Least Square Estimation

(Statistical Modelling I)

Week 1, Lecture 2

Outline

1. Revision
2. Estimating Unknown parameters β_0 and β_1
   - Least Square Estimation
   - Normal Equations
3. How to calculate β_0, β_1 using R
4. Exams Style Questions
5. Next Week targets

The simple linear regression model: Revision

Linear Regression Model: Given observation data (x_i, y_i) for $i = 1, 2, \cdots, n$ we can fit a straight line to describe the response variable Y in terms of the explanatory variable X where

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

where β_0 denotes the intercept and β_1 is the slope of the line .

Stochastic Linear Model: The stochastic linear model can be written either as

$$Y = E[Y_i|X = x_i] + \epsilon_i \text{ or as } Y_i = \beta_0 + \beta_1 X + \epsilon_i$$

for $i = 1, 2, \cdots, n$. Here ϵ_i is the random error.

The random error term is there since there will almost certainly be some variation in Y due strictly to random phenomenon that cannot be predicted or explained.

The simple linear regression model: Revision

The random error: We usually make 3 assumptions about the random error:

1. $E(\epsilon_i) = 0$ for all i
2. $\text{var}(\epsilon_i) = \sigma^2$ for all i
3. $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

About Y_i: Because ϵ_i is a random variable, Y_i must be a random variable.

1. $E(Y_i|X = x_i) = \mu_i = \beta_0 + \beta_1 x_i$ for all i (the dependence of Y on X is linear)
2. $\text{var}(Y_i|X = x_i = \sigma^2$ for all i (the variance of Y at each value of X is constant and does not depend on x_i)
3. $\text{cov}(Y_i|X = x_i, Y_j|x = x_j) = 0$ for all $i \neq j$, (Y_i and Y_j are uncorrelated)

Normal Assumption: We can write this in following 3 ways:

(A) $y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \beta_0 + \beta_1 x_i$
(B) $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
(C) $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where the ϵ_i and iid $\epsilon_i \sim N(0, \sigma^2)$.

Least Square Estimation

Residuals: $\epsilon_i = y_i - \hat{y}_i$ are the difference between the actual value y_i and the predicted value $\hat{y}_i$. Next page shows a hypothetical situation based on six data set points.

The model parameters β_0, β_1 are unknown. With a data we can estimate these parameters. We are interested to find values of the parameters that best explain the data we have observed.

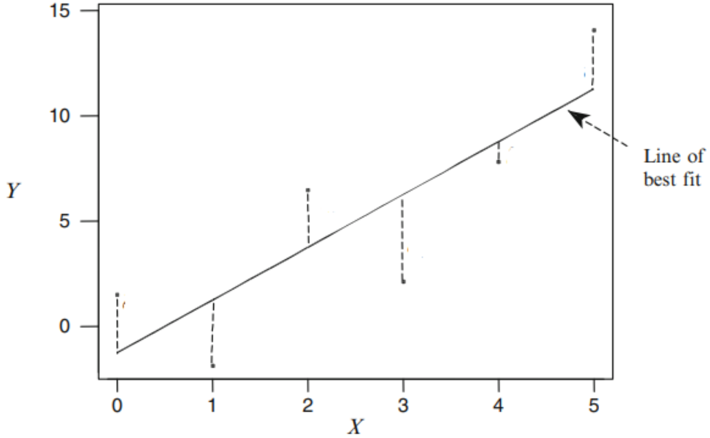
1. **Least Square Estimation**
2. **Maximum Likelihood Estimation**

Today we will consider **Least Square Estimation**. You can find Maximum likelihood estimation in Statistics books.

Queen Mary University of London

What is Least Square Estimation

As the name suggests β_0 and β_1 are chosen to minimize the sum of squared residuals.



This method is also known as **RSS (Residual sum of squares)**.

Least Square Estimation

The least squares estimators of the model parameters β_0 and β_1 are the parameter values that minimise the sum of the squares of the errors $S(\beta_0, \beta_1)$

$$S(\beta_0, \beta_1) = \sum_{i=0}^{n} \epsilon_i^2 = \sum_{i=0}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

To find a minimum:

- differentiate $S(\beta_0, \beta_1)$ with respect to both β_0 and β_1.
- we set each differential equal to zero
- solve the two simultaneous equations in β_0 and β_1
- the values of β_0 and β_1 that satisfy these simultaneous equations are $\hat{\beta}_0$ and $\hat{\beta}_1$

Normal Equations

$$\frac{dS}{d\beta_0} = -2\sum_{i=0}^{n}[y_i - (\beta_0 + \beta_1 x_i)] = 0 \qquad (A)$$

and

$$\frac{dS}{d\beta_1} = -2\sum_{i=0}^{n}[y_i - (\beta_0 + \beta_1 x_i)]x_i = 0 \qquad (B)$$

Now divide by -2 and separate out the terms in the () brackets

Normal Equations

Rearranging terms in the two equations gives

$$\sum_{i=0}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=0}^{n} x_i \qquad (C)$$

$$\sum_{i=0}^{n} x_i y_i = \beta_0 \sum_{i=0}^{n} x_i + \beta_1 \sum_{i=0}^{n} x_i^2 \qquad (D)$$

Finding $\hat{\beta}_0$

If we divide the first *normal equation by n*

$$\widehat{\beta_0} = \frac{1}{n}\sum_{i=0}^{n} y_i - \widehat{\beta_1}\,\frac{1}{n}\sum_{i=0}^{n} x_i$$

or

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

Finding $\hat{\beta}_1$

Substituting for β_0 in the second *normal equation* gives

$$\widehat{\beta_1} = \frac{\sum_{i=0}^{n} x_i y_i - \frac{1}{n} \sum_{i=0}^{n} x_i \sum_{i=0}^{n} y_i}{\sum_{i=0}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=0}^{n} x_i\right)^2}$$

or

$$\widehat{\beta_1} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$ often written in shorthand $$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

Finding $\hat{\beta}_1$

Now in calculus, to check that this is a minimum not a maximum for $S(\beta_0, \beta_1)$ at (β_0, β_1). We need to find all the second derivatives

$$\frac{d^2S}{d\beta_0^2}, \frac{d^2S}{d\beta_1^2}, \frac{d^2S}{d\beta_0\beta_1} \text{ and } \frac{d^2S}{d\beta_1\beta_0}$$

and check that the following determinant is > 0

$$\begin{bmatrix} \frac{d^2S}{d^2\beta_0} & \frac{d^2S}{d\beta_0 d\beta_1} \\ \frac{d^2S}{d\beta_0 d\beta_1} & \frac{d^2S}{d^2\beta_1} \end{bmatrix} = \begin{bmatrix} 2n & 2\sum_{i=0}^{n} x_i \\ 2\sum_{i=0}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Det ()$= 4n \sum_{i=1}^{n}(x_i - \overline{x})^2 > 0$.

The function $S(\beta_0, \beta_1)$ attains a minimum at $(\hat{\beta}_0, \hat{\beta}_1)$.

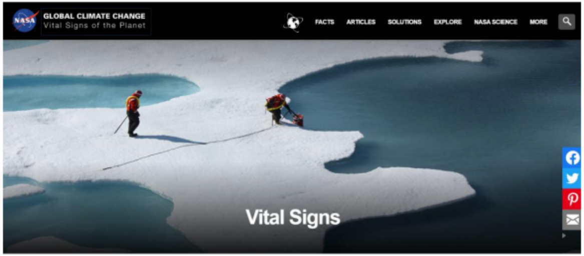
Important notes about β_0 and β_1

- In these equations, $\hat{\beta}_0$ and $\hat{\beta}_0$ are functions of Y as well as of X
- Y is random variable and is generally unknown
- Hence $\hat{\beta}_0$ and $\hat{\beta}_1$ are also random variables
- Because Y is not known, all we can do is calculate values for $\hat{\beta}_0$ and $\hat{\beta}_1$ given a particular set of observations (x_i, y_i)
- The estimator is the algebraic formula depending on the variables X_i and Y_i
- The estimate is that formula evaluated for a certain set if observations (x_i, y_i)
- With a different set of observation data we would expect different values for the estimates

The simple linear regression model

Example: Global average temperature over time

What is the evidence for climate change? How fast are average temperatures rising? NASA Goddard Institute for Space Studies records average surface temperature each year and compares to a baseline temperature of the average for period 1951 - 1980

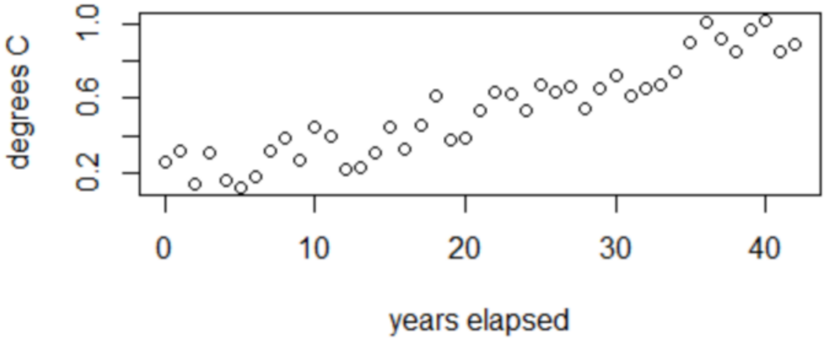


https://climate.nasa.gov/vital-signs/global-temperature/

The simple linear regression model

Scatterplot of 1980 – 2022 data

Global temperature compared to 1951- 1980 baseline



Global Average Temperature data

Using R we can create a vector by c(...) operator so

x=c(...)

y=c(...)

sum(x), sum(y), mean(x), median(x), sd(x), var(x), cor(x, y), cov(x, y)

We can use the "linear model" lm() function in R to find the least squared estimates of the simple linear regression model parameters

```
> model = lm(y~x)
> summary(model)
```

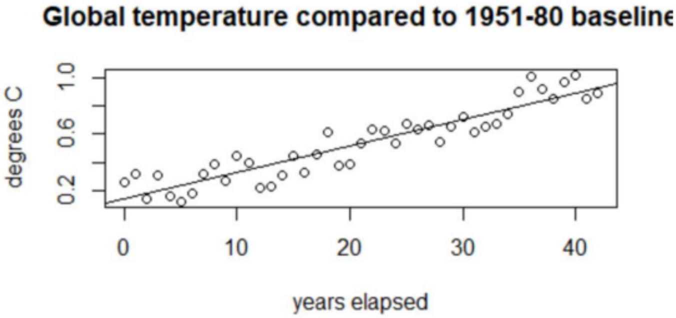
Global Average Temperature data, $\hat{\beta}_0$, $\hat{\beta}_1$

```
Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153268 -0.080700 -0.003953  0.080943  0.193511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.138404   0.028516   4.854 1.79e-05
x           0.018836   0.001169  16.112  < 2e-16

(Intercept) ***
x           ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09513 on 41 degrees of freedom
Multiple R-squared:  0.8636,  Adjusted R-squared:  0.8603
F-statistic: 259.6 on 1 and 41 DF,  p-value: < 2.2e-16
```

Global Average Temperature data, $\hat{\beta}_0$, $\hat{\beta}_1$

```
Call:
lm(formula = y ~ x)

Residuals:
      Min         1Q    Median        3Q       Max
-0.153268 -0.080700 -0.003953  0.080943  0.193511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.138404   0.028516   4.854 1.79e-05
x           0.018836   0.001169  16.112  < 2e-16

(Intercept) ***
x           ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09513 on 41 degrees of freedom
Multiple R-squared:  0.8636,  Adjusted R-squared:  0.8603
F-statistic: 259.6 on 1 and 41 DF,  p-value: < 2.2e-16
```

Model fitted to our temperature data, $\hat{y}_i$



Interpretation of the model result

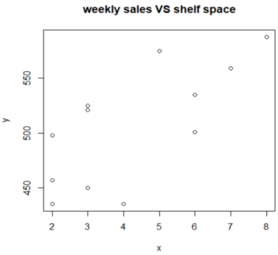
The simple linear regression model fitted may be interpreted as:

- by the end of the 1951 – 1980 baseline period, global average surface temperatures were already 0.138 degrees higher than the average for the 30-year baseline period
- since then annual average surface temperatures have increased by 0.0188 degrees Celsius per year on average based on this NASA GISS data.

Exams Style Question

Question (2022): A baker is interested to find the relationship between the width of the shelf-space for her brand of cookies (x, in feet) and monthly sales (y) of the product in a supermarket. Hence, she fits a model relating monthly sales y to the amount of shelf space x her cookies receive that month. That is, she is fitting the model in the following way $y = \beta_0 + \beta_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$.

x (shelf space)	y (weekly sales)
3	535
2	425
6	575
5	639
3	450
8	630
4	435
2	498
6	534
3	530
2	457
7	559



By looking at the summary output, write down the fitted model.

Exams Style Question

Using the R, we obtained the following output.

```
> mody <- lm(y ~ x)
> summary(mody)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-67.022 -31.346  -0.631  33.654  54.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  429.048     26.519  16.179 1.69e-08 ***
x             18.244      5.643   3.233  0.00898 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.2 on 10 degrees of freedom
Multiple R-squared:  0.511,     Adjusted R-squared:  0.4621
```

Exams Style Question

Question (2021): A life insurance company is examining the force of mortality, μ_x of a particular group of policyholders. It is thought that it is related to the age , x of the policyholders by the formula $\mu_{\mathcal{X}} = Bc^{\mathcal{X}}$.

It is decided to analyze this assumption by using the linear regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \text{where} \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

The results and summary statistics for eight ages were as follows where

Age, x	30	32	34	36	38	40	42	44
Force of mortality, μ_x ($\times 10^{-4}$)	5.84	6.1	6.48	7.05	7.87	9.03	10.56	12.66
$\ln \mu_x$	-7.45	-7.4	-7.34	-7.26	-7.15	-7.01	-6.85	-6.67

$$\sum_i x_i = 296 \qquad \sum_i x_i^2 = 11,120 \qquad \sum_i \ln \mu_{x_i} = -57.129$$

$$\sum_i (\ln \mu_{x_i})^2 = 408.5 \qquad \sum_i x_i \ln \mu_{x_i} = -2,104.5$$

Apply a translation to the original formuale, $\mu_{\mathcal{X}} = Bc^{\mathcal{X}}$, to make it suitable for analysis by linear regression. Hence write down expressions for Y, α and β in the terms of $\mu_{\mathcal{X}}$, B and c. Hence estimate the parameters $\hat{\alpha}$ and $\hat{\beta}$.

Queen Mary University of London

Exams Style Question

Taking log of the original expression gives:

$$\log \mu_x = \log B + x \log c$$

This expression is now linear in x- Comparing the expression with $Y = \alpha + \beta x$ gives:

$$Y = \log \mu_x, \qquad \alpha = \log B, \qquad \beta = \log c$$

Obtaining the estimates of α and β using

$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 11,120 - 8\left(\frac{296}{8}\right)^2 = 168$$

$$S_{xy} = \sum xy - n\bar{x}\bar{y} = -2,104.5 - 8\left(\frac{296}{8}\right)\left(\frac{-57.129}{8}\right) = 9.273$$

Thus, the estimates are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{9.273}{168} = 0.055196$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{-57.129}{8} - 0.0055196 \times \frac{296}{8} = -9.1834$$

Next Week Targets

1. Watch Weekly videos posted at QMplus page
2. Try Questions of Exercise Sheet 1
3. Try Questions of Introduction to R
4. Attend tutorial session to discuss problems from (1), (2) and (3) above.

Queen Mary
University of London