SIMPLE LINEAR REGRESSION:
OLS POINT ESTIMATORS OF $\hat{\beta}_0$ & $\hat{\beta}_1$ [DEVORE 12.2]

SIMPLE LINEAR REGRESSION MODEL (DEFINITION):

The **simple linear regression model** is: $Y_i = \beta_0 + \beta_1 x_i + E_i$ where $E_1, \cdots, E_n \overset{IID}{\sim} \text{Normal}(0, \sigma^2)$

$$\begin{array}{rcl} Y_i & \equiv & \text{rv for measurement of } i^{th} \text{ response} \\ x_i & \equiv & \text{Actual measurement of } i^{th} \text{ regressor} \\ \beta_0 & \equiv & \text{Expected value of response when regressor is zero} \\ \beta_1 & \equiv & \text{Expected change in response per unit increase in regressor} \\ E & \equiv & \text{Effect of random error on } i^{th} \text{ response} \end{array}$$

The **realized simple linear regression model** is: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \forall\ i = 1, \ldots, n$

$$\begin{array}{rcl} y_i & \equiv & \text{Actual measurement of } i^{th} \text{ response} \\ x_i & \equiv & \text{Actual measurement of } i^{th} \text{ regressor} \\ \hat{\beta}_0 & \equiv & \text{Estimated value of response when regressor is zero} \\ \hat{\beta}_1 & \equiv & \text{Estimated change in response per unit increase in regressor} \\ e_i & \equiv & \text{Actual error on } i^{th} \text{ response} \end{array}$$

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

$$S_x := \textstyle\sum_i x_i \qquad S_y := \sum_i y_i \qquad S_{xx} := \sum_i x_i x_i \qquad S_{yy} := \sum_i y_i y_i \qquad S_{xy} := \sum_i x_i y_i$$

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

$$(a)\ \ (S_{xy})^2 \le S_{xx} \cdot S_{yy} \qquad\qquad (b)\ \ (S_x)^2 \le n \cdot S_{xx}$$

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 $k \in \mathbb{R}$.

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

$$SC_{xx} := \textstyle\sum_i (x_i - \overline{x})^2 \qquad SC_{yy} := \sum_i (y_i - \overline{y})^2 \qquad SC_{xy} := \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

CENTERED SUMS LEMMA:

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

OLS ESTIMATORS:

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

GAUSS-MARKOV THEOREM: Suppose the following conditions are all satisfied:

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

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

SIMPLE LINEAR REGRESSION: SUMS OF SQUARES [DEVORE 12.2]

DEVIATION LEMMA:

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

RESIDUAL LEMMA:

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

SUMS OF SQUARES (DEFINITION):

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

VARIATION PARTITIONING:

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

SUMS OF SQUARES LEMMA:

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

DEGREES OF FREEDOM:

$$\begin{array}{ccccc}
\underbrace{\mathrm{SS}_{total}}_{Total\ Variation} & = & \underbrace{\mathrm{SS}_{reg}}_{Variation\ due\ to\ Regression} & + & \underbrace{\mathrm{SS}_{res}}_{Unexplained\ Variation} \\
\sum_i(y_i-\hat{\mu})^2 & = & \sum_i(\hat{y}_i-\hat{\mu})^2 & + & \sum_i(y_i^{res})^2 \\
\sum_i(y_i-\overline{y})^2 & = & \sum_i(\hat{y}_i-\overline{y})^2 & + & \sum_i(y_i-\hat{y}_i)^2 \\
\hline
\sum_i(y_i-\hat{\mu})^2 & = & \sum_i[(\hat{\beta}_0+\hat{\beta}_1x_i)-\hat{\mu}]^2 & + & \sum_i[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]^2 \\
\underbrace{\nu}_{Total\ dof's} & = & \underbrace{\nu_{reg}}_{Regression\ dof's} & + & \underbrace{\nu_{res}}_{Residual\ dof's} \\
\nu=n-1 & & \nu_{reg}=2-1=1 & & \nu_{res}=n-2
\end{array}$$

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

"... 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:
POINT ESTIMATION OF σ^2 [DEVORE 12.2]

$\hat{\beta}_1$ AS LINEAR COMBINATION OF RESPONSES:

$$\hat{\beta}_1 = \textstyle\sum_i \xi_i y_i \quad \text{where} \quad \xi_i := \dfrac{(x_i - \overline{x})}{SC_{xx}}$$

EXPECTATION & VARIANCE OF $\hat{\beta}_1$:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \qquad\qquad \mathbb{V}[\hat{\beta}_1] = \frac{\sigma^2}{SC_{xx}}$$

EE LEMMA:

$$(a)\ \mathbb{E}\left[\textstyle\sum_i (E_i - \overline{E})\right] = 0 \quad (b)\ \mathbb{E}[E_i \overline{E}] = \dfrac{\sigma^2}{n} \quad (c)\ \mathbb{E}\left[\textstyle\sum_i (E_i - \overline{E})^2\right] = (n-1)\sigma^2$$

EXPECTATION OF SS_{res}:

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

MEAN SQUARED RESIDUAL:

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

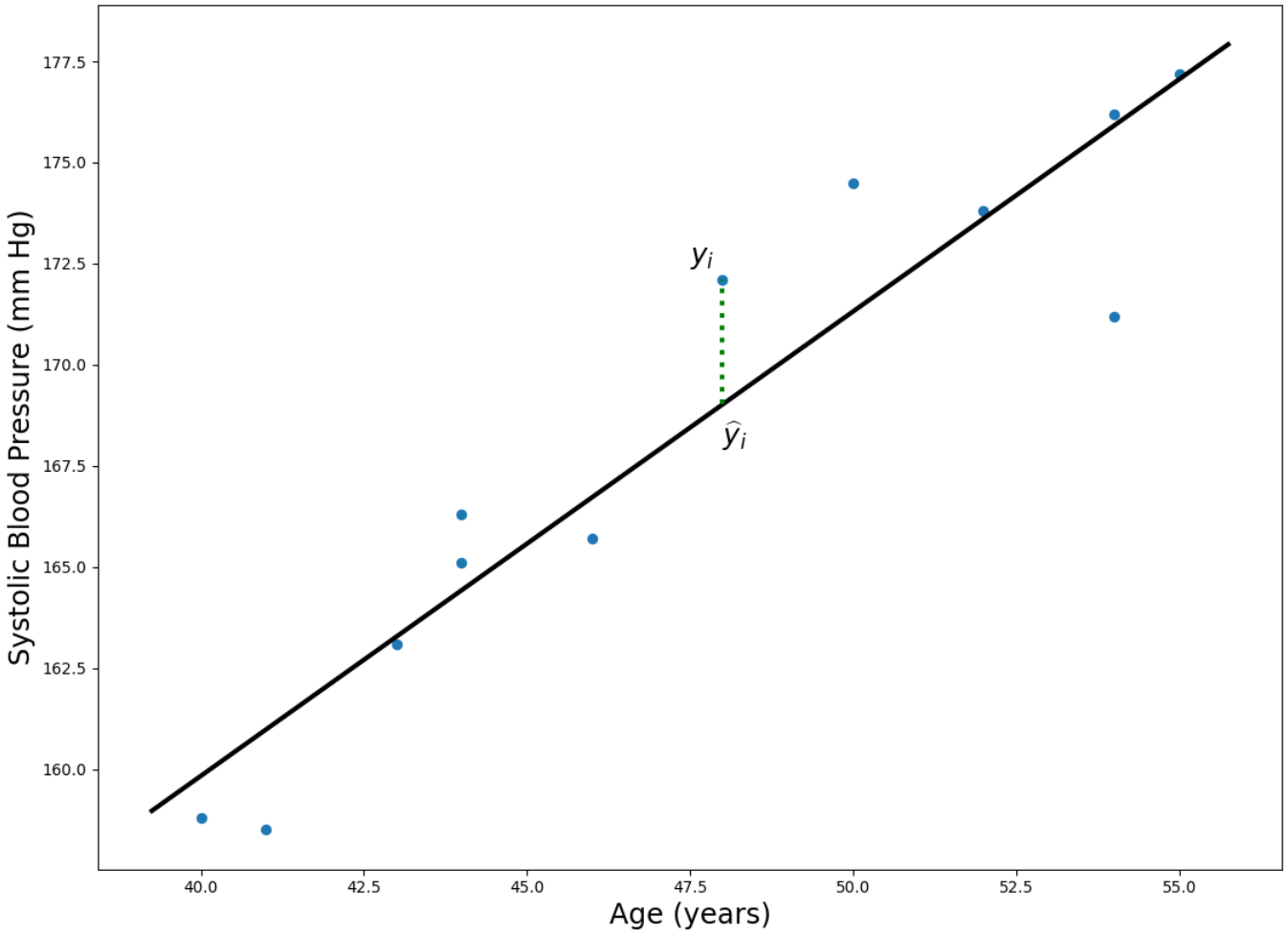
POINT ESTIMATOR OF σ^2:

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

VARIATION EXPLANATION (R^2):

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

SIMPLE LINEAR REGRESSION: VISUALIZATION [DEVORE 12.2]



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

EX 12.2.1: The systolic blood pressures of twelve adults were measured as a single group without intervention[†]:

SYSTOLIC BLOOD PRESSURE (in mmHg) versus Age (in years)													**TOTAL**
Index (i)	1	2	3	4	5	6	7	8	9	10	11	12	$n = 12$
Age (x_i)	40	41	43	44	44	46	48	50	52	54	54	55	$\sum_i x_i = 571$
BP (y_i)	158.8	158.5	163.1	165.1	166.3	165.7	172.1	174.5	173.8	176.2	171.2	177.2	$\sum_i y_i = 2022.5$

[†](Simplification of data) Y.H. Chan, "Biostatistics 201: Linear Regression Analysis", *Singapore Med. J.*, **45** (2004), 55-61.

(a) Identify the response and the sole regressor in this observational study.

(b) Formulate this observational study as a linear model.

(c) Compute the response sum and regressor sum: $S_y,\ S_x$.

(d) Compute the squared regressor sum, S_{xx}.

(e) Compute the cross-termed regressor-response sum, S_{xy}.

(f) Compute the cross-termed regressor-response centered sum, SC_{xy}.

(g) Compute the squared regressor centered sum, SC_{xx}.

(h) Compute the ordinary least-squares (OLS) estimates of the regression parameters $\hat{\beta}_0, \hat{\beta}_1$.

(i) Determine the OLS best-fit line.

(j) Compute the squared response sum, S_{yy}.

(k) Compute the squared response centered sum, SC_{yy}.

(ℓ) Compute the residual sum of squares, SS_{res}.

(m) Compute the linear model's estimated variance, $\hat{\sigma}^2$.

(n) Compute R^2.