• We typically assume the random errors balance out – they average zero.
• Then this is equivalent to assuming the mean of Y, denoted $E(Y)$, equals the deterministic component.

Straight-Line Regression Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

random component

Y = response variable (dependent variable)
X = predictor variable (independent variable)
ε = random error component

β_0 = Y-intercept of regression line
β_1 = slope of regression line

Note that the deterministic component of this model is
$E(Y) = \beta_0 + \beta_1 X$

Typically, in practice, β_0 and β_1 are unknown parameters. We estimate them using the sample data.

Response Variable (Y): Measures the major outcome of interest in the study.
Predictor Variable (X): Another variable whose value explains, predicts, or is associated with the value of the response variable.

Fitting the Model (Least Squares Method)

If we gather data (X, Y) for several individuals, we can use these data to estimate β_0 and β_1 and thus estimate the linear relationship between Y and X.

First step: Decide if a straight-line relationship between Y and X makes sense.

Plot the bivariate data using a scattergram (scatterplot).

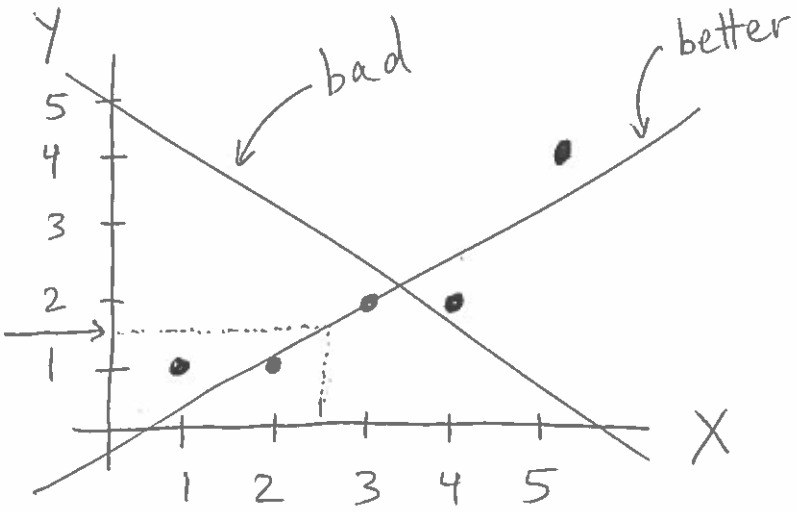


Table 11.1

X (drug amt)	Y (reaction time)
1	1
2	1
3	2
4	2
5	4

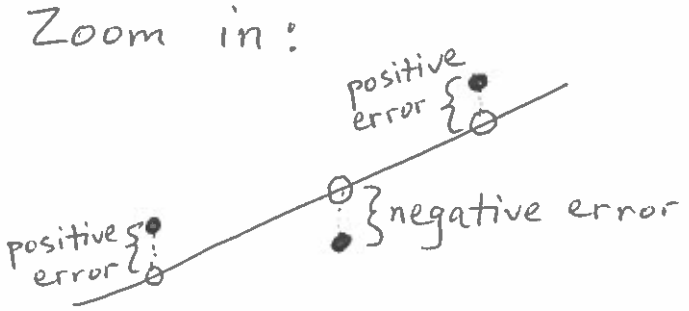
Once we settle on the "best-fitting" regression line, its equation gives a predicted Y-value for any new X-value.

$\hat{y}$ (Y-hat) = "predicted" Y-value

For drug amount 2.5 percent ($X=2.5$),

$\hat{y} \approx 1.7$ seconds.

How do we decide, given a data set, which line is the best-fitting line?



Note that usually, no line will go through all the points in the data set.

For each point, the <u>error</u> = $y - \hat{y}$ for each point (residual)
(Some positive errors, some negative errors)

We want the line that makes these errors as small as possible (so that the line is "close" to the points).

<u>Least-squares method</u>: We choose the line that minimizes the sum of all the <u>squared</u> errors (SSE).

Least squares regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$ ← equation of our estimated regression line.

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1 that produce the best-fitting line in the least squares sense.

Formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$:

Estimated slope and intercept:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \text{ and } \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

where $SS_{xy} = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$ **and**

$$SS_{xx} = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

and *n* = the number of observations.

Example (Table 11.3):

Y = response variable (reaction time)

X = predictor variable (drug amount)

SS_{xy} = $37 - \frac{(15)(10)}{5} = 37 - 30 = 7$

SS_{xx} = $55 - \frac{(15)^2}{5} = 55 - 45 = 10$

$\hat{\beta}_1 = \frac{7}{10} = 0.7, \quad \hat{\beta}_0 = \left(\frac{10}{5}\right) - (0.7)\left(\frac{15}{5}\right) = -0.1$

$\sum X_i = 1+2+3+4+5 = 15, \quad \sum Y_i = 1+1+2+2+4 = 10$

$\sum X_i Y_i = (1)(1) + (2)(1) + (3)(2) + (4)(2) + (5)(4) = 37$

$\sum X_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$

Least-squares estimated regression equation:

$$\hat{Y} = -0.1 + 0.7X$$

Interpretations:

Slope: $\hat{\beta}_1$ = predicted change in Y corresponding to a one-unit increase in X.

Intercept: $\hat{\beta}_0$ = predicted Y-value when $X=0$ (only when this makes sense)

Example: $\hat{\beta}_1 = 0.7$: For each one-percent (units of X) increase in drug amount, the predicted reaction time will increase by 0.7 seconds. (units of Y)

$\hat{\beta}_0 = -0.1$: When drug amount is 0 percent, the predicted reaction time is -0.1 seconds.

(DOESN'T MAKE SENSE HERE!)

Avoid extrapolation: predicting/interpreting the regression line for X-values outside the range of *X* in the data set. In the drug amount/reaction time example, we should avoid using the regression line to predict reaction times when drug amount is not between 1 percent and 5 percent.

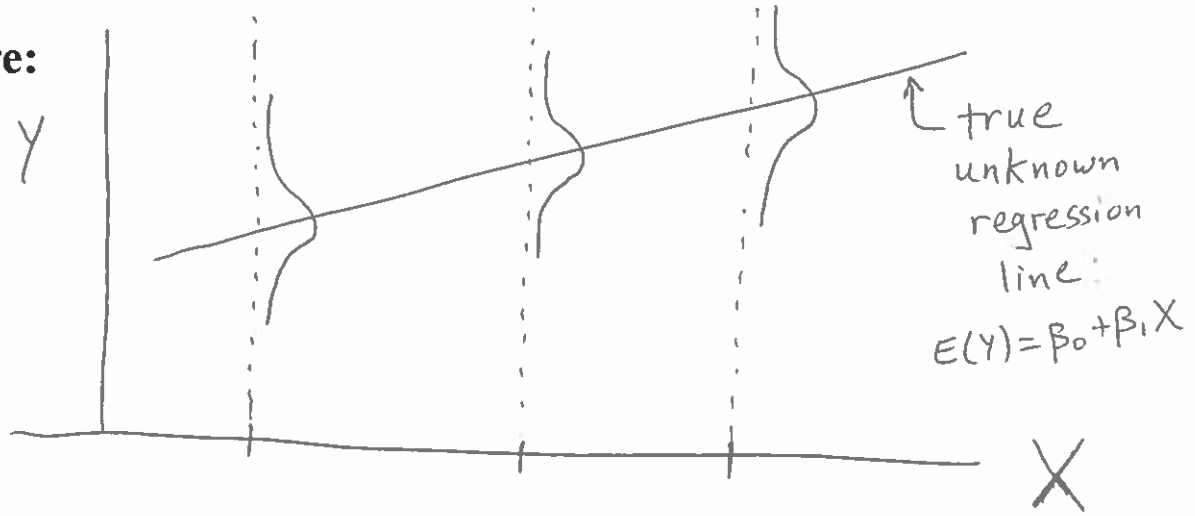
Model Assumptions

Recall model equation: $Y = \beta_0 + \beta_1 X + \varepsilon$

To perform inference about our regression line, we need to make certain assumptions about the random error component, ε. We assume:

(1) **The mean of the probability distribution of ε is 0. (In the long run, the values of the random error part average zero.)**
(2) **The variance of the probability distribution of ε is constant for all values of *X*. We denote the variance of ε by σ^2.**
(3) **The probability distribution of ε is normal.**
(4) **The values of ε for any two observed Y-values are independent – the value of ε for one Y-value has no effect on the value of ε for another Y-value.**

Picture:



- We check these assumptions using residual plots:
(Recall the residuals are the $(y - \hat{y})$ values for each observation)

① Plot residuals vs. X-values
② Normal Q-Q plot of the residuals.

Estimating σ^2

Typically the error variance σ^2 is unknown.

An unbiased estimate of σ^2 is the mean squared error (MSE), also denoted s^2 sometimes.

MSE = $\dfrac{\text{SSE}}{n-2}$

where SSE = SS_{yy} - $\hat{\beta}_1 SS_{xy}$

and $SS_{yy} = \sum Y_i^2 - \dfrac{(\sum Y_i)^2}{n}$

Note that an estimate of σ is

$s = \sqrt{MSE} = \sqrt{\dfrac{SSE}{n-2}}$

used to help us interpret $s = \sqrt{MSE}$

Since ε has a normal distribution, we can say, for example, that about 95% of the observed Y-values fall within 2*s* units of the corresponding values $\hat{Y}$.

↑ units of Y.

Example: In reaction time example, $s = \sqrt{MSE}$ = .606. We can say: Approximately 95% of the observed reaction times are within 1.212 seconds of the corresponding predicted times.