What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them?

Authored by

stats writer

June 30, 2024

Complete or quasi-complete separation in logistic/probit regression occurs when the independent variables perfectly or almost perfectly predict the dependent variable. This means that there is no overlap between the values of the independent variables for the different categories of the dependent variable. This can lead to issues in the regression model, such as infinite parameter estimates and unreliable predictions.

To deal with complete or quasi-complete separation, several options can be considered. One approach is to remove the problematic variables from the model. Another option is to use penalized regression techniques, such as Firth's penalized likelihood method, which can handle separation by shrinking the parameter estimates. Additionally, oversampling or undersampling techniques can be used to balance the categories of the dependent variable and reduce the impact of separation.

It is important to address complete or quasi-complete separation in logistic/probit regression in order to avoid biased results and improve the overall accuracy of the model. Careful consideration and appropriate techniques can help mitigate the effects of separation and improve the validity of the regression analysis.

FAQ

What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them?

Occasionally when running a logistic/probit regression we run into the problem of so-called complete separation or quasi-complete separation. On this page, we will discuss what complete or quasi-complete separation is and how to deal with the problem when it occurs.

Notice that the example data set used for this page is extremely small. It is for the purpose of illustration only.

What is complete separation and what do some of the most commonly used software packages do when it happens?

A complete separation happens when the outcome variable separates a predictor variable or a combination of predictor variables completely. Albert and Anderson (1984) define this as, "there is a vector α that correctly allocates all observations to their group." Below is a small example.

Y X1 X2
0 1 3
0 2 2
0 3 -1
0 3 -1
1 5 2
1 6 4
1 10 1
1 11 0

In this example, Y is the outcome variable, X1 and X2 are predictor variables. We can see that observations with Y = 0 all have values of X1<=3 and observations with Y = 1 all have values of X1>3. In other words, Y separates X1 perfectly. The other way to see it is that X1 predicts Y perfectly since X1<=3 corresponds to Y = 0 and X1 > 3 corresponds to Y = 1. By chance, we have found a perfect predictor X1 for the outcome variable Y. In terms of predicted probabilities, we have Prob(Y = 1 | X1<=3) = 0 and Prob(Y=1 X1>3) = 1, without the need for estimating a model.

Complete separation or perfect prediction can occur for several reasons. One common example is when using several categorical variables whose categories are coded by indicators. For example, if one is studying an age-related disease (present/absent) and age is one of the predictors, there may be subgroups (e.g., women over

55) all of whom have the disease.
Complete separation also may occur if there is a coding error or you mistakingly included
another version of the outcome as a predictor. For example, we might have dichotomized a
continuous variable X into a binary variable Y. We then wanted to study the
relationship between Y and some predictor variables. If we would include X as a
predictor variable, we would run into the problem of perfect prediction, since
by definition, Y separates X completely. . The other possible scenario for
complete separation to happen is when the sample size is very small. In our
example data above, there is no reason for why Y has to be 0 when X1 is
<=3. If the sample were large enough, we would probably have some observations
with Y = 1 and X1 <=3, breaking up the complete separation of X1.

What happens when we try to fit a logistic
or a probit regression model of Y on X1 and X2?

Mathematically the maximum likelihood estimate for X1 does not exist. In particular with this example, the larger the coefficient for X1, the larger the likelihood. In other words, the coefficient for X1 should be as large as it can be, which would be infinity!

In terms of the behavior of statistical software packages, below is what SAS (version 9.2), SPSS (version 18), Stata (version 11) and R (version 2.11.1) do when we run the model on the sample data. We present these results here in the hope that some level of understanding of the behavior of logistic/probit regression when using our familiar software package might help us identify the problem of complete separation more efficiently.

SAS

```
data t;
input Y X1 X2;
cards;
0 1 3
```

0 2 2

0 3 -1

0 3 -1

1 5 2

1 6 4

1 10 1

1 11 0

;

run;

proc logistic data = t descending;

model y = x1 x2;

run;

(some output omitted)

Model Convergence Status

Complete separation of data points detected.WARNING: The maximum likelihood estimate does not exist. WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Model Fit Statistics

Intercept

Intercept and

Criterion Only Covariates

AIC 13.090 6.005

SC 13.170 6.244

-2 Log L 11.090 0.005

WARNING: The validity of the model fit is questionable.

Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq

Likelihood Ratio 11.0850 2 0.0039

Score 6.8932 2 0.0319

Wald 0.1302 2 0.9370

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 -20.7083 73.7757 0.0788 0.7789

X1 1 4.4921 12.7425 0.1243 0.7244

X2 1 2.3960 27.9875 0.0073 0.9318

We can see that the first related message is that SAS detected complete
separation of data points, it gives further warning messages indicating
that the maximum likelihood estimate does not exist and continues to finish the
computation. Also notice that SAS does not tell us which variable is or which
variables are being separated completely by the outcome variable and the
parameter estimate for X1 is incorrect.

SPSS

```
data list list
/Y X1 X2.
begin data.
0 1 3
0 2 2
0 3 -1
0 3 -1
1 5 2
1 6 4
1 10 1
1 11 0
```

```
end data.
logistic regression variable Y
/method = enter X1 X2.

Logistic Regression

Warnings
|-----------------------------------------------------------------------
---------|
|The parameter covariance matrix cannot be computed.
Remaining statistics will be omitted.|
|-----------------------------------------------------------------------
---------|

(some output omitted)

Block 1: Method = Enter

Model Summary
|----|----------------|------------------|------------------|
|Step|-2 Log likelihood|Cox & Snell R
Square|Nagelkerke R Square|
|----|----------------|------------------|------------------|
|1 |.000a |.750 |1.000 |
|----|----------------|------------------|------------------|
```

a. Estimation terminated at iteration number 20 because a perfect fit is detected. This solution is not unique.

We see that SPSS detects a perfect fit and immediately stops the rest of the computation. It does not provide any parameter estimates. Neither does it provide us with any further information on the set of variables that gives the perfect fit.

Stata

```
clear
input Y X1 X2
0 1 3
0 2 2
0 3 -1
0 3 -1
1 5 2
1 6 4
1 10 1
1 11 0
end
logit Y X1 X2
```

outcome = X1 > 3 predicts data perfectly
r(2000);

We see that Stata detects the perfect prediction by X1 and stops computation
immediately.

R

```
y<- c(0,0,0,0,1,1,1,1)
x1<-c(1,2,3,3,5,6,10,11)
x2<-c(3,2,-1,-1,2,4,1,0)
m1<- glm(y~ x1+x2, family=binomial)

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(m1)
Call:
glm(formula = y ~ x1 + x2, family = binomial)

Deviance Residuals:
1 2 3 4 5 6 7
-2.107e-08 -1.404e-05 -2.522e-06 -2.522e-06 1.564e-05
2.107e-08 2.107e-08
```

8

2.107e-08

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -66.098 183471.722 -3.60e-04 1

x1 15.288 27362.843 0.001 1

x2 6.241 81543.720 7.65e-05 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.1090e+01 on 7 degrees of freedom

Residual deviance: 4.5454e-10 on 5 degrees of freedom

AIC: 6

Number of Fisher Scoring iterations: 24

The only warning message that R gives is right after fitting the logistic model. It

says that "fitted probabilities numerically 0 or 1 occurred".

Combining this piece of information with the parameter estimate for x1 being

really large (>15), we suspect that there is a problem of complete or quasi-complete separation. The standard errors
for the parameter estimates are way too large. This usually indicates a convergence issue or some degree of data separation.

What is quasi-complete separation and what do some of the most commonly used software packages do when it happens?

Quasi-complete separation in a logistic/probit regression happens when the outcome variable separates a predictor variable or a combination of predictor variables to certain degree. Here is an example.

Y X1 X2
0 1 3
0 2 0
0 3 -1
0 3 4
1 3 1
1 4 0
1 5 2

1 6 7

1 10 3

1 11 4

Notice that the outcome variable Y separates the predictor variable X1 pretty
well except for values of X1 equal to 3. In other words, X1 predicts Y perfectly
when X1 <3 (Y = 0) or X1 >3 (Y=1), leaving only when X1 = 3 as cases with
uncertainty. In terms of expected probabilities, we have Prob(Y=1 |
X1<3) = 0 and Prob(Y=1 | X1>3) = 1, nothing to be estimated, except for Prob(Y =
1 | X1 = 3).

What happens when we try to fit a logistic or a probit regression model of Y on X1 and X2
using the data above? It turns out that the maximum likelihood estimate for X1
does not exist. With this example, the larger the parameter for X1, the larger
the likelihood. In practice,
a value of 15 or larger does not make much difference

and they all basically
correspond to predicted probability of 1. The behavior of different statistical
software packages differ at how they deal with the issue of quasi-complete
separation. Below is what each package of SAS, SPSS, Stata
and R does with our sample data and the logistic regression model of Y on X1 and
X2. We present these results here in the
hope that some level of understanding of the behavior of logistic/probit regression
within our familiar software package might help us identify
the problem of separation more efficiently.

SAS

```
data t2;
input Y X1 X2;
cards;
0 1 3
0 2 0
0 3 -1
0 3 4
```

1 3 1

1 4 0

1 5 2

1 6 7

1 10 3

1 11 4

;

run;

proc logistic data = t2 descending;

model y = x1 x2;

run;

(some output omitted)

Response Profile

Ordered Total

Value Y Frequency

1 1 6

2 0 4

Probability modeled is Y=1.

Model Convergence Status
Quasi-complete separation of data points detected.

WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.
Model Fit Statistics

Intercept
Intercept and
Criterion Only Covariates

AIC 15.460 9.784
SC 15.763 10.691
-2 Log L 13.460 3.784

WARNING: The validity of the model fit is questionable.
Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq

Likelihood Ratio 9.6767 2 0.0079
Score 4.3528 2 0.1134

Wald 0.1464 2 0.9294

Analysis of Maximum Likelihood Estimates

Standard Wald
Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 -21.4542 64.5674 0.1104 0.7397
X1 1 6.9705 21.5019 0.1051 0.7458
X2 1 -0.1206 0.6096 0.0392 0.8431

We see that SAS used all 10 observations and it gave warnings at various
points. It informed us that it detected quasi-complete separation of the data
points. It is worth noticing that neither the parameter estimate for X1 or for
the intercept mean much at all.

Stata

```
clear
input y x1 x2
0 1 3
0 2 0
0 3 -1
```

0 3 4

1 3 1

1 4 0

1 5 2

1 6 7

1 10 3

1 11 4

end

logit y x1 x2

note: outcome = x1 > 3 predicts data perfectly except for

x1 == 3 subsample:

x1 dropped and 7 obs not used

Iteration 0: log likelihood = -1.9095425

Iteration 1: log likelihood = -1.8896311

Iteration 2: log likelihood = -1.8895913

Iteration 3: log likelihood = -1.8895913

Logistic regression Number of obs = 3

LR chi2(1) = 0.04

Prob > chi2 = 0.8417

Log likelihood = -1.8895913 Pseudo R2 = 0.0104

---

y | Coef. Std. Err. z P>|z|

----------+----------

x1 | (omitted)

x2 | -.1206257 .6098361 -0.20 0.843 -1.315883 1.074631

_cons | -.5427435 1.421095 -0.38 0.703 -3.328038 2.242551

Stata detected that there was a quasi-separation and informed us which
predict variable was part of the issue. It tells us that predictor variable x1
predicts the data perfectly except when x1 = 3. It therefore drops all the cases
when x1 predicts the outcome variable perfectly, keeping only the three
observations when x1 = 3. Since x1 is a constant (=3) on this small sample, it
is also dropped out of the analysis.

SPSS

```
data list list
/y x1 x2.
begin data.
```

```
0 1 3
0 2 0
0 3 -1
0 3 4
1 3 1
1 4 0
1 5 2
1 6 7
1 10 3
1 11 4
end data.
logistic regression variable y
/method = enter x1 x2.

(Some output omitted)

Block 1: Method = Enter

Model Summary
|----|----------------|------------------|------------------|
|Step|-2 Log likelihood|Cox & Snell R Square|Nagelkerke R Square|
|----|----------------|------------------|------------------|
|1 |3.779a |.620 |.838 |
```

|----|----------------|-------------------|------------------|

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table(a)

|------|----------------------|-------------------------------|

| |Observed |Predicted |

| |----|-------------|-----------------|

| |y |Percentage Correct|

| | |---------|----| |

| |.00 |1.00| |

|------|----------------|----|--------|---|-----------------|

|Step 1|y |.00 |4 |0 |100.0 |

| | |----|---------|----|-----------------|

| | |1.00|1 |5 |83.3 |

| |----------------|---|---------|----|-----------------|

| |Overall Percentage | | |90.0 |

|------|----------------------|---------|----|-----------------|

a. The cut value is .500

Variables in the Equation

|---------------|-------|---------|----|--|----|-------|

| |B |S.E. |Wald|df|Sig.|Exp(B) |

|-------|--------|-------|---------|----|--|----|-------|

|Step 1a|x1 |17.923 |5140.147 |.000|1 |.997|6.082E7|
| |--------|-------|---------|----|--|----|-------|
| |x2 |-.121 |.610 |.039|1 |.843|.886 |
| |--------|-------|---------|----|--|----|-------|
| |Constant|-54.313|15420.442|.000|1 |.997|.000 |
|-------|--------|-------|---------|----|--|----|-------|
a. Variable(s) entered on step 1: x1, x2.

SPSS tried to iterate to the default number of iterations and couldn't
reach a solution and thus stopped the iteration process. It didn't tell us
anything about quasi-complete separation. So it is up to us to figure out why
the computation didn't converge. One obvious evidence in this example is the
large magnitude of the
parameter estimate for x1. It is really large and its standard error is even
larger. Based on this piece of evidence, we should look at the relationship
between the outcome variable y and x1. For instance, we can take a look at

the cross tabulation of x1 by y as follows.

```
crosstabs
/tables = x1 by y.
```

x1 * y Crosstabulation

Count

	y	
	.00	1.00	Total
x1	1.00	1	0	1
	2.00	1	0	1
	3.00	2	1	3
	4.00	0	1	1
	5.00	0	1	1
	6.00	0	1	1
	10.00	0	1	1
	11.00	0	1	1
Total		4	6	10

The visual inspection reveals that there is a problem of quasi-complete separation involving x1. In practice, this process of identifying the issue could be very lengthy since there may

be multiple predictor variables involved.

R

```
y<- c(0,0,0,0,1,1,1,1,1,1)
x1<-c(1,2,3,3,3,4,5,6,10,11)
x2<-c(3,0,-1,4,1,0,2,7,3,4)
m1<- glm(y~ x1+x2, family=binomial)

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
summary(m1)

(some output omitted)
Call:
glm(formula = y ~ x1 + x2, family = binomial)

Deviance Residuals:
Min 1Q Median 3Q Max
-1.004e+00 -5.538e-05 2.107e-08 2.107e-08 1.469e+00

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -58.0761 17511.9030 -0.003 0.997
x1 19.1778 5837.3009 0.003 0.997
```

x2 -0.1206 0.6098 -0.198 0.843

The only warning we get from R is right after the glm command about
predicted probabilities being 0 or 1. From the parameter estimates we can see
that the coefficient for x1 is very large and its standard error is even larger,
an indication that the model might have some issues with x1. Based on this piece
of evidence, we should look at the relationship between the outcome variable y
and x1 descriptively as shown below. Visual inspection tells us that there is a problem with
quasi-complete separation involving variable x1.

table(x1, y)
y
x1 0 1
1 1 0
2 1 0
3 2 1
4 0 1
5 0 1

6 0 1

10 0 1

11 0 1

What are the techniques for dealing with complete separation or quasi-complete separation?

Now we have some understanding of what complete or quasi-complete separation is, an immediate question is what the techniques are for dealing with the issue. We will give a general and brief description about a few techniques for dealing with the issue with illustration sample code in SAS. Note that these techniques may be available in other packages, for example, Stata's user written firthlogit command. Let's say that the predictor variable involved in complete quasi-complete separation is called X.

References

Thanks to Maureen Lahiff for suggestions to improve this page.