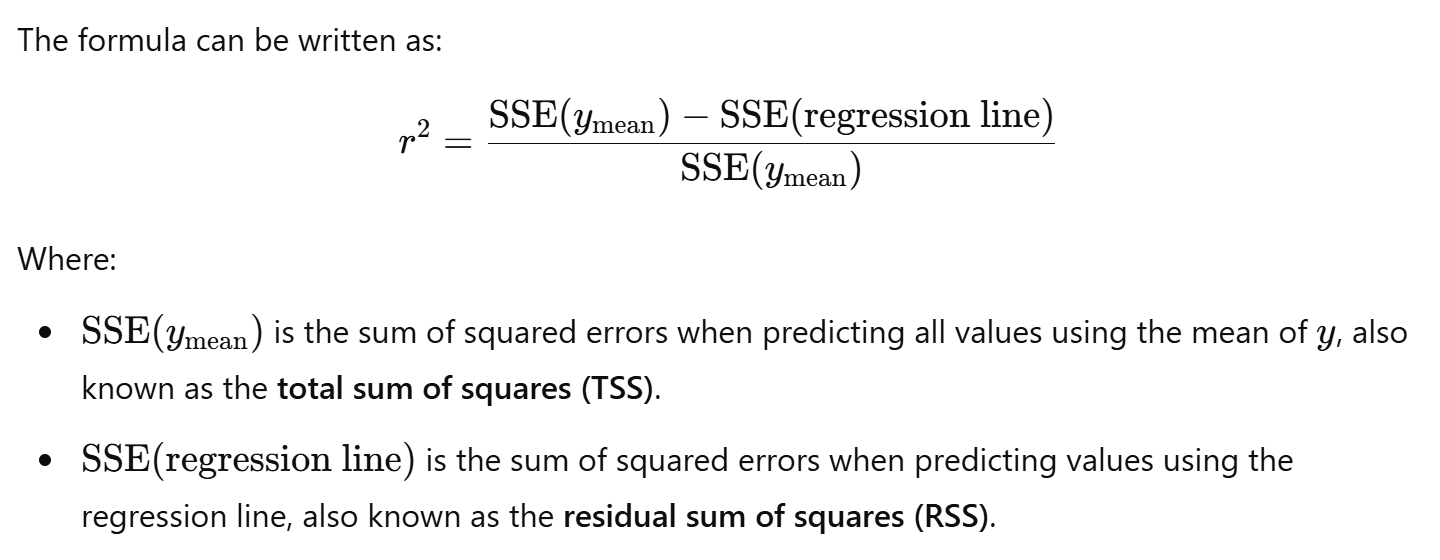
Ch7 (Regression) Lecture Note / Study Guide

Section 7.1

* Focus on the regression model. Make sure to be familiar with the definitions and notations:
  + What is a model? The model is formulated as:
  + is the predicted value.
  + X is an input feature(s).
  + is an output feature.
  + What is the residual of an instance? It is the difference between the observed and predicted value, which is

# Section 7.2



* Key differences between r and
  + r (**correlation coefficient)**
    - measure the strength and direction of the **linear relationship between two variables.**
    - Takes the value between -1 and 1.
    - r = 1 indicates a perfect positive relationship
    - r= -1 indicate a perfect negative relationship
  + (**coefficient of determination**)
    - Statistical measure used in regression analysis to assess how well a model fits the observed data. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. In simpler term, tell us how much of the variation in the outcome can be explained by the model.
    - For example, an R² of 0.8 means that 80% of the variance in the dependent variable is explained by the model, and the remaining 20% is unexplained or due to other factors.

Code analysis:

* X = crabs[['latitude']].values.reshape(-1, 1)
  + **crabs[['latitude']]**: This extracts the **'latitude'** column from the crabs DataFrame. The double brackets [['latitude']] ensure that the result is still a DataFrame (not a Series). This is useful when you want to work with the data in a way that maintains its 2D structure, such as for modeling purposes.
  + **values**: This converts the resulting DataFrame into a NumPy array, making it easier to manipulate for machine learning tasks (e.g., scikit-learn models often require NumPy arrays as input). As [['latitude']] is a DataFrame, .values will return a **2D NumPy array** with shape (n\_samples, 1)
  + **reshape(-1, 1)**: This reshapes the array into a 2D array with a single column. The -1 means "adjust the number of rows automatically," and 1 sets the number of columns to 1. This reshaping is needed because many machine learning models require the input features (X) and the target (y) to be in a 2D array format.

# Section 7.3

* In the **residual vs. predicted plot**, the behavior of residuals can help assess whether the **linearity** and **constant variance (homoscedasticity)** assumptions are met in a linear regression model.

**1. Linearity Assumption**

The **linearity assumption** means that the relationship between the independent variables (predictors) and the dependent variable is linear. In a residual vs. predicted plot:

* If the model meets the linearity assumption, the residuals (the vertical distance from the actual data points to the regression line) should be randomly scattered around zero, showing **no systematic pattern**.
* If you see a clear pattern (e.g., a curve, wave, or systematic shape 🡪so called non linear pattern), it suggests that the relationship between the predictors and the outcome is not purely linear. This indicates a violation of the linearity assumption and suggests that a different model (e.g., polynomial regression) might be more appropriate.

**What to look for in the plot:**

* **No clear pattern**: A "cloud" of points scattered randomly around the horizontal axis (zero line) suggests that the linearity assumption holds.
* **Systematic pattern** (e.g., curved or U-shaped pattern): Indicates that the model might need to account for a non-linear relationship between the predictors and the outcome.

**2. Constant Variance Assumption (Homoscedasticity)**

The **constant variance assumption** (also known as homoscedasticity) means that the spread of residuals should remain constant across all predicted values. In other words, the residuals should have the same variance regardless of the value of the predicted outcomes. Violating this assumption leads to **heteroscedasticity**, where the residuals' spread increases or decreases as a function of the predicted values.

In a residual vs. predicted plot:

* If the constant variance assumption is satisfied, the residuals should be evenly spread across the range of predicted values, showing a uniform "cloud."
* If you observe a **funnel shape** (e.g., residuals spread out more as predicted values increase or decrease), this indicates heteroscedasticity, meaning the model's errors vary with the level of the predicted variable.

**What to look for in the plot:**

* **Even spread of residuals**: **Residuals should have a similar spread above and below the zero line across all predicted values.**
* **Funnel shape**: If the residuals are tightly clustered for smaller predicted values and then widen as predicted values increase (or vice versa), this indicates heteroscedasticity and a violation of the constant variance assumption.

A graph of a function

Description automatically generated with medium confidence

# Violation of the linearity assumption

If the **linearity assumption holds**, the residuals vs fitted values plot will show **no discernible patterns**—just a **random cloud** of points centered around zero. This suggests that a linear model is appropriate for the data.

**Any pattern** (linear or non-linear) in the residuals vs fitted values plot means the **linearity assumption is violated**.

The presence of a pattern suggests the linear regression model is **inadequate** and needs adjustment:

* + For **non-linear patterns (e.g., Curve, U-Shape, or Waves),** , consider transformations or non-linear models.
  + For **linear patterns (e.g, linear line or two flat lines)**, check for missing interactions, variables, or groups in your data.

In short, the presence of any **pattern** in the residual plot indicates that the current model is not fully capturing the underlying relationship, and you should investigate the structure of your data or consider alternative modeling approaches.

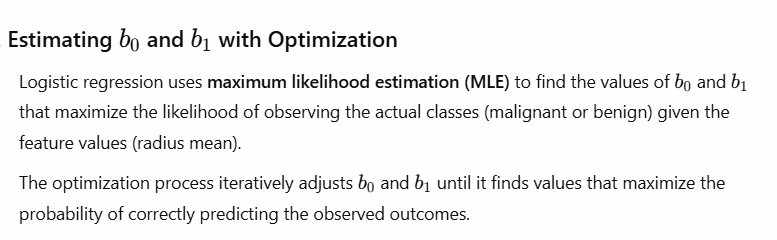
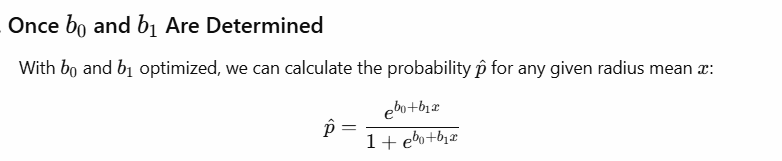
**If you observe a random scatter in the residuals vs fitted values plot, you can conclude:**

1. **Linearity assumption holds**:
   * A random scatter of residuals suggests that the linear model is appropriate for capturing the relationship between the dependent and independent variables. There are no systematic patterns that would indicate a non-linear relationship.
2. **Constant variance (Homoscedasticity) assumption holds**:
   * If the residuals are scattered evenly around the horizontal line (zero) and the spread of residuals is roughly constant across all fitted values, this suggests that the **constant variance (homoscedasticity)** assumption holds. If you see increasing or decreasing spread (a funnel shape), it would indicate **heteroscedasticity**.
3. **Maybe, the independence assumption holds, but there is no guarantee**:
   * **The independence assumption** means
     1. Residuals from one observation should not be correlated with residuals from another observation.
     2. Independence is especially important when observations are sequential (such as in time series data), where errors can potentially be correlated over time.
   * A random scatter **could suggest** that the independence assumption is not obviously violated, but this is not a definitive test for independence. To conclusively check the **independence of residuals** (especially for time-series or sequential data), you would need to apply additional tests such as Durbin Watson test, autocorrelation function plot, etc.

# Section 7.5 : Log Regression

A math equations with black text

Description automatically generated



Summary:

Typically, we do not start with the probability. Instead, log regression finds b0 and b1 using the actual binary outcome and input features. Once we have b0 and b1, we use them to calculate the probability for any given feature value x.

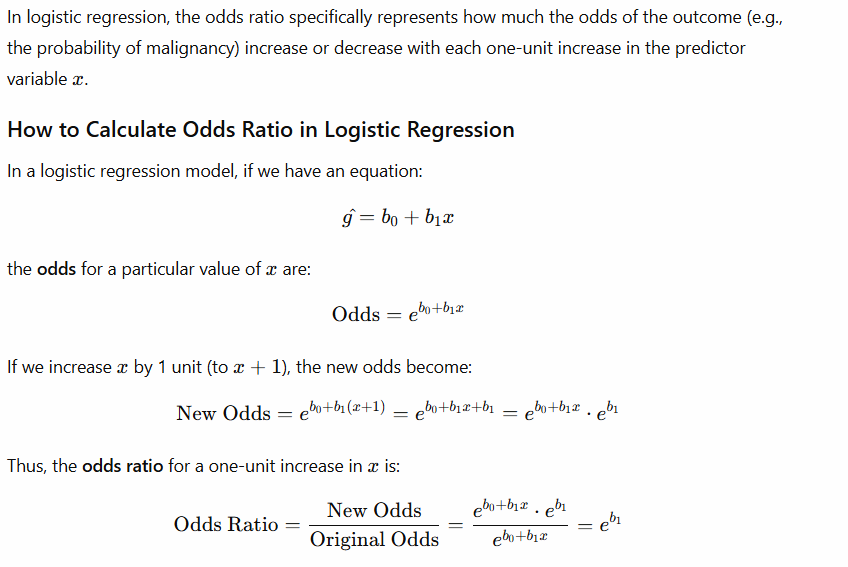
## Intercept of the curve

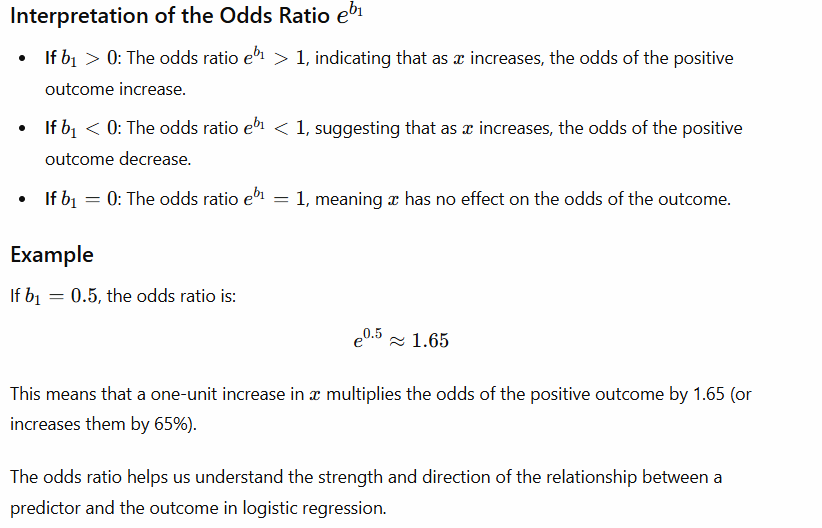
A white background with black text

Description automatically generated

Note: Sometime, the intercept of the curve is referred to baseline probability.

## Odds Ratio





Lecture note begins for the section 7.5:

A screenshot of a white paper with black text

Description automatically generated

Example: if p(head) = 0.1, then odds of heads = 0.1/0.9 = 1/9. This means that for every 1 expected occurrence of heads, we would expect 9 occurrences of tails over a large # of tosses (i.e., 1:9 ratio).

1. **Express the probability (helps bridge the understanding from log-odds back to the probability)**

A white background with black text

Description automatically generated

A graph on a white background

Description automatically generated

A graph of a logistic curve

Description automatically generated

A graph with a line

Description automatically generated

The above diagram demonstrates how a larger positive b1 results in a steeper S-curve, while a smaller positive b1 produces a more gradual slope.

**Odds Ratio**

**A white paper with black text and black text

Description automatically generated**

**A math equations and equations

Description automatically generated with medium confidence**