PPOL502-01, Spring 2016
Course Notes #6: R-Squared and Adjusted R-Squared in Multiple Regression

---

I. R-SQUARED FOR MULTIPLE REGRESSION (a.k.a. Coefficient of Multiple Determination)

- Recall R-squared from our discussion of bivariate regression:
- SST = Sum of Squares total = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ (i.e., the total variation in Y)
- SSE = *Explained* sum of Squares = $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ (i.e., the variation in Y that is explained by the regression)
- SSR = *Residual* Sum of Squares = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ (i.e., the variation in Y that is not explained by the regression)

- Total variation is equal to what the regression explains, plus what's left over (and is unexplained by the regression): SST = SSE + SSR

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{SSE}{SST} = \frac{\text{explained variation in } Y \text{ from model}}{\text{total variation in } Y} =$$

$$1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \overline{Y})^2} = 1 - \frac{SSR}{SST} = 1 - \frac{\text{unexplained variation}}{\text{total variation}} =$$

$$\frac{\sum(Y_i - \overline{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \overline{Y})^2} = \frac{SST - SSR}{SST} = \frac{Error\ without\ Xs - Error\ with\ Xs}{Error\ without\ Xs}$$

- These very same ideas transfer to multiple regression. The only difference is that for simple regression: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$
  and for multiple regression: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + ...\hat{\beta}_k X_k$

- R-squared still has a possible range between 0 (i.e., no explanatory power) and 1 (i.e., the variation in ***X*** perfectly explains the variation in *Y*)

- In bivariate regression, R-squared was exactly equal to the square of the correlation coefficient between Y and X_1. Remember that we said this was a special case for bivariate regression only. The sums of the squared correlation coefficients does not equal R-squared in the multivariate case.

- ******NOTE***:*** Remember, different texts and statistical packages refer to SSR and SSE differently (opposite in fact) (e.g., either "sum of squared residuals" or "sum of squares from the regression"; and "sum of squares explained from the estimation" or "sum of squared errors").

  "*Explained*" Sum of Squares may also be referred to as "Model" or "Regression"
  "*Unexplained*" Sum of Squares may also be referred to as "Error" or "Residual"

  So, you shouldn't memorize the formula for the terms "SSR" and/or "SSE" because these terms are easily confused. Instead, understand the formula in terms of $\sum(Y_i - \overline{Y})^2$, $\sum(\hat{Y} - \overline{Y})^2$, and $\sum(Y - \hat{Y})^2$.

- **Comparing regressions**. Recall the correlation matrix from CRIME1:

```
. corr crime1 police criminal
(obs=9)

             |   crime1   police criminal
-------------+---------------------------
      crime1 |   1.0000
      police |   0.7950   1.0000
    criminal |   0.9452   0.9487   1.0000
```

- What will R-squared be in the regression of CRIME1 on POLICE (X_1)?

  In the regression of CRIME1 on CRIMINAL (X_2)?

  In the regression of CRIME1 on POLICE *and* CRIMINAL?

MODEL 1:

```
. *** Model 1
. regress crime1 police

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =   12.03
       Model |  37.7643227     1  37.7643227           Prob > F      =  0.0104
    Residual |  21.9817902     7  3.14025575           R-squared     =  0.6321
-------------+------------------------------           Adj R-squared =  0.5795
       Total |  59.7461129     8  7.46826412           Root MSE      =  1.7721

------------------------------------------------------------------------------
      crime1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      police |   .7933507   .2287741     3.47   0.010     .2523859    1.334316
       _cons |   .9849016   1.287384     0.77   0.469    -2.059277    4.029081
------------------------------------------------------------------------------
```

MODEL 2:

```
. *** Model 2
. regress crime1 criminal

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =   58.62
       Model |  53.3729488     1  53.3729488           Prob > F      =  0.0001
    Residual |  6.37316412     7  .910452017           R-squared     =  0.8933
-------------+------------------------------           Adj R-squared =  0.8781
       Total |  59.7461129     8  7.46826412           Root MSE      =  .95418

------------------------------------------------------------------------------
      crime1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    criminal |    .994177   .1298469     7.66   0.000     .6871378    1.301216
       _cons |  -.0192298   .7229571    -0.03   0.980    -1.728752    1.690292
------------------------------------------------------------------------------
```

MODEL 3

```
. *** Model 3
. regress crime1 police criminal

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  2,     6) =  880.19
       Model |  59.5431697     2  29.7715849           Prob > F      =  0.0000
    Residual |  .202943222     6  .03382387            R-squared     =  0.9966
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  59.7461129     8  7.46826412           Root MSE      =  .18391

------------------------------------------------------------------------------
      crime1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      police |  -1.014086    .075082   -13.51   0.000    -1.197805   -.8303668
    criminal |   2.008263   .0791434    25.37   0.000     1.814606    2.20192
       _cons |  -.0192298   .1393464    -0.14   0.895    -.3601981    .3217384
------------------------------------------------------------------------------
```

- A fun fact about R-squared from multiple regressions: as you add variables to the model, *R-squared cannot get smaller*. Thus, regressions with additional variables typically provide a better fit of the model (i.e., explain more variation in Y). R-squared is hard-wired for this. Why? Think back to what R-squared is:

$$R^2 = \frac{\sum (Y_i - \bar{Y})^2 - \sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2}$$

Suppose that in **Model 1:** $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + u_1$

and in **Model 2**: $Y = \hat{\alpha}_0 + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + u_2$

- The *total* sum of squares is the same in each case, i.e., the denominator of R-squared, does not change.

- Why can't R-squared decrease when an additional *X* variable is added to the regression? Mirer (1995, p. 141) has a nice explanation:

  "Since OLS is acting to minimize [the sum of squared residuals - SSR], it need not allow an additional specified variable to increase the SSR; it could effectively ignore the new variable rather than let it increase the SSR."

- So, why not just add all the variables we can possibly think of to increase the explanatory power of the model? R-squared will just keep increasing until, "ideally," we get to R-squared = 1. What are the advantages and disadvantages to this approach?

II. ADJUSTED R-SQUARED

- The adjusted R-squared statistic accounts in a specific way for the fact that there's a price to pay for each additional variable added to the model: think of this as the "statistical cost" of adding a variable to the model:

Adjusted R-squared:
$$\bar{R}^2 = 1 - \left(\frac{\sum u_i^2 / (n-k-1)}{\sum (Y_i - \bar{Y})^2 / (n-1)} \right)$$

-- where *k* is the number of *Xs* in the model, *n* is the sample size.

EXAMPLE: Model 3: $CRIME1HAT = \hat{\beta}_0 + \hat{\beta}_1 POLICE + \hat{\beta}_2 CRIMINAL$

```
. *** Model 3
. regress crime1 police criminal

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  2,     6) =  880.19
       Model |  59.5431697     2  29.7715849            Prob > F      =  0.0000
    Residual |  .202943222     6   .03382387            R-squared     =  0.9966
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  59.7461129     8  7.46826412            Root MSE      =  .18391

------------------------------------------------------------------------------
      crime1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      police |  -1.014086    .075082   -13.51   0.000    -1.197805   -.8303668
    criminal |   2.008263   .0791434    25.37   0.000     1.814606    2.20192
       _cons |  -.0192298   .1393464    -0.14   0.895    -.3601981    .3217384
------------------------------------------------------------------------------
```

R-SQUARED:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{59.54317}{59.74611} = 0.9966$$

or

$$R^2 = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = \frac{(59.74611 - 0.20294)}{59.74611} = 0.9966$$

ADJUSTED R-SQUARED:

$$\bar{R}^2 = 1 - \left(\frac{\sum u_i^2 \big/ (n-k-1)}{\sum (Y_i - \bar{Y})^2 \big/ (n-1)} \right) = 1 - \left(\frac{0.20294/6}{59.74611/8} \right) = 0.9955$$

- Should you use "real" R-squared or adjusted R-squared?
  -- opinions differ: both are often helpful.

- Is a low R-squared inherently a "bad" thing?
  -- Not necessarily.

III. INTERPRETING R-SQUARED & ADJUSTED R-SQUARED

The interpretation of whether an R-squared value is "high" or "low" is context specific: Mirer (1995, p. 94) has a helpful thought:

"Our assessment of the magnitude of R-squared depends on the nature of the economic process being analyzed. The R-squared is often high in time-series work because Y and X often have a common trend. By contrast, the R-squared tends to be lower in cross-section work because there is no trend and because of the substantial variation in individual behavior. If R-squared = 0.443 were reported for a macroeconomic time-series saving function, it would be judged quite low; experienced researchers expect a regression of aggregate saving on income to have an R-squared of 0.95 or higher. However, an R-squared of 0.443 for a large-sample, cross-section saving function would be quite high compared with those for similar studies, and the regression would be judged to have a relatively good fit."

The following notes are from Introduction to Econometrics (p. 176) by Stock and Watson, 2003.

An R-squared or an adjusted R-squared near 1.0 means that the regressors are good at predicting the values of the dependent variable in the sample, and an R-squared or an adjusted R-squared near 0 means they are not. This makes these statistics useful summaries of the predictive ability of the regression. However, it is easy to read more into them than they deserve. Following are four pitfalls to guard against when using R-squared or adjusted R-squared:

- ***An increase in the R-squared or adjusted R-squared does not necessarily mean that an added variable is statistically significant.***

  You need to perform a hypothesis test using the t-statistic to ascertain whether an added variable is statistically significant.

- ***A high R-squared or adjusted R-squared does not mean that the regressors are a true cause of the dependent variable.***

  You could have a set of independent variables that are good predictors of the dependent variable but do not directly cause the outcome.

- ***A high R-squared or adjusted R-squared does not mean there is no omitted variable bias.***

  Omitted variable bias can occur in regressions with low, moderate, or high R-squareds. Conversely, a low R-squared does not imply that there necessarily is omitted variable bias.

- ***A high R-squared or adjusted R-squared does not necessarily mean you have the most appropriate set of regressors, nor does a low R-squared or adjusted R-squared necessarily mean you have an inappropriate set of regressors.***

  The question of what constitutes the right set of regressors is difficult and must weigh issues of omitted variable bias, data availability, data quality, economic theory and the nature of the substantive questions being addressed. It is inappropriate to judge the model simply based on the magnitude of the R-squared.

SUPPLEMENT I: INCREMENTS TO R-SQUARED FROM ADDING REGRESSORS

********** NOTICE **********

- In the formulas that follow, remember: the important thing is to realize that in multiple regression, the *Xs* can be correlated with each other and thus each one contributes a partial effect to the whole, not the exact same effect that we would get from a bivariate regression using that variable. This is just another way of understanding "holding criminals constant" or the partialling out interpretation.

  Don't worry about memorizing these formulas. Only dig into them if they help you to gain insight into what's going on in multiple regression.

- Here's a way to think of the buildup of the sum of squares that's explained by the regression, $\sum(\hat{Y} - \bar{Y})^2$:

Variable	***Sum of Squares Explained by the regression***
X_1	$r_{y1}^2 * \sum(Y - \bar{Y})^2$
Increment due to X_2	$r_{y1.2}^2(1 - r_{y1}^2) * \sum(Y - \bar{Y})^2$
X_1 and X_2	$R_{y.12}^2 * \sum(Y - \bar{Y})^2$

Where:

r_{y1}^2 is the square of the correlation between Y and X_1. In a simple regression of Y on X_1, $r_{y1}^2 = R^2$, i.e., the proportion of the variation in Y that is explained by the variation in X_1.

$r_{y1.2}^2$ is the proportion of the remaining variation in Y that is *uniquely* explained by *adding* X_2 to the model. Think of it as isolating X_2's "special" contribution to explaining Y, once X_1 is already in the model, where this "special" contribution is separate from some joint influence that both X_1 and X_2 might have in explaining Y. This term is defined as: $r_{y1.2}^2 = \dfrac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{y1}^2}\sqrt{1 - r_{12}^2}}$, where

r_{y2} is the correlation between Y and X_2

r_{y1} is the correlation between Y and X_1

r_{12} is the correlation between X_1 and X_2

For the example where X_1 = POLICE, X_2 = CRIMINALS, and Y=CRIME1:

- $r_{y1}^2 * \sum(Y - \bar{Y})^2 = 0.6320727 * 59.74611 = 37.764$. This is the sum of squares that is explained by the regression model. Compare this value to the "Model Sum of Squares" for MODEL 1 on the next page.

- $r^2_{y1.2} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1-r^2_{y1}}\sqrt{1-r^2_{12}}} = \frac{0.94516 - (0.79503 * 0.94868)}{\sqrt{1-0.6320727}\sqrt{1-0.899994}} = 0.9953648$

- $r^2_{y1.2}(1-r^2_{y1})$= [0.9953648* (1-0.6320727)] = 0.36622188. This is the *increment to* ***R-squared*** by adding X_2 to the model. (Compare the R-squared in Models 1 and 3 on the next page. The numbers are slightly different in the third decimal place due to rounding)

  $r^2_{y1.2}(1-r^2_{y1}) * \sum(Y-\bar{Y})^2 = 0.36622188 * 59.74611 = 21.88$ is the *increment* in the **Sum of Squares that is explained by the regression**, i.e., the additional variation in Y that is explained by adding X_2 to the model. (Compare the Model Sum of Squares in Models 1 and 3. Again, the numbers are slightly different due to rounding)

SUPPLEMENT II: Calculating Adjusted R-Squared:

- We know that the formula for $R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$.

  Is it possible to just calculate adjusted R-squared as: $\bar{R}^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2 \big/ (n-k-1)}{\sum(Y_i - \bar{Y})^2 \big/ (n-1)}$???

- Answer: No. Sections (A) and (B) below prove that this is the case:

A. Think Back to "Plain Old R-Squared"

- Specifically, focus on two of the definitions for **R-squared:**

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

- These implicitly have *(n-1)* in the denominator each element, and these terms cancel out in the formulas you see above, i.e.:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2 \big/ (n-1)}{\sum(Y_i - \bar{Y})^2 \big/ (n-1)} = \left(\frac{\sum(\hat{Y}_i - \bar{Y})^2}{(n-1)}\right) * \left(\frac{(n-1)}{\sum(Y_i - \bar{Y})^2}\right) = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

and

$$R^2 = \frac{\left(\sum(Y_i - \bar{Y})^2 \big/ (n-1)\right) - \left(\sum(Y_i - \hat{Y}_i)^2 \big/ (n-1)\right)}{\sum(Y_i - \bar{Y})^2 \big/ (n-1)} = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

B. Now think about Fancy (adjusted) R-squared in a similar way:

$$\bar{R}^2 = 1 - \left(\frac{\sum u_i^2 \big/ (n-k-1)}{\sum(Y_i - \bar{Y})^2 \big/ (n-1)}\right) = 1 - \left(\frac{\sum(Y_i - \hat{Y}_i)^2 \big/ (n-k-1)}{\sum(Y_i - \bar{Y})^2 \big/ (n-1)}\right)$$

Can this simply be estimated as $= \dfrac{\sum(\hat{Y}_i - \bar{Y})^2 \Big/ (n-k-1)}{\sum(Y_i - \bar{Y})^2 \Big/ (n-1)}$?

NO. To see why, start with the original $\bar{R}^2$ equation and substitute, multiply, divide, etc:

$$\bar{R}^2 = 1 - \left(\frac{\sum u_i^2 / (n-k-1)}{\sum (Y_i - \bar{Y})^2 / (n-1)} \right) = \left(\frac{\sum (Y_i - \bar{Y})^2 / (n-1)}{\sum (Y_i - \bar{Y})^2 / (n-1)} \right) - \left(\frac{\sum (Y_i - \hat{Y}_i)^2 / (n-k-1)}{\sum (Y_i - \bar{Y})^2 / (n-1)} \right) =$$

$$= \frac{\left(\sum (Y_i - \bar{Y})^2 / (n-1) \right) - \left(\sum (Y_i - \hat{Y}_i)^2 / (n-k-1) \right)}{\sum (Y_i - \bar{Y})^2 / (n-1)}$$

Look at the big numerator of the last line, and compare this to the equation at the top of this page, and compare the "proposed" equation for estimating adjusted R-squared above.

Yes, **it's true** that $\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2 = \sum(\hat{Y}_i - \bar{Y})^2$

but, it's *not* true that

$$\left(\sum (Y_i - \bar{Y})^2 / (n-1) \right) - \left(\sum (Y_i - \hat{Y}_i)^2 / (n-k-1) \right) = \left(\sum (\hat{Y}_i - \bar{Y})^2 / (n-k-1) \right)$$

****