Handling multicollinearity with PCR and PLSR

Thale Eliassen Fink, Andreas Raja Goklas Sitorus, Andreas Gudahl Tufte, Muhammad Tsaqif Wismadi, Harold Horsley, Irene Hofmann, Azimil Gani Alam

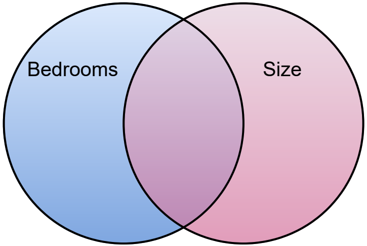
10.09.2024

Outline

1. What is Multicollinearity
2. Why is it a problem
3. How to measure collinearity
4. Principal Component Regression (PCR)
5. Partial Least Squares Regression (PLSR)
6. Discussion/Conclusion

What is Multicollinearity

- When **two or more predictor variables** in a regression model are **highly correlated**
  - When predicting a **house's price** using different predictor variables like the **size** of the house, the **number of bedrooms**, and the **age** of the house.
  - **Size** and **number of bedrooms** are highly related to each other, you have **multicollinearity**.



Source: Penn State

Multicollinearity - 3D plot

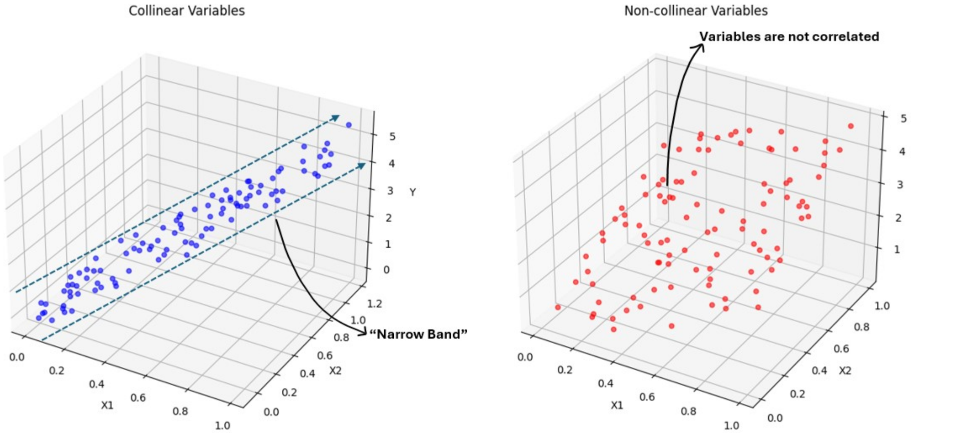


Figure: On the left, predictors that are highly correlated form a "Narrow Band". On the right, data without multicollinearity are dispersed and uncorrelated.

Fit the hyperplane to the data using MLR method

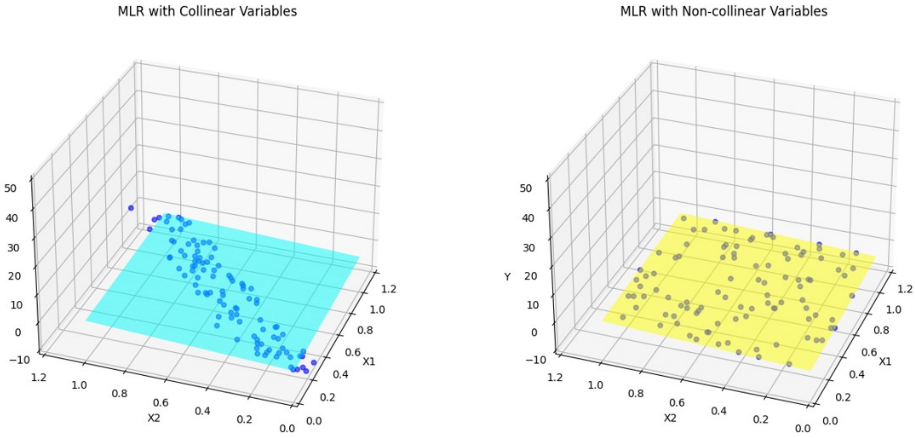


Figure: MLR fitting on both data, everything seems alright ...

Small data adjustment: Unstable hyperplane fitting

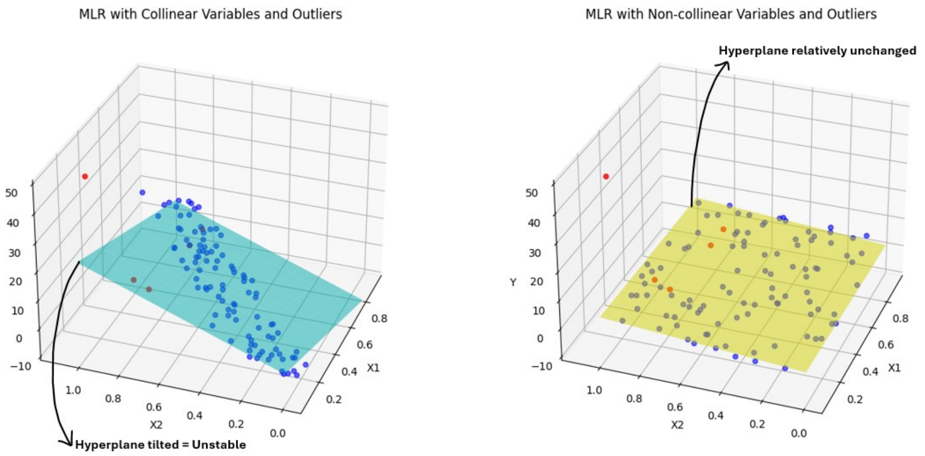
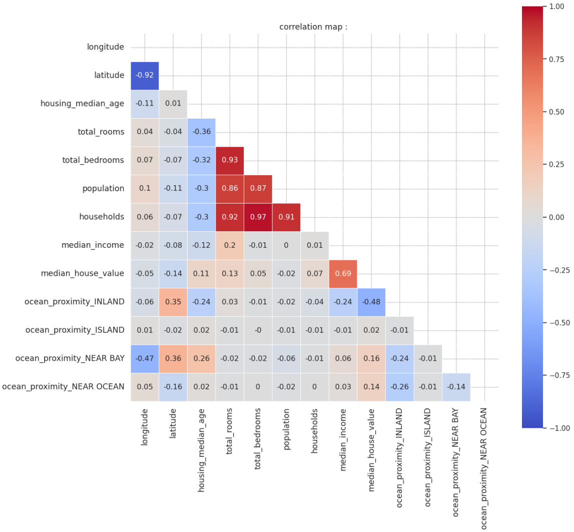


Figure: MLR is quite unstable when implemented to data with multicollinearity.

Why is it a problem

- **Inflated Standard Errors**
  - Makes it difficult to determine which predictors are significant.
- **Unstable Coefficients**
  - Coefficients change drastically with small data adjustments.
  - Coefficient change:
    - Data with multicollinearity, $[X_1, X_2] = [23.493,\ 21.637]$
    - Data without multicollinearity, $[X_1, X_2] = [3.518,\ 2.049]$
- **Complicates Interpretation**
  - Hard to identify each predictor's unique impact.

Python notebook here!
Source: Medium


correlation map :
longitude
latitude -0.92
housing_median_age -0.11 0.01
total_rooms 0.04 -0.04 -0.36
total_bedrooms 0.07 -0.07 -0.32 0.93
population 0.1 -0.11 -0.3 0.86 0.87
households 0.06 -0.07 -0.3 0.92 0.97 0.91
median_income -0.02 -0.08 -0.12 0.2 -0.01 0 0.01
median_house_value -0.05 -0.14 0.11 0.13 0.05 -0.02 0.07 0.69
ocean_proximity_INLAND -0.06 0.35 -0.24 0.03 -0.01 -0.02 -0.04 -0.24 -0.48
ocean_proximity_ISLAND 0.01 -0.02 0.02 -0.01 -0 -0.01 -0.01 -0.01 0.02 -0.01
ocean_proximity_NEAR BAY -0.47 0.36 0.26 -0.02 -0.02 -0.06 -0.01 0.06 0.16 -0.24 -0.01
ocean_proximity_NEAR OCEAN 0.05 -0.16 0.02 -0.01 0 -0.02 0 0.03 0.14 -0.26 -0.01 -0.14
longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value
ocean_proximity_INLAND
ocean_proximity_ISLAND
ocean_proximity_NEAR BAY
ocean_proximity_NEAR OCEAN
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

How to handle it?

Two solutions are:

- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)

Principal Component Regression

- Handles multicollinearity by **combining variables into principal components**
- *Example:* MLR with two variables, v_1 and v_2, can become PCR using one principal component, PC_1
  - PC_1 is some linear combination of v_1 and v_2. $PC_1 = \alpha_1 v_1 + \alpha_2 v_2$

MLR

$\hat{y} = \beta_0 + \beta_1 v_1 + \beta_2 v_2$

PCR

$\hat{y} = \beta_0 + \beta_1 PC_1$

- **Using** Principal Components ensures *non-collinearity* as components are orthogonal
- **Select** the number of Principal Components by using validation methods (next week) e.g. Root mean square error of prediction
- **Substitute** in $PC_1 = \alpha_1 v_1 + \alpha_2 v_2$ obtained from PCA to obtain variable coefficients

Principal Component Regression - Example: Predict student's abilities in untested areas

- **Dataset overview**: 234 students, 36 questions, numerical full rank but high condition number (105.3) indicating multicollinearity
- **Limitations of using total score**: Oversimplifies student abilities. Ignores variance among individual question performances
- **Challenges with individual predictors**: High multicollinearity among questions leads to unstable regression coefficients. Difficult to interpret due to overlapping content areas

Candidate ID	Q1	Q2	...	Q36	Score (0-100)	Prediction in untested area
$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$

$\rightarrow$ PCR might be a better predictor of individual's student abilities in untested areas

Data: Exam scores in TTK4105 spring 2024

Principal Component Regression - Problems

Values, α are chosen to optimise PCA (i.e maximum variance) and NOT to optimise prediction, $\hat{y}$. Therefore:

Problem

Resulting variable parameters ($\beta_1 \alpha_1$ etc) don't necessarily have meaning in the context of our **dependent variable** i.e. y

We need a way to optimise for the dependent variable so that learned parameters have meaning.

Solution

Partial Least Squares Regression

Partial Least Squares Regression

- Another way to ensure non-collinearity and lower dimensionality of predictor variables
- The method is the same as PCR, but in this case latent variable LV_1 is chosen to optimise for the dependent variable: $LV_1 = \alpha_1 v_2 + \alpha_2 v_2$

MLR

$\hat{y} = \beta_0 + \beta_1 v_1 + \beta_2 v_2$

PLSR

$\hat{y} = \beta_0 + \beta_1 LV_1$

- **Choose** different algorithms to find optimal set of latent variable components $\boldsymbol{\alpha}$
- **Continue** as you would with PCR

Partial Least Square Regression - Example: Predicting housing value based on house specification

- **Dataset overview**: 20640 house listed, 13 variables explaining housing specification
- **Multicollinearity indication**: 6 variables with high VIF scores, indicating multicollinearity
- **Objective**: PLSR is applied to improve the predictive accuracy of housing value despite the highly correlated variables
- **Results**: PLSR outperformed both PCR and MLR in terms of R-squared accuracy, achieving values of 0.626, 0.621, and 0.57, respectively.

Long	Lat	Housing age	...	No. of bedrooms	Ocean proximity	**House value**
⋮	⋮			⋮	⋮	⋮

Partial Least Square Regression - Problems

- **Higher Risk of Overlooking 'Real' Correlations**
  - *Example:* PLSR may cause the unique impact of important variables (e.g., house size) to disappear if they are highly correlated with others (e.g., number of bedrooms), as PLSR combines them into a single latent variable.
- **Sensitivity to the relative scaling of the descriptor variables**
  - **Normalization** only addresses scaling issue, but not distribution
  - PLSR can still be **sensitive if variables have different shapes of distribution**
  - *Example:* If one variable is **normally distributed** (e.g., house size) and another is **highly skewed** (e.g., distance to city center), the **skewed variable** can disproportionately influence the model

Source: Partial Least Squares (PLS): Its strengths and limitations

Conclusion/Discussion

- For handling multicollinearity, **PLSR** is generally preferred over **PCR** when your objective is **prediction**, as it directly links the predictor variables to the response while still reducing dimensionality