$$E(XY) = Cov(X,Y) + E(X)E(Y)$$
$$Cov(X,Y) = E[(X-E(X))(Y-E(Y))]$$
$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x}_i)^2 = \sum_{i=1}^{n}(x_i - \bar{x}_i)x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}_i^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x}_i)y_i = \sum_{i=1}^{n} x_i y_i - n\bar{x}_i\bar{y}.$$

Y_i's conditionally independent given x_i's
$$Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 X_i + \varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j)$$
$= \beta_1^2 \underline{Cov(X_i, X_j)}$ NOT necessarily 0

[linear approximation]
$$e^x = e^{x_0} + e^{x_0}(x-x_0) + \frac{1}{2}e^{x_0}(x-x_0)^2 + \cdots$$
if x is close enough to x_0,
then $e^x \approx e^{x_0} + e^{x_0}(x-x_0) = e^{x_0}(1-x_0) + e^{x_0}x$.
$$e^{x_0}|x-x_0| \gg \frac{1}{2}e^{x_0}|x-x_0|^2$$
$$2 \gg |x-x_0|$$
$\Rightarrow x \in (x_0-2, x_0+2)$ at least

$\boxed{\text{Optimal Linear Predictor}}$

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0, \beta_1)}{\operatorname{argmin}}\, E_{X,Y}[(Y-(\beta_0+\beta_1 X))^2]$$

$\beta_0^* = E(Y) - \beta_1^* E(X)$ make sure the line goes through $E(Y)$ at $E(X)$.

$\beta_1^* = \frac{Cov(X,Y)}{Var(X)}$ ↑ X and Y tend to fluctuate together. ↓ the more X fluctuate.

$\boxed{\text{Estimates}}$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}_i$$

Least squares estimate

$$S(\beta_0, \beta_1) = \hat{MSE}(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$
$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Matrix format

$$S(\beta) = \hat{MSE}(\beta) = \frac{1}{n}(Y - X\beta)^T(Y-X\beta)$$
$$\Rightarrow \hat{\beta} = (X^TX)^{-1}X^TY$$
$$\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$\boxed{\text{Bias \& Variance of Estimators}}$

$$E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_0) = \beta_0$$
$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}_i^2}{S_{xx}}\right)$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} \underbrace{\frac{x_i - \bar{x}_i}{S_{xx}}}_{c_i} y_i \qquad \sum_{i=1}^{n} c_i = 0 \qquad \sum_{i=1}^{n} c_i x_i = 1$$
$$\hat{\beta}_0 = \sum_{i=1}^{n} \underbrace{\left(\frac{1}{n} - \bar{x}_i c_i\right)}_{b_i} y_i \qquad \sum_{i=1}^{n} b_i = 1 \qquad \sum_{i=1}^{n} b_i x_i = 0$$

$$Var(\hat{\beta}_0 \mid X_1 \cdots X_n)$$
$$= Var(\bar{Y} - \hat{\beta}_1\bar{x}_i \mid X_1 \cdots X_n)$$
$$= Var(\bar{Y} \mid X_1 \cdots X_n) + \bar{x}_i^2 Var(\hat{\beta}_1 \mid X_1 \cdots X_n) - 2\bar{x}_i Cov(\bar{Y}, \hat{\beta}_1 \mid X_1 \cdots X_n)$$
$$= \frac{\sigma^2}{n} + \bar{x}_i^2\frac{\sigma^2}{S_{xx}}$$

in Matrix Form

$Y = X\beta + \varepsilon$ where $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$]

$$\hat{\beta} = \underbrace{(X^TX)^{-1}X^T}_{A}Y$$
$$Var(\hat{\beta} \mid X) = Var(AY \mid X) = A\,Var(Y \mid X)A^T = A\sigma^2 I A^T = \sigma^2 AA^T = \sigma^2(X^TX)^{-1}$$
$$\begin{bmatrix} Var(\hat{\beta}_0 \mid X) & Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X) \\ Cov(\hat{\beta}_1, \hat{\beta}_0 \mid X) & Var(\hat{\beta}_1 \mid X) \end{bmatrix} = \frac{1}{nS_{xx}}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

① The law of Total Expectation
$$E(\hat{\beta}_1) = E_X[E_{Y|X}(\hat{\beta}_1 \mid X_1 \cdots X_n)] = E_X(\beta_1) = \beta_1$$

② The law of Total Variance
$$Var(\hat{\beta}_1) = E_X[Var_{Y|X}(\hat{\beta}_1 \mid X_1 \cdots X_n)] + Var[E_{Y|X}(\hat{\beta}_1 \mid X_1 \cdots X_n)]$$
$$= \sigma^2 E\left(\frac{1}{S_{xx}}\right) + Var(\beta_1) = \frac{\sigma^2}{S_{xx}}$$

Residuals

$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}$

$$\begin{cases} \frac{\partial S}{\partial \beta_0} = \frac{1}{n}(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) = 0 \\ \frac{\partial S}{\partial \beta_1} = \frac{1}{n}(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})x_{i1} = 0 \end{cases}$$

① $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$
② $\sum_{i=1}^{n} e_i x_{i1} = 0$
③ $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i$
④ $\sum_{i=1}^{n}\hat{y}_i e_i = 0$

residual sum of squares

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 - \hat{\beta}_1 S_{xy}$$

explained sum of squares $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ SST total sum of squares.

The unbiased estimator of σ^2 is $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$ residual mean square

The standard error (se) of the estimators are the square root of the variance

$$se(\hat{\beta}_0) = \sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \qquad se(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}}$$

Replace σ^2 with $\hat{\sigma}^2$, then we get estimated standard error (ese)

$$ese(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \qquad ese(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{\text{hat matrix}} Y = HY$$

① H is systemmetic
② H is idempotent $H^TH = H$
③ $I_n - H$ is idempotent

$$\begin{aligned} SS_{Res} &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \\ &= (Y - HY)^T(Y - HY) \\ &= Y^T(I_n - H)^T(I_n - H)Y \\ &= Y^T(I_n - H) \end{aligned}$$

If v is a random vector, $E(v) = \mu$, $Var(v) = \Sigma$
For a constant matrix A, $E(v^TAv) = trace(A\Sigma) + \mu^TA\mu$

$$\begin{aligned} E(SS_{Res}|X) &= E[Y^T(I_n - H)Y|X] \\ &= trace((I_n - H)\sigma^2 I_n) + (X\beta)^T(I_n - H)(X\beta) \\ &= \sigma^2 trace(I_n - H) + \underbrace{\beta^TX^T(I_n - H)(X\beta)}_{0} \\ &= \sigma^2 trace(I_n) - \sigma^2 trace(H) + 0 \\ &= \sigma^2(n - trace(H)) \end{aligned}$$

$$\begin{aligned} X^T(I_n - H)X &= X^TX - X^THX \\ &= X^TX - X^T\underline{X(X^TX)^{-1}X^TX} \\ &= X^TX - X^TX \\ &= 0 \end{aligned}$$

$$\begin{aligned} trace(H) &= trace(X(X^TX)^{-1}X^T) \\ &= trace(X^TX(X^TX)^{-1}) \\ &= trace(I_p) \\ &= p \end{aligned}$$

where $p = 2$ for SLR

$\Rightarrow E(SS_{Res}|X) = \sigma^2(n-2)$

Gaussian-Noise SLR

same:
① The distribution of X is arbitrary (and perhaps X is even non-random
② $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. If $X_1 = x_1$, then $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ for some coefficients β_0, β_1 and random noise ε.

③ $\varepsilon \sim N(0, 1)$ $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$

④ ε is independent of X_1 and independent across observations. ε is uncorrelated with X_1.
Linear correlation $\Rightarrow$ those having 0 correlation might dependent non-linearly.

① Sampling distribution of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p} \quad (p = 2 \text{ for SLR})$$

② Confidence interval

$$\frac{\hat{\beta}_1 - \overset{\text{unknown}}{\beta_1}}{\underset{\text{unknown}}{se(\hat{\beta}_1)}} \sim N(0,1) \qquad \frac{\hat{\beta}_0 - \overset{\text{unknown}}{\beta_0}}{\underset{\text{unknown}}{se(\hat{\beta}_0)}} \sim N(0,1)$$

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{ese(\hat{\beta}_1)} \sim t_{n-p} \qquad T_0 = \frac{\hat{\beta}_0 - \beta_0}{ese(\hat{\beta}_0)} \sim t_{n-p}$$

where $p = 2$ for GN-SLR

A $1-\alpha$ level confidence interval

$$\begin{cases} CI(\beta_1) = [\hat{\beta}_1 - k\cdot ese(\hat{\beta}_1), \hat{\beta}_1 + k\cdot ese(\hat{\beta}_1)] \\ CI(\beta_0) = [\hat{\beta}_0 - k\cdot ese(\hat{\beta}_0), \hat{\beta}_0 + k\cdot ese(\hat{\beta}_0)] \end{cases}$$

$k \equiv t_{\frac{\alpha}{2}, n-2} > 0$

[Interpretation]

- The interval $CI(\beta_1)$ traps β_1 with probability $1-\alpha$.
- β_1 is non-random
- The width of the CI is $2\cdot k\, ese(\hat{\beta}_1)$
  As α shrinks, the interval widens
  As n grows, the interval shrinks.
  As σ^2 increases, the interval widens.
  As S_{xx} grows, the interval shrinks

③ Hypothesis Testing

$$\begin{cases} H_0: \beta_1 = c \\ H_1: \beta_1 \neq c \end{cases}$$

We reject H_0 if $|T_1| \geq k \equiv t_{\frac{\alpha}{2}, n-2}$
$P(|T| > |T_1|) < \alpha$.

$$T_1 = \frac{\hat{\beta}_1 - c}{ese(\hat{\beta}_1)}$$