How can I check for collinearity in survey regression?

Authored by

stats writer

July 1, 2024

Collinearity refers to a phenomenon where two or more independent variables in a regression model are highly correlated, making it difficult to determine the individual effects of each variable on the dependent variable. In survey regression, collinearity can occur due to questionnaire design or sampling methods, and it can significantly impact the accuracy and validity of the results.

To check for collinearity in survey regression, there are several steps that can be taken. The first step is to visually examine the correlation matrix of the independent variables. If there are strong positive or negative correlations between variables, it may indicate collinearity. Another approach is to calculate the variance inflation factor (VIF) for each independent variable. A high VIF, typically greater than 10, suggests collinearity.

In addition, conducting a regression analysis with a reduced set of variables, known as a stepwise regression, can also help identify collinearity. This method involves removing variables with high VIF values and re-running the regression until all remaining variables have a VIF below the threshold.

Lastly, conducting a principal component analysis (PCA) can also help detect collinearity. This method transforms the original variables into a smaller set of uncorrelated components, making it easier to identify collinearity.

Overall, checking for collinearity in survey regression is crucial to ensure the validity and reliability of the results. By following these steps, researchers can identify and address any collinearity issues in their regression model, leading to more accurate and meaningful conclusions.

How can I check for collinearity in survey regression? | Stata FAQ

Collinearity is a property of predictor variables and in OLS regression can easily be checked using the estat vif command after regress or by the user-written command, collin (see How can I use the search command to search for programs and get additional help? for more information about using search). The situation is a little bit trickier when using survey data.

We will illustrate this situation using the hsb2 dataset pretending that the variable math is the sampling weight (pweight) and that the sample is stratified on ses. Let's begin by running a survey regression with socst regressed on read, write and the interaction of read and write. We will create the interaction term, rw, by multiplying read and write together. Since rw is the product of two other predictors, it should create a situation with a high degree of collinearity.

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

generate rw = read*write /* create interaction of read and write */

svyset , strata(ses)

pweight: math
VCE: linearized
Single unit: missing
Strata 1: ses
SU 1:
FPC 1:
```

svy: regress socst read write rw
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10529
Design df = 197
F(3, 195) = 72.93
Prob > F = 0.0000
R-squared = 0.4816

--
| Linearized
socst | Coef. Std. Err. t P>|t|
-------------+--
read | .3061248 .3225945 0.95 0.344 -.330057 .9423066
write | .2870839 .3037079 0.95 0.346 -.3118521 .8860198
rw | .0023047 .0057398 0.40 0.688 -.0090146 .013624
_cons | 14.87601 16.18965 0.92 0.359 -17.05125 46.80327
--

Now, how can we tell if there is high collinearity among the three predictors? To answer this we will run three survey regressions using read, write and rw as the

response variables. After each regression we will manually compute the tolerance using the formula 1-R2 and the variance inflation factor (VIF) by 1/tolerance.

svy: regress read write rw

(running regress on estimation sample)

Survey: Linear regression

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10529
Design df = 197
F(2, 196) = 2609.52
Prob > F = 0.0000
R-squared = 0.9783

| Linearized
read | Coef. Std. Err. t P>|t|
---+---

write | -.8465658 .0264581 -32.00 0.000 -.8987433 -.7943884
rw | .017455 .0002706 64.49 0.000 .0169213 .0179888
_cons | 47.79902 .9873955 48.41 0.000 45.8518 49.74624

---

display "tolerance = " 1-e(r2) " VIF = " 1/(1-e(r2))

tolerance = .02165204 VIF = 46.185024

svy: regress write read rw
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10529
Design df = 197
F(2, 196) = 1811.99
Prob > F = 0.0000
R-squared = 0.9678

| Linearized
write | Coef. Std. Err. t P>|t|
---+---
read | -1.029195 .0304864 -33.76 0.000 -1.089316 -.9690733
rw | .019082 .0003452 55.28 0.000 .0184013 .0197628
_cons | 52.84414 .957194 55.21 0.000 50.95648 54.7318

---

display "tolerance = " 1-e(r2) " VIF = " 1/(1-e(r2))

tolerance = .03216943 VIF = 31.085416

svy: regress rw write read
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 3 Number of obs = 200
Number of PSUs = 200 Population size = 10529
Design df = 197
F(2, 196) = 5258.91
Prob > F = 0.0000
R-squared = 0.9916

| Linearized
rw | Coef. Std. Err. t P>|t|
---+---
write | 49.77855 .948907 52.46 0.000 47.90723 51.64987
read | 55.3573 .9117403 60.72 0.000 53.55928 57.15533
_cons | -2703.949 55.95981 -48.32 0.000 -2814.306 -2593.591

---

display "tolerance = " 1-e(r2) " VIF = " 1/(1-e(r2))

tolerance = .00843133 VIF = 118.60521

Note that we used each of the predictor variables, in turn, as the response variable for a survey regression. VIF values greater than 10 may warrant further examination.

In this example, all of the VIFs were problematic but the variable rw stands out with a VIF of 118.61. The high collinearity of the interaction term is not unexpected and probably is not going to cause a problem for our analysis.

This same approach can be used with survey logit (i.e., svy: logit) or any of the survey estimation procedures. To do this, replace the logit command with the regress command and then proceed as shown above. Running the regress command with a binary outcome variable will not be problem
because collinearity is a property of the predictors, not of the model.