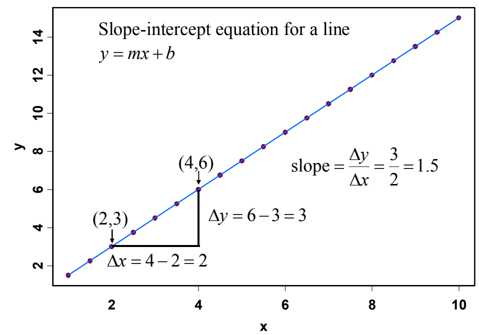
Simple Linear Regression

Regression equation—an equation that describes the average relationship between a response (dependent) and an explanatory (independent) variable.

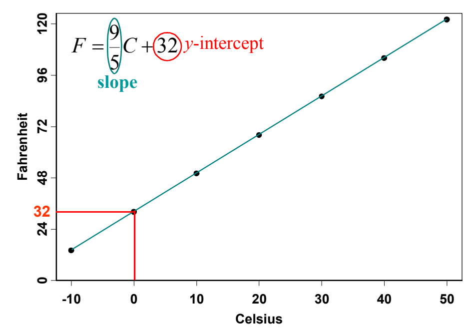


Deterministic Model

A model that defines an exact relationship between variables.

Example: $y = 1.5x$

There is no allowance for error in the prediction of y for a given x.



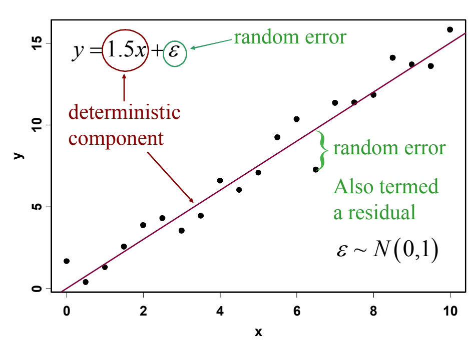
Probabilistic Model

A model that accounts for ***random error***.

Includes both a deterministic component and a random error component.

$$y = 1.5x + \text{ random error}$$

This model hypothesizes a probabilistic relationship between y and x.



Probabilistic Model—General Form

y = Deterministic component + Random component

where y is the "variable of interest".

Assume that the mean value of the random error is zero ➔ the mean value of y, $E(y)$, equals the deterministic component of the model

First-Order (Straight Line) Probabilistic Model

$y = \boxed{\beta_0 + \beta_1 x} + \varepsilon$ where y = ***Dependent variable***
x = ***Independent variable***

β_0 = ***population y-intercept of the line***—the point at which the line intersects or cuts through the y-axis

β_1 = ***population slope of the line***—the amount of increase (or decrease) in the deterministic component of y for every 1-unit increase (or decrease) in x.

ε = random error component

First-Order (Straight Line) Probabilistic Model

β_0 and β_1 are population parameters. They will only be known if the population of all (x, y) measurements are available.

β_0 and β_1, along with a specific value of the independent variable x determine the ***mean value*** of the dependent variable y.

Model Development

β_0 and β_1 will generally be unknown.

The process of developing a model, estimating model parameters, and using the model can be summarized in these 5-steps:

1. Hypothesize the deterministic component of the model that relates the mean, $E(y)$ to the independent variable x

$$E(y) = \beta_0 + \beta_1 x$$

2. Use sample data to estimate unknown model parameters

find estimates: $\hat{\beta}_0$ or b_0, $\hat{\beta}_1$ or b_1

Model Development (continued)

3. Specify the probability distribution of the random error term and estimate the SD of this distribution

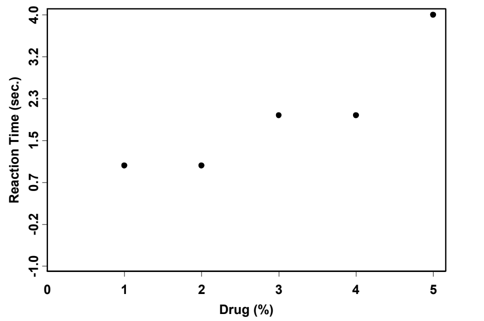
$\varepsilon \sim N(0,\sigma)$ −−will revisit this later

4. Statistically evaluate the usefulness of the model
5. Use model for prediction, estimation or other purposes

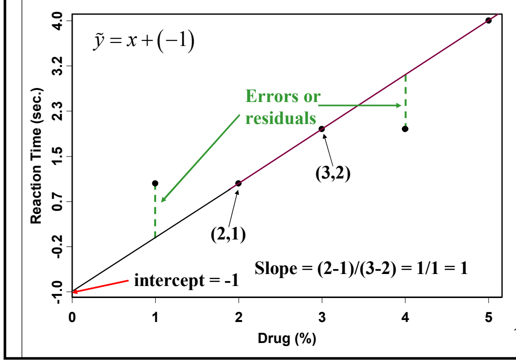
Example: Reaction time versus drug percentage

Subject	Amount of Drug (%) *x*	Reaction Time (seconds) *y*
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Example: Reaction time versus drug percentage



Example: Reaction time versus drug percentage



Example: Reaction time versus drug percentage

Errors of prediction---vertical differences between the observed and the predicted values of y

x	y	$\tilde{y} = -1 + x$	$(y - \tilde{y})$	$(y - \tilde{y})^2$
1	1	0	(1-0) = 1	1
2	1	1	(1-1) = 0	0
3	2	2	(2-2) = 0	0
4	2	3	(2-3) = -1	1
5	4	4	(4-4) = 0	0
			Sum of errors = 0	Sum of squared errors (SSE) = 2

Least Squares Line

Also called ***regression line***, or the ***least squares prediction equation***

Method to find this line is called the ***method of least squares***

For our example, we have a sample of n = 5 pairs of (x, y) values. The fitted line that we will calculate is written as $\hat{y} = b_0 + b_1 x$

$\hat{y}$ is an estimator of the mean value of y, $E(y)$;

b_0 and b_1 are estimators of β_0 and β_1

Least Squares Line (continued)

Define the sum of squares of the deviations of the y values about their predicted values for all n data points as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_i) \right]^2$$

We want to find b_0 and b_1 to make the SSE a minimum---termed ***least squares estimates***

$\hat{y} = b_0 + b_1 x$ is called the least squares line

Formulas for the Least Squares Estimates

Slope: $b_1 = \frac{\text{SS}_{xy}}{\text{SS}_{xx}}$ or $b_1 = r\frac{\text{SD}_y}{\text{SD}_x}$

$$s_{xy} = \text{SS}_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$s_{xx} = \text{SS}_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

y-intercept: $b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n}$

n = sample size

LS Calculations for Drug/Reaction Example

x_i	y_i	x_i^2	$x_i y_i$
1	1	1	1
2	1	4	2
3	2	9	6
4	2	16	8
5	4	25	20
$\sum x_i = 15$	$\sum y_i = 10$	$\sum x_i^2 = 55$	$\sum x_i y_i = 37$

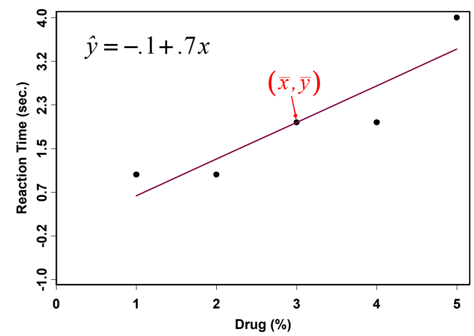
$$b_1 = \frac{7}{10} = 0.7$$

$$b_0 = \frac{10}{5} - (.7)\frac{15}{5} = 2 - (.7)(3) = 2 - 2.1 = -.1$$

$$SS_{xy} = 37 - \frac{(15)(10)}{5} = 37 - 30 = 7 \qquad SS_{xx} = 55 - \frac{(15)^2}{5} = 55 - 45 = 10$$

$$\hat{y} = -.1 + .7x$$

LS Line for Drug/Reaction Example



LS Calculations for Drug/Reaction Example

x	y	$\hat{y} = -.1 + .7x$	$(y - \hat{y})$	$(y - \hat{y})^2$
1	1	.6	(1-.6) = .4	.16
2	1	1.3	(1-1.3) = -.3	.09
3	2	2.0	(2-2.0) = 0	.00
4	2	2.7	(2-2.7) = -.7	.49
5	4	3.4	(4-3.4) = .6	.36
			Sum of errors = 0	Sum of squared errors (SSE) = 1.10

The LS line has a sum of errors = 0, but
SSE = 1.1 < 2.0 for visual model

Least Squares Line—Interpretation of $\hat{y} = -.1 + .7x$

Estimated intercept is negative
➔that the estimated ***mean reaction time*** is equal to -0.1 seconds when the amount of drug is 0%.

What does this mean since negative reaction times are not possible?

Model parameters should be interpreted only within the sampled range of the independent variable.

Least Squares Line—Interpretation of $\hat{y} = -.1 + .7x$

The slope of 0.7 implies that for every unit increase of x, the ***mean value*** of y is estimated to increase by 0.7 units.

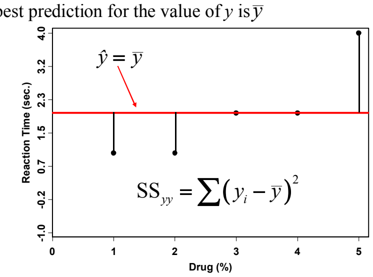
In the context of the problem:

For every 1% increase in the amount of drug in the bloodstream, the mean reaction time is estimated to increase by 0.7 seconds over the sampled range of drug amounts from 1% to 5%.

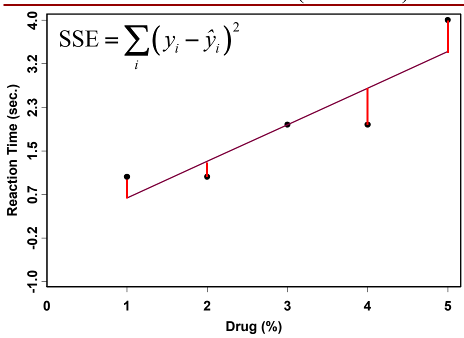
Coefficient of Determination

A measure of the contribution of x in predicting y

Assuming that x provides no information for the prediction of y, the best prediction for the value of y is $\bar{y}$



Coefficient of Determination (continued)



Coefficient of Determination (continued)

$\text{SS}_{yy} = \sum (y_i - \bar{y})^2$ --total sample variation around mean

$\text{SSE} = \sum (y_i - \hat{y}_i)^2$ --unexplained sample variability after fitting

$\text{SS}_{yy} - \text{SSE}$ --explained sample variability attributable to linear relationship

$\dfrac{\text{SS}_{yy} - \text{SSE}}{\text{SS}_{yy}} = \dfrac{\text{explained}}{\text{total}} =$ proportion of total sample variability explained by the linear relationship

Coefficient of Determination (continued)

$$r^2 = \frac{\text{SS}_{yy} - \text{SSE}}{\text{SS}_{yy}} = 1 - \underbrace{\frac{\text{SSE}}{\text{SS}_{yy}}}_{}$$ } Unexplained variability

In simple linear regression r^2 is computed as the square of the correlation coefficient, r

$0 \le r^2 \le 1$

Interpretation— $r^2 = .75$ means that the sum of squared deviations of the y values about their predicted values has been reduced by 75% by the use $\hat{y}$, instead of $\bar{y}$, to predict y of the least squares equation.