Printout

Thursday, March 24, 2016 3:51 PM

Lecture 17

Properties of OLS Estimators (Cont:)

(6)
$$var(\hat{\beta}_2) = \frac{1}{\sum x_{2i}^2} \cdot \frac{\sigma^2}{(1 - r_{23}^2)}$$

$$var(\hat{\beta}_3) = \frac{1}{\sum x_{3i}^2} \cdot \frac{\sigma^2}{(1 - r_{23}^2)}$$

- GIVEN $\sum x_{2i}^2$ AND σ^2, WHEN $r_{23} \uparrow \rightarrow var(\hat{\beta}_2) \uparrow \rightarrow$ ACCURACY OF ESTIMATION $\downarrow$:(
  IF $r_{23}^2 = 1$, THEN $var(\hat{\beta}_2)$ BECOMES INFINITE.
- GIVEN $\sum x_{3i}^2$ AND σ^2, WHEN $r_{23} \uparrow \rightarrow var(\hat{\beta}_3) \uparrow \rightarrow$ ACCURACY OF ESTIMATION $\downarrow$:(
  IF $r_{23}^2 = 1$, THEN $var(\hat{\beta}_3)$ BECOMES INFINITE.

(7) $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 \rightarrow$ ARE "BLUE" [= THEY SATISFY GAUSS-MARKOV THEOREM]

The Multiple Coefficient of Determination R^2 and the Multiple Coefficient of Correlation R

$R^2 = \frac{ESS}{TSS}$

In this section, we will study how to measure the proportion of the variation in Y explained by the variables X_2 and X_3 jointly. This is the same concept of r^2 that we have learned before.

The quantity that gives this information is known as the **the multiple coefficient of determination** and is denoted by R^2.

To derive R^2, we firstly write down the following equation:

$$\begin{aligned} Y_i &= \overbrace{\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}}^{=\hat{Y}_i} + \hat{u}_i \\ &= \hat{Y}_i + \hat{u}_i \end{aligned} \qquad (k=3) \qquad \text{(Eq.10)}$$

where $\hat{Y}_i$ is the estimated value of Y_i from the fitted regression line and is an estimator of true $E(Y_i|X_{2i}, X_{3i})$.

~~??~~ Eq 10 may be written as

$$\begin{aligned} \overbrace{y_i}^{(Y_i - \bar{Y})} &= \hat{\beta}_2 \overbrace{x_{2i}}^{= X_{2i} - \bar{X}_2} + \hat{\beta}_3 \overbrace{x_{3i}}^{= X_3 - \bar{X}_3} + \hat{u}_i \\ &= \hat{y}_i + \hat{u}_i \end{aligned} \qquad \text{(DEVIATION FORM)} \qquad \text{(Eq.11)}$$

Squaring ~~??~~ Eq 11 on both sides and summing over the sample values, we obtain

$$\boxed{\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i{}^2 + \hat{u}_i{}^2 + 2\overbrace{\cancel{\sum \hat{y}_i \hat{u}_i}}^{=0} \\ \Sigma y_i^2 &= \sum \hat{y}_i{}^2 + \sum \hat{u}_i{}^2 \\ \underbrace{\Sigma (Y_i - \bar{Y})^2}_{(TSS)} &= \underbrace{\Sigma (\hat{Y}_i - \bar{Y})^2}_{(ESS)} + \underbrace{\Sigma (Y_i - \hat{Y}_i)^2}_{(RSS)} \end{aligned}} \qquad \text{(Eq.12)}$$

FROM $y_i = \hat{y}_i + \hat{u}_i$

$$= \hat\beta_2 x_{2i} + \hat\beta_3 x_{3i} + \hat u_i$$

$$\hat u_i = y_i - \hat\beta_2 x_{2i} - \hat\beta_3 x_{3i}$$

$$\sum \hat u_i^2 = \sum \hat u_i (y_i - \hat\beta_2 x_{2i} - \hat\beta_3 x_{3i})$$

$$= \sum \hat u_i y_i - \hat\beta_2 \underbrace{\cancel{\sum \hat u_i x_{2i}}}_{=0} - \hat\beta_3 \underbrace{\cancel{\sum \hat u_i x_{3i}}}_{=0}$$

$$= \sum y_i (y_i - \hat\beta_2 x_{2i} - \hat\beta_3 x_{3i})$$

SO, $\boxed{\sum \hat u_i^2 = \sum y_i^2 - \hat\beta_2 \sum x_{2i} y_i - \hat\beta_3 \sum x_{3i} y_i}$

FROM $\sum y_i^2 = \sum \hat y_i^2 + \sum \hat u_i^2$ → (EQ. 12)

$$\cancel{\sum y_i^2} = \sum \hat y_i^2 + \cancel{\sum y_i^2} - \hat\beta_2 \sum x_{2i} y_i - \hat\beta_3 \sum x_{3i} y_i$$

SO $\sum \hat y_i^2 = \hat\beta_2 \sum x_{2i} y_i + \hat\beta_3 \sum x_{3i} y_i$ ⟹ ESS IN OTHER EXPRESSION!

$$R^2 = \frac{ESS}{TSS} = \frac{\sum \hat y_i^2}{\sum y_i^2}$$

$$R^2 = \frac{\hat\beta_2 \sum y_i x_{2i} + \hat\beta_3 \sum y_i x_{3i}}{\sum y_i^2}$$

→ ANOTHER VERSION OF R^2

(Eq.13)

$$var(\hat\beta_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$var(\hat\beta_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

CORRELATION COEFFICIENT BET. X_{2i} AND X_{3i}

The three-or-more-variable analogue of r is the coefficient of multiple correlation, denoted by R, and it is a measure of the degree of association between Y and all the explanatory variables jointly. Although r can be positive or negative, R is always taken to be positive.

$$Var(\hat\beta_j) = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1 - R_j^2}\right)$$

R_j^2 IS THE R^2 IN THE REGRESSION OF X_j ON THE REMAINING $(k-2)$ REGRESSORS.

Y_i x_{2i}

$-1-1$

Ex: $\underbrace{X_{2i} = \hat\alpha_1 + \hat\alpha_2 X_{3i}}_{\text{YOU GOT } R^2}$

7.3 R^2 and the Adjusted R^2

It should be noted that the R^2 is a nondecreasing function of the number of explanatory variables. Thus, when the number of regressors increases, R^2 almost invariably increases and never decreases. **In other words, an additional X variable will not decrease** R^2**!**

To explain this fact, let we write down the definition of R^2 again:

$$
\begin{aligned}
R^2 &= \frac{ESS}{TSS} \\
&= 1 - \frac{RSS}{TSS} \\
&= 1 - \frac{\sum \hat{u}_i{}^2}{\sum y_i^2}
\end{aligned}
\qquad \text{(Eq.14)}
$$

Therefore, in comparing two regression models **with the same dependent variable but differing number of X variables**, one should be very wary of choosing the model with the highest R^2.

In light of comparing two R^2 terms, we have to take into account the number of X variables present in the model. To achieve this goal , we can consider the alternative coefficient of determination, which is as follows:

$\hat\beta_1, \hat\beta_2, \hat\beta_3$

$$\bar R^2 = 1 - \frac{\sum \hat u_i^2 / (n-3)}{\sum y_i^2 (n-1)}$$

FOR 3-VARIABLE REGRESSION MODEL $(k=3)$

THE ADJUSTED R^2

$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-3)}{\sum y_i^2 (n-1)}$ FOR 3-VARIABLE REGRESSION MODEL ($k=3$)

THE ADJUSTED R^2

FOR $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots \beta_k X_{ki} + u_i$, $\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}$

k= the number of parameters in the model including the intercept term.
n= the number of observations in the sample data.

The above equation is known as **the adjusted** R^2, denoted by $\bar{R}^2$. The term adjusted means adjusted for the df associated with the sums of squares entering into EQ(14)

We can rewrite the the adjusted R^2 as:

IS VARIANCE OF RESIDUAL TERM OR RESIDUAL VARIANCE

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2}$$

WHERE $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{(n-k)}$ (UNBIASED ESTIMATOR FOR σ^2)

$S_Y^2 = \frac{\sum y_i^2}{(n-1)}$ (SAMPLE VARIANCE OF Y)

We can also get the equation which shows the relationship between $\bar{R}^2$ and R^2:

RECALL THAT $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2}$ —— (1)

$\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}$ —— (2)

SUBSTITUTING (1) INTO (2) GIVES

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

Besides R^2 and $\bar{R}^2$ as goodness of fit measures, other criteria are often used to judge the adequacy of a regression model. Two of these are **Akaike's Information criterion and Amemiya's Prediction criteria**, which are used to select between competing models. We will discuss these criteria in greater detail later.

(1) IF WE ADD MORE EXPLANATORY VARIABLES, R^2 WILL ALWAYS INCREASE BUT NOT NECESSARILY FOR $\bar{R}^2$!!!

SEE: FROM $\boxed{\bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}}$

$k \uparrow \rightarrow (n-k) \downarrow \rightarrow \underbrace{\frac{\sum \hat{u}_i^2}{(n-k)}}_{=A} \uparrow \rightarrow \bar{R}^2 = 1 - \frac{A \uparrow}{B} \Downarrow$

$\rightarrow \sum \hat{u}_i^2 \downarrow \longrightarrow \bar{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)} \Uparrow$

(2) FOR $k > 1$, $\bar{R}^2 < R^2$.

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

FOR $k > 1$, $\frac{(n-1)}{(n-k)} > 1$

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

Lecture 18
CHAPTER 8: Multiple Regression Analysis: The Problem of Inference

In this chapter, we will extend the ideas of interval estimation and hypothesis testing developed there to models involving three or more variables.

CHAPTER 8: Multiple Regression Analysis: The Problem of Inference

In this chapter, we will extend the ideas of interval estimation and hypothesis testing developed there to models involving three or more variables.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

We have already known that if our objective is to do interval estimation and hypothesis testing, we need to assume that the u_i follow the normal distribution with zero mean and constant variance σ^2

With the normality assumption and the CLRM assumptions, we know that:

[1] The OLS estimations of partial regression coefficients are best linear unbiased estimators (BLUE).

[2] The estimators $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are normally distributed with means equal to true β_1,β_2, and β_3 and variances are following:

$$var(\hat{\beta}_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2\bar{X}_3 \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i}x_{3i})^2}\right] * \sigma^2$$

$$se(\hat{\beta}_1) = +\sqrt{var(\hat{\beta}_1)}$$

$$var(\hat{\beta}_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} * \sigma^2$$

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\underbrace{\sum x_{2i}^2}_{\sum (X_{2i} - \bar{X}_2)^2}(1 - r_{23}^2)}$$

$$se(\hat{\beta}_2) = +\sqrt{var(\hat{\beta}_2)}$$

$$var(\hat{\beta}_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i}x_{3i})^2} * \sigma^2$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\underbrace{\sum x_{3i}^2}_{\sum (X_{3i} - \bar{X}_3)^2}(1 - r_{23}^2)}$$

$$se(\hat{\beta}_3) = +\sqrt{var(\hat{\beta}_3)}$$

	X_3 HIGH	X_3 LOW
X_2 HIGH	✓	✓
X_2 LOW	✓	✓

THE HIGHER THE $\sum x_{2i}^2$, THE LOWER THE $var(\hat{\beta}_2)$, GIVEN σ^2 AND r_{23}^2

" ——— " $\sum x_{3i}^2$, " ——— $var(\hat{\beta}_3)$, " ——— ".

FOR $k > 1$, $\frac{(n-1)}{(n-k)} > 1$

SO, $\bar{R}^2 < R^2$

③ $\bar{R}^2$ CAN BE NEGATIVE

EX: WHEN $R^2 = 0$

④ WHEN COMPARING TWO R-SQUARE VALUES, MAKE SURE THAT

① THE SAMPLE SIZE MUST BE THE SAME.

\+

② THE DEPENDENT VARIABLE MUST BE THE SAME.

EX) $\ln Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ ①

$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$ ②

$\beta_2 = \frac{d \ln Y}{d X_{2i}} = \frac{\Delta Y / Y}{\Delta X}$ (LOG-LIN MODEL)

$\alpha_2 = \frac{dY}{dX}$ (LINEAR MODEL)

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

Moreover, $\frac{(n-3)\hat{\sigma}^2}{\sigma^2}$ follows the χ^2 distribution with n-3 df. We can also show that, if we replace the true σ^2 by its unbiased estimator $\hat{\sigma}^2$ in the computation of the standard errors, we then get

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{se(\hat{\beta}_3)}$$

follows the t distribution with $\boxed{\text{n-3 df.}}$

Example Consider the following regression: $\log(salary) = \beta_1 + \beta_2 \log(sales) + \beta_3 ROE + \beta_4 ROS + u$

$$\widehat{\log(salary)} = 4.32 + \underset{\hat{\beta}_2}{0.280}\ \log(\text{sales}) + \underset{\hat{\beta}_3}{0.0174}\ \text{ROE} + \underset{\hat{\beta}_4}{0.00024}\ \text{ROS}$$

$$se = (0.32)\ (0.035)\ (0.0041)\ (0.00054) \qquad \text{(Eq.15)}$$

$n = 209$

$$R^2 = 0.283$$

where
salary =salary of CEO
sales=annual firm sales
ROE=return on equity in percent
ROS =return on firm's stock

interpreter the partial regression coefficients

$\hat{\beta}_2 = 0.28 \Rightarrow$ THE ELASTICITY OF CEO SALARY WRT. SALES

CETERIS PARIBUS, IF SALES RISE BY 1 PERCENT, ON AVERAGE,

CEO SALARY WOULD RISE BY 0.28 PERCENT

$\beta_2 = 0.28 \Rightarrow$ THE ELASTICITY OF CEO SALARY W.R.T. SALES, CETERIS PARIBUS, IF SALES RISE BY 1 PERCENT, ON AVERAGE, CEO SALARY WOULD RISE BY 0.28 PERCENT.

$\hat{\beta}_3 = 0.0174 \Rightarrow$ IF ROE RISES BY 1 UNIT, CEO SALARY WOULD RISE BY $100 \cdot 0.0174 = 1.74$ PERCENT

$$\left(= \frac{d \ln(salary)}{d\, ROE} \right) \quad \text{LIKE} \quad \frac{\frac{dY}{Y}}{dX} \rightarrow \frac{\frac{dY}{Y} \cdot 100}{dX} \quad \frac{\%\Delta Y}{\Delta X}$$

$\hat{\beta}_4 = 0.00024 \Rightarrow$ IF ROS RISES BY 1 UNIT, (= 1 PERCETAGE POINT) CEO SALARY WOULD RISE BY $100 \cdot 0.00024$ $= 0.024$ PERCENT.

Questions What about the statistical significance of the observed results?

For the coefficient of log(sales) of 0.280, Is this coefficient statistically significant different from zero?

For the coefficient of ROE of 0.0174, Is this coefficient statistically significant different from zero?

For the coefficient of ROS of 0.00024, Is this coefficient statistically significant different from zero?

Are these three coefficients statistically significant?
To answer these questions, we have to learn the kinds of hypothesis testing.

8.1 Hypothesis Testing About Individual Regression Coefficients

We can use the t-test to test a hypothesis about any individual partial regression coefficient.

8.1.1 Two-tail test:

Let us postulate that

① $H_0: \beta_2 = 0$ (SALES DOES NOT AFFECT CEO SALARY)
$H_1: \beta_2 \neq 0$ (SALES DOES AFFECT CEO SALARY)

② SET LEVEL OF SIGNIFICANCE $(\alpha) = 0.05$

③ COMPUTE t-STATISTIC :

$$\hat{t} = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{0.28 - 0}{0.035} = 8.00$$

④ FIND CRITICAL t :

$n = 209$
$df = n - k$ ← NUMBER OF EXPLANATORY VARIABLES INCLUDING INTERCEPT TERM
$= 209 - 4$
$= 205$

$$t_{\frac{\alpha}{2}, n-k} = t_{\frac{0.05}{2}, 205} = 1.96$$

⑤ COMPARE $\hat{t}$ WITH $t_{CRITICAL}$:

SINCE $\hat{t} > t_{CRITICAL}$, WE MAY REJECT THE NULL HYPOTHESIS THAT $H_0: \beta_2 = 0$. IN OTHER WORDS $\hat{\beta}_2$ IS STATISTICALLY SIGNIFICANT DIFFERENT FROM ZERO AT 95% !

OLS

DEPENDENT VARIABLE : log(salary)

EXPLANATORY VARIABLES : log(sales), ROE, ROS

Tuesday, March 29, 2016 6:35 PM

$n - k = 209 - 4 = 205$

```
. reg lsalary lsales roe ros
```

Source	SS	df	MS
Model	18.8613439 (ESS)	3	6.28711463
Residual	47.8608193 (RSS)	205	.233467411
Total	66.7221632 (TSS)	208	.320779631

Number of obs	=	209
F(3, 205)	=	26.93
Prob > F	=	0.0000
R-squared	=	0.2827
Adj R-squared	=	0.2722
Root MSE	=	.48318

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lsales	.2803149	.03532	7.94	0.000	.2106778	.349952
roe	.0174168	.0040923	4.26	0.000	.0093484	.0254852
ros	.0002417	.0005418	0.45	0.656	-.0008266	.0013099
_cons	4.311712	.3154329	13.67	0.000	3.689804	4.933621

$\hat\beta_2$ $\hat\beta_3$ $\hat\beta_4$

$$\hat t = \frac{\text{Coef.}}{\text{Std. Err.}}$$