Understanding Multicollinearity: Definition, Examples, and Implications

Authored by

Mohammed looti

November 2, 2025

Understanding Multicollinearity and the Concept of Perfect Correlation

In statistical modeling, particularly within the domain of regression analysis, a critical challenge known as Multicollinearity emerges when two or more predictor variables exhibit a strong correlation with one another. This high interdependency means the variables are not providing unique or independent information to the model, which significantly complicates the reliable estimation of coefficients.

While a high degree of correlation (imperfect multicollinearity) can lead to inflated standard errors and unstable coefficient estimates, the most severe and mathematically debilitating form of this phenomenon is known as **perfect multicollinearity**.

Perfect multicollinearity is defined as the exact linear dependence between at least two predictor variables. This implies that one variable can be derived precisely from the other using an exact linear relationship, such as one being a constant multiple of the other, or one being a simple mathematical transformation of the other. The presence of this perfect dependency renders the model fundamentally unsolvable using standard techniques.

To clearly visualize this foundational issue, consider a straightforward dataset where the variables are linked by a precise, deterministic rule:

y	x_1	x_2
6	2	4
6	2	4
8	2	4
12	3	6
13	4	8
14	5	10
15	5	10
15	7	14
13	9	18
17	19	20

In this illustrative example, observe that the values for predictor variable x2 are exactly double the values of x1 (x2 = 2 * x1). Because this mathematical link is flawless and absolute, the system suffers from **perfect multicollinearity**, meaning x2 offers zero unique information beyond what x1

already provides.

y	x_1	x_2
6	2 *2→	4
6	2 *2→	4
8	2 *2→	4
12	3 *2→	6
13	4	8
14	5	10
15	5	10
15	7	14
13	9	18
17	19	20

The Mathematical Impossibility of OLS Estimation

The central problem caused by **perfect multicollinearity** is its absolute incompatibility with the standard parameter estimation method used in linear models: Ordinary Least Squares (OLS). When a perfect linear dependency exists among the predictor variables, OLS is mathematically incapable of calculating stable or unique estimates for the regression coefficients.

The core objective of any regression is to isolate the independent contribution of each predictor variable on the response variable. If, for instance, x1 and x2 are perfectly correlated, it becomes theoretically impossible to calculate the marginal effect of x1 on the response variable (y) while holding x2 constant. This is because x2 cannot remain constant; it must change in direct, proportional lockstep with x1. They are, effectively, the same variable measured differently.

Technically speaking, **perfect multicollinearity** causes the design matrix (X'X) to be singular, meaning it is non-invertible. Since the OLS formula requires the inversion of this matrix to solve for the coefficients, the calculation breaks down completely, leading to undefined or nonsensical coefficient results for the dependent variable.

A Straightforward Solution: Removing Redundancy

Unlike the more nuanced issues posed by imperfect multicollinearity, resolving the perfect form is exceptionally straightforward and definitive. The most effective and simplest remedy is to identify

the redundant variable--the one that shares the exact linear relationship with another predictor--and eliminate it from the model entirely.

Returning to our initial illustration involving predictors x1 and x2, since x2 is perfectly predictable from x1 (x2 = 2 * x1), we must simply drop x2 as a predictor variable. By removing this redundancy, we ensure the remaining model matrix is full rank and invertible.

y	x_1
6	2
6	2
8	2
12	3
13	4
14	5
15	5
15	7
13	9
17	19

By simplifying the dataset in this manner, we can successfully fit a regression model using only x1 to predict y. This practical approach guarantees that all remaining variables contribute unique explanatory power, allowing the OLS algorithm to produce stable and identifiable coefficients.

Case Study 1: Redundant Scaling and Unit Conversion

A common real-world manifestation of **perfect multicollinearity** arises when researchers include variables that represent the same underlying measurement but use different units or scaling factors. These are inherently redundant measurements.

Consider a scenario where the goal is to predict an animal's weight using two different measurements of its physical height: "height in centimeters" and "height in meters."

The resultant structure of the dataset clearly illustrates the deterministic link between the two predictors, defined strictly by unit conversion:

weight	height (m)	height (cm)
400	1.3	130
460	0.7	70
470	0.6	60
475	1.3	130
490	1.2	120
440	1.5	150
430	1.2	120
490	1.6	160
500	1.1	110
540	1.4	140

The variable "height in centimeters" is precisely equal to "height in meters" multiplied by 100. This perfect, inherent exact linear relationship ensures that using both variables simultaneously results in **perfect multicollinearity**.

If we attempt to fit a multiple linear regression model in software like R using both variables, the program detects the singularity (the non-invertibility of the matrix) and is forced to refuse estimation for one of the coefficients (in this case, 'cm'), indicating the system is underdetermined:

#define data
df <- data.frame(weight=c(400, 460, 470, 475, 490, 440, 430, 490, 500, 540),
m=c(1.3, .7, .6, 1.3, 1.2, 1.5, 1.2, 1.6, 1.1, 1.4),
cm=c(130, 70, 60, 130, 120, 150, 120, 160, 110, 140))

#fit multiple linear regression model
model <- lm(weight~m+cm, data=df)

#view summary of model
summary(model)

Call:
lm(formula = weight ~ m + cm, data = df)

Residuals:
Min 1Q Median 3Q Max
-70.501 -25.501 5.183 19.499 68.590

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 458.676 53.403 8.589 2.61e-05 ***
m 9.096 43.473 0.209 0.839
cm NA NA NA NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.9 on 8 degrees of freedom
Multiple R-squared: 0.005442, Adjusted R-squared: -0.1189
F-statistic: 0.04378 on 1 and 8 DF, p-value: 0.8395

Case Study 2: Inclusion of Transformed Variables

Another classic case involves mistakenly including an original variable along with a mathematically transformed version of itself. While transformations (like standardization or normalization) are often beneficial steps in data preparation, including both the raw and the derived variable simultaneously guarantees **perfect multicollinearity**.

For instance, suppose we are constructing a regression model to predict basketball player ratings. We include the original metric "points" (pts) and a newly calculated variable, "scaled points" (scaled_pts).

The "scaled points" variable is calculated using the standard normalization formula: Scaled points = (points - μpoints) / σpoints. Since the mean (μ) and standard deviation (σ) are fixed constants for the dataset, the scaled variable is merely a linear transformation of the original variable.

The resulting data structure clearly shows this derived relationship, where one predictor is a perfect linear function of the other:

rating	points	scaled points
88	17	-0.884
83	19	-0.574
90	24	0.202
94	29	0.977
96	33	1.598
78	15	-1.194
79	14	-1.349
91	29	0.977
90	25	0.357
82	22	-0.109

Because "scaled points" is perfectly linearly dependent on "points," this setup constitutes **perfect multicollinearity**. Attempting to fit the OLS model confirms the inevitable failure, as the software cannot produce a coefficient estimate for the redundant predictor ("scaled_pts"):

#define data
df <- data.frame(rating=c(88, 83, 90, 94, 96, 78, 79, 91, 90, 82),
pts=c(17, 19, 24, 29, 33, 15, 14, 29, 25, 22))

df$scaled_pts <- (df$pts - mean(df$pts)) / sd(df$pts)

#fit multiple linear regression model
model <- lm(rating~pts+scaled_pts, data=df)

#view summary of model
summary(model)

Call:
lm(formula = rating ~ pts + scaled_pts, data = df)

Residuals:
Min 1Q Median 3Q Max
-4.4932 -1.3941 -0.2935 1.3055 5.8412

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.4218 3.5896 18.783 6.67e-08 ***

pts 0.8669 0.1527 5.678 0.000466 ***
scaled_pts NA NA NA NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.953 on 8 degrees of freedom
Multiple R-squared: 0.8012, Adjusted R-squared: 0.7763
F-statistic: 32.23 on 1 and 8 DF, p-value: 0.0004663

Case Study 3: Avoiding the Dummy Variable Trap

Perhaps the most frequent structural source of **perfect multicollinearity** in applied econometrics and social science modeling is the Dummy Variable Trap. This situation arises when researchers improperly handle categorical variables that are converted into multiple indicator columns for use in a regression model.

Imagine constructing a model where income is predicted by age and marital status:

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

To correctly incorporate a categorical variable like "marital status," it must be transformed into a set of dummy variable columns. The crucial step is selecting one category to serve as the reference or baseline, which is then excluded from the model.

For example, if "Single" is chosen as the baseline category, we create indicators only for "Married" and "Divorced." The status "Single" is then implicitly defined when both the "Married" and "Divorced" indicator variables are set to zero, thus maintaining the independence required for

estimation:

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

→

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

The error leading to the Dummy Variable Trap occurs when the researcher includes an indicator column for every single category (Single, Married, and Divorced):

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

→

Income	Age	Single	Married	Divorced
$45,000	23	1	0	0
$48,000	25	1	0	0
$54,000	24	1	0	0
$57,000	29	1	0	0
$65,000	38	0	1	0
$69,000	36	1	0	0
$78,000	40	0	1	0
$83,000	59	1	0	1
$98,000	56	1	0	1
$104,000	64	0	1	0
$107,000	53	0	1	0

If all three indicator variables are included, the variable "Single" becomes a perfect linear combination of the other two: Single = 1 - (Married + Divorced). This is the exact definition of **perfect multicollinearity** within the context of categorical data.

When attempting to fit this structurally flawed model in R, the software correctly identifies the

mathematical dependency and fails to estimate a coefficient for one of the dummy variable columns (here, 'married'), confirming the singularity:

```
#define data
df <- data.frame(income=c(45, 48, 54, 57, 65, 69, 78, 83, 98, 104, 107),
age=c(23, 25, 24, 29, 38, 36, 40, 59, 56, 64, 53),
single=c(1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0),
married=c(0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1),
divorced=c(0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0))

#fit multiple linear regression model
model <- lm(income~age+single+married+divorced, data=df)

#view summary of model
summary(model)

Call:
lm(formula = income ~ age + single + married + divorced, data = df)

Residuals:
Min 1Q Median 3Q Max
-9.7075 -5.0338 0.0453 3.3904 12.2454

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.7559 17.7811 0.942 0.37739
age 1.4717 0.3544 4.152 0.00428 **
single -2.4797 9.4313 -0.263 0.80018
married NA NA NA NA
divorced -8.3974 12.7714 -0.658 0.53187
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.391 on 7 degrees of freedom
Multiple R-squared: 0.9008, Adjusted R-squared: 0.8584
F-statistic: 21.2 on 3 and 7 DF, p-value: 0.0006865
```

Conclusion: Addressing Redundancy for Model Identifiability

Perfect multicollinearity represents a fatal mathematical flaw for standard OLS estimation methods. Its presence introduces mathematical redundancies that make it fundamentally

impossible for the algorithm to isolate the unique causal effects of the predictor variables, resulting in non-estimable coefficients.

It is crucial to differentiate this absolute problem from high but imperfect multicollinearity. The latter requires complex diagnostic tools and potentially advanced solutions like Ridge Regression. Perfect multicollinearity, however, demands a simple, definitive structural correction: the mandatory removal of the perfectly dependent variable.

By vigilantly ensuring that no predictor variable in the dataset is an exact linear combination of others--whether the dependency stems from unit measurement redundancy, inadvertent mathematical transformation, or the improper creation of a full set of dummy variable indicators--data scientists can guarantee that their models are identifiable and that valid coefficients can be successfully estimated.