STATISTICAL FOUNDATIONS FOR MACHINE LEARNING

Hypothesis Testing and Regression Analysis

Victor Lumumba

Mediacrest Training College
Bridge Module to Week 4 Machine Learnin

9 February 2026

Dataset: Hypertension Clinical Cohort (`htn_dat.csv`)

Why Statistics Before Machine Learning?

Machine Learning is applied statistics at scale.

Key statistical foundations:

- Hypothesis testing for decision-making
- Regression for prediction and inference
- p-values and confidence intervals for uncertainty

Healthcare relevance in Kenya:

- Logistic regression for hypertension risk prediction
- Survival analysis for HIV treatment outcomes
- Statistics ensures transparent and responsible AI

Machine learning without statistics becomes a black box.

Learning Objectives

By the end of this session, you will be able to:

1. Formulate and test statistical hypotheses
2. Interpret regression coefficients correctly
3. Apply OLS regression for continuous outcomes
4. Apply logistic regression for binary outcomes
5. Evaluate models using R², accuracy, and AUC
6. Connect regression concepts to ML algorithms

Dataset Overview

Dataset: `htn_dat.csv`
Records: 4,900 patients from Kenyan health facilities

Outcome variables:

- SBP (continuous systolic blood pressure)
- SBP_ge120 (binary hypertension indicator)

Predictor variables:

- Age, BMI, DBP
- Gender, marital status
- Urban clinic indicator
- HIV and ART status

Research question: Which factors significantly predict hypertension risk?

What is Hypothesis Testing?

Hypothesis testing evaluates claims about population parameters.

Null hypothesis (H): No effect or no relationship
Alternative hypothesis (H): An effect or relationship exists

Example:

- H: BMI is not associated with hypertension
- H: BMI is associated with hypertension

Decision Rule in Hypothesis Testing

Decision based on p-value:

- p ¡ 0.05 → Reject H
- p 0.05 → Fail to reject H

Important principle: We do not ”accept” the null hypothesis. We only reject or fail to reject it.

Understanding p-values

A p-value measures evidence against the null hypothesis.

Definition: Probability of observing the data (or more extreme) assuming the null hypothesis is true.

Interpretation:

- Small p-value → Strong evidence against H
- Large p-value → Weak evidence against H

Medical Example of a p-value

Hypothesis:

- H: Mean SBP is equal in urban and rural clinics
- H: Mean SBP differs between clinics

Result: $p = 0.003$

Interpretation: There is strong evidence that clinic location is associated with systolic blood pressure.

Confidence Intervals

A confidence interval provides a range of plausible values for a population parameter.

95% confidence interval: If the study were repeated many times, 95% of such intervals would contain the true value.

Confidence intervals show:

- Effect size
- Precision of estimates

Confidence Interval Example

Mean SBP difference (urban – rural) $= 4.2$ mmHg 95% CI: $[2.1, 6.3]$ mmHg
Interpretation:

- The true SBP difference lies within this range
- CI does not include zero → statistically significant
- Provides more information than p-value alone

Choosing the Right Statistical Test

Test selection depends on variable types:

- Two means → Independent t-test
- More than two means → ANOVA
- Two categorical variables → Chi-square test
- Two continuous variables → Correlation
- Continuous outcome → OLS regression
- Binary outcome → Logistic regression

Introduction to Regression Analysis

Regression models the relationship between predictors (X) and an outcome variable (Y).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Regression is the foundation of many ML algorithms.

Ordinary Least Squares (OLS) Regression

OLS regression predicts continuous outcomes.
Medical application: Predict systolic blood pressure using:

- BMI
- Age
- Gender

OLS minimizes the sum of squared prediction errors.

Interpreting OLS Coefficients

Coefficient interpretation:

- Sign (+/–): Direction of relationship
- Magnitude: Size of effect
- p-value: Statistical significance

Example: A BMI coefficient of 0.87 means SBP increases by 0.87 mmHg per unit BMI increase, holding other variables constant.

Model Fit in OLS Regression

Key evaluation metrics:

- R^2: Proportion of variance explained
- RMSE: Prediction error magnitude
- MAE: Average absolute error

R^2 does not imply causation, only explanatory power.

OLS Assumptions

OLS regression assumptions:

1. Linearity
2. Independence of observations
3. Homoscedasticity
4. Normality of residuals
5. No multicollinearity

Violations require transformations or alternative modeling approaches.

Logistic Regression

Logistic regression is used when the outcome is binary.
Medical application: Predict hypertension (yes/no)
Logistic regression models log-odds, not probabilities directly.

Logistic Regression Equation

Logit model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

Where: p = probability of the event occurring
This ensures predictions lie between 0 and 1.

Interpreting Logistic Regression Results

Two interpretations:

1. Coefficients → Change in log-odds
2. Odds ratios (e^{β}) → Multiplicative change in odds

- Odds ratio ¿ 1 → Increased odds
- Odds ratio ¡ 1 → Decreased odds

Logistic Regression Example

Advanced HIV status: Odds ratio $= 2.51$
Interpretation: Patients with advanced HIV have 2.5 times higher odds of hypertension compared to others.

Model Evaluation for Classification

Key metrics for binary outcomes:

- Accuracy
- Sensitivity (Recall)
- Specificity
- AUC–ROC

Medical screening prioritizes high sensitivity to reduce missed cases.

Confounding and Adjustment

Confounding occurs when a third variable distorts the relationship between X and Y.
Example: Urban clinics appear associated with higher SBP, but age explains the difference.
Solution: Multivariable regression adjustment.

Regression as Machine Learning

Regression models are machine learning algorithms.
Connections:

- Linear regression → Linear ML model
- Logistic regression → Classification ML model
- Neural networks → Stacked regression layers

Understanding regression enables ML mastery.

Common Statistical Pitfalls

Avoid these errors:

- Data leakage
- P-hacking
- Ignoring confounding
- Overfitting models

Statistical rigor protects patient safety and model credibility.

Lab Assignment Overview

Tasks:

- Hypothesis testing
- OLS regression for SBP
- Logistic regression for hypertension
- Model evaluation and interpretation

Deliverable: Python notebook (.ipynb)

Python Analysis Workflow

Steps:

1. Load and clean data
2. Conduct hypothesis tests
3. Fit OLS regression
4. Fit logistic regression
5. Evaluate model performance

Preparing for Week 4 Machine Learning

Before next session:

- Complete lab assignment
- Review regression concepts
- Understand train–test split
- Install required Python libraries

Week 3 focuses on scalable ML pipelines.

Conclusion

Key takeaways:

- Hypothesis testing guides evidence-based decisions
- Regression underpins machine learning algorithms
- Statistical literacy enables responsible AI
- Healthcare ML requires precision and ethics

Statistics is the foundation of trustworthy machine learning.