Further Model Check & Matrix Approach to Simple Linear Regression (Statistical Modelling I)

Dr Lubna Shaheen

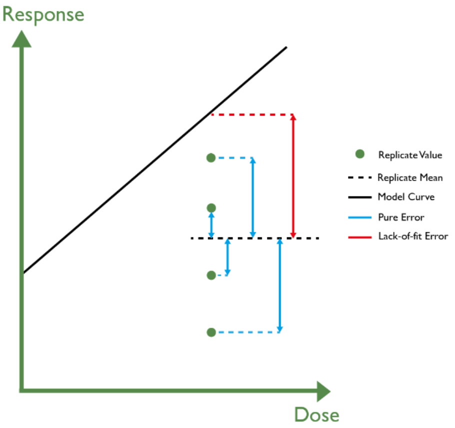
Week 5, Lecture 2

Outline

Residual Sum of Squares

In the simple linear regression model with replications

$$SS_{PE} + SS_{LoF} = SS_E$$



ANOVA

Source of variation	d.f.	SS	MS	VR
Regression	1	SS_R	MS_R	$\frac{MS_R}{MS_E}$
Residual	n – 2	SS_E	$MS_E = \frac{SS_E}{n-2}$	
Lack of Fit	$m-2$	SS_{LoF}	$MS_{LoF} = \frac{SS_{LoF}}{m-2}$	$\frac{MS_{LoF}}{MS_{PE}}$
Pure Error	$n-m$	SS_{PE}	$MS_E = \frac{SS_{PE}}{n-m}$	
Total	n – 1	SS_T		

Expanded ANOVA table

Exam Style Question

A chemist studied the concentration of a solution (Y) over time (x). Fifteen identical solutions were prepared. The solutions were randomly divided into five sets of three, and the five sets were measured, respectively after 1, 3, 5, 7, and 9 hours. Without making any plots the chemist entered the data into R, fitted a simple linear regression model and then carried out a goodness of fit test. The following is the Analysis of Variance table she produced but with some figures missing.

```
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq F value
x              1 12.5971
Residuals     13
  Lack of fit      2.770
  Pure error
Total         14 15.5218
```

(a) Copy and complete the Analysis of Variance Table without using R.

(b) Carry out two possible F tests, write down the corresponding null hypotheses and state your conclusions.

Exam Style Question

ANOVA TABLE:

Exam Style Question

Possible F tests:

Matrix Approach to Simple Linear Regression

Rewrite the model in Matrix form

Our data consists of n paired observations of the predictor variable $\mathbf{X}$ and the response variable $\mathbf{Y}$, i.e. $(x_1, y_1) \cdots (x_n, y_n)$. We wish to fit the model $\mathbf{Y} = \beta_{\mathbf{0}} + \beta_{\mathbf{1}}\mathbf{X} + \epsilon$ where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. We can write this in matrix formulation as

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_n + \epsilon$$

We can write this as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Where $\mathbf{Y}$ is a $(n \times 1)$ vectors of observation y_i, $\mathbf{X}$ is a $(n \times 2)$ matrix called the design matrix where the first column is series of 1 and the second column is the set of observations x_i and β is (2×1) vector of the unknown parameters β_0 and β_1.

Matrix Approach to Simple Linear Regression

Then the n equations can be rewritten

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

which is called **General Linear Model**. Now $\mathbf{Y}$ and ϵ here are random vectors.

Matrix Approach to Simple Linear Regression

The assumption about the random errors make us write $\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$ that is vector ϵ has n-dimensional normal distribution with

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathrm{E}\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}$$

and the variance-covariance matrix

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1, \varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_1, \varepsilon_n) \\ \mathrm{cov}(\varepsilon_2, \varepsilon_1) & \mathrm{var}(\varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\varepsilon_n, \varepsilon_1) & \mathrm{cov}(\varepsilon_n, \varepsilon_2) & \dots & \mathrm{var}(\varepsilon_n) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \boldsymbol{I}$$

Queen Mary
University of London

Matrix Approach to Simple Linear Regression

Remark: All the models we have considered so far can be written in this general form. The dimensions of the matrix $\mathbf{X}$ and of vector β depend on the number p of parameters in the model and respectively they are $n \times p$ and $p \times 1$.
In the full SLRM we have $p = 2$.

- **The null model** $(p = 1)$: $Y_i = \beta_0 + \varepsilon_i$ for $i = 1, \cdots, n$ is equivalent to $Y = 1\beta_0 + \varepsilon$ where 1 is an $(n \times 1)$ vector of 1'.
- **The no-intercept model** $(p = 1)$, $Y_i = \beta_1 x_i + \varepsilon_i$ for $i = 1, \cdots, n$ can be written as in matrix notation with,

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \boldsymbol{\beta} = (\beta_1)$$

Matrix Approach to Simple Linear Regression

- **Quadratic regression, (p=3)**

  $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ for $i = 1, \cdots, n$ can be written in matrix notation with

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & & \\ & & \\ 1 & x_n & x_n^2 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Matrix Approach to Simple Linear Regression

Expectations and Variances with Vectors and Matrices

Vectors $\mathbf{Y}$ and ϵ above are random vectors as their elements are random variables.
Definition: The expected value of a random vector is the vector of the respected values. Thats is for a random vector

$$\mathbf{z} = (\mathbf{z_1}, \cdots, \mathbf{z_n})^{\mathbf{T}}$$

we write

$$E(z) = E \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} E[z_1] \\ E[z_2] \\ \vdots \\ E[z_n] \end{bmatrix}$$

Matrix Approach to Simple Linear Regression

For a random vector $\mathbf{z}$, a constant scalar a, a constant vector $\mathbf{b}$ and for matrices of constants $\mathbf{A}$ and $\mathbf{B}$ we have

(i) $\mathrm{E}\left[az + \mathbf{b}\right] = aE[z] + \mathbf{b}$

(ii) $\mathrm{E}[\mathbf{A}z] = \mathbf{A}E[z]$

(iii) $\mathrm{E}[z^T\mathbf{B}] = E[z]^T\mathbf{B}$

With random vectors, variances and covariances of the random variables z_i together form the dispersion matrix sometimes called the variance-co variance matrix.

$$Var(z) = \begin{bmatrix} var(z_1) & cov(z_1, z_2) & \dots & \dots & cov(z_1, z_n) \\ & & & \dots & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(z_n, z_1) & cov(z_n, z_2) & & \dots & var(z_n) \end{bmatrix}$$

(iv) $Var(z)$ can also be expressed as $E\left[(z - E(z))(z - E(z))^T\right]$

Matrix Approach to Simple Linear Regression

(v) The dispersion matrix is symmetric since cov(z_i, z_j) =cov(z_j, z_i)

(vi) if all of the z_i are uncorrelated all $cov(z_i, z_j) = 0$ and hence the dispersion matrix is diagonal with the variances.

(vii) if **A** is a matrix of constants then Var$(\mathbf{A}z)$ $=\mathbf{A}$ var(z) $\mathbf{A}^T$.

Matrix Approach to Simple Linear Regression

The Multivariate Normal Distribution

A random vector $z = (z_1, z_2, \cdots, z_n)$ has a multivariate normal distribution if its probability density function (pdf) can be written in the form

$$f(z) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{det(\mathbf{V})}} exp^{-\frac{1}{2}(z-\mu)^T\mathbf{V}^{-1}(z-\mu)}$$

where,

- vector μ is the mean of the vector $z = (z_1, \cdots, z_n)$
- $\mathbf{V}$ is the variance-covariance or dispersion matrix of $z = (z_1, \cdots, z_n)$
- $\det(\mathbf{V})$ is the determinant of $\mathbf{V}$

with the multivariate normal distribution we typically use the notation $z \sim N_n(\mu, \mathbf{V})$.

Least Squares Estimation Using Matrices

Least Square Estimation

For the general linear model the normal equations are given by

$$\mathbf{Y} = \mathbf{X}\,\hat{\beta}$$
$$\mathbf{X}^{\mathbf{T}}\,\mathbf{Y} = \mathbf{X}^{\mathbf{T}}\,\mathbf{X}\,\hat{\beta}$$

as $\mathbf{X}^{\mathbf{T}}\,\mathbf{X}$ is invertible, i.e. its determinant is non-zero, the unique solution to the normal equations is given by

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

This matrix $\hat{\beta}$ is a linear combination of the elements of $\mathbf{Y}$. These estimates are normal if $\mathbf{Y}$ is normal. These estimates will be approximately normal in general.

Least Squares Estimation Using Matrices

Least Squares Estimation Using Matrices

Matrix Approach to Simple Linear Regression

Least Squares Estimation Using Matrices

The residual sum of square: $SS_E = \sum(Y_i - \hat{Y})^2$, **df**$_E = n - p$

$$SS_E = \boldsymbol{y}^T\boldsymbol{y} - \widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\boldsymbol{y}$$

The regressionl sum of square:$SS_R = \sum(\hat{Y}_i - \overline{Y})^2$, **df**$_R = p - 1$

$$SS_R = \widehat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{Y} - n\bar{y}^2$$

$E(\hat{\beta})$, $Var(\hat{\beta})$:

$$E(\hat{\beta}) = \hat{\beta} \qquad Var(\hat{\beta}) = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\sigma^2$$

Matrix Approach to Simple Linear Regression

Some Specific Models

- **The Null Model**

  As we have seen this can be written as

$$\mathbf{Y} = \mathbf{X}\widehat{\beta} + \epsilon$$

where $\mathbf{X} = 1$ is an $(n \times 1)$ vector of $1's$. So $\mathbf{X}^\mathbf{T}\mathbf{X} = n$, $\mathbf{X}^\mathbf{T}\mathbf{Y} = \sum Y_i$, which gives

$$\widehat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y} = \frac{1}{n}\sum Y_i = \overline{Y} = \widehat{\beta}_0$$

$$E(\hat{\beta}) = \hat{\beta}_0$$

$$Var(\hat{\beta}) = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\sigma^2 = \frac{\sigma^2}{n}$$

Matrix Approach to Simple Linear Regression

- **No Intercept Model**

  We sat that this example fits the General Linear Model with

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \beta = \beta_1$$

So $\mathbf{X}^\mathbf{T}\mathbf{X} = \sum x_i^2$ and $\mathbf{X}^\mathbf{T}\mathbf{Y} = \sum x_i Y_i$ and we can calculate

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}) = \frac{\sum x_i Y_i}{\sum x_i^2} = \hat{\beta}_1$$

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1} = \frac{\sigma^2}{\sum x_i^2}$$

Queen Mary
University of London

Matrix Approach to Simple Linear Regression

- **Example**

When fitting the model
$$E[Y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$
to a set of $n = 25$ observations, the following results were obtained using the general linear model notation:

$$\boldsymbol{X}^t\boldsymbol{X} = \begin{pmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{pmatrix}, \qquad \boldsymbol{X}^t\boldsymbol{Y} = \begin{pmatrix} 559.60 \\ 7375.44 \\ 337071.69 \end{pmatrix}$$

$$(\boldsymbol{X}^t\boldsymbol{X})^{-1} = \begin{pmatrix} 0.11321519 & -0.00444859 & -0.000083673 \\ -0.00444859 & 0.00274378 & -0.000047857 \\ -0.00008367 & -0.00004786 & 0.000001229 \end{pmatrix}$$

Also $\boldsymbol{Y}^t\boldsymbol{Y} = 18310.63$ and $\bar{Y} = 22.384$.

(a) Find the least squares estimated $\widehat{\boldsymbol{\beta}}$ and hence write down the fitted model;

(b) Use the results to construct the Analysis of Variance Table (Remember that the regression sum of squares is $\widehat{\boldsymbol{\beta}}^t\boldsymbol{X}^t\boldsymbol{Y} - n\bar{y}^2$)

Queen Mary
University of London

Matrix Approach to Simple Linear Regression

Based on the previous results:

(a) Test the null hypothesis that the overall regression is non-significant using a significance level of 5%.

(b) Find a 95% confidence interval for β_j with $j = 0, 1, 2$.

Matrix Approach to Simple Linear Regression

Matrix Approach to Simple Linear Regression

Least Squares Estimation Using Matrices

Exams Style Questions (2021):

Question 4 [17 marks]. We have the data for cigarette consumption for 46 US States for the year 1992 and we are interested in the relationship between the logarithm of cigarette consumption (in packs) per person of smoking age (> 16 years), the so-called $\mathbf{Y}$, the logarithm of real price of cigarettes in each state, $\mathbf{X}_1$, and the logarithm of real disposable income (per capita) in each state, $\mathbf{X}_2$. Data were collected for the 46 US States and the following computations for a multiple regression analysis of the model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

were obtained:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 30.930 & 4.811 & -6.679 \\ 4.811 & 3.945 & -1.177 \\ -6.679 & -1.177 & 1.449 \end{pmatrix}, \qquad \mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 223.001 \\ 45.428 \\ 1064.724. \end{pmatrix}$$

Also $\mathbf{Y}^T\mathbf{Y} = 1082.723$ and $\bar{\mathbf{Y}} = 4.848$ were computed.

(a) Find the least squares estimates $\widehat{\boldsymbol{\beta}}$ and hence write down the fitted model. **[4]**

(b) Use the results to construct the Analysis of Variance Table. **[9]**

Queen Mary
University of London

Least Squares Estimation Using Matrices

Exams Style Questions (2021):

Exams Style Questions (2019)

Question 4. [22 marks]

For the general linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a vector of errors assumed to be uncorrelated with zero mean and constant variance σ^2, the formula for the least squares estimator $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

(a) Prove that the expectation of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}$. [4]

(b) Derive a formula for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, quoting any necessary results. [6]

(c) Show that the vector of fitted values is given by $\boldsymbol{HY}$ where $\boldsymbol{H}$ is the hat matrix which you should define. [3]

(d) Show that $\boldsymbol{HH} = \boldsymbol{H}$. [3]

(e) Express the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i \qquad i = 1, 2, \ldots, 5$$

where the ε_i have mean zero, variance σ^2 and are uncorrelated, as a general linear model in matrix form by specifying $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. [6]

Queen Mary University of London

Exams Style Questions (2020)

Question 3 [19 marks]. For the general linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector of errors assumed to be uncorrelated with zero mean and constant variance σ^2, the formula for the least squares estimator $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

(a) Write the regression model

$$Y_i = \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \qquad i = 1, 2, \ldots, 5,$$

where the ε_i have mean zero, variance σ^2 and are uncorrelated, as a general linear model in matrix form by specifying $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. **[5]**

(b) Find expressions for the least squares estimators of β_1 and β_2,

(i) by minimising

$$S(\beta_1, \beta_2) = \sum_{i=1}^{5} \{Y_i - (\beta_1 x_i + \beta_2 z_i)\}^2,$$

[6]

(ii) by using the formula for $\hat{\boldsymbol{\beta}}$ above. **[5]**

(c) The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Find $\mathrm{Var}(\hat{\beta}_1)$ and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. **[3]**

Least Squares Estimation Using Matrices

Properties follows from the Matrix Approach

Hat Matrix

The vector of fitted values is given by

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\beta}$$
$$= \mathbf{X}(\mathbf{X}^{\mathsf{T}}X\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y}$$

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$ is callled the hat matrix. Note that

$$\mathbf{H}^{\mathsf{T}} = \mathbf{H}$$

and also

$$\mathbf{H}^2 = \mathbf{H}$$

Least Squares Estimation Using Matrices