Introductory Econometrics

Lecture 11: Goodness of fit, estimation of σ^2

Instructor: Ma, Jun

Renmin University of China

April 26, 2023

Fitted values

- Consider the multiple regression model with k regressors: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i$.
- Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimators.
- The fitted (or predicted) by the model value of Y is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \ldots + \hat{\beta}_k X_{k,i}$.
- The residual is: $\hat{U}_i = Y_i - \hat{Y}_i$.
- Consider the average of $\hat{Y}$:

$$\begin{aligned}
\bar{\hat{Y}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{U}_i\right) \\
&= \bar{Y} - \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i = \bar{Y}
\end{aligned}$$

because when there is an intercept, $\sum_{i=1}^{n}\hat{U}_i = 0$.

Sum-of-Squares

- The total variation of Y in the sample is:

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \text{ (Total Sum-of-Squares).}$$

- The explained variation of Y in the sample is:

$$SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 \text{ (Explained or Model Sum-of-Squares).}$$

- The residual (unexplained or error) variation of Y in the sample is:

$$SSR = \sum_{i=1}^{n} \hat{U}_i^2 \text{ (Residual Sum-of-Squares).}$$

- If the regression contains an intercept:

$$SST = SSE + SSR.$$

Proof of $SST = SSE + SSR$

- First,

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i + \hat{U}_i - \bar{Y}\right)^2 \\
&= \sum_{i=1}^{n} \left(\left(\hat{Y}_i - \bar{Y}\right) + \hat{U}_i\right)^2 \\
&= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \hat{U}_i^2 + 2 \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) \hat{U}_i .
\end{aligned}
$$

- Next, we will show that $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) \hat{U}_i = 0$.

Proof of $SST = SSE + SSR$

- Since $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \ldots + \hat{\beta}_k X_{k,i}$,

$$\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) \hat{U}_i = \sum_{i=1}^{n} \left(\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \ldots + \hat{\beta}_k X_{k,i}\right) - \bar{Y}\right) \hat{U}_i$$
$$= \hat{\beta}_0 \sum_{i=1}^{n} \hat{U}_i + \hat{\beta}_1 \sum_{i=1}^{n} X_{1,i}\hat{U}_i + \ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{k,i}\hat{U}_i - \bar{Y} \sum_{i=1}^{n} \hat{U}_i.$$

- The OLS normal equations for a model with an intercept:

$$\sum_{i=1}^{n} \hat{U}_i = \sum_{i=1}^{n} X_{1,i}\hat{U}_i = \ldots = \sum_{i=1}^{n} X_{k,i}\hat{U}_i = 0.$$

- It follows that $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right) \hat{U}_i = 0$.

R^2

- Consider the following measure of goodness of fit:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2} \\ &= \frac{SSE}{SST} \\ &= 1 - \frac{SSR}{SST} \\ &= 1 - \frac{\sum_{i=1}^n \hat{U}_i^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}. \end{aligned}$$

- $0 \leq R^2 \leq 1$.
- R^2 measures the proportion of variation in Y in the sample explained by the X's.

R^2 is a non-decreasing function of the number of the regressors

- Consider two models:

$$\begin{aligned} Y_i &= \tilde{\beta}_0 + \tilde{\beta}_1 X_{1,i} + \tilde{U}_i, \\ Y_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \hat{U}_i. \end{aligned}$$

- We will show that

$$\sum_{i=1}^{n} \tilde{U}_i^2 \geq \sum_{i=1}^{n} \hat{U}_i^2$$

  and therefore R^2 that corresponds to the regression with one regressor is less or equal than R^2 that corresponds to the regression with two regressors.
- This can be generalized to the case of k and $k+1$ regressors.

Proof

- Consider
$$\sum_{i=1}^{n} \left(\tilde{U}_i - \hat{U}_i\right)^2 = \sum_{i=1}^{n} \tilde{U}_i^2 + \sum_{i=1}^{n} \hat{U}_i^2 - 2\sum_{i=1}^{n} \tilde{U}_i \hat{U}_i.$$
- We will show that
$$\sum_{i=1}^{n} \tilde{U}_i \hat{U}_i = \sum_{i=1}^{n} \hat{U}_i^2.$$
- Then,
$$0 \leq \sum_{i=1}^{n} \left(\tilde{U}_i - \hat{U}_i\right)^2 = \sum_{i=1}^{n} \tilde{U}_i^2 - \sum_{i=1}^{n} \hat{U}_i^2,$$
or
$$\sum_{i=1}^{n} \tilde{U}_i^2 \geq \sum_{i=1}^{n} \hat{U}_i^2.$$

Proof

$$
\begin{aligned}
\sum_{i=1}^{n} \tilde{U}_i \hat{U}_i &= \sum_{i=1}^{n} \left(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1,i} \right) \hat{U}_i \\
&= \sum_{i=1}^{n} Y_i \hat{U}_i - \tilde{\beta}_0 \sum_{i=1}^{n} \hat{U}_i - \tilde{\beta}_1 \sum_{i=1}^{n} X_{1,i} \hat{U}_i \\
&= \sum_{i=1}^{n} Y_i \hat{U}_i - \tilde{\beta}_0 \cdot 0 - \tilde{\beta}_1 \cdot 0 \\
&= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \hat{U}_i \right) \hat{U}_i \\
&= \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 0 + \sum_{i=1}^{n} \hat{U}_i^2 \\
&= \sum_{i=1}^{n} \hat{U}_i^2 .
\end{aligned}
$$

Adjusted R^2

- Since R^2 cannot decrease when more regressors are added, even if the additional regressors are irrelevant, an alternative measure of goodness-of-fit has been developed.
- Adjusted R^2: the idea is to adjust SSR and SST for degrees of freedom:
$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}.$$
- $\bar{R}^2 < R^2$.
- $\bar{R}^2$ can decrease when more regressors are added.

Estimation of σ^2

- In the multiple linear regression model, we can estimate $\sigma^2 = \mathrm{E}\left[U_i^2\right]$ as follows:
  Let
  $$\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \hat{\beta}_2 X_{2,i} - \ldots - \hat{\beta}_k X_{k,i}.$$
  An estimator for σ^2 is
  $$\begin{aligned} s^2 &= \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{U}_i^2 \\ &= \frac{SSR}{n-k-1}. \end{aligned}$$
- The adjustment $k+1$ is for the number of parameters we have to estimate in order to construct $\hat{U}$'s:
  $$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k.$$

Estimation of σ^2

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{U}_i^2.$$

- s^2 is an unbiased estimator of σ^2 (i.e., $\mathrm{E}\left[s^2\right] = \sigma^2$) when the following conditions hold:
  1. $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + U_i$.
  2. Conditional on X's, $\mathrm{E}\left[U_i\right] = 0$ for all i's.
  3. Conditional on X's, $\mathrm{E}\left[U_i^2\right] = \sigma^2$ for all i's (homoskedasticity).
  4. Conditional on X's $\mathrm{E}\left[U_i U_j\right] = 0$ for all $i \neq j$.

Stata

```
. regress rent avginc pop enroll

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) =   29.05
       Model |  368241.042     3  122747.014           Prob > F      =  0.0000
    Residual |  253521.396    60  4225.35659           R-squared     =  0.5923
-------------+------------------------------           Adj R-squared =  0.5719
       Total |  621762.438    63  9869.24504           Root MSE      =  65.003

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   .0119416    .001318     9.06   0.000     .0093052     .014578
         pop |  -.0003538   .0001621    -2.18   0.033    -.0006781   -.0000296
      enroll |   .0025595    .001264     2.02   0.047     .0000311    .0050879
       _cons |    120.772   34.53081     3.50   0.001     51.70009    189.8439
------------------------------------------------------------------------------
```

- We have 64 observations ($n = 64$) and 3 regressors ($k = 3$).
- SSE is displayed under Model SS (Sum of Squares): 368241.042.
- The Model df (degrees of freedom) is $k = 3$.
- The Model MS (Mean Squares) is $SSE/k = 368241.042/3 = 122747.014$.

Stata

```
. regress rent avginc pop enroll

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) =   29.05
       Model |  368241.042     3  122747.014           Prob > F      =  0.0000
    Residual |  253521.396    60  4225.35659           R-squared     =  0.5923
-------------+------------------------------           Adj R-squared =  0.5719
       Total |  621762.438    63  9869.24504           Root MSE      =  65.003

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   .0119416    .001318     9.06   0.000     .0093052     .014578
         pop |  -.0003538   .0001621    -2.18   0.033    -.0006781   -.0000296
      enroll |   .0025595    .001264     2.02   0.047     .0000311    .0050879
       _cons |    120.772   34.53081     3.50   0.001     51.70009    189.8439
------------------------------------------------------------------------------
```

- SSR is displayed under Residual SS: 253521.396.
- The Residual df is $n - k - 1 = 64 - 3 - 1 = 60$.
- The Residual MS is $SSR/(n - k - 1) = s^2$.
- The Residual MS is $253521.396/60 = 4225.35659$.

Stata

```
. regress rent avginc pop enroll

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) =   29.05
       Model |  368241.042     3  122747.014           Prob > F      =  0.0000
    Residual |  253521.396    60  4225.35659           R-squared     =  0.5923
-------------+------------------------------           Adj R-squared =  0.5719
       Total |  621762.438    63  9869.24504           Root MSE      =  65.003

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   .0119416    .001318     9.06   0.000     .0093052     .014578
         pop |  -.0003538   .0001621    -2.18   0.033    -.0006781   -.0000296
      enroll |   .0025595    .001264     2.02   0.047     .0000311    .0050879
       _cons |    120.772   34.53081     3.50   0.001     51.70009    189.8439
------------------------------------------------------------------------------
```

- SST is displayed under Total SS: 621762.438.
- The Total df is $n - 1 = 64 - 1 = 63$.
- The Total MS is $SST/(n-1) = 621762.438/63 = 9869.24504$.

Stata

```
. regress rent avginc pop enroll

      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) =   29.05
       Model |  368241.042     3  122747.014           Prob > F      =  0.0000
    Residual |  253521.396    60  4225.35659           R-squared     =  0.5923
-------------+------------------------------           Adj R-squared =  0.5719
       Total |  621762.438    63  9869.24504           Root MSE      =  65.003

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   .0119416    .001318     9.06   0.000     .0093052     .014578
         pop |  -.0003538   .0001621    -2.18   0.033    -.0006781   -.0000296
      enroll |   .0025595    .001264     2.02   0.047     .0000311    .0050879
       _cons |    120.772   34.53081     3.50   0.001     51.70009    189.8439
------------------------------------------------------------------------------
```

- $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{253521.396}{621762.438} = 0.5923.$
- $\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{253521.396/60}{621762.438/63} = 0.5719.$
- Root MSE (Mean Squared Error) is $s = \sqrt{s^2} = \sqrt{4225.35659} = 65.003.$