Regression analysis is a powerful statistical method used to examine the relationship between one dependent variable and one or more independent variables.

**Types of Regressions**

1. **Linear Regression**:
   * **Simple Linear Regression**: Examines the relationship between two continuous variables (one independent and one dependent).
   * **Multiple Linear Regression**: Involves two or more independent variables predicting a single dependent variable.
2. **Polynomial Regression**: A form of linear regression where the relationship between the independent variable and the dependent variable is modeled as an nth degree polynomial.
3. **Logistic Regression**: Used when the dependent variable is binary (0 or 1). It predicts the probability of a binary outcome based on one or more predictor variables.
4. **Ridge Regression**: A type of linear regression that includes a regularization term (L2) to prevent overfitting by penalizing large coefficients.
5. **Lasso Regression**: Similar to ridge regression, but it uses L1 regularization, which can shrink some coefficients to zero, thus performing variable selection.
6. **Elastic Net Regression**: Combines the penalties of ridge regression (L2) and lasso regression (L1) to improve model performance, especially when dealing with highly correlated variables.
7. **Stepwise Regression**: Involves automatically selecting significant variables by adding or removing predictors based on specific criteria (e.g., p-values, AIC).
8. **Hierarchical Regression**: Used to examine the incremental value of adding sets of variables to a regression model. Often used in social sciences to assess the importance of different predictors.
9. **Quantile Regression**: Estimates the relationship between variables for specific quantiles (percentiles) of the dependent variable, providing a more comprehensive analysis of the data distribution.
10. **Nonlinear Regression**: Used when the relationship between the dependent and independent variables is nonlinear. This can involve various models such as exponential, logarithmic, or power functions.
11. **Robust Regression**: Designed to be less sensitive to outliers and violations of assumptions that standard linear regression relies on.
12. **Bayesian Regression**: Incorporates prior distributions and updates them with the data to form a posterior distribution, providing a probabilistic approach to regression.
13. **Poisson Regression**: Used for count data where the dependent variable is a count or rate, assuming the data follows a Poisson distribution.
14. **Ordinal Regression**: Used when the dependent variable is ordinal, i.e., categories with a meaningful order but not a precise distance between them.
15. **Probit Regression**: Similar to logistic regression but uses a probit link function, suitable for modeling binary outcome variables.

Each type of regression has its own assumptions, strengths, and appropriate use cases, depending on the nature of the data and the research question.

**Mutivariate Regression vs Multiple Linear Regression:**

**A white paper with black text and black text

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**Differences**

1. **Number of Dependent Variables**:
   * Multiple Linear Regression: One dependent variable.
   * Multivariate Regression: Multiple dependent variables.
2. **Complexity**:
   * Multiple Linear Regression: Simpler model focusing on the relationship between several predictors and one outcome.
   * Multivariate Regression: More complex, as it considers the relationships between several predictors and multiple outcomes, as well as the potential correlations among the dependent variables.
3. **Usage**:
   * Multiple Linear Regression: Ideal when the research question involves a single outcome.
   * Multivariate Regression: Necessary when the research question involves multiple outcomes that may be interrelated.
4. **Interpretation**:
   * Multiple Linear Regression: Easier to interpret as it involves one dependent variable.
   * Multivariate Regression: More complex interpretation due to multiple dependent variables and their interrelationships.

**Multicollinearity:**

Multicollinearity refers to a situation in regression analysis where two or more predictor variables (independent variables) are highly correlated. This means that one predictor variable can be linearly predicted from the others with a substantial degree of accuracy. Multicollinearity can create challenges in understanding the relationships between the predictors and the dependent variable, and it can affect the stability and interpretation of the regression coefficients.

**Consequences of Multicollinearity**

1. **Unstable Coefficients**: Regression coefficients can become very sensitive to changes in the model. Small changes in the data can lead to large changes in the coefficients.
2. **Inflated Standard Errors**: Standard errors of the regression coefficients can become large, leading to wider confidence intervals and less precise estimates of the coefficients.
3. **Reduced Statistical Power**: It becomes harder to detect the individual effect of each predictor variable on the dependent variable, making hypothesis tests less reliable.
4. **Interpretation Difficulties**: It becomes difficult to determine the unique contribution of each predictor variable to the dependent variable.

**Detecting Multicollinearity**

1. **Correlation Matrix**: Examine the correlation matrix of the predictor variables. High correlation coefficients (close to +1 or -1) indicate potential multicollinearity.
2. **Variance Inflation Factor (VIF)**: VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value greater than 10 is often considered indicative of significant multicollinearity.
3. **Tolerance**: Tolerance is the reciprocal of VIF (1/VIF). A tolerance value less than 0.1 indicates high multicollinearity.
4. **Condition Index**: Based on the eigenvalues of the predictor variables' correlation matrix. A condition index above 30 suggests severe multicollinearity.

**Addressing Multicollinearity**

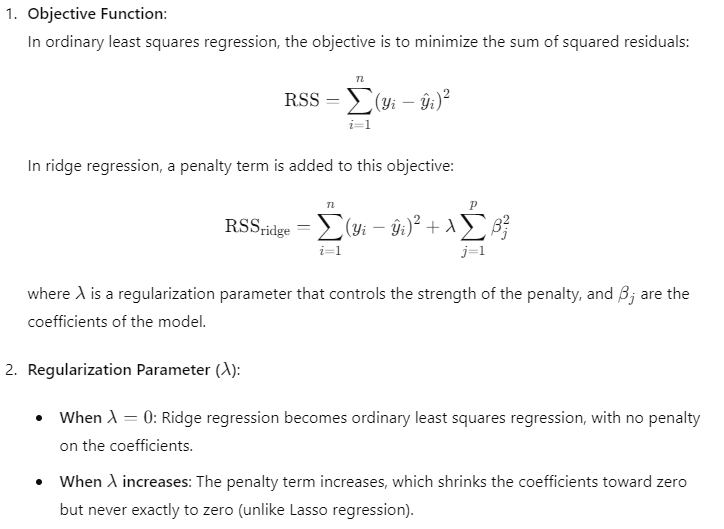
1. **Remove Highly Correlated Predictors**: If two or more predictors are highly correlated, consider removing one of them from the model.
2. **Combine Predictors**: Create a composite index or use principal component analysis (PCA) to combine highly correlated predictors into a single predictor.
3. **Regularization Techniques**: Use ridge regression or lasso regression, which include regularization terms that can help reduce the impact of multicollinearity.
4. **Centering the Variables**: Center the predictor variables by subtracting the mean. This can sometimes help reduce multicollinearity.
5. **Domain Knowledge**: Use domain knowledge to decide which predictors are most important and should be retained in the model.

**A white paper with black text

Description automatically generated**

**Ridge Regularization:**

Ridge regularization, also known as Ridge Regression or Tikhonov regularization, is a technique used in regression analysis to prevent overfitting by adding a penalty term to the least squares objective function. This penalty term is proportional to the square of the magnitude of the coefficients.

****

**A close-up of a white background

Description automatically generated**

**A white paper with black text

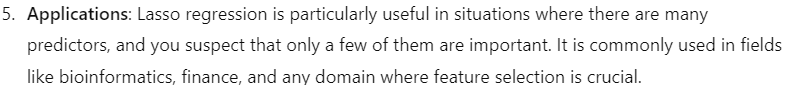
Description automatically generated**

**Lasso Regularization:**

Lasso regression, or Least Absolute Shrinkage and Selection Operator, is a type of linear regression that includes a regularization term. The regularization term is the sum of the absolute values of the coefficients, and it helps to prevent overfitting by shrinking some of the coefficients to zero. This can result in a more interpretable model with fewer predictors.

**A black and white text with a black and white text

Description automatically generated**

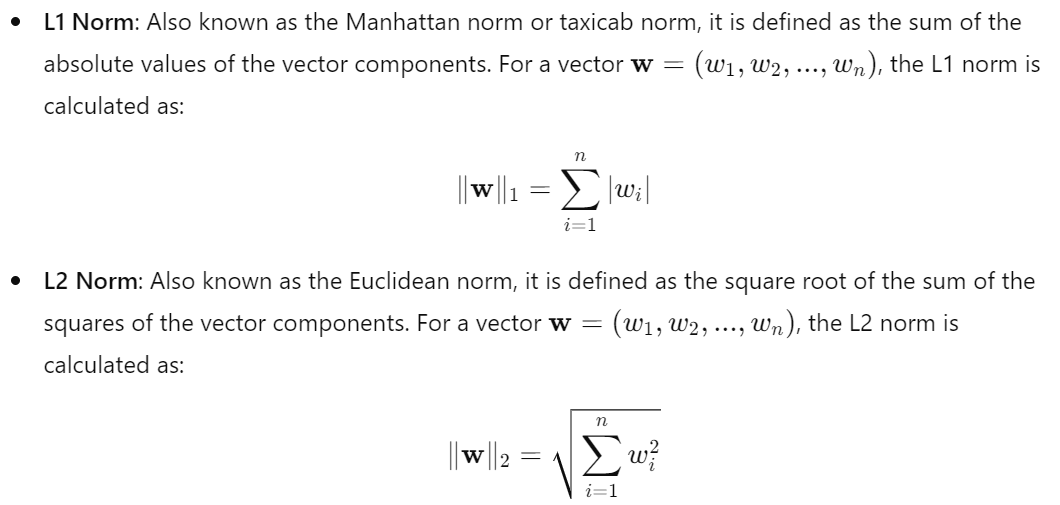
****

|  |  |
| --- | --- |
|  |  |
| **Effect on Coefficients**: Lasso can shrink some coefficients to exactly zero, effectively performing variable selection and simplifying the model. | **Effect on Coefficients**: Ridge regression shrinks coefficients but does not set any coefficients to zero. All features are retained in the model. |
| **Use Case**: Particularly useful when you have a high-dimensional dataset with many features, some of which are irrelevant or redundant. | **Use Case**: Useful when you have many correlated features, as it tends to distribute the coefficients more evenly across correlated variables. |
| **Interpretability**: By producing sparse models (with some coefficients exactly zero), lasso regression can be easier to interpret. | **Interpretability**: Ridge regression retains all features, so while it can handle multicollinearity well, it does not simplify the model as much as lasso. |
| **Optimization**: The L1 penalty makes the optimization problem non-differentiable at zero, which can complicate the optimization process. | **Optimization**: The L2 penalty keeps the optimization problem differentiable, making it easier to solve. |
| Lasso uses L1 (absolute value) | Ridge uses L2 (squared value) |
| Lasso can set some coefficients to zero | Ridge only shrinks coefficients to zero |
| Lasso for feature selection and simpler models | Ridge for dealing with multicollinearity and retaining all features. |
| Lasso's L1 penalty can lead to non-differentiable points | Ridge's L2 penalty is smooth and differentiable. |

**Lasso Regression vs Ridge Regression Differences:**

In practice, the choice between lasso and ridge regression depends on the specific characteristics of the dataset and the goals of the analysis. Sometimes, a combination of both techniques, called Elastic Net, is used to leverage the benefits of both L1 and L2 regularization.

Lasso regression is referred to as L1 regularization and ridge regression as L2 regularization due to the type of norm used in their penalty terms.

****

**Cross Validation:** When creating a predictive model, we usually split the data into training and testing parts. With cross-validation, on the other hand, we divide it into three – training, testing, and validation.

**A diagram of a data flow

Description automatically generated A diagram of training

Description automatically generated**

We go with the first fold for a validation set, so the quantity of the remaining folds would be ‘k -1. With the data divided this way, we need to pick a starting value for the tuning parameter – “lambda one”. Say we choose 0.1. Then, we fit the ‘k - 1’ training folds into our model, using lambda = 0.1 as a tuning parameter, and establish values for the coefficients in the ridge regression equation. Next, we use the obtained coefficients and the independent ‘X’ values from the validation set to estimate the predicted y values for the validation data. With the predicted y and the real y values in the validation fold, we can calculate the sum of squares error. We perform this operation with all other folds for validation sets. We choose to have five here, which means we can have five different options. Consequently, there will be five different results for the sum of squares error. We then must sum these results and measure how well the model works. After that, we will repeat the same operation with different values for lambda, depending on the dataset size – 0.1, 0.2, 0.3 till 10 for instance. The lambda leading to the lowest SSE would be the correct choice for our tuning parameter. Thus, we can fit the whole training set with the lambda value in question.