1.4 Assess Effectiveness of the Least Square Line

How effectively does the least squares line summarize the data?

- Residuals

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \Leftrightarrow \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Estimate of ε_i: $e_i = Y_i - (b_0 + b_1 X_i) = Y_i - \hat{Y}_i$ — *residual*

Observations		Fitted values $\hat{Y}_i = b_0 + b_1 X_i$	Residuals $e_i = Y_i - \hat{Y}_i$
X_1	Y_1	$\hat{Y}_1$	$Y_1 - \hat{Y}_1$
X_2	Y_2	$\hat{Y}_2$	$Y_2 - \hat{Y}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_n	Y_n	$\hat{Y}_n$	$Y_n - \hat{Y}_n$

— Residuals are building blocks for measuring how well the model fits the data.

Result: For the SLR, $\sum_{i=1}^{n} e_i = 0 \Leftrightarrow \bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0$.

Does the fitted line summarize the data effectively?

- Decompose the variability in Y

— If we had no information about an explanatory variable X, the observed variability in Y is summarized by the sample variance $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ or sample standard deviation s_Y.

— Decompose the deviation of Y_i from $\bar{Y}$

$$(Y_i - \bar{Y}) = \left(\hat{Y}_i - \bar{Y}\right) + (Y_i - \hat{Y}_i)$$

— Square both sides

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

— Sum over i

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \\ &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\end{aligned}$$

Note: $\sum_{i=1}^{n}(\hat{Y}_i \ -\bar{Y})(Y_i-\hat{Y}_i)=0.$

Result: $\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2.$

Total sum of squares(TSS) = Regression sum of squares($\mathrm{SS_{reg}}$) + Residual sum of squares(RSS).
Total variability in Y = Variability in Y explained by the LS line +Variability about the LS line.

Note: TSS = S_{YY}, $\mathrm{SS_{reg}}$ = S_{XY}^2/S_{XX}, RSS = TSS − $\mathrm{SS_{reg}}$.

Definition 1.1 (Degrees of freedom). *Degrees of freedom* (df) associated with a sum of squares (SS) is the number of distinct independent pieces of information that make up the SS, which can be calculated by the following formula.

df = #{distinct terms in SS} − #{linear constraints on deviations}

Ex. 1.2 Find the degrees of freedom of TSS, RSS, and $\mathrm{SS_{reg}}$

Definition 1.2 (Mean square). A sum of squares divided by its degrees of freedom is called a *mean square*, that is,

$$MS_x=\frac{\mathrm{SS}_x}{df\ \mathrm{of}\ \mathrm{SS}_x},\ e.g., \mathrm{MS\ of\ residual}=\frac{\mathrm{RSS}}{df\ \mathrm{of\ RSS}}\ .$$

Summarize the decomposition with an Analysis of Variance (ANOVA) table

Source of variation	df	SS	MS
Regression	1	$\mathrm{SS_{reg}}$	$\mathrm{MSreg}=\frac{\mathrm{SSreg}}{1}=\mathrm{SSreg}$
Residual	n−2	RSS	$\mathrm{MSresid}=\frac{RSS}{n-2}$
Total	n – 1	TSS	

- Estimating $\sigma^2 = Var(\varepsilon_i)$

Note that the errors ε_i are estimated by the residuals $e_i = Y_i - \hat{Y}_i$, $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}(e_i - \frac{\sum_{i=1}^{n} e_i}{n})^2$ and df associated with RSS is $n-2$. Therefore estimate σ^2 in SLR by

$$s^2 = \frac{RSS}{n-2} = \text{MSresid} \qquad \text{— the residual mean square.}$$

Estimate σ by $s = \sqrt{s^2}$.

s is called the *standard error of regression* (or *root mean square error*), which is interpreted as the typical amount by which an observation deviates from the fitted line (or the average amount by which observations deviate from the least squares line).

The smaller the s is, the more effective the LS line is.

Q: Why not estimate σ^2 by $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$?
A: Want to use the information of X?

- Coefficient of determination

$R^2 = \frac{SSreg}{TSS}$ — the proportion of variation in Y explained by the regression line.
100 R^2 — the percentage of variation in Y explained by the regression line.

The larger the R^2 is, the more effective the LS line is.

Note: A good least squares line should have a large R^2 and a small s.

Properties of R^2

(1) $0 \le R^2 \le 1$.
(2) $R^2 = r_{XY}^2$.

Q: Given R^2, how to find r_{XY}?

Ex. 1.1 (continued) The National Institute of Health is studying the relationship between the number of cigarettes smoked per day and the birthweight of babies born to mothers who smoke cigarettes. The following data are observed.

No. of cigarettes per day(X)	21	12	28	10	24	5
Birthweight (Y)	6.0	8.0	5.6	7.5	6.2	8.5

(5) Calculate and interpret R^2 and s.

1.5 Inferences about Regression Parameters

- Important distribution results

Distribution result 1: If $U_1, U_2, \cdots, U_k$ are independent and $U_i \sim N(\mu_i, \sigma_i^2)$.
(1) $c_0 + \sum_{i=1}^{k} c_i U_i \sim N\left(c_0 + \sum_{i=1}^{k} c_i \mu_i, \sum_{i=1}^{k} c_i^2 \sigma_i^2\right)$.
(2) $\frac{U_i - \mu_i}{\sigma_i}$ are iid $N(0, 1)$.
(3) $\sum_{i=1}^{k} (\frac{U_i - \mu_i}{\sigma_i})^2 \sim \chi_k^2$, i.e., the Chi-square distribution with k degrees of freedom (df).

Distribution result 2: If (i) $Z \sim N(0, 1)$, (ii) $U \sim \chi_k^2$, and (iii) Z and U are independent, then

$\frac{Z}{\sqrt{U/k}} \sim t_k$, i.e., the student's *t* distribution with *k* degrees of freedom (df).

Distribution result 3: (1) If (i) $U \sim \chi_p^2$, (ii) $V \sim \chi_q^2$, (iii) U and V are independent, then
$F = \frac{U/p}{V/q} \sim F_{p,q}$, i.e., the F distribution with p and q degrees of freedom.
(2) If $t \sim t_k$, then $t^2 \sim F_{1,q}$.

Distribution result 4: If $\varepsilon_i, i = 1, \cdots, n$, are iid $N(0, \sigma^2)$, then
(1) $Y_i's$ are independent and $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.
(2) $b_1 = \frac{S_{XY}}{S_{XX}} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} Y_i \sim N(\beta_1, \sigma^2 / S_{XX})$.
(3) $b_0 = \bar{Y} - b_1 \bar{X} = \sum_{i=1}^{n} [\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2}] Y_i \sim N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right))$.
(4) $\frac{(n-2)s^2}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi_{n-2}^2$, $\frac{SS_{reg}}{\sigma^2} \sim \chi_1^2$.
(5) $s^2 (= \frac{RSS}{n-2})$ is independent of b_0, b_1, and SS_{reg}.

Note: If random variables U and W are independent, then for any functions f and g, $f(U)$ and $g(W)$ are independent.

Definition 1.3 The *standard error* of a statistic is the estimated standard deviation of the statistic.

Ex. 1.3 The standard deviation of b_1 is

$$sd(b_1) = \sqrt{\frac{\sigma^2}{S_{XX}}} = \frac{\sigma}{\sqrt{S_{XX}}} \text{ estimated by}$$
$$se(b_1) = \frac{s}{\sqrt{S_{XX}}} \quad \text{— the standard error of } b_1$$

The standard error of b_0 is $se(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$.

- Hypothesis tests

(I) Tests for β_1

Hypotheses: H_0: $\beta_1 = 0$ (there is no linear relationship with X) vs.
H_a: $\beta_1 \neq 0$ (there is linear relationship with X)

- *t* test

Test statistic: $t^{(1)} = \frac{b_1 - 0}{se(b_1)}$.
Under H_0, $t^{(1)} \sim t_{n-2}$.

Reject H_0 if $\left|t_{obs}^{(1)}\right| \geq t_{n-2}(1-\frac{\alpha}{2})$ or *p*-value $= 2 \times P(t_{n-2} > \left|t_{obs}^{(1)}\right|) \leq \alpha$, where $t_{n-2}(1-\frac{\alpha}{2})$ is the $(1-\frac{\alpha}{2})th$ quantile of t_{n-2} and called the *critical value*, which can be found from the *t*-table on page 686.

- *F* test

Test statistic: $F = \frac{MS_{reg}}{MS_{resid}}$.
Under H_0, $F \sim F_{1,n-2}$.

Reject H_0 if $F_{obs} \geq F_{1,n-2}(1-\alpha)$ or *p*-value = $P(F_{1,n-2} \geq F_{obs}) \leq \alpha$, where $F_{1,n-2}(1-\alpha)$ is the $(1-\alpha)th$ quantile of $F_{1,n-2}$, which can be found from *F*-table on pages 688 - 693.

Notes: (1) Since $(t^{(1)})^2 = \frac{b_1^2}{s^2/S_{xx}} = \frac{(S_{XY}/S_{XX})^2 S_{XX}}{s^2} = \frac{S_{XY}^2/S_{XX}}{s^2} = \frac{MS_{reg}}{MS_{resid}} = F$, the above *t* test and *F* test are equivalent.
(2) The *t* test can be used for one-tailed tests (i.e., H_a: $\beta_1 > 0$ or $\beta_1 < 0$) but the *F* test cannot.
(3) The *t* test can be used to test for any hypothesized value β_{10} (i.e., test for H_0: $\beta_1 = \beta_{10}$ vs. H_a: $\beta_1 \neq \beta_{10}$) but the *F* test cannot.

(II) Test of β_0

— less common but occasionally useful.

Hypotheses: H_0: $\beta_0 = \beta_{00}$ vs. H_a: $\beta_0 \neq \beta_{00}$

Test statistic: $t^{(0)} = \frac{b_0 - \beta_{00}}{se(b_0)}$.
Under H_0, $t^{(0)} \sim t_{n-2}$.

Reject H_0 if $\left|t_{obs}^{(0)}\right| \geq t_{n-2}(1-\frac{\alpha}{2})$ or *p*-value = $2 \times P(t_{n-2} > \left|t_{obs}^{(0)}\right|) \leq \alpha$.

- Steps for hypothesis testing

1. Formulate null and alternative hypotheses (H_0, H_a).
2. Display the test statistic, compute the observed value of the test statistic, and give the critical value or *p*-value.
3. State the conclusion (the conclusion in statistics; why? and the conclusion in the context of the problem).

- Some basic concepts

Before selection: $X_1, X_2, \cdots, X_n$ — a random sample of size n. Each X_i is a random variable.
After selection: $x_1, x_2, \cdots, x_n$ — a realization of the random sample. Each x_i is a number.
Any function $\mathrm{T}(X_1, X_2, \cdots, X_n)$ of the random sample $\{X_1, X_2, \cdots, X_n\}$ is called a *statistic*. A statistic is a random variable.
$T(x_1, x_2, \cdots, x_n)$ — an observed value of the statistic $\mathrm{T}(X_1, X_2, \cdots, X_n)$, which is a number.

- Confidence Intervals

According to the above discussion, $t = \frac{b_i - \beta_i}{se(b_i)} \sim t_{n-2}, i = 0{,}1$.Thus,

$$1 - \alpha = P(-t_{n-2}(1 - \tfrac{\alpha}{2}) < \tfrac{b_i - \beta_i}{se(b_i)} < t_{n-2}(1 - \tfrac{\alpha}{2}))$$
$$= P(b_i - t_{n-2}(1 - \tfrac{\alpha}{2})se(b_i) < \beta_i < b_i + t_{n-2}(1 - \tfrac{\alpha}{2})se(b_i))$$

A 100(1-α)% confidence interval (CI) for β_1 is

$$b_1 \pm t_{n-2}(1 - \tfrac{\alpha}{2}) \frac{s}{\sqrt{S_{\mathrm{XX}}}} .$$

A 100(1-α)% confidence interval (CI) for β_0 is

$$b_0 \pm t_{n-2}(1 - \tfrac{\alpha}{2})\, s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}} .$$

Note: The general form of a confidence interval (CI) for a parameter θ is

(A point estimator $\hat{\theta}$ of θ) ± (critical value) (standard error of $\hat{\theta}$).

Ex. 1.1 (continued) The National Institute of Health is studying the relationship between the number of cigarettes smoked per day and the birthweight of babies born to mothers who smoke cigarettes. The following data are observed.

No. of cigarettes per day(X)	21	12	28	10	24	5
Birthweight (Y)	6.0	8.0	5.6	7.5	6.2	8.5

(6) Determine whether or not there is a linear relationship between X and Y by a hypothesis test at α = 0.05.

(7) Calculate a 95% confidence interval for β_1 and interpret the confidence interval. What is your conclusion based on the confidence interval?