9 – Linear Models

A major part of inferential statistics is to create *models* that represent relationships between variables. Models can be used to predict outcomes where we lack data.

Example A *linear* statistical model might be used to predict the **Pulse** rate of a student based on their **Age**, **Height**, and **Sex**.

	Sex	Wr.Hnd	NW.Hnd	W.Hnd	Fold	Pulse	Clap	Exer	Smoke	Height	M.I	Age
1	Female	18.5	18.0	Right	R on L	92	Left	Some	Never	173.00	Metric	18.250
2	Male	19.5	20.5	Left	R on L	104	Left	None	Regul	177.80	Imperial	17.583
3	Male	18.0	13.3	Right	L on R	87	Neither	None	Occas	*NA*	*NA*	16.917
4	Male	18.8	18.9	Right	R on L	*NA*	Neither	None	Never	160.00	Metric	20.333
5	Male	20.0	20.0	Right	Neither	35	Right	Some	Never	165.00	Metric	23.667
6	Female	18.0	17.7	Right	L on R	64	Right	Some	Never	172.72	Imperial	21.000
7	Male	17.7	17.7	Right	L on R	83	Right	Freq	Never	182.88	Imperial	18.833

...

235	Female	17.5	16.5	Right	R on L	*NA*	Right	Some	Never	170.00	Metric	18.583
236	Male	21.0	21.5	Right	R on L	90	Right	Some	Never	183.00	Metric	17.167
237	Female	17.6	17.3	Right	R on L	85	Right	Freq	Never	168.50	Metric	17.750

Age = 21
Height = 170 → predicted **Pulse** =
Sex = Female

Linear Correlation Coefficient

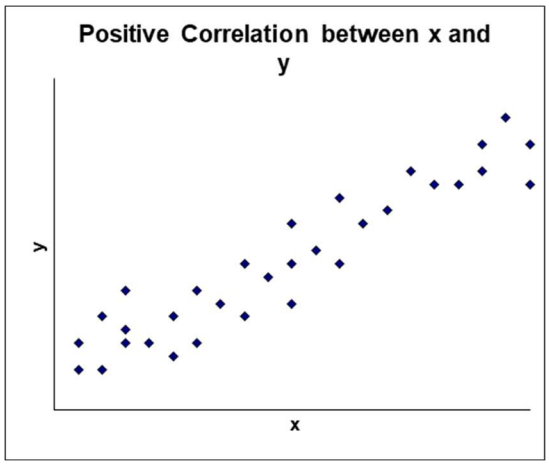
If X and Y are paired numerical variables with data $X = x_1, x_2, \dots, x_n$ and $Y = y_1, y_2, \dots, y_n$, then we define the *linear correlation coefficient*:

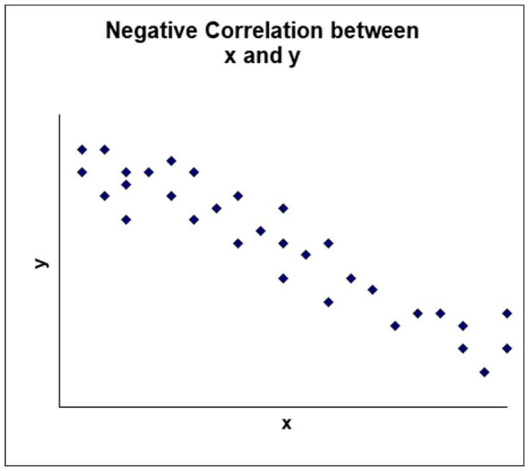
Pearson's Correlation Coefficient

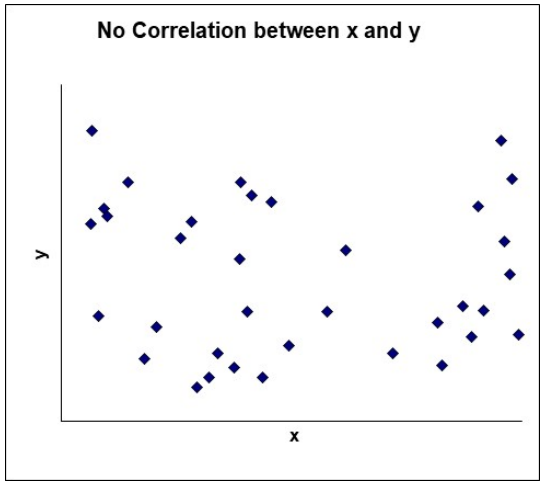
$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}}$$

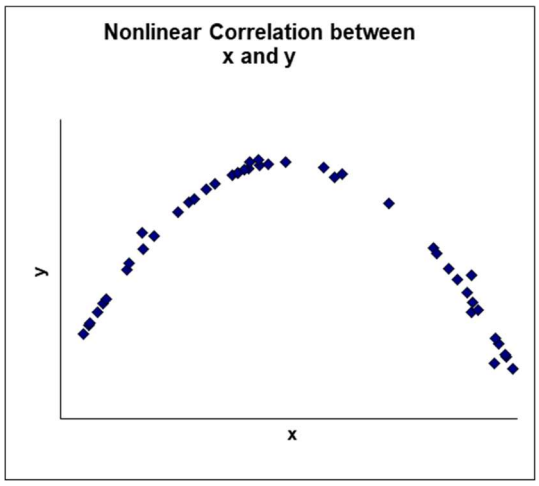
or using R:

```
r <- cor(X, Y)
```

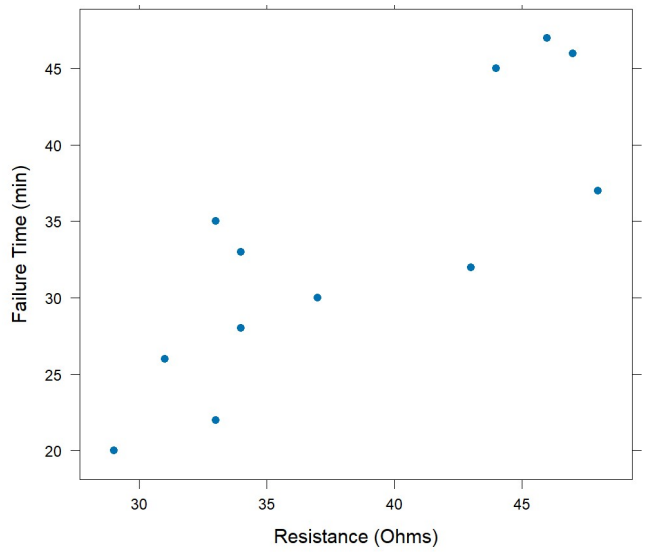








Example (Electric Circuit) Suppose we make $n = 12$ measurements of an electric circuit's Resistance (in Ohms) and Fail.Time (in minutes).



	Resistance	Fail.time
1	43	32
2	29	20
3	44	45
4	33	35
5	33	22
6	47	46
7	34	28
8	31	26
9	48	37
10	34	33
11	46	47
12	37	30

Let's calculate r using the data for **Resistance** and **Fail.Time**.

	Resistance	**Fail.Time**			
	x	y	xy	x^2	y^2
	43	32	1376	1849	1024
	29	20	580	841	400
	44	45	1980	1936	2025
	33	35	1155	1089	1225
	33	22	726	1089	484
	47	46	2162	2209	2116
	34	28	952	1156	784
	31	26	806	961	676
	48	37	1776	2304	1369
	34	33	1122	1156	1089
	46	47	2162	2116	2209
	37	30	1110	1369	900
Total	459	401	15907	18075	14301

Properties of the Linear Correlation Coefficient, *r*

Suppose X and Y are two numerical variables and we calculate r for a random sample.

1. The possible values of r are:
$$-1 \leq r \leq 1$$

2. The value of r does not change if X or Y are expressed in different *units*.

3. Swapping X and Y does not affect r.

4. Linear correlation coefficient r only measures the strength of *linear* relationships. It will not indicate the existence of non-linear relationships.

5. We use r to denote linear correlation for a *sample*.
We use ρ to denote linear correlation of a *population*.

$r = 1$	$\Rightarrow$ perfect positive linear correlation
r "close" to 1	$\Rightarrow$ strong positive linear correlation
$r = 0$	$\Rightarrow$ no linear correlation
r "close" to -1	$\Rightarrow$ strong negative linear correlation
$r = -1$	$\Rightarrow$ perfect negative linear correlation

IMPORTANT NOTE:

If X and Y are correlated ($r \neq 0$), it does not mean that higher values of X *cause* higher values of Y.

"Correlation does not imply causation"

Example The amount of ice cream consumed (X) and number of boating accidents (Y) each day are positively correlated. But eating ice cream does not cause boating accidents!

Hypothesis Testing for ρ

When we calculated $r = 0.8324$ for the **Resistance** and **Fail.Time** measurements, we were *certain* that the *sample* measurements were positively correlated.

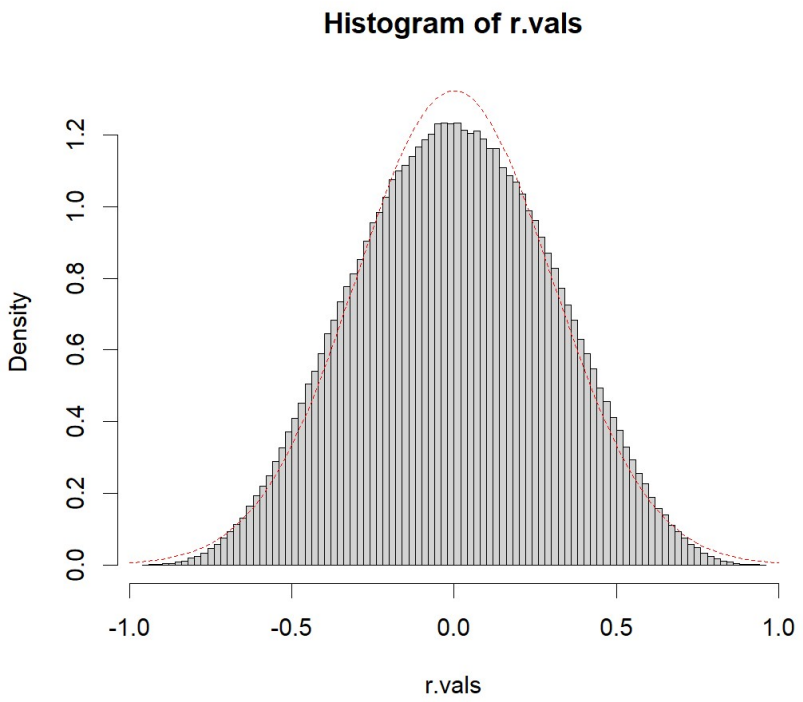
Can we conclude that **Resistance** and **Fail.Time** are correlated at the *population* level?

If they are not, we would have to say that $r = 0.8324$ occurred *by random chance*.

Sampling Distribution of r

Suppose X and Y are two normal variables that are *un*correlated ($\rho = 0$) and follow a *bivariate normal* distribution (more on this below).

If we take random samples of size n and calculate the correlation coefficient r, then r follows a distribution that is *roughly* normal.



In fact, it turns out that $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ follows a Student t-distribution with $\mathrm{df} = n - 2$.

[This is because $\sqrt{\frac{1-r^2}{n-2}}$ is an unbiased estimator of σ_r, the standard deviation of r.]

This allows us to perform a *hypothesis test* for the claim that $\rho \neq 0$.

Example Using the sample statistics for **Resistance** and **Fail.Time** ($n = 12, r = 0.8324$), test the claim that $\rho \neq 0$.

1. (Claim): The claim is that $\rho \neq 0$.

2. (Hypotheses):

3. (Test statistic): $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} =$

4. (P-value):

5. (Decision):

6. (Conclusion):

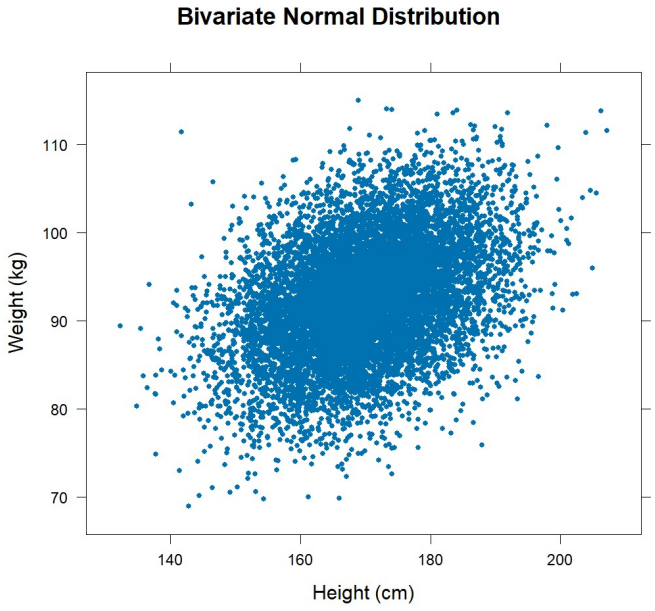
Assumptions for Hypothesis Test about ρ

Note that using the test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

is only correct under the assumption that X and Y are normally distributed for *each specific value* of the other variable. This is called a *bivariate normal* distribution.

Example Shown here is a scatter plot of variables X (Height) and Y (Weight) that follow a bivariate normal distribution.



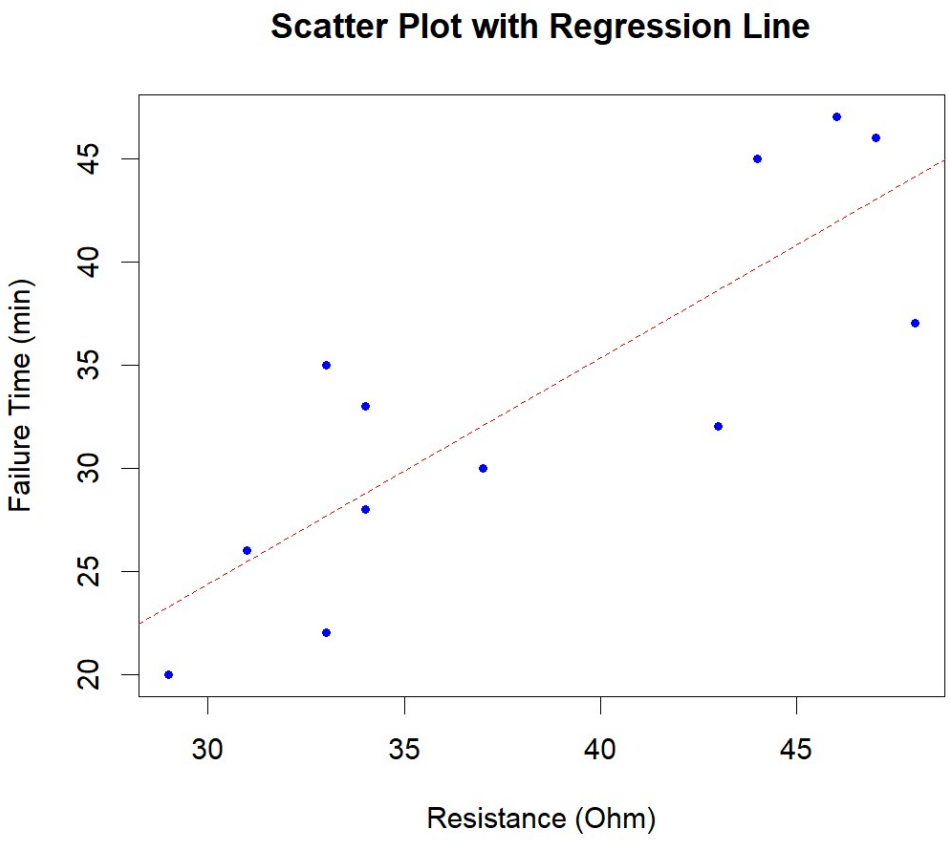
Linear Regression

Once we have determined that two variables X and Y are correlated at the population level, (i.e., they have a significant linear relationship), we typically describe that relationship using a *regression line* (i.e., "best fitting line").

In this context, we say:

X is the *independent* or *predictor* variable

Y is the *dependent* or *response* variable



We will write this regression line in the form

$$\hat{y} = a + bx$$

where $\hat{y}$ is the *predicted* Y value for a given value $X = x$. The coefficients are:

$a = y$-intercept of the line (α at population level)

$b =$ slope of the line (β at population level)

Linear Regression Coefficients: $\hat{y} = a + bx$

Using methods from calculus (or from linear algebra), it is possible to derive the formulas

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \qquad = r \cdot \frac{s_Y}{s_X}$$

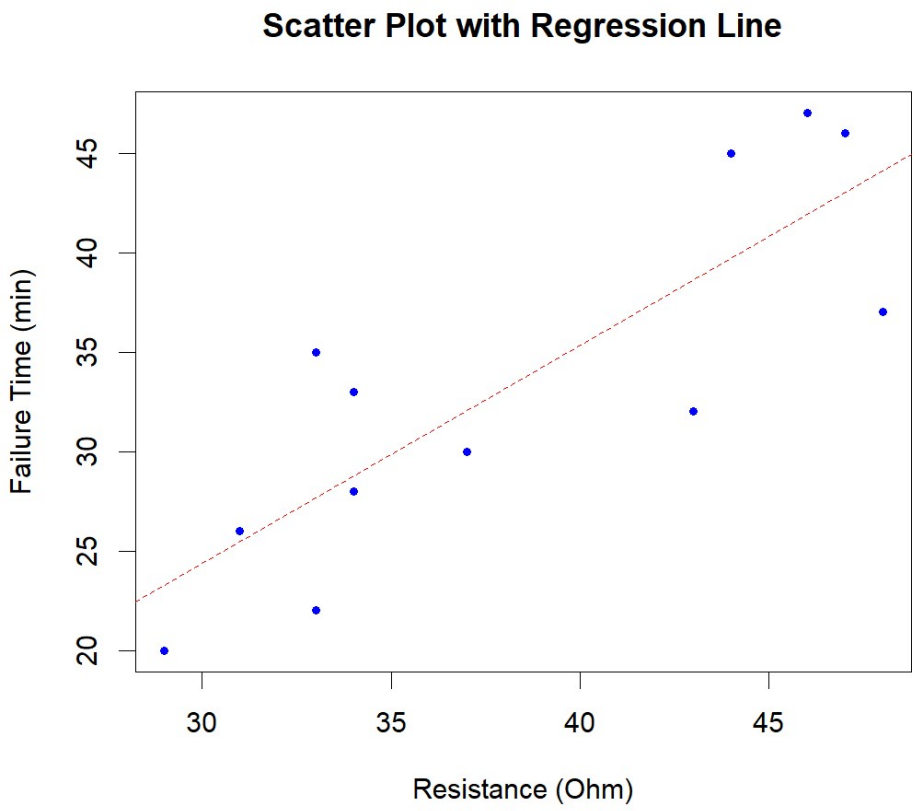
$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \qquad = \bar{Y} - b \cdot \bar{X}$$

where

s_X = sample std. dev. of X

s_Y = sample std. dev. of Y

The resulting line, $\hat{y} = a + bx$ is sometimes called the "least-squares" regression line, because it creates the least possible sum of the vertical errors (squared).



Example Use the sample data for **Resistance** and **Fail.time** to find the regression line.

	Resistance	**Fail.Time**			
	x	y	xy	x^2	y^2
	43	32	1376	1849	1024
	29	20	580	841	400
	44	45	1980	1936	2025
	33	35	1155	1089	1225
	33	22	726	1089	484
	47	46	2162	2209	2116
	34	28	952	1156	784
	31	26	806	961	676
	48	37	1776	2304	1369
	34	33	1122	1156	1089
	46	47	2162	2116	2209
	37	30	1110	1369	900
Total	459	401	15907	18075	14301

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} =$$

$$r \cdot \frac{s_Y}{s_X} =$$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

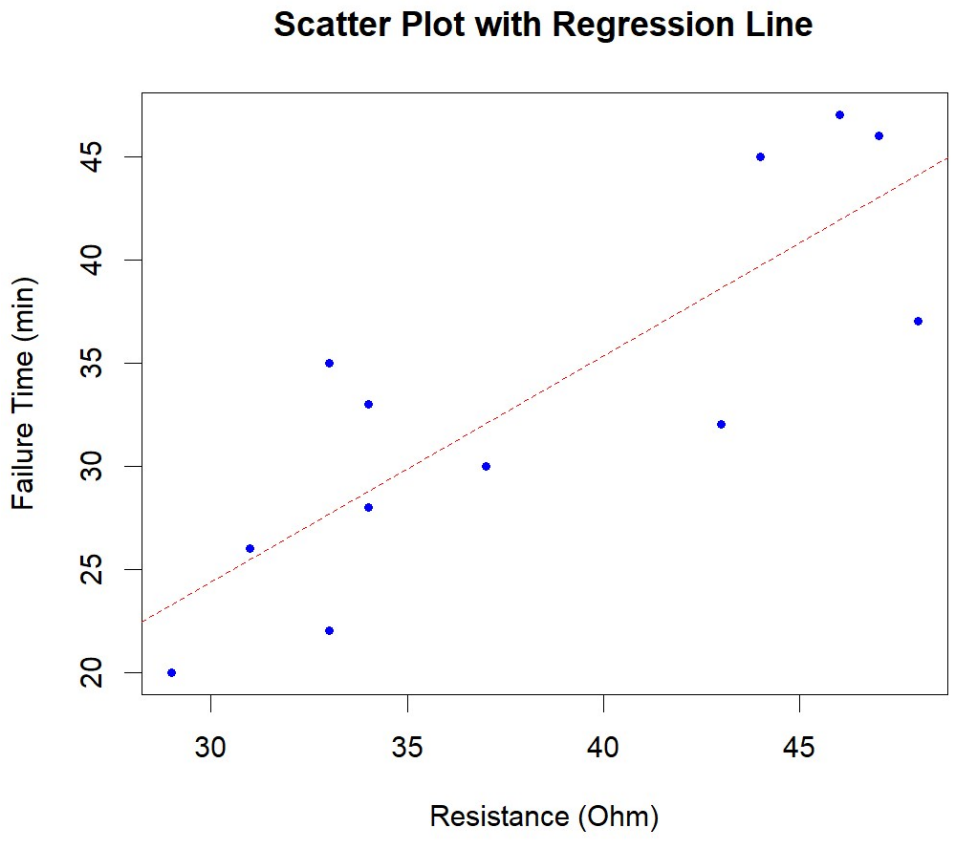
$$\bar{Y} - b \cdot \bar{X} =$$

Making Predictions

Knowing the model (equation) that best fits the data allows us to make predictions for values that were not measured.

Example Use the regression line to predict **Fail.Time** if **Resistance** is 40 Ohms.

$$\hat{y} = -8.56 + 1.10x$$



Note: It is NOT reliable to make predictions for X values that are outside the range of your data set. (This is called *extrapolation*.)

In this scenario, our X values go from 29 to 47. Using $X = 70$ to predict

$$\hat{y} = -8.56 + 1.10 \times 70 = 68.4 \text{ min}$$

would be unreliable, since we do not know if the linear relationship continues for $X > 47$.

Prediction Intervals

In the previous example we obtained a point estimate $\hat{y} = 35.4$ for $X = 40\ \text{Ohm}$.

It is even more useful to generate an *interval estimate* for Y. In this context, we call such an interval a *prediction interval*.

To do this, we need to make certain assumptions about X and Y.

Assumptions

- X and Y follow a bivariate distribution
- The variance of Y is the same for all specific values of X (*homoscedasticity)*
- The mean of Y at each X level lies along a line for different values of X (*linearity*)

Under these assumptions, it is possible to prove that the prediction error $(Y - \hat{y})$ follows a normal distribution with a standard deviation that we can estimate using:

$$S_e = \sqrt{\frac{\sum(Y - \hat{y})^2}{n-2}} = \sqrt{\frac{(\sum y^2) - a(\sum y) - b(\sum xy)}{n-2}}$$

The quantify S_e is called the *Standard Error of the Estimate*.

Think of S_e as the typical vertical distance between the regression line and a point (X, Y).

Example Calculate S_e using the sample data for **Resistance** and **Fail.Time**.

$$S_e = \sqrt{\frac{(\sum y^2) - a(\sum y) - b(\sum xy)}{n-2}}$$

$$= \sqrt{\frac{14301 - (-8.56041)(401) - (1.09744)(15907)}{12-2}} =$$

To construct a prediction interval for Y at $X = x_0$ with a confidence level $1 - \alpha$, we use

$$\hat{y} - E < Y < \hat{y} + E$$

where

$$\hat{y} = a + bx_0$$

$$E = t_{\alpha/2} \cdot S_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)s_X^2}}$$

The critical t value $t_{\alpha/2}$ has $n - 2$ degrees of freedom.

Example Calculate a 95% prediction interval for Y with $X = 40$ Ohm.

Confidence Intervals for α and β

We can construct confidence intervals for the parameters α and β in our regression model.

$$\hat{y} = \alpha + \beta X$$

As usual, the interval depends mainly on a formula for the *margin of error,* E.

The confidence interval (at level $1 - \alpha$) for the slope parameter β is

$$b - E < \beta < b + E$$

where

$$E = t_{\alpha/2} \frac{S_e}{\sqrt{n-1} \cdot s_X} \qquad (\text{where df} = n - 2)$$

Likewise, it is possible to perform a hypothesis test on the regression slope β using the test statistic

$$t = \frac{b - \beta}{\frac{S_e}{\sqrt{n-1} \cdot s_X}} = \frac{b - \beta}{S_e} \cdot \sqrt{n-1} \cdot s_X \quad (\text{df} = n - 2)$$

Example Continue our example involving time to failure of resistors. Find the 95% confidence interval for the regression slope β. Also test the hypothesis that the population regression slope is non-zero, at the 5% significance level.

```
> # Much of this can be read off of the output of the follow:
> model

Call:
lm(formula = Fail.time ~ Resistance, data = circuit.df)

Coefficients:
(Intercept)   Resistance
     -8.561        1.097

> summary(model)

Call:
lm(formula = Fail.time ~ Resistance, data = circuit.df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1167 -3.8628 -0.1064  4.4551  7.3449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.5605     8.9685  -0.955 0.362328
Resistance    1.0974     0.2311   4.749 0.000781 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.261 on 10 degrees of freedom
Multiple R-squared:  0.6928,    Adjusted R-squared:  0.6621
F-statistic: 22.55 on 1 and 10 DF,  p-value: 0.0007813
```