Simple Linear Regression: OLS Estimates

Engineering Statistics II
Section 12.2

Josh Engwer

TTU

2023

PART I

PART I:

Simple Linear Regression Analysis

OLS Estimators

Model Assumptions

Linear Model

Sums

Cauchy-Schwarz Inequality

Centered Sums

Predicted Responses

Residuals

Best Linear Unbiased Estimators (BLUE’s)

Gauss-Markov Theorem

Simple Linear Regression (Model Assumptions)

Proposition

(Simple Linear Regression Model Assumptions)

- *(**1 Numerical Response**) Response is not categorical.*
- *(**1 Numerical Regressor**) Regressor is not categorical.*

- *(**Regressors are Perfect**) No errors in regressor measurements.*
- *(**Balance around Fit Line**) Nearly equal scatter above & below fit line.*

- *(**Independence**) All measurements are independent.*
- *(**Normality**) All measurements are approximately normally distributed.*
- *(**Equal Variances**) All measurements have approx. same variance.*

Mnemonic: ***1NR(1NR)*** | ***RaP BaFL*** | ***I.N.EV***

Simple Linear Regression (Linear Model)

Definition

The **simple linear regression model** is:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

$Y_i \equiv$ rv for measurement of i^{th} response

$x_i \equiv$ Actual measurement of i^{th} regressor

$\beta_0 \equiv$ Expected value of response when regressor is zero

$\beta_1 \equiv$ Expected change in response per unit increase in regressor

$E_i \equiv$ Effect of random error on i^{th} response

The **realized simple linear regression model** is:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

$y_i \equiv$ Actual measurement of i^{th} response

$x_i \equiv$ Actual measurement of i^{th} regressor

$\hat{\beta}_0 \equiv$ Estimated value of response when regressor is zero

$\hat{\beta}_1 \equiv$ Estimated change in response per unit increase in regressor

$e_i \equiv$ Actual error on i^{th} response

Simple Linear Regression (Sums)

Several particular sums show up in Simple Linear Regression:

Definition

(Sums)

Let vectors $\mathbf{x} = (x_1, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, \cdots, y_n)^T$. Then:

Regressor Sum	$S_x := \sum_i x_i$
Response Sum	$S_y := \sum_i y_i$
Squared Regressor Sum	$S_{xx} := \sum_i x_i x_i$
Squared Response Sum	$S_{yy} := \sum_i y_i y_i$
Cross-termed Sum	$S_{xy} := \sum_i x_i y_i$

Cauchy-Schwarz Inequality for $\mathbb{R}^n$

Theorem

(Cauchy-Schwarz Inequality for $\mathbb{R}^n$)

Let vectors $\mathbf{x} = (x_1, \cdots, x_n)^T$ *and* $\mathbf{y} = (y_1, \cdots, y_n)^T$.
Let sums $S_x := \sum_i x_i, \quad S_y := \sum_i y_i, \quad S_{xx} := \sum_i x_i x_i, \quad S_{yy} := \sum_i y_i y_i, \quad S_{xy} := \sum_i x_i y_i$

Then: $(a) \quad (S_{xy})^2 \leq S_{xx} \cdot S_{yy} \qquad (b) \quad (S_x)^2 \leq n \cdot S_{xx}$

where equality in part (a) *holds when either* $\mathbf{x} = \vec{\mathbf{0}}$ *or* $\mathbf{y} = \vec{\mathbf{0}}$ *or* $\mathbf{y} = c\mathbf{x}$ *where* $c \neq 0$.
where equality in part (b) *holds when either* $\mathbf{x} = \vec{\mathbf{0}}$ *or* $\mathbf{x} = k \cdot \vec{\mathbf{1}}$ *where* $k \in \mathbb{R}$.

<u>PROOF:</u> Recall from trigonometry $(*)$: $\;|\cos\theta| \leq 1 \;\; \forall\theta \implies \cos^2\theta \leq 1 \;\; \forall\theta$

$$(a)\ \mathbf{x}\cdot\mathbf{y} = ||\mathbf{x}||_2||\mathbf{y}||_2\cos\theta \implies (\mathbf{x}\cdot\mathbf{y})^2 = ||\mathbf{x}||_2^2||\mathbf{y}||_2^2\cos^2\theta \overset{(*)}{\leq} ||\mathbf{x}||_2^2||\mathbf{y}||_2^2$$

$$\implies (\mathbf{x}\cdot\mathbf{y})^2 \leq ||\mathbf{x}||_2^2||\mathbf{y}||_2^2 \overset{SUMS}{\implies} \left(\textstyle\sum_i x_i y_i\right)^2 \leq \left(\textstyle\sum_i x_i x_i\right)\cdot\left(\textstyle\sum_i y_i y_i\right) \overset{SUMS}{\implies} (S_{xy})^2 \leq S_{xx}\cdot S_{yy}$$

Now: Let $\mathbf{x} = \vec{\mathbf{0}}$ or $\mathbf{y} = \vec{\mathbf{0}}$. Then $\mathbf{x}\cdot\mathbf{y} = 0$ and $||\mathbf{x}||_2^2 = 0$ or $||\mathbf{y}||_2^2 = 0$.

$$\implies \text{(LHS)} = (\mathbf{x}\cdot\mathbf{y})^2 = 0,\ \text{(RHS)} = ||\mathbf{x}||_2^2||\mathbf{y}||_2^2\cos^2\theta = 0\cdot\cos^2\theta = 0 \implies \text{(LHS)} = \text{(RHS)}$$

Now: Let $\mathbf{y} = c\mathbf{x}$. Then $\mathbf{x}\cdot\mathbf{y} = \mathbf{x}\cdot(c\mathbf{x}) = c(\mathbf{x}\cdot\mathbf{x}) = c||\mathbf{x}||_2^2$ and $||\mathbf{y}||_2^2 = ||c\mathbf{x}||_2^2 = c^2||\mathbf{x}||_2^2$.

$$\implies \text{(LHS)} = (\mathbf{x}\cdot\mathbf{y})^2 = c^2||\mathbf{x}||_2^4,\ \text{(RHS)} = ||\mathbf{x}||_2^2||\mathbf{y}||_2^2\cos^2\theta = c^2||\mathbf{x}||_2^4\cdot 1 \implies \text{(LHS)} = \text{(RHS)}$$

Cauchy-Schwarz Inequality for $\mathbb{R}^n$

Theorem

(Cauchy-Schwarz Inequality for $\mathbb{R}^n$)

Let vectors $\mathbf{x} = (x_1, \cdots, x_n)^T$ *and* $\mathbf{y} = (y_1, \cdots, y_n)^T$.
Let sums $S_x := \sum_i x_i, \quad S_y := \sum_i y_i, \quad S_{xx} := \sum_i x_i x_i, \quad S_{yy} := \sum_i y_i y_i, \quad S_{xy} := \sum_i x_i y_i$

Then: $(a) \quad (S_{xy})^2 \leq S_{xx} \cdot S_{yy} \qquad (b) \quad (S_x)^2 \leq n \cdot S_{xx}$

where equality in part (a) *holds when either* $\mathbf{x} = \vec{\mathbf{0}}$ *or* $\mathbf{y} = \vec{\mathbf{0}}$ *or* $\mathbf{y} = c\mathbf{x}$ *where* $c \neq 0$.
where equality in part (b) *holds when either* $\mathbf{x} = \vec{\mathbf{0}}$ *or* $\mathbf{x} = k \cdot \vec{\mathbf{1}}$ *where* $k \in \mathbb{R}$.

PROOF:

(b) Consider the particular case where $\mathbf{y}$ is the vector of all one's: $\mathbf{y} = (1, \cdots, 1)^T$. Then:

$$S_{xy} := \sum_i x_i y_i = \sum_i x_i \cdot 1 = \sum_i x_i := S_x$$
$$S_{yy} := \sum_i y_i y_i = \sum_i 1 \cdot 1 = \sum_i 1 = n$$

$\therefore$ Applying these simplified expressions to S_{xy} and S_{yy} in part (a) yields: $(S_x)^2 \leq n \cdot S_{xx}$

Now: Let $\mathbf{x} = \vec{\mathbf{0}}$. Then $S_x = \sum_i 0 = 0$ and $S_{xx} = \sum_i 0^2 = 0 \implies$ (LHS) = (RHS)
Now: Let $\mathbf{x}$ be a multiple of the vector of all one's: $\mathbf{x} = k \cdot (1, \cdots, 1)^T = (k, \cdots, k)^T$ where $k \in \mathbb{R}$.

$\implies S_x := \sum_i x_i = \sum_i k = nk$ and $S_{xx} := \sum_i x_i x_i = \sum_i k \cdot k = \sum_i k^2 = nk^2$

$\implies$ (LHS) $= (S_x)^2 = (nk)^2 = n^2k^2$, (RHS) $= n \cdot S_{xx} = n \cdot nk^2 = n^2k^2 \implies$ (LHS) = (RHS) □

Simple Linear Regression (Centered Sums)

Several particular centered sums show up in Simple Linear Regression:

Definition

(Centered Sums)

Let vectors $\mathbf{x} = (x_1, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, \cdots, y_n)^T$. Then:

Squared Regressor Centered Sum	$SC_{xx} := \sum_i (x_i - \overline{x})^2$
Squared Response Centered Sum	$SC_{yy} := \sum_i (y_i - \overline{y})^2$
Cross-termed Centered Sum	$SC_{xy} := \sum_i (x_i - \overline{x})(y_i - \overline{y})$

Simple Linear Regression (Centered Sums)

It is useful to express the centered sums in terms of sums:

Lemma

(Centered Sum Lemma – CSL)

$$(a)\ SC_{xx} = S_{xx} - \frac{1}{n} S_x S_x \qquad (b)\ SC_{yy} = S_{yy} - \frac{1}{n} S_y S_y \qquad (c)\ SC_{xy} = S_{xy} - \frac{1}{n} S_x S_y$$

Simple Linear Regression (Centered Sums)

It is useful to express the centered sums in terms of sums:

Lemma

(Centered Sum Lemma – CSL)

$$(a)\ SC_{xx} = S_{xx} - \frac{1}{n} S_x S_x \qquad (b)\ SC_{yy} = S_{yy} - \frac{1}{n} S_y S_y \qquad (c)\ SC_{xy} = S_{xy} - \frac{1}{n} S_x S_y$$

PROOF:

$$\begin{aligned}
(a)\ \ SC_{xx} \quad &:= \quad \textstyle\sum_i (x_i - \bar{x})^2 \\
&= \quad \textstyle\sum_i (x_i x_i - 2\bar{x} x_i + \overline{xx}) \\
&= \quad \textstyle\sum_i x_i x_i - 2\bar{x} \sum_i x_i + \sum_i \overline{xx} \\
&\overset{\bar{x}}{=} \quad \textstyle\sum_i x_i x_i - \frac{2}{n} \sum_i x_i \sum_i x_i + \sum_i \overline{xx} \\
&\overset{S.}{=} \quad S_{xx} - \frac{2}{n} S_x S_x + \textstyle\sum_i \frac{S_x}{n} \frac{S_x}{n} \\
&= \quad S_{xx} - \frac{2}{n} S_x S_x + \frac{1}{n} S_x S_x \\
&= \quad S_{xx} - \frac{1}{n} S_x S_x
\end{aligned}$$

Simple Linear Regression (Centered Sums)

It is useful to express the centered sums in terms of sums:

Lemma

(Centered Sum Lemma – CSL)

$$(a)\ SC_{xx} = S_{xx} - \frac{1}{n}S_xS_x \qquad (b)\ SC_{yy} = S_{yy} - \frac{1}{n}S_yS_y \qquad (c)\ SC_{xy} = S_{xy} - \frac{1}{n}S_xS_y$$

PROOF:

$$\begin{aligned}
(b)\ \ SC_{yy} \quad &:= \quad \textstyle\sum_i (y_i - \bar{y})^2 \\
&= \quad \textstyle\sum_i (y_iy_i - 2\bar{y}y_i + \bar{y}\bar{y}) \\
&= \quad \textstyle\sum_i y_iy_i - 2\bar{y}\sum_i y_i + \sum_i \bar{y}\bar{y} \\
&\overset{\bar{y}}{=} \quad \textstyle\sum_i y_iy_i - \frac{2}{n}\sum_i y_i \sum_i y_i + \sum_i \bar{y}\bar{y} \\
&\overset{S.}{=} \quad S_{yy} - \frac{2}{n}S_yS_y + \textstyle\sum_i \frac{S_y}{n}\frac{S_y}{n} \\
&= \quad S_{yy} - \frac{2}{n}S_yS_y + \frac{1}{n}S_yS_y \\
&= \quad S_{yy} - \frac{1}{n}S_yS_y
\end{aligned}$$

Simple Linear Regression (Centered Sums)

It is useful to express the centered sums in terms of sums:

Lemma

(Centered Sum Lemma – CSL)

$$(a)\ SC_{xx} = S_{xx} - \frac{1}{n} S_x S_x \qquad (b)\ SC_{yy} = S_{yy} - \frac{1}{n} S_y S_y \qquad (c)\ SC_{xy} = S_{xy} - \frac{1}{n} S_x S_y$$

PROOF:

$$\begin{aligned}
(c)\ \ SC_{xy} &:= \textstyle\sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
&= \textstyle\sum_i (x_i y_i - x_i \bar{y} - \bar{x} y_i + \overline{xy}) \\
&= \textstyle\sum_i x_i y_i - \bar{y} \sum_i x_i - \bar{x} \sum_i y_i + \sum_i \overline{xy} \\
&= \textstyle\sum_i x_i y_i - \left(\frac{1}{n} \sum_i y_i\right) \sum_i x_i - \left(\frac{1}{n} \sum_i x_i\right) \sum_i y_i + \sum_i \overline{xy} \\
&= S_{xy} - \tfrac{1}{n} S_y S_x - \tfrac{1}{n} S_x S_y + \textstyle\sum_i \frac{S_x}{n} \frac{S_y}{n} \\
&= S_{xy} - \tfrac{2}{n} S_x S_y + \tfrac{1}{n} S_x S_y \\
&= S_{xy} - \tfrac{1}{n} S_x S_y \qquad \square
\end{aligned}$$

Simple Linear Regression (Predicted Responses)

Definition

(Predicted Responses)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \dots, n$$

Then the corresponding **predicted responses**, denoted $\hat{y}_i$, are:

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$$

SYNONYMS: Predicted values, fitted values

Simple Linear Regression (Residuals)

Definition

(Residuals)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\ i = 1, \ldots, n$$

Then the corresponding predicted responses, denoted $\hat{y}_i$, are:

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Moreover, the corresponding **residuals**, denoted y_i^{res}, are:

$$y_i^{res} := y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Simple Linear Regression (OLS Derivation)

We wish to find the OLS best-fit line to the n data points (with at least two distinct regressors) such that the sum of the squares of the residuals (vertical errors in predicted values) are minimized.

$Q(\beta_0, \beta_1) := \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$ (OLS ≡ **O**rdinary **L**east-**S**quares)

$(\hat{\beta}_0, \hat{\beta}_1) = \text{arg min}_{\beta_0,\beta_1} Q(\beta_0, \beta_1)$

$$\implies \begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \end{cases} \implies \begin{cases} \sum_i -2\,[y_i - (\beta_0 + \beta_1 x_i)] = 0 \\ \sum_i -2x_i\,[y_i - (\beta_0 + \beta_1 x_i)] = 0 \end{cases}$$

$$\implies \begin{cases} 2\sum_i y_i - 2n\beta_0 - 2\beta_1 \sum_i x_i = 0 \\ 2\sum_i x_i y_i - 2\beta_0 \sum_i x_i - 2\beta_1 \sum_i x_i x_i = 0 \end{cases} \overset{S.}{\implies} \underbrace{\begin{cases} n\beta_0 + S_x\beta_1 = S_y \\ S_x\beta_0 + S_{xx}\beta_1 = S_{xy} \end{cases}}_{2x2\ linear\ system\ in\ \beta_0,\beta_1}$$

$$\implies \beta_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x S_x} = \frac{S_{xy} - \frac{1}{n} S_x S_y}{S_{xx} - \frac{1}{n} S_x S_x} \overset{CSL}{=} \frac{SC_{xy}}{SC_{xx}}$$

$$\implies n\beta_0 + S_x\beta_1 = S_y \implies \beta_0 = \frac{1}{n} S_y - \frac{1}{n} S_x \beta_1 \implies \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$\therefore$ A possible minimum or maximum of Q occurs at point $(\beta_0^*, \beta_1^*) := \left(\bar{y} - \beta_1^* \bar{x},\ \dfrac{SC_{xy}}{SC_{xx}}\right)$

Simple Linear Regression (OLS Derivation)

We wish to find the OLS best-fit line to the n data points (with at least two distinct regressors) such that the sum of the squares of the residuals (vertical errors in predicted values) are minimized.

$Q(\beta_0, \beta_1) := \sum_i \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2$ (OLS $\equiv$ **O**rdinary **L**east-**S**quares)

$(\hat{\beta}_0, \hat{\beta}_1) = \text{arg min}_{\beta_0, \beta_1}\ Q(\beta_0, \beta_1)$

A possible minimum or maximum of Q occurs at point $(\beta_0^*, \beta_1^*) := \left(\bar{y} - \beta_1^* \bar{x},\ \dfrac{SC_{xy}}{SC_{xx}}\right)$

Now, use second derivative test to determine whether (β_0^*, β_1^*) is a min or max or neither:

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \sum_i -2\left[y_i - (\beta_0 + \beta_1 x_i)\right] &&\implies \quad \frac{\partial^2 Q}{\partial \beta_0^2} = \sum_i 2 = 2n \\
\frac{\partial Q}{\partial \beta_0} &= \sum_i -2\left[y_i - (\beta_0 + \beta_1 x_i)\right] &&\implies \quad \frac{\partial^2 Q}{\partial \beta_0 \beta_1} = \sum_i 2x_i = 2S_x \\
\frac{\partial Q}{\partial \beta_1} &= \sum_i -2x_i\left[y_i - (\beta_0 + \beta_1 x_i)\right] &&\implies \quad \frac{\partial^2 Q}{\partial \beta_1 \beta_0} = \sum_i 2x_i = 2S_x \\
\frac{\partial Q}{\partial \beta_1} &= \sum_i -2x_i\left[y_i - (\beta_0 + \beta_1 x_i)\right] &&\implies \quad \frac{\partial^2 Q}{\partial \beta_1^2} = \sum_i 2x_i x_i = 2S_{xx}
\end{aligned}$$

$$\implies H(Q) := \begin{bmatrix} Q_{\beta_0\beta_0} & Q_{\beta_0\beta_1} \\ Q_{\beta_1\beta_0} & Q_{\beta_1\beta_1} \end{bmatrix} = \begin{bmatrix} 2n & 2S_x \\ 2S_x & 2S_{xx} \end{bmatrix}$$

$$\implies \det\left[H(Q)\right] := \frac{\partial^2 Q}{\partial \beta_0^2} \cdot \frac{\partial^2 Q}{\partial \beta_1^2} - \frac{\partial^2 Q}{\partial \beta_0 \beta_1} \cdot \frac{\partial^2 Q}{\partial \beta_1 \beta_0} = (2n)(2S_{xx}) - (2S_x)(2S_x) = 4nS_{xx} - 4(S_x)^2$$

Simple Linear Regression (OLS Derivation)

We wish to find the OLS best-fit line to the n data points (with at least two distinct regressors) such that the sum of the squares of the residuals (vertical errors in predicted values) are minimized.

$Q(\beta_0, \beta_1) := \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$ (OLS $\equiv$ **O**rdinary **L**east-**S**quares)

$(\hat{\beta}_0, \hat{\beta}_1) = \text{arg min}_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$

A possible minimum or maximum of Q occurs at point $(\beta_0^*, \beta_1^*) := \left(\bar{y} - \beta_1^* \bar{x},\ \dfrac{SC_{xy}}{SC_{xx}}\right)$

Now, use second derivative test to determine whether (β_0^*, β_1^*) is a min or max or neither:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0, \quad \frac{\partial^2 Q}{\partial \beta_0 \beta_1} = 2S_x, \quad \frac{\partial^2 Q}{\partial \beta_1 \beta_0} = 2S_x, \quad \frac{\partial^2 Q}{\partial \beta_1^2} = 2S_{xx} \implies H(Q) = \begin{bmatrix} 2n & 2S_x \\ 2S_x & 2S_{xx} \end{bmatrix}$$

$\implies$ det $[H(Q)] = (2n)(2S_{xx}) - (2S_x)(2S_x) = 4nS_{xx} - 4(S_x)^2 > 0$ by Cauchy-Schwarz Inequality

(SUBTLE: det $[H(Q)]$ cannot be zero since there are at least two distinct regressors.)

$\therefore$ Since $\dfrac{\partial^2 Q}{\partial \beta_0^2} > 0$ and det $[H(Q)] > 0$, The only critical point (β_0^*, β_1^*) is a minimum.

$$\therefore\ (\hat{\beta}_0, \hat{\beta}_1) = (\beta_0^*, \beta_1^*) = \left(\bar{y} - \hat{\beta}_1 \bar{x},\ \frac{SC_{xy}}{SC_{xx}}\right)$$

Simple Linear Regression (OLS Estimators)

Theorem

(OLS Estimators – OLS)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \ \text{ where } \ E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

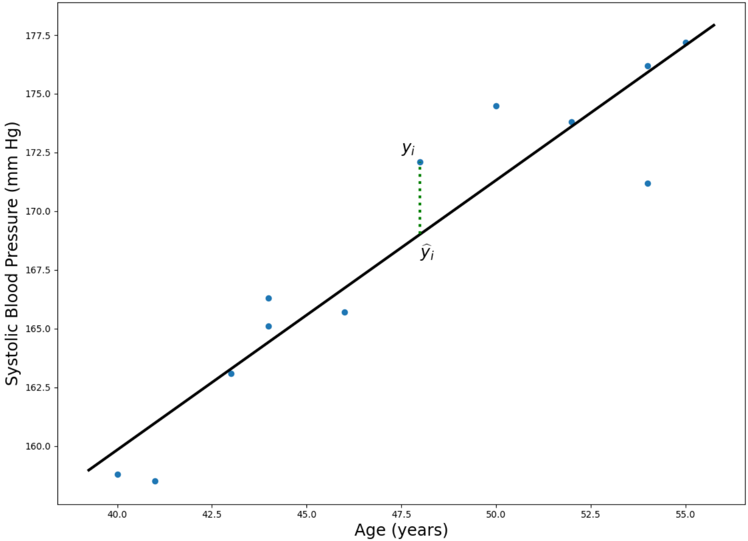
The ***OLS estimators***♠♣ ***(LSE's)*** *for the model parameters are:*

$$\begin{cases} \hat{\beta}_1 & = & \dfrac{SC_{xy}}{SC_{xx}} & = & \dfrac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})(x_i - \overline{x})} & = & \dfrac{S_{xy} - \frac{1}{n} S_x S_y}{S_{xx} - \frac{1}{n} S_x S_x} \\ \hat{\beta}_0 & = & \overline{y} - \hat{\beta}_1 \overline{x} & = & \dfrac{\sum_i y_i - \hat{\beta}_1 \sum_i x_i}{n} & = & \dfrac{S_y - \hat{\beta}_1 S_x}{n} \end{cases}$$

♠A.M. Legendre, *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*, 1806.

♣Gauss, *Theoria Motus Corporum Coelestrium in Sectionibus Conicis Solem Ambientium*, 1809.

Simple Linear Regression (Visualization)



The i^{th} residual (y_i^{res}) is represented by the vertical dashed green line's length.

Simple Linear Regression (BLUE's)

Point estimators for a simple linear regression model should be ideal ones:

Definition

(Best Linear Unbiased Estimators – BLUE's)

A point estimator $\hat{\theta}$ is called a **best linear unbiased estimator (BLUE)** if:

- It estimates a parameter θ of a linear model.
- It is a linear combination of the data points: $\hat{\theta} := \sum_{k=1}^{n} c_k x_k$
- It is an unbiased estimator: $\mathbb{E}[\hat{\theta}] = \theta$
- It has minimum variance of all such unbiased estimators.

Simple Linear Regression (Gauss-Markov Theorem)

Theorem

(Gauss[†]-Markov[‡] Theorem)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

Moreover, suppose the following conditions are all satisfied:

$$\begin{array}{rcll} \mathbb{E}[E_i] & = & 0 & (\textit{errors are all centered at zero}) \\ \mathbb{V}[E_i] & = & \sigma^2 & (\textit{errors all have the same finite variance}) \\ \mathbb{C}[E_i, E_{i'}] & = & 0 & (\textit{errors are uncorrelated when } i \neq i') \end{array}$$

Then, the OLS estimators (LSE's) $\hat{\beta}_0, \hat{\beta}_1$ are both BLUE's.

PROOF: Omitted due to time.

[†] C.F. Gauss, "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae", (1823), 1-58.

[‡] A.A. Markov, *Calculus of Probabilities*, 1^{st} Edition, 1900.

PART II

PART II:

Simple Linear Regression Analysis

Sums of Squares

Deviation Lemma

Residual Lemma

Variation Partitioning

Sums of Squares Lemma

Degrees of Freedom

Simple Linear Regression (Deviation Lemma)

Lemma

(Deviation Lemma – DL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then: (a) $\sum_i (x_i - \overline{x}) = 0$ $\quad (b)$ $\sum_i (x_i - \overline{x}) x_i = SC_{xx}$ $\quad (c)$ $\sum_i (x_i - \overline{x}) y_i = SC_{xy}$

Simple Linear Regression (Deviation Lemma)

Lemma

(Deviation Lemma – DL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then: $(a)\ \sum_i (x_i - \bar{x}) = 0 \quad (b)\ \sum_i (x_i - \bar{x}) x_i = SC_{xx} \quad (c)\ \sum_i (x_i - \bar{x}) y_i = SC_{xy}$

<u>PROOF:</u>

$$(a) \quad \sum_i (x_i - \bar{x}) = \sum_i x_i - \sum_i \bar{x} = S_x - n\bar{x} = S_x - n \cdot \left(\frac{S_x}{n}\right) = S_x - S_x = 0$$

$$(b) \quad \sum_i (x_i - \bar{x}) x_i = \sum_i x_i x_i - \bar{x} \sum_i x_i = S_{xx} - \frac{1}{n} S_x S_x \overset{CSL}{=} SC_{xx}$$

$$(c) \quad \sum_i (x_i - \bar{x}) y_i = \sum_i x_i y_i - \bar{x} \sum_i y_i = S_{xy} - \frac{1}{n} S_x S_y \overset{CSL}{=} SC_{xy} \qquad \square$$

Simple Linear Regression (Residual Lemma)

Lemma

(Residual Lemma – RL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \ \text{ where } \ E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then: $(a)\ \sum_i (y_i - \hat{y}_i) = 0 \quad (b)\ \sum_i (y_i - \hat{y}_i) x_i = 0 \quad (c)\ \sum_i (y_i - \hat{y}_i)\hat{y}_i = 0$

Simple Linear Regression (Residual Lemma)

Lemma

(Residual Lemma – RL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then: $(a)\ \sum_i (y_i - \hat{y}_i) = 0 \quad (b)\ \sum_i (y_i - \hat{y}_i) x_i = 0 \quad (c)\ \sum_i (y_i - \hat{y}_i)\hat{y}_i = 0$

<u>PROOF:</u> (a)

$$\sum_i (y_i - \hat{y}_i) \overset{\hat{y}_i}{=} \sum_i \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right] \overset{\hat{\beta}.}{=} \sum_i \left[y_i - \left(\bar{y} - \frac{SC_{xy}}{SC_{xx}} \bar{x} \right) - \left(\frac{SC_{xy}}{SC_{xx}} \right) x_i \right]$$

$$= \sum_i y_i - \sum_i \bar{y} + \frac{SC_{xy}}{SC_{xx}} \sum_i \bar{x} - \frac{SC_{xy}}{SC_{xx}} \sum_i x_i = \sum_i y_i - n\bar{y} - \frac{SC_{xy}}{SC_{xx}} \sum_i (x_i - \bar{x})$$

$$\overset{S.}{=} S_y - n \cdot \left(\frac{1}{n} S_y \right) - \frac{SC_{xy}}{SC_{xx}} \sum_i (x_i - \bar{x}) = 0 - \frac{SC_{xy}}{SC_{xx}} \sum_i (x_i - \bar{x}) \overset{DL(a)}{=} 0 - \frac{SC_{xy}}{SC_{xx}} \cdot 0 = 0$$

Simple Linear Regression (Residual Lemma)

Lemma

(Residual Lemma – RL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \ \textit{ where } \ E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then: $(a)\ \sum_i (y_i - \hat{y}_i) = 0 \quad (b)\ \sum_i (y_i - \hat{y}_i) x_i = 0 \quad (c)\ \sum_i (y_i - \hat{y}_i)\hat{y}_i = 0$

PROOF: (b)

$$\sum_i (y_i - \hat{y}_i) x_i \overset{\hat{y}_i}{=} \sum_i \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right] x_i \overset{\hat{\beta}.}{=} \sum_i \left[y_i - \left(\overline{y} - \frac{SC_{xy}}{SC_{xx}} \overline{x} \right) - \left(\frac{SC_{xy}}{SC_{xx}} \right) x_i \right] x_i$$

$$= \sum_i x_i y_i - \overline{y} \sum_i x_i - \frac{SC_{xy}}{SC_{xx}} \sum_i (x_i - \overline{x}) x_i \overset{S.}{=} \left(S_{xy} - \frac{1}{n} S_x S_y \right) - \frac{SC_{xy}}{SC_{xx}} \sum_i (x_i - \overline{x}) x_i$$

$$\overset{CSL}{=} SC_{xy} - \frac{SC_{xy}}{SC_{xx}} \sum_i (x_i - \overline{x}) x_i \overset{DL(b)}{=} SC_{xy} - \frac{SC_{xy}}{SC_{xx}} \cdot SC_{xx} = SC_{xy} - SC_{xy} = 0$$

Simple Linear Regression (Residual Lemma)

Lemma

(Residual Lemma – RL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \;\; \textit{where} \;\; E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\; i = 1, \ldots, n$$

Then: $(a)\ \sum_i (y_i - \hat{y}_i) = 0 \quad (b)\ \sum_i (y_i - \hat{y}_i) x_i = 0 \quad (c)\ \sum_i (y_i - \hat{y}_i)\hat{y}_i = 0$

PROOF: (c)

$$\sum_i (y_i - \hat{y}_i)\hat{y}_i \overset{\hat{y}_i}{=} \sum_i (y_i - \hat{y}_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 \sum_i (y_i - \hat{y}_i) + \hat{\beta}_1 \sum_i (y_i - \hat{y}_i) x_i$$

$$\overset{RL(a)}{=} \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \sum_i (y_i - \hat{y}_i) x_i \overset{RL(b)}{=} \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 = 0 + 0 = 0 \qquad \square$$

Simple Linear Regression (Sums of Squares)

Definition

(Sums of Squares for Simple Linear Regression)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then, there are three key sums of squares:

(Total Variation) $SS_{total} := \sum_i (y_i - \bar{y})^2$

(Unexplained Variation) $SS_{res} := \sum_i (y_i - \hat{y}_i)^2$

(Explained Variation) $SS_{reg} := \sum_i (\hat{y}_i - \bar{y})^2$

Simple Linear Regression (Partitioning Variation)

Theorem

(Sums of Squares Partitioning Variation Theorem – SSPVT)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\ i = 1, \ldots, n$$

Then, the three key sums of squares are partitioned as follows:

$$\underbrace{\textstyle\sum_i (y_i - \overline{y})^2}_{SS_{total}} = \underbrace{\textstyle\sum_i (y_i - \hat{y}_i)^2}_{SS_{res}} + \underbrace{\textstyle\sum_i (\hat{y}_i - \overline{y})^2}_{SS_{reg}}$$

Simple Linear Regression (Partitioning Variation)

Theorem

(Sums of Squares Partitioning Variation Theorem – SSPVT)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

Then, the three key sums of squares are partitioned as follows:

$$\underbrace{\sum_i (y_i - \overline{y})^2}_{SS_{total}} = \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SS_{res}} + \underbrace{\sum_i (\hat{y}_i - \overline{y})^2}_{SS_{reg}}$$

<u>PROOF:</u> $(CIZ \equiv$ **C**lever **I**nsertion of **Z**ero)

$$SS_{total} := \sum_i (y_i - \overline{y})^2 \overset{CIZ}{=} \sum_i \left[y_i - \overline{y} + (\hat{y}_i - \hat{y}_i)\right]^2 = \sum_i \left[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\right]^2$$

$$= \sum_i \left[(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + (\hat{y}_i - \overline{y})^2\right]$$

$$= \sum_i (y_i - \hat{y}_i)^2 + 2 \cdot \sum_i (y_i - \hat{y}_i)\hat{y}_i - 2\overline{y} \cdot \sum_i (y_i - \hat{y}_i) + \sum_i (\hat{y}_i - \overline{y})^2$$

$$\overset{RL}{=} \sum_i (y_i - \hat{y}_i)^2 + 2 \cdot 0 - 2\overline{y} \cdot 0 + \sum_i (\hat{y}_i - \overline{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \overline{y})^2 := SS_{res} + SS_{reg} \qquad \square$$

Simple Linear Regression (Sums of Squares Lemma)

Lemma

(Sums of Squares Lemma – SSL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \ \text{ where } \ E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

...where parameters β_0, β_1 are estimated by OLS point estimators $\hat{\beta}_0, \hat{\beta}_1$.

Then:

$$(a)\ SS_{total} = SC_{yy} \qquad (b)\ SS_{res} = SC_{yy} - \hat{\beta}_1^2 \cdot SC_{xx} \qquad (c)\ SS_{reg} = \hat{\beta}_1^2 \cdot SC_{xx}$$

Simple Linear Regression (SS_{total})

PROOF:

$$
\begin{aligned}
(a)\ \mathrm{SS}_{total} &:= \textstyle\sum_i (y_i - \bar{y})^2 \\
&= \textstyle\sum_i (y_i y_i - 2\bar{y}y_i + \overline{yy}) \\
&= \textstyle\sum_i y_i y_i - \sum_i 2\bar{y}y_i + \sum_i \overline{yy} \\
&= \textstyle\sum_i y_i y_i - 2\bar{y}\sum_i y_i + \overline{yy}\sum_i 1 \\
&\overset{S.}{=} S_{yy} - 2\bar{y}S_y + n\overline{yy} \\
&\overset{\bar{y}}{=} S_{yy} - 2S_y\left(\frac{S_y}{n}\right) + n\cdot\left(\frac{S_y}{n}\right)\cdot\left(\frac{S_y}{n}\right) \\
&= S_{yy} - \frac{2}{n}S_yS_y + \frac{1}{n}S_yS_y \\
&= S_{yy} - \frac{1}{n}S_yS_y \\
&\overset{CSL}{=} SC_{yy}
\end{aligned}
$$

Simple Linear Regression $\left(\text{SS}_{res}\right)$

PROOF:

(b) SS$_{res}$

$$:= \sum_i (y_i^{res})^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 = \sum_i \left[y_i^2 - 2y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2 \right]$$

$$= \sum_i y_i^2 - 2\hat{\beta}_0 \sum_i y_i - 2\hat{\beta}_1 \sum_i x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 \sum_i x_i + \hat{\beta}_1^2 \sum_i x_i^2 + \sum_i \hat{\beta}_0^2$$

$$\overset{S.}{=} S_{yy} - 2\hat{\beta}_0 S_y - 2\hat{\beta}_1 S_{xy} + 2\hat{\beta}_0 \hat{\beta}_1 S_x + \hat{\beta}_1^2 S_{xx} + \hat{\beta}_0^2 n$$

$$\overset{\hat{\beta}_0}{=} S_{yy} - 2S_y \left(\frac{S_y - \hat{\beta}_1 S_x}{n} \right) - 2\hat{\beta}_1 S_{xy} + 2\hat{\beta}_1 S_x \left(\frac{S_y - \hat{\beta}_1 S_x}{n} \right) + \hat{\beta}_1^2 S_{xx} + n \left(\frac{S_y - \hat{\beta}_1 S_x}{n} \right)^2$$

$$= S_{yy} - \frac{2}{n} S_y S_y + \frac{2}{n} \hat{\beta}_1 S_x S_y - 2\hat{\beta}_1 S_{xy} + \frac{2}{n} \hat{\beta}_1 S_x S_y - \frac{2}{n} \hat{\beta}_1^2 S_x S_x + \hat{\beta}_1^2 S_{xx} + \frac{1}{n} S_y S_y - \frac{2}{n} \hat{\beta}_1 S_x S_y + \frac{1}{n} \hat{\beta}_1^2 S_x S_x$$

$$= \left(S_{yy} - \frac{1}{n} S_y S_y \right) + \hat{\beta}_1^2 \left(S_{xx} - \frac{1}{n} S_x S_x \right) - 2\hat{\beta}_1 \left(S_{xy} - \frac{1}{n} S_x S_y \right)$$

$$\overset{CSL}{=} SC_{yy} + \hat{\beta}_1^2 \cdot SC_{xx} - 2\hat{\beta}_1 \cdot SC_{xy} \overset{\hat{\beta}_1}{=} SC_{yy} + \left(\frac{SC_{xy}}{SC_{xx}} \right)^2 \cdot SC_{xx} - 2 \left(\frac{SC_{xy}}{SC_{xx}} \right) \cdot SC_{xy}$$

$$= SC_{yy} - \frac{(SC_{xy})^2}{SC_{xx}} \overset{CIO}{=} SC_{yy} - \frac{(SC_{xy})^2}{(SC_{xx})^2} \cdot SC_{xx} = SC_{yy} - \left(\frac{SC_{xy}}{SC_{xx}} \right)^2 \cdot SC_{xx}$$

$$\overset{\hat{\beta}_1}{=} SC_{yy} - \hat{\beta}_1^2 \cdot SC_{xx}$$

$(CIO \equiv$ **C**lever **I**nsertion of **O**ne$)$

Simple Linear Regression (SS_{reg})

PROOF:

$$
\begin{aligned}
(c)\ \mathsf{SS}_{reg} &= \mathsf{SS}_{total} - \mathsf{SS}_{res} \\
&= SC_{yy} - (SC_{yy} - \hat{\beta}_1^2 \cdot SC_{xx}) \\
&= \hat{\beta}_1^2 \cdot SC_{xx} \qquad \square
\end{aligned}
$$

Simple Linear Regression (Degrees of Freedom)

"... we can think of degrees of freedom as a form of currency."[†]

"In a broad sense, statistics deals with the marketplace of knowledge. Labor, in this framework, corresponds to the task of gathering information. More precisely, our job is to draw a random sample from a population. Each observation we obtain results in a degree of freedom, much like every few minutes of work at a job results in a dollar earned. In statistics, we spend our degrees of freedom to estimate parameters, to increase the probability of reaching correct decisions, or to form models of the way the world behaves..."[†]

"In short, degrees of freedom buy knowledge."[†]

[†] R.S. Schulman, *Statistics in Plain English with Computer Applications*, 1992. (§2.7)

Simple Linear Regression (Degrees of Freedom)

“This number indicates how many independent pieces of information involving the n independent numbers are needed to compile the sum of squares.”[†]

$$\underbrace{\mathrm{SS}_{total}}_{Total\ \ Variation} \quad = \quad \underbrace{\mathrm{SS}_{reg}}_{Variation\ \ due\ \ to\ \ Regression} \quad + \quad \underbrace{\mathrm{SS}_{res}}_{Unexplained\ \ Variation}$$

$$\left(\begin{array}{c}\text{\# dof's in}\\ \text{SS expr}\end{array}\right) = \left(\begin{array}{c}\text{\# total responses}\\ \text{or parameter estimates}\\ \text{in left difference term}\end{array}\right) - \left(\begin{array}{c}\text{\# parameter estimates}\\ \text{in right difference term}\end{array}\right)$$

[†] N.R. Draper, H. Smith, *Applied Regression Analysis*, 3^{rd} Ed., Wiley, 1998. (§1.3)

Simple Linear Regression (Degrees of Freedom)

$$\underbrace{\mathsf{SS}_{total}}_{Total\ \ Variation} = \underbrace{\mathsf{SS}_{reg}}_{Variation\ \ due\ \ to\ \ Regression} + \underbrace{\mathsf{SS}_{res}}_{Unexplained\ \ Variation}$$

$$\sum_i (y_i - \hat{\mu})^2 = \sum_i (\hat{y}_i - \hat{\mu})^2 + \sum_i (y_i^{res})^2$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\sum_i (y_i - \hat{\mu})^2 = \sum_i [(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \hat{\mu}]^2 + \sum_i [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$$\underbrace{\nu}_{Total\ \ dof's} = \underbrace{\nu_{reg}}_{Regression\ \ dof's} + \underbrace{\nu_{res}}_{Residual\ \ dof's}$$

$$\nu = n - 1 \qquad \nu_{reg} = 2 - 1 = 1 \qquad \nu_{res} = n - 2$$

$$\begin{pmatrix} \text{\# dof's in} \\ \text{SS expr} \end{pmatrix} = \begin{pmatrix} \text{\# total responses} \\ \text{or parameter estimates} \\ \text{in left difference term} \end{pmatrix} - \begin{pmatrix} \text{\# parameter estimates} \\ \text{in right difference term} \end{pmatrix}$$

Simple Linear Regression (Degrees of Freedom)

Seriously, why does SS$_{reg}$ have only one degree of freedom???

$$\text{SS}_{reg} := \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

"Although there are n deviations $(\hat{y}_i - \overline{y})$, all fitted values $\hat{y}_i$ are calculated from the same estimated regression line. Two degrees of freedom are associated with a regression line, ..."$^\diamondsuit$

SS$_{reg}$ "has 1 degree of freedom associated with it. There are two parameters in the regression equation, but the deviations $(\hat{y}_i - \overline{y})$ are subject to the constraint $\sum_i (\hat{y}_i - \overline{y}) = 0$."†

$^\diamondsuit$J. Neter, *Applied Linear Regression Models*, 3^{rd} Ed., McGraw-Hill, 1996. (§2.7)

†J. Neter, W. Wasserman, *Applied Linear Statistical Models*, Irwin Inc., 1974. (§3.8)

PART III

PART III:

Simple Linear Regression Analysis

Estimation of β_1 & σ^2

$\hat{\beta}_1$ as Linear Combo of y_i's

Expectation of $\hat{\beta}_1$

Variance of $\hat{\beta}_1$

EE Lemma

Expectation of SS_{res}

Mean Squared Residual, MS$_{res}$

Estimation of σ^2

Variation Explanation (R^2)

Simple Linear Regression ($\hat{\beta}_1$ as Linear Combo of y_i's)

Prior to determination of the Expectation & Variance of $\hat{\beta}_1$,
it is convenient to first rewrite $\hat{\beta}_1$ as a linear combination of the responses (y_i):

Lemma

($\hat{\beta}_1$ as Linear Combination of Responses – BH1LCR)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

...where parameters β_0, β_1 are estimated by OLS point estimators $\hat{\beta}_0, \hat{\beta}_1$.

Then, $\hat{\beta}_1$ can be written as so: $\hat{\beta}_1 = \sum_i \xi_i y_i$ *where* $\xi_i := \dfrac{(x_i - \overline{x})}{SC_{xx}}$

<u>PROOF:</u> $\hat{\beta}_1 \overset{OLS}{=} \dfrac{SC_{xy}}{SC_{xx}} \overset{DL(c)}{=} \dfrac{\sum_i (x_i - \overline{x}) y_i}{SC_{xx}} = \sum_i \left(\dfrac{x_i - \overline{x}}{SC_{xx}} \right) \cdot y_i := \sum_i \xi_i y_i \qquad \square$

Simple Linear Regression (Expectation of $\hat{\beta}_1$)

The expectation & variance of $\hat{\beta}_1$ are needed in the estimation of σ^2.
The expectation & variance of $\hat{\beta}_1$ are also used in inference later.

Theorem

(Expectation & Variance of $\hat{\beta}_1$ Theorem – EVB1HT)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \;\; \textit{where} \;\; E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

...where parameters β_0, β_1 are estimated by OLS point estimators $\hat{\beta}_0, \hat{\beta}_1$.

$$\textit{Then:} \qquad (a)\ \mathbb{E}[\hat{\beta}_1] = \beta_1 \qquad (b)\ \mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{SC_{xx}}$$

PROOF:

$$(a)\ \mathbb{E}[\hat{\beta}_1] \overset{BH1LCR}{=} \mathbb{E}\left[\sum_i \xi_i Y_i\right] = \sum_i \mathbb{E}\left[\xi_i Y_i\right] = \sum_i \xi_i \cdot \mathbb{E}\left[Y_i\right] = \sum_i \xi_i \cdot (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_i \xi_i + \beta_1 \sum_i \xi_i x_i = \frac{\beta_0}{SC_{xx}} \sum_i (x_i - \bar{x}) + \frac{\beta_1}{SC_{xx}} \sum_i (x_i - \bar{x}) x_i \overset{DL}{=} \frac{\beta_0}{SC_{xx}} \cdot 0 + \frac{\beta_1}{SC_{xx}} \cdot SC_{xx} = \beta_1$$

Simple Linear Regression (Variance of $\hat{\beta}_1$)

The expectation & variance of $\hat{\beta}_1$ are needed in the estimation of σ^2.
The expectation & variance of $\hat{\beta}_1$ are also used in inference later.

Theorem

(Expectation & Variance of $\hat{\beta}_1$ Theorem – EVB1HT)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \textit{where} \quad E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

...where parameters β_0, β_1 are estimated by OLS point estimators $\hat{\beta}_0, \hat{\beta}_1$.

$$\textit{Then:} \qquad (a)\ \mathbb{E}[\hat{\beta}_1] = \beta_1 \qquad\qquad (b)\ \mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{SC_{xx}}$$

PROOF:

$$(b)\ \mathbb{V}[\hat{\beta}_1] \overset{BH1LCR}{=} \mathbb{V}\left[\sum_i \xi_i Y_i\right] = \sum_i \mathbb{V}\left[\xi_i Y_i\right] = \sum_i \xi_i^2 \cdot \mathbb{V}\left[Y_i\right] = \sum_i \xi_i^2 \cdot \sigma^2$$

$$= \sigma^2 \sum_i \xi_i^2 = \sigma^2 \sum_i \frac{(x_i - \overline{x})^2}{SC_{xx}^2} = \frac{\sigma^2}{SC_{xx}^2} \sum_i (x_i - \overline{x})^2 \overset{SC.}{=} \frac{\sigma^2}{SC_{xx}^2} \cdot SC_{xx} = \frac{\sigma^2}{SC_{xx}} \qquad \square$$

Simple Linear Regression (EE Lemma)

Lemma

(EE Lemma – EEL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

Then: $(a)\ \mathbb{E}\left[\sum_i (E_i - \overline{E})\right] = 0 \qquad (b)\ \mathbb{E}[E_i\overline{E}] = \frac{\sigma^2}{n} \qquad (c)\ \mathbb{E}\left[\sum_i (E_i - \overline{E})^2\right] = (n-1)\sigma^2$

PROOF:

$(1)\ E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2) \implies \mathbb{E}[E_i] = 0$ and $\overline{E} \sim \text{Normal}(0, \sigma^2/n) \implies \mathbb{E}[\overline{E}] = 0$

$(2)\ E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2) \implies \mathbb{E}[E_iE_j] \overset{IND}{=} \mathbb{E}[E_i] \cdot \mathbb{E}[E_i] = \delta_{ij}$

$$(a)\ \mathbb{E}\left[\sum_i (E_i - \overline{E})\right] = \sum_i \left[\mathbb{E}[E_i] - \mathbb{E}[\overline{E}]\right] \overset{(1)}{=} \sum_i (0 - 0) = n \cdot 0 = 0$$

$$(b)\ \mathbb{E}[E_i\overline{E}] = \mathbb{E}\left[E_i \cdot \frac{1}{n}\sum_j E_j\right] \overset{(2)}{=} \frac{1}{n} \cdot \mathbb{E}\left[E_i^2\right] = \frac{1}{n} \cdot \left(\mathbb{V}[E_i] + \mathbb{E}[E_i]^2\right) = \frac{1}{n} \cdot \left(\sigma^2 + 0^2\right) = \frac{\sigma^2}{n}$$

$$(c)\ \mathbb{E}\left[\sum_i (E_i - \overline{E})^2\right] = \sum_i \mathbb{E}\left[(E_i - \overline{E})^2\right] = \sum_i \left[\mathbb{E}\left[E_i^2\right] - 2 \cdot \mathbb{E}[E_i\overline{E}] + \mathbb{E}[\overline{E}^2]\right]$$

$$= \sum_i \left[\left(\mathbb{V}[E_i] + \mathbb{E}[E_i]^2\right) + \left(\mathbb{V}[\overline{E}] + \mathbb{E}[\overline{E}]^2\right) - 2 \cdot \mathbb{E}[E_i\overline{E}]\right] \overset{EEL(b)}{=} \sum_i \left[(\sigma^2 + 0^2) + \left(\frac{\sigma^2}{n} + 0^2\right) - \frac{2\sigma^2}{n}\right]$$

$$= \sum_i \left[\sigma^2 + \frac{\sigma^2}{n} - \frac{2\sigma^2}{n}\right] = n\sigma^2 + \sigma^2 - 2\sigma^2 = (n-1)\sigma^2 \qquad \square$$

Simple Linear Regression (Expectation of SS_{res})

Lemma

(Estimation of SS_{res} Lemma – ESSRESL)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \ \textit{ where } \ E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

Then:

$$\mathbb{E}\left[\textit{SS}_{res}\right] = (n-2)\sigma^2$$

Simple Linear Regression (Expectation of SS_{res})

PROOF:

$$\begin{aligned} & Y_i = \beta_0 + \beta_1 x_i + E_i && \text{where} && E_i \overset{IID}{\sim} \text{Normal}(0,\sigma^2) \\ \Longrightarrow\ & \overline{Y} = \beta_0 + \beta_1 \overline{x} + \overline{E} && \text{where} && \overline{E} \sim \text{Normal}(0,\sigma^2/n) \end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\text{SS}_{res}\right] &= \mathbb{E}\left[\sum_i (Y_i - \hat{Y}_i)^2\right] \overset{SSL}{=} \mathbb{E}\left[SC_{yy} - \hat{\beta}_1^2 \cdot SC_{xx}\right] \overset{SC_{yy}}{=} \mathbb{E}\left[\sum_i (Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \cdot SC_{xx}\right] \\
&\overset{y}{=} \mathbb{E}\left[\sum_i \left[(\beta_0 + \beta_1 x_i + E_i) - (\beta_0 + \beta_1 \overline{x} + \overline{E})\right]^2\right] - \mathbb{E}\left[\hat{\beta}_1^2 \cdot SC_{xx}\right] \\
&= \mathbb{E}\left[\sum_i \left[\beta_1 (x_i - \overline{x}) + (E_i - \overline{E})\right]^2\right] - \mathbb{E}\left[\hat{\beta}_1^2 \cdot SC_{xx}\right] \\
&= \mathbb{E}\left[\beta_1^2 \cdot SC_{xx} + 2\beta_1 \cdot \sum_i (x_i - \overline{x})(E_i - \overline{E}) + \sum_i (E_i - \overline{E})^2\right] - \mathbb{E}\left[\hat{\beta}_1^2 \cdot SC_{xx}\right] \\
&= \mathbb{E}\left[\beta_1^2 \cdot SC_{xx}\right] + 2\beta_1 \cdot \mathbb{E}\left[\sum_i (x_i - \overline{x})(E_i - \overline{E})\right] + \mathbb{E}\left[\sum_i (E_i - \overline{E})^2\right] - \mathbb{E}\left[\hat{\beta}_1^2 \cdot SC_{xx}\right] \\
&= \beta_1^2 \cdot SC_{xx} + 2\beta_1 \cdot \sum_i (x_i - \overline{x}) \mathbb{E}\left[(E_i - \overline{E})\right] + \sum_i \mathbb{E}\left[(E_i - \overline{E})^2\right] - \mathbb{E}\left[\hat{\beta}_1^2 \cdot SC_{xx}\right] \\
&\overset{EEL}{=} \beta_1^2 \cdot SC_{xx} + 2\beta_1 \cdot 0 + (n-1)\sigma^2 - SC_{xx} \cdot \mathbb{E}\left[\hat{\beta}_1^2\right] \\
&= \beta_1^2 \cdot SC_{xx} + (n-1)\sigma^2 - SC_{xx} \cdot \left[\mathbb{V}\left[\hat{\beta}_1\right] + \mathbb{E}\left[\hat{\beta}_1\right]^2\right] \\
&\overset{EVB1HT}{=} \beta_1^2 \cdot SC_{xx} + (n-1)\sigma^2 - SC_{xx} \cdot \left[\frac{\sigma^2}{SC_{xx}} + \beta_1^2\right] \\
&= \beta_1^2 \cdot SC_{xx} + (n-1)\sigma^2 - \sigma^2 - \beta_1^2 \cdot SC_{xx} \\
&= (n-2)\sigma^2 \qquad \square
\end{aligned}$$

D. Wackerly *et al*, *Mathematical Statistics with Applications*, 5th Ed, Duxbury Press, 1996. (§11.4)

Simple Linear Regression (Mean Squared Residual)

Definition

(Mean Squared Residual)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \dots, n$$

...where parameters β_0, β_1 are estimated by OLS estimators $\hat{\beta}_0, \hat{\beta}_1$.

Then:

$$\text{MS}_{res} := \frac{\text{SS}_{res}}{n-2}$$

Simple Linear Regression (Point Estimator of σ^2)

Proposition

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \;\; \textit{where} \;\; E_1, \cdots, E_n \overset{IID}{\sim} \textit{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\, i = 1, \ldots, n$$

...where parameters β_0, β_1 are estimated by OLS point estimators $\hat{\beta}_0, \hat{\beta}_1$.

Then:

$$\mathbb{E}\left[\mathit{MS}_{res}\right] = \sigma^2 \implies \hat{\sigma}^2 = \mathit{MS}_{res}$$

i.e. MS_{res} is always an unbiased estimator of σ^2.

PROOF:

$$\mathbb{E}[\mathrm{MS}_{res}] := \mathbb{E}\left[\frac{\mathrm{SS}_{res}}{n-2}\right] = \frac{1}{n-2} \cdot \mathbb{E}[\mathrm{SS}_{res}] \overset{ESSRESL}{=} \frac{1}{n-2} \cdot (n-2)\sigma^2 = \sigma^2 \qquad \square$$

Simple Linear Regression (Variation Explanation, R^2)

Definition

(Coefficient of Determination, R^2)

Given a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \text{ where } E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$$

and the corresponding realized model:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\ i = 1, \ldots, n$$

...where parameters β_0, β_1 are estimated by OLS point estimators $\hat{\beta}_0, \hat{\beta}_1$.

Then the **coefficient of determination** is: $R^2 := \dfrac{\text{SS}_{reg}}{\text{SS}_{total}} = \dfrac{SC_{xy} \cdot SC_{xy}}{SC_{xx} \cdot SC_{yy}}$

Fin

Fin.