Lab 8 - Analysis of Variance, Standard Errors

Spring 2018

Contents

- Today we will use a dataset on Congressional campaign expenditures **vote1**
- We will explore the relationship between share of votes received for Candidate A ,`voteA`, and the share of total campaign expenditure spent by Candidate A, `shareA`.
- We hypothesize that the relative campaign expenditure of one candidate should impact the share of votes she receives. As a result, if we were to have a cross-section of observations of `voteA` and `shareA` we expect there to be the following relation (for *each* observation i):

$$voteA_i = \beta_0 + \beta_1 shareA_i + \varepsilon_i$$

  Note that $E\left(\partial voteA/\partial shareA\right) = \beta_1$. What is the interpretation of β_1?

1 Analysis of Variance

Analysis of Variance (sometimes called ANOVA) is the first step towards assessing the success of the relationship that you hypothesized as existing in the data. ANOVA will help us determine whether the variations in our regressors match the variations in our dependent variable.

1.1 ANOVA - Theory

- As suggested by the title, ANOVA looks at the variance in our model. In this section we provide a simple formulation of the ideas that we will work with.
- Start with the following model specification that holds for each observation i in the data:

$$y_i = \alpha + \beta x_i + u_i \tag{1}$$

- Estimate the specification using the data you have. The estimated relationship is

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i \tag{2}$$

  This condition holds exactly in your data by the construction of the OLS estimates. All of the OLS assumptions should also hold.
- Alternatively we can write the above equation as

$$y_i = \hat{y}_i + \hat{u}_i \tag{3}$$

  this simply states that the observed value for y is the sum of two components: the first is the predicted value, $\hat{y}_i$ i.e. the part of y explained by our model: the second is the residual, $\hat{u}_i$ the part of y that our model cannot explain.
- Recall two properties of our OLS estimator:
  1. the sample average of the predicted values, $\hat{y}_i$ is the same as the sample average of the y_i, or $\bar{\hat{y}} = \bar{y}$. Notice that this and the following result in equation (4) are only true if you include a constant in your regression.
  2. the sample covariance between $\hat{y}_i$ and $\hat{u}_i$ is equal to zero.

- Now computing the variance of both sides of equation (3), using the two properties above and omitting the division by $N-1$ on both sides:

$$\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{N}\hat{u}_i^2 \tag{4}$$

Let us now define the objects above:

- $SST \equiv \sum_{i=1}^{N}(y_i-\bar{y})^2$ is the **total sum of squares**
- $SSE \equiv \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2$ is the **explained sum of squares**
- $SSR \equiv \sum_{i=1}^{N}\hat{u}_i^2$ is the **residual sum of squares**
- SST measure the total sample variation in y_i (Notice that if you divide SST by $N-1$ you obtain the sample variance of y_i). SSE measures the sample variation in $\hat{y}_i$ and SSR measures the sample variation of the residual $\hat{u}_i$. Equation (4) is telling you that the variation in y_i can always be expressed as the sum of the variation explained by your model and the variation unexplained by it.

$$SST = SSE + SSR \tag{5}$$

1.2 ANOVA - Application

- Estimate $voteA_i = \beta_0 + \beta_1 shareA_i + \varepsilon_i$ using `regress`. Focus on the top panel of results, which is sometimes called ANOVA table

```
      Source |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  1,   171) = 1017.70
       Model |  41486.4749     1  41486.4749           Prob > F      =  0.0000
    Residual |  6970.77363   171  40.7647581           R-squared     =  0.8561
-------------+------------------------------           Adj R-squared =  0.8553
       Total |  48457.2486   172  281.728189           Root MSE      =  6.3847
```

- The second column reports the calculated sums as we defined them above - in order SSE, SSR, SST. The third column reports the number of regressors (not including the intercept) and the number of residuals minus the total number of coefficients estimated. The fourth column is just the second divided by the third.
- As we discussed above the size of SSE is a good measure of how much variation there is in $\hat{y}_i$. Recall that $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. This is telling you that the variation that you observe in $\hat{y}_i$ comes directly from the variation of variable x_i in the sample.
- It is then natural to take as a measure of *how much sample variation in y is explained by variation in x* the ratio of SSE to SST[1]. We call this object the **R-squared or coefficient of determination** of the regression:

$$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{6}$$

[1] As long as SST $\neq 0$

- You can find the R-squared on the fourth line of the right side of the top panel. It provides a better measure for the success of our model compared to just looking at the sums in the ANOVA. By construction R-squared is between 0 and 1.
- We use R^2 to compare one specification of a model to another. A higher R^2 is usually interpreted as evidence for a "better" model.
- Since the OLS estimates come from minimizing SSR, and since SSR cannot increase if you add explanatory variables to the model, R-squared cannot decrease, and will typically increase, when you add explanatory variables to the model.
- On the other hand, the more variables you include, the more coefficients you have to estimate and the more degrees of freedom you lose in the process, ending up with more imprecise estimates. Adjusted R-squared accounts for this and only increases when the additional variables have significant additional explanatory power.

2 Standard Errors and Statistical Significance

For this part we estimate the augmented model

$$voteA_i = \beta_0 + \beta_1 \times shareA_i + \beta_2 \times democA + \varepsilon_i \tag{7}$$

- Note that the R^2 of the augmented model with an additional regressor has increased while the Adj. R^2 has actually declined implying that our first model is better in terms of predicting the dependent variable.
- As we discussed earlier, instead of using the Adj. R^2 as a criterion for which model is better we can examine the statistical significance of the additional regressor. We proceed with this strategy.
- The second panel of the regression output for the second model is:

```
------------------------------------------------------------------------------
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      shareA |   .4634151   .0151651    30.56   0.000      .433479    .4933512
      democA |   .0999448   1.018794     0.10   0.922    -1.911172    2.111062
       _cons |   26.77795   .9570672    27.98   0.000     24.88869    28.66722
------------------------------------------------------------------------------
```

- The second column reports the estimated OLS coefficients: $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$.
- The third column reports in a very loose sense: $\sqrt{\text{Var}(\hat{\alpha})}$, $\sqrt{\text{Var}(\hat{\beta}_1)}$, and $\sqrt{\text{Var}(\hat{\beta}_2)}$. Pay particular attention to the hats in the notation. The interpretation is the standard deviation of the *OLS-ESTIMATED* coefficients. In other words they tell us how wrong are these estimates from the true coefficients α, β_1, and β_2 (note the lack of hats in the notation).
- You can think of the coefficient estimates as the best guesses of the true coefficients. The Std. Err. tells you how wrong the guess could be.

- Since a point-guess is very restrictive by construction we can create an interval-guess i.e. estimate a range of possible estimates that should contain the true coefficient with very high degree of confidence. Note that while it is very easy for the point-guess to be wrong, an interval-guess can have a much higher probability of being right. The interval guess is what we formally refer to as a **confidence interval**.
- The confidence interval can be found in the last two columns of the second panel of the regression output.
- Note that our confidence interval for the `shareA` is $[0.43, 0.49]$. Given our OLS assumptions this interval is highly likely (in fact 95% likely) to include the true coefficient value.
- Note that our confidence interval for `democA` is $[-1.9, 2.1]$. This implies that the true coefficient can in fact be zero. If this is indeed the case then this regressor will not affect the dependent variable.

2.1 Statistical Significance

- We saw that sometimes the OLS point-guess is not zero but the OLS interval-guess includes zero as the true coefficient. To further investigate the possibility of the true coefficient being zero we proceed with a test of statistical significance.
- Let us assume that the true coefficient is indeed zero (as could be in the case of the second regressor). Even in this case our point-guess will not necessarily be zero as well since we always have some degree of error in our estimates.
- To get around this point we standardize the estimated coefficient using the estimated standard error. In loose terms the standardized coefficient is reported in the fourth column. The benefit of using a standardized coefficient is that we can tell how likely it is for the specific value to occur. For example values above 1.96 will occur only 2.5% of the times and values below -1.96 will occur only 2.5% of the times.
- The fourth column gives us the standardized values while the fifth column will tell the p-value. The p-value answers the following question:
  - Under the assumption that the true β is equal zero, what is the probability of $\hat{\beta}$ occurring?
- The following output:

```
------------------------------------------------------------------------------
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      democA |   .0999448   1.018794     0.10   0.922    -1.911172   2.111062
------------------------------------------------------------------------------
```

  tells us that the standardized value of the coefficient is .10 and $\hat{\beta} = .0999448$ would occur with probability 0.92 if the true coefficient were indeed zero.
- We conclude that the true coefficient is zero and that `democA` is not statistically significant.

- The following output:

```
------------------------------------------------------------------------------
       voteA |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      shareA |   .4634151   .0151651    30.56   0.000      .433479    .4933512
------------------------------------------------------------------------------
```

  tells us that the standardized value of the coefficient is 30.56 and that this value has almost 0 probability of occurring if the true coefficient were indeed zero.