# Chapter 6

# Simple Regression

We look at scatter diagrams, linear correlation and linear and nonlinear regression for bivariate and multivariate quantitative data sets.

## 6.1 Introduction

**Exercise 6.1 (Introduction)**

1. *Scatter Diagram: Reading Ability Versus Brightness.*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| brightness, x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ability to read, y | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |

```
brightness <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
reading.ability <- c(70, 70, 75, 88, 91, 94, 100, 92, 90, 85)
```

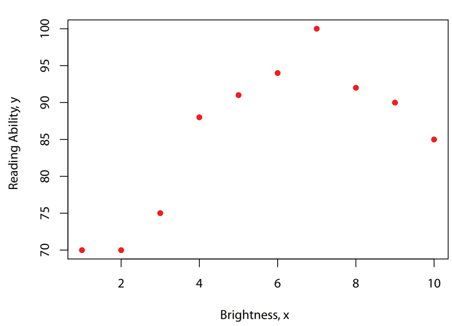


Figure 6.1: Scatter Diagram, Reading Ability Versus Brightness

```
plot(brightness,reading.ability,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y")
```

(a) There are (i) **10** (ii) **20** (iii) **30** data points.
One particular data point is (i) $\mathbf{(70, 75)}$ (ii) $\mathbf{(75, 2)}$ (iii) $\mathbf{(2, 70)}$.
Data point (9,90) means

i. for brightness 9, reading ability is 90.

ii. for reading ability 9, brightness is 90.

(b) Reading ability (i) **positively** (ii) **not** (ii) **negatively** *associated* to brightness.
As brightness increases, reading ability (i) **increases** (ii) **decreases**.

(c) Association (i) **linear** (ii) **nonlinear (curved)** because straight line cannot be drawn on graph where all points of scatter fall on or near line.

(d) "Reading ability" is (i) **response** (ii) **predictor** variable and "brightness" is (i) **response** (ii) **predictor** variable because reading ability depends on brightness, not the reverse

(e) Scatter diagrams drawn for quantitative data, not qualitative data because (circle one or more)

i. qualitative data has no order,

ii. distance between qualitative data points is not meaningful.

(f) Another sampled ten individuals gives (i) **same** (ii) **different** scatter plot. Data here is a (i) **sample** (ii) **population**.

2. *Scatter Diagram: Grain Yield (tons) versus Distance From Water (feet).*

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dist, x | 0 | 10 | 20 | 30 | 45 | 50 | 70 | 80 | 100 | 120 | 140 | 160 | 170 | 190 |
| yield, y | 500 | 590 | 410 | 470 | 450 | 480 | 510 | 450 | 360 | 400 | 300 | 410 | 280 | 350 |

```
distance <- c(0, 10, 20, 30, 45, 50, 70, 80, 100, 120, 140, 160, 170 190)
grain.yield <- c(500, 590, 410, 470, 450, 480, 510, 450, 360, 400, 300, 410, 280, 350)
```

(a) Scatter diagram has
(i) **a pattern** (ii) **no pattern (randomly scattered)** with
(i) **positive** (ii) **negative** association,
which is (i) **linear** (ii) **nonlinear**, that is a
(i) **weak** (ii) **moderate** (iii) **strong** (non)linear relationship,
where grain yield is (i) **response** (ii) **predictor** variable.

(b) *Review.* Second random sample would be (i) **same** (ii) **different** scatter plot of (distance, yield) points. Any statistics calculated from second plot would be (i) **same** (ii) **different** from statistics calculated from first plot.

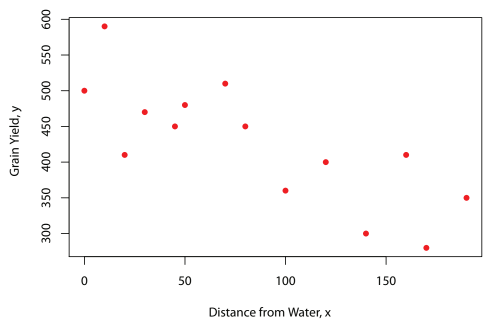


Figure 6.2: Scatter Diagram, Grain Yield Versus Distance from Water

```
plot(distance,yield,pch=16,col="red",xlab="Distance from Water, x",ylab="Grain Yield, y")
```

3. *Scatter Diagram: Pizza Sales (\$1000s) versus Student Number (1000s).*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| student number, $x$ | 2 | 6 | 8 | 8 | 12 | 16 | 20 | 20 | 22 | 26 |
| pizza sales, $y$ | 58 | 105 | 88 | 118 | 117 | 137 | 157 | 169 | 149 | 202 |

```
distance <- c(0, 10, 20, 30, 45, 50, 70, 80, 100, 120, 140, 160, 170 190)
grain.yield <- c(500, 590, 410, 470, 450, 480, 510, 450, 360, 400, 300, 410, 280, 350)
plot(students,sales,pch=16,col="red",xlab="Number of Students, x (1000s)",ylab="Pizza Sales, y ($1000)")
```

Scatter diagram has
(i) **a pattern** (ii) **no pattern (randomly scattered)** with
(i) **positive** (ii) **negative** association,
which is (i) **linear** (ii) **nonlinear**, that is a
(i) **weak** (ii) **moderate** (iii) **strong** (non)linear relationship,
where student number is (i) **response** (ii) **predictor** variable.

4. *More Scatter Diagrams*

   (a) Scatter diagram (a) has
   (i) **a pattern** (ii) **no pattern (randomly scattered)**.

   (b) Scatter diagram (b) has
   (i) **pattern** (ii) **no pattern (randomly scattered)**
   with (i) **positive** (ii) **negative** association,
   which is (i) **linear** (ii) **nonlinear**, that is a
   (i) **weak** (ii) **moderate** (iii) **strong** (non)linear relationship.

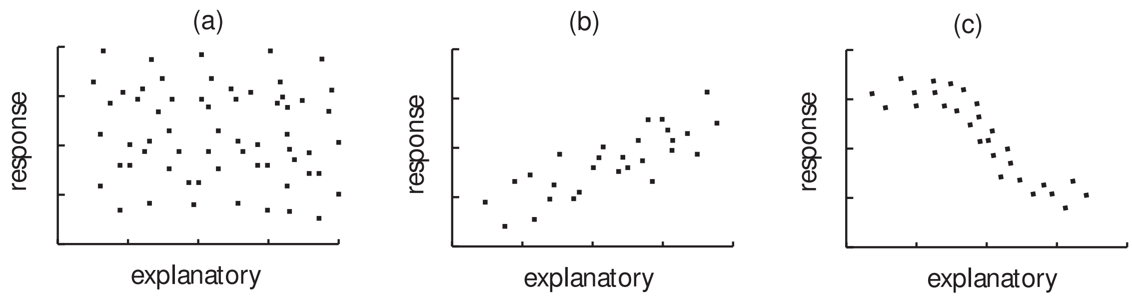


Figure 6.3: More Scatter Diagrams

(c) Scatter diagram (c) has
(i) **pattern** (ii) **no pattern (randomly scattered)**
with (i) **positive** (ii) **negative** association,
which is (i) **linear** (ii) **nonlinear**, that is a
(i) **weak** (ii) **moderate** (iii) **strong** (non)linear relationship.

## 6.2 Covariance and Correlation

We look at *covariance*, a measure of the *strength of association* between two random variables and also the closely related *correlation* which measures the *strength of linear association* between two variables. (Population parameter) covariance is defined by

$$\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y,$$

and has the following properties,

- $\text{Cov}(X,Y) = \text{Cov}(Y,X)$,
- $\text{Cov}(X,X) = V(X) = \sigma_X^2$,
- $\text{Cov}(aX,bY) = ab\text{Cov}(X,Y)$, where $a, b$ are constants,
- $\text{Cov}(X,Y) = 0$ if $X, Y$ are independent.

The units for covariance are (units $X$) $\cdot$ (units $Y$), which is a problem because then covariance, strength of association between two variables, is sensitive to seemingly unrelated changes in units; for example, covariance of two weight variables would be different if the units for these variables were measured in tons or kilograms. Consequently, we often use the *unitless (population parameter) correlation* $\rho$, $-1 \leq \rho \leq 1$, given by

$$\rho(x,y) = \rho = \frac{\text{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{E(XY) - \mu_X\mu_Y}{\sigma_1\sigma_2},$$

which has the following properties,

- $\rho(X,Y) = 0$ if $X,Y$ are independent,
- $|\rho(X,Y)| = 1$ if $P(Y = mX + b) = 1$ ($X,Y$ linearly related), $a,b$ constants.

The population parameter $\rho(X,Y)$ is estimated by *Pearson's sample correlation*,

$$\begin{aligned} r &= \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \\ &= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \end{aligned}$$

To test $H_0 : \rho = 0$ requires assuming $X,Y$ has a bivariate normal distribution,

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{h(x,y)}{2}}$$

where

$$h(x,y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]$$

with marginal pdfs

$$f_X(x) = \frac{1}{\sqrt{2}\sigma_X}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}, \quad f_Y(y) = \frac{1}{\sqrt{2}\sigma_Y}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\},$$

and, in particular, the test statistic is

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

which has a student's $t$-distribution with $n-2$ degrees of freedom and a $100(1-\alpha)\%$ confidence interval of $\rho$ is $l \leq \rho \leq u$ where hyperbolic tangents are used

$$l = \tanh(z-k), \quad u = \tanh(z+k),$$

where

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}, \quad k = \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}.$$

**Exercise 6.2 (Covariance and Correlation)**

1. *Covariance and Correlation: Waiting Times To Catch Fish.*
   The joint density, $f(x,y)$, of the number of minutes waiting to catch the *first* fish, $x$, *and* the number of minutes waiting to catch the *second* fish, $y$, is given below.

| $y \downarrow x \rightarrow$ | 1 | 2 | 3 | $f_Y(y) = P(Y = y)$ |
|---|---|---|---|---|
| 1 | 0.01 | 0.01 | 0.07 | 0.09 |
| 2 | 0.02 | 0.02 | 0.08 | 0.12 |
| 3 | 0.08 | 0.08 | 0.63 | 0.79 |
| $f_X(x) = P(X = x)$ | 0.11 | 0.11 | 0.78 | 1.00 |

(a) Calculate $E(XY)$.

$$\begin{aligned} E[XY] &= \sum_{x=1}^{3}\sum_{y=1}^{3} (xy)\, f(x,y) \\ &= (1\times 1)(0.01) + (1\times 2)(0.02) + (1\times 3)(0.08) \\ &\quad + (2\times 1)(0.01) + (2\times 2)(0.02) + (2\times 3)(0.08) \\ &\quad + (3\times 1)(0.07) + (3\times 2)(0.08) + (3\times 3)(0.63) = \end{aligned}$$

(choose one) (i) **5.23** (ii) **6.23** (iii) **7.23**.

```
x <- c(1,1,1,2,2,2,3,3,3)
y <- c(1,2,3,1,2,3,1,2,3)
f <- c(0.01,0.02,0.08,0.01,0.02,0.08,0.07,0.08,0.63)
EXY <- sum(x*y*f); EXY
```

```
[1] 7.23
```

(b) Calculate $\text{Cov}(X,Y)$.

$$\begin{aligned} E[X] = \mu_X &= \sum_{x=1}^{3} x f_X(x) = (1)(0.11) + (2)(0.11) + (3)(0.78) = 2.67 \\ E[Y] = \mu_Y &= \sum_{y=1}^{3} y f_Y(y) = (1)(0.09) + (2)(0.12) + (3)(0.79) = 2.7, \end{aligned}$$

so

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = 7.23 - (2.67)(2.7) \approx$$

(i) $\mathbf{-0.039}$ (ii) **0.021** (iii) **0.139**.

(c) Calculate $\rho(X,Y)$.

$$\begin{aligned} E[X^2] &= \sum_{x=1}^{3} x^2 f_X(x) = (1^2)(0.11) + (2^2)(0.11) + (3^2)(0.78) = 7.57, \\ V[X] = \sigma_X^2 &= E\left[X^2\right] - \left[E[X]\right]^2 = 7.57 - 2.67^2 = 0.4411, \\ E[Y^2] &= \sum_{y=1}^{3} y^2 f_Y(y) = (1^2)(0.09) + (2^2)(0.12) + (3^2)(0.79) = 2.7, \\ V[Y] = \sigma_Y^2 &= E\left[Y^2\right] - \left[E[Y]\right]^2 = 7.68 - 2.7^2 = 0.39, \end{aligned}$$

then the correlation is

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{0.021}{\sqrt{0.4411 \times 0.39}} \approx$$

(i) **0.235** (ii) **0.139** (iii) **0.051**.
There is little linear relationship between the two waiting times.

(d) Let $U_1 = X + Y$ and $U_2 = X - Y$. Then

$$\begin{aligned}
\text{Cov}(U_1, U_2) &= E(U_1U_2) - E(U_1)E(U_2) \\
&= E\left[(X+Y)(X-Y)\right] - E\left((X+Y)\right)E\left((X-Y)\right) \\
&= E\left[X^2 - Y^2\right] - \left[E(X)+E(Y)\right]\left[E(X)-E(Y)\right] \\
&= E\left[X^2\right] - E\left[Y^2\right] - \left\{[E(X)]^2 - [E(Y)]^2\right\} \\
&= E\left[X^2\right] - [E(X)]^2 - \left\{E\left[Y^2\right] - [E(Y)]^2\right\} \\
&= V(X) - V(Y) = 0.4411 - 0.39 \approx
\end{aligned}$$

(i) **0.0355** (ii) **0.0392** (iii) **0.0511**.

(e) *Relationship between covariance and variance.*

$$\text{Cov}(X,X) = E(XX) - E(X)E(X) = E(X^2) - [E(X)]^2 = E(X^2) - \mu_X^2 =$$

(i) $\boldsymbol{\sigma_X^2}$ (ii) $\boldsymbol{\sigma_Y^2}$ (iii) $\mathbf{Cov(Y, Y)}$

and

$$\text{Cov}(Y,Y) = E(YY) - E(Y)E(Y) = E(Y^2) - [E(Y)]^2 = E(Y^2) - \mu_Y^2 =$$

(i) $\boldsymbol{\sigma_X^2}$ (ii) $\boldsymbol{\sigma_Y^2}$ (iii) $\mathbf{Cov(X, X)}$

2. *Linear Correlation Coefficient Using R.*
*Linear correlation coefficient* statistic, $r$, measures *linearity* of scatter diagram. The larger $|r|$, the closer $r$ is to $\pm 1$, the more linear the scatterplot.

(a) *Reading ability versus brightness*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| brightness, $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| reading ability, $y$ | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |

In this case, $r \approx$ (i) **0.704** (ii) **0.723** (iii) **0.734**.

```
brightness <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
reading.ability <- c(70, 70, 75, 88, 91, 94, 100, 92, 90, 85)
cor(brightness,reading.ability)
```

```
[1] 0.7043218
```

So, association between reading ability and brightness is
(i) **positive strong linear**
(ii) **negative moderate linear**
(iii) **positive moderate linear**

(b) *Grain yield versus distance from water*

| dist, x | 0 | 10 | 20 | 30 | 45 | 50 | 70 | 80 | 100 | 120 | 140 | 160 | 170 | 190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| yield, y | 500 | 590 | 410 | 470 | 450 | 480 | 510 | 450 | 360 | 400 | 300 | 410 | 280 | 350 |

In this case, $r \approx$ (i) $\mathbf{-0.724}$ (ii) $\mathbf{-0.785}$ (iii) $\mathbf{-0.950}$.

```
distance <- c(0, 10, 20, 30, 45, 50, 70, 80, 100, 120, 140, 160, 170, 190)
grain.yield <- c(500, 590, 410, 470, 450, 480, 510, 450, 360, 400, 300, 410, 280, 350)
cor(distance,grain.yield)
```

```
[1] -0.7851085
```

So, association between grain yield and distance from water is
(i) **positive strong linear**
(ii) **negative moderate linear**
(iii) **positive moderate linear**

(c) *Annual pizza sales versus student number*

| student number, $x$ | 2 | 6 | 8 | 8 | 12 | 16 | 20 | 20 | 22 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|
| pizza sales, $y$ | 58 | 105 | 88 | 118 | 117 | 137 | 157 | 169 | 149 | 202 |

In this case, $r \approx$ (i) $\mathbf{0.724}$ (ii) $\mathbf{0.843}$ (iii) $\mathbf{0.950}$.

```
student.number <- c(2, 6, 8, 8, 12, 16, 20, 20, 22, 26)
pizza.sales <- c(58, 105, 88, 118, 117, 137, 157, 169, 149, 202)
cor(student.number,pizza.sales)
```

```
[1] 0.950123
```

So, association between pizza sales and student number is
(i) **positive strong linear**
(ii) **negative moderate linear**
(iii) **positive moderate linear**

3. *More linear correlation coefficient*
Match correlation coefficients with scatter plots.

(a) scatter diagram (a): (i) $\boldsymbol{r = -0.7}$ (ii) $\boldsymbol{r = 0}$ (iii) $\boldsymbol{r = 0.3}$

(b) scatter diagram (b): (i) $\boldsymbol{r = -0.7}$ (ii) $\boldsymbol{r = 0.1}$ (iii) $\boldsymbol{r = 1}$

(c) scatter diagram (c): (i) $\boldsymbol{r = -0.7}$ (ii) $\boldsymbol{r = 0}$ (iii) $\boldsymbol{r = 0.7}$

(d) scatter diagram (d): (i) $\boldsymbol{r = -0.7}$ (ii) $\boldsymbol{r = 0}$ (iii) $\boldsymbol{r = 0.7}$

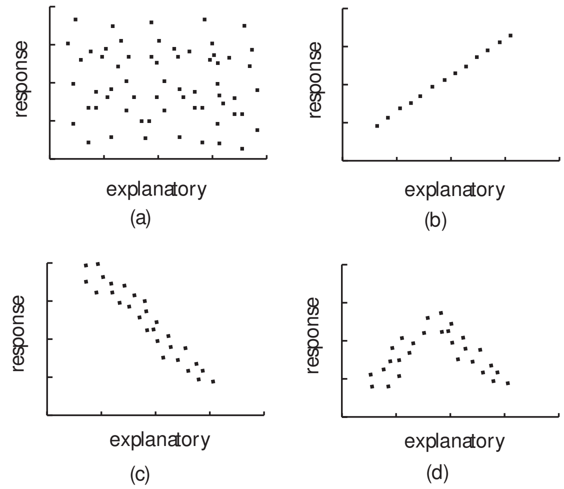


Figure 6.4: Scatter Diagrams and Possible Correlation Coefficients

When $r \neq 0$, $x$ and $y$ are *linearly* related to one another. If $r = 0$, $x$ and $y$ are *nonlinearly* related to one another, which *often* means diagram (a) or sometimes means diagram (d) where positive and negative associated data points cancel one another out. Always show scatter diagram with correlation $r$.

4. *Inference for correlation, $\rho$: reading ability versus brightness.*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| brightness, $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ability to read, $y$ | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |

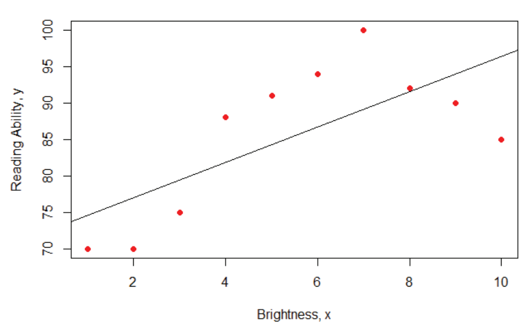


Figure 6.5: Scatterplot, correlation, reading vs brightness

```
brightness <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
reading.ability <- c(70, 70, 75, 88, 91, 94, 100, 92, 90, 85)
plot(brightness,reading.ability,pch=16,col="red",xlab="Diameter, x",ylab="Volume, y")
```

Use sample correlation $r \approx 0.704$ to test if population correlation, $\rho$, is *positive* at a level of significance of 5%. Also, calculate a 95% confidence interval.

(a) *Hypothesis test, right–sided, p-value vs level of significance.*

i. *Statement.*

A. $H_0 : \rho = 0$ versus $H_1 : \rho < 0$

B. $H_0 : \rho = 0$ versus $H_1 : \rho > 0$

C. $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$

ii. *Test.*

Chance $r = 0.704$ or more, *if* $\rho = 0$, is, with $n - 2 = 10 - 2 = 8$ df,

$$\text{p–value} = P(r \geq 0.704) = P\left(r\sqrt{\frac{n-2}{1-r^2}} \geq 0.704\sqrt{\frac{10-2}{1-0.704^2}}\right) \approx P\left(t \geq 2.806\right) \approx$$

(i) **0.002** (ii) **0.011** (iii) **0.058**

Level of significance $\alpha =$ (i) **0.01** (ii) **0.05** (iii) **0.10**.

```
cor1.test <- function(x, y, cor.null, signif.level, type) {
  r <- cor(x,y); n <- length(x); df <- n-2; r; n; df
  t.test.statistic <- r*sqrt((n-2)/(1-r^2))
  if(type=="right") {
    t.crit <- -1*qt(signif.level,df)
    p.value <- 1-pt(t.test.statistic,df)
  }
  dat <- c(cor.null, r, t.crit, t.test.statistic, p.value)
  names(dat) <- c("cor.null", "r", "t crit value", "t test stat", "p value")
  return(dat)
}
cor1.test(brightness, reading.ability, 0, 0.05, "right") # approx t-test for correlation, right-sided

    cor.null            r t crit value  t test stat      p value
  0.00000000   0.70432178   1.85954804   2.80627774   0.01148723
```

iii. *Conclusion.*

Since p–value $= 0.011 < \alpha = 0.050$,
(i) **do not reject** (ii) **reject** null $H_0 : \rho = 0$.
Data indicates population correlation
(i) **smaller than** (ii) **equals** (iii) **greater than** zero (0).
In other words, according to correlation test, reading ability
(i) **is** (ii) **is not** positively linearly associated with brightness.

iv. *Comment.*

The scatterplot (i) **agrees** (ii) **disagrees** with test,
the data is clearly curved.

(b) *Hypothesis test, right–sided, test statistic versus critical value.*

i. *Statement.*

A. $H_0 : \rho = 0$ versus $H_1 : \rho < 0$

B. $H_0 : \rho = 0$ versus $H_1 : \rho > 0$

C. $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$

ii. *Test.*

Test statistic of statistic $r = 2.42$ is

$$t = r\sqrt{\frac{n-2}{1-r^2}} \approx 0.704\sqrt{\frac{10-2}{1-0.704^2}} \approx$$

(i) **2.31** (ii) **2.51** (iii) **2.81**

degrees of freedom, $n - 2 =$ (i) **8** (ii) **9** (iii) **10**

so critical value at $\alpha = 0.05$ is

$t^*_{n-2} = t^*_8 \approx$ (i) **1.86** (ii) **2.31** (iii) **3.31**

```
   cor.null           r t crit value  t test stat      p value
 0.00000000  0.70432178   1.85954804   2.80627774   0.01148723
```

iii. *Conclusion.*

Since $t = 2.81 > t^*_8 \approx 1.86$,

(i) **do not reject** (ii) **reject** null $H_0 : \rho = 0$.

Data indicates population slope

(i) **smaller than** (ii) **equals** (ii) **greater than** zero (0).

In other words, reading ability

(i) **is** (ii) **is not** positively associated with brightness.

iv. *Comment*

Both p-value approach and critical value approach

(i) **agree** (ii) **disagree** with one another.

(c) *95% Confidence interval for $\rho$.*

The 95% CI for correlation of all (reading ability, brightness) $\rho$, is

(i) $\mathbf{(0.034, 0.924)}$ (ii) $\mathbf{(0.134, 0.824)}$ (iii) $\mathbf{(0.134, 0.924)}$.

```
cor1.z.interval <- function(x, y, conf.level) {
  r <- cor(x,y); n <- length(x)
  z.crit <- -1*qnorm((1-conf.level)/2)
  z <- 0.5 * log((1+r)/(1-r), base=exp(1))
  k <- z.crit/sqrt(n-3)

  ci.lower <- tanh(z-k)
  ci.upper <- tanh(z+k)

  dat <- c(r, z.crit, ci.lower, ci.upper)
  names(dat) <- c("r", "Critical Value", "lower bound", "upper bound")
  return(dat)
}
cor1.z.interval(brightness, reading.ability, 0.95) # t-interval for correlation
```

```
            r Critical Value    lower bound    upper bound
    0.7043218      1.9599640      0.1342139      0.9241327
```

## 6.3 Method of Least Squares

The goal is to a create (sample-based) line, $\hat{y} = \hat{m}x_i + \hat{b}$, to fit the scatterplot as best as possible and then use this line to predict values of $y$ for given values of $x_i$. Estimate the values of $m$ and $b$ by using the least-squares criterion, specifically, find numbers $\hat{m}$ and $\hat{b}$ which minimize sum of squared *residuals*, distances between observed $y_i$ and the line, $y_i - \hat{y}_i$,

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left( y_i - \left( \hat{m}x_i + \hat{b} \right) \right)^2.$$

Take the derivative of $S$ with respect to both $\hat{m}$ and $\hat{b}$, set them equal to 0, and solve for both $\hat{m}$ and $\hat{b}$ to find a formula for $\hat{m}$ and $\hat{b}$ in terms of the $x$ and $y$ coordinates:

$$\begin{aligned} \frac{dS}{d\hat{m}} &= \sum 2(y_i - \hat{m}x_i - \hat{b})(-x_i) = 0, \\ \frac{dS}{d\hat{b}} &= \sum 2(y_i - \hat{m}x_i - \hat{b})(-1) = 0, \end{aligned}$$

so

$$\hat{m} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}, \quad \hat{b} = \bar{y} - \hat{m}\bar{x}.$$

**Exercise 6.3 (Method of Least Squares)**

1. *Reading ability versus brightness.*
   Create scatter diagram, calculate least-squares regression line and superimpose line on scatter diagram.

| brightness, $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| reading ability, $y$ | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |

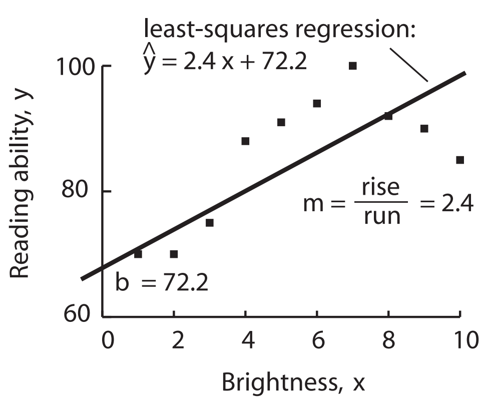


Figure 6.6: Least–squares Line, reading ability versus brightness

```
brightness <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
reading.ability <- c(70, 70, 75, 88, 91, 94, 100, 92, 90, 85)
plot(brightness,reading.ability,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y")
abline(lm(reading.ability~brightness),col="black")
```

(a) *Least-squares regression line.* Choose *two.*
(i) $\boldsymbol{\hat{y} = 72.2 + 2.418x}$
(ii) $\boldsymbol{\hat{y} = 2.418x + 72.2}$
(iii) $\boldsymbol{\hat{y} = 72.2x + 2.418}$
(iv) $\boldsymbol{\hat{y} = 47.04x + 2.944}$

```
linear.regression.predict <- function(x, y, x.zero) {
  n <- length(x)
  sx <- sum(x); sx2 <- sum(x^2)
  sy <- sum(y); sy2 <- sum(y^2); sxy <- sum(x*y)
  slope <- (n*sxy - sx*sy)/(n*sx2-sx^2)
  intercept <- mean(y) - slope*mean(x)
  y <- intercept + slope*x.zero
  regress <- c(intercept,slope,x.zero,y)
  names(regress) <- c("intercept","slope","x","y.predict(x)")
  return(regress)
}
linear.regression.predict(brightness, reading.ability, x.zero=6.5)
```

```
   intercept        slope            x y.predict(x)
   72.200000     2.418182     6.500000    87.918182
```

(b) *Slope and y-intercept of least-squares regression line,* $\hat{y} = 2.418x + 72.2$.
*Slope* is $b_1 =$ (i) **72.2** (ii) **2.418**.
Slope, $b_1 = 2.418$, means, on average, reading ability increases 2.418 units for an increase of *one* unit of brightness.

The *y–intercept* is $b_0 =$ (i) **72.2** (ii) **2.418**.
The $y$–intercept, $b_0 = 72.2$, means average reading ability is 72.2, if brightness is zero.

(c) *Prediction.*
At brightness $x = 6.5$, predicted reading ability is
$\hat{y} \approx 2.418x + 72.2 = 2.418(6.5) + 72.2 \approx$ (i) **84.9** (ii) **85.5** (iii) **87.9**.

(d) *More Prediction.*
At $x = 5.5$, $\hat{y} \approx 2.418(5.5) + 72.2 \approx$ (i) **84.9** (ii) **85.5** (iii) **87.6**.
At $x = 7.5$, $\hat{y} \approx 2.418(7.5) + 72.2 \approx$ (i) **84.9** (ii) **89.5** (iii) **90.4**.

```
linear.regression.predict(brightness, reading.ability, x.zero=5.5)
linear.regression.predict(brightness, reading.ability, x.zero=7.5)
```

```
> linear.regression.predict(brightness, reading.ability, x.zero=5.5)
   intercept        slope            x y.predict(x)
   72.200000     2.418182     5.500000    85.500000
> linear.regression.predict(brightness, reading.ability, x.zero=7.5)
   intercept        slope            x y.predict(x)
   72.200000     2.418182     7.500000    90.336364
```

(e) *Residual.*
At $x = 7$, $\hat{y} \approx 2.418(7) + 72.2 \approx$ (i) **87.9** (ii) **89.1** (iii) **120.6**.

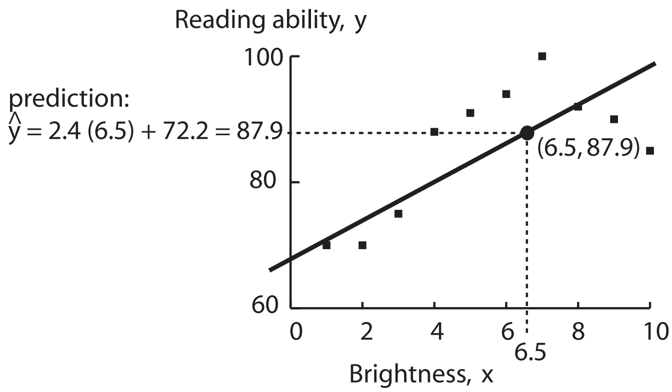


Figure 6.7: Least–Squares Line: Prediction

*Observed* value, $y = 100$ compared to predicted $\hat{y} = 89.1$;
difference between two is *residual*:
$y - \hat{y} = 100 - 89.1 =$ (i) **9.2** (ii) **10.9** (iii) **12.6**.

```
linear.regression.residuals <- function(x, y) {
  n <- length(x)
  sx <- sum(x); sx2 <- sum(x^2)
  sy <- sum(y); sy2 <- sum(y^2); sxy <- sum(x*y)
  slope <- (n*sxy - sx*sy)/(n*sx2-sx^2)
  intercept <- mean(y) - slope*mean(x)
  y.pred <- intercept + slope*x
  residuals <- y - y.pred
  data.stats <- rbind(x, y, y.pred, residuals)
  return(data.stats)
}
linear.regression.residuals(brightness, reading.ability)
```

```
x          1.000000  2.000000  3.000000  4.000000  5.000000  6.000000   7.00000  8.0000000  9.000000  10.00000
y         70.000000 70.000000 75.000000 88.000000 91.000000 94.000000 100.00000 92.0000000 90.000000  85.00000
y.pred    74.618182 77.036364 79.454545 81.872727 84.290909 86.709091  89.12727 91.5454545 93.963636  96.38182
residuals -4.618182 -7.036364 -4.454545  6.127273  6.709091  7.290909  10.87273  0.4545455 -3.963636 -11.38182
```

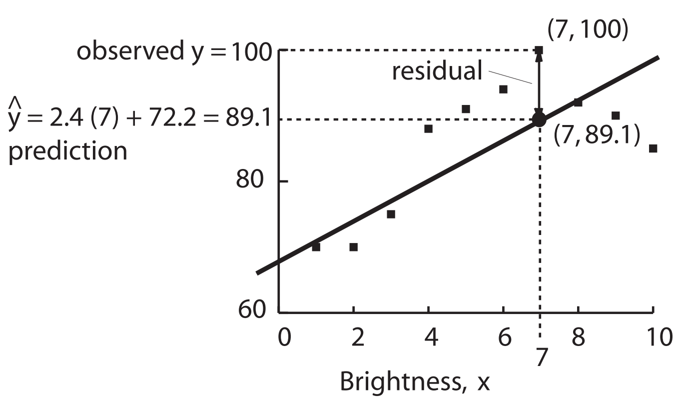


Figure 6.8: Least–Squares Line: Residual

Residual for $x = 7$ is vertical distance between observed (7,100) and predicted (7, 89.1) on least-squares regression line.

(f) *More Residuals.*
At $x = 8$, $y - \hat{y} \approx 92 - 91.5 =$ (i) $\mathbf{-0.5}$ (ii) **0.5** (iii) **1.5**.
At $x = 3$, $y - \hat{y} \approx 75 - 79.5 =$ (i) $\mathbf{-4.5}$ (ii) $\mathbf{-4.5}$ (iii) $\mathbf{-1.5}$.
There are (i) **1** (ii) **5** (iii) **10** residuals on scatter diagram.

(g) *Review.* Second random sample gives (i) **same** (ii) **different** scatter diagram. Statistics calculated from second plot (i) **same** (ii) **different** from statistics calculated from first plot. So, slope, $\hat{m}$, and $y$–intercept, $\hat{b}$, are both (i) **statistics** (ii) **parameters**.

(h) Identify statistical items in example.

| terms | grain yield/water example |
|---|---|
| **(a)** population | **(a)** all (yield, distance) amounts |
| **(b)** sample | **(b)** $\hat{m}, \hat{b}$ |
| **(c)** statistics | **(c)** $m, b$ |
| **(d)** parameters | **(d)** 14 (yield, distance) amounts |

| terms | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| example | | | | |

## 6.4 The Simple Linear Model

The *simple linear model* assumes random variables $X$ and $Y$ are related by

$$Y = mX + b + \epsilon,$$

where $m$ and $b$ are constants, the residual $\epsilon$, also a random variable, is $N(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon^2$ is a constant variance. These assumptions imply

$$Y \text{ is } N(mx + b, \sigma_\epsilon^2).$$

The residual $\epsilon$ is approximated by

$$y_i - \hat{y}_i = y_i - \left(\hat{m}x_i + \hat{b}\right), i = 1, 2, \ldots, n.$$

where statistics $\hat{m}$ and $\hat{b}$ are calculated as given in the previous section and where, notice, mean $\hat{y}_i = E(Y|X = x) = mx + b$. Just like the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

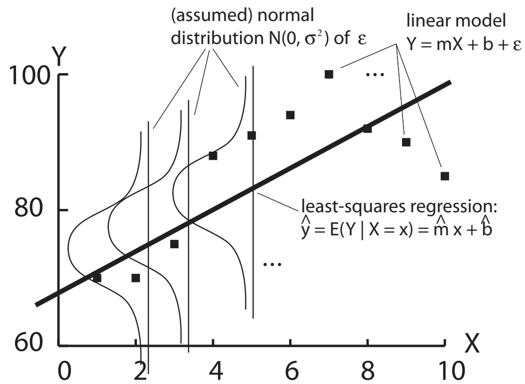


Figure 6.9: Linear regression model

is an estimate of population variance $\sigma^2$ of random variable $X$, one possible sample variance estimate of $\sigma_\epsilon^2$ is, remember $E(\epsilon) = 0$,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left( (y_i - \hat{y}_i) - \hat{E(\epsilon)} \right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left( \left(y_i - \left(\hat{m}x_i + \hat{b}\right)\right) - 0 \right)^2,$$

and an *unbiased* estimate of $\sigma_\epsilon^2$ is given by the *standard error of estimate*,

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left( y_i - \hat{m}x_i - \hat{b} \right)^2.$$

A $100(1-\alpha)\%$ confidence interval and prediction interval of $Y|X = x_0$ of model $Y = mX + b + \epsilon$ are, respectively,

$$\hat{y} \pm t_{\frac{\alpha}{2}, n-2} s_e \sqrt{\frac{1}{n} + \frac{x_0 - \bar{x}}{\sum (x_i - \bar{x})^2}}, \quad \hat{y} \pm t_{\frac{\alpha}{2}, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{x_0 - \bar{x}}{\sum (x_i - \bar{x})^2}}.$$

Test statistic for slope, $H_0 : m = m_0$, and $100(1-\alpha)\%$ confidence interval of slope $m$ of model $Y = mX + b + \epsilon$, are, respectively,

$$t = \frac{1}{s_e} (\hat{m} - m_0) \sqrt{\sum (x_i - \bar{x})^2}, \qquad \hat{m} \pm t_{\frac{\alpha}{2}, n-2} \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

Test statistic for intercept, $H_0 : b = b_0$, and $100(1-\alpha)\%$ confidence interval of intercept $b$ of model $Y = mX + b + \epsilon$, are, respectively,

$$t = \frac{1}{s_e} \left( \hat{b} - b_0 \right) \sqrt{\frac{n \sum (x_i - \bar{x})^2}{\sum x_i^2}}, \qquad \hat{b} \pm t_{\frac{\alpha}{2}, n-2} s_e \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}.$$

**Exercise 6.4 (The Simple Linear Model)**
Consider the reading ability versus brightness data.

| illumination, $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| ability to read, $y$ | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |

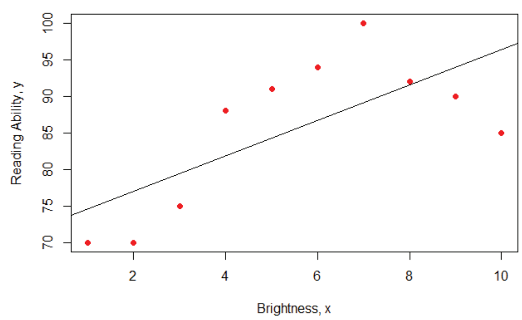


Figure 6.10: Scatterplot, regression, reading vs brightness

```
brightness <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
reading.ability <- c(70, 70, 75, 88, 91, 94, 100, 92, 90, 85)
plot(brightness,reading.ability,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y")
abline(lm(reading.ability~brightness),col="black")
```

Based on $n = 10$ data points, we find sample slope $\hat{m} \approx 2.418$. Check if linear model assumptions true for this data, whether it is possible to perform tests and calculate confidence intervals. Calculate $s_e$. Calculate both a prediction interval and confidence interval of the response (fit, predicted) value $Y$ at $x_0 = 3.5$ and also $x_0 = 6.5$. Test if population slope, $m$, is *positive* at a level of significance of 5%. Also, calculate a 95% confidence interval for the slope $m$.

1. *Check assumptions: does the data fit a linear model?*

| brightness, x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| ability to read, y | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |
| predicted, $\hat{y}$ | 74.6 | 77.0 | 79.5 | 81.9 | 84.3 | 86.7 | 89.1 | 91.5 | 94.0 | 96.4 |
| residual, $y - \hat{y}$ | -4.6 | -7.0 | -4.5 | 6.1 | 6.7 | 7.3 | 10.9 | 0.5 | -4.0 | -8.6 |

```
linear.regression.residuals(brightness, reading.ability)
```

```
x            1.000000   2.000000   3.000000   4.000000   5.000000   6.000000    7.00000   8.0000000   9.000000   10.00000
y           70.000000  70.000000  75.000000  88.000000  91.000000  94.000000  100.00000  92.0000000  90.000000   85.00000
y.pred      74.618182  77.036364  79.454545  81.872727  84.290909  86.709091   89.12727  91.5454545  93.963636   96.38182
residuals   -4.618182  -7.036364  -4.454545   6.127273   6.709091   7.290909   10.87273   0.4545455  -3.963636  -11.38182
```

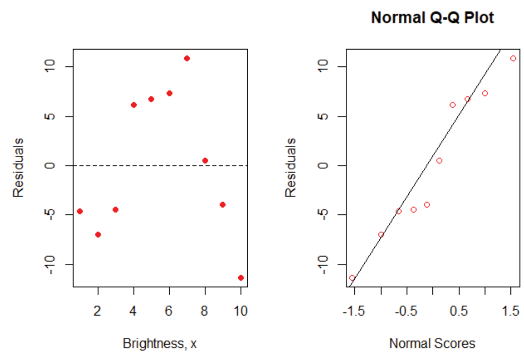


Figure 6.11: Diagnostics of residuals, reading vs brightness

(a) *Linearity assumption/condition?*
According to either scatter diagram or residual plot,
there (i) **is a** (ii) **is no** pattern (around line): points are curved.

(b) *Independence assumption?*
Subjects act (i) **independently** (ii) **dependently** of one another.

(c) *Constant (equal) variance condition?*
According to residual plot, residuals vary -10 and 10 over entire range of brightness; that is, data variance is (i) **constant** (ii) **variable**.

(d) *Nearly normal condition?*
Normal probability plot indicates residuals
(i) **normal** (ii) **not normal**.

```
output <- linear.regression.residuals(brightness, reading.ability); residuals <- output[4,] # residuals 4th row

par(mfrow=c(1,2))
plot(brightness,residuals,pch=16,col="red",xlab="Brightness, x",ylab="Residuals")
abline(h=0,lty=2,col="black")
qqnorm(residuals, col="red", ylab="Residuals", xlab="Normal Scores")
qqline(residuals) # Q-Q (normal probability plot) of residuals check for normality
par(mfrow=c(1,1))
```

2. *Important statistic: residual standard error (deviation),* $s_e$.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| brightness, x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ability to read, y | 70 | 70 | 75 | 88 | 91 | 94 | 100 | 92 | 90 | 85 |
| predicted, $\hat{y}$ | 74.6 | 77.0 | 79.5 | 81.9 | 84.3 | 86.7 | 89.1 | 91.5 | 94.0 | 96.4 |
| residual, $y - \hat{y}$ | -4.6 | -7.0 | -4.5 | 6.1 | 6.7 | 7.3 | 10.9 | 0.5 | -4.0 | -8.6 |
| residuals$^2$, $(y - \hat{y})^2$ | 21.4 | 49.8 | 20.1 | 37.1 | 44.3 | 52.3 | 116.7 | 0.1 | 16.4 | 131.8 |

Total residuals$^2$, $\sum(y - \hat{y})^2 \approx 490.1$, measures how close points are to least-squares line. Residual standard error, $s_e$, measures "average" distance observed data is from least–squares line,

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} \approx \sqrt{\frac{490.1}{10-2}} \approx$$

(i) **1.4** (ii) **6.3** (iii) **7.8**.
Residual standard error is related to least–squares line in much same way standard deviation is related to (i) **average** (ii) **variance**.

```
reading.pred <- 72.2 + 2.428*brightness; n <- 10
residual2 <- (reading.ability - reading.pred)^2; residual2
se <- sqrt(sum(residual2)/(n-2)); sum(residual2); se
```

```
> residual2 <- (reading.ability - reading.pred)^2; residual2
[1]  21.418384  49.787136  20.106256  37.063744  44.355600  52.301824 116.726416   0.141376
[9]  16.418704 131.790400
se <- sqrt(sum(residual2)/(n-2)); sum(residual2); se
[1] 490.1098
[1] 7.827115
```

3. *Inference for fit, $Y|X = x$*

   (a) *Confidence interval (CI) and prediction interval (PI) at $x_0 = 3.5$*
   95% CI for $\hat{y}$ at $x_0 = 3.5$ is

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} s_e \sqrt{\frac{1}{n} + \frac{x_0 - \bar{x}}{\sum(x_i - \bar{x})^2}} \approx$$

   (i) **(54.23, 102.32)** (ii) **(61.32, 100.01)** (iii) **(73.71, 87.62)**.
   95% PI for $\hat{y}$ at $x_0 = 3.5$ is

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{x_0 - \bar{x}}{\sum(x_i - \bar{x})^2}} \approx$$

   (i) **(54.23, 102.32)** (ii) **(61.32, 100.01)** (iii) **(73.71, 87.62)**.

```
simple.reg.fit.t.interval(brightness, reading.ability, 3.5, 0.95, type="prediction")
simple.reg.fit.t.interval(brightness, reading.ability, 3.5, 0.95, type="confidence")
```

```
> simple.reg.fit.t.interval(brightness, reading.ability, 3.5, 0.95, type="prediction")
     x_0          y-hat  Critical Value Margin of Error     lower bound     upper bound
3.500000      80.663636        2.306004       19.342294       61.321342      100.005931
> simple.reg.fit.t.interval(brightness, reading.ability, 3.5, 0.95, type="confidence")
     x_0          y-hat  Critical Value Margin of Error     lower bound     upper bound
3.500000      80.663636        2.306004        6.954829       73.708808       87.618465
```

   CI is **longer** (ii) **shorter** than PI.

(b) *Confidence interval (CI) and prediction interval (PI) at $x_0 = 6.5$*
95% CI for $\hat{y}$ at $x_\nu = 6.5$ is

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} s_e \sqrt{\frac{1}{n} + \frac{x_0 - \bar{x}}{\sum (x_i - \bar{x})^2}} \approx$$

(i) $\mathbf{(81.88, 93.96)}$ (ii) $\mathbf{(68.88, 106.95)}$ (iii) $\mathbf{(66.54, 108.11)}$. 95% PI for $\hat{y}$ at $x_0 = 6.5$ is

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{x_0 - \bar{x}}{\sum (x_i - \bar{x})^2}} \approx$$

(i) $\mathbf{(81.88, 93.96)}$ (ii) $\mathbf{(68.88, 106.95)}$ (iii) $\mathbf{(66.54, 108.11)}$.
CI is **longer** (ii) **shorter** than PI.

```
simple.reg.fit.t.interval(brightness, reading.ability, 6.5, 0.95, type="prediction")
simple.reg.fit.t.interval(brightness, reading.ability, 6.5, 0.95, type="confidence")
```

```
> simple.reg.fit.t.interval(brightness, reading.ability, 6.5, 0.95, type="prediction")
    x_0          y-hat  Critical Value Margin of Error     lower bound     upper bound
6.500000       87.918182        2.306004       19.033621       68.884561      106.951803
> simple.reg.fit.t.interval(brightness, reading.ability, 6.5, 0.95, type="confidence")
    x_0          y-hat  Critical Value Margin of Error     lower bound     upper bound
6.500000       87.918182        2.306004        6.043510       81.874672       93.961692
```

(c) *Confidence (prediction) band, from confidence (prediction) intervals.*

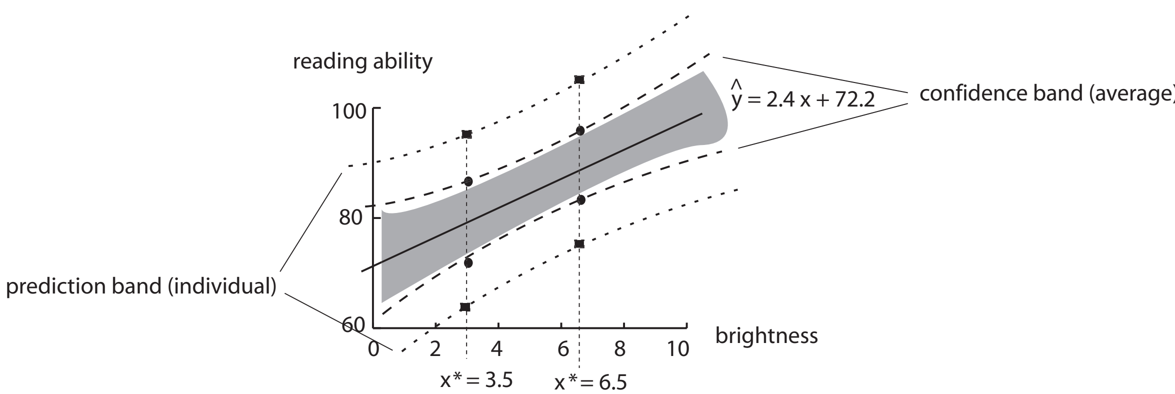


Figure 6.12: CI, PI, confidence and prediction bands

(i) **True** (ii) **False** CIs (PIs) change for different $x_0$ and, together, create a confidence (prediction) *band* of intervals. Confidence (prediction) *band* narrowest at point of averages $(\bar{x}, \bar{y})$.

4. *95% Confidence interval for slope m.*

$$\hat{m} \pm t_{\frac{\alpha}{2},n-2} \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} \approx$$

(i) $\mathbf{2.42 \pm 0.99}$ (ii) $\mathbf{2.42 \pm 1.99}$ (iii) $\mathbf{2.42 \pm 2.99} \approx (0.43, 4.41)$

```
simple.reg.slope.t.interval(brightness, reading.ability, 0.95) # t-interval 95% CI for slope

intercept             s_e  Critical Value Margin of Error     lower bound     upper bound
2.418182        7.826819        2.306004        1.987094        0.431088        4.4052
```

5. *Hypothesis test slope m, right–sided, p-value vs level of significance.*

   (a) *Statement.*

      i. $H_0 : m = 0$ versus $H_1 : m < 0$
      ii. $H_0 : m = 0$ versus $H_1 : m > 0$
      iii. $H_0 : m = 0$ versus $H_1 : m \neq 0$

   (b) *Test.*
   Chance $\hat{m} = 2.42$ or more, *if* $m_0 = 0$, is

   $$\text{p–value} = P(\hat{m} \geq 2.42) = P\left(t \geq \frac{1}{s_e}(\hat{m} - m_0)\sqrt{\sum(x_i - \bar{x})^2}\right) \approx P(t \geq 2.81) \approx$$

   (i) **0.002** (ii) **0.011** (iii) **0.058** (with $n - 2 = 10 - 2 = 8$ df)
   Level of significance $\alpha =$ (i) **0.01** (ii) **0.05** (iii) **0.10**.

   ```
   simple.reg.slope.t.test(brightness, reading.ability, 0, 0.05, "right")

   slope.null        slope t crit value  t test stat      p value
   0.00000000    2.41818182   1.85954804   2.80627774   0.01148723
   ```

   (c) *Conclusion.*
   Since p–value $= 0.011 < \alpha = 0.050$,
   (i) **do not reject** (ii) **reject** null $H_0 : m = 0$.
   Data indicates population slope
   (i) **smaller than** (ii) **equals** (iii) **greater than** zero (0).
   In other words, reading ability
   (i) **is** (ii) **is not** positively associated with brightness.

6. *Hypothesis test slope m, right–sided, test statistic versus critical value.*

   (a) *Statement.*

      i. $H_0 : m = 0$ versus $H_1 : m < 0$
      ii. $H_0 : m = 0$ versus $H_1 : m > 0$
      iii. $H_0 : m = 0$ versus $H_1 : m \neq 0$

   (b) *Test.*
   Test statistic of statistic $\hat{m} = 2.42$ is

   $$t = \frac{1}{s_e}(\hat{m} - m_0)\sqrt{\sum(x_i - \bar{x})^2} \approx$$

   (i) **2.31** (ii) **2.51** (iii) **2.81**
   degrees of freedom, $n - 2 =$ (i) **8** (ii) **9** (iii) **10**
   critical value of level of significance at $\alpha = 0.05$ is
   $t^*_{n-2} = t^*_8 \approx$ (i)**1.31** (ii) **1.86** (iii) **3.31**

```
simple.reg.slope.t.test(brightness, reading.ability, 0, 0.05, "right")

slope.null        slope t crit value  t test stat      p value
0.00000000   2.41818182   1.85954804   2.80627774   0.01148723
```

(c) *Conclusion.*
Since $t = 2.81 > t_8^* \approx 1.86$,
(i) **do not reject** (ii) **reject** null $H_0 : m = 0$.
Data indicates population slope
(i) **smaller than** (ii) **equals** (ii) **greater than** zero (0).
In other words, reading ability
(i) **is** (ii) **is not** positively associated with brightness.