

6.3 More on Goodness-of-Fit and Selection of Regressors

1. What is R-squared? (Chapter 3)

It is a ratio of the explained sum of squares (SSE) to total sum of squares (SST) or could be expressed as a measure of what is potentially left unexplained through the residual sum of squares (SSR)

$$SST = \sum_{i=0}^n (y_i - \bar{y})^2 \quad SSE = \sum_{i=0}^n (\hat{y}_i - \bar{y})^2$$
$$SSR = \sum_{i=0}^n \hat{u}_i^2 \quad R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Interpreted as the proportion of the sample variation in y_i that is explained by the OLS regression line

2. **Caution:** But it has nothing to do with biased-ness of the estimates or the accuracy of the functional form.
- ▶ Could potentially increase sample size and reduce error variance, increasing the precision of the estimates. But the model could still be faulty!
 - ▶ Could have a variable whose true explanatory power (in population) is only 25% so R-squared will not go above 25% no matter how large the sample is!
 - ▶ With a carefully designed random experiment, you could potentially obtain a $\hat{\beta}$ which is true to its population counterpart, despite a very low R^2
3. Low R-squared will make prediction difficult. If our estimates are precise through random experimental data, but they only explain 10% of total dependent variable behavior, then we cannot predict accurately.

6.3.1 Adjusted R-Squared

1. Adjusted R-square: is nothing but R-squared adjusted for the size of the data (n) and the number of independent variables (k) in the model

$$\bar{R}^2 = 1 - \frac{SSR/(n - k - 1)}{SST/(n - 1)} = 1 - \frac{\hat{\sigma}^2}{SST/(n - 1)}$$

2. Note: \bar{R}^2 is not generally known to be a better estimator than R^2 . Then why do we use \bar{R}^2 ?

3. \bar{R}^2 tell us how many x 's are too many x 's?

- ▷ Since adding independent variables increases explanatory power of the model, should we go on adding x 's?

No, since you gain explanatory power (perhaps), but you loose degree of freedom which is reflected in \bar{R}^2 .

- ▷ How do we know which addition is good?

If we add a new independent variable to a regression equation, \bar{R}^2 increases if, and only if, the t statistic on the new variable is greater than one in absolute value.

4. Relation between \bar{R}^2 and R^2

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

6.3.2 Choosing between Non-nested models

1. Remember: usage of F-stat for comparing strength of nested models (chapter 4)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + u$$

Second model is obtained when you estimate the first model under the restriction $\beta_2 = 0$. These are called nested models. (Do exercise C5 in chapter 4: next week Monday)

2. **Example of non nested models:** consider the problem of whether an increase in sales causes firms to invest more in research and development as follows:

$$rdintens = \beta_0 + \beta_1 \log(sales) + u \qquad R^2 = 0.061 \quad \bar{R}^2 = 0.030$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u \qquad R^2 = 0.148 \quad \bar{R}^2 = 0.090$$

Strange example: because the second model is better by looking at both the measures.

3. Bottom line:

- (a) when you suspect how many x 's, then use \bar{R}^2
- (b) \bar{R}^2 cannot help you choose the type of functional form - total variation in y and $\log(y)$ are not same
- (c) most likely, \bar{R}^2 might just be a double check rather than your go to measure as you would not have a lot of independent variables in most applied work.

6.3.3 Over-controlling and Adding Too Much

1. Consider the following example where we want to explain traffic fatalities (fatalities) as a function of tax on alcohol (tax) controlling for other co-variates. We do not use the following model:

$$\begin{aligned} fatalities = & \beta_0 + \beta_1 tax + \beta_2 beercons + \beta_3 miles \\ & + \beta_4 perc male + \beta_5 perc 16 - 21 \end{aligned}$$

Why is this wrong: β_1 measures the difference in fatalities due to a one percentage point increase in tax, holding *beercons* fixed.

2. **Over-controlling:** We want the effect of tax on fatalities and not tax on beer consumption itself.
 - (a) Potentially unsure of how the effects of the independent variable are affecting the dependent variable – overcontrolling by accident
 - (b) Potentially unsure of isolating effect of desired independent variable of interest – overcontrolling on intent

3. **Caution:** If we remember that different models serve different purposes, and we focus on the *ceteris paribus* interpretation of regression, then we will not include the wrong factors in a regression model.
- (a) Add variables when they are sure to affect y and are uncorrelated with all other x 's – no harm, but how much does it contribute?
 - ▷ Read the beer explanation (page 200)
 - (b) Adding extra x 's will not ensure an unbiased estimation, but might just ensure more precise estimates
 - (c) We need not worry about whether some of our explanatory variables are “endogenous”—provided these variables themselves are not affected by the policy