Linear Regression - Prediction and Analysis

Detailed Study Notes

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It allows us to predict values and understand relationships.

Definition:
Linear regression finds the best-fit straight line through data points that minimizes the sum of squared distances from points to the line.

Simple Linear Regression Formula:
$y = mx + b$

where:
- y = predicted value (dependent variable)
- x = independent variable (predictor)
- m = slope (rate of change)
- b = y-intercept (value when x = 0)

Calculating Slope and Intercept:
$m = \Sigma[(x - x\blacksquare)(y - \blacksquare)] / \Sigma(x - x\blacksquare)^2$
$b = \blacksquare - m \cdot x\blacksquare$

Detailed Example:
Study hours (x): 2, 4, 6, 8, 10
Test scores (y): 60, 70, 80, 85, 95

Step 1: Calculate means
$x\blacksquare = 6$, $\blacksquare = 78$

Step 2: Calculate slope (m)
Numerator: $\Sigma[(x - x\blacksquare)(y - \blacksquare)]$
$(2-6)(60-78) + (4-6)(70-78) + (6-6)(80-78) + (8-6)(85-78) + (10-6)(95-78)$
$= 72 + 16 + 0 + 14 + 68 = 170$

Denominator: $\Sigma(x - x\blacksquare)^2$
$(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2$
$= 16 + 4 + 0 + 4 + 16 = 40$

m = 170 / 40 = 4.25

Step 3: Calculate y-intercept (b)
b = 78 - 4.25(6) = 78 - 25.5 = 52.5

Step 4: Regression equation
y = 4.25x + 52.5

Interpretation:
- For each additional study hour, test score increases by 4.25 points
- A student with 0 study hours would score 52.5 (theoretical baseline)

Making Predictions:
If a student studies 7 hours:
y = 4.25(7) + 52.5 = 29.75 + 52.5 = 82.25 points

Coefficient of Determination (R^2):
R^2 measures how well the regression line fits the data
Range: $0 \leq R^2 \leq 1$

$R^2 = 1 - (SS_{residual} / SS_{total})$

Interpretation:
- $R^2 = 0.80$ means 80% of variance in Y is explained by X
- $R^2 = 1$ means perfect fit
- $R^2 = 0$ means X doesn't explain Y

Assumptions of Linear Regression:

1. Linearity: Relationship is linear
2. Independence: Observations are independent
3. Homoscedasticity: Constant variance of residuals
4. Normality: Residuals are normally distributed
5. No multicollinearity (for multiple regression)

Residuals:
Residual = Actual Y - Predicted Y
- Should be randomly distributed around zero
- Patterns indicate model problems

Real-World Applications:

Business:
- Sales forecasting based on advertising spend
- Price optimization
- Demand prediction

Healthcare:
- Predicting patient outcomes
- Dosage recommendations
- Disease progression modeling

Education:
- Predicting student performance
- Identifying at-risk students
- Resource allocation

Finance:
- Stock price prediction
- Risk assessment
- Credit scoring

Science:
- Calibration curves in chemistry
- Growth models in biology
- Climate modeling

Multiple Linear Regression:
y = b■ + b■x■ + b■x■ + ... + b■x■

Example: Predicting house prices
Price = b■ + b■(Size) + b■(Bedrooms) + b■(Age)

Limitations:
1. Only models linear relationships
2. Sensitive to outliers
3. Extrapolation beyond data range is risky
4. Assumes causation is directional (X $\rightarrow$ Y)

Common Mistakes to Avoid:
- Extrapolating far beyond observed data
- Assuming correlation implies causation
- Ignoring assumption violations
- Over-interpreting R^2
- Not checking residual plots

Best Practices:
- Always visualize data with scatter plots

- Check residual plots for patterns
- Validate assumptions
- Report both slope and R^2
- Consider confidence intervals
- Test on new data when possible