Properties of linear regression model estimators

Susan Thomas
IGIDR, Bombay

2 October, 2008

Recap

- The linear regression model is one of the form:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_J X_{J,i} + u_i$$

- Or, $Y = X\beta + U$
  where $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_J)'$
- The linear regression model estimates are those which minimise the sum of squared errors:

$$\begin{aligned} \sum \epsilon_i = \sum (Y_i - \beta_0 - \ldots - \beta_J X_{J,i}) &= 0 \\ \min_{\beta,\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_{1,i} - \ldots - \beta_J X_{J,i})^2 &= \min_{\beta,\sigma^2} \sum u_i^2 \end{aligned}$$

- Jargon: The model is referred to as the "Data Generating Process" (DGP).

Recap

- For a simple one-exogenous variable model,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$$

- β_0 is the intercept on the "regression line" and β_1 is the slope.
- The above equation is called the "population regression line".
- After estimation, we have $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + u_i$ which is called the "estimated/sample regression line"
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  the "regression line" passes through the mean of the dataset.
- $\hat{\beta}_1 = S_{xy}/S_{xx}$
  where S_{xy} is the sample covariance and S_{xx} is the sample variance of the exogenous data X.

Properties of linear regression estimators

Questions after the regression

- $(\hat{\beta}_0, \hat{\beta}_1)$ are the estimated parameters that provide the "best fit" to the data.
- But we also ask other questions:
  1. What are their sample statistical properties?
  2. What reliability/ precision they have?
  3. How can we use the estimates to test a hypothesis?
  4. How can we use the estimates when forecasting the Y_i?
- These are questions about how a sample estimate can be used to capture knowledge about the population estimate.

Sample properties of regression estimators

- Sample statistical features will be the distribution of the estimator.
- This distribution will have a mean and a variance, which in turn, leads to the following properties of estimators:
  1. Unbiasedness: $E(\hat{\beta}) = \beta$
  2. Consistency: As $N \rightarrow \infty$, $\hat{\beta} = \beta$, $var(\hat{\beta}) \rightarrow 0$.
  3. Efficiency: $E(\hat{\beta})^2$ is minimum among all other estimators

Testing unbiasedness: Mean of (β_0, β_1)

- Each estimated value may differ from β_0, β_1, but their expected/ average value would be equal to β_0, β_1.
- Thus, if these estimators are "unbiased", then

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0 \\ E(\hat{\beta}_1) &= \beta_1 \end{aligned}$$

Assumptions needed for unbiasedness of β_0, β_1

- $E(u_i|X_i) = 0$

$$\Rightarrow \quad E(u_i) = 0$$

  (by law of iterated expectations)

- X_i's are fixed and non- stochastic.

$$\Rightarrow Cov(X_i, u_i) = 0$$

- Working this out:

$$\begin{aligned} Cov(X_i, u_i) &= E\left[(X_i - E(X_i))(u_i - E(u_i))\right] \\ &= E(X_i u_i) - E(X_i)E(u_i) \\ &= X_i E(u_i) - E(X_i)E(u_i) \\ &= X_i E(u_i) - X_i E(u_i) \\ &= 0 \end{aligned}$$

Proof that $E(\hat{\beta}_1) = \beta_1$

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$
- S_{XY} works out to be:

$$
\begin{aligned}
&= \sum X_{1,i} Y_i - \frac{1}{N} \sum X_{1,i} \sum Y_i \\
&= \sum X_{1,i} \left[\beta_0 + \beta_1 X_{1,i} + u_i\right] - \frac{1}{N} \sum X_i \sum (\beta_0 + \beta_1 X_{1,i} + u_i) \\
&= \beta_0 \sum X_{1,i} + \beta_1 \sum X_{1,i}^2 + \sum X_{1,i} u_i \\
&\quad - \frac{1}{N} \sum X_{1,i} \left[N\beta_0 + \beta_1 \sum X_{1,i} + \sum u_i\right] \\
&= \beta_0 \sum X_{1,i} + \beta_1 \sum X_{1,i}^2 + \sum X_{1,i} u_i \\
&\quad - \beta_0 \sum X_{1,i} + \beta_1 \frac{1}{N} \left[\sum X_{1,i}\right]^2 - \frac{1}{N} \sum X_{1,i} \sum u_i \\
&= \beta_1 \left[\sum X_{1,i}^2 - \frac{1}{N} \left[\sum X_{1,i}\right]^2\right] + \left[\sum X_{1,i} u_i - \frac{1}{N} \sum X_{1,i} \sum u_i\right]
\end{aligned}
$$

- So,

$$S_{XY} = \beta_1 S_{XX} + S_{Xu}$$

Proof that $E(\hat{\beta}_1) = \beta_1$

Thus:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \\
&= \frac{\beta S_{XX} + S_{Xu}}{S_{XX}} = \beta \frac{\cancel{S_{XX}}}{\cancel{S_{XX}}} + \frac{S_{Xu}}{S_{XX}} \\
\hat{\beta}_1 = \beta_1 + \frac{S_{Xu}}{S_{XX}} E(\hat{\beta}_1) &= \beta_1 + E\left(\frac{S_{Xu}}{S_{XX}} | X\right) \\
&= \beta_1 + \frac{1}{S_{XX}}. \; E(S_{Xu}|X) \\
\text{Or, } E(\hat{\beta}_1) &= \beta_1 + \frac{1}{S_{XX}} E(S_{Xu}|X)
\end{aligned}
$$

Proof that $E(\hat{\beta}_1) = \beta_1$

- Then,

$$\begin{aligned} E(S_{Xu}|X) &= E\left[\sum (X_{1,i} - \bar{X}_1)(u_i - \bar{u})\right] \\ &= \sum E(X_{1,i} - \bar{X}_1)(u_i - \bar{u}) \\ &= \sum (X_{1,i} - \bar{X}_1) E(u_i - \bar{u}) \\ &= 0 \end{aligned}$$

Therefore $E(\hat{\beta}_1|X) = \beta_1$ $\rightarrow$ Unbiased

By law of iterated expectation $E(\hat{\beta}_1) = \beta_1$.

HW

Show that $E(\hat{\beta}_0) = \beta_0$

Property 2: Consistency

- If we have two estimators, $\hat{\beta}_{0,n}, \hat{\beta}_{1,n}$

$$(\hat{\beta}_{0,n=1}, \hat{\beta}_{1,n=1}) \cdots (\hat{\beta}_{0,n=m}, \hat{\beta}_{1,n=m})$$

- As (n = m) becomes large $\hat{\beta}_{0,n=m}, \hat{\beta}_{1,n=m}$ converge to the true parameters, β_0, β_1.

$\hat{\beta}_0, \hat{\beta}_1$ are consistent

- We know that:

$$\hat{\beta}_1 = \beta_1 + \frac{S_{Xu}}{S_{XX}} = \beta_1 + \frac{S_{Xu}/N}{S_{XX}/N}$$

- As $N \rightarrow \beta_0$,

$$S_{Xu} \rightarrow Cov(X, u)$$

  and by the Law of large numbers $S_{Xu} = 0$
- As $N \rightarrow \infty$,

$$S_{XX} \rightarrow Var(X) = \sigma^2 X$$

- Therefore, $\text{Plim}\hat{\beta} = \beta$

Assumptions needed for efficiency

- Assumptions on the error, u_i.
  1. Assumption of *homoskedasticity*: $u_i \sim \text{i.i.d}$ in conditional variance σ^2

$$V(u_i|X_{1,i}) = E(u_i^2|X_{1,i}) = \sigma^2 \quad \forall i$$

  2. Assumption of serial independence.

$$CovE(u_i, u_j|X_{1,i}) = E(u_i u_j|X_{1,i}) = 0 \quad i \neq j$$

Precision of estimators and model

Step 1: How precisely are the estimators observed?

- The "precision of estimators" is captured by the estimator's variance.

$$
\begin{aligned}
\text{var}(\hat{\beta}_1) &= \text{var}\left[\beta_1 + \frac{S_{Xu}}{S_{XX}}\right] \\
&= 0 + \frac{1}{[S_{XX}]^2} \quad \text{var}(S_{Xu}) \\
\text{But} \quad \text{var}(S_{Xu}) &= \text{var}[\sum(X_{1,i} - \bar{X})u_i] \\
&= [\sum \text{var}(X_{1,i} - \bar{X})u_i] \text{ by serial independence} \\
&= \sum(X_i - \bar{X})^2 \text{var}(u_i) \\
&= \sigma_u^2 \sum(X_{1,i} - \bar{X})^2 \\
&= \sigma_u^2 S_{XX}
\end{aligned}
$$

Variance of OLS estimators

- The variance of the estimators are:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{S_{XX}} \quad (1)$$

$$\text{var}(\hat{\beta}_0) = \frac{\sum X_i^2 . \sigma_u^2}{N S_{XX}} \quad (2)$$

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}_1 \sigma_u^2}{S_{XX}} \quad (3)$$

Where $S_{XX} = \sum (X_{1,i} - \bar{X}_1)^2$

- The higher the S_{XX} (more variation in X) and larger the sample size, N, the lower is the *se* of the estimated parameters.
  With larger samples and higher variability in X, we estimate β more precisely.

σ_u^2 estimator

- The above are the "true" or population variances, which is unknown because σ_u^2 is unknown.
- We estimate σ_u^2 as:

$$
\begin{aligned}
\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} \\
\Rightarrow \quad \hat{u}_i &= Y_i - \hat{Y}_i \\
\Rightarrow \quad \hat{u}_i &= Y_i - \hat{\beta}_0 - \hat{\beta}_0 X_{1,i} \quad \text{estimate errors} \\
\tilde{\sigma}_u^2 &= \frac{\sum \hat{u}_i^2}{N} \quad \text{Law of large numbers} \\
&\Leftarrow \tilde{\sigma}_u^2 \rightarrow \sigma_u^2
\end{aligned}
$$

- But the above is biased. An unbiased estimator is:

$$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{(N-2)}$$

- $(N-2)$, not N, because there are two parameters to estimate: (β_0, β_1), using which we got $\hat{u}_i$. $N > 2$ for the σ_u^2 to be positive

Using σ_u^2 to calculate var(β)

- $\hat{u}_i$ satisfy $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is
  - Standard error of the disturbances, or
  - Standard error of regression
- The estimated variances are

$$\begin{aligned}
\hat{V}(\hat{\beta}_1) &= s^2_{\hat{\beta}_1} = \hat{\sigma}^2_u / S_{XX} \\
\hat{V}(\hat{\beta}_0) &= s^2_{\hat{\beta}_0} = \hat{\sigma}^2_u \sum X^2_{1,i} / (N S_{XX}) \\
\widehat{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= s_{(\hat{\beta}_0 \hat{\beta}_1)} = -\bar{X}_1 \hat{\sigma}^2_u / S_{XX}
\end{aligned}$$

- The square root of the estimated variances above, are called the “standard errors” of the regression coefficients.

Goodness of Fit

- Overall Goodness of F_{it}: many straight lines can pass through the observation paris. The OLS fitted line is the "best".
- Yet it is imprecise: it does not pass through all the points, because of the errors u_i.
- A quantification of how "good" is the fit of the relationship is captured by the R^2 measure.

Goodness of Fit for $Y_i = \beta_0 + \epsilon_i$

- If we knew only Y_i's then our best predictor would be $\bar{Y}$. Then error would be $(Y_i - \bar{Y})$.
- If we square and sum all the errors we get

$$\sum (Y_i - \bar{Y})^2 = \text{TSS, Total sum of squared errors}$$

- Sample standard deviation:

$$\hat{\sigma}_Y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{N - 1}}$$

Goodness of Fit for $Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$

- If we know $X_{1,i}, \hat{\beta}_0, \hat{\beta}_1$, and the linear relation between Y_i and $X_{1,i}$, we can compute

$$
\begin{aligned}
\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} \\
\Rightarrow \hat{u}_i &= Y_i - \hat{Y}_i, \text{Estimated errors} \\
\sum \hat{u}_i^2 &= \text{ESS, Error sum of squares} \\
\hat{\sigma}_u &= \sqrt{\frac{\text{ESS}}{(N-2)}}, \text{ Standard deviation, dispersion of errors}
\end{aligned}
$$

$Y_i = \beta_0 + \epsilon_i$ vs $Y_i = \beta_0 + \beta_1 X_{1,i} + u_i$

- Compare $\hat{\sigma}_u$ to $\hat{\sigma}_\epsilon$.
- Large reduction means good fitted relation. But $\hat{\sigma}_u$ to $\hat{\sigma}_\epsilon$ depend on unit of measurement.

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y})^2$$

$$TSS = RSS + ESS$$

- Dividing both sides by TSS:

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

- Thus, $Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y})$
  The same holds true for sum of square also: TSS = ESS + RSS, or $R^2 = 1 - \frac{ESS}{TSS}$
- $R^2 \rightarrow$ "coefficient of multiple determination".
  Jargon: In single variable case the word "multiple" does not apply. R^2 is instead called "coefficient of determination".

Characteristics of R^2

- Better the fit, the close are the scatter points to the fitted line.
  Then, the lower would be $\sum \hat{u}_i^2$, or ESS, and greater would be RSS.
  Thus R^2 is a measure of goodness of fit.
- ESS $\longrightarrow$ Unexplained variation.
- RSS $\longrightarrow$ Explained variation.
- Thus R^2 is the percentage of total variation explained by model.
- Low R^2 means that a lot of variation in Y_i is unexplained by model.

Features of R^2

- It is obvious that $0 \leq R^2 \leq 1$

$$\begin{aligned} \text{We know that } TSS &= RSS + ESS \\ \frac{TSS}{TSS} &= \frac{RSS}{TSS} + \frac{ESS}{TSS} \\ 1 &= R^2 + \frac{ESS}{TSS} \longrightarrow 1 - \frac{ESS}{TSS} \\ \text{Since } 0 \leq \frac{ESS}{TSS} \leq 1 &\longrightarrow 0 \leq R^2 \leq 1 \end{aligned}$$

- How to decide if R^2 is high or low?
  $\Rightarrow$ No unique answer

HW

- Show that $E(\hat{Y}_i) = \bar{Y}$.

The OLS model

- Under the assumptions of:
  1. $Y = X\beta + u$ (linear in parameters)
  2. $E(\hat{\beta}) = \beta$ (unbiased)
  3. $E(u|X) = 0$
  4. X are fixed for i, $\text{Cov}(X_i, u_i) = 0$
  5. $u_i \sim iid$, such that $\sigma_i^2 = \sigma^2$ (homoskedasticity)
  6. $\text{Cov}(u_i, u_j|X_i) = 0$, (serial independence)

  $\beta = (X'X)^{-1}(X'Y)$
- These parameters are called the *Ordinary Least Squares* or “OLS” parameters.
- Note: no distribution assumption on u_i.

OLS is “BLUE”

- The parameters minimising the SSE ($\sum u_i^2$) are *most* efficient among other linear, unbiased, estimators.
- Jargon: OLS parameters are *BLUE*: Best Linear Unbiased Estimators.
- Key to note here: it’s only “best” amongst linear and unbiased estimators.

Getting there

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$

$$\begin{aligned}
S_{XY} &= \sum X_i Y_i - \frac{1}{N} \sum X_i \sum Y_i \\
&= \sum X_i Y_i - \bar{X} \sum Y_i \\
&= \sum (X_i - \bar{X}) Y_i \\
\text{Rewrite } \hat{\beta}_1 &= \frac{\sum (X_i - \bar{X}) Y_i}{S_{XX}} = \sum \frac{(X_i - \bar{X})}{S_{XX}} . Y_i \\
&= \sum \omega_i Y_i \\
\text{Now, } Y_i &= \beta_0 + \beta_1 X_i + u_i \\
\text{With independence} \Rightarrow \text{var}(Y) &= \text{var}(u_i) = \sigma^2 \\
\text{Therefore, var}(\hat{\beta}_1) &= \sum \omega_i^2 \text{var}(Y_i) = \sigma^2 \sum \omega_i^2
\end{aligned}$$

Compare with alternative linear unbiased estimators

- $\tilde{\beta}_1 = \sum a_i Y_i$

$$
\begin{aligned}
E(\tilde{\beta}_1) &= \sum a_i E(Y_i) \\
E(\tilde{\beta}_1) &= \sum a_i [\beta_0 + \beta_1 X_i] \\
\tilde{\beta}_1 &= \sum a_i Y_i \\
\text{Set } a_i &= \omega_i + d_i \\
\text{Then } \tilde{\beta}_1 &= \sum (\omega_i + d_i) Y_i \\
\tilde{\beta} &= \sum \omega_i Y_i + \sum d_i Y_i = \hat{\beta}_1 + \sum d_i Y_i \\
E(\tilde{\beta}_1) &= \beta_1 + \sum d_i E(Y_i) \\
&= \beta_1 + \sum d_i [\beta_0 + \beta_1 X_i] \\
E(\tilde{\beta})_1 &= \beta_1 + \beta_0 \sum d_i + \beta_1 \sum d_i X_i
\end{aligned}
$$

- For unbiasedness, we need:

$$\sum d_i = 0, \text{ and } \sum d_i X_i = 0$$

Checking for efficiency of $\tilde{\beta}_1$

- What is the var$(\tilde{\beta}_1)$

$$\begin{aligned}
\text{var}(\tilde{\beta}_1) &= \sum(\omega_i + d_i)^2\sigma^2 \\
&= \sigma^2\sum(\omega_i^2 + d_i^2 + 2\omega_i d_i) \\
\text{Therefore, var}(\tilde{\beta}) &= \sigma^2\sum\omega_{t^2} + \sigma^2\sum d_{t^2} + 2\sigma^2\sum\omega_i d_i
\end{aligned}$$

- But we know that:

$$\begin{aligned}
\sum\omega_i d_i &= \sum\left(\frac{X_i - \bar{X}}{S_{XX}}\right)d_i \\
&= \frac{\sum X_i d_i - \bar{X}\sum d_i}{S_{XX}} = 0
\end{aligned}$$

- Which means:

$$\begin{aligned}
\text{var}(\tilde{\beta}) &= \sigma^2\sum\omega_i^2 + \sigma^2\sum d_i^2 + 2\sigma^2\sum\omega_i d_i \\
&= \sigma^2(\sum\omega_i^2 + \sum d_i^2) > \sigma^2\sum\omega_i^2
\end{aligned}$$

- Therefore, $\hat{\beta}_1$ is BLUE.