

## R-Squared Notes:

So far, we have not focused on the R-squared value to evaluate how “well” our model fits the data. Why? Because too much emphasis can be placed on this particular measure, and if you go on to study “time-series” data, you will see that the R-squared value can be extremely misleading.

Things to note:

- there is no value that R-squared should be for you to claim that your model does a good job at explaining the variation in the dependent variable. It is simply an estimate of how much variation can be explained.
- a small R-squared value implies that the error variance is large relative to the variance of  $y$ , which means that we may have a hard time precisely estimating the  $\beta$  coefficients. BUT, this can be offset by a large sample size. This is true even if we have not controlled for many unobserved factors – which leads to the large error term. EXAMPLE: suppose that some incoming students at a large university are RANDOMLY given grants to buy computer equipment. If the amount of the grant is truly randomly determined, we can estimate the ceteris paribus effect of the grant amount on subsequent college grade point average by using simple regression analysis. Because of the random assignment, all of the other factors affecting GPA would be UNCORRELATED with the grant size. Now, it seems pretty unlikely that grant size would explain very much of the variation in GPA, so the R-squared from this simple regression would probably be pretty low, BUT we might still (with a large enough  $N$ ) get a reasonably precise estimator for the effect on the grant. (NOTE: we don't need to worry about omitted variable bias since all the omitted variables would be uncorrelated with the grant size!)
- The relative CHANGE in the R-squared value when variables are added to an equation provides A LOT OF USEFUL INFORMATION. This is related to the joint F-tests that we talked about earlier in testing joint restrictions.

R-squared and Adjusted R-squared Value: what happens when we add regressors to our equation.

- Recall that R-squared is the ratio between the explained SS/total SS, or:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{RSS/N}{TSS/N}$$

Now, why is it helpful to write R-squared in this fashion? Think about the following: let  $\sigma_y^2$  be the population variance of  $y$  (unobserved by us) and  $\sigma_\epsilon^2$  be the population variance on the random disturbance term (again, unobserved by us). Define the POPULATION R-squared to be:

$$pop R^2 = 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2}$$

which tells us the proportion of the variation of y in the population explained by the independent variables. But we don't observe the population variances. So, we can use estimators for them:

Okay: so  $RSS/N$  is our ESTIMATOR for  $\sigma_\epsilon^2$  and  $TSS/N$  is our estimator for  $\sigma_y^2$  in the "usual" R-squared. That is, the usual R-squared is an estimator for the POPULATION R-squared. BUT WE KNOW THAT BOTH OF THESE ESTIMATORS ARE BIASED (numerator and denominator). We can, instead use unbiased estimators for  $\sigma_\epsilon^2$  and  $\sigma_y^2$ . In particular, we could use:

$RSS/(N-k-1)$  and  $TSS/(N-1)$ .

If we do this, we can get an ADJUSTED-R-squared value that is given by:

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{TSS/(N-1-k)}$$

BUT: something to keep in mind is that the ratio of unbiased estimators DOES NOT LEAD TO AN UNBIASED ESTIMATOR. And, in fact, the adjusted R-squared estimator is not generally thought to be a better estimator for the population R-squared over the usual R-squared value.

(Recalling that our UNBIASED estimator for the variance on the error term is  $RSS/(N-k-1)$ .)

So, how does the adjusted and regular R-squared differ?

1. Adjusted R-squared value takes into account the number of INDEPENDENT variables in the model, whereas the regular R-squared does not.
2. In fact, if we add new independent variables to our model, the adjusted R-squared value will ONLY go up if the t-statistic on the coefficient estimator of the new variable is GREATER THAN ONE in absolute value. (If you add MORE THAN ONE independent variable, the adjusted R-squared will only go up if the F-statistic for the JOINT SIGNIFICANCE of all the new variables is greater than one). SO: this is a little bit different than if you were to look at the individual t-stat or the F-stat, alone (since we would only reject the null if the test statistic is usually LARGER than one...at the usual levels of significance).
3.  $\bar{R}^2 = 1 - (1 - R^2)(N-1)/(N-k-1)$  is the relationship between the adjusted and regular R-squared values.

Okay: so, now why would we ever look at the adjusted R-squared value and not the R-squared value?

Using the Adjusted R-squared to Choose Between Non-nested Models.

R-squared will ALWAYS go up if you add RHS variables. Why? Because the RSS can never go up when you add additional variables to your equation. And, if that's so, looking at the R-squared alone and whether it goes up doesn't tell you if you've got a "better" model.

So, what can you do?

- when we computed F-statistics for testing joint significance of a group of variables, this allowed us to decide, at a particular significance level, whether at least ONE of the variables in the group affects the dependent variable. This test does NOT tell you which of the variables has an effect.
- in some cases, you may want to choose a model WITHOUT redundant indep vars, and the adjusted R-squared can help you do this.

Consider the following NON-NESTED models:

$$(1) \quad \log(\text{salary}) = \beta_0 + \beta_1 \text{years} + \beta_2 \text{games} + \beta_3 \text{bavg} + \beta_4 \text{hruns} + \epsilon$$

$$(2) \quad \log(\text{salary}) = \beta_0 + \beta_1 \text{years} + \beta_2 \text{games} + \beta_3 \text{bavg} + \beta_4 \text{rbis} + \epsilon$$

These models are non-nested as one is NOT a special case of the other (in the joint F-test when you ran restricted and unrestricted models, the restricted model was a special case of the unrestricted model – that was a nested model).

Not surprisingly, home runs and rbis are highly correlated. You may just use one to use equation (1) or (2) but not an equation that contains BOTH home runs and rbis. (When you run that regression with both of them in there, neither was individually statistically significant.)

Suppose that the adjusted R-squared for (1) is 0.6211 and for (2) is 0.6266. This says that according to the adjusted R-squared value, it marginally prefers equation (2). (Because both models have the same number of indep vars, you could have used the regular R-squared to come to this conclusion as well...)

Consider the following two non-nested models on R&D intensity to firm sales:

$$(1) \quad \text{rdintensity} = \beta_0 + \beta_1 \log(\text{sales}) + \epsilon$$

$$(2) \quad \text{rdintensity} = \beta_0 + \beta_1 \text{sales} + \beta_2 \text{sales}^2 + \epsilon$$

First model captures diminishing returns by including the log of sales. The second model captures diminishing returns by including the squared value of sales. We have TWO DIFFERENT FUNCTIONAL FORMS. How do you choose between them?

- comparing R-squared values isn't really fair because (2) has more RHS variables than (1). So, you can compare the adjusted R-squared value, instead.

NOTE: you can't use the adjusted or regular R-squared value to compare functional forms when the dependent variable is NOT IDENTICAL. The LHS var must be in the same functional form to compare it. Why? Suppose you had  $y$  and  $\ln(y)$ . The R-squared value (adjusted and regular) is

trying to show how much of the variation in the LHS variable is explained by the data. But the  $\text{Var}(y)$  and the  $\text{Var}(\ln y)$  are going to be DIFFERENT. So, this just doesn't make sense.