What is Multicollinearity

06 May 2023 08:23

Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a multiple regression model are highly correlated. In other words, these variables exhibit a strong linear relationship, making it difficult to isolate the individual effects of each variable on the dependent variable.

$\boxed{0.9}$ or $\boxed{0.8}$

$\boxed{100}$

cgpa	iq	lpa
8	80	8

iq ↑ → cgpa ↑

$\boxed{\text{Corr}}$ → linear relation

iq | #backlogs | lpa

iq ↑ backlo ↓ $\boxed{\text{multicoll}}$

cgpa | dob | lpa (3)

$$lpa = \beta_0 + \beta_1 \, cgpa + \beta_2 \, iq$$

? ?

$\beta_0 \; \beta_1 \; \beta_2$ → unreliable

interpret

When is Multicollinearity bad?

06 May 2023 14:33

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

corr

$Y \to X_1$

$Y \to X_2$

1. Inference:
   - Inference focuses on understanding the relationships between the variables in a model.
   - It aims to draw conclusions about the underlying population or process that generated the data.
   - Inference often involves hypothesis testing, confidence intervals, and determining the significance of predictor variables.
   - The primary goal is to provide insights about the structure of the data and the relationships between variables.
   - Interpretability is a key concern when performing inference, as the objective is to understand the underlying mechanisms driving the data.
   - Examples of inferential techniques include linear regression, logistic regression, and ANOVA.
2. Prediction: linear reg → predict → mult
   - Prediction focuses on using a model to make accurate forecasts or estimates for new, unseen data.
   - It aims to generalize the model to new instances, based on the patterns observed in the training data.
   - Prediction often involves minimizing an error metric, such as mean squared error or cross-entropy loss, to assess the accuracy of the model.
   - The primary goal is to create an accurate and reliable model for predicting outcomes, rather than understanding the relationships between variables.
   - Interpretability may be less important in predictive modelling, as the main objective is to create accurate forecasts rather than understanding the underlying structure of the data.
   - Examples of predictive techniques include decision trees, support vector machines, neural networks, and ensemble methods like random forests and gradient boosting machines.

In summary, inference focuses on understanding the relationships between variables and interpreting the underlying structure of the data, while prediction focuses on creating accurate forecasts for new, unseen data based on the patterns observed in the training data.

100 → | cgpa | iq | lpa |

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Finance → banking → model

Customer data → loan

User activity → ban/allow

| X_1 | X_2 | Y |

$$X_1 = a_0 + a_1 X_2 + \eta$$

error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

reducible error

$$Y = \beta_0 + \beta_1 (a_0 + a_1 X_2 + \eta) + \epsilon$$

$$Y = \beta_0 + \beta_1 a_0 + \beta_1 a_1 X_2 + \beta_1 \eta + \epsilon$$

$$Y = (\beta_0 + \beta_1 a_0) + \beta_1 a_1 X_2 + (\beta_1 \eta + \epsilon)$$

$$\boxed{Y \to X_2}$$

X_1

option → $$\boxed{Y = c_0 + c_1 X_2 + \gamma}$$

What exactly happens in Multicollinearity(Mathematically?)

06 May 2023 08:29

When multicollinearity is present in a model, it can lead to several issues, including:

1. Difficulty in identifying the most important predictors: Due to the high correlation between independent variables, it becomes challenging to determine which variable has the most significant impact on the dependent variable.

2. Inflated standard errors: Multicollinearity can lead to larger standard errors for the regression coefficients, which decreases the statistical power and can make it challenging to determine the true relationship between the independent and dependent variables.

3. Unstable and unreliable estimates: The regression coefficients become sensitive to small changes in the data, making it difficult to interpret the results accurately.

$lpa = \beta_0 + \beta_1 cgpa + \beta_2 iq$

$SE(\beta) \rightarrow$ high value

statsmodel $\rightarrow$ summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Sat, 29 Apr 2023   Prob (F-statistic):           1.58e-96
Time:                        07:32:56   Log-Likelihood:                -386.18
No. Observations:                 200   AIC:                             780.4
Df Residuals:                     196   BIC:                             793.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.9389      0.312      9.422      0.000       2.324       3.554
TV             0.0458      0.001     32.809      0.000       0.043       0.049
Radio          0.1885      0.009     21.893      0.000       0.172       0.206
Newspaper     -0.0010      0.006     -0.177      0.860      -0.013       0.011
==============================================================================
Omnibus:                       60.414   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
Skew:                          -1.327   Prob(JB):                     1.44e-33
Kurtosis:                       6.332   Cond. No.                         454.
==============================================================================
```

TV | radio | news | sales

x_1 x_2 x_3 y

$\boxed{SE} \rightarrow$ precision

$\leftarrow$ reg analysis impact

multicollinearity (30) >

multicollinearity
- $\rightarrow$ unstable coefficients
- $\rightarrow$ high SE

$$\boxed{\beta = (X^T X)^{-1} X^T y}$$

cgpa | iq | lpa $\longleftrightarrow$ sample (100)

$\boxed{lpa = \beta_0 + \beta_1 cgpa + \beta_2 iq}$

$Var(\beta) = SE(\beta)$ (100)

$\rightarrow \boxed{\beta_0 \; \beta_1 \; \beta_2}$

$\boxed{10 \text{ datasets}}$

β_0 β_1 β_2

β_0 β_0 β_0

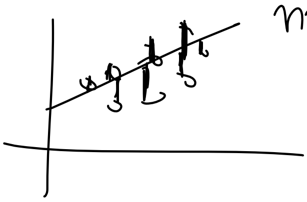
β_0 β_1 β_2

$\boxed{SE(\beta_0) \quad SE(\beta_1) \quad SE(\beta_2)}$

$Var(\beta) = SE$

linear reg

$\sqrt{Var(\beta)} = SE$ linear reg



normally dist $\rightarrow N \sim (0, \sigma^2)$

$\beta \rightarrow 3$

variance cov

$Var(\beta) = \sigma^2 \underbrace{(X^T X)^{-1}}_{3 \times 3}$

$$\sigma^2 \begin{bmatrix} a & - & - \\ - & b & - \\ - & - & c \end{bmatrix}$$

SE $a \rightarrow \sqrt{var(\beta_0)}$

SE $b \rightarrow \sqrt{var(\beta_1)}$

SE $c \rightarrow \sqrt{var(\beta_2)}$

$SE(\beta) = \sqrt{diag(\sigma^2 (X^T X)^{-1})}$

perfect multicollin $\rightarrow det(X^T X) = 0$

$\beta = (X^T X)^{-1} X^T Y$

sensitive to input

strong multi

$det(X^T X)$ very small

$(X^T X)^{-1}$

inve $\rightarrow \frac{1}{det}$

very small

$$\begin{bmatrix} - & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \rightarrow \text{high}$$

inflate

(β) β_0 β_1 β_2

sensitive to data

$(X^T X)^{-1}$ σ^2

cgpa

$8.1 \rightarrow 8.2$

$(\beta) \rightarrow (X^T X)^{-1}$

Perfect Multicollinearity

06 May 2023 08:35

Perfect multicollinearity occurs when one independent variable in a multiple regression model is an exact linear combination of one or more other independent variables. In other words, there is an exact linear relationship between the independent variables, making it impossible to uniquely estimate the individual effects of each variable on the dependent variable.

corr

linear

corr $\quad x_1 = a_1 x_2 + a_0 + \boxed{\text{error}}$

$\boxed{x_1 = a_1 x_2 + a_0}$ ← perfect multicollinearity

cgpa	percent	lpa
8.5	85	7
9.12	91.2	6

percent = $10 \times$ cgpa + 0

10×8.5

$a_1 = 10 \quad a_0 = 0$

$10 \times$ cgpa + 0 + error

cgpa	percent	lpa
8.5	83	
9.12	95	

cgpa	percent	lpa
8	80	3
6	60	4

$\beta \rightarrow$

lpa = $\beta_0 + \beta_1$ cgpa + β_2 percent + error

$\beta_0 \quad \beta_1 \quad \beta_2$

OLS / GD

$$\beta = (X^T X)^{-1} X^T y$$

X ← design matrix

$$X = \begin{bmatrix} 1 & 8 & 80 \\ 1 & 6 & 60 \end{bmatrix}$$

inverse ✗

$[X^T X]$ → singular matrix

$$\underbrace{\underset{3\times 2}{\begin{bmatrix} 1 & 1 \\ 8 & 6 \\ 80 & 60 \end{bmatrix}} \underset{2\times 3}{\begin{bmatrix} 1 & 8 & 80 \\ 1 & 6 & 60 \end{bmatrix}}}_{3\times 3} = \begin{bmatrix} 2 & 14 & 140 \\ 14 & 84 & 8400 \\ 140 & 8400 & 840000 \end{bmatrix}$$

X^T

$\boxed{\text{Det}} \rightarrow 2(0) - 14(0) + 140(0) = \boxed{0} \rightarrow$

Types of Multicollinearity

06 May 2023 08:30

1. **Structural multicollinearity**: Structural multicollinearity arises due to the way in which the variables are defined or the model is constructed. It occurs when one independent variable is created as a linear combination of other independent variables or when the model includes interaction terms or higher-order terms (such as polynomial terms) without proper scaling or centering.

2. **Data-driven multicollinearity**: Data-driven multicollinearity occurs when the independent variables in the dataset are highly correlated due to the specific data being analysed. In this case, the high correlation between the variables is not a result of the way the variables are defined or the model is constructed but rather due to the observed data patterns.

One hot encoding

cat

city

D

M

K

OHE

D	M	K
1	0	0
0	1	0
0	0	1

$(X^T X)^{-1} X$

det = 0

perfect multicollin

sqft | #wal

data

$K = 1 - D - M$

$1 - 1 - 0 = 0$

$1 - 0 - 1 = 0$

$1 - 0 - 0 = 1$

$\boxed{\beta}$ cal X

x | y

$\boxed{x^0 \quad x^1 \quad x^2}$ y

How to Detect Multicollinearity

06 May 2023 08:30

Corr

VIF

Variance inflation factor

Condition No

Correlation

06 May 2023 14:23

Correlation is a measure of the linear relationship between two variables, and it is commonly used to identify multicollinearity in multiple linear regression models. Multicollinearity occurs when two or more predictor variables in the model are highly correlated, making it difficult to determine their individual contributions to the output variable.

To detect multicollinearity using correlation, you can calculate the correlation matrix of the predictor variables. The correlation matrix is a square matrix that shows the pairwise correlations between each pair of predictor variables. The diagonal elements of the matrix are always equal to 1, as they represent the correlation of a variable with itself. The off-diagonal elements represent the correlation between different pairs of variables.

In the context of multicollinearity, you should look for off-diagonal elements with high absolute values (e.g., greater than 0.8 or 0.9, depending on the specific application and the level of concern about multicollinearity). High correlation values indicate that the corresponding predictor variables are highly correlated and may be causing multicollinearity issues in the regression model.

It's important to note that while correlation can be a useful tool for detecting multicollinearity, it doesn't provide a complete picture of the severity of the issue or its impact on the regression model. Other diagnostic measures, such as Variance Inflation Factor (VIF) and condition number, can also be used to assess the presence and severity of multicollinearity in a regression model.

$\underline{X_1}$ $\underline{X_2}$

$\rightarrow \underline{X_1} = \underbrace{a_1 X_2}_{r} + \underline{a_0} + \underline{error}$

$\underline{corr()}$

Variance Inflation Factor

06 May 2023 08:30

Variance Inflation Factor (VIF) is a metric used to quantify the severity of multicollinearity in a multiple linear regression model. It measures the extent to which the variance of an estimated regression coefficient is increased due to multicollinearity.

For each predictor variable in the regression model, VIF is calculated by performing a separate linear regression using that predictor as the response variable and the remaining predictor variables as the independent variables. The VIF for the predictor variable is then calculated as the reciprocal of the variance explained by the other predictors, which is equal to $1 / (1 - R^2)$. Here, R^2 is the coefficient of determination for the linear regression using the predictor variable as the response variable.

The VIF calculation can be summarized in the following steps:

1. For each predictor variable X_i in the regression model, perform a linear regression using X_i as the response variable and the remaining predictor variables as the independent variables.
2. Calculate the R^2 value for each of these linear regressions.
3. Compute the VIF for each predictor variable X_i as $VIF_i = 1 / (1 - R^2_i)$.

A VIF value close to 1 indicates that there is very little multicollinearity for the predictor variable, whereas a high VIF value (e.g., greater than 5 or 10, depending on the context) suggests that multicollinearity may be a problem for the predictor variable, and its estimated coefficient might be less reliable.

Keep in mind that VIF only provides an indication of the presence and severity of multicollinearity and does not directly address the issue. Depending on the VIF values and the goals of the analysis, you might consider using techniques like variable selection, regularization, or dimensionality reduction methods to address multicollinearity.

x_1 y x_2

y ← TV | NP | Radio | Sales → LR R^2

x_1 x_2 y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

3 input

3 LR → R^2 score

$R^2 = 0.1$

$$VIF = \frac{1}{1 - R^2}$$

15 1 bi·

TV | NP → Radio

R^2 → 90

Condition No

06 May 2023 14:23

→ Eigen value
Eigen vectors

In the context of multicollinearity, the condition number is a diagnostic measure used to assess the stability and potential numerical issues in a multiple linear regression model. It provides an indication of the severity of multicollinearity by examining the sensitivity of the linear regression to small changes in the input data.

The condition number is calculated as the ratio of the largest eigenvalue to the smallest eigenvalue of the matrix X^TX, where X is the design matrix of the regression model (each row representing an observation and each column representing a predictor variable). A high condition number suggests that the matrix X^TX is ill-conditioned and can lead to numerical instability when solving the normal equations for the regression coefficients.

In the presence of multicollinearity, the design matrix X has highly correlated columns, which can cause the eigenvalues of X^TX to be very different in magnitude (one or more very large eigenvalues and one or more very small eigenvalues). As a result, the condition number becomes large, indicating that the regression model may be sensitive to small changes in the input data, leading to unstable coefficient estimates.

Typically, a condition number larger than 30 (or sometimes even larger than 10 or 20) is considered a warning sign of potential multicollinearity issues. However, the threshold for the condition number depends on the specific application and the level of concern about multicollinearity.

It's important to note that a high condition number alone is not definitive proof of multicollinearity. It is an indication that multicollinearity might be a problem, and further investigation (e.g., using VIF, correlation matrix, or tolerance values) may be required to confirm the presence and severity of multicollinearity.

$[X^T X]$ → linear tran

condi number

↓

ill conditioning of the matrix

$\det(A) \simeq 0$

> 30

↓

ill condit

↓

multi

How to remove multicollinearity

06 May 2023 08:31

VIF = 3 4 18

X_1 X_2

sqft | # wash

size → large
- med
- small

1. Collect more data: In some cases, multicollinearity might be a result of a limited sample size. Collecting more data, if possible, can help reduce multicollinearity and improve the stability of the model.

2. Remove one of the highly correlated variables: If two or more independent variables are highly correlated, consider removing one of them from the model. This step can help eliminate redundancy in the model and reduce multicollinearity. Choose the variable to remove based on domain knowledge, variable importance, or the one with the highest VIF.

3. Combine correlated variables: If correlated independent variables represent similar information, consider combining them into a single variable. This combination can be done by averaging, summing, or using other mathematical operations, depending on the context and the nature of the variables.

4. Use partial least squares regression (PLS): PLS is a technique that combines features of both principal component analysis and multiple regression. It identifies linear combinations of the predictor variables (called latent variables) that have the highest covariance with the response variable, reducing multicollinearity while retaining most of the predictive power.

PCA → LR

X_1 X_2 → X_1' X_2'