Statistical Properties of LS Estimators

Lecture 3

Alexandra Chronopoulou

ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will:

- study properties of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ as an estimate of the true coefficient vector (β_0, β_1).
- construct confidence/prediction intervals for (β_0, β_1).

Notation Remark

- Uppercase letters are normally used for Random Variables, and lowercase letters for observed values of the random variables.
- Uppercase letter will be also reserved for matrices.
- In some occasions lowercase letter will also be used for random variables.

LS Estimators Properties

$$\textbf{Model:} \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

Assumptions

The **errors** $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to

- have **mean zero**: $\mathbb{E}(\varepsilon_i) = 0$
- be **uncorrelated**: $\mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0,\ \mathrm{i} \neq \mathrm{j}$,
- be **homoscedastic**: $\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$ does not depend on i.

⇒ We can combine the last two and write it as

$$Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$$

where $\delta_{ij} = 0$ if $i \neq j$.

Assumptions on Y

Moments of $(Y|X)$

Based on the SLR model moment assumptions on the error terms, we have the following assumptions for the moments of Y conditioning on X:

$$\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$$
$$\mathrm{Var}(\mathrm{y_i}|\mathrm{x_i}) = \sigma^2$$
$$\mathrm{Cov}(\mathrm{y_i}, \mathrm{y_j}|\mathrm{x_i}, \mathrm{x_j}) = 0, \;\; i \neq j$$

Remark:
When we evaluate expectation, only y_i's are random and x_i's are treated as known, non-random constants.

ILLINOIS

Unbiasedness of LS Estimators

Proposition

Both LS estimators $\hat{\beta}_1$, $\hat{\beta}_0$ are *unbiased*, i.e. $\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1$, $\mathbb{E}\left(\hat{\beta}_0\right) = \beta_0$.

Proof for Slope

Recall that $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x}) \cdot y_i}{\sum_i (x_i - \bar{x})^2}$. So, we have

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_1\right) &= \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}\left(y_i\right)}{\sum_i (x_i - \bar{x})^2}, \qquad \text{since the } x_i' s \text{ are known} \\
&= \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}\left(\beta_0 + \beta_1 x_i\right)}{\sum_i (x_i - \bar{x})^2} = \sum_i c_i \left(\beta_0 + \beta_1 x_i\right), \ \text{where } c_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\
&= \beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i = \beta_1
\end{aligned}$$

where the last result is true since $\sum_i c_i = 0$, and $\sum_i c_i x_i = 1$.

Unbiasedness of LS Estimators

Proof for Intercept

Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. So, we have

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_0\right) &= \mathbb{E}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\
&= \mathbb{E}\left(\bar{y}\right) - \bar{x} \cdot \mathbb{E}\left(\hat{\beta}_1\right) = \frac{1}{n} \sum_i \mathbb{E}(y_i) - \bar{x} \cdot \beta_1 \\
&= \frac{1}{n} \sum_i \mathbb{E}(\beta_0 + \beta_1 x_i) - \bar{x} \cdot \beta_1 \\
&= \beta_0 + \bar{x} \cdot \beta_1 - \bar{x} \cdot \beta_1 = \beta_0
\end{aligned}$$

MSE of LS Estimators

(*) Note that since both estimators are unbiased $\Rightarrow MSE =$ Variance.

MSE for Slope

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}\right] = \mathrm{Var}\left(\sum_i c_i y_i\right) \quad \text{(}c_i\text{ as before)} \\
&= \sum_i c_i^2 \cdot \mathrm{Var}(y_i) = \sum_i c_i^2 \sigma^2 \quad \text{(from model assumption)} \\
&= \sigma^2 \cdot \left(\frac{\sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right)^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \sigma^2 \frac{1}{S_{xx}}
\end{aligned}$$

MSE for Intercept

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \mathrm{Var}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Normal Error Regression Model

SLR Model

$$\textbf{Model:} \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

Normality Assumption

Additionally, we assume that

$$\varepsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$$

Equivalently, $y_i \sim^{iid} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. (Why?)

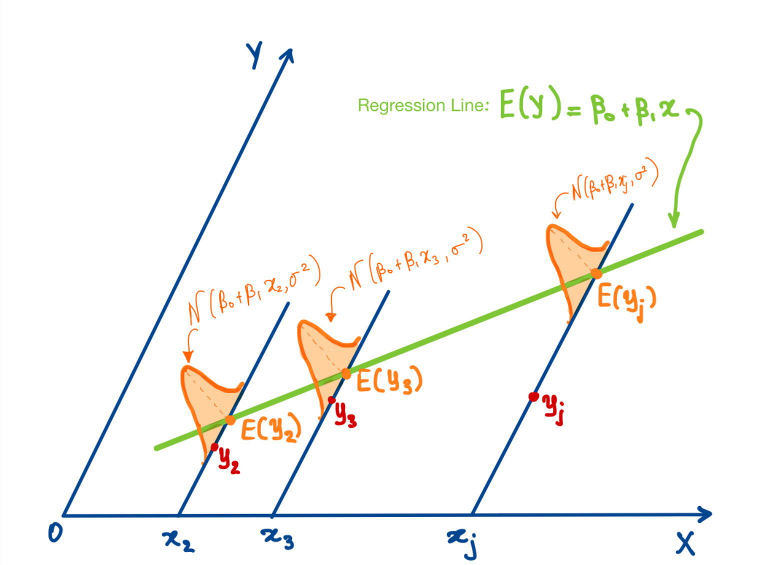
Normality Assumption

Recall that the error terms ε_i are *independent*, *normally distributed* with mean 0 and variance σ^2. Based on that, we can prove the following properties for the y_i's.

Properties of y_i

- $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$, since the ε_i's have mean zero.
- y_i's are independent, since ε_i's are independent.
- $\mathrm{Var(y_i)} = \mathrm{Var(\varepsilon_i)} = \sigma^2$.
- y_i's are a linear shift of the ε_i's, so they are also *normally distributed*.
- The y_i's are *jointly normal*, and so are linear combinations of the y_i's, since the errors are normally distributed and uncorrelated/independen.

Normal Regression Model Illustration



Distribution of LS Estimators

- $\hat{\beta}_1$ and $\hat{\beta}_0$ are jointly normally distributed with

$$\mathbb{E}(\hat{\beta}_1) = \beta_1 \quad \mathsf{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\mathsf{S}_{\mathsf{XX}}}$$

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \quad \mathsf{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{\mathsf{n}} + \frac{\bar{\mathsf{x}}^2}{\mathsf{S}_{\mathsf{XX}}} \right)$$

$$\mathsf{Cov}(\hat{\beta}_1,\ \hat{\beta}_0) = -\sigma^2 \frac{\bar{\mathsf{x}}}{\mathsf{S}_{\mathsf{xx}}}.$$

- $RSS = \sum_i (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-2}$ which implies that

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{RSS}{n-2} \right) = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$$

- $(\hat{\beta}_0, \hat{\beta}_1)$ and RSS are *independent*.

Testing for the Slope

$$\begin{cases} H_0 : \beta_1 = c \ (\textit{null}) \\ H_\alpha : \beta_1 \neq c \ (\textit{alternative}) \end{cases}$$

where c is an known constant.

- The test statistics is

$$t = \frac{\hat{\beta}_1 - c}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{xx}}}$$

- The distribution of t *under the null* is T_{n-2}.
- The *p*-value is twice the area under the T_{n-2} distribution more extreme than the observed statistic t.

R ouputs the *p*-value for testing β_1 against 0, i.e. $c = 0$.

Hypothesis Testing

Testing for the Intercept

$$\begin{cases} H_0 : \beta_0 = c \ (\mathit{null}) \\ H_\alpha : \beta_0 \neq c \ (\mathit{alternative}) \end{cases}$$

- The test statistics is

$$t = \frac{\hat{\beta}_0 - c}{\sqrt{\mathrm{Var}(\hat{\beta}_0)}}$$

- The distribution of t *under the null* is T_{n-2}.
- The p-value is twice the area under the T_{n-2} distribution more extreme than the observed statistic t.

R ouputs the p-value for testing β_0 against 0, i.e. $c = 0$.

ANOVA Table & F-Test

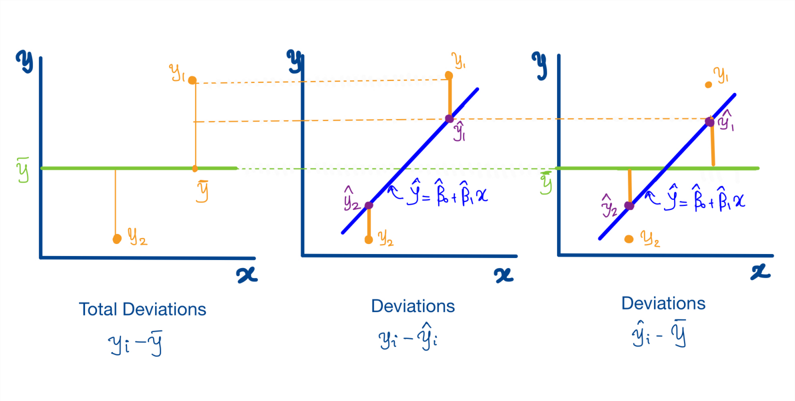
Partitioning the Total Variation (Revisited)

Recall the decomposition of the Total Sum of Squares (TSS)

$$\underset{\substack{\uparrow \\ \text{Total variation} \\ \text{in } y}}{TSS} = \underset{\substack{\uparrow \\ \text{Variation accounted} \\ \text{through the model}}}{FSS} + \underset{\substack{\uparrow \\ \text{Variation left}}}{RSS}$$

- $TSS = \sum_i (y_i - \bar{y})^2$ is the measure of the total variation in ys: the greater TSS is, the more variation there is in the y values.
- $RSS = \sum_i (y_i - \hat{y}_i)^2$ measures the variation in the data using the stated regression model: the larger RSS is, the more y_is vary around the *estimated regression line*.
- $FSS = \sum_i (\hat{y}_i - \bar{y})^2$ measures how far the predicted center of each probability distribution is from the overall center of all y's together.

Partitioning the Total Variation (Revisited)



Breakdown of Degrees of Freedom

- $df_{TSS} = n - 1$: one df is lost, because the sample mean is used to estimate the population mean.
- $df_{RSS} = n - 2$: two df are lost, because the two parameters are estimated in obtaining the fitted values $\hat{y}$.
- $df_{FSS} = 1$: there are n deviations $\hat{y}_i - \bar{y}$, but all the fitted values are associated with the same regression line.

$$df_{TSS} = df_{RSS} + df_{FSS}$$

Sum of Squares	Expression	df
TSS	$\sum_i (y_i - \bar{y})^2$	$n - 1$
FSS	$\sum_i (\hat{y}_i - \bar{y})^2$	1
RSS	$\sum_i (y_i - \hat{y}_i)^2$	$n - 2$

ANOVA Table

Source	SS	df	MS	F
Regression (model)	FSS	1	$MSReg = \frac{FSS}{1}$	$F = \frac{MSReg}{MSE}$
Error	RSS	$n-2$	$MSE = \frac{RSS}{n-2}$	
Total	TSS	$n-1$		

(*) The Mean Squares are *not* additive.

F- Test

An alternative way to test for the model parameters is using the F test:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_\alpha : \beta_1 \neq 0 \end{cases}$$

- Under H_0, the F-test statistic is

$$F = \frac{MSReg}{MSE} = \frac{FSS}{RSS/(n-2)} \sim F_{1,n-2}$$

- It can be shown that the F-test statistic is equal to the square of the t-test statistic and their p-values are the same. So, this test is *equivalent* to the t-test before.

Estimation and Prediction

Estimation & Prediction at a New Case

The LS line can be used to obtain values of the response ($Y^\star$) for given values of the predictor ($X = x^\star$). There are two variants of this problem[1]:

1. Estimation: We want to estimate the mean response at $x\star$. This is equivalent to estimate: $\beta_0 + \beta_1 x^\star$
2. Prediction of an outcome of random variable $Y^\star$ at a given value $x^\star$, where

$$Y^\star \sim N(\beta_0 + \beta_1 x^\star, \sigma^2)$$

The fitted value (or point estimate) for estimation and prediction are the same: $\hat{\beta}_0 + \hat{\beta}_1 x^\star$. However the accuracy for estimation and the one for prediction are different.

[1] Estimation looks to get information from the data about a fixed but parameter, while prediction looks to get information about a random variable

Estimation of the Mean Response

– Accuracy of the estimation is measured by the expected value of the squared difference between the point estimate and the target.

– For estimation the target is $\beta_0 + \beta_1 x^\star$:

$$
\begin{aligned}
&\mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star\right)^2 \\
=&\mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}^\star\right) \\
=&\mathrm{Var}\left(\hat{\beta}_0\right) + (\mathrm{x}^\star)^2 \mathrm{Var}\left(\hat{\beta}_1\right) + 2\mathrm{x}^\star \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
=&\sigma^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right) \\
=&\sigma^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}\right)
\end{aligned}
$$

[2]

[2] Recall that all our calculations are done *conditionally* on $x^\star$

ILLINOIS

Confidence Interval for $\mathbb{E}(Y^\star)$

– A confidence interval is always reported for a parameter. An $(1-\alpha)100\%$ Confidence Interval for the *Mean Response* when $x = x_\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

Prediction

- For prediction the target is $Y^\star = \beta_0 + \beta_1 x^\star + e^\star$, where $e^\star \sim N(0, \sigma^2)$. This new error $e^\star$ is independent of the previous n data points, i.e. is independent of $(\hat{\beta}_0, \hat{\beta}_1)$

$$
\begin{aligned}
&\mathbb{E}[(\hat{\beta}_0 + \hat{\beta}_1 x^\star - Y^\star)^2] \\
&= \mathbb{E}[(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star - e^\star)^2] \\
&= \mathbb{E}[(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star)^2] + \mathbb{E}[(e^\star)^2] \\
&= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{XX}}\right)
\end{aligned}
$$

Prediction Interval

- A prediction interval is reported for the value of a random variable, for example, $Y^\star$. An $(1-\alpha)100\%$ Prediction Interval for $\hat{Y}^\star$ when $x = x^\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

Remarks

- Based upon the $\mathrm{Var}(\hat{Y}^\star)$, the prediction interval is wider than the interval used to estimate the mean response at fixed $x = x^\star$.
- So far, we have assumed that the x-levels are known constants. So, all the previous results hold if:
    1. $f(y|x)$ are independent and normally distributed with mean $\beta_0 + \beta_1 x$ and variance σ^2 conditionally on x.
    2. x are independent with distribution $g(x_i)$ that does not depend on β_0, β_1, or σ^2.

Association/Correlation vs Causation

- The statement "X causes Y" means that changing the value of X will change the distribution of Y. When X causes Y , X and Y will be associated, but the reverse is not, in general, true.

Association does not necessarily imply causation.

- If the data are from a *randomized study*, then the causal interpretation is correct.
- If the data are from a *observational study,* then the association interpretation is correct.