Logistic Regression (LR)

Prof. Manoel Gadi (mfalonso@faculty.ie.edu)

This material covers logistic regression. It's a **supervised classification** technique, meaning we train the model on labeled data to predict the likelihood of a specific outcome. It is a powerful statistical method used to model the probability of a binary outcome based on one or more predictor variables. It is a versatile technique widely employed in various fields, including marketing, finance, healthcare, and social sciences. Logistic regression is particularly useful for tasks such as predicting customer churn, assessing credit risk, identifying disease risk factors, and understanding the impact of social factors on behavior.

Key Concepts in Logistic Regression:

- Sigmoid Function: This function maps any real value to a value between 0 and 1, representing the probability of the positive class.
- Log Odds: The log of the odds ratio, which is the ratio of the probability of success to the probability of failure, Log(odds) ranges between -infinity and +infinity.
- Maximum Likelihood Estimation (MLE): A statistical method used to find the parameters of a model that maximize the likelihood of observing the data

Discriminant Analysis vs. Logistic Regression: Key Similarities and Differences

Similarities:

- **Both are classification techniques:** They are used to predict the category or class of a data point based on its features or attributes.
- **Both handle multiple independent variables:** They can consider multiple factors to make predictions.

Differences:

- **Underlying Assumptions:**
  - **Discriminant Analysis:** Assumes that the independent variables follow a multivariate normal distribution and that the covariance matrices of the different classes are equal.
  - **Logistic Regression:** Makes fewer assumptions and is more robust to violations of these assumptions.
- **Model Estimation:**

  - **Discriminant Analysis:** Estimates the parameters of the model using a statistical approach based on sample means and covariance matrices.
  - **Logistic Regression:** Estimates the parameters using maximum likelihood estimation, which is more flexible and can handle non-linear relationships between the predictors and the outcome.
- **Model Interpretation:**
  - **Discriminant Analysis:** Provides a linear discriminant function that can be used to classify new observations.
  - **Logistic Regression:** Provides probabilities of class membership, which can be more informative in some cases.

Regarding the specific question:

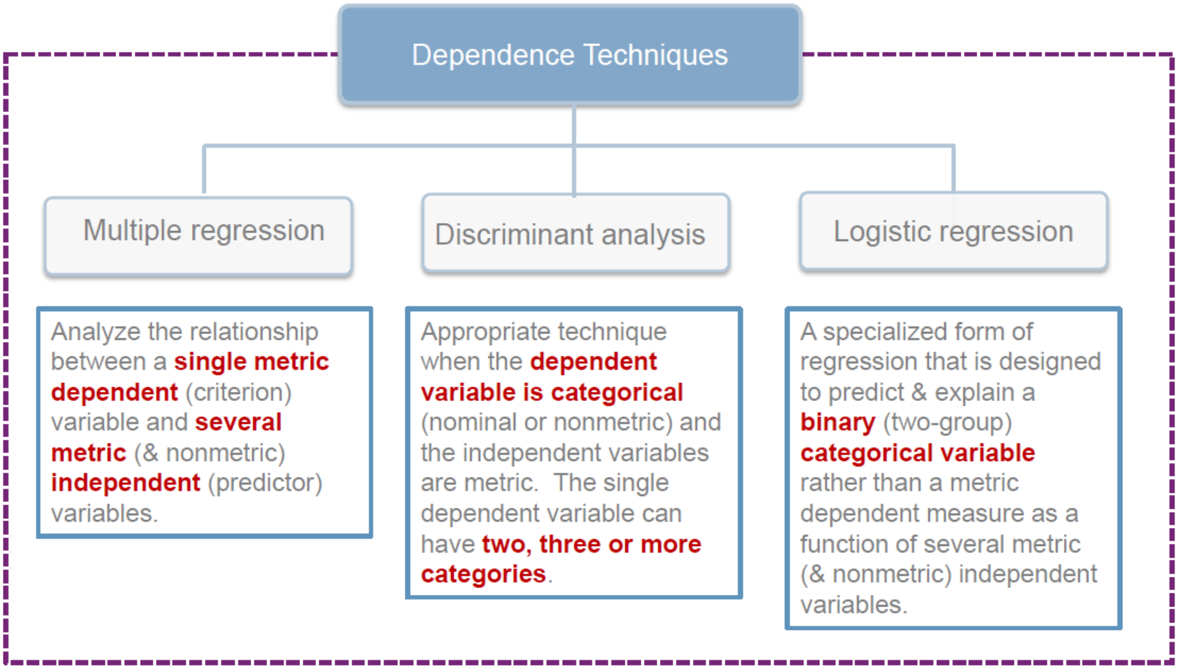
In practice, logistic regression is often preferred over discriminant analysis because it is more flexible and robust to violations of assumptions. However, discriminant analysis can be a useful tool when the assumptions are met and when the goal is to obtain a simple linear discriminant function. From now on, we will cover Logistic Regression more in detail.

When to use Logistic Regression?

- Binary Classification Problems: LR is well-suited for handling classification problems where the dependent variable has two possible outcomes (e.g., yes/no, 1/0, true/false).

LR are well-suited for handling:

- Predicting Probabilities: LR can estimate the probability of an event occurring, allowing for nuanced decision-making.
- Identifying Key Predictors: LR helps identify the most important factors influencing the outcome.
- Understanding Relationships: LR can reveal the nature of the relationship between independent and dependent variables.



Dependence Techniques

Multiple regression

Analyze the relationship between a **single metric dependent** (criterion) variable and **several metric** (& nonmetric) **independent** (predictor) variables.

Discriminant analysis

Appropriate technique when the **dependent variable is categorical** (nominal or nonmetric) and the independent variables are metric. The single dependent variable can have **two, three or more categories**.

Logistic regression

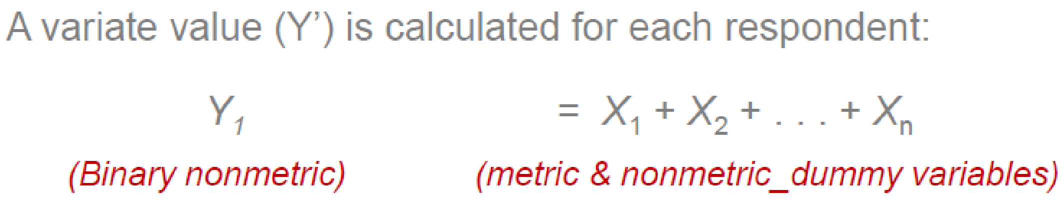
A specialized form of regression that is designed to predict & explain a **binary** (two-group) **categorical variable** rather than a metric dependent measure as a function of several metric (& nonmetric) independent variables.

The image presents a hierarchical classification of statistical techniques known as **Dependence Techniques**. These techniques are used to analyze relationships between variables.

Breakdown:

- **Multiple Regression:**
  - Analyzes the relationship between a **single metric dependent variable** (like income or height) and **several metric or nonmetric independent variables** (like age, education, gender).
  - It aims to predict the value of the dependent variable based on the independent variables.
- **Discriminant Analysis:**
  - Appropriate when the **dependent variable is categorical** (nominal or nonmetric), meaning it can have two or more categories (like gender, product preference).
  - The independent variables can be metric or nonmetric.
  - It seeks to classify observations into the correct category based on their values on the independent variables.
- **Logistic Regression:**

- A specialized form of regression designed for predicting and explaining a **binary (two-group) categorical dependent variable**.



A variate value (Y') is calculated for each respondent:

$$Y_1 = X_1 + X_2 + . . . + X_n$$

(Binary nonmetric) *(metric & nonmetric_dummy variables)*

- The independent variables can be metric or nonmetric.
- It models the probability of the outcome belonging to one of the two categories based on the values of the independent variables.

Key Points:

- The choice of technique depends on the nature of the dependent variable.
- Multiple Regression is used for continuous dependent variables.
- Discriminant Analysis is used for categorical dependent variables with two or more categories.
- Logistic Regression is used for binary categorical dependent variables.

In Summary:

The image provides a clear overview of three key dependence techniques: Multiple Regression, Discriminant Analysis, and Logistic Regression. By understanding their distinctions and applications, you can select the appropriate technique for your specific research question and data.

Use Cases and Industries:

Logistic Regression: A Deeper Dive into Binary Predictions

Logistic Regression is a powerful statistical method used to model the probability of a binary outcome based on one or more predictor variables. It's a versatile technique with a wide range of applications.

Key Applications of Logistic Regression:

1. Predicting Gender:

- Input Features: Demographic information (age, location, occupation), browsing behavior, purchase history, etc.
- Output: Probability of being male or female.
- Use Cases: Targeted marketing, personalization, and understanding customer preferences.

2. Identifying Heavy vs. Light Users:

- Input Features: Usage frequency, time spent, purchase volume, etc.
- Output: Probability of being a heavy user.
- Use Cases: Customer segmentation, loyalty programs, and product recommendations.

3. Predicting Purchase Behavior:

- Input Features: Demographic information, browsing history, cart abandonment rate, etc.
- Output: Probability of making a purchase.
- Use Cases: Targeted advertising, email campaigns, and inventory management.

4. Assessing Credit Risk:

- Input Features: Credit history, income, debt, employment status, etc.
- Output: Probability of defaulting on a loan.
- Use Cases: Credit scoring, risk management, and fraud detection.

5. Identifying Potential Members:

- Input Features: Demographic information, interests, browsing behavior, etc.
- Output: Probability of becoming a member.

- Use Cases: Membership recruitment, targeted marketing, and community building.

How Logistic Regression Works:

1. Data Collection: Gather relevant data for both the independent (predictor) variables and the dependent (binary) variable.
2. Model Training: Use statistical techniques to estimate the coefficients of the logistic regression model.
3. Probability Prediction: For a given set of input features, the model calculates the probability of the positive outcome (e.g., being male, being a heavy user).
4. Classification: Based on a predefined threshold (e.g., 0.5), the model classifies the observation as belonging to one of the two categories.

Advantages of Logistic Regression:

- Interpretability: The coefficients of the model can be interpreted to understand the impact of each predictor variable on the outcome.
- Efficiency: It can handle large datasets and multiple predictor variables.
- Versatility: It can be applied to various fields, including finance, marketing, healthcare, and social sciences.

By understanding the underlying principles and applications of logistic regression, you can effectively leverage this powerful tool to make data-driven decisions and gain valuable insights from your data.

Assumptions

Key Assumptions of Logistic Regression:

1. Independence of Observations: Each observation should be independent of the others. This means that the outcome for one observation should not influence the outcome for another.
2. Binary Dependent Variable: The dependent variable should be binary, meaning it can take on only two values (e.g., 0 or 1, yes or no).

While it's true that logistic regression doesn't require:

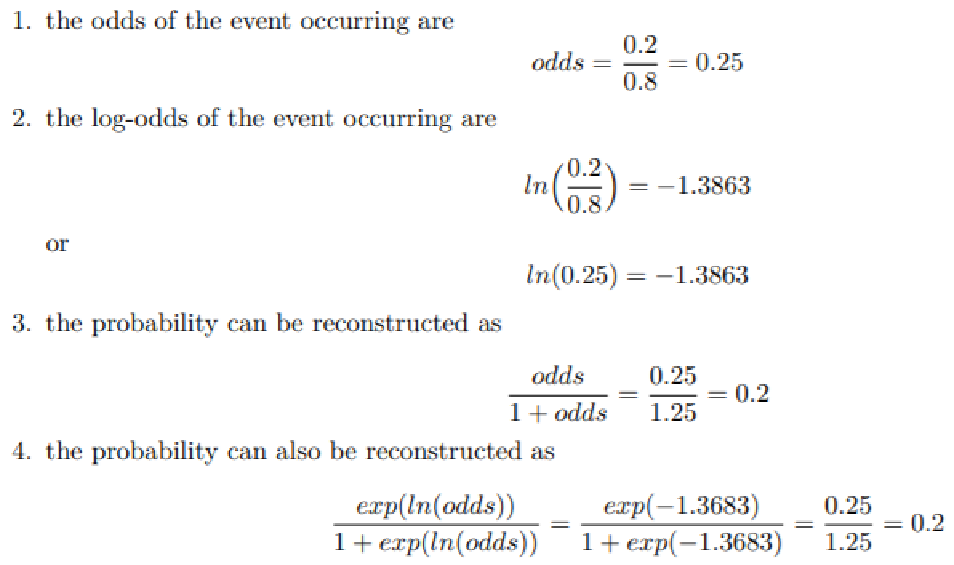
- Heteroscedasticity of Independent Variables: This is a common misconception. Logistic regression inherently assumes heteroscedasticity, as the variance of the error term changes with the predicted probability.
- Linear Relationships Between Dependent and Independent Variables: Logistic regression models a non-linear relationship between the log-odds of the outcome and the linear combination of the independent variables.

Other Important Considerations:

- Multicollinearity: High correlation among independent variables can affect the model's stability and interpretation.
- Outliers: Outliers can significantly influence the model's results.
- Sample Size: A sufficient sample size is necessary to ensure reliable estimates.

By understanding and addressing these assumptions, you can build more robust and accurate logistic regression models.

Odds and Probability relationships



1. the odds of the event occurring are

$$odds = \frac{0.2}{0.8} = 0.25$$

2. the log-odds of the event occurring are

$$ln\left(\frac{0.2}{0.8}\right) = -1.3863$$

or

$$ln(0.25) = -1.3863$$

3. the probability can be reconstructed as

$$\frac{odds}{1 + odds} = \frac{0.25}{1.25} = 0.2$$

4. the probability can also be reconstructed as

$$\frac{exp(ln(odds))}{1 + exp(ln(odds))} = \frac{exp(-1.3683)}{1 + exp(-1.3683)} = \frac{0.25}{1.25} = 0.2$$

1. Odds of the Event Occurring

- The odds of an event occurring are calculated as the ratio of the probability of the event happening (success) to the probability of the event not happening (failure).
- In this case, the odds are given as 0.2/0.8, which simplifies to 0.25.

2. Log-Odds of the Event Occurring

- The log-odds (or logit) is the natural logarithm of the odds.
- It transforms the odds ratio into a linear scale, making it easier to work with in statistical models.
- Here, the log-odds is calculated as the natural logarithm of 0.25, which is approximately -1.3863.

3. Reconstructing the Probability from Odds

- The probability of the event can be recovered from the odds using the following formula:

Probability = Odds / (1 + Odds)

- In this case, plugging in the odds of 0.25, we get:

Probability = 0.25 / (1 + 0.25) = 0.2

4. Reconstructing the Probability from Log-Odds

- Alternatively, we can reconstruct the probability directly from the log-odds using the inverse logit function (also known as the logistic function):

Probability = exp(Log-Odds) / (1 + exp(Log-Odds))

- Using the log-odds of -1.3863, we get:

Probability = exp(-1.3863) / (1 + exp(-1.3863)) ≈ 0.2

In essence, the image demonstrates the relationship between odds, log-odds, and probability, and how they can be transformed into each other.

Python example:

```
# Import necessary libraries
%matplotlib inline
import pandas as pd
import statsmodels.api as sm
from patsy import dmatrices

# Load the Titanic dataset
df = pd.read_csv('/content/titanic_data.csv')

# Split the data into training and testing sets
df_train = df.iloc[0:600, :]  # First 600 rows for training
df_test = df.iloc[600:, :]     # Remaining rows for testing

# Define the logistic regression formula
formula = 'Survived ~ C(Pclass) + C(Sex) + Age + SibSp + C(Embarked) + Parch'

# Create design matrices for training and testing data
y_train, x_train = dmatrices(formula, data=df_train, return_type='dataframe')
y_test, x_test = dmatrices(formula, data=df_test, return_type='dataframe') 1

# Create a logistic regression model
model = sm.Logit(y_train, x_train)

# Fit the model to the training data
res = model.fit()

# Print the model summary
res.summary() ()
```

Interpreting the code above:

The code sets up a logistic regression model to predict passenger survival on the Titanic, trains the model on a portion of the data, and then prints a summary of the model's performance.

Interpreting the output:

```
Optimization terminated successfully.

         Current function value: 0.464182

         Iterations 6

Logit Regression Results

Dep. Variable: Survived       No. Observations:      473

Model:  Logit   Df Residuals:  464

Method: MLE     Df Model:      8

Date:   Thu, 28 Nov 2024       Pseudo R-squ.: 0.3150

Time:   15:32:53       Log-Likelihood:        -219.56

converged:     True    LL-Null:       -320.54

Covariance Type:       nonrobust      LLR p-value:   2.467e-39

coef    std err z       P>|z|   [0.025  0.975]

Intercept      3.7123  0.552   6.720   0.000   2.630   4.795

C(Pclass)[T.2] -0.8144 0.363   -2.242  0.025   -1.526  -0.102

C(Pclass)[T.3] -1.9752 0.349   -5.655  0.000   -2.660  -1.291

C(Sex)[T.male] -2.5689 0.256   -10.034 0.000   -3.071  -2.067

C(Embarked)[T.Q]       0.2204  0.724   0.304   0.761   -1.199  1.639

C(Embarked)[T.S]       -0.2345 0.334   -0.703  0.482   -0.888  0.419

Age     -0.0354 0.010   -3.598  0.000   -0.055  -0.016

SibSp   -0.3609 0.148   -2.445  0.014   -0.650  -0.072

Parch   0.0311  0.160   0.194   0.846   -0.283  0.346
```

Interpreting the Logistic Regression Output

Optimization Termination

- Optimization terminated successfully: Indicates that the model converged to an optimal solution.
- Current function value: This is the value of the log-likelihood function at the optimal solution. Lower values indicate a better fit.
- Iterations 6: The model converged after 6 iterations of the optimization algorithm.

Model Summary

- Dep. Variable: Survived: The dependent variable is whether a passenger survived or not.
- No. Observations: 473: The model was trained on 473 observations.
- Model: Logit: This is a logistic regression model.
- Method: MLE: Maximum Likelihood Estimation was used to fit the model.
- Pseudo R-squared: 0.3150: This is a pseudo R-squared value, which is a measure of model fit, similar to R-squared in linear regression. Higher values indicate a better fit.
- Log-Likelihood: -219.56: The log-likelihood of the model.
- LL-Null: -320.54: The log-likelihood of a null model (intercept only).
- LLR p-value: 2.467e-39: The likelihood ratio test p-value, which tests the overall significance of the model. A very small p-value indicates that the model is statistically significant.

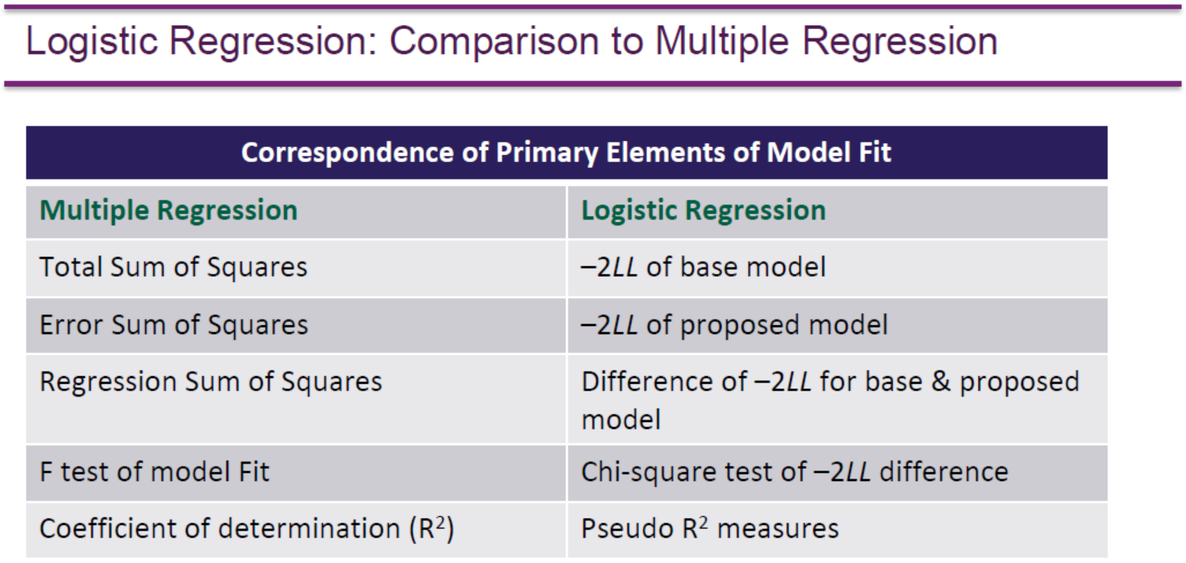
Coefficients and Significance

- Intercept: 3.7123: This is the baseline log-odds of survival.
- C(Pclass)[T.2], C(Pclass)[T.3]: These coefficients represent the log-odds difference in survival for passengers in class 2 and 3 compared to class 1 (the reference category). Negative coefficients indicate lower odds of survival.
- C(Sex)[T.male]: This coefficient represents the log-odds difference in survival for males compared to females (the reference category). A negative coefficient indicates lower odds of survival for males.
- C(Embarked)[T.Q], C(Embarked)[T.S]: These coefficients represent the log-odds difference in survival for passengers who embarked at Queenstown and Southampton, respectively, compared to those who embarked at Cherbourg (the reference category).
- Age, SibSp, Parch: These coefficients represent the impact of age, number of siblings/spouses, and number of parents/children on the log-odds of survival.

The p-values associated with each coefficient indicate the statistical significance of that variable in predicting survival. Lower p-values (typically less than 0.05) indicate that the variable is statistically significant.

Metrics and Measures for model fit and validation

Metric for stage 4, model adjustment or fitting:



Logistic Regression: Comparison to Multiple Regression

Correspondence of Primary Elements of Model Fit

Multiple Regression	Logistic Regression
Total Sum of Squares	$-2LL$ of base model
Error Sum of Squares	$-2LL$ of proposed model
Regression Sum of Squares	Difference of $-2LL$ for base & proposed model
F test of model Fit	Chi-square test of $-2LL$ difference
Coefficient of determination (R^2)	Pseudo R^2 measures

The image above provided compares the key elements of model fit in multiple regression and logistic regression. Let's break down the correspondence between the two:

Multiple Regression

- Total Sum of Squares: Represents the total variation in the dependent variable.
- Error Sum of Squares: Represents the unexplained variation in the dependent variable.
- Regression Sum of Squares: Represents the variation in the dependent variable explained by the model.
- F test of model fit: Tests the overall significance of the model.
- Coefficient of determination (R^2): Measures the proportion of variance in the dependent variable explained by the model.

Logistic Regression

- -2LL of base model: Represents the -2 log-likelihood of a null model (intercept only). The concept corresponding to the Total Sum of Squares in multiple regression is the -2 log-likelihood of the base model. It does not use squares because it measures the goodness of fit based on the likelihood function, which involves logarithms.
- -2LL of proposed model: Represents the -2 log-likelihood of the full model with predictors.
- Difference of -2LL for base & proposed model: Represents the improvement in model fit due to the addition of predictors.
- Chi-square test of -2LL difference: Tests the significance of the improvement in model fit. The key difference between the F test and the Chi-square test of -2LL difference, more known as Hosmer-Lemeshow test, is that the F test is based on least squares estimation, while the Hosmer-Lemeshow test is based on maximum likelihood estimation.

Key Points:

- -2LL: The -2 log-likelihood is used in logistic regression because the dependent variable is binary (0 or 1). It measures the goodness of fit of the model to the data.

Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a model by maximizing the likelihood function. In the context of[1] logistic regression, the likelihood function represents the probability of observing the data given a specific set of parameter values.

How MLE Works for Logistic Regression:

1. Likelihood Function: The likelihood function, L(β), measures how likely the observed data is given a particular set of parameter values (β). For logistic regression, the likelihood function is the product of the probabilities of each observation:
2. L(β) = ∏ P(yi | xi, β)

where:

   - P(yi | xi, β) is the probability of the observed outcome yi (0 or 1) given the predictor variables xi and the parameters β.
   - β is the vector of parameters to be estimated.

3. Log-Likelihood Function: To simplify calculations and avoid numerical underflow, the log-likelihood function is often used:

4. log L(β) = ∑ log P(yi | xi, β)
5. Maximizing the Log-Likelihood: The goal of MLE is to find the values of β that maximize the log-likelihood function. This is typically done using iterative optimization algorithms like Newton-Raphson or gradient descent.
6. Interpreting the Parameters: The estimated parameters (coefficients) represent the impact of each predictor variable on the log odds of the outcome. By exponentiating these coefficients, we can obtain odds ratios, which quantify the change in the odds of the outcome for a one-unit increase in the predictor variable.

In Summary:

MLE is a powerful technique for estimating the parameters of logistic regression models. By maximizing the likelihood of observing the data, it provides a robust and efficient method for understanding the relationship between predictor variables and the binary outcome.

These metrics measure for stage 6, model validation:

Below we have the most common metrics used to evaluate the performance of classification models. Let's break down each one:

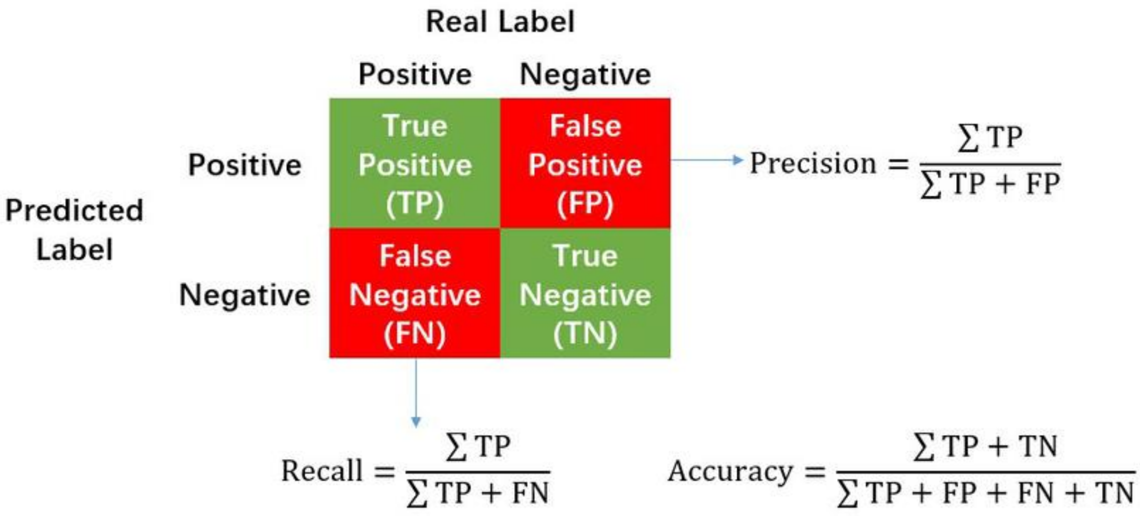
- Pseudo R^2 measures: Measures the proportion of variance in the log odds explained by the model. Pseudo R^2: Unlike R^2 in linear regression, pseudo R^2 measures in logistic regression are not directly comparable to R^2. They provide a relative measure of model fit. Accuracy and all its derivates below are generally more used in classification than Pseudo R-squared and they are more intuitive: They directly measures the proportion of correct predictions, making it easy to understand. Widely used: They are common metrics used across various classification algorithms and domains. Pseudo R-squared: Less intuitive: It's a measure of explained variation in the log-odds, which can be less straightforward to interpret. Limited range: It often has a smaller range compared to R-squared in linear regression, making it less sensitive to model improvements. Multiple versions: There are different versions of pseudo R-squared, each with its own interpretation and limitations.

More commonly used metrics in Classification Problems:

- Accuracy: The proportion of correctly classified data points.
- Precision: The proportion of true positives among predicted positives.
- Recall: The proportion of true positives identified correctly.

- F1-Score: A harmonic mean of precision and recall.
- AUC (Area Under the ROC Curve): A performance metric for classification models that uses the ROC curve.
- The KS statistic is a non-parametric metric used to compare the cumulative distribution functions (CDFs) of two samples. It's often used to determine if two samples come from the same distribution.

Breaking it down:


Real Label
Positive
Negative
Predicted Label
Positive
Negative
True Positive (TP)
False Positive (FP)
False Negative (FN)
True Negative (TN)
$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}}$
$\text{Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}}$
$\text{Accuracy} = \frac{\sum \text{TP} + \text{TN}}{\sum \text{TP} + \text{FP} + \text{FN} + \text{TN}}$

1. Accuracy

- Definition: The proportion of correct predictions made by the model.
- Formula: (TP + TN) / (TP + TN + FP + FN)
- Interpretation: A high accuracy score indicates that the model is making correct predictions most of the time. However, it can be misleading in imbalanced datasets.

2. Precision

- Definition: The proportion of positive predictions that are actually correct.
- Formula: TP / (TP + FP)
- Interpretation: A high precision score indicates that the model is making fewer false positive predictions.

3. Recall

- Definition: The proportion of actual positive cases that the model correctly identifies.
- Formula: TP / (TP + FN)
- Interpretation: A high recall score indicates that the model is good at identifying all positive cases.

4. F1-Score

- Definition: The harmonic mean of precision and recall.
- Formula: 2 * (Precision * Recall) / (Precision + Recall)
- Interpretation: A high F1-score indicates a balance between precision and recall. It's particularly useful when there's an imbalance between positive and negative classes.

5. AUC (Area Under the ROC Curve)

- Definition: The area under the Receiver Operating Characteristic (ROC) curve.
- Interpretation: A higher AUC indicates a better model. An AUC of 1.0 represents a perfect classifier, while an AUC of 0.5 represents a random classifier.

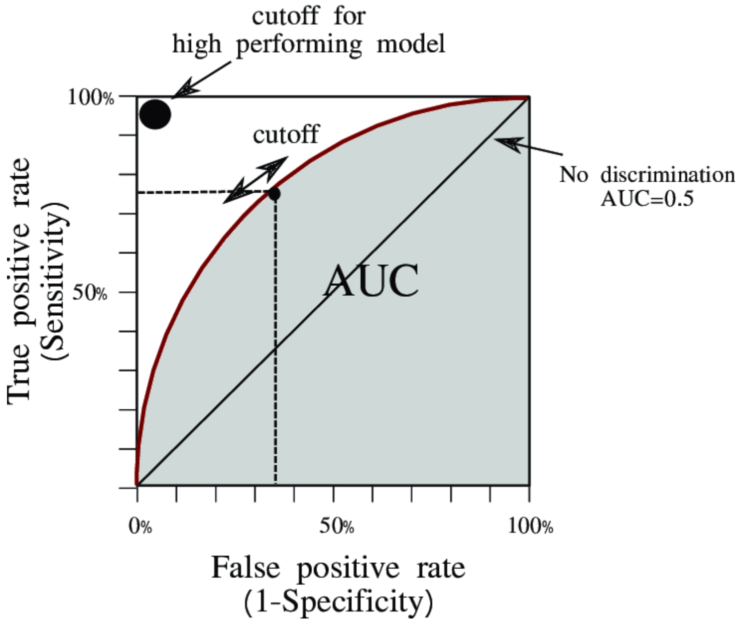
Understanding the ROC Curve and AUC

How the ROC Curve is Created:

1. Model Predictions: A classification model, like a decision tree, outputs a probability score for each data point, indicating the likelihood of it belonging to the positive class.
2. Thresholding: We set different threshold values to classify data points. For instance, if the probability score is above the threshold, it's classified as positive; otherwise, it's negative.
3. Calculating Metrics: For each threshold, we calculate:
    - True Positive Rate (TPR) or Sensitivity: The proportion of actual positive cases correctly identified as positive.
    - False Positive Rate (FPR) or 1-Specificity: The proportion of actual negative cases incorrectly identified as positive.
4. Plotting the Points: Each threshold value corresponds to a point on the ROC curve. The x-axis represents the FPR, and the y-axis represents the TPR.
5. Connecting the Points: The points are connected to form the ROC curve.

Interpreting the ROC Curve:

- Ideal Curve: A perfect classifier would have a curve that goes straight up to the top-left corner. This means it would correctly classify all positive instances and no negative instances.
- Random Guessing: A random classifier would have a diagonal line from the bottom-left corner to the top-right corner, indicating no discrimination between positive and negative classes.
- AUC Calculation: The Area Under the Curve (AUC) is the area under the ROC curve. It quantifies the overall performance of the model. A higher AUC indicates a better-performing model.



In the Image above:

- The red curve represents the ROC curve of a model.
- The shaded area represents the AUC.
- The two black dots on the curve indicate different threshold settings.

Key Points:

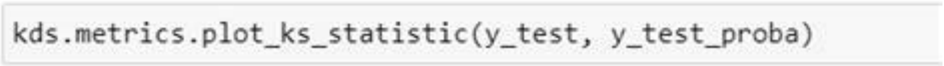
- The ROC curve helps visualize the trade-off between sensitivity and specificity at various threshold levels.
- A higher AUC indicates a better model, as it can discriminate between positive and negative classes more effectively.
- The choice of the optimal threshold depends on the specific use case and the desired balance between sensitivity and specificity.

By understanding the ROC curve and AUC, you can assess the performance of classification models specially when using imbalanced data where the false positive and negative costs are very different or unknow.

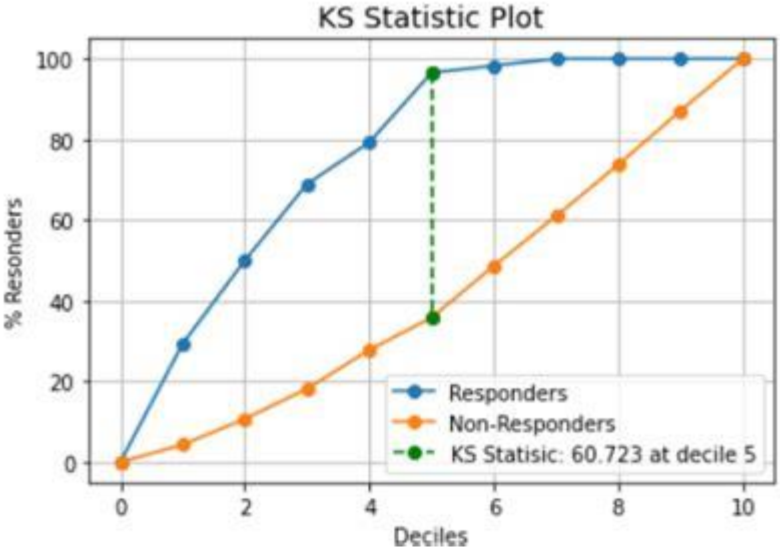
6. KS (Kolmogorov-Smirnov) Statistic

The KS statistic measures the maximum difference between the cumulative distribution functions (CDFs) of the positive and negative classes.

Interpretation: A higher KS statistic indicates a better model's ability to discriminate between the positive and negative classes. A KS value of 1.0 represents a perfect model, while a value of 0.0 represents a random model.



```
kds.metrics.plot_ks_statistic(y_test, y_test_proba)
```



Note: The KS statistic is often used in credit risk modeling and other financial applications to assess the predictive power of a model.

The image you provided shows a Kolmogorov-Smirnov (KS) statistic plot. This plot is used to evaluate the performance of a classification model, specifically its ability to discriminate between positive and negative classes.

Here's how to interpret the plot:

1. Cumulative Distribution Functions (CDFs):

- The blue line represents the cumulative distribution function (CDF) of the positive class (responders). It shows the percentage of responders at each decile.

- The orange line represents the CDF of the negative class (non-responders). It shows the percentage of non-responders at each decile.

2. KS Statistic:

- The KS statistic is the maximum vertical distance between the two CDFs. In this plot, the maximum distance occurs at decile 5, and the KS statistic is 60.723.

3. Interpretation:

- A higher KS statistic indicates better model performance. A higher KS value means the model can more effectively distinguish between positive and negative classes.
- In this case, a KS statistic of 60.723 at decile 5 suggests that the model can differentiate between responders and non-responders quite well.

Additional Notes:

- The decile axis divides the data into ten equal parts.
- The KS statistic is often used in credit risk modeling and other financial applications.

In Summary:

The KS statistic plot is a useful tool for assessing the discriminatory power of a classification model. A higher KS statistic indicates a better model, as it can more accurately identify positive and negative cases.

Choosing the Right Metric

The choice of metric depends on the specific data, use case, and the importance of different error types.

Accuracy

- Suitable for balanced datasets where false positives and false negatives are equally important.
- Requires a threshold to classify instances.

Precision

- Useful when false positives are more costly or critical (e.g., spam detection).
- Requires a threshold to classify instances.

Recall

- Important when false negatives are more costly or critical (e.g., medical diagnosis).
- Requires a threshold to classify instances.

F1-score

- Useful for imbalanced datasets where both precision and recall are important.
- Balances precision and recall, giving more weight to the lower-scoring metric.
- Requires a threshold to classify instances.

AUC

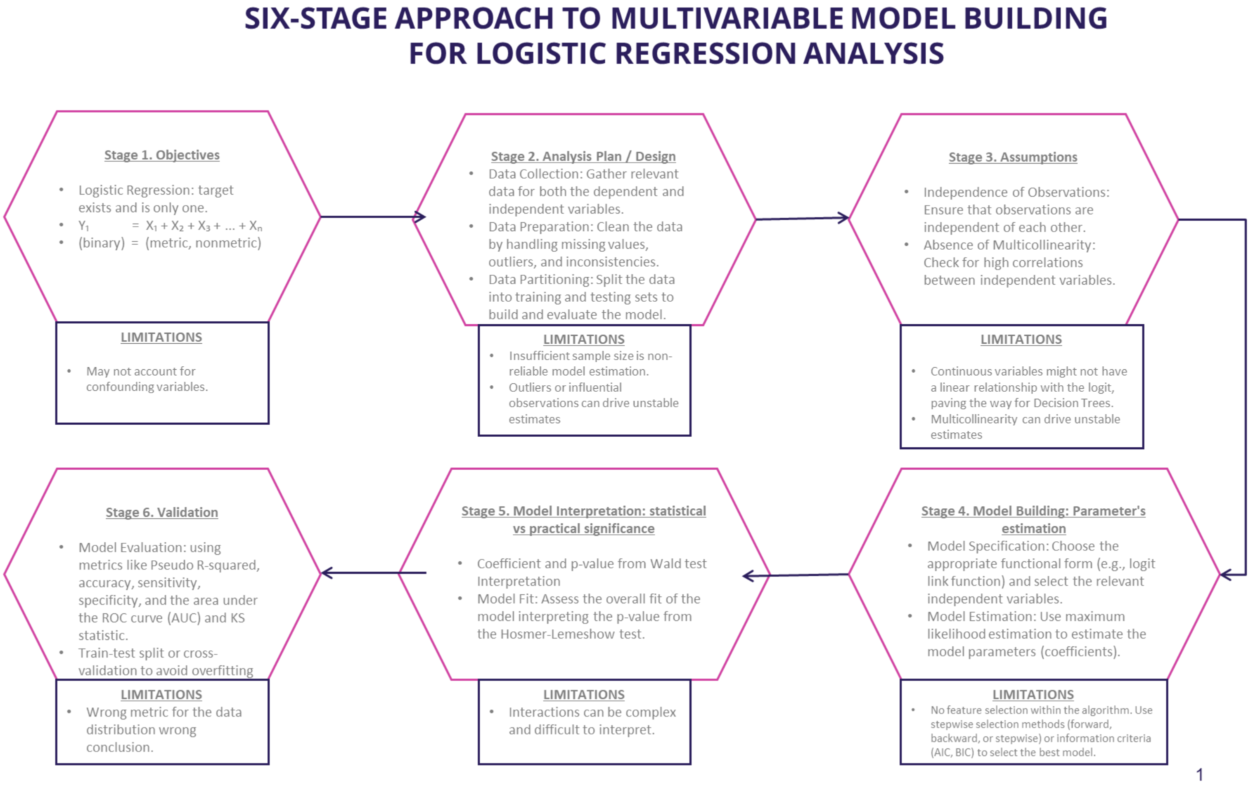
- Evaluates a model's overall performance, especially in imbalanced datasets.
- Considers the model's ability to rank instances correctly, regardless of the specific classification threshold.
- Does not require a fixed threshold, making it suitable when the optimal threshold is unknown or may vary later, example: when the decision of the threshold is taking after the model is deployed according to budget (e.g. Fraud Detection or Marketing campaigns).
- As well as F1-score, it is useful when both precision and recall are important but in a scenario where the threshold is unknown or may vary later.

KS

- Evaluates a model's overall performance, especially in imbalanced datasets.
- Considers the model's ability to rank instances at the optimal cut-off, the one with maximum difference between the two curves.
- Does not require a fixed threshold, making it suitable when the optimal threshold is unknown but stable, example: when the decision of the threshold is taking after the model is deployed around a small or sable range (e.g. Credit Risk or Profit Models).
- As well as F1-score, it is useful when both precision and recall are important but in a scenario where the threshold is unknown or may vary later.

It's often helpful to consider multiple metrics to get a comprehensive understanding of a model's performance.

Six-Stage Approach to Multivariable Model Building for Logistic Regression



SIX-STAGE APPROACH TO MULTIVARIABLE MODEL BUILDING FOR LOGISTIC REGRESSION ANALYSIS

Stage 1. Objectives

- Logistic Regression: target exists and is only one.
- $Y_1 = X_1 + X_2 + X_3 + ... + X_n$
- (binary) = (metric, nonmetric)

LIMITATIONS

- May not account for confounding variables.

Stage 2. Analysis Plan / Design

- Data Collection: Gather relevant data for both the dependent and independent variables.
- Data Preparation: Clean the data by handling missing values, outliers, and inconsistencies.
- Data Partitioning: Split the data into training and testing sets to build and evaluate the model.

LIMITATIONS

- Insufficient sample size is non-reliable model estimation.
- Outliers or influential observations can drive unstable estimates

Stage 3. Assumptions

- Independence of Observations: Ensure that observations are independent of each other.
- Absence of Multicollinearity: Check for high correlations between independent variables.

LIMITATIONS

- Continuous variables might not have a linear relationship with the logit, paving the way for Decision Trees.
- Multicollinearity can drive unstable estimates

Stage 4. Model Building: Parameter's estimation

- Model Specification: Choose the appropriate functional form (e.g., logit link function) and select the relevant independent variables.
- Model Estimation: Use maximum likelihood estimation to estimate the model parameters (coefficients).

LIMITATIONS

- No feature selection within the algorithm. Use stepwise selection methods (forward, backward, or stepwise) or information criteria (AIC, BIC) to select the best model.

Stage 5. Model Interpretation: statistical vs practical significance

- Coefficient and p-value from Wald test Interpretation
- Model Fit: Assess the overall fit of the model interpreting the p-value from the Hosmer-Lemeshow test.

LIMITATIONS

- Interactions can be complex and difficult to interpret.

Stage 6. Validation

- Model Evaluation: using metrics like Pseudo R-squared, accuracy, sensitivity, specificity, and the area under the ROC curve (AUC) and KS statistic.
- Train-test split or cross-validation to avoid overfitting

LIMITATIONS

- Wrong metric for the data distribution wrong conclusion.

1

The image above summarizes the Six-Stage Approach to Multivariable Model Building for Logistic Regression.

let's break down the Six-Stage Approach to Multivariate Model Building for Logistic Regression based on the image above.

Stage 1: Objectives

- **Define the target variable:** Ensure it's binary (e.g., 0/1, yes/no).
- **Identify relevant predictor variables:** These should be both categorical and continuous.

Stage 2: Analysis Plan/Design

- **Data Collection:** Gather relevant data for both the dependent and independent variables.
- **Data Preparation:** Clean the data by handling missing values, outliers, and inconsistencies.
- **Data Partitioning:** Split the data into training and testing sets to build and evaluate the model.

Stage 3: Assumptions

- **Independence of Observations:** Ensure that observations are independent of each other.
- **Absence of Multicollinearity:** Check for high correlations between independent variables.
- **Linearity of the Logit:** Continuous variables should have a linear relationship with the log odds of the outcome.

Stage 4: Model Building (Parameter Estimation)

- **Model Specification:** Choose the appropriate functional form (e.g., logit link function) and select the relevant independent variables.
- **Model Estimation:** Use maximum likelihood estimation to estimate the model parameters (coefficients).

Stage 5: Model Interpretation (Statistical vs. Practical Significance)

- **Coefficient Interpretation:** Interpret the coefficients and their p-values from the Wald test.
- **Model Fit:** Assess the overall fit of the model using the Hosmer-Lemeshow test.

Stage 6: Validation

- **Model Evaluation:** Use metrics like pseudo R-squared, accuracy, sensitivity, specificity, and the area under the ROC curve (AUC) to assess the model's performance.
- **Train-Test Split or Cross-Validation:** Avoid overfitting by using these techniques to assess the model's performance on unseen data.

Limitations and Considerations:

- **Univariate Analysis:** May not account for confounding variables.
- **Multivariable Model Comparisons:** Risk of overfitting.

- **Linearity Assumption:** Non-linear relationships might not be captured adequately.
- **Interactions Among Covariates:** Can lead to complex models and interpretation challenges.
- **Model Fit:** Model fit can be influenced by factors like sample size and data quality.
- **Model Validation:** Validation results might be sensitive to the specific validation method used.

Conclusion

This document has comprehensively covered Logistic Regression, a powerful statistical method for classifying data points into two categories based on one or more predictor variables. We explored its key concepts, including the sigmoid function, log-odds, and maximum likelihood estimation.

We compared Logistic Regression to Discriminant Analysis, highlighting the advantages of Logistic Regression in terms of flexibility and robustness to assumption violations.

We then delved into the specific applications of Logistic Regression, showcasing its versatility in predicting probabilities, identifying key predictors, and understanding relationships. We provided a real-world Python example demonstrating how to implement Logistic Regression with popular libraries like pandas and statsmodels.

Following that, we offered a detailed explanation of how to interpret the logistic regression output, including the optimization termination, model summary, coefficients, significance, and metrics for model fit and validation. This explanation covered key metrics like pseudo R-squared, accuracy, precision, recall, F1-score, AUC (Area Under the ROC Curve), and the KS statistic.

In conclusion, this document provides a thorough understanding of Logistic Regression, making it a valuable resource for anyone seeking to leverage this powerful technique for binary classification tasks. By understanding its concepts, applications, and interpretation, you can effectively utilize Logistic Regression to gain valuable insights from your data and make informed decisions.

Quiz Question

Which of the following is NOT a common metric used to evaluate the performance of a logistic regression?

- A. Accuracy
- B. Precision
- C. Recall
- D. Mean Squared Error
- E. F1-Score

The correct answer is D. Mean Absolute Error (MAE).

Explanation:

MAE is a metric primarily used for regression problems, not classification. Regression trees aim to predict a continuous numerical value, while classification trees predict a categorical label.

Therefore, metrics like Accuracy, Precision, Recall, and F1-Score, which are specifically designed for classification tasks, are more appropriate for evaluating the performance of classification trees.

Further reading / watching material:

Videos:

- StatQuest: Logistic Regression: https://www.youtube.com/watch?v=yIYKR4sgzI8
- Logistic Regression Details Pt1: Coefficients: https://www.youtube.com/watch?v=vN5cNN2-HWE
- Logistic Regression Details Pt 2: Maximum Likelihood: https://www.youtube.com/watch?v=BfKanl1aSG0
- Logistic Regression Details Pt 3: R-squared and p-value: https://www.youtube.com/watch?v=xxFYro8QuXA
- KS and GINI video: https://www.youtube.com/watch?v=MiBUBVUC8kE
- Odds and Log(Odds), Clearly Explained!!: https://www.youtube.com/watch?v=ARfXDSkQf1Y