

Understanding Partial Least Squares Regression: A Guide to Overcoming Multicollinearity

Authored by
Mohammed loot

November 6, 2025

RECOMMENDED CITATION

Mohammed loot (2025). *Understanding Partial Least Squares Regression: A Guide to Overcoming Multicollinearity*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=11763>

The Challenge of Multicollinearity in Predictive Modeling

In the complex landscape of [predictive modeling](#) and statistical analysis, a fundamental obstacle frequently encountered is [multicollinearity](#). This statistical phenomenon describes a situation where two or more predictor variables (also known as independent variables) within a dataset are highly linearly correlated with one another. While correlation among predictors is common, severe multicollinearity can profoundly destabilize traditional regression techniques.

The presence of this high correlation introduces significant threats to both the stability and the interpretability of foundational regression models, most notably **Ordinary Least Squares (OLS)**. When predictors are nearly redundant, the model struggles to isolate the unique effect of each variable. Mathematically, this manifests as an inflation of the variance associated with the regression coefficients. Such inflated variance renders the coefficient estimations highly unreliable, causing them to be overly sensitive to minor fluctuations or errors within the training data.

When coefficients become unstable, the resulting model faces a critical generalization problem. It may exhibit an excellent fit--a low error rate--on the specific training dataset used for calibration, but it often fails dramatically when applied to new, unseen data. This failure to generalize is the classic symptom of model [overfitting](#), where the algorithm has learned the noise and specific correlated structure of the training set rather than the underlying signal. Addressing multicollinearity is thus a necessary step toward building robust and reliable statistical models.

Limitations of Principal Components Regression (PCR)

To effectively mitigate the detrimental effects of [multicollinearity](#), data scientists often turn to methods centered on **dimensionality reduction**. One such historically significant and popular approach is [Principal Components Regression](#) (PCR). The methodology behind PCR involves transforming the original P predictor variables (X) into a smaller, orthogonal set of M latent variables, known as "principal components." These components are linear combinations of the original predictors, designed to capture the maximum possible variance within the predictor space.

Once these M principal components have been calculated, PCR proceeds by utilizing the standard method of [Least Squares](#) to construct a linear regression model. Critically, these newly derived, uncorrelated components are used as the predictors in place of the original, highly correlated variables. This transformation successfully eliminates the multicollinearity problem, achieving a streamlined, orthogonal predictor space.

Despite its effectiveness in data compression and orthogonalization, PCR suffers from a fundamental conceptual flaw: it is an unsupervised technique concerning the target variable. Specifically, the construction of the principal components focuses exclusively on maximizing the magnitude of variance captured among the predictor variables (X). It **does not** incorporate any

information from the [response variable](#) (Y) during the component calculation stage.

This oversight means that components explaining the greatest variance in the input data (X) might be completely irrelevant or hold minimal predictive power regarding the output variable (Y). Since the components are optimized for describing the structure of X rather than predicting Y, PCR can often result in a suboptimal predictive model. This inherent limitation paved the way for more sophisticated, supervised dimensionality reduction techniques.

Introducing Partial Least Squares Regression (PLS)

To address the critical shortcomings of PCR, the technique of **Partial Least Squares** (PLS) regression was developed. PLS is a powerful, supervised statistical method that also functions as a dimensionality reduction technique. Like PCR, PLS calculates M linear combinations, referred to as "PLS components" or "latent variables," from the original set of P predictor variables. However, the mechanism for constructing these components is fundamentally different and intentionally optimized for prediction.

The definitive advantage of PLS lies in its objective function. Unlike PCR, which maximizes variance only within the predictor space (X), Partial Least Squares seeks to identify latent variables that maximize the **covariance** between the predictors (X) and the response variable (Y). By focusing on this shared variation, PLS ensures that the components extracted are those most relevant to the prediction task at hand.

This methodological shift means that PLS components are inherently optimized for predictive performance, as they capture the underlying structure in X that is most strongly correlated with Y. By simultaneously explaining variation in **both** the response variable and the predictor variables, PLS often yields models that are significantly more accurate and robust than those produced by PCR or standard OLS, especially in high-dimensional datasets or scenarios plagued by severe [multicollinearity](#).

Detailed Methodology: The Implementation of PLS

Implementing the **Partial Least Squares** algorithm involves a rigorous, multi-step iterative process designed to systematically extract components that possess maximal predictive information. This process ensures that the latent variables are derived in a manner that maximizes their correlation with the dependent variable.

Data Standardization: The critical initial step involves standardizing all input data. Both the predictor variables (X) and the response variable (Y) must be centered and scaled, typically to possess a mean of 0 and a standard deviation of 1. This necessary transformation mitigates the risk that variables with naturally larger scales or magnitudes might disproportionately influence the

weighting coefficients and the subsequent component calculation.

Iterative Component Calculation: The core of the PLS method is the calculation of $Z_1, \dots, Z_M$, which represent the M PLS components (linear combinations) derived from the original P predictors. The general form of a component is expressed as $Z_m = \sum \Phi_{jm} X_j$, for a set of weighting constants $\Phi_{1m}, \Phi_{2m}, \dots, \Phi_{pm}$, where $m = 1, \dots, M$.

To calculate the first component, **Z_1** , the weighting coefficients (Φ_{j1}) are determined by performing a simple linear regression of the response variable (Y) onto each predictor (X_j). This strategic choice ensures that Z_1 is oriented in the direction that maximizes the covariance between the predictor block (X) and the response variable (Y).

For calculating subsequent components (e.g., **Z_2**), the variation that was already explained by the previous component (Z_1) must be rigorously removed from both the predictor and response variables. This is achieved through deflation: regressing each variable onto Z_1 and utilizing the resulting **residuals** (the orthogonalized data) for the next step. Z_2 is then calculated from this residual data using the exact same procedure employed for calculating Z_1 .

This iterative deflation and calculation process is repeated M times until the full, orthogonal set of M PLS components is obtained.

Final Model Fitting: Once the M PLS components ($Z_1, \dots, Z_M$) are established, the final stage involves applying the method of [Least Squares](#). A linear regression model is fitted using these new, uncorrelated components as the predictive features, resulting in a stable and robust final model.

Optimizing Performance: Selecting the Number of Components (M)

A crucial aspect of successfully deploying **Partial Least Squares** regression is the determination of the appropriate number of components (M) to retain in the final model structure. This decision represents a classic bias-variance trade-off: retaining too few components may result in **underfitting**, failing to capture essential predictive information, while retaining too many components risks capturing noise and leading to model instability and [overfitting](#).

The industry standard and most robust technique for optimizing this parameter (M) is the application of [k-fold cross-validation](#). This statistically sound validation method involves systematically partitioning the training data into K distinct subsets. The model is trained K times, each time leaving out one fold for testing, and the performance is evaluated across a range of potential component numbers ($M=1, M=2, \dots, M=P$).

The "optimal" number of PLS components is consistently identified as the value of M that successfully minimizes the test **Mean-Squared Error (MSE)** across the validation folds. By selecting the model complexity that achieves the lowest prediction error on data it has not been

trained on, we ensure the best possible balance between variance explained and predictive accuracy, leading to superior generalization on truly unseen data.

PLS in Practice and Comprehensive Model Comparison

In real-world data science applications, particularly those characterized by high-dimensional data or strong inter-predictor correlation, **Partial Least Squares** regression consistently demonstrates superior performance and coefficient stability when compared directly against standard ordinary [Least Squares](#) regression. PLS provides an exceptionally effective methodology for seamlessly integrating powerful dimensionality reduction techniques with supervised learning objectives, making it a staple tool in chemometrics and related fields.

However, expert data science practice dictates that relying solely on any single modeling technique is suboptimal. A rigorous and comprehensive approach requires fitting and comparing a diverse portfolio of distinct models to empirically identify the one that generalizes most effectively to future data. Model selection should always be driven by objective performance metrics derived from validation sets.

Practitioners routinely evaluate a broad suite of linear models, including PLS, [Principal Components Regression](#) (PCR), penalized methods such as [Ridge Regression](#) and [Lasso Regression](#), and standard Multiple Linear Regression. By employing robust validation strategies like [k-fold cross-validation](#) across all candidates, we can pinpoint the specific model that consistently yields the lowest test MSE. This comparative approach ensures maximal predictive robustness and trustworthiness across the entire modeling landscape.