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 (LM1) or (LM2) is exactly the same as

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

- 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

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

- 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 ith 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

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

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

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

- 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
```

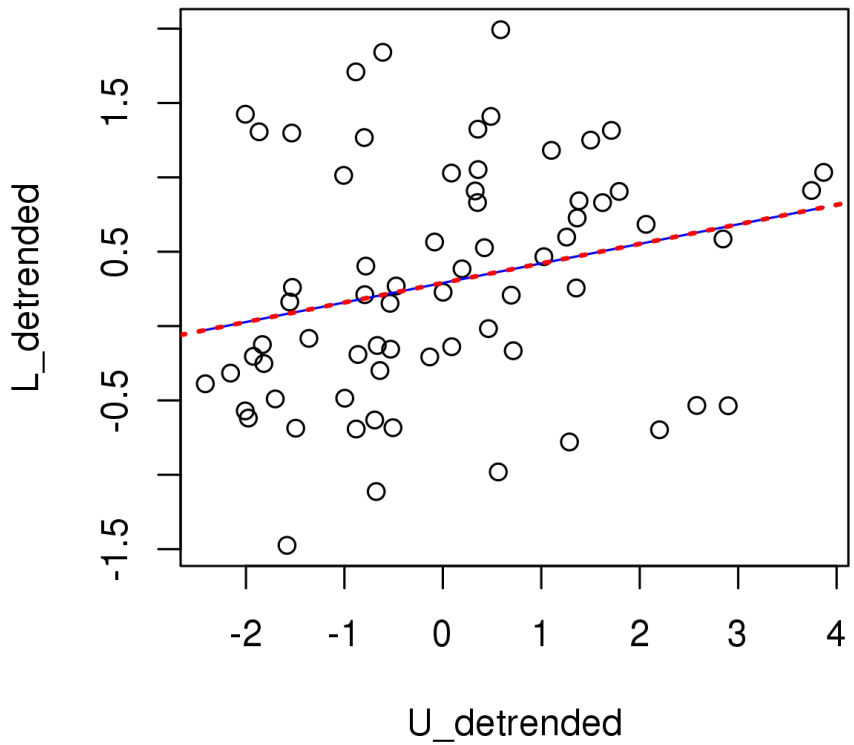
```
my_fitted_values[1:2]

## [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.

Review of summation signs

- To do statistics, we often want to sum things up over all data points so the **summation sign** $\sum_{i=1}^{n}$ comes up frequently.
- The basic trick to understand $\sum_{i=1}^{n}$ is that anything written using a summation sign can be written as a usual sum.
- As long as you can expand from $\sum_{i=1}^{n} z_i$ to $z_1 + z_2 + \cdots + z_n$, and then contract back again from $z_1 + z_2 + \cdots + z_n$ to $\sum_{i=1}^{n} z_i$, then you can use what you already know about $+$ to work with $\sum_{i=1}^{n}$.

Worked example 1. Can we take a constant outside a sum sign? Is it true that

$$\sum_{i=1}^{n} cy_i = c \sum_{i=1}^{n} y_i.$$

Solution 1. Expand the sum, apply the distributive rule, and contract again.

$$\sum_{i=1}^{n} cy_i \;=\; cy_1 + cy_2 + \cdots + cy_n \;=\; c(y_1 + \cdots + y_n) \;=\; c \sum_{i=1}^{n} y_i.$$

More worked examples using summation signs

Worked example 2. What happens if we sum a constant? Show that

$$\sum_{i=1}^{n} c = nc.$$

Solution 2. Expand the sum, and apply basic addition.

$$\sum_{i=1}^{n} c \;=\; c + c + \cdots + c \;(n \text{ times}) \;=\; nc$$

Worked example 3. Can we take a derivative through a sum sign? Is it true that

$$\frac{d}{dm}\Big(\sum_{i=1}^{n} m^2 x_i^2\Big) = \sum_{i=1}^{n} 2mx_i^2$$

Solution. Start by expanding the sum sign.

$$\Big(\sum_{i=1}^{n} m^2 x_i^2\Big) = m^2x_1^2 + m^2x_2^2 + \cdots + m^2x_n^2.$$

Worked example 3 continued

The derivative of a sum is the sum of the derivatives (from Calc I) so we differentiate both sides of the previous expression to give

$$\frac{d}{dm}\Big(\sum_{i=1}^{n} m^2 x_i^2\Big) = 2mx_1^2 + 2mx_2^2 + \cdots + 2mx_n^2.$$

Now, we must recognize that the $+$ expression can be rewritten using $\sum_{i=1}^{n}$. This gives two possible contractions, one of which includes an application of the distributive rule:

$$2mx_1^2 + 2mx_2^2 + \cdots + 2mx_n^2 = \sum_{i=1}^{n} 2mx_i^2 = 2m\sum_{i=1}^{n} x_i^2.$$

Putting these steps together, we obtain

$$\frac{d}{dm}\Big(\sum_{i=1}^{n} m^2 x_i^2\Big) = \sum_{i=1}^{n} 2mx_i^2 = 2m\sum_{i=1}^{n} x_i^2.$$

We have checked that we can pass d/dm through the summation sign.

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

$$\mathrm{RSS} = \sum_{i=1}^{n} \Big(y_i - (m x_i + c) \Big)^2$$

- To find m and c minimizing RSS, we differentiate with respect to m and c and set the derivatives equal to zero.
- When we differentiate RSS with respect to m treating c as a constant, this is called a **partial derivative** and is written as $\partial \mathrm{RSS}/\partial m$.
- If we can find m and c with $\partial \mathrm{RSS}/\partial m = 0$ and $\partial \mathrm{RSS}/\partial c = 0$ we have found a **minimum or maximum** of RSS.
- RSS is non-negative and can get arbitrarily large for bad choices of m and c. It must have a minimum but not a maximum.

Expanding the square in the definition of RSS

- It may be helpful to first expand the square in the definition of RSS to give

$$\mathrm{RSS} = \sum_{i=1}^{n} y_i^2 + m^2 \sum_{i=1}^{n} x_i^2 + nc^2 - 2m \sum_{i=1}^{n} x_i y_i - 2c \sum_{i=1}^{n} y_i + 2mc \sum_{i=1}^{n} x_i$$

Question 3.2. Check that you can work out the expansion of the square. This is an exercise in working with summation signs, similar to worked examples 1 and 2 above.

Differentiating RSS with respect to m and c

Question 3.3. Check that $\dfrac{\partial}{\partial m}\mathrm{RSS} = 2m\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + 2c\sum_{i=1}^{n} x_i.$

Question 3.4. Check that $\dfrac{\partial}{\partial c}\mathrm{RSS} = 2nc - 2\sum_{i=1}^{n} y_i + 2m\sum_{i=1}^{n} x_i.$

Now to minimize RSS, giving the least squares values of m and c, we set the derivatives to zero and solve the resulting simultaneous linear equations for m and c. Canceling the common factor of 2, we get

$$\text{(LS1)} \quad \begin{cases} m\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ m\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases}$$

We want to know if the **least squares equations** (LS1) match (LM4).

Simple linear regression in matrix form

- The matrix form of $y_i = mx_i + c + e_i$, for $i = 1, \ldots, n$, is $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ where $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{b} = (m, c)$, $\mathbf{e} = (e_1, \ldots, e_n)$, $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{1} = (1, 1, \ldots, 1)$ are column vectors, and $\mathbb{X} = [\mathbf{x}\ \mathbf{1}]$.
- Written out in full, this is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Question 3.5. For this $\mathbb{X}$, check that $\mathbb{X}^{\mathrm{T}}\mathbb{X} = \begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}$

Question 3.6. Also, check that $\mathbb{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix}$

The least squares equations in matrix form

- Rather than directly showing that (LS1) has solution $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$, it is easier to check that (LS1) is written in matrix form as

(LS2) $$\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)\mathbf{b} = \mathbb{X}^{\mathrm{T}}\mathbf{y}.$$

Question 3.7. Check that (LS1) in matrix form is

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

- Combining Questions 3.5, 3.6 and 3.7 together, we have found that (LS2) and (LS1) are the same equations. Therefore they must have the same solution, which is $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$.
- We have proved, using differentiation, that $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$ is the least squares coefficient vector for simple linear regression.