Lecture Note 8: Multiple Linear Regression

1. Introduction

Multiple linear regression (MLR) model is a natural progression from the simple linear regression model, expanding its capabilities to accommodate and analyze the relationship between a single dependent variable and two or more independent variables. While simple linear regression maps the relationship between a single predictor and a response variable using a straight line, MLR extends this concept to a higher-dimensional space, allowing for a more nuanced understanding of how various predictors collectively influence the outcome.

The transition from simple to multiple linear regression marks a significant step in modeling complexity, enabling analysts and researchers to account for multiple influencing factors simultaneously. This is particularly useful in real-world scenarios where outcomes are rarely the result of a single factor. For example, predicting the sales of a product could depend on factors such as price, marketing spend, seasonality, and competitor actions, among others. MLR allows for the inclusion of all these variables in a single model, providing a richer, more accurate representation of the underlying phenomena.

Moreover, MLR facilitates the exploration of the relative importance of each predictor in explaining the variance in the dependent variable, offering insights into which factors have the most significant impact. This extension from simple to multiple regression enhances the analytical toolkit available for tackling complex, multifaceted problems, making it a cornerstone of predictive modeling in various fields such as economics, finance, health sciences, and beyond.

2. MLR Model

At its core, MLR seeks to fit a linear equation to observed data where the response variable is thought to be a linear combination of the independent variables, each contributing to the prediction according to its own coefficient. This model is represented by the equation:

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

- Y: the dependent variable or target variable you try to predict or explain.
- $X_1, \cdots X_p$: independent variables or features or predictors that you use to predict Y.
- β_0: Y - intercept, representing the expected value of Y when all features are zero.
- $\beta_1, \cdots \beta_p$: coefficients that represent the change in Y for a one-unit change in X, holding other variables constant.
- ε: error term capturing the unexplained variability in Y.

Ordinary Least Squares (OLS) is a method used in linear regression to estimate the coefficients (beta values) of the model, ensuring the best possible fit between the observed data and the model's predictions. In the context of multiple linear regression, just like with simple linear regression, the goal of OLS is to minimize SSE: the sum of the squared errors (residuals) between the observed values and the values predicted by the model. The optimal estimated value of the beta coefficients in a multiple linear regression (MLR) model is commonly represented by the symbol b. Consequently, the equation for the estimated MLR model can be expressed as follows:

$$\hat{Y} = b_0 + b_1X_1 + \cdots + b_pX_p$$

In essence, b_i is the estimated change in the dependent variable, Y, resulting from a one-unit increase in the independent variable X_i , while keeping all other variables constant.

In practice, alternative optimization techniques like gradient descent are commonly employed by software to computationally determine the estimated coefficients in a model. Unlike the analytical approach of ordinary least squares, which directly computes the coefficient estimates, gradient descent iteratively adjusts the coefficients to minimize the cost function—the measure of error between the model's predictions and the actual data, such as SSE or MSE. This method is particularly useful for large datasets or complex models where the analytical solution is computationally intensive or infeasible. Through iterative refinement, gradient descent seeks the set of coefficients that result in the lowest possible error, effectively finding the optimal values that best fit the model to the data.

3. Assumptions

Linear regression, with its simplicity and ease of interpretation, has enjoyed widespread application across various industries and academic fields for many decades. However, the effectiveness and validity of linear regression models hinge critically on several key assumptions. These assumptions, which are foundational to the model's integrity, unfortunately, are not always met in real-world data. The assumptions in question include ***linearity***, which posits a linear relationship between the independent and dependent variables; ***no multicollinearity***, which requires that the predictor variables are not too highly correlated with each other; ***homoscedasticity***, ensuring that the variance of error terms is constant across all levels of the independent variables; ***normality of residuals***, which assumes that the error terms are normally distributed; and ***independence of errors***, stating that the residuals (errors) are independent of each other. Each of these assumptions plays a crucial role in the model's ability to accurately capture and predict the underlying relationship in the data.

3.1 Linearity

The linearity assumption in multiple linear regression posits that there is a linear relationship between each independent variable and the dependent variable. This means that the change in the dependent variable is expected to be a linear function of changes in any independent variable, holding all other independent variables constant. Essentially, it suggests that the effect of altering an independent variable on the predicted outcome is constant, which is a cornerstone for the linear regression model to provide accurate predictions.

To assess whether the linearity assumption holds, you can employ several methods:

1. **Scatter Plots**: Plotting scatter plots of the dependent variable against each independent variable can visually indicate whether a linear relationship exists. Non-linear patterns, such as curves or clusters, may suggest violations of the linearity assumption.
2. **Residual Plots**: Plotting the residuals (the differences between the observed and predicted values) against the predicted values or against each independent variable can also help assess linearity. Ideally, these plots should show no clear pattern; the presence of a systematic structure (e.g., a curve) indicates that the relationship might not be linear.
3. **Partial Regression Plots**: For multiple linear regression, partial regression plots can help visualize the relationship between the dependent variable and each independent variable, controlling for

the presence of other independent variables. These can give a clearer picture of linearity between each predictor and the outcome.

If the linearity assumption does not hold, there are several approaches to address the issue:

1. **Transformation of Variables**: Applying transformations to the independent and/or dependent variables can sometimes linearize relationships. Common transformations include logarithmic, square root, or power transformations.
2. **Polynomial Regression**: Extending the model to include polynomial terms allows it to capture non-linear relationships while still using the linear regression framework.
3. **Adding Interaction Terms**: Sometimes, the non-linearity arises from the interaction between variables rather than from the variables themselves. Including interaction terms can capture these effects.
4. **Non-Linear Models**: When linear approaches are insufficient to model the relationship accurately, it may be necessary to consider non-linear models that are more flexible in capturing complex relationships, such as generalized additive models (GAMs) or machine learning algorithms like decision trees or neural networks.

It's important to address violations of the linearity assumption, as failing to do so can lead to biased or inaccurate estimates, affecting the model's predictive performance and the validity of any inferences drawn from it. By carefully assessing linearity and applying appropriate remedies, you can improve the model's fit and ensure more reliable results.

3.2 No Multicollinearity

The no multicollinearity assumption in multiple linear regression stipulates that the independent variables should not be too highly correlated with each other. Multicollinearity can be problematic because it undermines the statistical significance of the independent variables, making it difficult to distinguish their individual effects on the dependent variable. High multicollinearity can lead to inflated standard errors for the coefficient estimates, resulting in wider confidence intervals and less reliable statistical tests.

There are several ways to assess multicollinearity in your data:

1. **Correlation Matrix**: A correlation matrix among all independent variables can provide a quick visual inspection of potential multicollinearity. High correlation coefficients between pairs of variables indicate multicollinearity.
2. **Variance Inflation Factor (VIF)**: VIF quantifies the extent of multicollinearity in an ordinary least squares regression analysis. It provides a measure of how much the variance of an estimated regression coefficient increases if your predictors are correlated. A VIF value greater than 5 or 10 (depending on the source) suggests significant multicollinearity that needs to be addressed.
3. **Tolerance**: Tolerance is the inverse of VIF and measures the amount of variability of the selected independent variable not explained by the other independent variables. A low tolerance value (close to 0) indicates multicollinearity.

If multicollinearity is detected, several strategies can be employed to address it:

1. **Remove Highly Correlated Predictors**: One approach is to remove one of the predictors from the model if two predictors are highly correlated with each other. The choice of which variable to remove can be based on theoretical considerations or the variable's relevance to the analysis.
2. **Combine Correlated Variables**: If the correlated variables convey similar information, consider combining them into a single predictor through methods such as averaging or summing.

3. **Principal Component Analysis (PCA)**: PCA is a dimensionality reduction technique that can be used to transform the correlated variables into a set of uncorrelated variables (principal components) that can be used as new predictors in the regression model.
4. **Ridge Regression**: Ridge regression is a type of regularization technique that can handle multicollinearity by adding a penalty term to the regression model. This method shrinks the coefficients of correlated predictors, thereby reducing their impact on the model.

It's important to address multicollinearity because it can significantly affect the interpretability and predictive power of your regression model. By identifying and mitigating multicollinearity, you ensure more reliable and meaningful statistical inferences.

3.3 Homoscedasticity

The homoscedasticity assumption in multiple linear regression implies that the variances of the residuals (the differences between the observed and predicted values) are constant across all levels of the independent variables. This means that as the value of the predictors changes, the spread of the residuals remains the same. Homoscedasticity is crucial for the reliability of the regression model's standard errors, confidence intervals, and hypothesis tests. If the residuals exhibit heteroscedasticity (the opposite of homoscedasticity), where their variances are not constant, it can lead to inefficiencies in the model estimates and weaken the statistical conclusions that can be drawn from the model.

There are several methods to assess whether the homoscedasticity assumption holds:

1. **Residual vs. Fitted Values Plot**: Plotting the residuals against the fitted values (predicted values) is a straightforward visual method. In a homoscedastic model, this plot should not display any systematic patterns or funnel shapes; the spread of the residuals should be roughly the same across all levels of the fitted values.
2. **Breusch-Pagan Test**: This is a statistical test specifically designed to detect the presence of heteroscedasticity. If the test indicates the presence of heteroscedasticity, it suggests that the homoscedasticity assumption may not hold.

If heteroscedasticity is detected, there are several approaches to address it:

1. **Transforming the Dependent Variable**: Applying transformations to the dependent variable, such as logarithmic, square root, or inverse transformations, can help stabilize the variance of the residuals.
2. **Using Weighted Least Squares (WLS)**: WLS is an extension of ordinary least squares (OLS) that assigns weights to each data point based on the variance of its residuals. By giving less weight to observations with higher variance, WLS can effectively address heteroscedasticity.
3. **Adding Missing Variables**: Sometimes, heteroscedasticity arises because important predictors or interaction terms are missing from the model. Including these variables can sometimes solve the problem by capturing the unexplained variation in the residuals.
4. **Robust Standard Errors**: Another approach is to use robust standard errors, which are designed to be valid even in the presence of heteroscedasticity. This does not solve the underlying problem but allows for more reliable hypothesis testing despite heteroscedasticity.

Addressing heteroscedasticity is important for ensuring the accuracy and reliability of the regression model's inferences. By carefully assessing for homoscedasticity and applying appropriate remedies when needed, you can improve the robustness of your regression analysis.

3.4 Normality of Residuals

The normality of residuals assumption in multiple linear regression posits that the residuals (errors) from the model are normally distributed. This assumption is crucial for the validity of various inferential statistics associated with the regression, such as t-tests and F-tests, which rely on normality to produce accurate significance levels for hypothesis testing regarding regression coefficients.

To check if the normality assumption holds, you can use the following methods:

1. **Histogram of Residuals**: Plotting a histogram of the residuals can provide a visual assessment of normality. If the residuals are normally distributed, the histogram should resemble a bell-shaped curve.
2. **Q-Q Plot (Quantile-Quantile Plot)**: This plot compares the quantiles of residuals to the quantiles of a normal distribution. If the points lie approximately along a straight line, it suggests that the residuals are normally distributed.
3. **Statistical Tests**: Formal tests for normality include the Shapiro-Wilk test and the Anderson-Darling test. These tests provide a p-value, with a high p-value (typically >0.05) indicating that there is no evidence against normality.

If the normality assumption does not hold, consider the following remedies:

1. **Transforming the Dependent Variable**: Applying a transformation (e.g., log, square root, inverse) to the dependent variable can help achieve normality in the residuals. The choice of transformation depends on the pattern of non-normality.
2. **Adding Missing Variables**: Non-normal residuals can sometimes result from omitting relevant predictors or interaction terms that capture systematic information in the residuals. Adding these variables can help.
3. **Robust Regression Methods**: If normality cannot be achieved through transformations or model adjustments, robust regression techniques that do not rely on the assumption of normality can be used. These methods are less sensitive to non-normal residuals.
4. **Bootstrapping**: This is a resampling technique that can be used to derive more accurate estimates of the standard errors and confidence intervals for the regression coefficients without relying on the assumption of normality.

It's important to address deviations from normality because they can affect the reliability of hypothesis tests and confidence intervals derived from the regression model. By carefully assessing the distribution of residuals and applying appropriate remedies, you can ensure that your regression analysis remains robust and reliable, even when the normality assumption is challenged.

3.5 Independence of Errors

The independence of errors assumption in multiple linear regression states that the residuals (errors) from the regression model should be independent of each other. This means that the value of one error term should not predict the value of another error term. The assumption is crucial for ensuring the validity of statistical tests, as correlated errors can lead to biased estimates of regression coefficients and undermine the reliability of confidence intervals and hypothesis tests.

Assessing the independence of observations can be challenging, but some methods include:

1. **Plotting Residuals**: Inspecting a plot of residuals versus time or the order of data collection can help detect patterns indicating dependence. For time series data, this might show up as autocorrelation, where residuals at one time are correlated with residuals at another time.

2. **Durbin-Watson Test**: This statistical test specifically checks for autocorrelation in the residuals of a regression model. A Durbin-Watson statistic close to 2 suggests no autocorrelation, while values deviating far from 2 indicate positive or negative autocorrelation.
3. **Examining Study Design**: Sometimes, the study design itself can suggest potential violations. For example, clustered or repeated measures data inherently violate the independence assumption.

If the independence assumption does not hold, consider the following remedies:

1. **Time Series Analysis**: For data that are ordered in time and show autocorrelation, time series analysis techniques, such as ARIMA models, can be used to model the data appropriately, taking into account the time-dependent structure.
2. **Generalized Estimating Equations (GEE)**: GEE can be used for clustered or longitudinal data where observations within clusters may be correlated.
3. **Mixed Effects Models**: Also known as hierarchical linear models, these can account for both fixed effects (observed variables) and random effects (unobserved variables that contribute to the correlation among observations).
4. **Robust Standard Errors**: Adjusting standard errors to account for clustering or serial correlation can mitigate the impact of non-independence on inference, though it does not address the root cause.

Ensuring the independence of observations is key to the integrity of a regression analysis. By properly assessing and addressing any violations of this assumption, you can enhance the credibility and reliability of your statistical findings.

3.6 "Independence" in Linear Regression

The term "independence" can often lead to confusion or misinterpretation, even within the realm of linear regression. Previously, we delved into the concept of independence concerning errors. In the context of linear regression models, when we speak of independence as an assumption, it usually pertains to the independence of errors. Occasionally, however, individuals may allude to the absence of multicollinearity as the linear independence of independent variables or features, ensuring a solution to the ordinary least squares method. Low multicollinearity further requires that the correlation among features is low. This ensures a more robust linear regression model with a low standard error.

Moreover, there exists yet another assumption known as independence of observations. It refers to the premise that each observation in the dataset is collected without being influenced by the other observations. In other words, the data points are not related to each other in any way that affects the analysis. This assumption is often implicitly assumed in regression models without explicit mention. Independence of observations pertains to the data collection process and whether the data points themselves are independent, while independence of errors is concerned with the residuals of a model being independent after accounting for the relationship modeled. Independence of errors is a more specific condition that needs to be met to ensure reliable regression analysis results, even if the original observations are independent. Dependence among observations can indeed be a cause of dependence among errors in statistical modeling, particularly in regression analysis. Essentially, while independent observations are a broader requirement for many statistical analyses, independent errors are specifically critical for the validity of regression models.

4. Example (forthcoming)