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 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 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} :

$$\bar{\hat{Y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i$$
$$= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{U}_i)$$
$$= \bar{Y} - \frac{1}{n} \sum_{i=1}^{n} \hat{U}_i = \bar{Y}$$

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:

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

• The explained variation of *Y* in the sample is:

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

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

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

► If the regression contains an intercept:

$$SST = SSE + SSR.$$

Proof of SST = SSE + SSR

► First,

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

=
$$\sum_{i=1}^{n} (\hat{Y}_i + \hat{U}_i - \bar{Y})^2$$

=
$$\sum_{i=1}^{n} ((\hat{Y}_i - \bar{Y}) + \hat{U}_i)^2$$

=
$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{U}_i^2 + 2\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) \hat{U}_i.$$

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

Proof of SST = SSE + SSR

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

$$\sum_{i=1}^{n} (\hat{Y}_{i} - \bar{Y}) \hat{U}_{i} = \sum_{i=1}^{n} ((\hat{\beta}_{0} + \hat{\beta}_{1}X_{1,i} + \dots + \hat{\beta}_{k}X_{k,i}) - \bar{Y}) \hat{U}_{i}$$
$$= \hat{\beta}_{0} \sum_{i=1}^{n} \hat{U}_{i} + \hat{\beta}_{1} \sum_{i=1}^{n} X_{1,i} \hat{U}_{i} + \dots + \hat{\beta}_{k} \sum_{i=1}^{n} X_{k,i} \hat{U}_{i} - \bar{Y} \sum_{i=1}^{n} \hat{U}_{i}.$$

► 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 $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) \hat{U}_i = 0.$

 R^2

Consider the following measure of goodness of fit:

$$R^{2} = \frac{\sum_{i=1}^{n} (\hat{Y}_{i} - \bar{Y})^{2}}{\sum_{i=1}^{n} (Y_{i} - \bar{Y})^{2}}$$

= $\frac{SSE}{SST}$
= $1 - \frac{SSR}{SST}$
= $1 - \frac{\sum_{i=1}^{n} \hat{U}_{i}^{2}}{\sum_{i=1}^{n} (Y_{i} - \bar{Y})^{2}}.$

- ▶ $0 \le R^2 \le 1$.
- *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:

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

► We will show that

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

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,

$$0 \le \sum_{i=1}^{n} \left(\tilde{U}_{i} - \hat{U}_{i} \right)^{2} = \sum_{i=1}^{n} \tilde{U}_{i}^{2} - \sum_{i=1}^{n} \hat{U}_{i}^{2},$$

or

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

Proof

$$\begin{split} \sum_{i=1}^{n} \tilde{U}_{i} \hat{U}_{i} &= \sum_{i=1}^{n} \left(Y_{i} - \tilde{\beta}_{0} - \tilde{\beta}_{1} X_{1,i} \right) \hat{U}_{i} \\ &= \sum_{i=1}^{n} Y_{i} \hat{U}_{i} - \tilde{\beta}_{0} \sum_{i=1}^{n} \hat{U}_{i} - \tilde{\beta}_{1} \sum_{i=1}^{n} X_{1,i} \hat{U}_{i} \\ &= \sum_{i=1}^{n} Y_{i} \hat{U}_{i} - \tilde{\beta}_{0} \cdot 0 - \tilde{\beta}_{1} \cdot 0 \\ &= \sum_{i=1}^{n} \left(\hat{\beta}_{0} + \hat{\beta}_{1} X_{1,i} + \hat{\beta}_{2} X_{2,i} + \hat{U}_{i} \right) \hat{U}_{i} \\ &= \hat{\beta}_{0} \cdot 0 + \hat{\beta}_{1} \cdot 0 + \hat{\beta}_{2} \cdot 0 + \sum_{i=1}^{n} \hat{U}_{i}^{2} \\ &= \sum_{i=1}^{n} \hat{U}_{i}^{2}. \end{split}$$

Adjusted R^2

- Since R² 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 R²: the idea is to adjust SSR and SST for degrees of freedom:

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

 $\blacktriangleright \ \bar{R}^2 < R^2.$

• \bar{R}^2 can decrease when more regressors are added.

Estimation of σ^2

► In the multiple linear regression model, we can estimate $\sigma^2 = E\left[U_i^2\right]$ 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

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

• 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., $E[s^2] = \sigma^2$) 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, $E[U_i] = 0$ for all *i*'s.
 - 3. Conditional on X's, E $[U_i^2] = \sigma^2$ for all *i*'s (homoskedasticity).
 - 4. Conditional on *X*'s $\mathbb{E}\left[U_i U_j\right] = 0$ for all $i \neq j$.

df Number of obs = Source | SS MS F(3, 60) = 29.05 368241.042 3 122747.014 Proh > F= 0.0000 Model | Residual | 253521.396 60 4225.35659 R-squared = 0.5923 Adj R-squared = 0.5719 -------621762.438 63 9869.24504 Root MSE Total | 65.003 Coef. Std. Err. t P>|t| [95% Conf. Interval] rent | .0119416 .001318 9.06 0.000 .0093052 avginc | .014578 pop | -.0003538 .0001621 -2.18 0.033 -.0006781 -.0000296.001264 2.02 0.047 .0000311 enroll | .0025595 .0050879 _cons | 120.772 34.53081 3.50 0.001 51,70009 189 8439 -------

. regress rent avginc pop enroll

- 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.

64

SS	df	MS		Number of $obs = 66$ F(3, 60) = 29.01						
368241.042 253521.396	3 1227	47.014		Prob > F = 0.000 R-squared = 0.592						
	63 9869	.24504		Adj R-squared = 0.571 Root MSE = 65.00						
Coef.			P> t	[95% Conf. Interval						
.0119416 0003538 .0025595 120.772	.001318 .0001621 .001264 34.53081	9.06 -2.18 2.02 3.50	0.000 0.033 0.047 0.001	.0093052 .01457 0006781000029 .0000311 .005087 51.70009 189.843						
	SS 368241.042 253521.396 621762.438 Coef. .0119416 0003538 .0025595	SS df 368241.042 3 1227 253521.396 60 4225 621762.438 63 9865 Coef. Std. Err. .0119416 .001318 0003538 .0001621 .0025595 .001264	SS df MS 368241.042 3 122747.014 253521.396 60 4225.35659 621762.438 63 9869.24504 Coef. Std. Err. t .0119416 .001318 9.06 .0003538 .0001621 -2.18 .0025595 .001264 2.02	SS df MS 368241.042 3 122747.014 253521.396 60 4225.35659 621762.438 63 9869.24504 Coef. Std. Err. t P> t .0119416 .001318 9.06 0.000 0003538 .0001621 -2.18 0.033 .0025595 .001264 2.02 0.047						

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

regress rent avging pop enroll

► The Residual MS is 253521.396/60 = 4225.35659.

df Source | SS MS Number of obs = 64 F(3. 60) = 29.05 Model | 368241.042 3 122747.014 Prob > F0.0000 = 253521.396 60 4225.35659 R-squared = 0.5923Residual | Adj R-squared = 0.5719Total | 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-.0003538.0001621-2.180.033-.0006781coll.0025595.0012642.020.047.0000311 -.0000296 enroll | .0050879 _cons | 3.50 120.772 34.53081 0.001 51,70009 189.8439

. regress rent avginc pop enroll

- ► 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.

. regress rent avginc pop enroll

Source	SS	df	MS		Number of obs : F(3, 60) :	
Model Residual	368241.042 253521.396	3 122 60 422	747.014 5.35659		Prob > F :	= 0.0000 = 0.5923
Total	621762.438		9.24504			= 65.003
rent	Coef.	Std. Err.		P> t	[95% Conf.]	-
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 $s = \sqrt{s^2} = \sqrt{4225.35659} = 65.003.$