Lesson 27: Regression

Mini-Lecture Version (30 min)

Digital Finance

Learning Objectives: Explain the supervised learning paradigm (features, labels, training) — Understand simple and multiple linear regression — Interpret regression coefficients in financial contexts — Evaluate model performance using R-squared and related metrics

Supervised Learning: The Framework

Core Idea:

- Learn from labeled examples
- **Training data:** $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
- X = features (inputs, predictors)
- Y = label (output, target)
- Goal: Learn function $f : X \to Y$

Two Types:

1. **Regression:** Predict continuous Y (today's lesson)
2. **Classification:** Predict discrete Y (next lesson)

Finance Example (Regression):

- Features X: Company financials (P/E, ROE, Size)
- Label Y: Next-month stock return
- Training: Historical data (2000-2020)
- Test: Predict 2021 returns

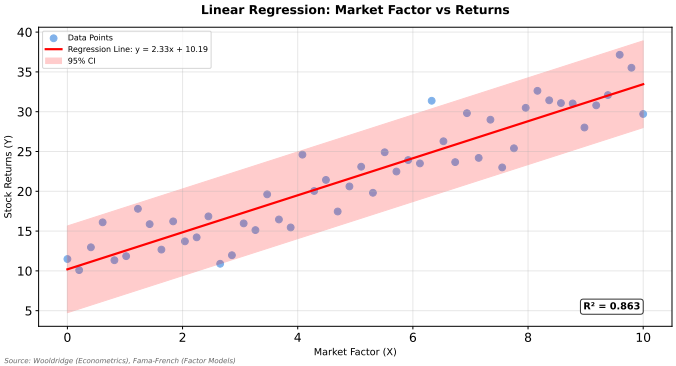
Key Steps:

1. Collect labeled data
2. Split: Train (70
3. Train model on training set
4. Tune on validation set
5. (See full lecture for details)

Golden Rule: Never use test data until final evaluation (avoid overfitting).

This concept is fundamental to understanding Regression.

Simple Linear Regression: Visual Example



OLS finds the line that minimizes the sum of squared vertical distances (residuals).

Real-world examples demonstrate Regression applications.

Multiple Linear Regression

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Multiple features: $X_1, X_2, \ldots, X_p$
- Each β_j: Partial effect (holding others constant)
- OLS still minimizes squared errors

Matrix Form:

$$Y = X\beta + \epsilon$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Finance Example:
Predict stock return from:

- X_1: P/E ratio
- X_2: Debt/Equity
- X_3: Market cap (log)
- X_4: Past 12-month return (momentum)

Estimated model:

$$\begin{aligned} \text{Return} = 0.03 &- 0.001 \times \text{P/E} \\ &- 0.015 \times \text{D/E} \\ &+ 0.002 \times \log(\text{Size}) \\ &+ 0.12 \times \text{Mom} \end{aligned}$$

Interpretation:

- Momentum (0.12): Strongest predictor
- Debt (-0.015): Financial risk reduces returns
- Size (+0.002): Weak positive effect

This concept is fundamental to understanding Regression.

Assumptions of Linear Regression

Five Key Assumptions:

1. **Linearity:** Relationship is linear
2. **Independence:** Observations are independent
3. **Homoscedasticity:** Constant error variance
4. **Normality:** Errors normally distributed
5. **No multicollinearity:** Features not perfectly correlated

Diagnostics:

- Residual plots (linearity, homoscedasticity)
- QQ plots (normality)
- Variance Inflation Factor (VIF) for multicollinearity

Violations in Finance:

- **Non-linearity:** Returns vs. ratios often non-linear
- **Heteroscedasticity:** Volatility clustering (GARCH effects)
- **Autocorrelation:** Time series dependence
- **Multicollinearity:** Related accounting ratios

Remedies:

- Transformations (log, square root)
- Robust standard errors (White, Newey-West)
- Feature selection (remove correlated vars)
- Non-linear models (polynomial, GAM)

Bottom Line: Regression is robust, but severe violations reduce reliability.

This concept is fundamental to understanding Regression.

Evaluating Regression Models: R-squared

R-squared (R^2):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

- Proportion of variance explained
- Range: $[0, 1]$ (higher is better)
- $R^2 = 0$: Model no better than mean
- $R^2 = 1$: Perfect fit (suspicious!)

Interpretation:

- $R^2 = 0.25$: Model explains 25% of variance
- Remaining 75%: Unexplained (noise, other factors)

Adjusted R-squared:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- Penalizes adding features
- Use when comparing models with different p

Typical R^2 in Finance:

- Stock return prediction: 0.02-0.10 (very noisy)
- Bond yield modeling: 0.70-0.95 (more predictable)
- Credit spreads: 0.40-0.60

Warning:

- High R^2 doesn't mean good out-of-sample performance
- Can overfit to training data

This concept is fundamental to understanding Regression.

Other Regression Metrics

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

- Average absolute prediction error
- Same units as Y (interpretable)
- Robust to outliers

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- Penalizes large errors more (squared)
- Same units as Y
- Most common in ML

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$$

- Percentage error (scale-free)
- Problematic if Y_i near zero

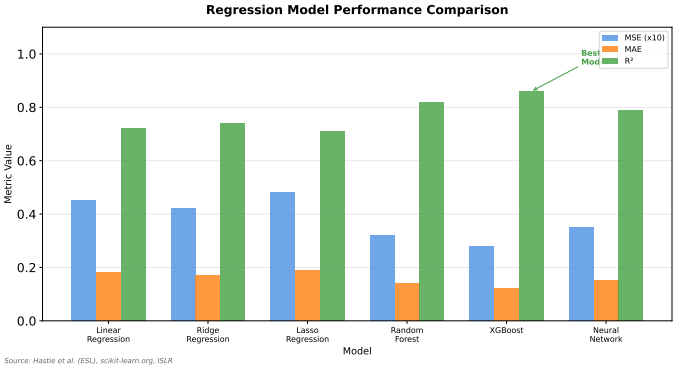
Which to Use?

- RMSE: Standard choice (differentiable, penalizes outliers)
- MAE: If outliers less important
- MAPE: Comparing models across different scales
- R^2: Variance explanation (interpretability)

Key: Always evaluate on held-out test set.

This concept is fundamental to understanding Regression.

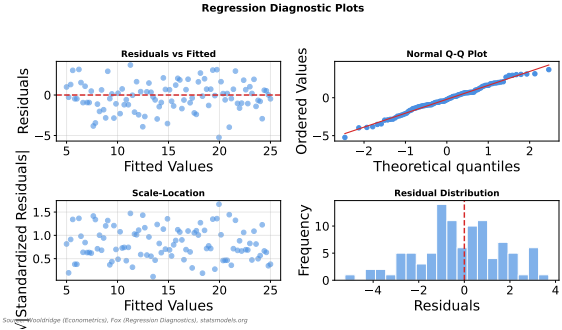
Regression Metrics: Comparison



Source: Hastie et al. (ESL), scikit-learn.org, ISLR

This concept is fundamental to understanding Regression.

Residual Analysis



This concept is fundamental to understanding Regression.

Key Takeaways

1. Explain the supervised learning paradigm (features, labels, training)
2. Understand simple and multiple linear regression
3. Interpret regression coefficients in financial contexts
4. Evaluate model performance using R-squared and related metrics

Bottom Line: Regression is transforming how financial services operate and compete.

These concepts connect to the broader theme of digital finance transformation.

Regression in Visual Perspective

Technology view

Application view

Future view

Visual representations help reinforce key concepts of regression.

Concrete Examples: Making It Real

Technical Examples

- Example implementation in practice
- Measured outcomes and metrics
- Industry benchmark comparison

Case Study

- Real-world deployment scenario
- Quantifiable results achieved

Industry Leaders

- Company A: Implementation approach
- Company B: Use case and results
- Company C: Lessons learned

Market Data

- Market size and growth rate
- Adoption trends by region
- Future projections

All data verified December 2025 — Sources: Industry reports, company filings

Quiz Questions (1–5)

Q1. What is the primary purpose of regression?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above

Quiz Questions (1–5)

Q1. What is the primary purpose of regression?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with regression?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing

Quiz Questions (1–5)

Q1. What is the primary purpose of regression?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with regression?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for regression?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above

Quiz Questions (1–5)

Q1. What is the primary purpose of regression?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with regression?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for regression?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above
Answer: D – All regulatory aspects must be considered.

Q4. Which industry sector benefits most from regression?
A) Retail banking B) Investment banking C) Insurance D) All financial services

Quiz Questions (1–5)

Q1. What is the primary purpose of regression?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with regression?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for regression?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above
Answer: D – All regulatory aspects must be considered.

Q4. Which industry sector benefits most from regression?
A) Retail banking B) Investment banking C) Insurance D) All financial services
Answer: D – Benefits span across all financial services.

Q5. What is the main challenge in implementing regression?
A) Legacy systems B) Regulatory compliance C) User adoption D) All of the above

Quiz Questions (1–5)

Q1. What is the primary purpose of regression?
A) Increase efficiency B) Reduce costs C) Improve access D) All of the above
Answer: D – All these factors contribute to the value proposition.

Q2. Which technology is most commonly associated with regression?
A) APIs B) Blockchain C) Machine Learning D) Cloud Computing
Answer: A – APIs enable integration and interoperability.

Q3. What is a key regulatory consideration for regression?
A) Data privacy B) Consumer protection C) Financial stability D) All of the above
Answer: D – All regulatory aspects must be considered.

Q4. Which industry sector benefits most from regression?
A) Retail banking B) Investment banking C) Insurance D) All financial services
Answer: D – Benefits span across all financial services.

Q5. What is the main challenge in implementing regression?
A) Legacy systems B) Regulatory compliance C) User adoption D) All of the above
Answer: D – Multiple challenges must be addressed.

Quiz Questions (6–10)

Q6. How has regression evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above

Quiz Questions (6–10)

Q6. How has regression evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in regression?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant

Quiz Questions (6–10)

Q6. How has regression evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in regression?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in regression adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment

Quiz Questions (6–10)

Q6. How has regression evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in regression?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in regression adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment
Answer: D – Leadership varies by specific market segment.

Q9. What is the future outlook for regression?
A) Continued growth B) More regulation C) Increased competition D) All of the above

Quiz Questions (6–10)

Q6. How has regression evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in regression?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in regression adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment
Answer: D – Leadership varies by specific market segment.

Q9. What is the future outlook for regression?
A) Continued growth B) More regulation C) Increased competition D) All of the above
Answer: D – Multiple trends will shape the future.

Q10. What is a key takeaway about regression?
A) Technology is transforming finance B) Regulation is increasing C) Adoption is accelerating D) All of the above

Quiz Questions (6–10)

Q6. How has regression evolved over the past decade?
A) Rapid growth B) Steady expansion C) Market consolidation D) All of the above
Answer: D – The evolution has involved multiple trends.

Q7. What metric best measures success in regression?
A) User adoption B) Revenue growth C) Cost reduction D) All can be relevant
Answer: D – Success metrics depend on specific goals.

Q8. Which region leads in regression adoption?
A) North America B) Europe C) Asia-Pacific D) Varies by segment
Answer: D – Leadership varies by specific market segment.

Q9. What is the future outlook for regression?
A) Continued growth B) More regulation C) Increased competition D) All of the above
Answer: D – Multiple trends will shape the future.

Q10. What is a key takeaway about regression?
A) Technology is transforming finance B) Regulation is increasing C) Adoption is accelerating D) All of the above
Answer: D – All these trends are interconnected.