Why ANOVA and linear regression are the same

Christian Peters, Victor van Pelt
Feb 26, 2021

If your doctoral training was anything like ours, you have learned to conduct ANOVAs for one class of experiments (i.e., behavioral experiments), while running linear regressions for another class of experiments (i.e., economic experiments). However, there was not a single class that put it all together. When asked, people may advocate the use of one analysis over the other, but give vague reasons when pressed. It was not until we started comparing ANOVA and linear regression ourselves that we realized how closely-related they are.

In this post, we use a simple example to show you the equivalence between ANOVA and a linear regression. Our example features a two-way interaction effect between two binary categorical variables. Therefore, it should be representative for the vast majority of accounting experiments.

When comparing ANOVA and linear regression, it is essential to distinguish two types of of coding: Effect coding and dummy coding. The first type of coding contrasts group means with the grand mean, whereas the last type of coding contrasts group means with a specified reference group. If you have multiple categorical variables in your model, often the choice between effect coding and

dummy coding for your interpretation does not matter. However, when you have an interaction with two categorical variables, the main effects either represent “true” main effects (for effect coding) or simple effects (for dummy coding).

We use publicly available data from the Stata 16 manual (i.e., fitness.dta). We do not aim to use this dataset to test some hypothesis. Instead, our goal is to estimate a basic empirical model predicting the continuous dependent variable “hours,” which are the number of hours an individual exercises per day, using two binary categorical variables (i.e., “smoke” and “single”) and their interaction.

We first load the data in Stata 16 and tabulate the means to get an impression of how “hours” varies across the two categorical variables “smoke” and “single.”

```
. webuse fitness, clear

. table smoke single, c(mean hours) row col

----------------------------------------
dummy for  |    dummy for not married
smoking    |  married    single     Total
-----------+----------------------------
nonsmoking | .6644013  1.075681  .9116104
   smoking | .4489347  .6688333  .5891583
     Total | .6450533  1.033642  .8800172
----------------------------------------
```

The output displays the means for each category, sub-category, and even sub-sub-category. Also, the output reports a grand mean of `0.8800172`.

We run a two-way ANOVA immediately followed by its underlying linear regression model. The ANOVA automatically uses effect coding for the categorical variables “smoke” and “single.” This means that the categories are coded with 1’s and -1’s so that each category’s mean is compared to the grand mean of the sample. In the linear regression, the categorical variables are dummy coded, which means that each category’s mean is compared to the reference group’s mean. In this case, the reference group is the group that does not smoke and that are not single.

```
. anova hours smoke##single

                          Number of obs =     19,831    R-squared     =  0.0424
                          Root MSE      =    1.02876    Adj R-squared =  0.0423

                 Source | Partial SS         df         MS        F    Prob>F
             -----------+----------------------------------------------------
                  Model |   929.4697          3   309.82323    292.74  0.0000
                        |
                  smoke |  157.38411          1   157.38411    148.71  0.0000
                 single |  161.89951          1   161.89951    152.97  0.0000
          smoke#single |  14.884657          1   14.884657     14.06  0.0002
                        |
               Residual |   20983.97     19,827   1.0583533
             -----------+----------------------------------------------------
                  Total |   21913.44     19,830    1.105065

. regress

      Source |       SS           df       MS      Number of obs   =    19,831
-------------+----------------------------------   F(3, 19827)     =    292.74
       Model |    929.4697         3  309.823233   Prob > F        =    0.0000
    Residual |  20983.9702    19,827  1.05835327   R-squared       =    0.0424
-------------+----------------------------------   Adj R-squared   =    0.0423
       Total |  21913.4399    19,830  1.10506505   Root MSE        =    1.0288

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -.2154666   .0406406    -5.30   0.000    -.2951254   -.1358077
             |
      single |   .4112795   .0157081    26.18   0.000     .3804902    .4420687
             |
smoke#single |  -.1913809   .0510322    -3.75   0.000    -.2914083   -.0913535
             |
       _cons |   .6644013   .0121783    54.56   0.000     .6405307    .6882719
------------------------------------------------------------------------------

. test 1.smoke#1.single

 ( 1)  1.smoke#1.single = 0

       F(  1, 19827) =   14.06
            Prob > F =    0.0002
```

Both the ANOVA and linear regression output report an identical F-statistic for the empirical model (`F(3,19827) = 292.74, two-tailed p-value < 0.001`). This should already be a sign that the analyses are essentially equivalent. Specifically, they present the same information but in different ways.

We can use the linear regression output to retrieve each of the means by departing from the mean of the reference group, which is the constant of `0.6644013`, and adding regression coefficients. For instance, we can retrieve the mean for the category of people who smoke but are not single by adding the regression coefficient of “smoke” to the regression's constant (`0.6644013-0.2154666 = 0.4489347`).

It is not very complicated to also make the statistical tests reported in the linear regression equivalent to the tests reported in the ANOVA (or at least show that they are the same). To achieve this, we will estimate the linear regression not

with dummy variables but with variables that are effect coded (just like the ANOVA). We include the extension "_rec" after these new variables to highlight that they are effect coded.

```
. regress hours smoke_rec single_rec smoke_single_rec

      Source |       SS           df       MS      Number of obs   =    19,831
-------------+----------------------------------   F(3, 19827)     =    292.74
       Model |    929.4697         3  309.823233   Prob > F        =    0.0000
    Residual |  20983.9702    19,827  1.05835327   R-squared       =    0.0424
-------------+----------------------------------   Adj R-squared   =    0.0423
       Total |  21913.4399    19,830  1.10506505   Root MSE        =    1.0288

------------------------------------------------------------------------------
           hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
       smoke_rec |  -.1555785   .0127581   -12.19   0.000    -.1805854   -.1305716
      single_rec |   .1577945   .0127581    12.37   0.000     .1327877    .1828014
smoke_single_rec |  -.0478452   .0127581    -3.75   0.000    -.0728521   -.0228384
           _cons |   .7144625   .0127581    56.00   0.000     .6894557    .7394694
------------------------------------------------------------------------------

.
. test smoke_rec

 ( 1)  smoke_rec = 0

       F(  1, 19827) =  148.71
            Prob > F =    0.0000

. test single_rec

 ( 1)  single_rec = 0

       F(  1, 19827) =  152.97
            Prob > F =    0.0000

. test smoke_single_rec

 ( 1)  smoke_single_rec = 0

       F(  1, 19827) =   14.06
            Prob > F =    0.0002
```

When we run the test command for each coefficient, which just tests for linear effects after the estimation, we obtain the same F-statistics reported in the original ANOVA.

Our example suggests it may be commendable not to cling too strongly to one's preference for either linear regression or ANOVA. At least for relatively simple, two-way interaction models, which are popular in experimental accounting, the two approaches are econometrically the same. However, for other, often more complicated, analyses, things may be different and some experts suggest using one type of analysis over another (e.g., Gelman, 2005). Thus, we suggest you try this little exercise yourself with your experimental data.

References

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *Annals of statistics*, *33*(1), 1-53.

How to reference this online article?

Peters, C., & Van Pelt, V. F. J. (2021, February 26). Why anova and linear regression are the same. *Accounting Experiments*, Available at: https://www.accountingexperiments.com/post/anova-regression/.

ANOVA regression analyses data

Christian Peters

PhD Researcher in Accounting

Related

- 'Effect sizes don't matter in experiments.' Or do they?
- When and how to cluster standard errors in experimental data?