# 3. Fitting a linear model to a sample by least squares

- Recall the sample version of the linear model. Data are $y_1, y_2, \ldots, y_n$ and on each individual $i$ we have $p$ explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip}$.

$$\text{(LM1)} \qquad y_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i \quad \text{for } i = 1, 2, \ldots, n$$

- Using summation notation, we can equivalently write

$$\text{(LM2)} \qquad y_i = \sum_{j=1}^{p} x_{ij} b_j + e_i \qquad \text{for } i = 1, 2, \ldots, n$$

- We can also use matrix notation. Define column vectors $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_p)$. Define the matrix of explanatory variables, $\mathbb{X} = [x_{ij}]_{n \times p}$. In matrix notation, writing $\text{(LM1)}$ or $\text{(LM2)}$ is exactly the same as

$$\boxed{\text{(LM3)} \qquad \mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e}}$$

- Matrices give a compact way to write the linear model, and also a good way to carry out the necessary computations.

# The least squares formula

- We seek the **least squares** choice of $\mathbf{b}$ that minimizes the sum of squared error, $\sum_{i=1}^n e_i^2$.
- Since $n$ is usually much bigger than $p$, there is usually no value of $\mathbf{b}$ for which we can exactly explain the data using the explanatory matrix $\mathbb{X}$.
- In other words, there is no choice of $\mathbf{b}$ which solves $\mathbb{X}\,\mathbf{b} = \mathbf{y}$.
- The least squares choice of $\mathbf{b}$ turns out to be

$$\boxed{\text{(LM4)} \qquad \mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}}$$

- We will check that this is the formula R uses to fit a linear model.
- We will also gain understanding of (LM4) by studying the **simple linear regression** model $y_i = b_1 x_i + b_2 + e_i$ for which $p = 2$.
- In the simple linear regression model, $b_1$ and $b_2$ are called the slope and the intercept. In general, $b_1, \ldots, b_p$ are called the **coefficients** of the linear model. We call $\mathbf{b}$ the coefficient vector.
- In R, we obtain $\mathbf{b}$ using the `coef()` function as demonstrated below.

# Checking the coefficient estimates from R

- Consider the example from Chapter 1, where `L_detrended` is life expectancy for each year, after subtracting a linear trend, and `U_detrended` is the corresponding detrended unemployment.

```
lm1 <- lm(L_detrended~U_detrended)
coef(lm1)


## (Intercept) U_detrended
##   0.2899928   0.1313673
```

- Now, we can construct the $\mathbb{X}$ matrix corresponding to this linear model and ask R to compute the coefficients using the formula (LM4).

```
X <- cbind(U_detrended,intercept=rep(1,length(U_detrended)))

solve( t(X) %*% X ) %*% t(X) %*% L_detrended


##                  [,1]
## U_detrended 0.1313673
## intercept   0.2899928
```

# Checking the $\mathbb{X}$ matrix we constructed

- The matrix calculation on the previous slide matches the coefficients produced by `lm()`.
- Take some time to check that our R implementation matches the formula (LM4).
- We're fairly sure we got the computation right, because we exactly matched `lm()`, but it is a good idea to look at the $\mathbb{X}$ matrix we constructed.

```
head(X)

##   U_detrended intercept
## 1  -1.0075234         1
## 2   1.1027941         1
## 3   0.4881116         1
## 4  -1.5349043         1
## 5  -1.8662535         1
## 6  -2.0059360         1
```

```
length(U_detrended)

## [1] 68

dim(X)

## [1] 68  2
```

# Naming the $\mathbb{X}$ matrix in the linear model $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$

- "The $\mathbb{X}$ matrix" is not a great name since we would have the same model if we had called it $\mathbb{Z}$.
- Many names are used for $\mathbb{X}$ for the many different purposes of linear models.
- Sheather's textbook calls $\mathbb{X}$ the **matrix of predictor variables** or **matrix of explanatory variables**.
- We call $\mathbb{X}$ the **design matrix** in situations where $x_{ij}$ is the setting of adjustable variable $j$ for the $i$th run of an experiment. For example, $y_i$ could be the stregth of an alloy made up of a fraction $x_{ij}$ of metal $j$ for $j = 1, \ldots, p-1$. We would also want to include an intercept, $x_{ip} = 1$.
- $\mathbb{X}$ can also be called the **matrix of covariates**.
- Sometimes, $\mathbf{y}$ is called the **dependent variable** and $\mathbb{X}$ is the **matrix of independent variables**. Scientifically, an independent variable is one that can be set by the scientist. However, independence has a different technical meaning in statistics.

# Fitted values

- The **fitted values** are the estimates of the data based on the explanatory variables. For our linear model, these fitted values are

$$\hat{y} = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}, \qquad \text{for } i = 1, 2, \ldots, n.$$

- The vector of least squares fitted values $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ is given by

$$\boxed{(\text{LM5}) \qquad \hat{\mathbf{y}} = \mathbb{X}\mathbf{b} = \mathbb{X}\big(\mathbb{X}^{\mathrm{T}}\mathbb{X}\big)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}.}$$

- It is worth checking we now understand how R produces the fitted values for predicting detrended life expectancy using unemployment:

```
my_fitted_values<-X %*% solve(t(X)%*%X) %*% t(X) %*% L_detrended
```

```
lm1$fitted.values[1:2]

## [1] 0.1576371 0.4348639
```

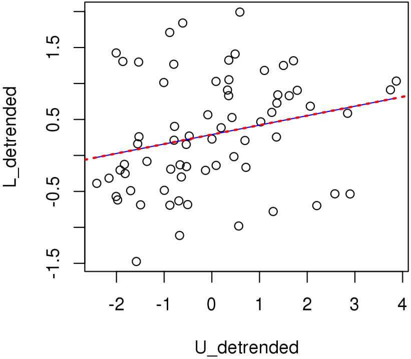
```
my_fitted_values[1:2]

## [1] 0.1576371 0.4348639
```

- We see that the matrix calculation (LM5) exactly matches the fitted values of the `lm1` model that we built earlier using `lm()`.

# Plotting the data

- We have already seen plots of the life expectancy and unemployment data before. When you fit a linear model you should look at the data and the fitted values. We plot the fitted values two different ways.

```
plot(L_detrended~U_detrended)
lines(U_detrended,my_fitted_values,lty="solid",col="blue")
abline(coef(lm1),lty="dotted",col="red",lwd=2)
```



**Question 3.1**. Learn about the `abline()` and `lines()` functions. Explain to yourself why the solid blue line and the dotted red line coincide.

# Deriving the formula for the least squares coefficient vector

**This material will not be tested in the exam. It is presented to show you an application of differentiation and to explain where the formula (LM4) for $\mathbf{b}$ comes from.**

- We derive (LM4) for the simple linear regression model (SLR1).
- The **sum of squared error** is also the **sum of the squared residuals** and is known as the **residual sum of squares (RSS)**. For simple linear regression, this is

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - a x_i - b)^2$$

- To find $a$ and $b$ minimizing RSS, we differentiate with respect to $a$ and $b$ and set the derivatives equal to zero.

  **Question**. Check that

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