Why do R and statsmodels give slightly different ANOVA results?

Using a small R sample dataset and the ANOVA example from statsmodels, the degrees of freedom for one of the variables are reported differently, & the F-values results are also slightly different. Perhaps they have slightly different default approaches? Can I set up statsmodels to use R's defaults?

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols


##R code on R sample dataset

#> anova(with(ChickWeight, lm(weight ~ Time + Diet)))
#Analysis of Variance Table
#
#Response: weight
#           Df  Sum Sq Mean Sq  F value    Pr(>F)
#Time        1 2042344 2042344 1576.460 < 2.2e-16 ***
#Diet        3  129876   43292   33.417 < 2.2e-16 ***
#Residuals 573  742336    1296
#write.csv(file='ChickWeight.csv', x=ChickWeight, row.names=F)

cw = pd.read_csv('ChickWeight.csv')
cw_lm=ols('weight ~ Time + Diet', data=cw).fit()

print(sm.stats.anova_lm(cw_lm, typ=2))
#                 sum_sq   df           F         PR(>F)
#Time      2024187.608511    1  1523.368567  9.008821e-164
#Diet       108176.538530    1    81.411791   2.730843e-18
#Residual   764035.638024  575          NaN            NaN
```

Head and tail of the datasets are the same*, also mean, min, max, median of weight and time.

r pandas statsmodels anova

edited Aug 11 '15 at 6:39

asked Feb 27 '15 at 0:45

cphlewis

3,537 1 10 26

What versions of statsmodels and pandas were used for this example? I'm getting an error from the anova_lm function with pandas 0.18.0, statsmodels 0.6.1 – Alex Hasha Apr 8 at 17:11

Just checked my current system; pandas 0.17.1, statsmodels 0.6.1, had to re-install patsy but then it was fine. – cphlewis Apr 8 at 22:59

Thanks for checking. I realized I was running into this issue because my design matrix had missing values. – Alex Hasha Apr 9 at 20:31

1 Answer

Looks like "Diet" only has one degree of freedom in the statsmodels call which means it was probably treated as a continuous variable whereas in R it has 3 degrees of freedom so it probably was a factor/discrete random variable.

To make ols() treat "Diet" as a categorical random variable, use

```
cw_lm=ols('weight ~ C(Diet) + Time', data=cw).fit()
```

answered Feb 27 '15 at 1:17

MrFlick
70k 4 42 75