Further OLS Assumptions

Building on the previous section, let's cover the other key assumptions of Ordinary Least Squares (OLS) Linear Regression, which are particularly important for the validity of statistical inference (like hypothesis tests on coefficients).

4. Error Terms are Uncorrelated to each other (Independence of Errors)

* **Statement:** The error term (ε) for one observation should not be correlated with the error term of any other observation. There should be no pattern of dependency between the residuals.
* **Interpretation:** The value of an individual random discrepancy (error) between an observed and predicted value must not provide any information about the value of another discrepancy.
* **Violation (Autocorrelation):** This assumption is often violated in **time-series data**, where the error in one period might be correlated with the error in the previous period (this is called autocorrelation or serial correlation).
* **Impact:** While OLS coefficients remain unbiased, the standard errors become unreliable, leading to incorrect p-values and confidence intervals. Efficiency is also reduced.

5. The Error Term has a Constant Variance (Homoscedasticity)

* **Statement:** The variance of the error terms (ε) should be consistent across all levels of the predictor variables (or across all predicted values).
* **Concept:** This condition is called **Homoscedasticity** (meaning constant variance).
* **Violation (Heteroscedasticity):** If the variance of the errors changes for different predicted values (e.g., errors are larger for higher predicted prices), the condition is violated, and we have **Heteroscedasticity**.
* **Detection:** Often checked by plotting the residuals (or standardized residuals) against the fitted (predicted) values. A random scatter of points around zero suggests homoscedasticity, while a pattern (like a cone shape) indicates heteroscedasticity.
* **Impact:** Similar to autocorrelation, OLS coefficients remain unbiased, but standard errors are incorrect, invalidating hypothesis tests and confidence intervals.

6. Absence of Perfect Multicollinearity

* **Statement:** In a multiple linear regression model, there shouldn't be perfect linear relationships between the predictor variables. One predictor variable should not be a perfect linear combination of other predictor variables.
* **Concept:** This phenomenon is called **Multicollinearity** (or collinearity if between two variables). While *perfect* multicollinearity makes estimation impossible, high (but not perfect) multicollinearity is also problematic.
* **Interpretation:** If predictors are highly correlated, it becomes difficult for the model to isolate the individual effect of each predictor on the target variable. The coefficient estimates become unstable and highly sensitive to small changes in the data or model specification. Standard errors inflate, making coefficients appear statistically insignificant even if they are relevant.
* **Example:** Including a person's weight in both pounds and kilograms as separate predictors in the same model. Including 'total rooms' and also 'bedrooms' and 'other rooms' might also cause high multicollinearity.
* **Impact:** Unreliable coefficient estimates and standard errors, difficulty interpreting the individual importance of predictors.

7. Residual Normality (Optional but Recommended)

* **Statement:** The error terms (ε) should be normally distributed.
* **Importance:** This assumption is technically **not required** for OLS to produce unbiased and consistent coefficient estimates (especially in large samples due to the Central Limit Theorem). However, it **is required** for valid hypothesis testing (t-tests, F-tests) and the construction of accurate confidence intervals, particularly in smaller samples.
* **Detection:** Often checked using histograms of residuals, Q-Q plots, or P-P plots, or statistical tests like the Shapiro-Wilk test.

Interpreting Model Results (Statistical Output)

When you fit a linear regression model using libraries like statsmodels, you typically get a detailed summary table. Here's how to interpret some key components:

* **R-Squared (Coefficient of Determination):**
  + **Recap:** Measures how well the regression model approximates the observed values (goodness-of-fit). It represents the proportion of the variance in the target variable explained by the predictor variables in the model.
  + **Range:** 0 to 1. Higher is generally better (closer to 1 indicates more variance explained).
* **Adjusted R-Squared:**
  + **Problem with R-Squared:** R-Squared will *always* increase or stay the same whenever you add *any* predictor variable to the model, even if that variable has no real explanatory power. This can be misleading.
  + **Solution:** Adjusted R-Squared penalizes the R-Squared value based on the number of predictors (p) in the model relative to the sample size (N). It only increases if the added variable improves the model more than would be expected by chance.
  + **Formula:** Adjusted R² = 1 - [ ((1 - R²) \* (N - 1)) / (N - p - 1) ]
  + **Interpretation:** Generally considered a better measure than R-Squared for comparing models with different numbers of predictors.
* **F-Statistic:**
  + **Purpose:** Tests the overall significance of the entire regression model. It compares the variance explained by your model (mean regression sum of squares) to the unexplained variance (mean error sum of squares).
  + **Null Hypothesis (H₀):** All regression coefficients (β₁, β₂, ..., β<0xE2><0x82><0x99>) are simultaneously equal to zero (i.e., the model has no explanatory power).
  + **Interpretation:** A large F-statistic suggests that at least one predictor variable is significantly related to the target variable, allowing you to reject the null hypothesis.
* **Prob (F-Statistic):**
  + **Purpose:** This is the p-value associated with the F-statistic.
  + **Interpretation:** It represents the probability of observing an F-statistic as large as (or larger than) the one calculated, *assuming the null hypothesis is true*.
  + **Decision:** If Prob (F-Statistic) is very small (typically < 0.05), we reject the null hypothesis and conclude that the overall model is statistically significant (i.e., at least one predictor is useful). A value close to 0 (like in the slide example) strongly indicates the model has explanatory power.
* **'coeff' (Coefficients - β₀, β₁, ...):**
  + **Recap:** These are the estimated values for the intercept (β₀) and the slopes (β₁, β₂, ...) for each individual predictor variable.
  + **Interpretation (Slope βᵢ):** Represents the estimated change in the target variable (y) for a one-unit increase in the predictor xᵢ, holding all other predictors constant.
* **P-Value (P > |t|):**
  + **Purpose:** Associated with each individual coefficient (βᵢ). It tests the statistical significance of *that specific predictor variable*.
  + **Null Hypothesis (H₀):** The coefficient for that predictor is zero (βᵢ = 0), meaning the predictor has no statistically significant linear relationship with the target variable, after accounting for other predictors in the model.
  + **Interpretation:** The p-value is the probability of observing a t-statistic (which measures how many standard errors the coefficient is away from zero) as extreme as (or more extreme than) the one calculated, *assuming the null hypothesis (βᵢ = 0) is true*.
  + **Decision:** If the P-value is less than a chosen significance level (commonly α = 0.05), we reject the null hypothesis. This indicates that the predictor is statistically significant in explaining the target variable's values (its relationship is unlikely due to random chance).

Eliminating Multicollinearity using VIF

As mentioned in the assumptions, high multicollinearity (strong linear relationships between predictor variables) is undesirable.

* **Definition:** Multicollinearity is the phenomenon of how well a predictor variable can be explained (predicted) by the *other* predictor variables in the dataset.
* **Why it's bad:** It confounds the model, making it difficult to determine and isolate the independent influence of each correlated variable on the target. Coefficient estimates become unstable and unreliable.
* **Common Causes:**
  + Including variables that are naturally associated (e.g., household income and monthly expenditure).
  + Creating derived variables and retaining the original ones (e.g., having length in 'cm' and also length in 'inches').
  + Not removing one dummy variable for each categorical feature after one-hot encoding (as discussed previously).

Variance Inflation Factor (VIF)

A common method to detect and quantify multicollinearity is the **Variance Inflation Factor (VIF)**.

* **Method:** VIF calculates how much the variance of an estimated regression coefficient is increased due to its correlation with other predictors.
  1. Take one predictor variable (Xⱼ).
  2. Regress Xⱼ against *all other* predictor variables in the model.
  3. Calculate the R² value from *this* regression (let's call it R²ⱼ). This R²ⱼ tells you how much of the variance in Xⱼ is explained by the other predictors.
  4. Calculate the VIF for Xⱼ using the formula:
  5. VIFⱼ = 1 / (1 - R²ⱼ)
* **Interpretation:**
  1. VIF = 1: No correlation between the predictor Xⱼ and any others (Ideal). R²ⱼ = 0.
  2. VIF between 1 and 5: Generally considered acceptable or moderate multicollinearity.
  3. VIF > 5 or VIF > 10: Often considered indicative of problematic multicollinearity. It means the variance of the coefficient estimate for Xⱼ is inflated by a factor of VIF due to its correlations with other predictors. (Note: These thresholds are rules of thumb).
* **Usage:** Calculate VIF for each predictor. If high VIFs are found, iteratively remove the predictor with the highest VIF (above the chosen threshold) and recalculate VIFs for the remaining predictors until all are within acceptable limits.

Feature Selection: Recursive Feature Elimination (RFE)

Often, not all features collected are relevant or necessary for building a good predictive model. Including irrelevant features can increase model complexity, potentially introduce noise, and sometimes worsen predictive performance. **Feature selection** aims to select a subset of the most relevant features.

**Recursive Feature Elimination (RFE)** is one such algorithm:

* **Goal:** To find an optimal subset of features by starting with all features and successively removing the least important ones.
* **Process:**
  1. **Fit Model:** Train the chosen machine learning algorithm (e.g., Linear Regression) using the *current* set of features.
  2. **Rank Features:** Rank the features based on some measure of importance derived from the model (e.g., the magnitude of the coefficients in Linear Regression, although other importance measures exist for different models).
  3. **Eliminate:** Remove the least important feature(s).
  4. **Repeat:** Repeat steps 1-3 with the reduced feature set until a specified number of features remains, or until performance (evaluated via cross-validation) starts to degrade.
* **Benefit:** Helps simplify the model, potentially improve performance by removing noise, and reduce computational cost. sklearn.feature\_selection.RFE implements this in Scikit-learn.

Testing your Linear Regression Model: Residual Analysis

A critical step after training a model is to perform **Residual Analysis** to check whether the OLS assumptions appear to hold. This involves examining the residuals (yᵢ - ŷᵢ).

Residual Normality Check

* **Assumption Recap:** Errors should ideally be normally distributed for valid hypothesis testing.
* **Method 1: Histogram:** Plot a histogram of the residuals. If the assumption holds, the histogram should resemble a bell shape (normal distribution curve), centered around zero.
* **Method 2: P-P Plot (Probability-Probability Plot):**
  + This plot compares the cumulative probability distribution of the residuals against the cumulative probability distribution of a theoretical normal distribution.
  + If the residuals are normally distributed, the points on the P-P plot should fall approximately along the 45-degree diagonal line. Significant deviations suggest non-normality.

Homoscedasticity Check

* **Assumption Recap:** The variance of the residuals should be constant across all levels of predicted values.
* **Method: Residuals vs. Fitted Plot:** Create a scatter plot with the predicted (fitted) values (ŷ) on the x-axis and the residuals (or standardized residuals) on the y-axis.
  + **Homoscedasticity:** If the assumption holds, the plot should show a random scatter of points around the horizontal line at zero, with roughly constant vertical spread across the range of fitted values.
  + **Heteroscedasticity:** If the plot shows a pattern (e.g., a funnel or cone shape where the spread increases or decreases with fitted values), the assumption is violated.

Performing these checks helps validate the model and understand the reliability of its statistical outputs. If assumptions are significantly violated, transformations of variables or using different modeling techniques (like Weighted Least Squares) might be necessary.