Honors - ML & DS (U-1) Notes by OMG

1. Visualizing Fitting Logistic Regression Curves

Visualizing the fit of a logistic regression curve is essential for understanding how the model predicts a binary outcome. Unlike linear regression which fits a straight line, logistic regression fits an **"S"-shaped" or sigmoid curve**. This curve represents the probability of the outcome being a '1' (or success) for each value of the independent variable.

The visualization process involves plotting the independent variable on the x-axis and the probability of the outcome (from 0 to 1) on the y-axis. The actual data points for the binary outcome are also plotted, typically at y=0 and y=1. The logistic regression model then fits the sigmoid curve that best separates these two sets of points.

Key elements to observe in the visualization are:

- **The Sigmoid Shape**: This "S" shape is the hallmark of the logistic function. It demonstrates how the predicted probability changes with the independent variable—changing slowly at the extremes and more rapidly in the middle.
- **The Decision Boundary**: This is a threshold, usually set at a probability of 0.5, that is used to classify the outcome. If the model predicts a probability above this threshold for a given observation, it's classified as '1'; otherwise, it's classified as '0'. This threshold can be adjusted depending on the specific application.
- **Data Points**: The raw data points (the actual 0s and 1s) are also shown on the plot. A well-fitting curve will be closer to 1 for the data points where the actual outcome was 1 and closer to 0 for those where the outcome was 0.

---

2. Estimation in Logistic Regression and Poisson Regression

i) Estimation in Logistic Regression

Estimation in logistic regression is the process of finding the optimal values for the model's coefficients. The most common method used is **Maximum Likelihood Estimation (MLE)**.

MLE works by finding the coefficient values that maximize the probability of observing the actual outcomes in the dataset. In essence, it answers the question: "What coefficient values would make the observed data most likely?"

Example: Suppose we want to predict if a student will pass (1) or fail (0) an exam based on the hours they studied. With data from several students, MLE would test different coefficient values for the 'hours studied' variable until it finds the value that makes the observed pass/fail results most probable.

ii) Poisson Regression

Poisson regression is a statistical method used for modeling **count data**. The dependent variable in a Poisson regression model is a count of events, such as the number of occurrences of an event within a specific time frame or area. This model assumes that the response variable follows a Poisson distribution.

Example: A city's traffic department might want to model the number of accidents at a particular intersection per day. The number of accidents is count data. They could use Poisson regression to analyze how factors like traffic volume, weather conditions, and the time of day influence the number of accidents.

---

3. Statistical Linear Model and Its Types

A **statistical linear model** is a mathematical approach to modeling the relationship between a dependent variable and one or more independent variables. The term "linear" signifies that the model is linear in its parameters, not necessarily in the relationship between the variables themselves.

The general form of a simple linear model is:

$Y=\beta 0+\beta 1X+\epsilon$

Where:

- Y is the dependent variable.
- X is the independent variable.
- β0 is the intercept.
- β1 is the slope.

- ϵ is the error term, representing the variability in Y not explained by X.

Types of Statistical Linear Models:

1. **Simple Linear Regression**: Involves a single independent variable predicting a continuous dependent variable.
   - **Example**: Predicting a person's salary based on their years of experience.
2. **Multiple Linear Regression**: Uses two or more independent variables to predict a continuous dependent variable.
   - **Example**: Predicting a home's price based on its size, number of bedrooms, and age.
3. **Analysis of Variance (ANOVA)**: Used to compare the means of two or more groups, where the independent variables are categorical.
   - **Example**: A medical researcher comparing the effectiveness of three different drugs on blood pressure.
4. **Analysis of Covariance (ANCOVA)**: A combination of ANOVA and regression, used to compare group means while controlling for the effects of other continuous variables (covariates).
   - **Example**: In the drug effectiveness study, ANCOVA could be used to compare the drugs' effects while controlling for the patients' initial blood pressure levels.
5. **Multivariate Linear Regression**: Involves more than one dependent variable.
   - **Example**: Predicting a student's scores in both math and science based on their study hours and attendance.

---

4. Difference Between a Regression Model and an Estimated Regression Equation

Feature	Regression Model	Estimated Regression Equation
Nature	A theoretical representation of the relationship between variables for an entire population.	A concrete equation derived from a sample of data.
Components	Includes a random error term (ϵ) to account for unexplained variation.	Does not have a random error term.

Parameters	Contains unknown population parameters (e.g., β0, β1).	Contains estimated parameters (e.g., β^0, β^1).
Purpose	To describe the true, underlying relationship.	To estimate the true relationship and make predictions.
Notation	Uses Greek letters for population parameters (e.g., $Y=\beta0+\beta1X+\epsilon$).	Uses "hats" to denote estimates (e.g., $\hat{Y}=\hat{\beta}0+\hat{\beta}1X$).
Variability	The error term signifies that actual values will vary around the true line.	Provides a single predicted value for a given set of inputs.
Derivation	A theoretical construct based on assumptions.	Calculated from sample data using methods like Ordinary Least Squares.
Uniqueness	There is only one true regression model for a given population.	Different samples from the same population will yield different estimated equations.

---

5. Difference Between Simple Linear and Multiple Linear Regression

Simple and multiple linear regression are both statistical techniques used to model the relationship between variables. The primary distinction lies in the number of independent variables used in the model. [11]

Here are 8 key differences:

Feature	Simple Linear Regression	Multiple Linear Regression
Independent Variables	Uses a **single** independent variable (X) to explain or predict the outcome of a dependent variable (Y).	Uses **two or more** independent variables (X_1, X_2, ..., X_n) to explain or predict the outcome of a dependent variable (Y).
Equation	The relationship is modeled by the equation: $Y=\beta0+\beta1X+\epsilon$	The relationship is modeled by the equation: $Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$
Objective	To understand and quantify the linear relationship between two continuous variables.	To understand the linear relationship between a dependent variable and *several* independent variables simultaneously.
Visualization	The model can be visualized as a **straight line** on a two-dimensional scatter plot.	The model is represented by a **plane** (for two independent variables) or a **hyperplane** (for more than two) in a multi-dimensional space, which is difficult to visualize directly.
Coefficient Interpretation	The slope coefficient ($\beta1$) represents the change in the dependent variable for a one-unit change in the single independent variable.	Each coefficient (βi) represents the *average* change in the dependent variable for a one-unit change in one independent variable, **while holding all other independent variables constant**.
Complexity	Simpler to calculate, understand, and interpret.	More complex due to the potential interactions between multiple independent variables.

Key Concern	The primary focus is on the strength and direction of the relationship between the two variables.	A major additional concern is **multicollinearity**, which occurs when independent variables are highly correlated with each other, making it difficult to isolate their individual effects.
Goodness-of-Fit	**R-squared** (R2) is commonly used to measure the proportion of variance in the dependent variable explained by the model.	**Adjusted R-squared** is often preferred because it accounts for the number of predictors in the model, providing a more accurate measure of model fit as more variables are added.

---

6. What is a Residual?

A **residual** is the difference between the actual observed value of the dependent variable and the value predicted by the regression model. In simpler terms, it is the

prediction error for a single data point. [222]

Each data point in your dataset has a residual. A positive residual means the model underpredicted the actual value, while a negative residual means it overpredicted the value. The goal of a regression model is to find the line (or plane) that minimizes the sum of these squared errors.

Computation:

The formula to compute the residual (ei) for the i-th observation is:

ei=yi−y^i
Where:

- ei is the residual for the i-th observation.
- yi is the **actual (observed)** value of the dependent variable for the i-th observation.
- y^i is the **predicted (fitted)** value of the dependent variable for the i-th observation, calculated from the regression equation.

---

7. Odds and Odds Ratio

Odds and **Odds Ratios** are concepts typically used in logistic regression to understand the relationship between variables when the outcome is binary (e.g., success/failure, yes/no). [3]

What are Odds?

The **Odds** of an event occurring is the ratio of the probability that the event will happen to the probability that it will not happen.

Formula:

Odds=1−P(event)P(event)

- Example: If the probability of a team winning a match is 80% (or 0.80), the probability of them not winning is 20% (or 0.20). The odds of them winning are:

  Odds=0.200.80=4

  This is often expressed as "4 to 1 odds".

What is an Odds Ratio (OR)?

The **Odds Ratio** compares the odds of an event occurring in one group to the odds of it occurring in another group. It is a measure of the strength of association between an exposure (e.g., a treatment, a risk factor) and an outcome.

Formula:

Odds Ratio=Odds of event in Group BOdds of event in Group A
Interpretation of the Odds Ratio:

The interpretation depends on whether the OR is greater than, less than, or equal to 1:

- **OR > 1**: Indicates that the exposure is associated with **higher odds** of the outcome. For instance, if the OR is 3, the odds of the outcome occurring in the exposed group are three times the odds of it occurring in the unexposed group.
- **OR = 1**: Indicates that the exposure **does not affect** the odds of the outcome. There is no association.
- **OR < 1**: Indicates that the exposure is associated with **lower odds** of the outcome. This

suggests a "protective" effect. For instance, if the OR is 0.4, the odds of the outcome in the exposed group are 60% lower than in the unexposed group.

---

8. Regression, Residuals, and Regression Inference

These three terms are fundamental concepts in regression analysis. [4]

Regression

Regression analysis is a powerful statistical method used to model and investigate the relationship between a **dependent variable** (the outcome you want to predict) and one or more **independent variables** (the factors that are used to make the prediction). The primary goals of regression are:

1. **To understand** the nature and strength of the relationship between variables.
2. To predict the value of the dependent variable based on the values of the independent variables.
   The output of this analysis is a regression equation that mathematically describes this relationship.

Residuals

As detailed in question 2, a **residual** ($e=y-\hat{y}$) is the error in a prediction, specifically the difference between the observed value and the value predicted by the regression model. Analyzing the pattern of residuals is a critical diagnostic step. A good model will have residuals that are randomly scattered around zero with no clear pattern. Patterns in residuals can indicate that the model is not a good fit for the data or that certain assumptions of the regression have been violated.

Regression Inference

Regression inference is the process of using the results from a sample to draw conclusions about the true relationship between variables in the entire population. It addresses the uncertainty in the model's estimates. Key components include:

- **Hypothesis Testing:** This is used to determine if the relationship observed in the sample data is statistically significant or if it could have occurred by chance. For each independent variable, we test the null hypothesis that its coefficient is zero ($H0:\beta i=0$), meaning it has no effect on the dependent variable. A small p-value (typically < 0.05) allows us to reject this hypothesis.
- **Confidence Intervals:** A confidence interval provides a range of plausible values for a

population regression coefficient, based on the sample data. For example, a 95% confidence interval for a coefficient suggests we are 95% confident that the true population coefficient lies within that range. If the interval does not contain zero, it provides further evidence that the variable has a significant effect.

---

9. Poisson Regression and Logistic Regression

These are both types of Generalized Linear Models, used when the assumptions of ordinary linear regression are not met, particularly with respect to the dependent variable. [5]

Logistic Regression

- **When to Use:** Logistic regression is used when the dependent variable is **binary** or **categorical**. The outcome can have only two (or a limited number of) possible values, such as Yes/No, Pass/Fail, Spam/Not Spam, or Healthy/Diseased.
- **What it Models:** It doesn't model the value of the outcome directly. Instead, it models the **probability** that the outcome belongs to a particular category. Specifically, it models the *log-odds* of the outcome as a linear combination of the independent variables.
- The Equation:

  $$\ln(1-PP) = \beta 0 + \beta 1 X 1 + \cdots + \beta n Xn$$

  Where P is the probability of the event occurring. The term $\ln(1-PP)$ is called the logit.
- **Interpretation:** The coefficients (β) are interpreted in terms of the change in log-odds. By exponentiating a coefficient ($e\beta$), we obtain the **Odds Ratio**, which is a more intuitive measure of how a one-unit change in an independent variable affects the odds of the outcome occurring.

Poisson Regression

- **When to Use:** Poisson regression is used when the dependent variable is a **count** of events occurring over a fixed interval of time or space. The data must be non-negative integers (0, 1, 2, 3, ...). Examples include the number of customer complaints per day, the number of accidents at an intersection per month, or the number of weeds per square meter of a field.
- **What it Models:** It assumes the dependent variable follows a Poisson distribution. It models the *log of the expected count* (or rate) as a linear combination of the independent variables.
- The Equation:

  $$\ln(\lambda) = \beta 0 + \beta 1 X 1 + \cdots + \beta n Xn$$

Where λ (lambda) is the expected count or rate of the event.

- **Key Assumption:** A core assumption of the Poisson model is that the **mean and variance of the count variable are equal** (a property called equidispersion). When the variance is much larger than the mean (overdispersion), a different model like Negative Binomial Regression may be more appropriate.

---

10. Linear Models in Statistics

A **linear model** in statistics is any model that describes a dependent variable as a linear combination of predictor variables and parameters. The term "linear" refers to the fact that the model is linear in its parameters (β), not necessarily in its variables.

Here are two commonly used linear models explained with examples: [6]

1. Analysis of Variance (ANOVA)

- **Purpose:** ANOVA is used to compare the means of **three or more groups** to determine if there is a statistically significant difference between them. The independent variable is categorical (defining the groups), and the dependent variable is continuous.
- **How it Works:** It analyzes the variance in the data by partitioning the total variability into two parts: the variability *between* the groups and the variability *within* the groups. It then calculates an **F-statistic**, which is the ratio of the between-group variance to the within-group variance. If this ratio is significantly large, it implies that the difference between the group means is not due to random chance.
- **Suitable Example:** An agricultural scientist wants to determine if different types of fertilizer lead to different crop yields.
    - **Dependent Variable (Continuous):** Crop yield (in kilograms per acre).
    - **Independent Variable (Categorical):** Fertilizer type (Type A, Type B, Type C, Control Group).
    - **Application:** The scientist would apply each fertilizer to different plots of land and measure the resulting yield. ANOVA would then be used to compare the mean yield across the four groups. A significant F-test would indicate that at least one fertilizer type results in a different average yield compared to the others.

2. Multiple Linear Regression

- **Purpose:** As described earlier, multiple linear regression is used to model the relationship between a **continuous dependent variable** and **two or more independent variables**. It helps in understanding the collective effect of several variables on an outcome and can be used for prediction.

- **How it Works:** The model fits a linear equation (a hyperplane in multi-dimensional space) to the data that best predicts the dependent variable from the independent variables. The method of "least squares" is used to find the coefficients (β) that minimize the sum of the squared residuals.
- **Suitable Example:** A financial analyst wants to predict the stock price of a tech company.
    - **Dependent Variable (Continuous):** Stock Price.
    - **Independent Variables (Continuous/Categorical):**
        - Quarterly Revenue (in millions)
        - Profit Margin (in percentage)
        - Number of Active Users (in millions)
        - Competitor's Stock Price
    - Application: Using historical data for these variables, the analyst can build a multiple regression model:
      $\text{Stock Price} = \beta_0 + \beta_1(\text{Revenue}) + \beta_2(\text{Profit Margin}) + \cdots + \epsilon$
      This model can help determine which factor has the most significant impact on the stock price and can be used to forecast future prices based on predictions for the independent variables.