What is Perfect Multicollinearity? (Definition & Examples)

Authored by

stats writer

May 6, 2024

Perfect multicollinearity refers to a situation in which two or more independent variables in a statistical model are highly correlated, making it difficult to distinguish their individual effects on the dependent variable. This can occur when one variable is a linear combination of the other(s), resulting in a perfect correlation between them. In other words, there is no unique solution for the regression coefficients in the model, making it impossible to accurately estimate the impact of each variable. An example of perfect multicollinearity may be seen in a model that includes both height and weight as independent variables, as these two variables are highly correlated and one can be calculated from the other. This can lead to unreliable and misleading results in statistical analysis, highlighting the importance of identifying and addressing multicollinearity in research.

What is Perfect Multicollinearity? (Definition & Examples)

In statistics, occurs when two or more predictor variables are highly correlated with each other, such that they do not provide unique or independent information in the regression model.

If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the regression model.

The most extreme case of multicollinearity is known as perfect multicollinearity. This occurs when at least two predictor variables have an exact linear relationship between them.

For example, suppose we have the following dataset:

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

Notice that the values for predictor variable x2 are simply the values of x1 multiplied by 2.

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

This is an example of perfect multicollinearity.

The Problem with Perfect Multicollinearity

When perfect multicollinearity is present in a dataset, the method of ordinary least squares is unable to produce estimates for regression coefficients.

This is because it's not possible to estimate the marginal effect of one predictor variable (x1) on the response variable (y) while holding another predictor variable (x2) constant because x2 always moves exactly when x1 moves.

In short, perfect multicollinearity makes it impossible to estimate a value for every coefficient in a regression model.

How to Handle Perfect Multicollinearity

The simplest way to handle perfect multicollinearity is to drop one of the variables that has an exact linear relationship with another variable.

For example, in our previous dataset we could simply drop x2 as a predictor variable.

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

We would then fit a regression model using x1 as a predictor variable and y as the response variable.

Examples of Perfect Multicollinearity

The following examples show the three most common scenarios of perfect multicollinearity in practice.

1. One Predictor Variable is a Multiple of Another

Suppose we want to use "height in centimeters" and "height in meters" to predict the weight of a certain species of dolphin.

Here's what our dataset might look like:

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

Notice that the value for "height in centimeters" is simply equal to "height in meters" multiplied by 100. This is a case of perfect multicollinearity.

If we attempt to fit a multiple linear regression model in R using this dataset, we won't be able to produce a coefficient estimate for the "meters" predictor variable:

```
#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
```

2. One Predictor Variable is a Transformed Version of Another

Suppose we want to use "points" and "scaled points" to predict the rating of basketball players.

Let's assume that the variable "scaled points" is calculated as:

Scaled points = (points - μpoints) / σpoints

Here's what our dataset might look like:

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

Notice that each value for "scaled points" is simply a standardized version of "points." This is a case of perfect multicollinearity.

If we attempt to fit a multiple linear regression model in R using this dataset, we won't be able to produce a coefficient estimate for the "scaled points" predictor variable:

```
#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

3. The Dummy Variable Trap

Another scenario where perfect multicollinearity can occur is known as the . This is when we want to use a categorical variable in a regression model and convert it into a "dummy variable" that takes on values of 0, 1, 2, etc.

For example, suppose we would like to use predictor variables "age" and "marital status" to predict income:

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 use "marital status" as a predictor variable, we need to first convert it to a dummy variable.

To do so, we can let "Single" be our baseline value since it occurs most often and assign values of 0 or 1 to "Married" and "Divorce" as follows:

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

A mistake would be to create three new dummy variables as follows:

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

In this case, the variable "Single" is a perfect linear combination of the "Married" and "Divorced" variables.

This is an example of perfect multicollinearity.

If we attempt to fit a multiple linear regression model in R using this dataset, we won't be able to produce a coefficient estimate for every predictor variable:

```
#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