Inference in Simple Regression 2

Class 7

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang Sik Kim's lecture notes.

Coefficient of Determination

How Good Is the Fitted Regression Line?

- So far, we were concerned with the problem of estimating regression coefficients.

- We now consider the goodness of fit of the fitted regression line to data.
  - That is, we will find out how "well" the sample regression line $(\hat{Y}_i)$ fits the data (Y_i).
  - **Question:** Is the variation in the dependent variable largely explained by the variation in the independent variable ?
  - If yes, we have a "good fit"!!

How Good Is the Fitted Regression Line? [cont'd]

$$
\begin{aligned}
Y_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{e}_i \\
&= \hat{Y}_i + \hat{e}_i \\
\Rightarrow \quad Y_i - \bar{Y} &= \hat{Y}_i - \bar{Y} + \hat{e}_i
\end{aligned}
$$

- We want to know whether $Y_i - \bar{Y}$ (variation in Y) is largely explained by $\hat{Y}_i - \bar{\hat{Y}}$ (variation in $\hat{Y}$) or not.
- Note:

$$
\begin{aligned}
Y_i &= \hat{Y}_i + \hat{e}_i \\
\Rightarrow \quad \sum Y_i &= \sum \hat{Y}_i + \sum \hat{e}_i \\
\Rightarrow \quad \frac{1}{n} \sum Y_i &= \frac{1}{n} \sum \hat{Y}_i \\
\Rightarrow \quad \bar{Y} &= \bar{\hat{Y}}
\end{aligned}
$$

How Good Is the Fitted Regression Line? [cont'd]

- Therefore,

$$\left(Y_i - \bar{Y}\right) = \left(\hat{Y}_i - \bar{Y}\right) + \hat{e}_i$$

  - $Y_i - \bar{Y}$: variation in Y_i around its mean
  - $\hat{Y}_i - \bar{Y} \left(= \hat{Y}_i - \bar{\hat{Y}}\right)$: variation in Y_i explained by X_i around its mean
  - $\hat{e}_i$: variation in Y_i not explained by X_i

- For a **"good"** fit, $\hat{Y}_i - \bar{Y}$ should have **"big"** proportion. Then, what would be an overall measure of fit?
  - Consider

$$\frac{\sum\left(\hat{Y}_i - \bar{Y}\right)}{\sum\left(Y_i - \bar{Y}\right)}$$

  - We **cannot** use the above because it has zeros in both numerator and denominator.

Coefficient of Determination (R^2)

$$\begin{aligned}\sum\left(Y_i-\bar{Y}\right)^2 &= \sum\left(\hat{Y}_i-\bar{Y}+\hat{e}_i\right)^2 \\ &= \sum\left(\hat{Y}_i-\bar{Y}\right)^2+\sum\hat{e}_i^2+2\sum\left(\hat{Y}_i-\bar{Y}\right)\hat{e}_i \\ &= \sum\left(\hat{Y}_i-\bar{Y}\right)^2+\sum\hat{e}_i^2 \quad \text{(Why?)}\end{aligned}$$

- Denote:
  - $\sum\left(Y_i-\bar{Y}\right)^2$: **TSS (Total Sum of Squares)**, total variation of Y
  - $\sum\left(\hat{Y}_i-\bar{Y}\right)^2$: **ESS (Explained Sum of Squares)**, total variation of $\hat{Y}$
  - $\sum\hat{e}_i^2$: **RSS (Residual Sum of Squares)**, total unexplained variation of Y

$$TSS = ESS + RSS$$

Coefficient of Determination (R^2) [cont'd]

- Coefficient of Determination (R^2):

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Note:

$$0 \leq R^2 \leq 1$$

  - $R^2 = 0$: $ESS = 0$, which means $\hat{Y}_i - \bar{Y} = 0$.
    - Variation in X does not help predicting variation in Y
    - There is no relationship between the regressand and the regressor (*i.e.* $\hat{\beta}_2 = 0$).
  - $R^2 = 1$: $RSS = 0$, which means $\hat{e}_i = 0$.
    - It means a perfect fit, *i.e.* all data lie on SRF.

Coefficient of Determination (R^2) [cont'd]

- R^2 just compares the values of the $\left(\hat{Y}_i - \bar{Y}\right)$'s to the $\hat{e}_i$'s.
- R^2 is just a descriptive statistic.
  - R^2 does **never** measures the quality of regressions.
  - It is **never** objective of regression to increase R^2.
- The values of R^2 can be easily manipulated.
  - For example, adding any regressors in the regression will increase R^2, which is meaningless.

Example: Food Expenditure and Income

$$\widehat{\text{food\_exp}} = \underset{(43.410)}{83.4160} + \underset{(2.0933)}{10.2096}\,\text{income}$$

$$T = 40 \qquad R^2 = 0.3850$$

(standard errors in parentheses)

- Variation of income about its mean explains about 38.5% of the variation of food expenditure in the linear regression model.

R^2 and Correlation Coefficient

- **Sample correlation coefficient**: Using sample analogues of covariance and variances, the sample correlation coefficient is given by

$$r_{X,Y} = \frac{\frac{1}{n-1}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2} \cdot \sqrt{\frac{1}{n-1}\sum\left(Y_i - \bar{Y}\right)^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$$

where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$.

- The sample correlation coefficient has a value between -1 and 1.
  - It measures the strength of the linear association.
  - The sign of $r_{X,Y}$ is the same as that of OLS estimator in the linear regression model.

R^2 and Correlation Coefficient [cont'd]

- We can show that

$$R^2 = r^2_{X,Y}$$

(Why?)

$$\begin{aligned} ESS &= \sum \left(\hat{Y}_i - \bar{Y}\right)^2 = \sum \left(\hat{\beta}_1 + \hat{\beta}_2 X_i - \hat{\beta}_1 - \hat{\beta}_2 \bar{X}\right)^2 \\ &= \hat{\beta}_2^2 \sum \left(X_i - \bar{X}\right)^2 = \hat{\beta}_2^2 \sum x_i^2 \\ &= \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)^2 \sum x_i^2 = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2} \end{aligned}$$

$$\implies \quad R^2 = \frac{ESS}{TSS} = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} = r^2_{X,Y}$$

- R^2 can be thought as a measure of the strength of the **"linear"** relationship between two variables X and Y.

Functional Forms of Regression Models

Functional Forms of Regression Models

- Usually, a linear model implies **"linear in parameters"** in most cases.
  - In this sense, the linear regression models are not necessarily linear in variables.

- Variables (both dependent and independent) can be transformed in any convenient way (e.g. take logs, the reciprocal of data, etc.)

- Transformation in variables should be based on economic theories and models.

- In particular, we discuss the following regression models:

  1. Log-Linear Model
  2. Semi-Log Model
  3. Reciprocal Model

Log-Linear Model

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + e_i$$

- One attractive feature of the log-linear model is that the slope coefficient β_2 measures **the elasticity of Y with respect to X**, that is, the percentile change in Y for given small percentile change in X.

$$\begin{aligned}
\beta_2 &= \frac{d \ln Y}{d \ln X} \\
&= d \ln Y \cdot \frac{dY}{dY} \cdot \frac{dX}{dX} \cdot \frac{1}{d \ln X} \\
&= \frac{d \ln Y}{dY} \cdot dY \cdot \frac{dX}{d \ln X} \cdot \frac{1}{dX} \\
&= \frac{\frac{d \ln Y}{dY} \cdot dY}{\frac{d \ln X}{dX} \cdot dX} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} \\
&= \frac{\%\ \text{change in } Y}{\%\ \text{change in } X} = \text{Elasticity of } Y \text{ w.r.t. } X
\end{aligned}$$

Log-Linear Model [cont'd]

- **Example:**

$$\widehat{\ln Y_i} = 0.7774 - 0.2530 \ln X_i$$

  - Y: Coffee consumption, cups per person a day
  - X: Real price of coffee, dollars per pound

- The price elasticity of coffee demand is -0.25!
  - That is, for 1% increase in the real price of coffee, the demand for coffee on the average decreases by about 0.25%.

Other Functional Forms

- Semi-Log model
  - $\ln Y_i = \beta_1 + \beta_2 X_i + e_i$ (Log-Lin model)
  - $Y_i = \beta_1 + \beta_2 \ln X_i + e_i$ (Lin-Log model)
- Reciprocal model
  - $Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + e_i$

Choice of Functional Form

- We discussed several functional forms an empirical model can assume within the confines of linear (“linear-in-parameter”) models.
- It is important that we choose an appropriate model for empirical estimation.
  - The underlying theory (e.g. consumption theory, Philips curve, etc.) may suggest a particular functional form.
- In most cases, a simple linear model can be the best specification.
  - Nevertheless, be sure you are able to justify the functional form you have chosen.
  - For example, spend some time examining the sensitivity of your results by making modifications to the variables included in the model.
  - If your results are stable to these types of variations, that provides justification for your conclusion.
- Note that, there is no denying that a great deal of skill and experience are required in choosing an appropriate model!

Scaling and Units of Measurement

The Effects of Scaling the Data

- Consider the original regression model is

$$Y_i = \beta_1 + \beta_2 X_i + e_i$$

  - Then, the fitted model is

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{e}_i$$

- **Changing the scale of X and Y:** Define new variables

$$Y_i^* = \omega_1 Y_i$$
$$X_i^* = \omega_2 X_i$$

  - For example, when Y_i is food expenditure **(measured in \$100)** and X_i is income **(measured in \$100)**, you can change the unit from \$100 to \$1,000 by defining

$$Y_i^* = \frac{1}{10} Y_i, \quad X_i^* = \frac{1}{10} X_i$$

  - Then Y_i^* is food expenditure **(measured in \$1,000)** and X_i^* is income **(measured in \$1,000)**.

The Effects of Scaling the Data [cont'd]

- Since, $Y_i = \frac{1}{\omega_1} Y_i^*$ and $X_i = \frac{1}{\omega_2} X_i^*$,

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + e_i \\ \implies \quad \frac{1}{\omega_1} Y_i^* &= \beta_1 + \beta_2 \left(\frac{1}{\omega_2} X_i^* \right) + e_i \\ \implies \quad Y_i^* &= \omega_1 \beta_1 + \frac{\omega_1}{\omega_2} \beta_2 X_i^* + \omega_1 e_i \\ \implies \quad Y_i^* &= \beta_1^* + \beta_2^* X_i^* + e_i^* \end{aligned}$$

  where $\beta_1^* = \omega_1 \beta_1$, $\beta_2^* = (\omega_1 / \omega_2)\, \beta_2$ and $e_i^* = \omega_1 e_i$.

- The fitted regression model is:

$$Y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^* + \hat{e}_i^*$$

  where $\hat{e}_i^* = \omega_1 \hat{e}_i$.

The Effects of Scaling the Data [cont'd]

- We can apply OLS methods, and we can obtain OLS estimator $\hat{\beta}_1^*, \hat{\beta}_2^*$.

$$\hat{\beta}_2^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}}, \quad \hat{\beta}_1^* = \bar{Y}^* - \hat{\beta}_2^* \bar{X}^*$$

- It can be verified that:

$$\hat{\beta}_2^* = \left(\frac{\omega_1}{\omega_2}\right) \hat{\beta}_2, \quad \hat{\beta}_1^* = \omega_1 \hat{\beta}_1$$

$$Var\left(\hat{\beta}_2^*\right) = \left(\frac{\omega_1}{\omega_2}\right)^2 Var\left(\hat{\beta}_2\right), \quad Var\left(\hat{\beta}_1^*\right) = \omega_1^2 Var\left(\hat{\beta}_1\right)$$

$$\hat{\sigma^2}^* = \omega_1^2 \hat{\sigma^2}$$

- It is clear that, given the regression results based on one scale of measurement, we can derive another scale of measurement once the **scaling factors** (ω_1 and ω_2) are known.

- Note that $R^2 = R^{2*}$!

The Effects of Scaling the Data [cont'd]

- Changing the scale of X only
  - Only slope coefficient and its variance are multiplied by the factor $\left(\frac{1}{\omega_2}\right)$.
- Changing the scale of Y only
  - Slope coefficient, intercept, and their standard errors are all multiplied by the same factor ω_1.
- The same scale to X and Y
  - No change in the slope parameter and its variance, but intercept and its standard error are both multiplied by ω_1.

A Word about Interpretation

- Since the slope coefficient β_2 is simply the rate of change, it is measured in the units of the ratio:

$$\frac{\text{Units of the dependent variable}}{\text{Units of the explanatory variable}}$$

- For example, using our example of food expenditure and household income

$$(\textbf{Model 1}:\ Y_i \text{ in } \$100,\ X_i \text{ in } \$100)\ \hat{Y}_i = 83.42 + 10.21X_i$$

$$(\textbf{Model 2}:\ Y_i \text{ in } \$100,\ X_i \text{ in } \$1{,}000)\ \hat{Y}_i = 83.42 + 102.1X_i$$

  - Interpretation of Model 1: \$100 change in income leads to 10.21 hundred dollar change in food expenditure.
  - Interpretation of Model 2: \$1,000 change in income leads to 102.1 hundred dollar change in food expenditure.
  - Note that the two results are of course identical in the effects of income on food expenditure.