Summary of Chapter 3: Linear Regression

Introduction to Linear Regression

Chapter 3 of the textbook introduces linear regression, a foundational statistical learning method used for predicting a quantitative response. Despite being a basic method compared to modern statistical learning techniques, linear regression remains a powerful and widely used tool. The chapter highlights its importance as a stepping stone for understanding more advanced machine learning and statistical methods.

The chapter begins by discussing a practical application: predicting product sales based on advertising budgets for TV, radio, and newspapers. It outlines key questions that linear regression can help answer, such as:

- Whether a relationship exists between advertising and sales.
- The strength of this relationship.
- The contribution of each advertising medium.
- The accuracy of future sales predictions.

Simple Linear Regression

Simple linear regression models a response variable Y as a linear function of a single predictor X:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- β_0 is the intercept.
- β_1 is the slope.
- ϵ is an error term.

Estimating Coefficients

The coefficients β_0 and β_1 are unknown and must be estimated using data. The most common approach is the **least squares method**, which minimizes the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \beta_0 + \beta_1 x_i$ is the predicted value of Y.

Using calculus, the least squares estimates of the coefficients are:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y.

Assessing the Accuracy of the Model

To assess the accuracy of estimated coefficients, standard errors (SE) are computed:

$$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

where σ is the standard deviation of the error term. Confidence intervals for the coefficients can be computed as:

$$\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$$

Hypothesis Testing

A hypothesis test can determine whether a predictor is significantly related to the response. The null hypothesis is:

$$H_0 : \beta_1 = 0$$

The test statistic follows a t-distribution:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

A small **p-value** indicates strong evidence against H_0, meaning X significantly affects Y.

Assessing Model Fit

Two important measures of model fit are:

- **Residual Standard Error (RSE)**: Measures the model's average prediction error.
- R^2 Statistic: Measures the proportion of variance in Y explained by X:

  $$R^2 = 1 - \frac{RSS}{TSS}$$

  where TSS (Total Sum of Squares) represents total variation in Y. Higher R^2 values indicate a better fit.

Multiple Linear Regression

Multiple linear regression extends simple regression to include multiple predictors:

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

where $X_1, X_2, \ldots, X_p$ are multiple predictors.

Estimating Coefficients

Similar to simple regression, the coefficients are estimated using the least squares method, minimizing:

$$RSS = \sum (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \beta_0 + \sum \beta_j X_{ij}$.

Assessing Significance

The **F-test** is used to test whether at least one predictor is significantly related to Y:

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

A high F-statistic with a low p-value suggests at least one predictor significantly contributes to the model.

Variable Selection

To determine which predictors to keep, variable selection methods include:

- **Forward selection**: Start with no predictors, add them one by one based on their significance.
- **Backward selection**: Start with all predictors, remove the least significant one iteratively.
- **Mixed selection**: Combines both forward and backward selection.

Collinearity

Collinearity occurs when predictors are highly correlated, making it difficult to isolate their effects. The **Variance Inflation Factor (VIF)** detects collinearity:

$$VIF(X_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

High VIF values (above 5 or 10) indicate problematic collinearity.

Extensions to the Linear Model

Interaction Effects

Interaction terms capture synergistic effects between predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \epsilon$$

If β_3 is significant, the effect of X_1 on Y depends on X_2.

Non-linearity

The standard model assumes a linear relationship, but polynomial regression can capture non-linearity:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \epsilon$$

which introduces curvature to the model.

Qualitative Predictors

Categorical variables can be included using **dummy variables**. If a predictor has k levels, we create $k-1$ dummy variables.

For example, if "region" has three levels (East, West, South), we create:

$$X_1 = \begin{cases} 1, & \text{if South} \\ 0, & \text{otherwise} \end{cases}, \quad X_2 = \begin{cases} 1, & \text{if West} \\ 0, & \text{otherwise} \end{cases}$$

One category (East) is the **baseline**.

Common Problems in Regression

1. **Non-linearity**: Addressed using polynomial transformations or other modeling techniques.
2. **Correlation of error terms**: Often found in **time series data**.
3. **Non-constant variance (heteroscedasticity)**: Residual plots help detect this; transformations like $\log(Y)$ can help.
4. **Outliers**: Large residuals suggest unusual data points.
5. **High-leverage points**: Have extreme predictor values and can disproportionately affect the model.
6. **Collinearity**: Addressed by removing correlated predictors or using principal component analysis.

Comparison of K-Nearest Neighbors (KNN) with Linear Regression

The chapter provides a detailed comparison between **linear regression** (a parametric method) and **K-Nearest Neighbors (KNN)** (a non-parametric method).

Key Differences

1. **Assumption on Functional Form**:
   - **Linear regression** assumes a fixed functional form $f(X)$, which is beneficial when the true relationship is close to linear.
   - **KNN regression** does not assume any parametric form, making it more flexible in capturing complex relationships.
2. **Interpretability vs. Flexibility**:
   - **Linear regression** is highly interpretable, allowing for hypothesis testing and confidence intervals.
   - **KNN** is more flexible but lacks interpretability; it does not provide explicit coefficient estimates or statistical inference.
3. **Bias-Variance Tradeoff**:

- **Linear regression** has **low variance** but may have **high bias** if the true relationship is non-linear.
- **KNN** can have **low bias** but tends to have **high variance**, especially when K is small.

4. **Performance in Low vs. High Dimensions**:
   - When p (number of predictors) is small, **KNN may outperform linear regression if the true relationship is highly non-linear**.
   - However, as p increases, **KNN suffers from the "curse of dimensionality," leading to poor performance**, while **linear regression remains stable**.

Illustrative Findings from the Chapter

- **When the true relationship is linear**, **linear regression performs better** than KNN, as KNN introduces unnecessary variance.
- **When the true relationship is non-linear**, **KNN can outperform linear regression**, particularly when K is chosen optimally.
- **When there are many irrelevant predictors (high p)**, **KNN struggles** because neighbors are no longer close in high-dimensional space, making linear regression the better choice.

Final Takeaway

- **Linear regression** is the preferred choice when the relationship is approximately linear or when interpretability is important.
- **KNN** is more flexible and useful when the true relationship is complex and non-linear, but it requires careful tuning of K and suffers in high dimensions.

This comparison underscores the importance of understanding the nature of the data before selecting a modeling approach.

Conclusion

Chapter 3 provides a comprehensive guide to linear regression, covering:

- **Model formulation, estimation, and interpretation**.
- **Assessing model fit and hypothesis testing**.
- **Extensions such as interactions and polynomial regression**.
- **Practical issues such as collinearity and outliers**.
- **Comparison with KNN for different data scenarios**.

This foundation is essential for understanding more advanced regression techniques and machine learning models.