Chapter 3: Linear Regression Study Notes (Based on ISLP)

Tanmay Chopra

Linear regression is a foundational tool in **statistical learning**. It is a **supervised learning** approach used for predicting a **quantitative response**. A strong understanding of linear regression is crucial because many advanced statistical learning methods are extensions or generalisations of it.

1 Simple Linear Regression (SLR)

SLR predicts a quantitative response Y based on a single predictor X.

1.1 The Model and Estimation

1. **Model Form:** Assumes an approximately linear relationship:

$$Y \approx \beta_0 + \beta_1 X$$

   Or, including error:

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

   - β_0 is the **intercept**, and β_1 is the **slope**; together they are the **model coefficients** or **parameters**.
   - ϵ (the error term) accounts for the true relationship not being linear, unmeasured variables, and measurement error.

2. **Estimation using Least Squares:** The coefficients $(\hat{\beta}_0, \hat{\beta}_1)$ are estimated by minimizing the **Residual Sum of Squares (RSS)**:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

   - The values that minimize RSS define the **least squares line**.
   - The least squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ are **unbiased**; their average over many datasets would equal the true parameters (β_0, β_1).

1.2 Assessing the Accuracy of Coefficients

1. **Standard Errors (SE):** The SE of a coefficient estimate measures the average amount that the estimate deviates from the actual value.
   - $\text{SE}(\hat{\beta}_1)$ is smaller when the predictor values x_i are more spread out.

2. **Confidence Intervals (CI):** A 95% CI defines a range of values that, with 95% probability, contains the true unknown parameter value.

$$\text{CI for } \beta_1 \approx \hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1)$$

3. **Hypothesis Testing:** Used to test the null hypothesis (H_0) that there is **no relationship** between X and Y ($H_0 : \beta_1 = 0$).
   - The **t-statistic** is computed as:

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

   - The **p-value** is the probability of observing a t-statistic equal to or larger than $|t|$ assuming H_0 is true. A **small p-value** (e.g., $< 5\%$) suggests an association exists, leading to the rejection of H_0.

1.3 Assessing the Accuracy of the Model

1. **Residual Standard Error (RSE):** This is an estimate of the standard deviation (σ) of the error term (ϵ).
   - It is the **average amount** that the response will deviate from the true regression line.
   - $$\text{RSE} = \sqrt{\text{RSS}/(n-2)}$$
     It is a measure of the lack of fit.
2. R^2 **Statistic:** The **proportion of variance explained**.
   - It takes a value between 0 and 1, independent of the scale of Y.
   - $$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$
     where TSS (Total Sum of Squares) measures total variance in Y before regression.
   - A value close to 1 indicates a large proportion of variability is explained. In SLR, R^2 is the square of the **correlation** between X and Y.

2 Multiple Linear Regression (MLR)

MLR extends SLR to accommodate **multiple predictors** (p predictors) simultaneously.

2.1 Model Form and Interpretation

- $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$
- **Crucial Interpretation:** β_j is the average effect on Y of a one-unit increase in X_j, **holding all other predictors fixed**.
- MLR coefficients may differ greatly from SLR coefficients if predictors are **correlated** (collinearity/confounding).

2.2 Important Questions in MLR

1. **Is at least one predictor useful?**
   - Test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.
   - This is done using the **F-statistic**. If H_0 is true, F is expected to be close to 1. If F is significantly greater than 1, we reject H_0.
2. **Which predictors are useful?**
   - Individual **t-statistics** and associated **p-values** are calculated to assess if each predictor is related to Y after adjusting for others.
   - **Variable Selection Methods:** When dealing with many predictors, approaches like **Forward Selection**, **Backward Selection**, or examining statistics like **Mallow's** C_p, **AIC**, **BIC**, and **Adjusted** R^2 are used.
3. **How well does the model fit?**
   - R^2 **always increases** when more variables are added to the model.
   - In MLR, R^2 is the square of the correlation between the response Y and the fitted values $\hat{Y}$: $R^2 = \text{Cor}(Y, \hat{Y})^2$.
4. **Prediction Accuracy:**
   - **Prediction intervals** estimate uncertainty for an **individual response** ($Y = f(X) + \epsilon$) and are always wider than **confidence intervals**, which estimate uncertainty for the **average response** ($f(X)$).

3 Other Considerations in the Regression Model

3.1 Qualitative Predictors

- Qualitative (categorical) variables are incorporated using **dummy variables**.
- For a qualitative variable with K levels, $K-1$ **dummy variables** are created. The excluded level is the **baseline**.

3.2 Extensions of the Linear Model

The standard linear model makes two key assumptions that can be relaxed: **additivity** and **linearity**.

1. **Interaction Terms (Removing Additivity):**
   - An **interaction term** (e.g., X_1X_2) is introduced:
   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$
   - This allows the effect of X_1 on Y to depend on the value of X_2 (i.e., $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$).
   - **Hierarchical Principle:** If an interaction term is included, the **main effects** must also be included.
2. **Polynomial Regression (Removing Linearity):**
   - This is relaxed by including **transformed versions of the predictors** (e.g., X^2, X^3).
   - Example:
   $$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

3.3 Potential Problems

1. **Non-linearity:** Detectable by a **pattern** (e.g., U-shape) in the **residual plot**.
2. **Correlated Errors:** Often found in time series data.
3. **Non-constant Variance (Heteroscedasticity):** Detected by a **funnel shape** in residual plots. Solution: transform the response variable Y (e.g., $\log(Y)$).
4. **Outliers:** Observations far from the estimated line. Identified by large **studentized residuals**.
5. **High Leverage Points:** Observations with unusual predictor values (x_i). Quantified using the **leverage statistic** (h_i).
6. **Collinearity:** Two or more predictors are highly correlated.
   - **Detection:** Calculate the **Variance Inflation Factor (VIF)**:
   $$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{\setminus j}}}$$

4 Comparison of Linear Regression with K-Nearest Neighbors

Feature	**Linear Regression**	**K-Nearest Neighbo**
Type	**Parametric** (assumes $f(X)$ is linear)	**Non-parametric** (ma
Advantages	Easy to fit, coefficients interpretable, better when true $f(X)$ is close to linear.	Highly flexible, perforr
Disadvantages	Poor performance if true $f(X)$ is non-linear (**high bias**).	Performance degrades

4.1 Key Takeaway

If the true relationship $f(X)$ is nearly linear, the parametric approach (linear regression) is usually superior. If $f(X)$ is highly non-linear and p is small, the non-parametric approach (KNN) may be superior.

5 Lab: Linear Regression (Practical Skills)

- **Model Fitting and Summarizing:** Using `sm.OLS()` to fit models and view coefficients, standard errors, and p-values.
- **Diagnostic Tools:** Plotting leverage statistics and calculating **VIFs**.
- **Model Extensions:** Including interaction terms and non-linear transformations (e.g., `poly('horsepower', 2)`).