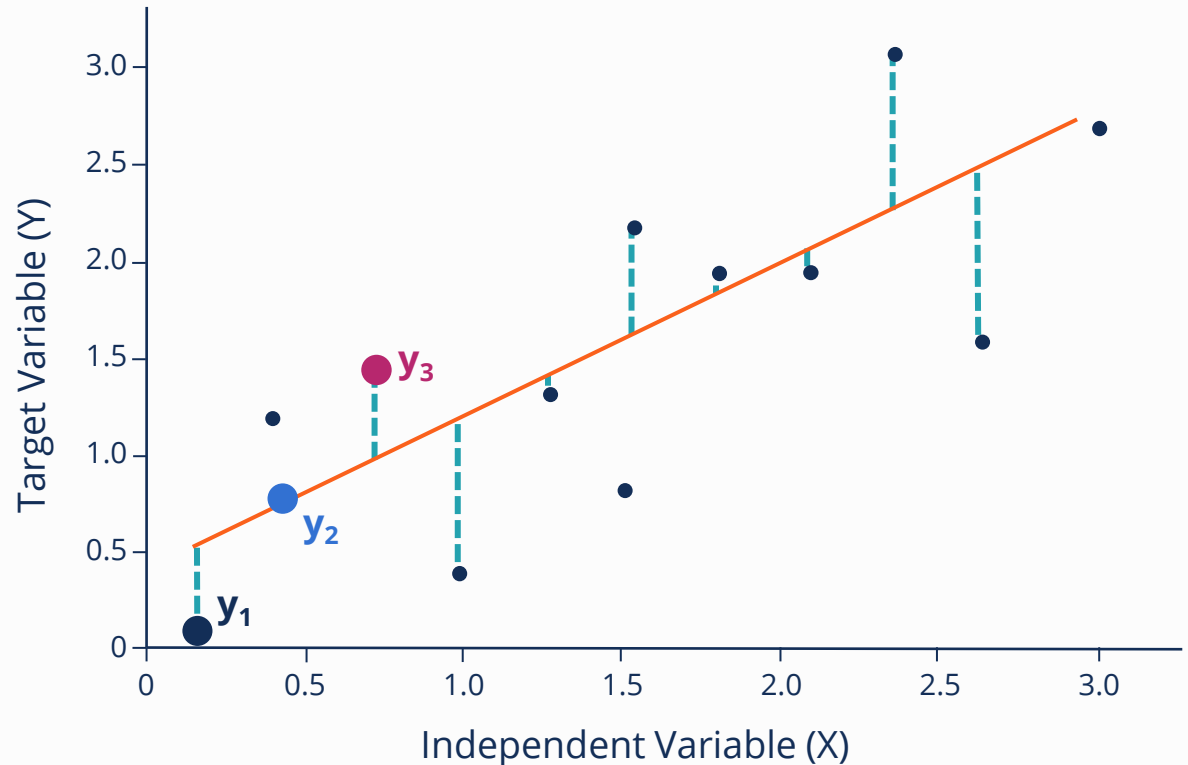
$$y_3 - \hat{y}_3 = 1.50 - ((0.7 \times 0.7) + 0.45) = +0.56$$

... and so on

- The sum of square errors for these points is then:

$$SSE = (-0.49)^2 + 0.02^2 + 0.56^2 + \dots$$

- Summing over all the points gives us an SSE of 4.02



Fitting the Parameters


- The goal is to **minimize the total error (SSE)** produced by the line of best fit.
- To find the optimal parameters, we can:
 1. Calculate the derivatives of the sum of the square errors
 2. Evaluate these where they equal zero

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Intercept

$$\theta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

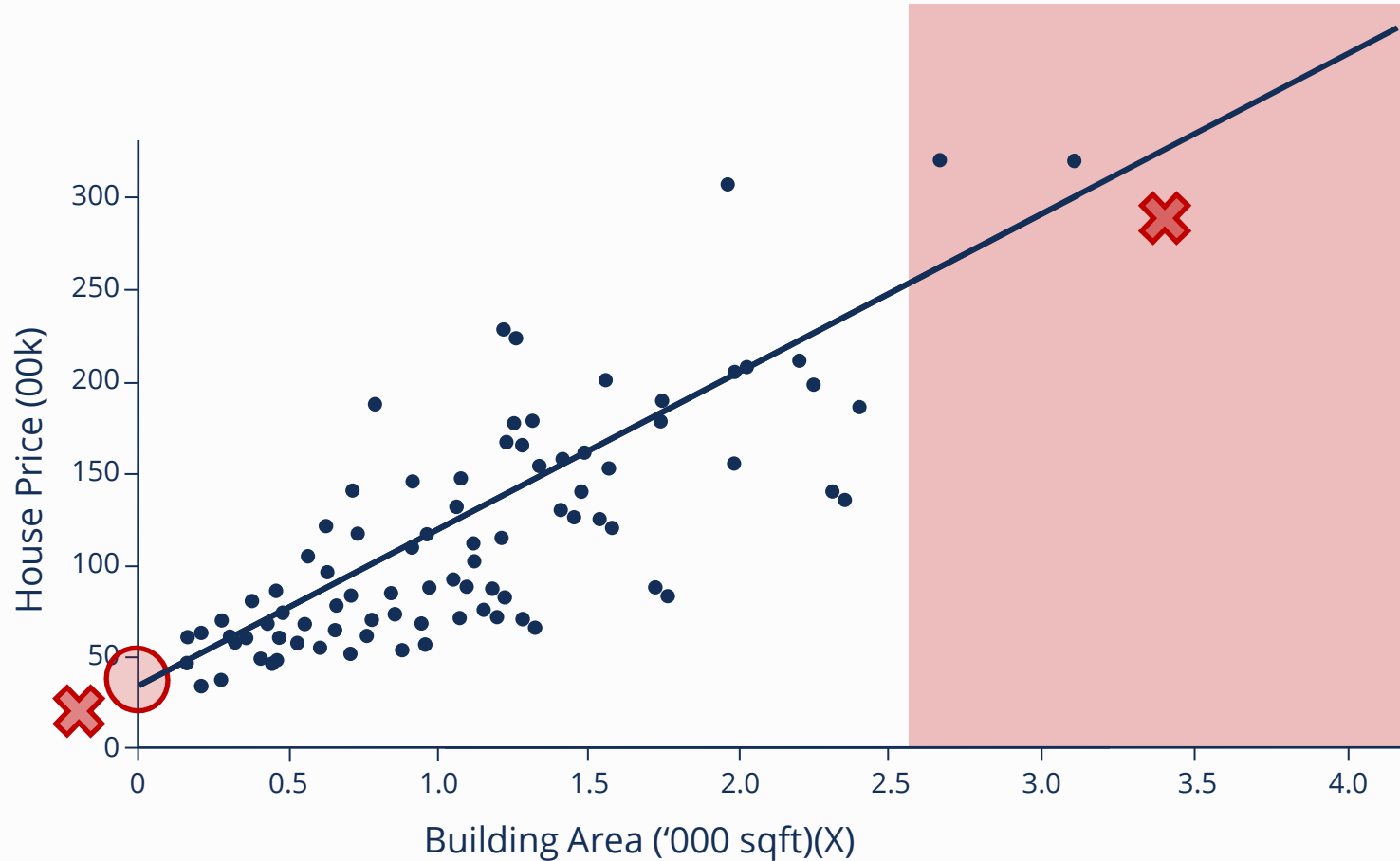
Slope



- \bar{x} and \bar{y} are the **averages of the all the observed x_i and y_i values.**
- Different values of θ_0 and θ_1 give different predictions and hence different values for the SSE.

Caution - Extrapolation

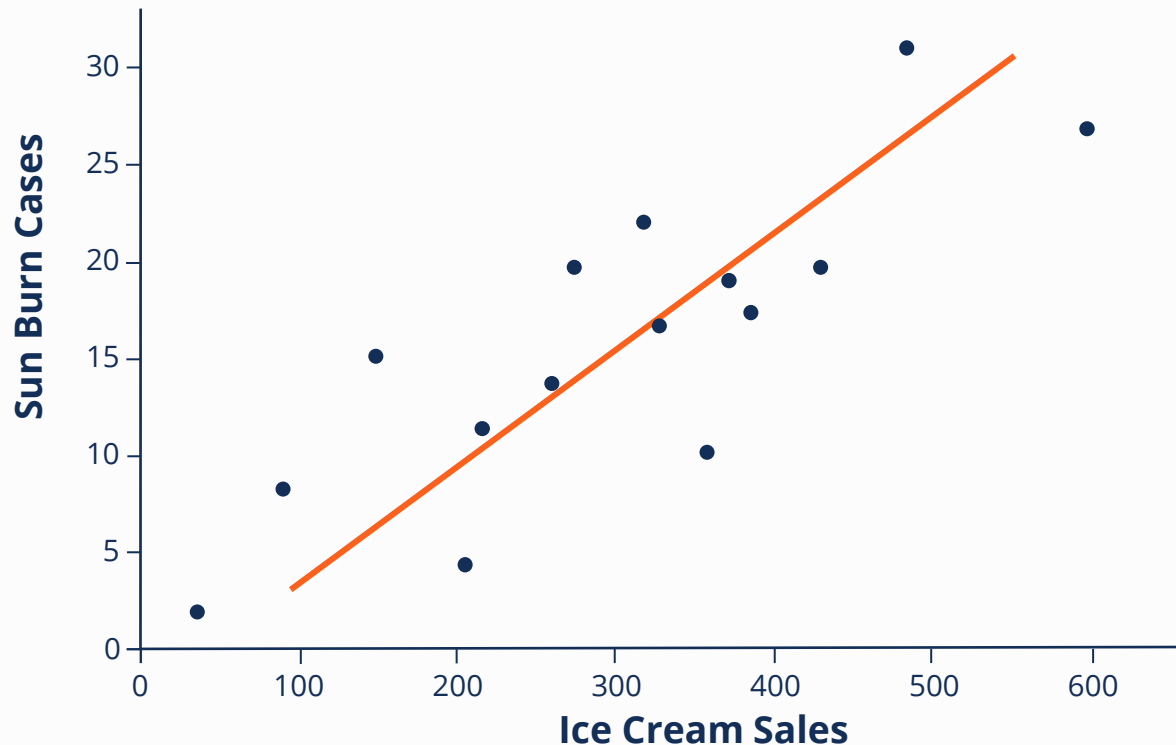
We should be careful not to extrapolate our results beyond the sample range of data.



- The model may perform poorly.
- A house of zero area still has a positive value. This prediction (intercept) is not reliable.
- Looking beyond the range of observations may lead us to false conclusions.
- **Conclusion:** The model is only reliable for the observed sample space.

Caution – Correlation Vs Causation

A relationship or **correlation does not necessarily mean causation**.



- Ice cream sales are **correlated** to the number of sun burn cases
- Ice cream **does not cause** not burn
- In fact, the **weather (sun) causes both**.
- Additionally, we can find many correlated variables with no real relationship. These are called Spurious Regressions.

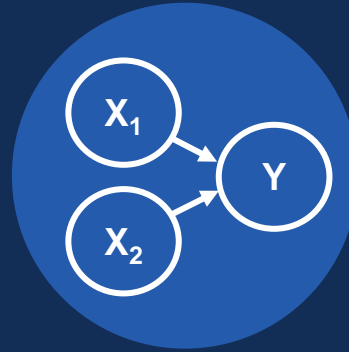
We should carefully consider if our independent variable is truly **causing** the effect on the target variable.

Linear Regression in Practice

- Ordinary least squares will **always produce the same results** (analytic algorithm)
- We must **test our model thoroughly** using new data, before applying it to real world decisions.

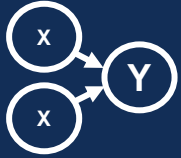


- Using code we can automate a lot of regression calculations.
- We will use **StatsModels and Scikit-Learn in Python** to practice our skills.
- We will also **fit a regression model in Excel**.



Multiple Linear Regression

Multiple Linear Regression



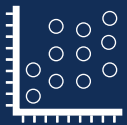
Introduce and define Multiple Linear Regression



Compare Multiple Linear Regression to Simple Linear Regression



Explore the parameters of Multiple Linear Regression



Introduce Multicollinearity and why it is important to Multiple Linear Regression



Identify how the common issue of overfitting a Multiple Linear Regression model can occur



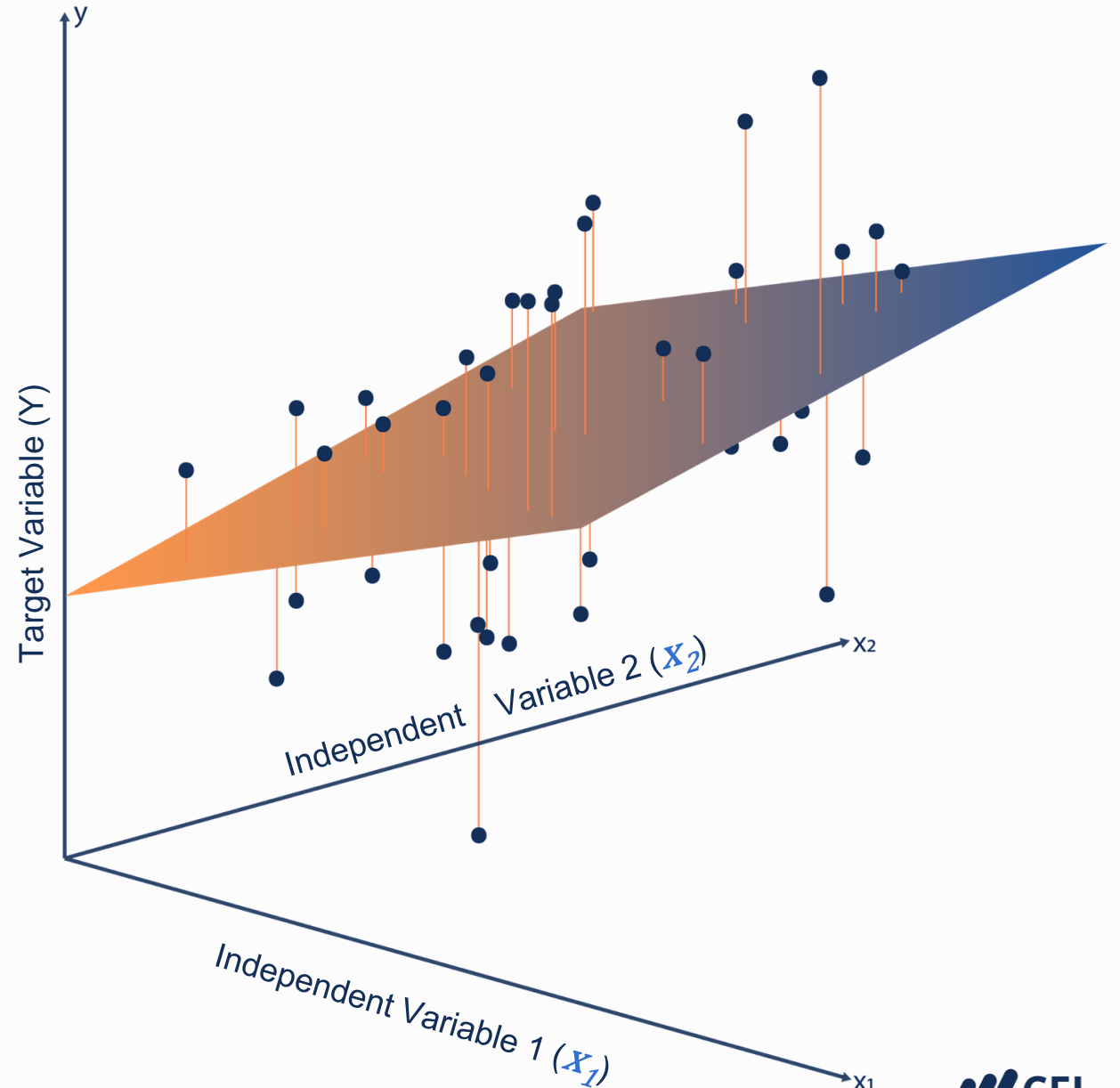
Practice an in-depth Multiple Linear Regression scenario in Python

What is Multiple Linear Regression?

- It is rare that a single independent variable fully describes the variation in the target variable
- **Multiple linear regression** allows us to predict a target variable using **any number of independent variables**

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p$$

- When we have two independent variables, we are fitting a plane to the data instead of a straight line



How the Algorithm Differs from Simple Linear Regression

- When we have one input variable, we have two parameters: slope and intercept

$$\hat{y} = \theta_1 x + \theta_0$$

- θ_0 and θ_1 are the **parameters** that define the regression line.

- θ_0 represents the intercept

- θ_1 represents the slope

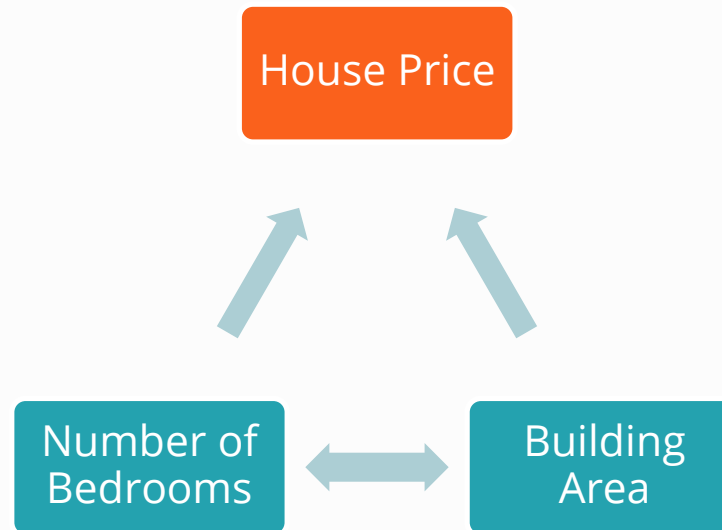
- For p independent variables plus an intercept, we have a total of $p + 1$ parameters

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

- We use ordinary least squares again and solve $p + 1$ simultaneous equations using matrix algebra for the best fit parameter values

Multicollinearity

- **Multicollinearity** occurs when our input variables are strongly correlated with each other



- With multicollinear variables, the algorithm would be unable to separate the effects of these two variables
- There's a high chance multicollinear variables will cause errors

Caution - Multiple Linear Regression

- Adding uncorrelated independent variables to the model often helps better explain the target variable but must be done with caution
- If a new variable is added that has **no predictive value**, the algorithm will calculate that and assign a **parameter value close to zero**

$$\hat{y} = \theta_0 + 0.001x_1 + \theta_2x_2 + \cdots + \theta_px_p$$

- **Overfitting** can occur when the model captures too much of the detail in the training data that doesn't exist in the test data
- This can be caused by random effects in the data reducing the SSE and hence being detected by the algorithm
- It is best to balance the number of independent variables so that we can adequately describe the data without overfitting



Interpreting Linear Regression

Interpreting Linear Regression



Review the concepts of residuals and residual plots



Analyze datasets to ensure they meet the Ordinary Least Squares assumptions



Evaluate Linear Regression metrics to measure the error in models



Analyze model output with Linear Regression coefficients & p-values



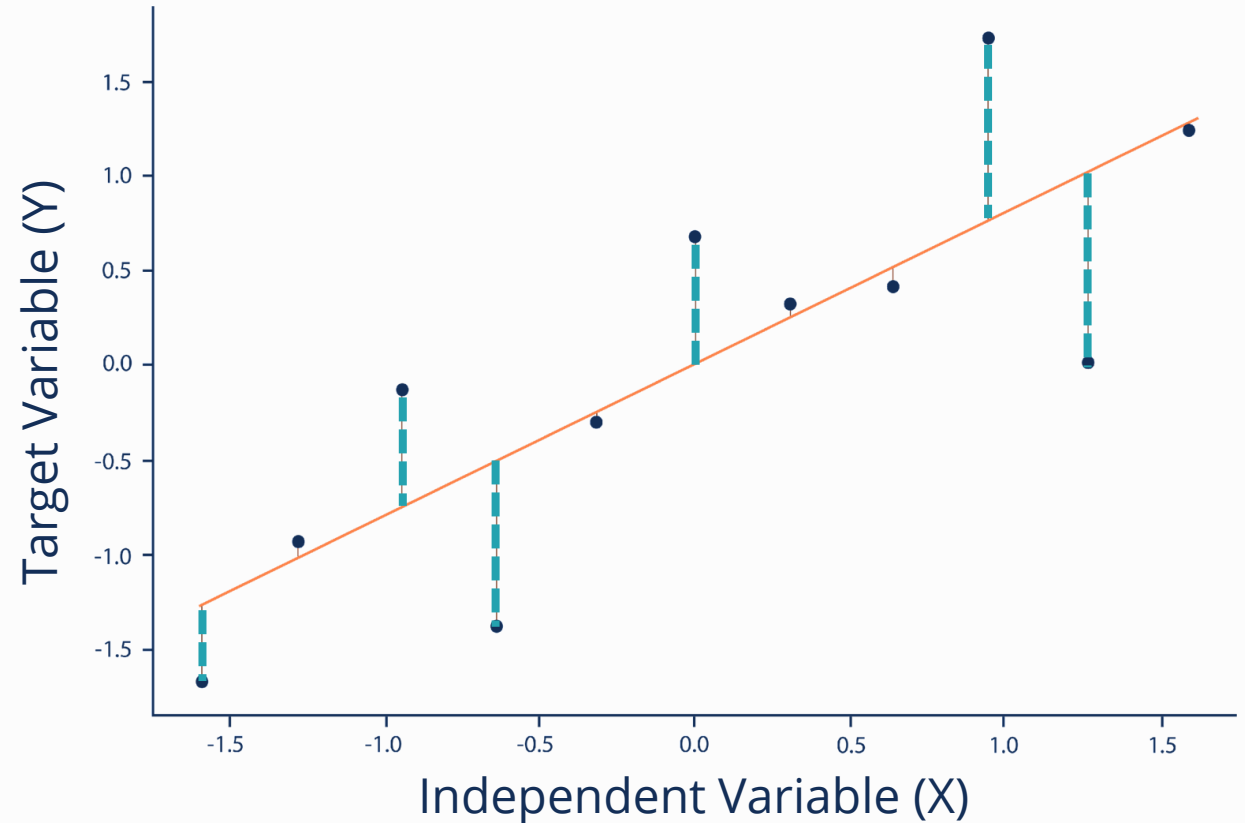
Apply knowledge with an interactive exercise and interpretation scenarios



Practice interpreting the results of a Linear Regression model in Python

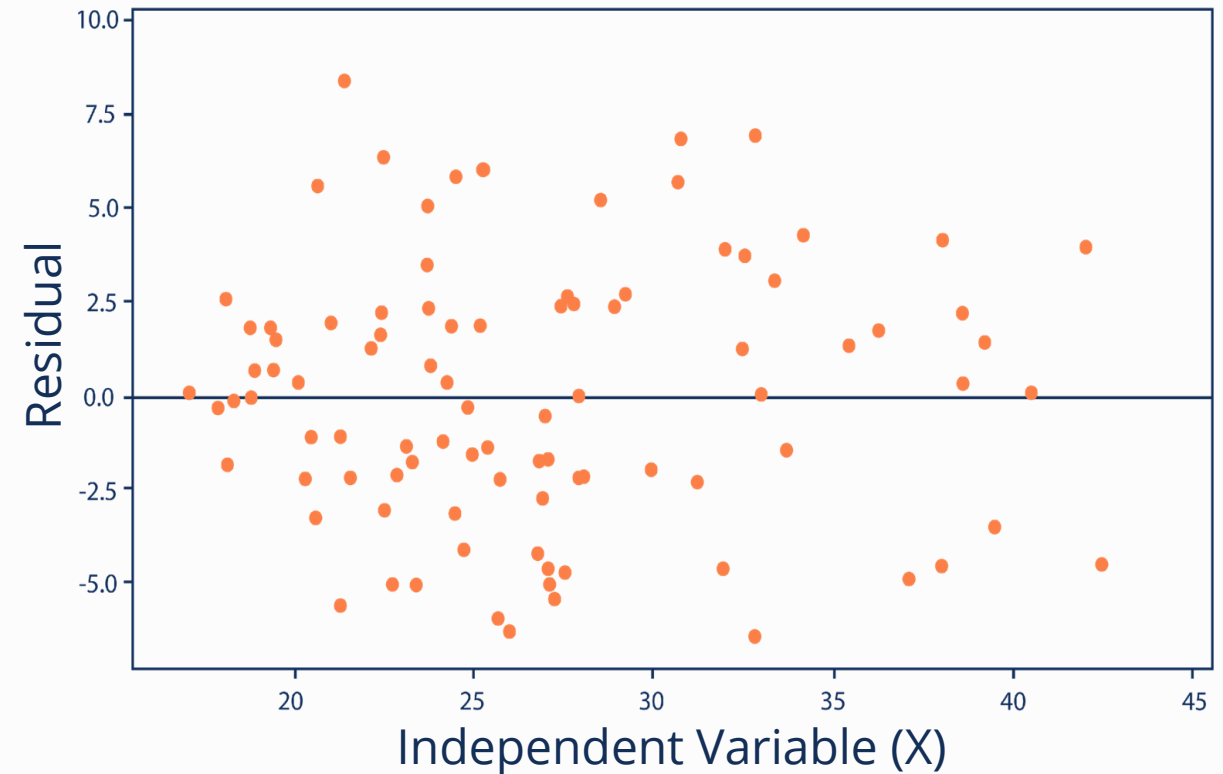
Residuals Recap

- Residuals, or errors, are the differences between the **predicted values** and the observed values ($y_i - \hat{y}_i$)
- The residual quantifies the variation in the target variable that is unexplainable by our line of best fit
- If our model captures all the independent variables, the residuals must be down to simple random errors that we cannot predict



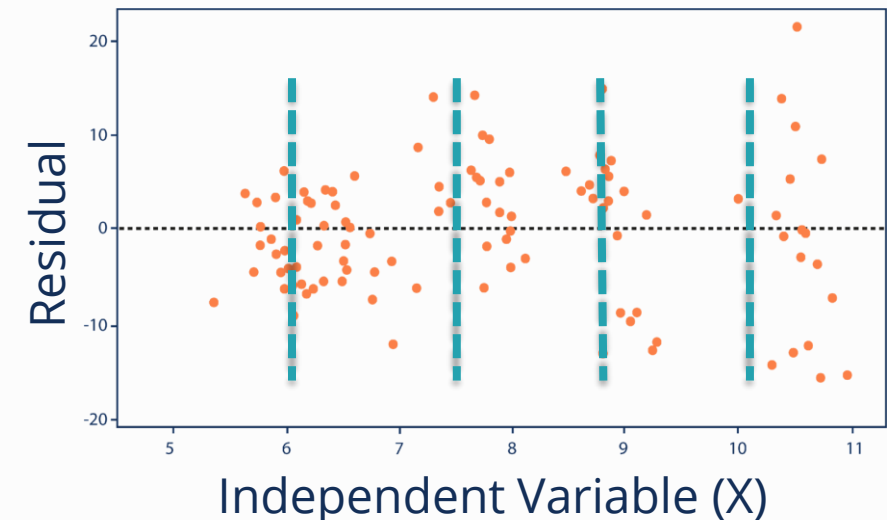
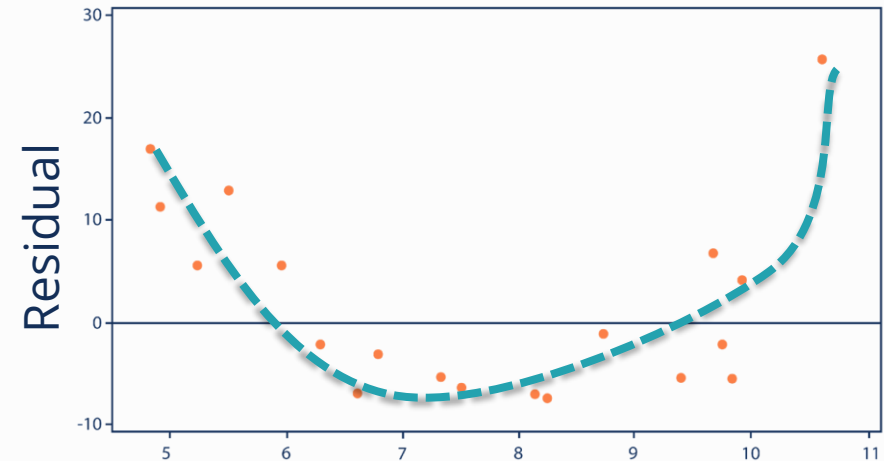
Plotting Residuals

- Plotting residuals is a useful tool for evaluating the trustworthiness of our model at describing the data
- A reliable model will have
 - Residuals randomly scattered around the x-axis
 - Residuals with an average value of zero
- **Note:** A reliable model does not necessarily mean an overall good model – the SSE could still be large



Non-Random Residuals

- **Non-random plots can be problematic** for our model
- Non-random patterns in the residuals could be caused by one or more of the following factors:
 - Linear model is not appropriate
 - Omitted variable bias
 - Data does not satisfy OLS assumptions
- In these cases, we may be able to transform the data to make it fit the model or otherwise use a different model

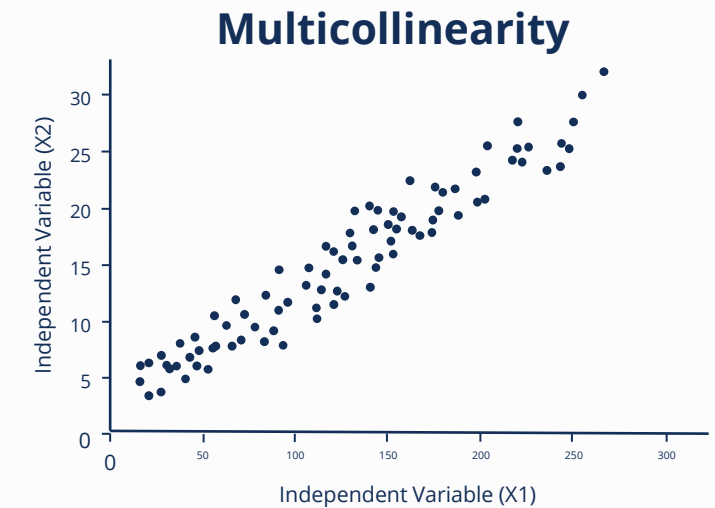
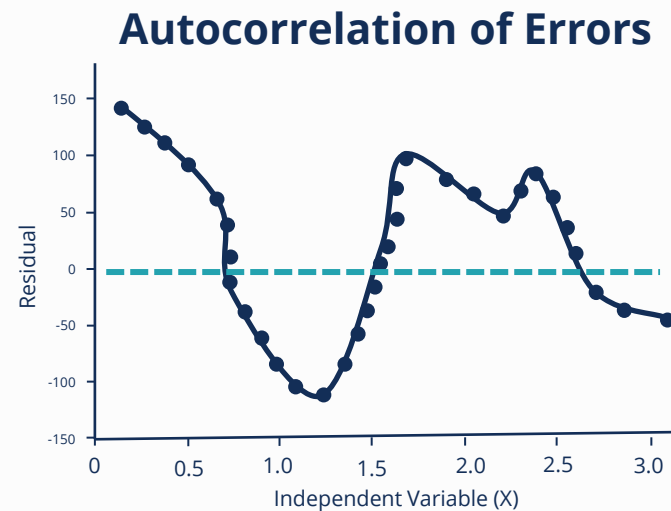
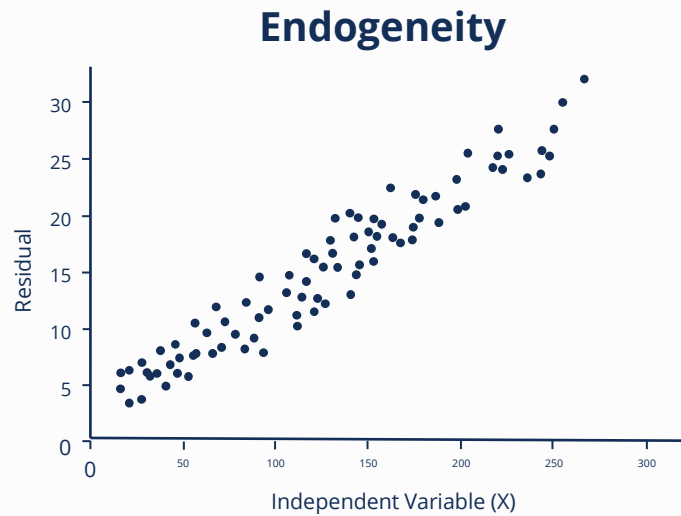
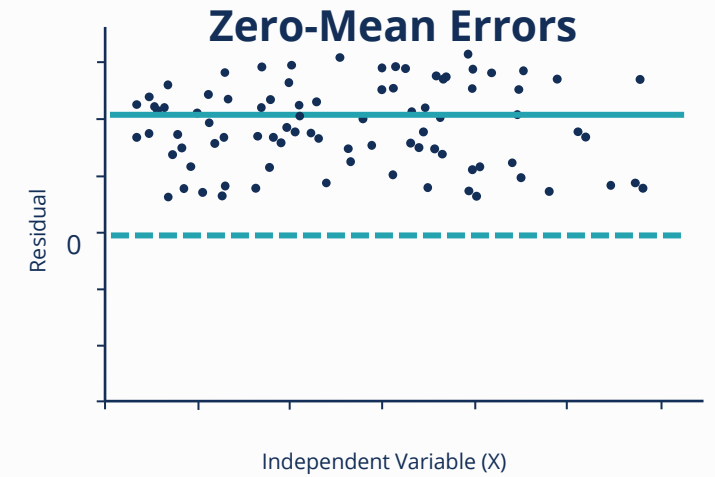
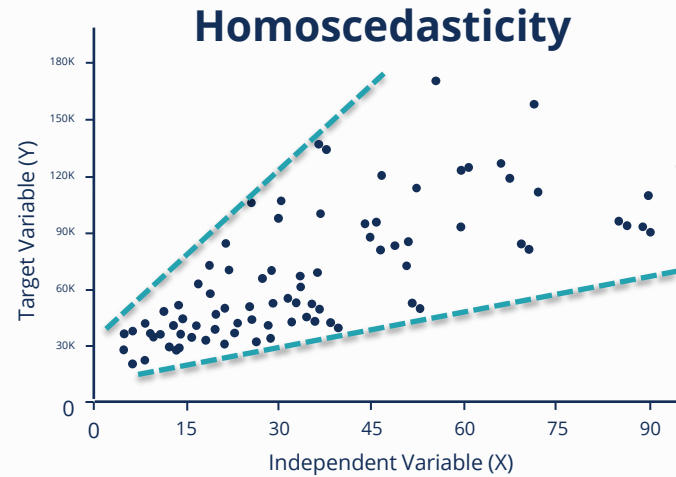
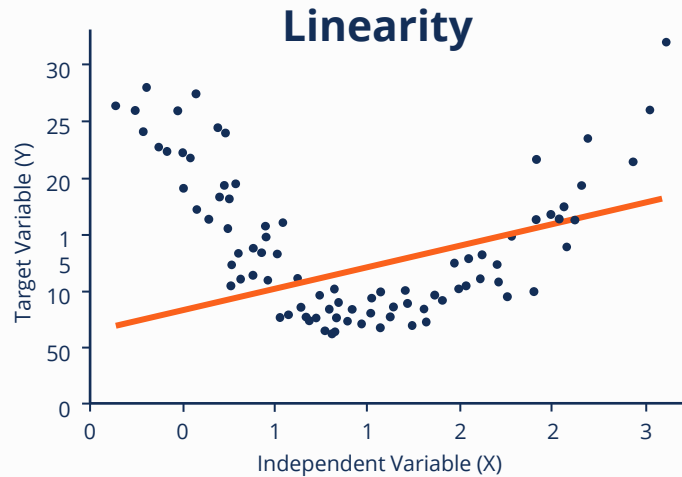


Summary

- Plotting **residuals** offers a powerful tool to assess the model fit
- If we observe the residuals to be randomly distributed, then we can be reasonably content that our model is unbiased
- **Non-random patterns** in the residuals provide the opportunity to improve the model by introducing new independent variables or transformations
- If these methods do not resolve the problems, then potentially a non-linear model is more appropriate

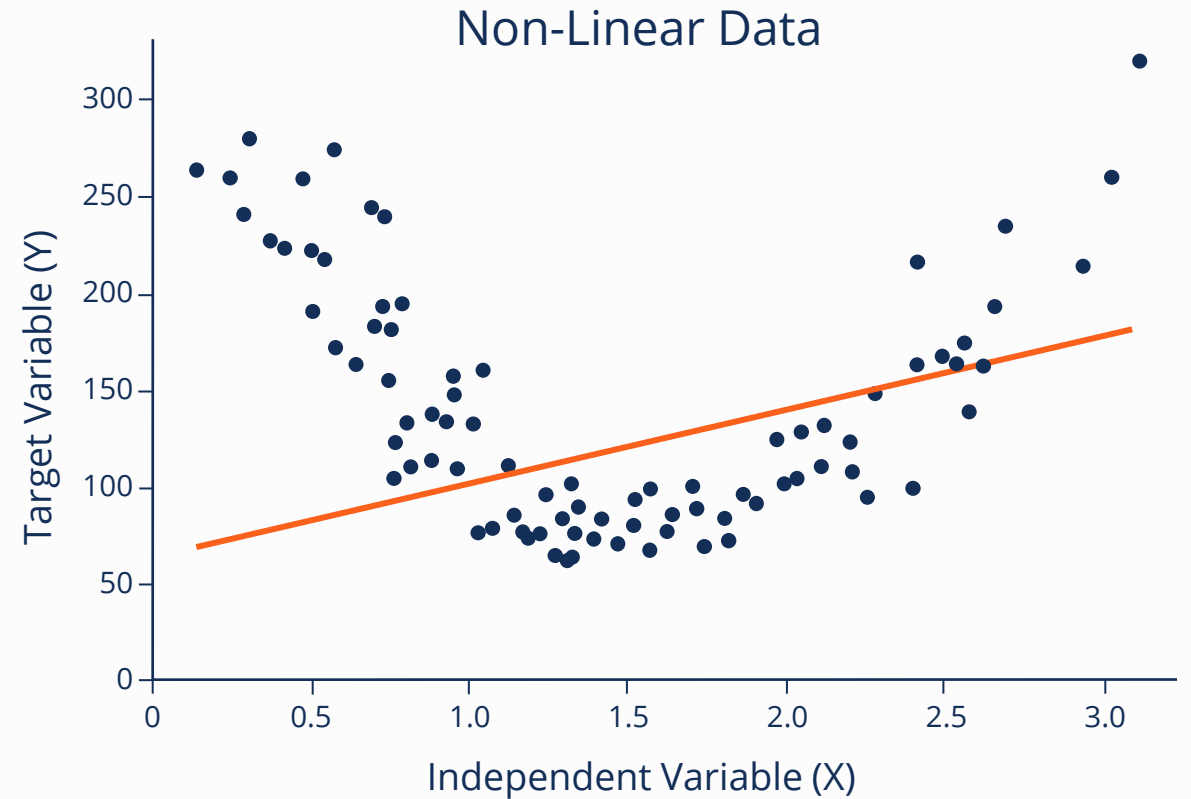
OLS Assumptions

The **Ordinary Least Squares** method relies on 6 assumptions about the data:



Assumption 1: Linearity

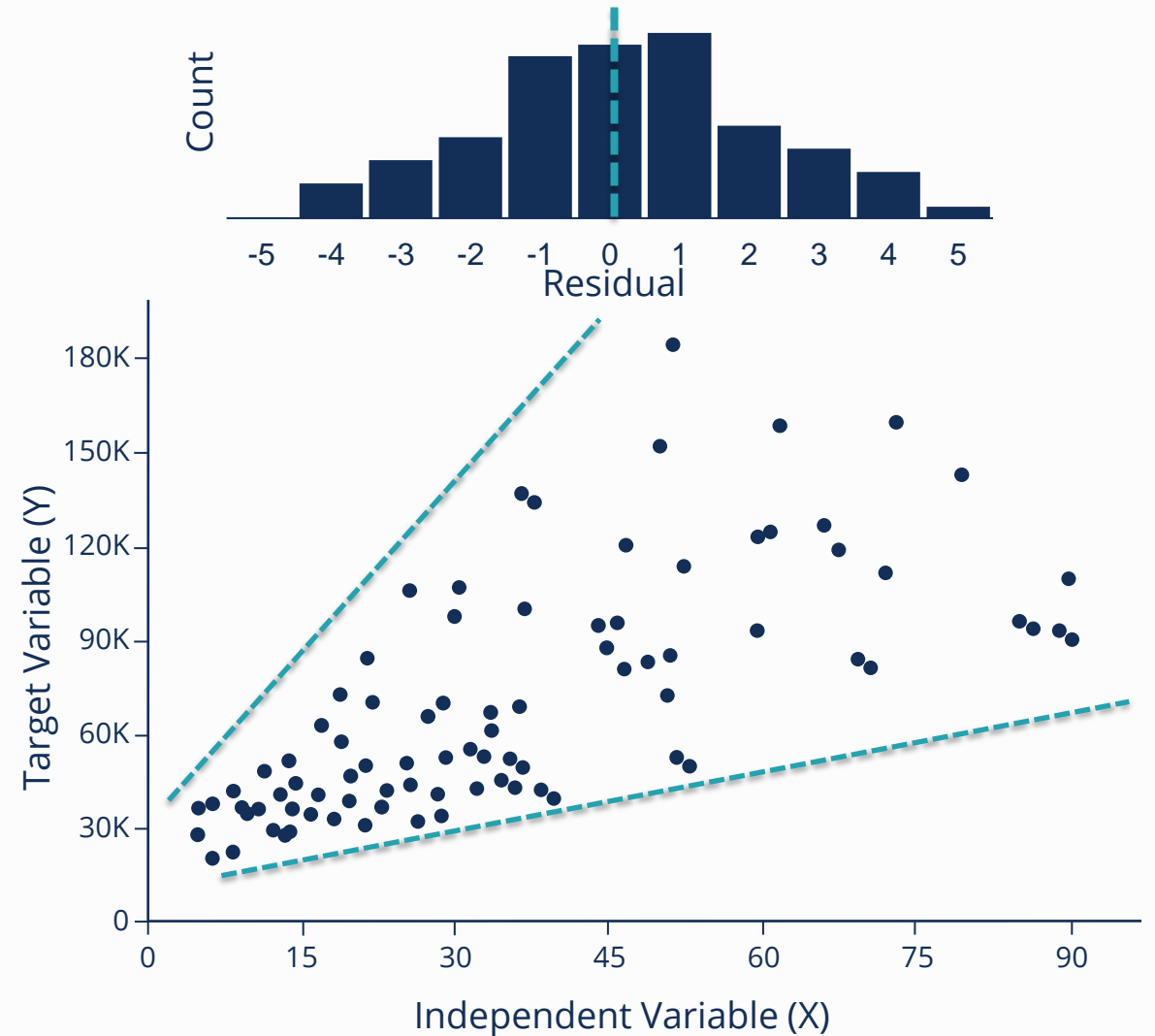
- For a linear regression model to be appropriate, the data must be **linear** in nature
- The target variable should increase by a fixed amount for an increase in each of the independent variables
- We can test for linearity by ensuring the errors in a residual plot are **randomly and symmetrically** scattered about the x-axis
- Non-linear variables can be made linear using transformations such as a log-log transformation (more on this later)



Assumption 2 – Normality and Homoscedasticity of Errors

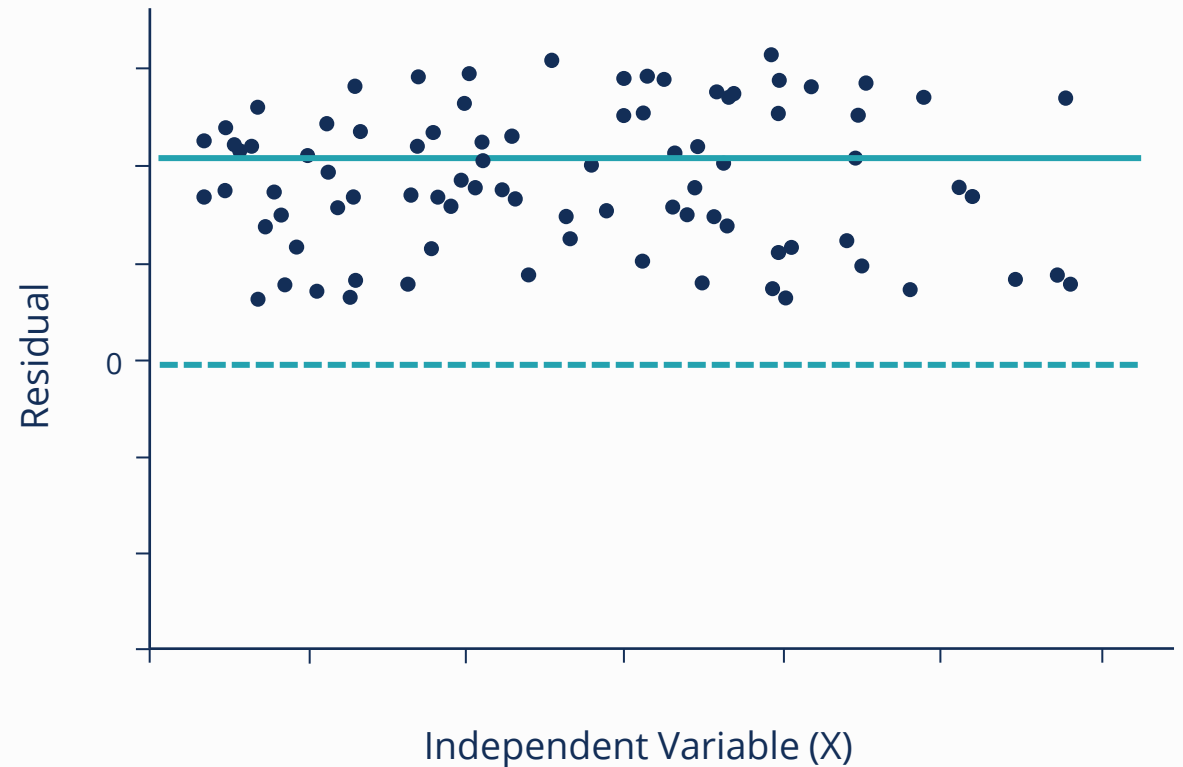
- **Normality** - plot the magnitude and direction of each error on a chart to see if it fits a normal distribution
- **Homoscedasticity** – the spread of errors should not change with the independent variables
- **Heteroskedasticity** – errors that do not have constant variance
- We can test for homoscedasticity by inspecting the scatter of the residual plot

Note: Normality is not strictly required for OLS, but it is required if we want to make statistical inferences about parameters



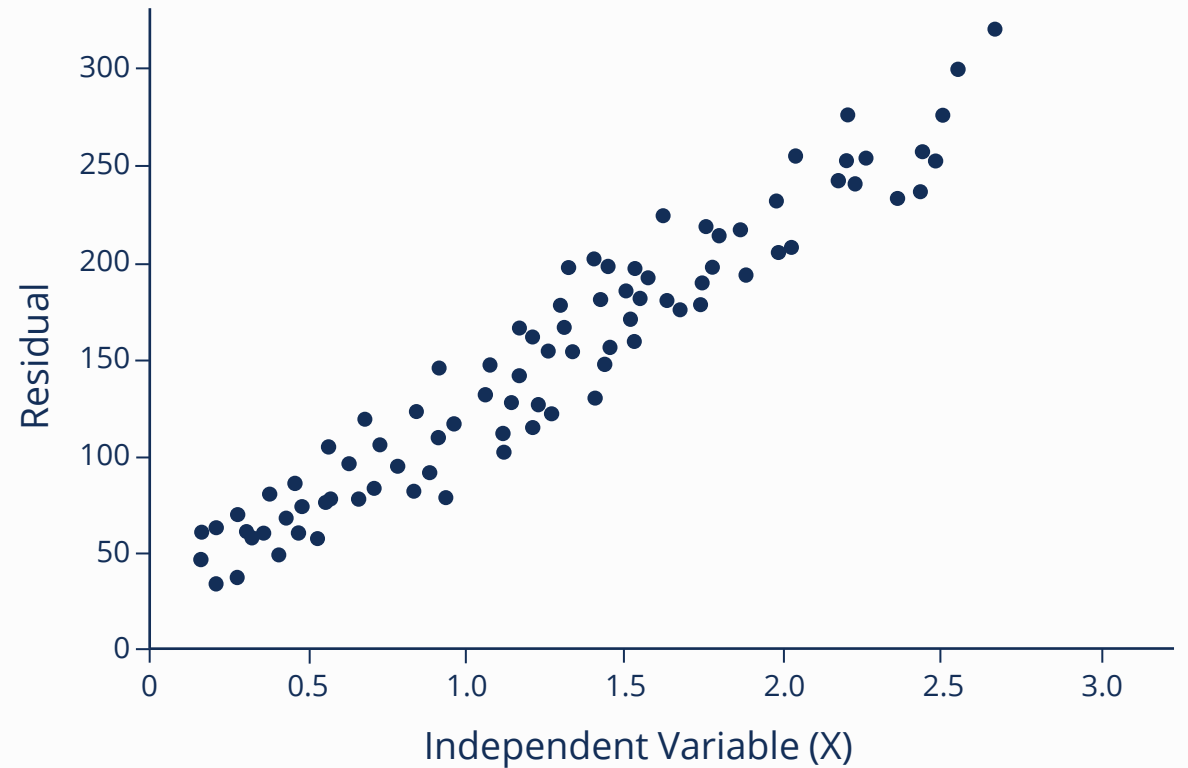
Assumption 3 – Zero Mean Errors

- For an unbiased model, the **average of all the errors should be zero**
- The errors should only capture the random variation in the data that the independent variables cannot explain
- A non-zero average error means the model has a systematic offset between the predicted and observed values, known as **bias**
- When present, the intercept parameter will absorb any systematic offset, forcing the average error to be zero
- Can test for this by checking the average of the errors is close to zero



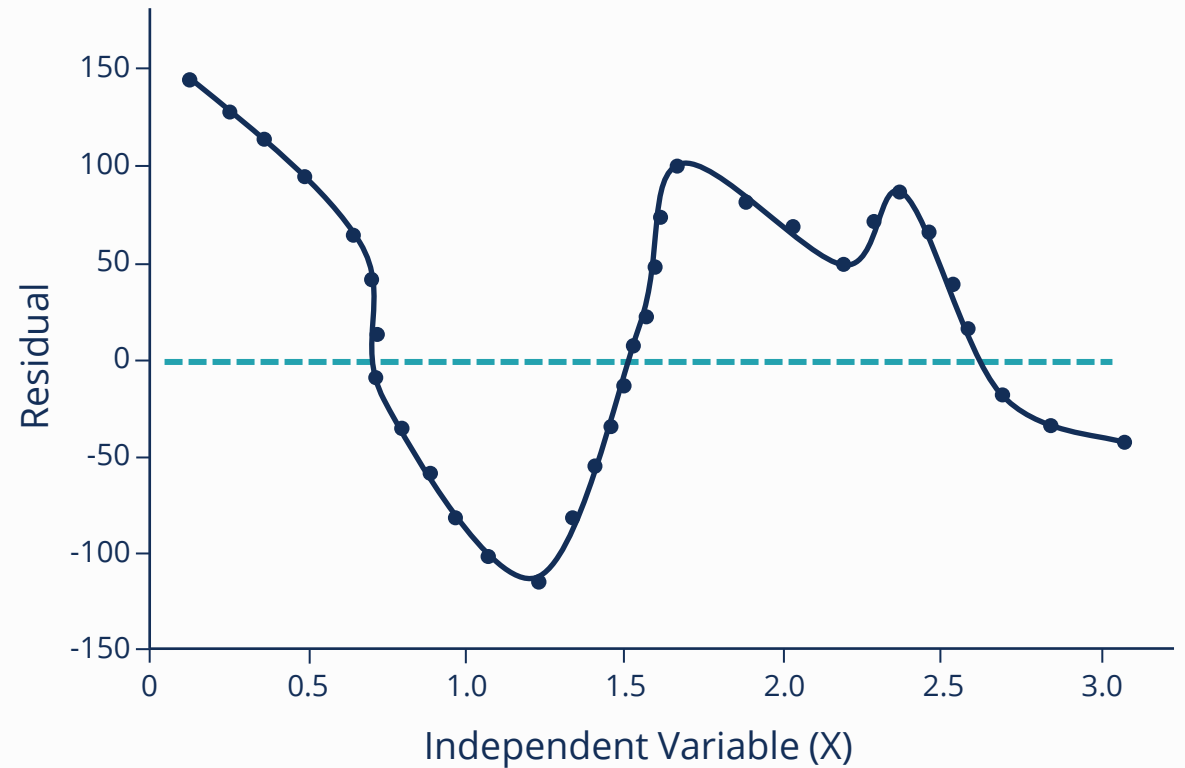
Assumption 4 - Endogeneity

- **Endogeneity** - there is no correlation between the error and the independent variables
- Correlation suggests that the random error can be predicted by independent variables
- Endogeneity leads to biased parameter estimates
- **Omitted-variable bias** – an important independent variable is not included in the model and is correlated with other independent variables. The omitted variable will be included in the error and will cause endogeneity



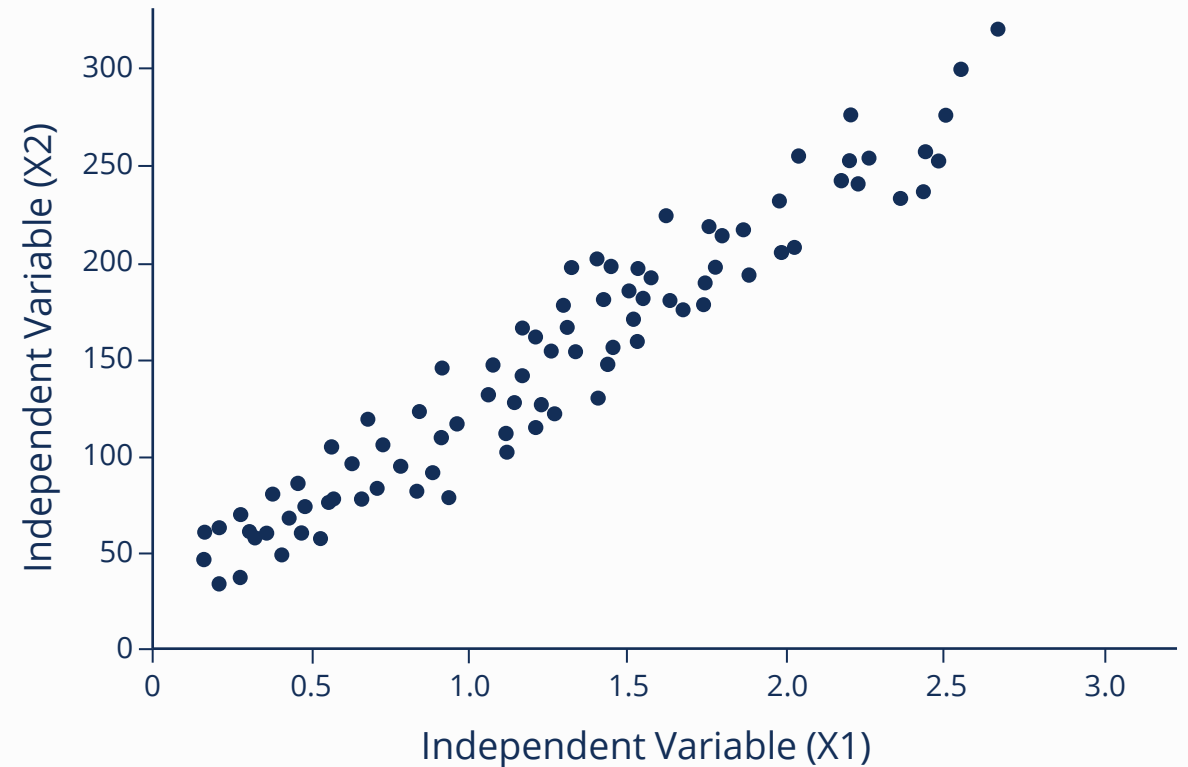
Assumption 5 – Autocorrelation of Errors

- Errors should not be correlated with previous errors or themselves
- There is **autocorrelation** of the errors when an error value is dependent on the previous error value
- We often observe autocorrelation in time series data
- We can identify autocorrelation of errors through patterns in the **residual plot** or using the Durbin-Watson statistic



Assumption 6 – Multicollinearity

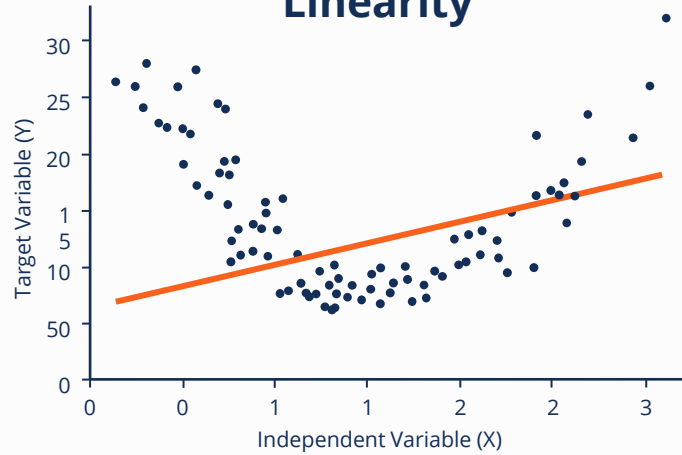
- **Multicollinearity** - independent variables are highly correlated with each other
- The OLS algorithm cannot separate the effects of the correlated independent variables on the target variable and so will produce unstable parameter estimates
- We can prevent multicollinear independent variables by calculating the correlations between them and removing one from every highly correlated pair



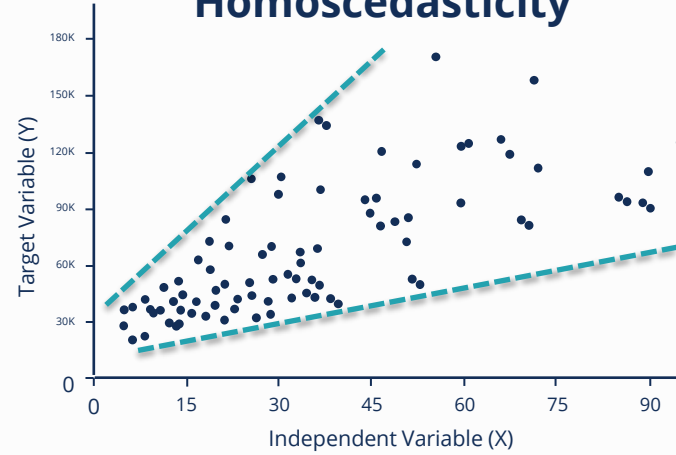
OLS Assumptions

When all six assumptions are met, the **Ordinary Least Squares** method is the best linear regression method

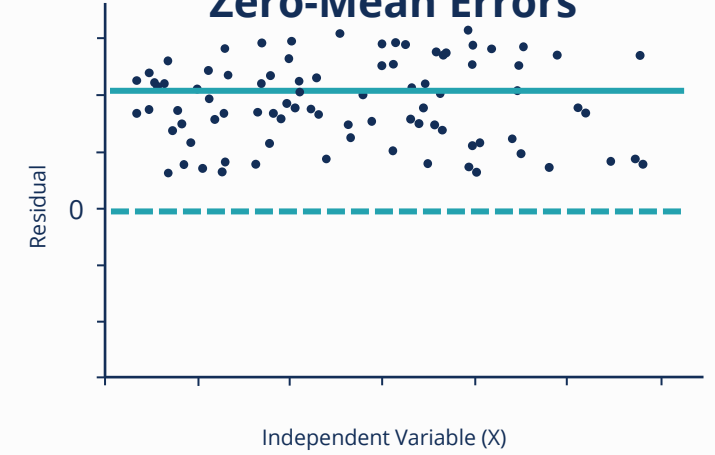
Linearity



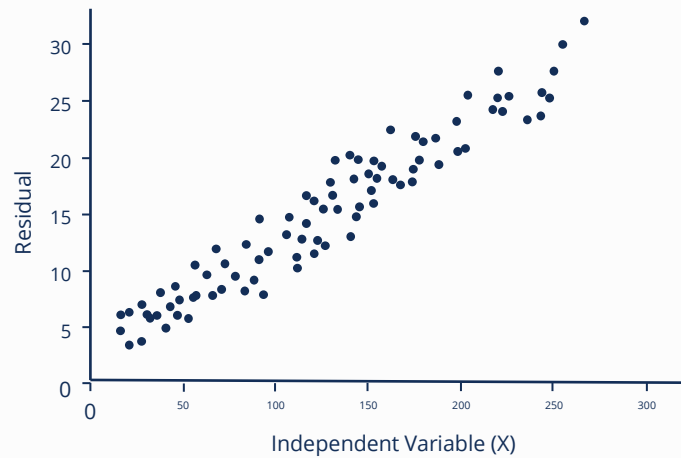
Homoscedasticity



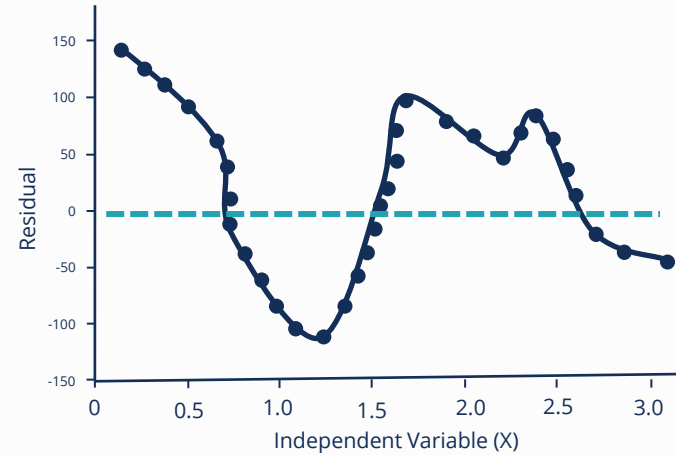
Zero-Mean Errors



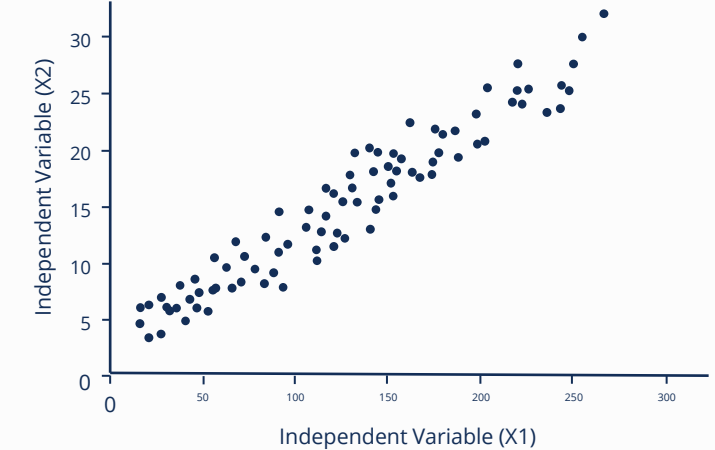
Endogeneity



Autocorrelation of Errors

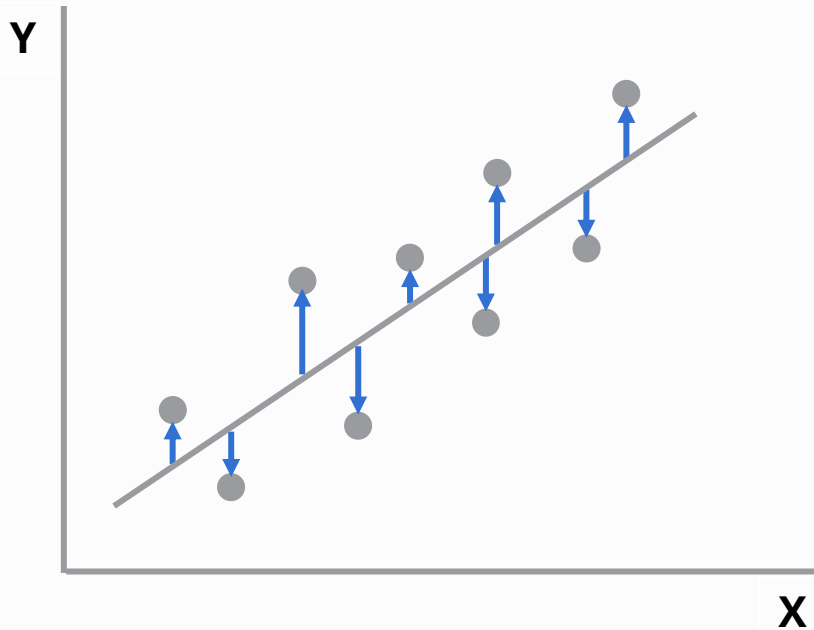


Multicollinearity



Linear Regression Evaluation

- We can measure the model performance by how good our model is at making predictions



- These metrics allow us to compare the performance of different models

Squared Metrics

Sum of **Squared** Error (SSE)

Mean **Squared** Error (MSE)

Root Mean **Squared** Error (RMSE)

Absolute Metrics

Sum **Absolute** Error (SAE)

Mean **Absolute** Error (MAE)

Linear Regression Evaluation – Squared Metrics

- **Sum of square errors** (SSE) sums the squared distance from the residuals to a **line of best fit**

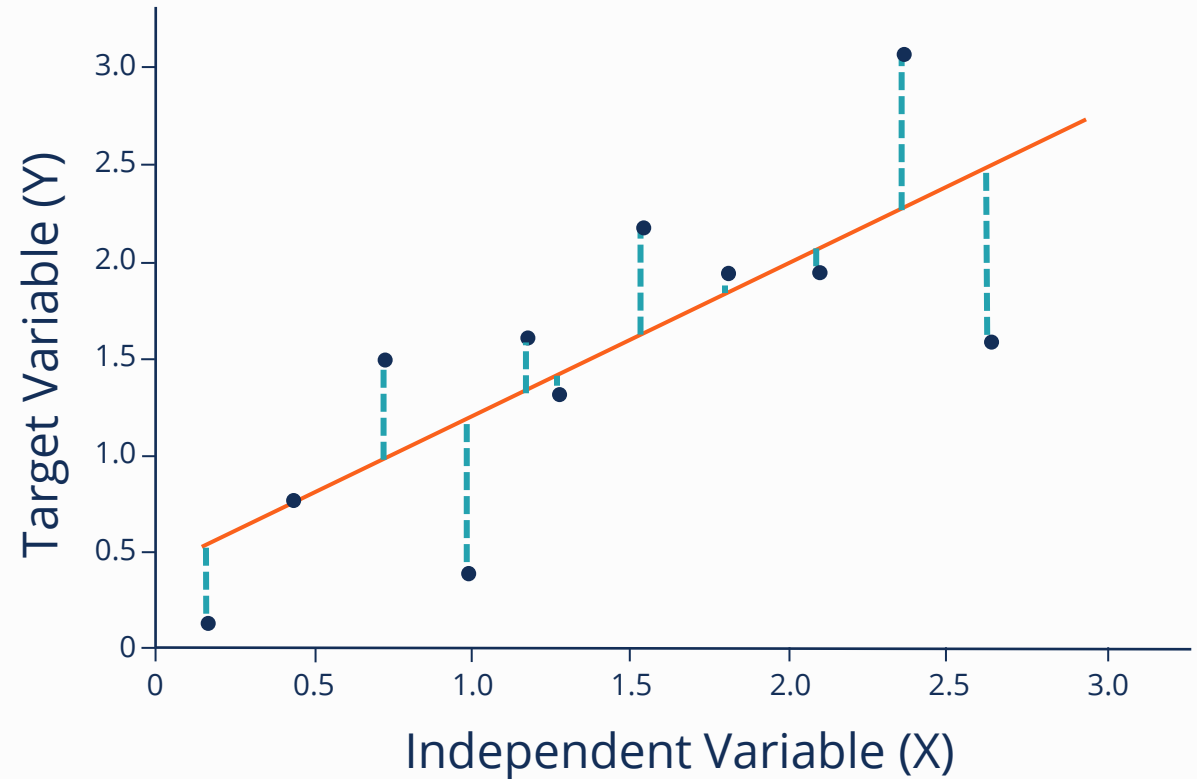
$$SSE = \sum (y_i - \hat{y}_i)^2$$

- **Mean square error** (MSE) divides the sum of square errors by the **number of data points**

$$MSE = \frac{SSE}{N}$$

- **Root mean square error** (RMSE) takes the square root of the mean square error

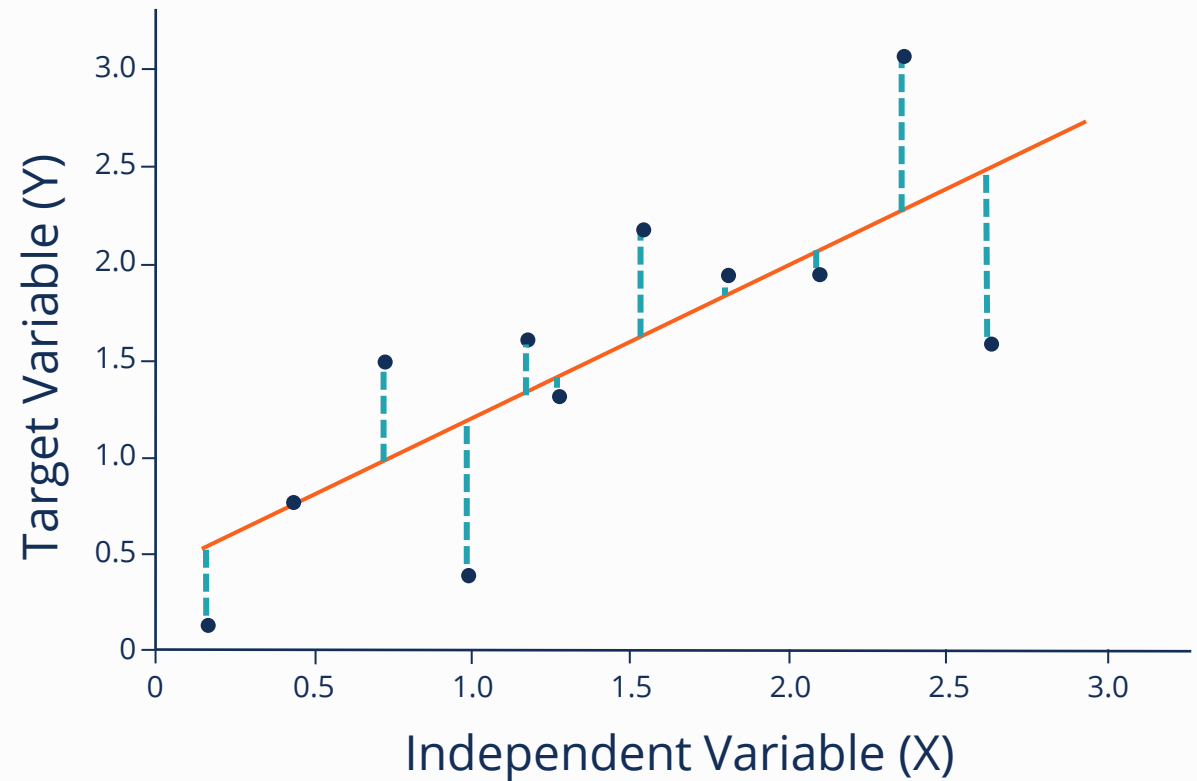
$$RMSE = \sqrt{MSE}$$



Linear Regression Evaluation – Absolute Metrics

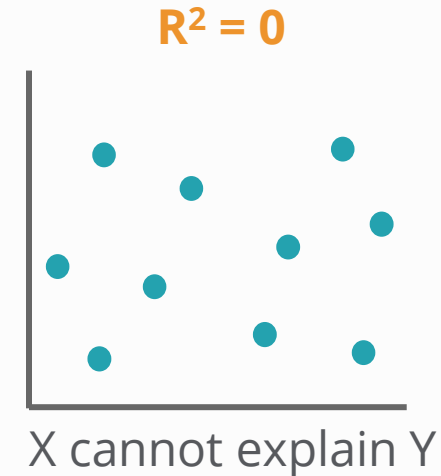
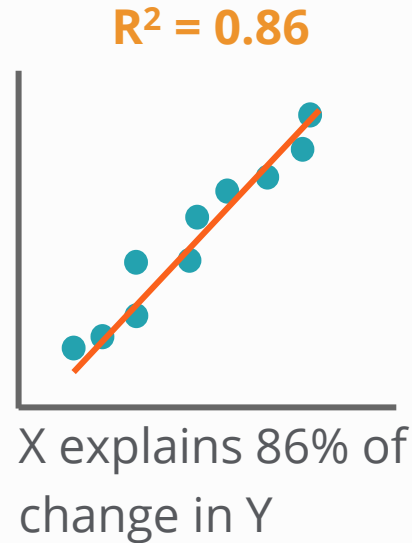
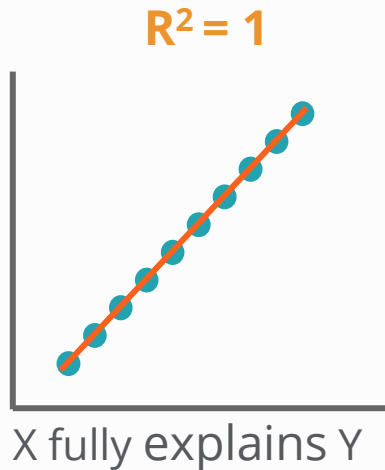
- **Sum of square errors** (SSE) sums the squared distance from the residuals to a **line of best fit**
- **Mean square error** (MSE) divides the sum of square errors by the **number of data points**
- **Root mean square error** (RMSE) takes the square root of the mean square error
- **Mean absolute error** (MAE) sums the distance from the residuals to a **line of best fit** divided by the **number of data points**

$$MAE = \frac{1}{N} \sum |y_i - \hat{y}_i|$$



Linear Regression Evaluation – R²

- **Coefficient of Determination (R²)** calculates to what extent the independent variables explain the changes in the target variable



- Coefficient of Determination:
- Sum of square errors:
- Total sum of squares:

$$R^2 = 1 - \frac{SSE}{TSS}$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$TSS = \sum (y_i - \bar{y})^2$$

Linear Regression Evaluation – R^2

- SSE/TSS tells us how much variance there is in our model as a fraction of the total variation in the target
- One minus this value tells us how much **variation has been explained by the model**

$$R^2 = 1 - \frac{SSE}{TSS}$$

- The independent variables in models with high R^2 values explain most of the variance present in the data
- As with the SSE, the coefficient of determination will never decrease by adding in new variables; however, adding in unnecessary variables can lead to **overfitting** the data
- We can see the effects of overfitting when we compare training performance with test data performance

Linear Regression Evaluation - Adjusted R²

- We can modify the R² to account for the number of independent variables – this is known as the **adjusted R²** or \bar{R}^2

$$\bar{R}^2 = 1 - \frac{SSE / (n - p - 1)}{TSS / (n - 1)}$$

- n is the number of observed datapoints
 - p is the number of independent variables in the model (excluding the intercept term)
-
- The adjusted R² value only increases when a new variable is added that improves the model by more than would be expected due to random chance
 - It can therefore decrease if a variable is added that does not sufficiently explain the data
 - This adjusted R² will always be lower than the unadjusted R²
 - It allows comparison between models with different numbers of independent variables

Overview of Linear Regression Evaluation

- Evaluating these metrics after building a model provides insight into the model performance
- MSE, RMSE and MAE metrics tell us how far our predictions are from their true values on average
- MSE or RMSE can be used when we want to give greater importance to individual large errors
- The coefficient of determination tells us how much of the variability in the data the independent variables are explaining
- The adjusted coefficient of determination helps prevent overfitting by decreasing if a variable is added that does not sufficiently explain the data
- The MSE, RMSE and MAE are absolute measures of fit, while R^2 is relative to the total sum of squares

Regression Coefficients

- **Regression coefficients** help us understand the interaction between variables
- The following model predicts house prices (P) using the building area (A) and a measure of the overall house condition (C):

$$P = 77000 + 80A + 9200C$$

Three Parameters

- A base price of £ 77,000
- On average for every extra sqft we add, the house price increases by £80
- Similarly for every condition point we add, the price will increase by £9,200

Regression Coefficients

- **Regression coefficients** help us understand the interaction between variables
- The following model predicts house prices (P) using the building area (A) and a measure of the overall house condition (C):

$$P = 77000 + 80A + 9200C$$

- In general, our parameter values or coefficients tell us **how much the target variable changes** on average when we **change the corresponding independent variables**
- Positive coefficients mean that the target variable is positively correlated with the independent variable while negative coefficients tell us that it is negatively correlated

Compare Coefficients

- We can scale the independent variables before fitting the regression in order to compare coefficients
- **Standardization** – transforming data to follow normal distribution with mean zero and unit variance by subtracting the **mean value** and dividing by the **standard deviation**

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

- Our scaled house price model then becomes:

$$P = 113200 + 41766Z_A + 30047Z_C$$

- We can compare relatively which independent variable contributes more to the target variable for a single unitless increment in the independent variable
- This produces the same results for our 800sqft/score 6 house and we observe that area contributes more to the price relative to the condition score

p-Values

- We can test whether coefficients are statistically significant by calculating **p-values**
- A p-value tells us the probability that we would get the results we observe when we assume the null hypothesis is true
- If this p-value is below 0.05 then it is unlikely that we observe such a distant from zero value for the coefficient by chance and we therefore assume that **there is a significant relationship** between the independent variable and the target variable
- We can calculate p-values for all the coefficients in our model and discard the independent variables that are not significant
- Removing independent variables that are not statistically significant helps prevent overfitting

Calculate p-Values

Process to calculate p-values:

Calculate the standard error on each coefficient

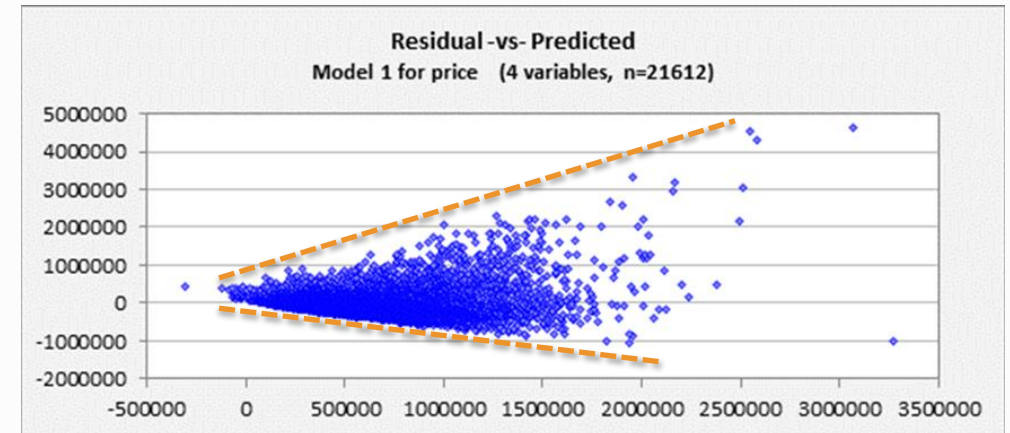
Combine the standard error with the coefficient value to produce a t-statistic value

Calculate the p-value as the probability of obtaining this t-value from the student's t-distribution

- A significant p-value does not necessarily imply a high R^2 value and vice-versa

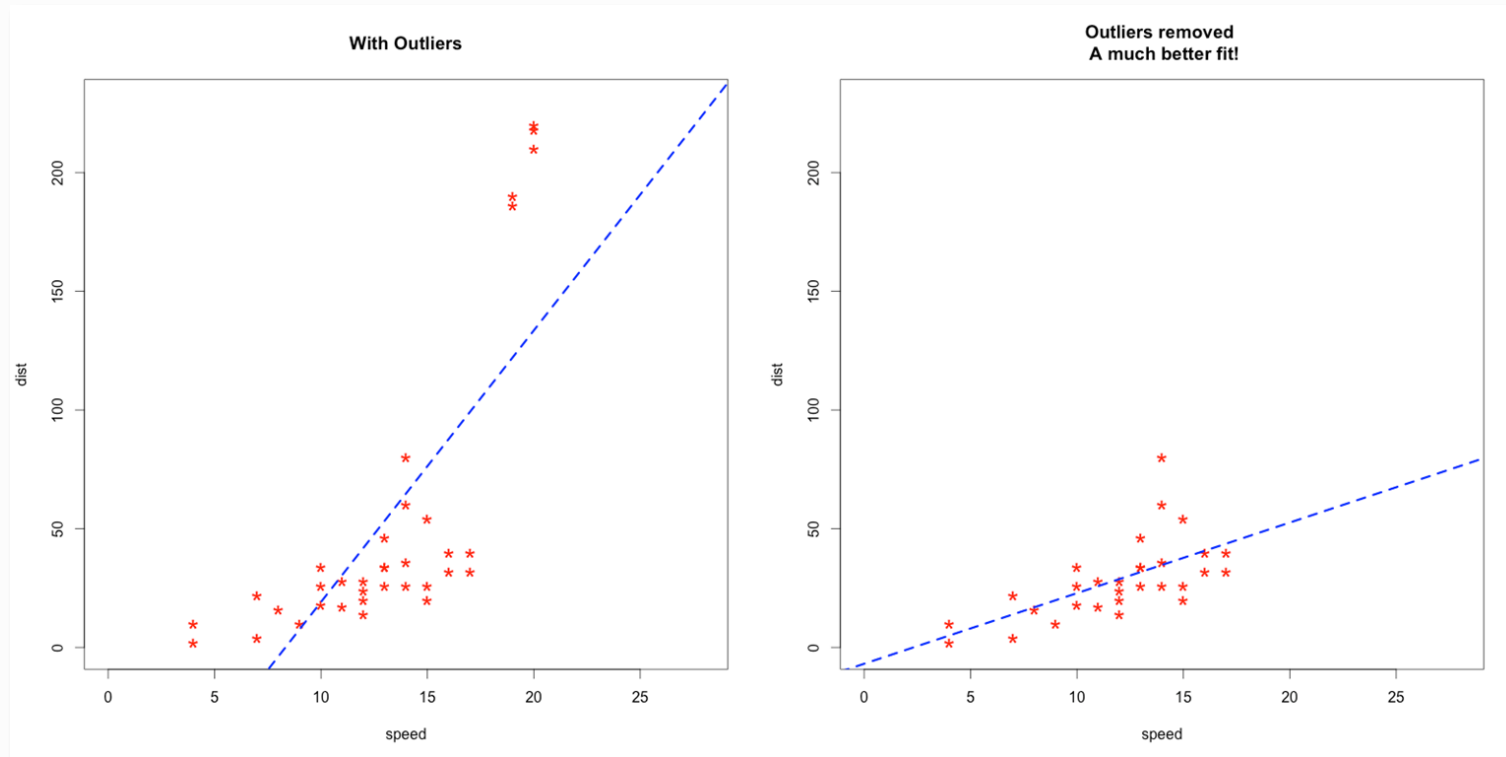
Interpreting Regression Model for House Price

- In the practical exercise, we used RegressIt to predict house price.
- A good model will have a tight grouping of residuals that are consistent across all values.
- We can see that our model works well for low prices but begins to widen at higher values.
- This means that if our model predicts a high house price we cannot be as sure of its accuracy.



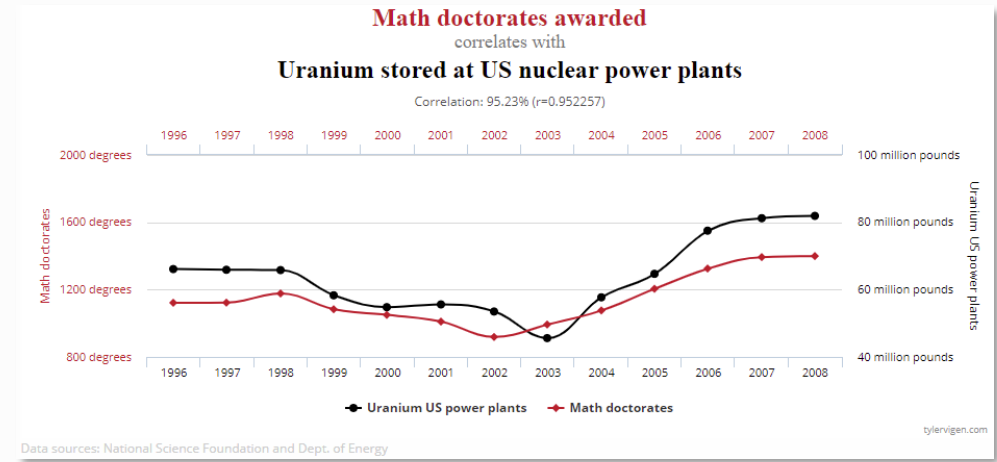
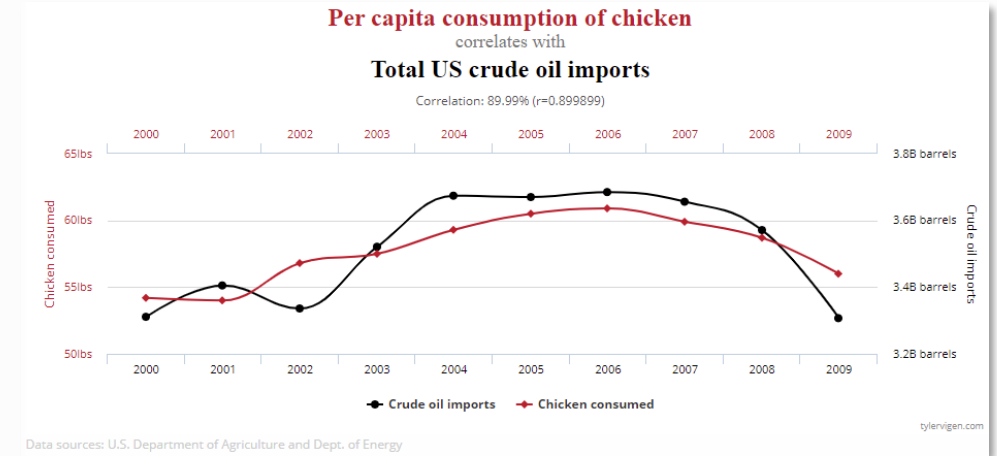
Outliers

- Outliers can skew the prediction of our regression model.
- We can see from below that the two models give very different predictions.
- Should we remove the outliers, leave them there, or use a different kind of model?



Correlation/Causation

- An important final point to remember is the difference between correlation and causation.
- As we can see from these charts, there can be very strong correlations between data, with no realistic connection.
- It is important to remember the distinction between correlation and causation before making recommendations and decisions.
- It is essential to first understand data and to question whether your hypothesis are plausible.





Advanced Regression

Advanced Regression



An introduction to more advanced regression techniques



Know what you don't know and give you a few ideas of what to explore next



Practice a more advanced regression in Excel

Techniques we will outline:

Log Log Linear Regression

Polynomial Regression

Random Effect Models

Repeated Measure Regression

Logistic Regression

Bayesian Regression

Segmented & GAM Regression

Lasso Regression

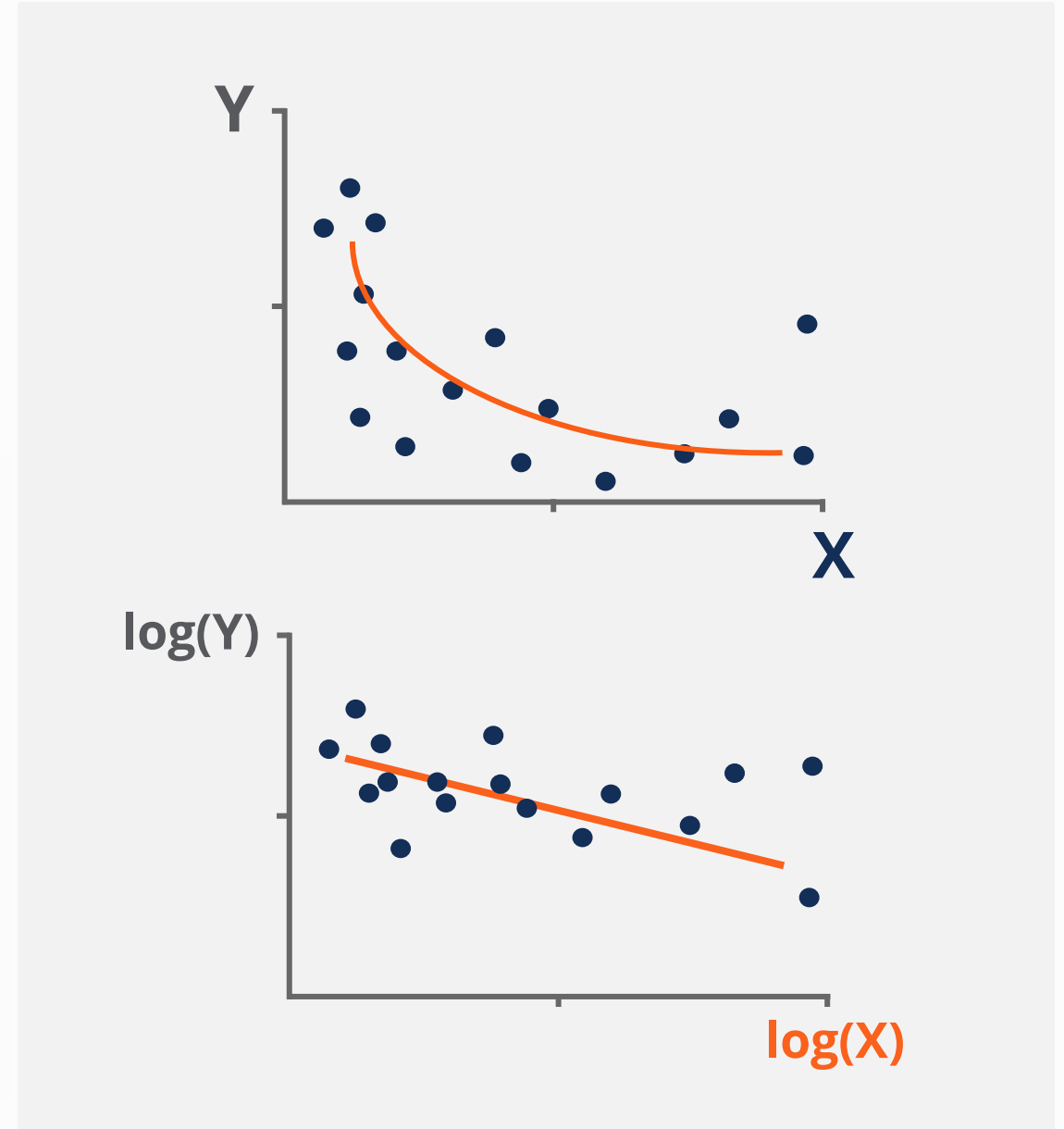
Poisson Regression

Log-Log Linear Regression

Non-linear relationships are hard to interpret.
For that reason, linear models are desirable.

Log-Log Linear Regression

- Log-log regression is used when our data appears non-linear or we have heteroscedastic errors.
- When Y and at least one X are transformed with a log, this is log-log linear regression.
- Useful when we have linear percentage increases between target and independent variables.

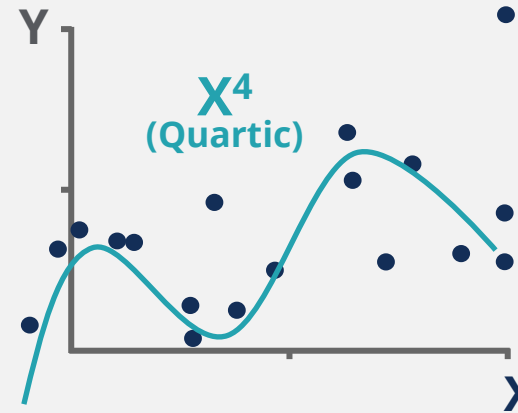
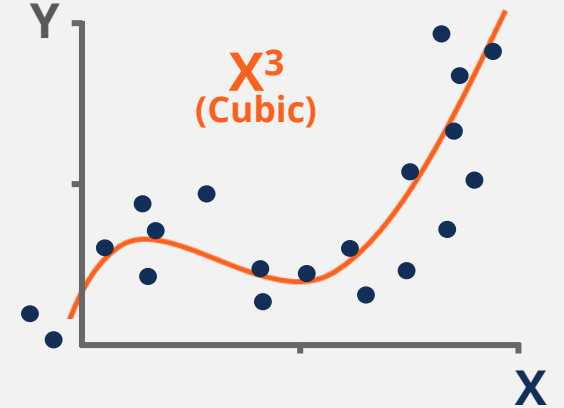
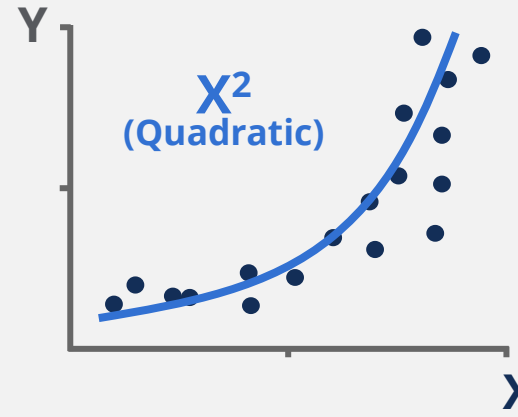


Polynomial Regression

- We use polynomial regression when we have a smooth, non-linear relationship between variables.
- Each input (x) variable is transformed with an exponent:

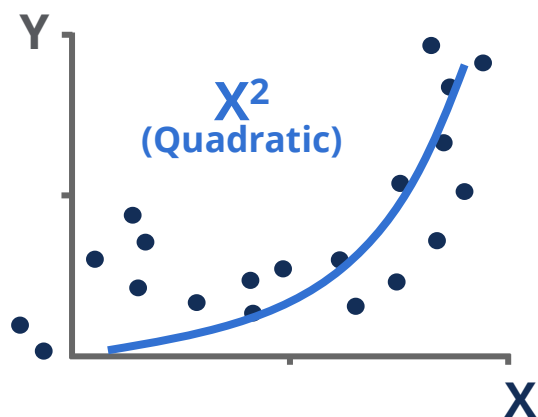
$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_n x^n$$

- A higher exponent suggests a more complex relationship.
- **Benefits:** Easy to add complexity.



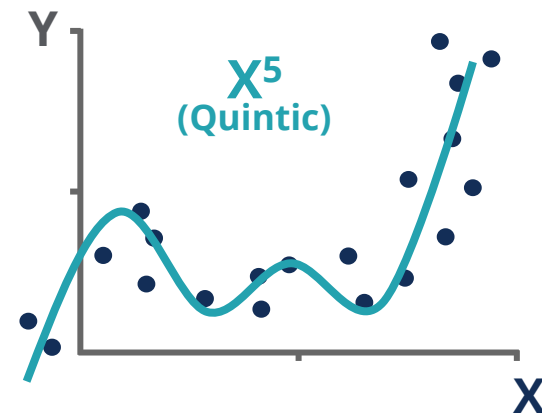
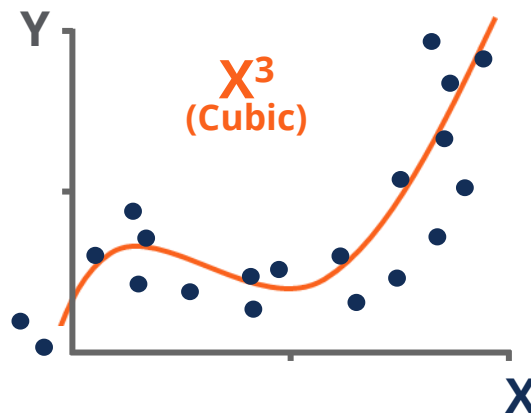
Polynomial Regression Continued

The polynomial exponent indicates the complexity of the regression line.



Underfitting Polynomial

When the exponent is **too low**, the relationship is **over simplified**.



Overfitting Polynomial

When the exponent is **too high**, the relationship is **too specific**.

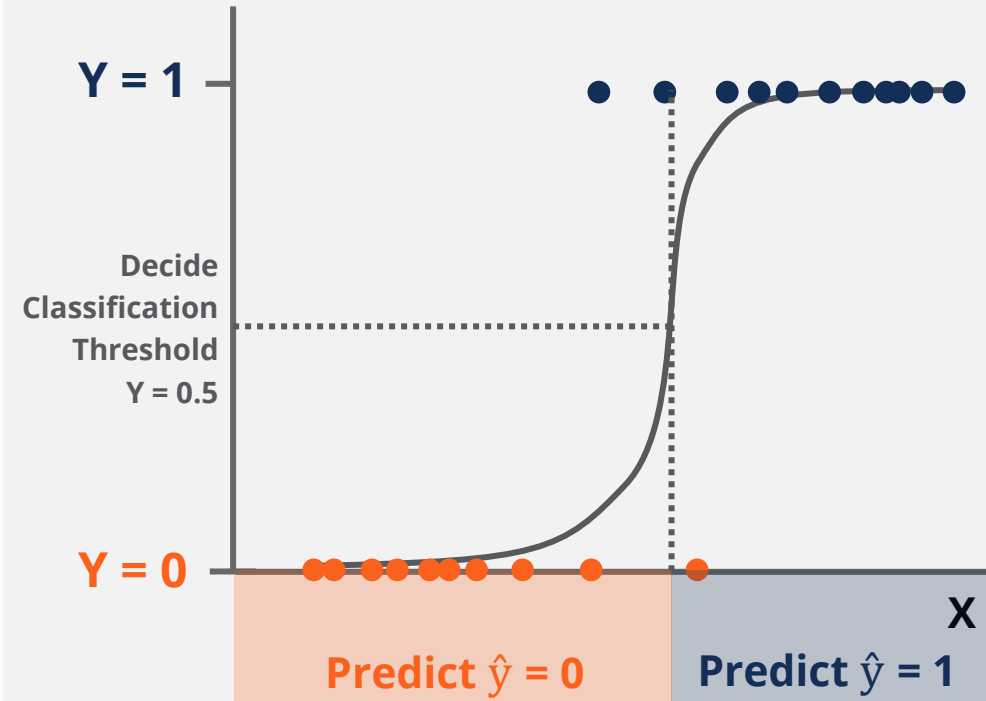
Logistic Regression

- Logistic regression is typically used for classification problems
- In logistic regression we fit the logistic curve function to the data:

$$y = \frac{1}{1 + e^{-\beta x}}$$

- The y-value increases from zero to one as x increases
- **Benefit:** Easy way to apply classification decisions to a continuous outcome.

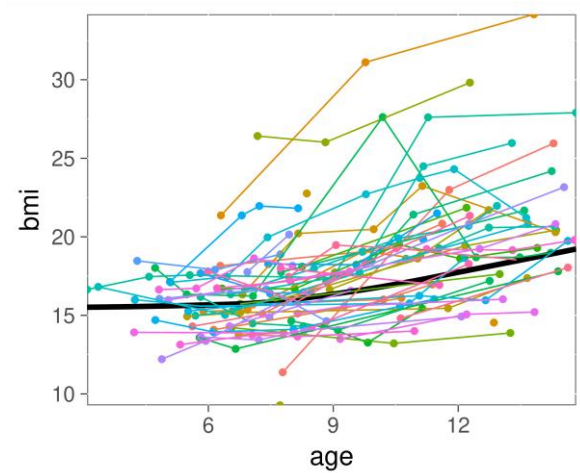
Logistic Classification



- New datapoints are classified based on whether their predicted y value is above or below a certain threshold
- Typically, we start with 0.5 or 50% as the decision threshold.

Repeated Measure Regression

- Repeated measure regression is used when measurements are repeated over time.
- Common in medical fields where the same patient is measured at regular intervals.
- Repeat measures are therefore not independent.



In this example the patient is measured 3 times

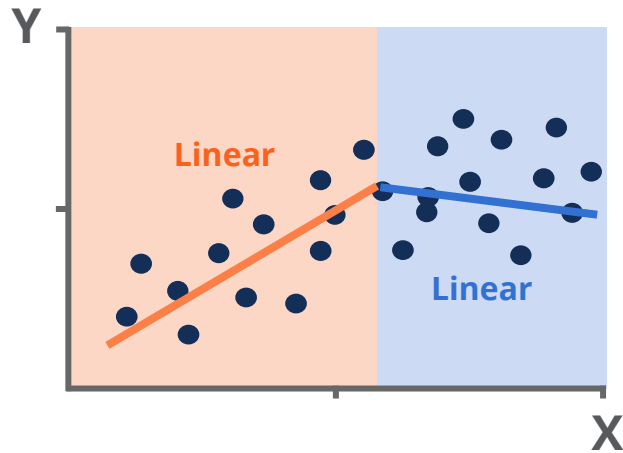


Patient	Y1	Y2	Y3
1	2	3	4
2	0	3	1
3	1	4	3

- ANOVA techniques are popular for repeated measure regression
- ANOVA requires the intervals (Age in this case) to be regular.

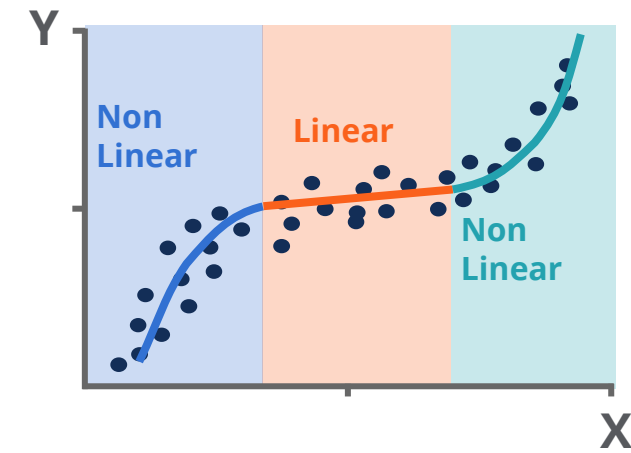
Segmented Regression

Basic Segmented Regression



- Applies unique **regression lines** for different regions of X.
- Both regions have the same form.
- **Benefit:** Low complexity, easy to understand.

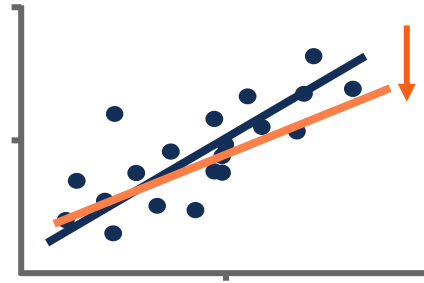
Generalized Additive Models



- Applies unique **regression lines** for different regions of X.
- Each region may exhibit a different **regression form**.
- **Benefit:** Can help model more complex relationships.

Other Advanced Models

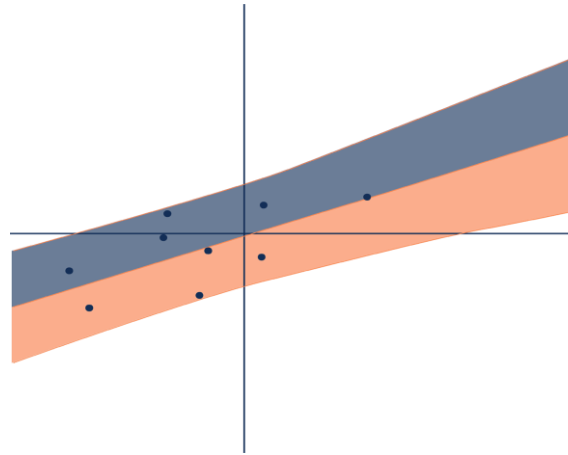
Lasso Regression



- Penalises less informative variables by reducing their coefficients.
- Effectively adds bias, in order to reduce variance.
- **Benefit:** Low complexity, easy to understand.

Other Advanced Models

Bayesian Regression



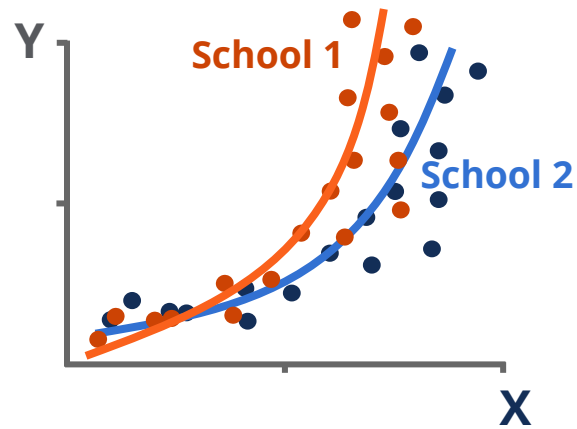
- Used to provide a level of certainty with regression output.
- Useful in aiding decision making. I.e. an output with low certainty should be treated with caution.

Other Advanced Models

Random Effect Models

- Attempt to capture underlying correlations between observations.

e.g. nationwide student results where students may correlate by school



Other Advanced Models

Poisson Regression

- Used to model counts of something in a given time or area.

e.g. Vehicles passing in a minute

e.g. Number of positive cases in each town

