Model Assumptions and Residual Analysis

Overview

1. Model Assumptions

2. Variation in the model

Motivation

We have build a linear model for predicting a response variable,

- Why do we need assumptions? The answer is that we need probabilistic assumptions to ground statistical inference about the model parameters.
- Are the assumptions met for the model we built? We need to diagnose the model to ensure the assumptions are met.

Linear Model

The linear model:

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

where ε is a random variable and β_i are unknown parameters.

Model Assumptions

We need an approach to quantify the variability of the estimator $\hat{\beta}$ to infer properties about the unknown population parameters/coefficients β from the sample data.

The assumptions of the linear model are:

1. Linearity Assumption: Is there a linear relationship between response and predictors?
2. Independence Assumption: are the $\varepsilon_i's$ independent (uncorrelated)?
3. Equal Variance Assumption (or Homoscedasticity): is $V(\$\varepsilon) = \sigma^2$?
4. Normal Population Assumption: $\varepsilon \sim N(0, \sigma^2)$?

Checking the Model Assumptions

Models are useful only when specific assumptions are reasonable. We check conditions that provide information about the assumptions.

1. Linearity Assumption: Check each of the predictors against the response for linearity. Also check the residual plots. Any patterns, especially bends or nonlinearities, are signs that the condition is not met.
2. Independence Assumption: The sample data should be randomly selected from the population. Also check the residual plot for lack of patterns or clumping.
3. Equal Variance Assumption (or Homoscedasticity): $V(\$\varepsilon) = \sigma^2$. The variability in the errors for a given predictor should be consistent. A scatterplot of residuals versus predicted values should not display any discernible pattern, such as a cone-shaped distribution, which would indicate heteroscedasticity. Addressing heteroscedasticity might involve data transformations.
4. Normal Population Assumption: Examine the QQ plot of the residuals to check for normality.

Variation in the model: Setting up for overall model usefulness

There are two variations in our linear regression model: Explained variation and Unexplained Variation.

Total Variation (Sum of Square Total, SST) = Explained Variation (Sum of Square Regression, SSR) + Unexplained Variation (Sum of Square Error, SSE)

1. Total Variation: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
2. Explained Variation: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
3. Unexplained Variation: $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$

Coefficient of determination, R^2, is the proportion of variation in the response, Y, that can be explained by the linear model. That is,

$$R^2 = \frac{SSR}{SST}.$$

Comparing R^2 and Adjusted R^2

1. R^2:
   - $R^2 = \frac{SSR}{SST}$: the proportion of variation in the response, Y, that can be explained by the linear model
   - R^2 tends to increase as more predictors are added to the model, even if those predictors do not improve the model's performance meaningfully
2. Adjusted R^2:
   - Adjusted $R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$: provides a more accurate measure by adjusting for the number of predictors, helping to evaluate the model's performance more reliably and prevent overfitting
   - adjusts the regular R^2 for the number of predictors in the model. It accounts for the fact that adding more predictors to the model can artificially inflate R^2 even if those predictors are not meaningfully improving the model.
   - Adjusted R^2: proportion of the variability in the response variable that can be explained by the linear model after adjusting for the number of predictors.

More on the Independent Assumption...

Inherent in the independence assumption is we do not want the predictor variables are not too highly correlated with each other, a condition known as Multicollinearity. Multicollinearity may be tested with three central criteria:

1. Correlation matrix - correlation coefficients for pairwise comparisons between predictors should ideally be below 0.80.
2. Tolerance - the tolerance measures the influence of one predictor variable on all other predictor variables. Tolerance is defined as $T = 1 – R^2$ for the first step regression analysis. If T $<$ 0.1 there might be multicollinearity issues and with T $<$ 0.01 there certainly is multicolinearity.
3. Variance Inflation Factor (VIF) - the variance inflation factor of the linear regression is defined as VIF $= 1/$T. With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the predictor variables.

Simple solution: Remove predictor variables with high VIF values.