SIMPLE LINEAR REGRESSION (Chapter 11)

Review of some inference and notation: A common population mean model

We begin this chapter by reviewing an important topic: inference regarding a common population mean based on a random sample. We then modify this topic so as to introduce the topic of regression, in particular simple linear regression.

Consider a random sample, $y_1,...,y_n$, from the normal distribution with unknown mean $Ey_i = \mu$ and known variance $Vy_i = \sigma^2$. The model for these data may be written

$$y_i \sim iid\ N(\mu,\sigma^2), \quad i=1,...,n,$$

or equivalently,

$$y_i = \mu + e_i, \quad i=1,...,n,$$

where $e_1,...,e_n \sim iid\ N(0,\sigma^2)$. We refer to $e_1,...,e_n$ as the error terms.

The above model may be called the common population mean (CPM) model. In the context of this model, a key result is that $\dfrac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

This implies that an unbiased estimate of μ is the sample mean, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, an exact $1-\tau$ confidence interval (CI) for μ is $\left(\bar{y} \pm z_{\tau/2}\sigma/\sqrt{n}\right)$, and an exact p-value for testing $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ is $2P\left(Z > \left|\dfrac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}\right|\right)$, where $Z \sim N(0,1)$.

Note 1: For notational simplicity, we will in this chapter not usually distinguish between random variables and their realised values using upper and lower case letters. For example, y_i may refer to both what were denoted Y_i and y_i in previous chapters.

Note 2: We use the symbol τ because we wish to reserve α for another quantity.

Note 3: Suppose that the normal variance σ^2 is unknown. In that case, an important result is that the sample variance

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \left(= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right) \right)$$

is unbiased and consistent for σ^2. It can also be shown that $\frac{(n-1)s_y^2}{\sigma^2} \sim \chi^2(n-1)$, and that s_y^2 is independent of $\bar{y}$. From these results, it follows that $\frac{\bar{y}-\mu}{s_y/\sqrt{n}} \sim t(n-1)$.

This fact implies that an exact $1-\tau$ CI for μ is $\left(\bar{y} \pm t_{\tau/2}(n-1)s_y/\sqrt{n}\right)$, and an exact *p*-value for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ is $2P\left(T_{n-1} > \left| \frac{\bar{y}-\mu_0}{s_y/\sqrt{n}} \right| \right)$, where $T_{n-1} \sim t(n-1)$, and where $t_{\tau/2}(n-1)$ denotes the upper τ-quantile of T_{n-1}.

Note 4: We use the symbol s_y^2 because we wish to reserve s^2 for another quantity.

Note 5: Suppose that the sample values are not normally distributed. Then all of the above inferences are still valid, with an understanding that they are only approximate, but that the approximations improve as the sample size *n* increases, and that the inferences are asymptotically exact, meaning exact in the limit as *n* tends to infinity.

This is true even if σ is unknown and replaced in the relevant formulae by s_y and/or if $t_{\tau/2}(n-1)$ and T_{n-1} are replaced by $z_{\tau/2}$ and *Z*. It is true even if the error terms $e_1, ..., e_n$ are not independent and/or not identically distributed. All that is required is that these terms are uncorrelated with mean zero and finite variance σ^2.

These facts follow by the central limit theorem (CLT) and other results in probability theory. As a very rough rule of thumb, the approximations may be considered 'good' if *n* is 'large', meaning $n \geq 30$ (say).

Introduction to regression: Simple linear regression

Consider the above CPM model, but now instead suppose that the *n* sample values $y_1,...,y_n$ do not all have the same mean μ, but rather that the *i*th value y_i has mean

$$\mu_i = Ey_i = \alpha + \beta x_i, \quad i = 1,...,n,$$

where α and β are unknown constants and $x_1,...,x_n$ are known constants.

This new model may be written

$$y_i \sim \perp N(\mu_i, \sigma^2), \ i = 1,...,n$$

(where $\perp$ signifies independence) or

$$y_i = \mu_i + e_i, \ i = 1,...,n,$$

where $e_1,...,e_n \sim iid\ N(0,\sigma^2)$. As before, we refer to $e_1,...,e_n$ as the error terms.

The idea here is that we believe there to be a linear relationship between two variables *x* and *y* which can be expressed in the form of the equation

$$\mu = Ey = \alpha + \beta x,$$

with $x_1,...,x_n$ being examples of *x,* and with $y_1,...,y_n$ being examples of *y*. In this equation, μ is implicitly a function of α, β and *x*.

The model just described is called the simple linear regression (SLR) model. In the context of this model, we call *y* the dependent variable (since it depends on another variable, namely *x*), and we call *x* the independent variable. Another term for *x* is the covariate variable, and $x_1,...,x_n$ may be referred to as the sample covariate values.

The focus of inference now are the two parameters α and β (and functions thereof), rather than the single parameter μ as previously in the CPM model. We call α and β the SLR parameters. More specifically, α is the intercept parameter, and β is the slope parameter. The next example should explain why this terminology is used.

Note 1: If β = 0 then the SLR model reduces to the CPM model, with $\mu = \alpha$.

Note 2: In some books, the symbols β_0 and β_1 are used instead of α and β, respectively. We will use the latter notation since it is easier to write and say.

Note 3: For definiteness, we have defined the SLR with iid normally distributed errors that have a known variance σ^2. As for the CPM model, we first treat this 'basic' version of the model, and later consider variations (e.g. unknown variance).

Example 1

We are interested in the effects of a particular fertiliser on wheat yield.
To this end, we divide a large field into 7 plots of equal size, similar soil quality, etc.
We then plant wheat in these 7 plots after adding varying amounts of the fertiliser.
At harvest time the yields are observed and recorded, as shown in the following table.

Field, i	Quantity of fertiliser (kg), x_i	Yield (tonnes), y_i
1	0.0	2.0
2	0.5	3.1
3	1.0	3.0
4	1.5	3.8
5	2.0	4.1
6	2.5	4.3
7 (= n)	3.0	6.0

Produce a plot of these data and discuss the relationship between fertiliser and yield.

Solution

The required plot is shown below. There appears to be a positive linear relationship between fertiliser and yield, since we can imagine drawing a straight line which passes roughly through the points. The equation of this imaginary line has the form

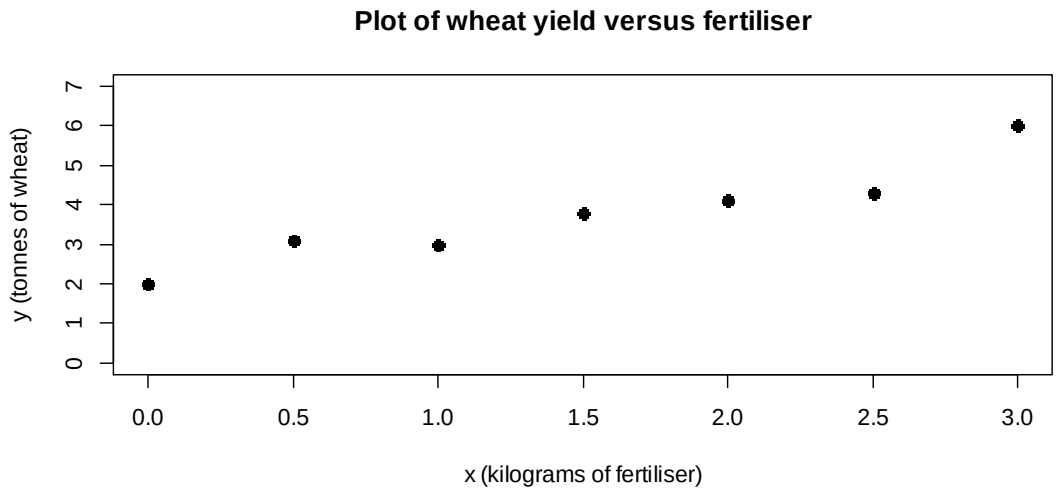
$$Ey = \alpha + \beta x,$$

where α is the intercept and β is the slope of the line. Thus it is plausible that the data follow the SLR model

$$y_i = \mu_i + e_i,\ i = 1,...,n,$$

where $\mu_i = Ey_i = \alpha + \beta x_i$ and $e_1,...,e_n \sim iid\ N(0,\sigma^2)$.

We suppose that the two parameters α and β have some 'true' values which may never be known exactly but which could be estimated.



R Code

```
x = c(0, 0.5, 1, 1.5, 2, 2.5, 3)
y = c(2.0, 3.1, 3.0, 3.8, 4.1, 4.3, 6.0)
X11(w=8,h=4); par(mfrow=c(1,1))
plot(x,y,xlim=c(0,3),ylim=c(0,7), main="Plot of wheat yield versus fertiliser",
     xlab="x (kilograms of fertiliser)",ylab="y (tonnes of wheat)",pch=16,cex=1.2)
```

Least squares estimation

Consider the SLR model given by

$$y_i = \mu_i + e_i, \quad i = 1,...,n$$

where $\mu_i = Ey_i = \alpha + \beta x_i$

and $e_1,...,e_n \sim iid\ N(0,\sigma^2)$, with σ^2 known.

We will now derive formulae for suitable estimates *a* and *b* of α and β, respectively. These estimates will be functions of the observed data pairs, $(x_1, y_1),...,(x_n, y_n)$.

Note: The estimates of α and β could also be denoted $\hat{\alpha}$ and $\hat{\beta}$ (or $\hat{\beta}_0$ and $\hat{\beta}_1$). However, for ease of writing, and speaking, we choose to use the symbols *a* and *b*.

First, define the *i*th fitted mean as

$$\hat{\mu}_i = a + bx_i .$$

Note: This quantity also provides an estimate of $y_i = \alpha + \beta x_i + e_i$ (since $Ee_i = 0$). Therefore, it may also be denoted $\hat{y}_i$ and be referred to as the *i*th fitted value or *i*th predicted value or *i*th predictor. Another notation for $\hat{\mu}_i$ is $\hat{E}y_i$ (since $\mu_i = Ey_i$).

Next define the *i*th error as

$$e_i = y_i - \mu_i = y_i - \alpha - \beta x_i ,$$

and the *i*th fitted error as

$$\hat{e}_i = y_i - \hat{\mu}_i = y_i - \hat{y}_i = y_i - a - bx_i .$$

Now, intuitively, a straight line with equation $y = a + bx$ will provide a good fit to the *n* data points if the sum of the squares of the *n* fitted errors is small. Therefore, one reasonable approach is to choose *a* and *b* so as to make that sum as small as possible.

To formalise this idea, we define the sum of squares for error (SSE) as

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \ .$$

We next write down the partial derivatives of the SSE with respect to *a* and *b*:

$$\frac{\partial SSE}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)^1(-1)$$

$$\frac{\partial SSE}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i)^1(-x_i) \ .$$

Setting these derivatives to zero, respectively, we get:

$$0 = \sum_{i=1}^{n} (y_i - a - bx_i) = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} 1 - b\sum_{i=1}^{n} x_i = n\overline{y} - an - bn\overline{x} \Rightarrow a = \overline{y} - b\overline{x}$$

$$0 = \sum_{i=1}^{n} (y_i - a - bx_i)x_i = \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 \Rightarrow a = \frac{\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2}{n\overline{x}} \ .$$

Equating these two different expressions for *a*, we get

$$\overline{y} - b\overline{x} = \frac{\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2}{n\overline{x}} \qquad \Rightarrow n\overline{x}\overline{y} - nb\overline{x}^2 = \sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2$$

$$\Rightarrow b\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}$$

$$\Rightarrow b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \ .$$

Thus, the required formulae are given by

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \qquad \text{and} \qquad a = \overline{y} - b\overline{x} \ .$$

We call *a* and *b* the least squares estimates (LSEs) of the SLR parameters α and β.

Note 1: Another way to express the LSEs are as

$$b = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad a = \bar{y} - b\bar{x},$$

where: $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \left(= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right)$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \qquad \left(= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right).$$

Yet another way to express b is as $\frac{s_{xy}}{s_{xx}}$, where:

$s_{xy} = \frac{S_{xy}}{n-1}$ (the sample covariance for the two variables)

$s_{xx} = \frac{S_{xx}}{n-1} = s_x^2$ (the sample variance for the covariate variable).

Note 2: The quantities a and b here may also be called the least squares estimates in any context with data of the form $(x_1, y_1), ..., (x_n, y_n)$. The formulae for a and b do not involve σ^2 or depend on any assumptions about the distribution of the errors $e_1, ..., e_n$.

Example 2

For the data in Example 1, find the least squares estimates of the simple linear regression parameters. Then draw the associated line of best fit. Also calculate and display in your graph the fitted values. Also, calculate the fitted errors and the SSE.

Solution

Here: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}(0 + 0.5 + 1 + 1.5 + 2 + 2.5 + 3) = 1.5$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{7}(2 + 3.1 + 3 + 3.8 + 4.1 + 4.3 + 6) = 3.757$

$$\sum_{i=1}^{n} x_i^2 = 0^2 + 0.5^2 + ... + 3^2 = 22.75$$

$$\sum_{i=1}^{n} x_i y_i = 0 \times 2 + 0.5 \times 3.1 + ... + 3 \times 6 = 47.2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 7$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = 7.75 .$$

So the least squares estimates are:

$$b = \frac{S_{xy}}{S_{xx}} = 1.107 \qquad \text{and} \qquad a = \overline{y} - b\overline{x} = 2.096 .$$

The fitted values are:

$$\hat{y}_1 = a + bx_1 = 2.096 + 1.107 \times 0 = 2.096$$

$$\hat{y}_2 = a + bx_2 = 2.096 + 1.107 \times 0.5 = 2.650$$

$$\hat{y}_3 = ... = 3.204 , \qquad \hat{y}_4 = ... = 3.757 , \qquad \hat{y}_5 = ... = 4.311$$

$$\hat{y}_6 = ... = 4.864 , \qquad \hat{y}_7 = ... = 5.418 .$$

So the fitted errors are:

$$\hat{e}_1 = y_1 - \hat{y}_1 = 2 - 2.096 = -0.096$$

$$\hat{e}_2 = y_2 - \hat{y}_2 = 3.1 - 2.65 = 0.45$$

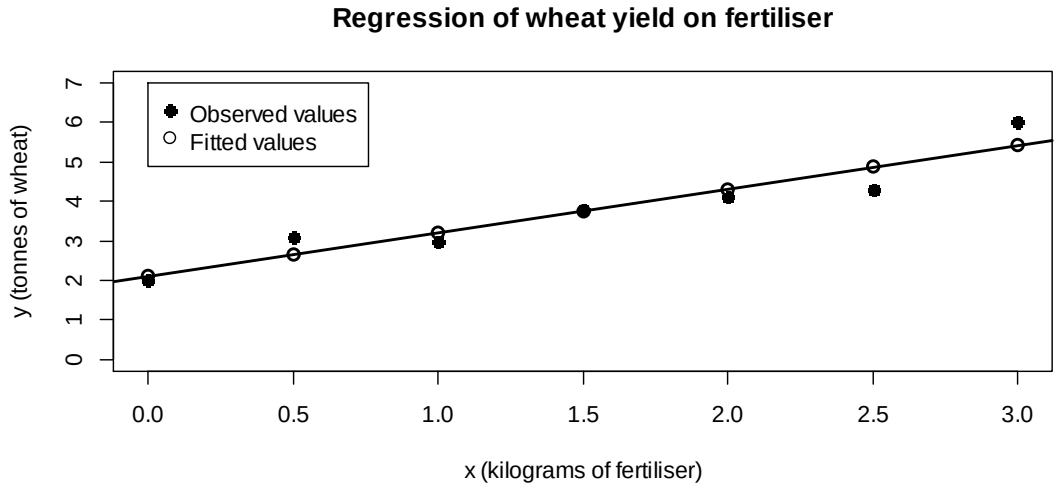
$$\hat{e}_3 = ... = -0.203 , \qquad \hat{e}_4 = ... = 0.043 , \qquad \hat{e}_5 = ... = -0.211$$

$$\hat{e}_6 = ... = -0.564 , \qquad \hat{e}_7 = ... = 0.582 .$$

Finally, the sum of squares for error is

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = (-0.096)^2 + 0.45^2 + ... + 0.582^2 = 0.957 .$$

Below is the required figure, showing the observed values $y_1, ..., y_n$, the estimated regression line $y = a + bx$, and the fitted values, $\hat{y}_i = a + bx_i$, $i = 1,...,n$.



R Code

```
options(digits=4); x = c(0, 0.5, 1, 1.5, 2, 2.5, 3); y = c(2.0, 3.1, 3.0, 3.8, 4.1, 4.3, 6.0)
xbar=mean(x); sumx2=sum(x^2);   ybar=mean(y); sumxy = sum(x*y);  n=length(y)
c(xbar,sumx2,ybar,sumxy) # 1.500 22.750  3.757 47.200
Sxx=sumx2-n*xbar^2; Sxy=sumxy-n*xbar*ybar; c(Sxx, Sxy) # 7.00 7.75
b=Sxy/Sxx; a =ybar-b*xbar; c(a,b) # 2.096 1.107
yhat=a+b*x;  ehat=y-yhat; rbind(yhat, ehat)
# yhat  2.09643 2.65  3.2036 3.75714  4.3107  4.8643 5.4179
# ehat -0.09643 0.45 -0.2036 0.04286 -0.2107 -0.5643 0.5821
SSE = sum(ehat^2); SSE # 0.9568

X11(w=8,h=4)
plot(x,y,xlim=c(0,3),ylim=c(0,7), main="Regression of wheat yield on fertiliser",
      xlab="x (kilograms of fertiliser)",ylab="y (tonnes of wheat)",pch=16,cex=1.2)
abline(a,b,lwd=2);  points(x,yhat,lwd=2,cex=1.2)
legend(0,7,c("Observed values","Fitted values"),
      pch=c(16,1), pt.lwd=c(1,1.5),pt.cex=c(1.2,1.2))
```

Properties of the least squares estimators

We will now determine some properties of the LSEs under the SLR model.
Unless otherwise indicated, all summations will be over $i = 1,...,n$.

Theorem 1: $b = S_{xy} / S_{xx}$ is an unbiased estimator of β.

Proof: First, observe that $S_{xx} = \sum (x_i - \bar{x})^2$ is a constant.

Next, note that $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i - \bar{y} \sum (x_i - \bar{x})$,

where $\sum (x_i - \bar{x}) = \sum x_i - \bar{x} \sum 1 = n\bar{x} - \bar{x}n = 0$.

Then also, $S_{xx} = \sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x}) x_i - \bar{x} \sum (x_i - \bar{x}) = \sum (x_i - \bar{x}) x_i - 0$.

Therefore $ES_{xy} = \sum (x_i - \bar{x}) Ey_i = \sum (x_i - \bar{x})(\alpha + \beta x_i) = \alpha \sum (x_i - \bar{x}) + \beta \sum (x_i - \bar{x}) x_i$

$= 0 + \beta S_{xx}$.

It follows that $Eb = \frac{ES_{xy}}{S_{xx}} = \frac{\beta S_{xx}}{S_{xx}} = \beta$.

Theorem 2: $a = \bar{y} - b\bar{x}$ is an unbiased estimator of α.

Proof: Observe that $E\bar{y} = \frac{1}{n} \sum Ey_i = \frac{1}{n} \sum (\alpha + \beta x_i) = \frac{1}{n} (n\alpha + \beta n\bar{x}) = \alpha + \beta \bar{x}$.

Therefore $Ea = E(\bar{y} - b\bar{x}) = E\bar{y} - \bar{x}Eb = (\alpha + \beta \bar{x}) - \bar{x}\beta = \alpha$ (using Theorem 1).

Note: Theorems 1 and 2 are true under much less restrictive assumptions than those in the SLR model (as we have defined it). These results are true so long as all the error terms $e_1,...,e_n$ have mean 0. These error terms need not be uncorrelated or normal; they don't even need to be identically distributed or have the same variance.

Theorem 3: $Vb = \sigma^2 / S_{xx}$.

Proof: First, $VS_{xy} = V\sum(x_i - \bar{x})y_i = \sum(x_i - \bar{x})^2 Vy_i = S_{xx}\sigma^2$.

Therefore $Vb = V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{S_{xx}\sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$.

Theorem 4: $C(\bar{y}, b) = 0$.

Proof: $C(\bar{y}, b) = C\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{S_{xx}}\sum_{j=1}^{n}(x_j - \bar{x})y_j\right) = \frac{1}{nS_{xx}}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_j - \bar{x})C(y_i, y_j)$.

Now, $C(y_i, y_j) = 0$ for all $i \neq j$, and $C(y_i, y_j) = Vy_i = \sigma^2$ if $i = j$.

Therefore $C(\bar{y}, b) = \frac{1}{nS_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\sigma^2 = \frac{\sigma^2}{nS_{xx}} \times 0 = 0$.

Theorem 5: $Va = \frac{\sigma^2 \sum x_i^2}{nS_{xx}}$.

Proof: First note that $V\bar{y} = \frac{1}{n^2}\sum Vy_i = \frac{1}{n^2}\sum\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$.

So, by Theorems 3 and 4, $Va = V(\bar{y} - b\bar{x}) = V\bar{y} + \bar{x}^2 Vb - 2\bar{x}C(\bar{y}, b)$

$$= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{xx}} - 2\bar{x} \times 0 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$= \sigma^2\left(\frac{S_{xx} + n\bar{x}^2}{nS_{xx}}\right) = \sigma^2\left(\frac{\left(\sum x_i^2 - n\bar{x}^2\right) + n\bar{x}^2}{nS_{xx}}\right) = \frac{\sigma^2\sum x_i^2}{nS_{xx}}.$$

Theorem 6: $C(a,b) = \dfrac{-\bar{x}\sigma^2}{S_{xx}}$.

Proof: $C(a,b) = C(\bar{y} - b\bar{x}, b) = C(\bar{y}, b) - \bar{x}C(b,b) = 0 - \bar{x}Vb = \dfrac{-\bar{x}\sigma^2}{S_{xx}}$.

Theorem 7: Let $\lambda = u + v\alpha + w\beta$, where *u*, *v* and *w* are finite constants.
(Thus, λ is any linear combination of the regression parameters.)

Then: **(i)** an unbiased estimate of λ is $\hat{\lambda} = u + va + wb$

(ii) $V\hat{\lambda} = \dfrac{\sigma^2}{S_{xx}}\left\{v^2 \dfrac{1}{n}\sum x_i^2 + w^2 - 2vw\bar{x}\right\}$.

Proof: (i) $E\hat{\lambda} = u + vEa + wEb = u + v\alpha + w\beta = \lambda$.

(ii) $V\hat{\lambda} = v^2Va + w^2Vb + 2vwC(a,b) = v^2\left(\dfrac{\sigma^2\sum x_i^2}{nS_{xx}}\right) + w^2\left(\dfrac{\sigma^2}{S_{xx}}\right) + 2vw\left(\dfrac{-\bar{x}\sigma^2}{S_{xx}}\right)$

$= \dfrac{\sigma^2}{S_{xx}}\left\{v^2 \dfrac{1}{n}\sum x_i^2 + w^2 - 2vw\bar{x}\right\}$.

Note: Theorems 3 to 7 are true under much less restrictive assumptions than all of those in the SLR model (as we have defined it). These results are true so long as the error terms $e_1, ..., e_n$ are uncorrelated and have the same variance σ^2. The error terms need not be independent or normal; they don't even need to be identically distributed.

Making inferences under the SLR model

Under the assumption in the SLR model that $e_1,...,e_n \sim iid\ N(0,\sigma^2)$, it is true that:

$$\frac{a-\alpha}{\sqrt{Va}} \sim N(0,1), \qquad \frac{b-\beta}{\sqrt{Vb}} \sim N(0,1) \qquad \text{and} \qquad \frac{\hat{\lambda}-\lambda}{\sqrt{V\hat{\lambda}}} \sim N(0,1),$$

where *a*, *b*, $\hat{\lambda}$, *Va*, *Vb* and $V\hat{\lambda}$ are as above. This follows because each of *a*, *b*, and so also $\hat{\lambda}$, is a linear combination of the normally distributed data values, $y_1,...,y_n$.

These results allow us to perform inference on the regression parameters α and β, as well as on any linear combination of the parameters having the form $\lambda = u + v\alpha + w\beta$.

For example, an exact $1-\tau$ CI for β is $\left(b \pm z_{\tau/2}\sqrt{Vb}\right)$, where $Vb = \sigma^2 / S_{xx}$, and an exact *p*-value for testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ is $2P\left(Z > \left|\frac{b-\beta_0}{\sqrt{Vb}}\right|\right)$.

As another example, suppose that we are interested in $\mu = \alpha + x\beta$, the mean (or mathematical expectation) of a *y*-value with some specified covariate value *x*.

Now, μ is just a special case of the general linear combination $\lambda = u + v\alpha + w\beta$ (with *u* = 0, *v* = 1 and $w = x$). Therefore, an unbiased estimate of μ is

$$\hat{\mu} = 0 + 1a + xb = a + xb,$$

and also, by Theorem 7,

$$V\hat{\lambda} = \frac{\sigma^2}{S_{xx}}\left\{v^2\frac{1}{n}\sum x_i^2 + w^2 - 2vw\bar{x}\right\} \quad = \frac{\sigma^2}{S_{xx}}\left\{1^2\frac{1}{n}\sum x_i^2 + x^2 - 2\times 1\times x\times \bar{x}\right\}$$

$$= \frac{\sigma^2}{S_{xx}}\left\{\frac{1}{n}\sum x_i^2 - \frac{n\bar{x}^2}{n} + \bar{x}^2 - 2x\bar{x} + x^2\right\} = \frac{\sigma^2}{S_{xx}}\left\{\frac{1}{n}S_{xx} + (x-\bar{x})^2\right\} = \sigma^2\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\}.$$

It follows that an exact $1-\tau$ CI for μ is $\left(\hat{\mu} \pm z_{\tau/2}\sqrt{V\hat{\mu}}\right)$, and an exact *p*-value for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ is $2P\left(Z > \left|\frac{\hat{\mu}-\mu_0}{\sqrt{V\hat{\mu}}}\right|\right)$.

Example 3

Suppose that the data in Example 1 follow the SLR model with normal errors and standard deviation 0.15. Estimate the slope parameter and the mean of a *y*-value with covariate 2.2. For each quantity, report a point estimate and suitable 95% CI.

Solution

Here (as in Example 2):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}(0+0.5+1+1.5+2+2.5+3) = 1.5$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{7}(2+3.1+3+3.8+4.1+4.3+6) = 3.757$$

$$\sum_{i=1}^{n} x_i^2 = 0^2 + 0.5^2 + ... + 3^2 = 22.75, \quad \sum_{i=1}^{n} x_i y_i = 0\times 2 + 0.5\times 3.1 + ... + 3\times 6 = 47.2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 7, \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = 7.75.$$

$$b = \frac{S_{xy}}{S_{xx}} = 1.107, \quad \sigma = 0.15, \quad Vb = \frac{\sigma^2}{S_{xx}} = 0.003214.$$

Thus, we estimate the slope parameter β by b = 1.107, and a 95% CI for β is

$\left(b \pm z_{\tau/2}\sqrt{Vb}\right)$ = (0.996, 1.218) (using $z_{\tau/2} = z_{0.025} = 1.96$).

Next, we wish to estimate $\mu = \alpha + x\beta$, where x = 2.2. To this end:

$$a = \overline{y} - b\overline{x} = 2.096, \quad \hat{\mu} = a + xb = 4.532,$$

$$V\hat{\mu} = \sigma^2\left\{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right\} = 0.004789.$$

Thus, we estimate the mean μ as $\hat{\mu}$ = 4.532, and a 95% CI for μ is

$\left(\hat{\mu} \pm z_{\tau/2}\sqrt{V\hat{\mu}}\right)$ = (4.397, 4.668).

R Code

```
options(digits=4); sig=0.15; xval = 2.2
x = c(0, 0.5, 1, 1.5, 2, 2.5, 3); y = c(2.0, 3.1, 3.0, 3.8, 4.1, 4.3, 6.0)
xbar=mean(x); sumx2=sum(x^2);   ybar=mean(y); sumxy = sum(x*y);  n=length(y)
c(xbar,sumx2,ybar,sumxy) # 1.500 22.750  3.757 47.200
Sxx=sumx2-n*xbar^2; Sxy=sumxy-n*xbar*ybar; c(Sxx, Sxy) # 7.00 7.75
b=Sxy/Sxx; Vb=sig^2/Sxx; CIbeta=b+c(-1,1)*qnorm(0.975)*sqrt(Vb)
c(b, Vb, CIbeta) # 1.107143 0.003214 0.996023 1.218262
a =ybar-b*xbar; muhat = a + xval*b
Vmuhat = sig^2 * ( 1/n + (xval-xbar)^2 / Sxx )
CImu = muhat + c(-1,1)*qnorm(0.975)*sqrt(Vmuhat)
c(a, muhat, Vmuhat, CImu) # 2.096429 4.532143 0.004789 4.396504 4.667782
```

Prediction

We have already seen how to perform inference on $\mu = \alpha + x\beta$, the mean of a *y*-value with covariate value *x*. This quantity may also be thought of as the average of a hypothetically infinite number of independent 'new' *y*-values, all with covariate *x*.

But what if we wish to perform inference on just a single such value itself (and not on its mean or expectation)?

To this end, we may write the new independent single value of interest as

$$y = \alpha + \beta x + e,$$

where $e \sim N(0, \sigma^2)$ is an error term which is independent of $e_1, ..., e_n$.

Now, $Ee = 0$, and so we may estimate the new value $y = \alpha + \beta x + e$ by $\hat{y} = a + bx$.

Notably, this is exactly the same as the estimate $\hat{\mu} = a + xb$ of $\mu = \alpha + x\beta$.

To construct an interval estimate for $y = \alpha + \beta x + e$, we consider the error in estimation (or prediction), $\hat{y} - y$, whose mean is zero and which has variance

$$V(\hat{y} - y) = V\{(a + bx) - (\alpha + \beta x + e)\} = V\{(a + bx) - e\}.$$
$$= V(a + bx) + Ve - 2C(a + bx, e).$$

Now, *a* and *b* are functions of $y_1, ..., y_n$, which are independent of the new error term, *e*. Therefore $C(a + bx, e) = 0$, and so

$$V(\hat{y} - y) = V\hat{\mu} + Ve - 2 \times 0$$
$$= \sigma^2 \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\} + \sigma^2 + 0 = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right\}.$$

Since $\hat{y} - y$ is a linear combination of normal random variables, we now have that

$$\frac{\hat{y} - y}{\sqrt{V(\hat{y} - y)}} \sim N(0,1),$$

from which it follows that an exact $1 - \tau$ prediction interval (PI) for the new value *y* is

$$\left(\hat{y} \pm z_{\tau/2} \sqrt{V(\hat{y} - y)} \right) = \left(a + bx \pm z_{\tau/2} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right).$$

Note that this interval for $y = \alpha + \beta x + e$ is somewhat wider than the exact $1 - \tau$ CI for $\mu = \alpha + x\beta$, namely

$$\left(\hat{\mu} \pm z_{\tau/2} \sqrt{V\hat{\mu}} \right) = \left(a + bx \pm z_{\tau/2} \sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \right).$$

This makes sense, since there is obviously more variability associated with a single value *y* than with the average of a hypothetically infinite number of such values.

As an exercise, the reader may wish to find an exact $1 - \tau$ PI for the average of *m* independent *y*-values all having covariate *x*. (This should be 'between' the $1 - \tau$ CI for μ and the $1 - \tau$ PI for *y*, and it should converge to the CI as *m* tends to infinity.)

The case of unknown variance

Consider the SLR model above (with $e_1,...,e_n \sim iid\ N(0,\sigma^2)$), but now suppose the normal variance σ^2 is unknown. In that case, an important result is that the quantity

$$s^2 = \frac{SSE}{n-2} \qquad \left(= \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-a-bx_i)^2 \right)$$

is unbiased and consistent for σ^2. See Note 2 below for a proof that $Es^2 = \sigma^2$.

Some other important results in that case are that

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$$

and that s^2 is independent of both *a* and *b*. The proof of these results is omitted.

Note 1: Here, s^2 is not the same as the sample variance, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

Note 2: The following is a proof that $Es^2 = \sigma^2$.

First observe that $E\hat{e}_i = Ey_i - Ea - x_iEb = (\alpha+\beta x_i) - E\alpha - x_i\beta = 0$.

Therefore, $E\hat{e}_i^2 = V\hat{e}_i = Vy_i + Va + x_i^2Vb - 2C(y_i,a) - 2x_iC(y_i,b) + 2x_iC(a,b)$.

Now, $C(y_i,b) = C\left(y_i, \frac{1}{S_{xx}}\sum_{j=1}^{n} y_j(x_j-\bar{x})\right) = C\left(y_i, \frac{1}{S_{xx}} y_i(x_i-\bar{x})\right) = \left(\frac{x_i-\bar{x}}{S_{xx}}\right)\sigma^2$.

Also, $C(y_i,a) = C(y_i, \bar{y}-\bar{x}b) = C\left(y_i, \frac{1}{n}\sum_{j=1}^{n} y_j - \bar{x}\frac{1}{S_{xx}}\sum_{j=1}^{n} y_j(x_j-\bar{x})\right)$

$$= C\left(y_i, y_i\left[\frac{1}{n} - \bar{x}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)\right]\right) = \left[\frac{1}{n} - \bar{x}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)\right]\sigma^2.$$

Therefore, using previous results, we have that

$$E\hat{e}_i^2 = \sigma^2 + \frac{\sigma^2 \sum_{j=1}^{n} x_j^2}{nS_{xx}} + x_i^2 \frac{\sigma^2}{S_{xx}} - 2\left[\frac{1}{n} - \bar{x}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)\right]\sigma^2 - 2x_i\left(\frac{x_i - \bar{x}}{S_{xx}}\right)\sigma^2 + 2x_i\left(\frac{-\bar{x}\sigma^2}{S_{xx}}\right)$$

$$= \frac{\sigma^2}{S_{xx}}\left\{S_{xx} + \frac{1}{n}\sum_{j=1}^{n} x_j^2 + x_i^2 - \frac{2}{n}S_{xx} + 2\bar{x}(x_i - \bar{x}) - 2x_i(x_i - \bar{x}) - 2x_i\bar{x}\right\}$$

$$= \frac{\sigma^2}{S_{xx}}\left\{\left(\frac{n-2}{n}\right)S_{xx} + \frac{1}{n}\sum_{j=1}^{n} x_j^2 - x_i^2 + 2\bar{x}x_i + 2\bar{x}^2\right\}.$$

So $E(SSE) = \sum_{i=1}^{n} E\hat{e}_i = \frac{\sigma^2}{S_{xx}}\left\{n\left(\frac{n-2}{n}\right)S_{xx} + \sum_{j=1}^{n} x_j^2 - \sum_{i=1}^{n} x_i^2 + 2n\bar{x}\bar{x} + 2n\bar{x}^2\right\} = (n-2)\sigma^2,$

and so it follows that $Es^2 = E\left(\frac{SSE}{n-2}\right) = \sigma^2,$ as required.

* * *

The above results can be used to show that:

$$\frac{a-\alpha}{\sqrt{\hat{V}a}} \sim t(n-2), \qquad \frac{b-\beta}{\sqrt{\hat{V}b}} \sim t(n-2) \qquad \text{and} \qquad \frac{\hat{\lambda}-\lambda}{\sqrt{\hat{V}\hat{\lambda}}} \sim t(n-2),$$

where $\hat{V}a = \frac{s^2 \sum x_i^2}{nS_{xx}}$, $\hat{V}b = \frac{s^2}{S_{xx}}$ and $\hat{V}\hat{\lambda} = \frac{s^2}{S_{xx}}\left\{v^2 \frac{1}{n}\sum x_i^2 + w^2 - 2vw\bar{x}\right\}$

(i.e. where $\hat{V}a$, $\hat{V}b$ and $V\hat{\lambda}$ are the same as *Va*, *Vb* and $V\hat{\lambda}$ with σ^2 replaced by s^2).
(Keep in mind that $\lambda = u + v\alpha + w\beta$ and $\hat{\lambda} = u + va + wb$.)

On the basis of these facts, inference regarding α, β and λ can proceed exactly as before, but with σ and $z_{\tau/2}$ everywhere changed to *s* and $t_{\tau/2}(n-2)$, respectively.

For example, an exact $1-\tau$ CI for β is $\left(b \pm t_{\tau/2}(n-2)\sqrt{\hat{V}b}\right)$, where $\hat{V}b = s^2 / S_{xx}$,

and an exact *p*-value for testing $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$ is $2P\left(T_{n-2} > \left|\frac{b-\beta_0}{\sqrt{\hat{V}b}}\right|\right)$.

Also, an exact $1-\tau$ PI for a single new value *y* with covariate value *x* is

$$\left(\hat{y} \pm t_{\tau/2}(n-2)\sqrt{\hat{V}(\hat{y}-y)}\right) = \left(a + bx \pm t_{\tau/2}(n-2)s\sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}}\right).$$

Example 4

Consider Example 1. Assuming that the data follow the SLR model with normal errors, but with the normal variance σ^2 unknown, find a 95% CI for the quantity $\mu = \alpha + x\beta$, where x = 2.2, and also a 95% PI for a new *y*-value with covariate *x*.

Solution

Using results from previous examples, we have that:

$\hat{\mu} = \hat{y} = a + xb = 2.096 + 2.2 \times 1.107$ = 4.532 (same as before)

$s^2 = \frac{SSE}{n-2} = \frac{0.9568}{5}$ = 0.1914

$\hat{V}\hat{\mu} = s^2\left\{\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right\}$ = 0.04073

$t_{\tau/2}(n-2) = t_{0.025}(5) = 2.571.$

So a 95% CI for μ is $\left(\hat{\mu} \pm t_{\tau/2}(n-2)\sqrt{\hat{V}\hat{\mu}}\right)$ = (4.013, 5.051).

Also, $\hat{V}(\hat{y}-y) = s^2\left\{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right\}$ = 0.2321.

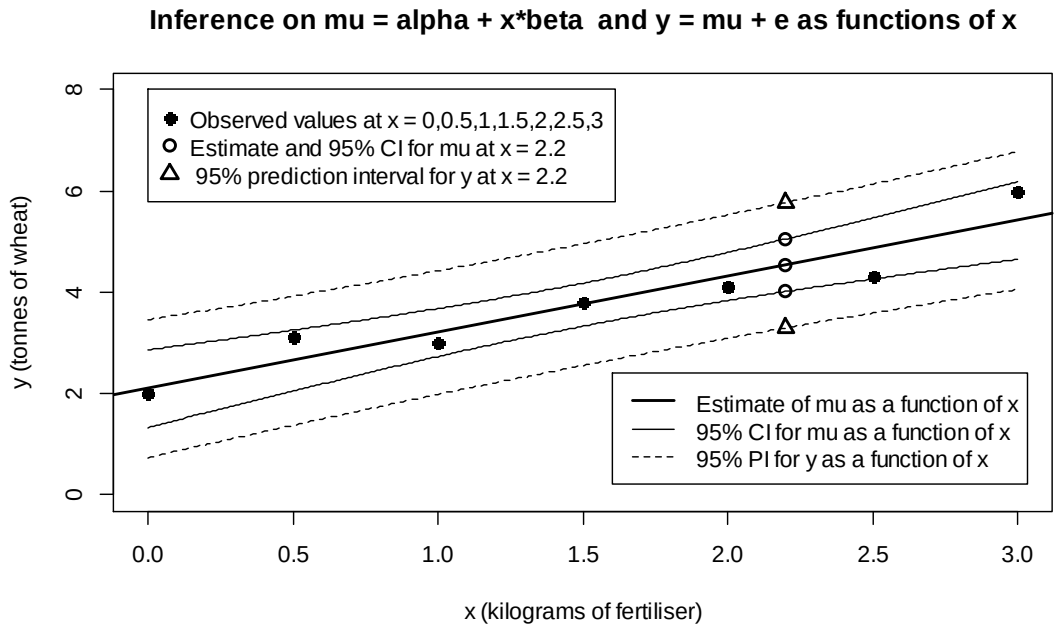
Therefore, a 95% PI for *y* is $\left(\hat{y} \pm t_{\tau/2}(n-2)\sqrt{\hat{V}(\hat{y}-y)}\right)$ = (3.294, 5.771).

Note that the 95% PI has the same centre as, but is wider than, the 95% CI.

Below is a figure showing:

- the observed data pairs in the sample, namely $(x_1, y_1),...,(x_n, y_n)$
- the estimated regression line, $y = a + bx$, where a = 2.096 and b = 1.107
- the estimate of $\mu = \alpha + 2.2\beta$, namely $\hat{\mu} = a + 2.2b$ = 4.532
- the 95% CI for $\mu = \alpha + 2.2\beta$, namely (4.013, 5.051)
- the estimate and 95% CI for $\mu = \alpha + x\beta$ for all x over the range $0 \le x \le 3$
- the 95% PI for $y = \alpha + 2.2\beta + e$, namely (3.294, 5.771)
- the 95% PI for $y = \alpha + x\beta + e$ for all x over the range $0 \le x \le 3$.

Note that one line not shown in the figure is the unknown 'true' regression line given by the equation $Ey = \alpha + x\beta$. In most situations, this will never be known exactly.



R Code

```
options(digits=4); x = c(0, 0.5, 1, 1.5, 2, 2.5, 3); y = c(2.0, 3.1, 3.0, 3.8, 4.1, 4.3, 6.0)
xbar=mean(x); sumx2=sum(x^2);   ybar=mean(y); sumxy = sum(x*y);  n=length(y)
c(xbar,sumx2,ybar,sumxy) # 1.500 22.750  3.757 47.200
Sxx=sumx2-n*xbar^2; Sxy=sumxy-n*xbar*ybar; c(Sxx, Sxy) # 7.00 7.75
b=Sxy/Sxx; a =ybar-b*xbar; c(a,b) # 2.096 1.107
yhat=a+b*x;  ehat=y-yhat; rbind(yhat, ehat)
SSE = sum(ehat^2); SSE # 0.9568
xval=2.2; muhat=a+xval*b; muhat  # 4.532
s2=SSE/(n-2); c(s2, sqrt(s2))  # 0.1914 0.4374
Vhatmuhat=s2*(1/n + (xval-xbar)^2/Sxx); Vhatmuhat # 0.04073
tval=qt(0.975,n-2); tval # 2.571
CI = muhat + c(-1,1)*tval*sqrt(Vhatmuhat); CI # 4.013 5.051

ypred=muhat
Vhatypred=s2*(1+1/n + (xval-xbar)^2/Sxx); Vhatypred   # 0.2321
PI = ypred + c(-1,1)*tval*sqrt(Vhatypred); PI # 3.294 5.771

X11(w=8,h=5)
plot(x,y,xlim=c(0,3),ylim=c(0,8),
        main="Inference on mu = alpha + x*beta  and y = mu + e as functions of x",
        xlab="x (kilograms of fertiliser)",ylab="y (tonnes of wheat)",pch=16,cex=1.2)
abline(a,b,lwd=2);
points(rep(xval,3),c(muhat,CI),lwd=2,cex=1.2)
points(rep(xval,2),PI,pch=2,lwd=2,cex=1.2)

legend(0,8,c("Observed values at x = 0,0.5,1,1.5,2,2.5,3",
        "Estimate and 95% CI for mu at x = 2.2",
        " 95% prediction interval for y at x = 2.2"),
        pch=c(16,1,2), pt.lwd=c(1,2,2),pt.cex=c(1.2,1.2,1.2))
```

```
xvec=seq(0,3,0.01); J=length(xvec); CILBvec= rep(NA,J); CIUBvec= rep(NA,J)
for(j in 1:J){    xvalue=xvec[j]; muhatvalue=a+xvalue*b
        CI = muhatvalue + c(-1,1)*tval*sqrt(s2*(1/n + (xvalue-xbar)^2/Sxx))
        CILBvec[j]=CI[1]; CIUBvec[j]=CI[2]   }
lines(xvec,CILBvec); lines(xvec,CIUBvec)

PILBvec= rep(NA,J); PIUBvec= rep(NA,J)
for(j in 1:J){    xvalue=xvec[j]; muhatvalue=a+xvalue*b
        PI = muhatvalue + c(-1,1)*tval*sqrt(s2*(1+1/n + (xvalue-xbar)^2/Sxx))
        PILBvec[j]=PI[1]; PIUBvec[j]=PI[2]   }
lines(xvec, PILBvec,lty=2); lines(xvec, PIUBvec,lty=2)

legend(1.6,2.4,c("Estimate of mu as a function of x",
        "95% CI for mu as a function of x",
        "95% PI for y as a function of x"),lty=c(1,1,2), lwd=c(2,1,1))
```

The case of non-normality

Suppose that the sample values are not normally distributed. Then all of the above inferences are still valid, with an understanding that they are only approximate, but that the approximations improve as the sample size *n* increases, and that the inferences are asymptotically exact, meaning exact in the limit as *n* tends to infinity.

This is true even if σ is unknown and replaced in the relevant formulae by *s* and/or if $t_{\tau/2}(n-2)$ and T_{n-2} are replaced by $z_{\tau/2}$ and *Z*. It is true even if the error terms $e_1,...,e_n$ are not independent and/or not identically distributed. All that is required is that these terms are uncorrelated with mean zero and finite variance σ^2.

These facts follow by the central limit theorem (CLT) and other results in probability theory. As a very rough rule of thumb, the approximations may be considered 'good' if *n* is 'large', meaning $n \geq 30$ (say).

The relationship between simple linear regression and correlation analysis

We have seen that simple linear regression can be a useful tool for exploring the relationship between two variables *x* and *y*, where *y* is a random variable and *x* is a non-random variable (the covariate). The strength of that relationship is reflected by the least squares estimate $b = \frac{S_{xy}}{S_{xx}}$ of the simple linear regression slope parameter β.

Another useful tool for exploring the relationship between two variables *x* and *y*, but in the context where both are random variables, is correlation analysis. In particular, the strength of that relationship is reflected by the sample correlation, $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$, which provides a consistent estimate of the correlation, $\rho = Corr(x, y) = \frac{C(x, y)}{\sqrt{Vx}\sqrt{Vy}}$.

We see that there is a relationship between the two estimates, namely $r = b\sqrt{\frac{S_{xx}}{S_{yy}}}$.

Another relationship between simple linear regression and correlation analysis is that

$$r^2 = \frac{S_{yy} - SSE}{S_{yy}} = 1 - \frac{SSE}{S_{yy}},$$

where $SSE = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ and $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

(For a proof of this, see Section 11.8 in the textbook).

Here, r^2 may be interpreted as the proportion of the 'total' variation in the *y*-values (around their average, $\bar{y}$) which is 'explained' by the *x*-variable in a simple linear regression (i.e. by the variation of the *y*-values around their fitted values, $a + bx_i$).

We call r^2 the coefficient of determination. The idea of analysing the 'total' variation in a regression (such as of *y* on *x* in our examples) and attributing part of that variation to some variable (like *x*) (or set of variables) is called analysis of variance (ANOVA). This is a topic in more advanced courses on regression.