“Why don’t my ANOVA and regression results agree?”

Authored by

stats writer

July 1, 2024

ANOVA (Analysis of Variance) and regression are both statistical methods used to analyze relationships between variables and determine the significance of these relationships. However, it is possible for the results obtained from these two methods to differ. This can occur due to several reasons such as the type of data being analyzed, the assumptions made by each method, and the underlying nature of the relationship between the variables. Additionally, differences in the sample size, the level of measurement of the variables, and the presence of outliers can also affect the results. Therefore, it is important to carefully consider the assumptions and limitations of each method and to interpret the results in the context of the specific research question.

Why don't my anova and regression results agree? | Stata FAQ

We recently received a question asking why the results from the same model specified as anova versus a regression would not agree. The model in question had both categorical and continuous predictors. This question is really just a variation of questions concerning dummy (zero/one) coding versus effect coding. There are several FAQs that address this issue: How can get anova main-effects with dummy coding?, How can I get anova simple main effects with dummy coding?, How can I understand a three-way interaction in anova? and others.

Here is an example that is similar to the question asked by our client. It involves a model that

has a categorical by continuous interaction.

```
use https://stats.idre.ucla.edu/stat/data/hsbdemo, clear
```

anova write c.socst##i.female

Number of obs = 200 R-squared = 0.4299

Root MSE = 7.21161 Adj R-squared = 0.4211

Source | Partial SS df MS F Prob > F

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

Model | 7685.43528 3 2561.81176 49.26 0.0000

|

socst | 6242.19751 1 6242.19751 120.03 0.0000

female | 450.252986 1 450.252986 8.66 0.0036

female#socst | 239.648735 1 239.648735 4.61 0.0331

|

Residual | 10193.4397 196 52.0073455

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

Total | 17878.875 199 89.843593

```
regress write c.socst##i.female
```

Source | SS df MS Number of obs = 200

-------------+------------------------------ F(3, 196) = 49.26

Model | 7685.43528 3 2561.81176 Prob > F = 0.0000

Residual | 10193.4397 196 52.0073455 R-squared = 0.4299

-------------+------------------------------ Adj R-squared = 0.4211

Total | 17878.875 199 89.843593 Root MSE = 7.2116

--

write | Coef. Std. Err. t P>|t|

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

socst | .6247968 .0670709 9.32 0.000 .4925236 .7570701

1.female | 15.00001 5.09795 2.94 0.004 4.946132 25.05389

|

female#|

c.socst |

1 | -.2047288 .0953726 -2.15 0.033 -.3928171 -.0166405

|

_cons | 17.7619 3.554993 5.00 0.000 10.75095 24.77284

--

`test socst`

(1) socst = 0

F(1, 196) = 86.78

Prob > F = 0.0000

**As you can see the F-ratio for `socst` in `anova` is 120.03 and in `regress`
86.78. They are very different. What is going on here?**

**The answer is, of course, that the anova and the regression F-ratios are testing two different things. The `anova` F-ratio is computed from the partial sum of squares for `socst` with all
of the other effects partialed out. The sum of squares is divided by its degrees of freedom (one) and
is in turn divided by the mean square residual (the pooled within cell variance). Although the
`anova` F-ratio is significant, you wouldn't want to spend much effort trying to interpret it
since `socst` is also part of the significant `socst#female` interaction.**

**This particular regression model has a categorical variable, `female`, that is dummy coded (zero/one)
using the built_in factor variables notation.
The F-ratio in the regression is testing the slope of `write` on `socst` for the reference
group, in this case `female` = 0 (males). In fact, the regression coefficient (.6247968) is**

the slope of `write` on `socst` for the males.

So, how can you get the `anova` F-ratio from the `regress` model. We will demonstrate
three ways of doing this.

Method 1: using the test command:

`quietly regress write c.socst##i.female` /* rerun regression model */

`test c.socst + 1.female#c.socst/2 = 0` /* divide by 2 because there are two levels of female */

(1) socst + .5*1.female#c.socst = 0

F(1, 196) = 120.03
Prob > F = 0.0000

This method shows that the "main" effect for `socst` is made of of the effect for `socst`
plus the average of the interaction effect over the two levels of `female`.

Method 2: using the margins command:

```
margins, dydx(socst) asbalanced post
```

Average marginal effects Number of obs = 200

Model VCE : OLS

Expression : Linear prediction, predict()

dy/dx w.r.t. : socst

at : female (asbalanced)

--

| Delta-method

| dy/dx Std. Err. z P>|z|

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

socst | .5224324 .0476863 10.96 0.000 .428969 .6158959

--

`test socst`

(1) socst = 0

chi2(1) = 120.03

Prob > chi2 = 0.0000

For the `margins` command we need to use both the `post` and `asbalanced` options.

The `post` option allows us to use the `test` command after `margins` and the

`asbalanced` is needed both because the categorical variable

(`female`) have unequal cell size and also because we have a continuous predictor in the model.

Method 3: using a sum-to-zero coding:

You indicate categorical variables for `regress` using the `i.` prefix. This indicates that Stata should use factor variables. Stata uses dummy (zero-one) coding for its factor variables. The use of dummy coding is the reason that the `anova` and `regress` results are different. If you were to use a sum-to-zero coding then the results would be the same. We will demonstrate this using effect coding in which the reference group is coded as minus one (-1). Technically, this coding scheme does not actually sum to zero in an unbalanced design but it still works the way we want it to.

```
recode female (0 = -1), gen(fem)
```
/* effect coding for female */

```
regress write c.socst##c.fem
```

Source | SS df MS Number of obs = 200

-------------+------------------------------ F(3, 196) = 49.26

Model | 7685.43528 3 2561.81176 Prob > F = 0.0000

Residual | 10193.4397 196 52.0073455 R-squared = 0.4299

-------------+------------------------------ Adj R-squared = 0.4211

Total | 17878.875 199 89.843593 Root MSE = 7.2116

--

write | Coef. Std. Err. t P>|t|

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

socst | .5224324 .0476863 10.96 0.000 .4283883 .6164766

fem | 7.500004 2.548975 2.94 0.004 2.473066 12.52694

|

c.socst#|

c.fem | -.1023644 .0476863 -2.15 0.033 -.1964085 -.0083203

|

_cons | 25.2619 2.548975 9.91 0.000 20.23496 30.28884

--

`test c.socst`

(1) socst = 0

F(1, 196) = 120.03

Prob > F = 0.0000

For the sake of completeness, we need to mention that if there is no interaction then the `anova` and `regress` results agree perfectly, as shown below.

```
anova write c.socst i.female
```

Number of obs = 200 R-squared = 0.4165
Root MSE = 7.27735 Adj R-squared = 0.4105

Source | Partial SS df MS F Prob > F
-----------+--
Model | 7445.78654 2 3722.89327 70.30 0.0000
|
socst | 6269.5727 1 6269.5727 118.38 0.0000
female | 906.143844 1 906.143844 17.11 0.0001
|
Residual | 10433.0885 197 52.9598399
-----------+--
Total | 17878.875 199 89.843593

```
regress write c.socst i.female
```

Source | SS df MS Number of obs = 200
-------------+------------------------------ F(2, 197) = 70.30

Model | 7445.78654 2 3722.89327 Prob > F = 0.0000

Residual | 10433.0885 197 52.9598399 R-squared = 0.4165

-------------+------------------------------ Adj R-squared = 0.4105

Total | 17878.875 199 89.843593 Root MSE = 7.2774

--

write | Coef. Std. Err. t P>|t|

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

socst | .5235458 .0481182 10.88 0.000 .428653 .6184386

1.female | 4.280318 1.034786 4.14 0.000 2.239637 6.320998

_cons | 23.00581 2.606248 8.83 0.000 17.86608 28.14554

--

`test socst`

(1) socst = 0

F(1, 197) = 118.38

Prob > F = 0.0000