# Advanced Linear Regression: Comprehensive Notes

## 1. Introduction to Linear Regression

Linear regression is a foundational statistical learning method used to model the relationship between a quantitative response variable and one or more predictor variables. It forms the basis for many sophisticated statistical learning techniques and provides an interpretable framework for understanding relationships in data.

## 2. Simple Linear Regression

### 2.1 Model Formulation

Simple linear regression models the relationship between a response variable and a single predictor variable using a linear function:

Where: - is the **intercept** (the expected value of when ) - is the **slope** (the expected change in for a one-unit increase in )

Together, these parameters define a straight line that best approximates the relationship between and .

### 2.2 Parameter Estimation

In practice, the true parameters β₀ and β₁ are unknown and must be estimated from data. The most common approach is the **method of least squares**, which finds estimates β̂₀ and β̂₁ that minimize the sum of squared differences between observed and predicted values.

The **residual** for the ith observation is:

eᵢ = yᵢ - ŷᵢ = yᵢ - (β̂₀ + β̂₁xᵢ)

The goal is to minimize the **Residual Sum of Squares (RSS)**:

RSS = Σ eᵢ² = Σ (yᵢ - β̂₀ - β̂₁xᵢ)²

Using calculus, the least squares estimates can be derived as:

β̂₁ = Σ(xᵢ - x̄)(yᵢ - ȳ) / Σ(xᵢ - x̄)²

β̂₀ = ȳ - β̂₁x̄

Where x̄ and ȳ are the sample means of X and Y, respectively.

### 2.3 Statistical Properties of Linear Regression

The complete statistical model for simple linear regression is:

Where represents the error term, which captures: - Non-linearity in the true relationship - Effects of unmeasured variables - Measurement error

**Key assumptions about the error term include:** - Independence from - Zero mean: - Constant variance: (homoscedasticity) - Normality: (for inference purposes)

### 2.4 Assessing Parameter Estimate Accuracy

#### 2.4.1 Standard Errors

The standard errors of the coefficient estimates quantify their uncertainty:

SE(β̂₁)² = σ²/Σ(xᵢ - x̄)²

SE(β̂₀)² = σ²[1/n + x̄²/Σ(xᵢ - x̄)²]

Where σ² = Var(ε) is estimated by the **Residual Standard Error (RSE)**:

RSE = √(RSS/(n-2))

The degrees of freedom adjustment (n-2) accounts for estimating two parameters.

#### 2.4.2 Confidence Intervals

95% confidence intervals for the parameters can be constructed as:

β̂ⱼ ± t₀.₉₇₅,ₙ₋₂ × SE(β̂ⱼ)

Where t₀.₉₇₅,ₙ₋₂ is the 97.5th percentile of the t-distribution with n-2 degrees of freedom (approximately 2 for large sample sizes).

#### 2.4.3 Hypothesis Testing

Hypothesis tests can be conducted to determine if there is a relationship between X and Y:

* Null hypothesis: H₀: β₁ = 0 (no relationship)
* Alternative hypothesis: H₁: β₁ ≠ 0 (relationship exists)

The test statistic is:

t = β̂₁/SE(β̂₁)

This follows a t-distribution with n-2 degrees of freedom under the null hypothesis. The p-value is P(|T| > |t|), where T ~ t\_{n-2}.

### 2.5 Prediction

Once parameters are estimated, predictions for new observations can be made:

There are two types of intervals for predictions: - **Confidence intervals**: Quantify uncertainty in estimating the mean response - **Prediction intervals**: Quantify uncertainty in predicting an individual response at

Prediction intervals are always wider than confidence intervals because they include both the uncertainty in estimating the parameters and the inherent variability in the response.

### 2.6 Assessing Model Fit

Two key metrics for evaluating the fit of a linear regression model are:

#### 2.6.1 Residual Standard Error (RSE)

RSE estimates the standard deviation of the error term and represents the average amount by which the response deviates from the true regression line.

#### 2.6.2 R-squared Statistic

Where: - TSS (Total Sum of Squares) = (total variance in the response) - RSS (Residual Sum of Squares) = (unexplained variance)

represents the proportion of variance in explained by and ranges from 0 to 1, with higher values indicating better fit.

For simple linear regression, equals the square of the correlation between and :

## 3. Multiple Linear Regression

### 3.1 Model Formulation

Multiple linear regression extends simple linear regression to incorporate multiple predictors:

Where: - is the intercept - is the coefficient for predictor , representing the average effect of a one-unit increase in on when all other predictors are held constant

### 3.2 Parameter Estimation

The least squares principle extends to multiple regression by minimizing:

The solution is best expressed in matrix form:

Where: - is the vector of coefficient estimates - is the design matrix with a column of 1s and the predictor variables - is the vector of response values

### 3.3 Inference in Multiple Regression

#### 3.3.1 Standard Errors

The variance-covariance matrix of the coefficient estimates is:

The standard errors are the square roots of the diagonal elements of this matrix.

#### 3.3.2 F-Statistic for Overall Significance

To test the overall significance of the model:

* Null hypothesis: (none of the predictors are related to the response)
* Alternative hypothesis: at least one

The F-statistic is:

Under the null hypothesis, follows an F-distribution with and degrees of freedom.

#### 3.3.3 Partial F-Tests

To test the significance of a subset of predictors:

Where is the residual sum of squares for the model without the predictors of interest.

### 3.4 Variable Selection

In practice, including too many predictors can lead to overfitting, while including too few can result in underfitting. Variable selection methods aim to identify the best subset of predictors.

#### 3.4.1 Forward Selection

1. Begin with the null model (intercept only)
2. Add the predictor that most improves the model fit
3. Continue adding predictors one by one until some stopping criterion is met

#### 3.4.2 Backward Selection

1. Begin with the full model including all predictors
2. Remove the least significant predictor (highest p-value)
3. Continue removing predictors one by one until all remaining predictors are significant

#### 3.4.3 Mixed (Stepwise) Selection

Combines forward and backward approaches: 1. Begin with the null model 2. Add the most significant predictor 3. Add or remove predictors based on significance thresholds 4. Continue until no additions or removals improve the model

### 3.5 Model Fit Assessment in Multiple Regression

The R² and RSE metrics extend to multiple regression:

R² = 1 - RSS/TSS

RSE = √(RSS/(n-p-1))

In multiple regression, R² = Cor(Y, Ŷ)², the squared correlation between the observed and predicted responses.

A limitation of R² is that it always increases when predictors are added, even if they have little relationship with the response. The **adjusted R²** addresses this by penalizing the number of predictors:

R²ₐdⱼ = 1 - (RSS/(n-p-1)) / (TSS/(n-1))

## 4. Qualitative Predictors

### 4.1 Dummy Variables

Categorical predictors are incorporated into regression models using dummy variables.

For a binary predictor with levels A and B, we can create a dummy variable:

The resulting model is:

Where: - is the mean response for category B - is the mean response for category A - represents the difference in mean response between categories A and B

### 4.2 Categorical Predictors with Multiple Levels

For a categorical predictor with k levels, we need k-1 dummy variables. With levels A, B, and C, we might have:

X₁ᵢ = {  
 1 if observation i is in category A  
 0 otherwise  
}

X₂ᵢ = {  
 1 if observation i is in category B  
 0 otherwise  
}

The model becomes:

Y = β₀ + β₁X₁ + β₂X₂ + ε

Where: - β₀ is the mean response for category C (baseline) - β₀ + β₁ is the mean response for category A - β₀ + β₂ is the mean response for category B - β₁ represents the difference between categories A and C - β₂ represents the difference between categories B and C

## 5. Extensions to Linear Models

### 5.1 Interaction Terms

The standard linear model assumes that the effect of each predictor is additive and independent of other predictors. Interaction terms allow for more complex relationships where the effect of one predictor depends on the value of another.

A model with an interaction between X₁ and X₂ is:

Y = β₀ + β₁X₁ + β₂X₂ + β₃X₁X₂ + ε

The interaction coefficient β₃ represents the change in the effect of X₁ on Y for a one-unit increase in X₂ (and vice versa).

The **hierarchical principle** states that when including an interaction term, the main effects should also be included in the model, even if their individual p-values are not significant.

### 5.2 Non-Linear Transformations

Linear regression can accommodate non-linear relationships through transformations of the predictors:

**1. Polynomial regression**: Including polynomial terms of predictors

**2. Log transformations**: Taking the logarithm of predictors or the response

**3. Other transformations**: Square root, reciprocal, etc.

Choosing appropriate transformations often involves exploratory data analysis and residual plots.

## 6. Potential Problems in Linear Regression

### 6.1 Non-linearity of the Response-Predictor Relationship

If the true relationship is non-linear, linear regression may provide poor predictions and misleading inferences.

**Detection**: Residual plots showing systematic patterns (e.g., U-shaped pattern)

**Solution**: Apply non-linear transformations to predictors or use more flexible models

### 6.2 Correlation of Error Terms

Linear regression assumes independent errors. Correlated errors often occur in time series data or when observations are related.

**Detection**: Plot residuals against time or order of collection

**Solution**: Time series methods, mixed-effects models, or generalized estimating equations

### 6.3 Non-constant Variance (Heteroscedasticity)

Unequal error variance across the range of predictors violates the homoscedasticity assumption.

**Detection**: Funnel-shaped residual plots

**Solution**: Variance-stabilizing transformations (e.g., log or square root of response), weighted least squares, or robust regression

### 6.4 Outliers

Observations with unusual response values can disproportionately influence the model.

**Detection**: Studentized residuals (values > |3| are potential outliers)

**Solution**: Robust regression methods, outlier removal (if justified), or transformation

### 6.5 High-Leverage Points

Observations with unusual predictor values can have undue influence on coefficient estimates.

**Detection**: Leverage statistic (high if greatly exceeds )

**Solution**: Robust regression, influence diagnostics, or sensitivity analysis

### 6.6 Collinearity

High correlation among predictors makes it difficult to separate their individual effects.

**Detection**: - Correlation matrix - Variance Inflation Factor (VIF): VIF(β̂ⱼ) = 1/(1-R²\_Xⱼ|X₍₋ⱼ₎) (where R²\_Xⱼ|X₍₋ⱼ₎ is the R² from regressing Xⱼ on all other predictors)

**Solution**: - Drop one of the collinear predictors - Combine collinear predictors (e.g., through principal components) - Ridge regression or other regularization methodsR^2\_{X\_j|X\_{-j}}}$ (where is the from regressing on all other predictors)

**Solution**: - Drop one of the collinear predictors - Combine collinear predictors (e.g., through principal components) - Ridge regression or other regularization methods

## 7. Comparison with Non-Parametric Methods

### 7.1 K-Nearest Neighbors Regression

K-nearest neighbors (KNN) regression estimates by averaging the responses of the training observations closest to :

Where is the set of observations closest to .

### 7.2 Parametric vs. Non-Parametric Trade-offs

* **Bias-Variance Trade-off**: Parametric methods have higher bias but lower variance compared to flexible non-parametric methods
* **Sample Size**: Parametric methods perform better with smaller sample sizes
* **Dimensionality**: Non-parametric methods suffer from the "curse of dimensionality" in high-dimensional spaces
* **Interpretability**: Linear regression provides more interpretable results

## 8. Advanced Topics in Linear Regression

### 8.1 Ridge Regression and Lasso

These regularization methods add penalties to the least squares objective function:

* **Ridge Regression**: Minimize
* **Lasso**: Minimize

These methods help reduce overfitting and handle collinearity.

### 8.2 Cross-Validation for Model Selection

K-fold cross-validation provides a systematic way to evaluate model performance:

1. Divide the data into K equal-sized folds
2. For each fold, train the model on the remaining K-1 folds and test on the held-out fold
3. Average the performance metric (e.g., MSE) across all K iterations

Cross-validation helps in selecting optimal models, tuning parameters, and comparing different modeling approaches.

### 8.3 Bootstrapping for Uncertainty Estimation

Bootstrapping resamples the data with replacement to estimate the sampling distribution of statistics:

1. Generate B bootstrap samples by randomly sampling n observations with replacement
2. Calculate the statistic of interest for each bootstrap sample
3. Use the distribution of bootstrap statistics to estimate confidence intervals or standard errors

This approach provides robust uncertainty estimates without assuming normality.

## 9. Conclusion

Linear regression is a powerful and interpretable framework for modeling relationships between variables. Despite its simplicity, it provides the foundation for many advanced statistical methods and continues to be widely applied in various fields. Understanding its principles, assumptions, and limitations is essential for effective data analysis and interpretation.