What's so bad about the R-Square?
It always gets larger as you add more variables to the model!

In this example I took a 1% sample of the National Medical Expenditure Survey (just for illustration purposes). Suppose you are interested in predicting a person's total **positive** medical expenditure based on their age and gender alone. One possible model is below (simplest one).

```
gen inter = LASTAGE * MALE
regress TOTALEXP LASTAGE MALE inter if TOTALEXP > 0
```

```
      Source |       SS       df       MS              Number of obs =     215
-------------+------------------------------           F(  3,   211) =    5.56
       Model |   159824534     3  53274844.5           Prob > F      =  0.0011
    Residual |  2.0207e+09   211  9576914.43           R-squared     =  0.0733
-------------+------------------------------           Adj R-squared =  0.0601
       Total |  2.1806e+09   214  10189502.2           Root MSE      =  3094.7

------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LASTAGE |   49.05184   15.13983     3.24   0.001     19.20714    78.89653
        MALE |   297.7409   1189.562     0.25   0.803    -2047.208    2642.689
       inter |   -24.6929   23.03988    -1.07   0.285    -70.11073    20.72493
       _cons |  -310.2091   778.7501    -0.40   0.691    -1845.336    1224.918
------------------------------------------------------------------------------
```

Suppose we have three additional variables in the dataset, but no real knowledge about how they are associated with medical expenditures, they all happen to be normally distributed with various means and variances.

```
gen ran1 = invnorm(uniform())*4 + 20
gen ran2 = invnorm(uniform())*2 - 4
gen ran3 = invnorm(uniform())*10+100
```

Now sequentially add these additional variables to your model: Which model would you select for prediction purposes?

```
regress TOTALEXP LASTAGE MALE inter ran1 if TOTALEXP > 0
```

```
      Source |       SS       df       MS              Number of obs =     215
-------------+------------------------------           F(  4,   210) =    4.46
       Model |   170584704     4  42646175.9           Prob > F      =  0.0018
    Residual |  2.0100e+09   210  9571279.88           R-squared     =  0.0782
-------------+------------------------------           Adj R-squared =  0.0607
       Total |  2.1806e+09   214  10189502.2           Root MSE      =  3093.7

------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LASTAGE |   48.70697   15.13887     3.22   0.001     18.86334    78.55059
        MALE |   265.9122   1189.591     0.22   0.823    -2079.158    2610.982
       inter |  -23.74023   23.05062    -1.03   0.304    -69.18048    21.70002
        ran1 |   53.55534    50.5101     1.06   0.290    -46.01647    153.1271
       _cons |  -1381.186   1275.286    -1.08   0.280    -3895.189    1132.818
------------------------------------------------------------------------------
```

```
regress TOTALEXP LASTAGE MALE inter ran1 ran2 if TOTALEXP > 0
```

```
      Source |       SS       df       MS              Number of obs =     215
-------------+------------------------------           F(  5,   209) =    3.82
       Model |   182691062     5  36538212.3           Prob > F      =  0.0025
    Residual |  1.9979e+09   209  9559150.32           R-squared     =  0.0838
-------------+------------------------------           Adj R-squared =  0.0619
```

```
       Total |  2.1806e+09   214  10189502.2           Root MSE      =  3091.8

------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LASTAGE |   49.47856    15.1448     3.27   0.001     19.62241    79.33471
        MALE |   211.2277   1189.829     0.18   0.859    -2134.378    2556.833
       inter |  -23.58146   23.03644    -1.02   0.307    -68.99502    21.83209
        ran1 |   50.39465   50.55615     1.00   0.320    -49.27071      150.06
        ran2 |  -113.0473    100.453    -1.13   0.262    -311.0783    84.98374
       _cons |  -1767.653   1319.934    -1.34   0.182    -4369.744    834.4378
------------------------------------------------------------------------------

regress TOTALEXP LASTAGE MALE inter ran1 ran2 ran3 if TOTALEXP > 0

      Source |       SS       df       MS              Number of obs =     215
-------------+------------------------------           F(  6,   208) =    3.21
       Model |   184756347     6  30792724.4           Prob > F      =  0.0049
    Residual |  1.9958e+09   208  9595178.51           R-squared     =  0.0847
-------------+------------------------------           Adj R-squared =  0.0583
       Total |  2.1806e+09   214  10189502.2           Root MSE      =  3097.6

------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LASTAGE |   49.24802   15.18145     3.24   0.001     19.31879    79.17726
        MALE |   203.2154   1192.195     0.17   0.865    -2147.118    2553.549
       inter |  -23.54172   23.07997    -1.02   0.309    -69.04237    21.95893
        ran1 |   47.68723    50.9864     0.94   0.351    -52.82912    148.2036
        ran2 |  -115.0753    100.737    -1.14   0.255    -313.6718    83.52116
        ran3 |   10.33581   22.27826     0.46   0.643    -33.58431    54.25594
       _cons |  -2728.032   2456.393    -1.11   0.268    -7570.651    2114.587
------------------------------------------------------------------------------
```

Our results can be summarized as follows:

Model	MSE	Model df	R-squared	Adj. R-squared
Age, gender, interaction	9576914.43	3	0.0733	0.0601
Age, gender, interaction, ran1	9571279.88	4	0.0782	0.0607
Age, gender, interaction, ran1, ran2	9559150.32	5	0.0838	0.0619
Age, gender, interaction, ran1, ran2, ran3	9595178.51	6	0.0847	0.0583

NOTE: The R-squared value continues to increase with the addition of each “noise” variable. What is the Adjusted R-squared calculating?