Simultaneous Inferences and Other Topics in Regression Analysis

Paweł Polak

October 4, 2017

Linear Regression Models - Lecture 5

Content: ALRM Book Chapter 4 (Sec. 4.1-4.4) and Chapter 5 (Sec. 5.1-5.13)

Chapter 4:

- Joint Inference on β_0 and β_1
  - Bonferroni Joint Confidence Intervals
- Simultaneous Inference on Mean Response
  - Working-Hotteling Procedure
  - Bonferroni Procedure
- Simultaneous Prediction Intervals for New Observations
- Regression through Origin

Chapter 5:

- Matrices (5.1-5.8)
  - addition, subtraction, and multiplication, special types of matrices, linear dependence, rank, and inverse of a matrix;
  - random vectors and matrices.
- Simple Linear Regression in Matrix Terms (5.9-5.13)
  - least square estimation, fitted values and residuals, analysis of variance results, inferences in regression analysis.

Simultaneous Inferences

- From chapter 2 we know how to construct two separate confidence intervals for β_0 and β_1.
- Now, we will discuss what to do if we want a confidence level of 95% jointly for both β_0 and β_1.
- One could construct a separate confidence interval for β_0 and β_1. BUT, then the probability of both happening is below 95%.
  - E.g., even if the inferences on β_0 and β_1 were independent, the probability of both being correct would be $(0.95)^2 = 0.9025$
- How to create a joint confidence interval?

Bonferroni Joint Confidence Intervals

- Calculation of Bonferroni joint confidence intervals is a general technique
- We highlight its application in the regression setting
  - Joint confidence intervals for β_0 and β_1
- Intuition
  - Set each separate confidence level to larger than $1 - \alpha$ so that the family coefficient is at least $1 - \alpha$
  - BUT how much larger?

Ordinary Confidence Intervals

- Start with ordinary confidence intervals for β_0 and β_1

$$b_0 \pm t(1-\alpha/2; N-2)s\{b_0\},$$

$$b_1 \pm t(1-\alpha/2; N-2)s\{b_1\},$$

where

$$s^2\{b_0\} = MSE\left[\frac{1}{N} + \frac{\bar{X}^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right],$$

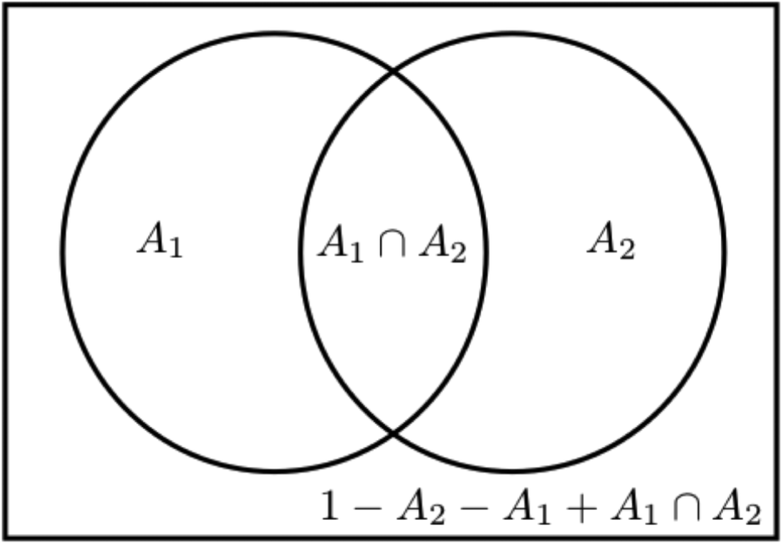
$$s^2\{b_1\} = \frac{MSE}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

- And ask what is the probability that one or both of these intervals are incorrect.

General Procedure

- Let A_1 denote the event that the first confidence interval does not cover β_0, i.e. $P(A_1) = \alpha$
- Let A_2 denote the event that the second confidence interval does not cover β_1, i.e. $P(A_2) = \alpha$
- We want to know the probability that both estimates fall in their respective confidence intervals, i.e. $P(\bar{A}_1 \cap \bar{A}_2)$
- How do we get there from what we know?

Venn Diagram



Bonferroni inequality

- We can see that $P(\bar{A}_1 \cap \bar{A}_2) = 1 - P(A_2) - P(A_1) + P(A_1 \cap A_2)$
  - In a Venn diagram, sizes of sets are equal to their areas and their areas are equal to the probabilities.
- It is also clear that $P(A_1 \cap A_2) \geq 0$
- So,

$$\begin{aligned} P(\bar{A}_1 \cap \bar{A}_2) &\geq 1 - P(A_2) - P(A_1) \\ &= 1 - 2\alpha \end{aligned}$$

Using the Bonferroni inequality cont.

- To achieve a $1-\alpha$ *family* confidence interval for β_0 and β_1 (for example) using the Bonferroni procedure we know that both individual α's must be smaller.

- Returning to our confidence intervals for β_0 and β_1 from before

$$b_0 \pm t(1-\alpha/2; N-2)s\{b_0\}$$

$$b_1 \pm t(1-\alpha/2; N-2)s\{b_1\}$$

- To achieve a $1-\alpha$ *family* confidence interval these intervals must *widen* to

$$b_0 \pm t(1-\alpha/4; N-2)s\{b_0\}$$

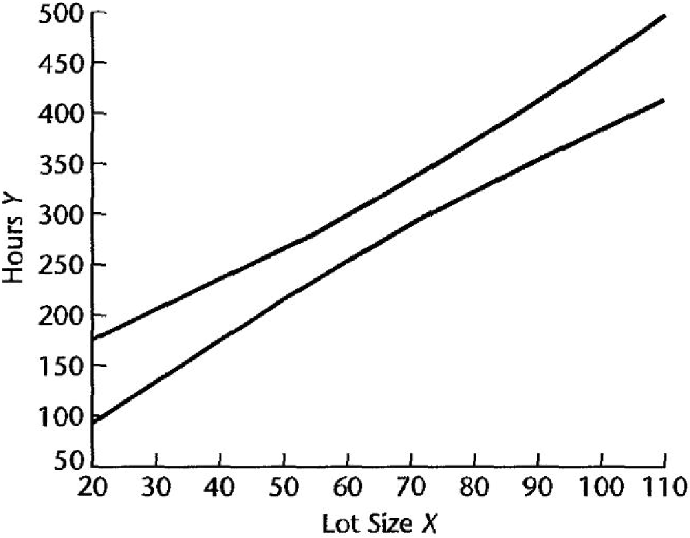
$$b_1 \pm t(1-\alpha/4; N-2)s\{b_1\}$$

- Then $P(\bar{A}_1 \cap \bar{A}_2) \geq 1 - P(A_2) - P(A_1) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha$

Confidence Band for Regression Line

- Remember in Chapter 2.5, we get the confidence interval for $E\{Y_h\}$ to be
$$\hat{Y}_h \pm t(1-\alpha/2; N-2)s\{\hat{Y}_h\}$$
- Now, we want to get a confidence band for the entire regression line $E\{Y\} = \beta_0 + \beta_1 X$
- Bonferroni Procedure
$$\hat{Y}_h \pm B \times s\{\hat{Y}_h\}$$
where $B = t(1 - \frac{\alpha}{2g}; N-2)$, and g is the number of confidence intervals in the family.
- The Working-Hotelling $1-\alpha$ confidence band is
$$\hat{Y}_h \pm W \times s\{\hat{Y}_h\}$$
where $W^2 = 2F(1-\alpha; 2; N-2)$.
  - Same form as Bonferroni, except the B multiple is replaced with the W multiple.

Example confidence band



Bonferroni v.s. Working-Hotelling

- Bonferroni
$$\hat{Y}_a \pm t(1 - \frac{\alpha}{2g}; N - 2)s\{\hat{Y}_h\}$$
- Working-Hotelling
$$\hat{Y}_h \pm W \times s\{\hat{Y}_h\}$$
- In larger families (more X variables) to be considered simultaneously, Working-Hotelling is always tighter, since W stays the same for any number of statements but B becomes larger.
$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{N} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right].$$

Using the Bonferroni inequality cont.

- The Bonferroni procedure is very general. To make joint confidence statements about multiple simultaneous predictions remember that

$$\hat{Y}_h \pm t(1-\alpha/2; N-2)s\{pred\}$$

$$s^2\{pred\} = MSE\left[1 + \frac{1}{N} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right]$$

- If one is interested in a $1-\alpha$ confidence statement about g predictions then Bonferroni says that the confidence interval for each individual prediction must get wider (for each h in the g predictions)

$$\hat{Y}_h \pm t(1-\frac{\alpha}{2g}; N-2)s\{pred\}$$

Note: if a sufficiently large number of simultaneous predictions are made, the width of the individual confidence intervals may become so wide that they are no longer useful.

Simultaneous Prediction Intervals for g New Observations

1. Scheffe procedure
$$\hat{Y} \pm Ss\{pred\}, \tag{1}$$
where $S^2 = gF(1-\alpha; g; N-2)$,

2. Bonferroni procedure
$$\hat{Y} \pm Bs\{pred\}, \tag{2}$$
where $B = t(1-\alpha/(2g); N-2)$.

Both use the same

$$s^2\{pred\} = MSE\left[1 + \frac{1}{N} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right].$$

Regression Through the Origin

Model

$$Y_i = \beta_1 X_i + \varepsilon_i$$

- Sometimes it is known that the regression function is linear and that it must go through the origin,
- then the so called regression through the origin can be used.
- It requires that
    - the range of available data is around zero, and
    - one does not anticipate discontinuity at the origin.
- These are very restrictive conditions which seldom hold in practice.
- X production output, Y labor costs. (What about fixed employment costs?)
- X population, Y GDP. (You usually observe data for populations far away from zero, and the linearity of a GDP by population model is going to break down way before population hits 0)

Regression Through the Origin

Model

$$Y_i = \beta_1 X_i + \varepsilon_i$$

- β_1 is parameter
- X_i are known constants
- ε_i are i.i.d $N(0, \sigma^2)$
- as before the least squares and maximum likelihood estimators for β_1 coincide
- the estimator is $b_1 = \frac{\sum_{i=1}^N X_i Y_i}{\sum_{i=1}^N X_i^2}$

Regression Through the Origin

- In regression through the origin there is only one free parameter (β_1 so the number of degrees of freedom of the MSE

$$s^2 = MSE = \frac{\sum_{i=1}^{N} e_i^2}{N-1} = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-1}$$

is increased by one.

- This is because this is a "reduced" model in the general linear test sense and because the number of parameters estimated from the data is less by one.

Estimate of	Estimated Variance	Confidence Limits	
β_1	$s^2\{b_1\} = \frac{MSE}{\sum X_i^2}$	$b_1 \pm ts\{b_1\}$	(4.18)
$E\{Y_h\}$	$s^2\{\hat{Y}_h\} = \frac{X_h^2 MSE}{\sum X_i^2}$	$\hat{Y}_h \pm ts\{\hat{Y}_h\}$	(4.19)
$Y_{h(\text{new})}$	$s^2\{\text{pred}\} = MSE\left(1 + \frac{X_h^2}{\sum X_i^2}\right)$	$\hat{Y}_h \pm ts\{\text{pred}\}$	(4.20)
		where: $t = t(1-\alpha/2; n-1)$	

A few notes on regression through the origin

- $\sum_{i=1}^{N} e_i \neq 0$ in general now. Only constraint is $\sum_{i=1}^{N} X_i e_i = 0$.
- In case of a curvilinear pattern or linear pattern with a intercept away from the origin, $SSE = \sum_{i=1}^{N} e_i^2$ may exceed the total sum of squares $SSTO = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$.
- Therefore, $R^2 = 1 - SSE/SSTO$ may be negative!
- Generally, it is safer to use the original model opposed with regression-through-the-origin model.

Linear Algebra Review

Definition of Matrix

- Rectangular array of elements arranged in rows and columns

$$\begin{bmatrix} 16000 & 23 \\ 33000 & 47 \\ 21000 & 35 \end{bmatrix}$$

- A matrix has dimensions
- The dimension of a matrix is its number of rows and columns
- It is expressed as 3×2 (in this case)

Indexing a Matrix

- Rectangular array of elements arranged in rows and columns

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

- Matrix $\mathbf{A}$ can also be notated

$$\mathbf{A} = [a_{ij}]\,, i = 1, 2; j = 1, 2, 3$$

Square Matrix and Column Vector

- A square matrix has equal number of rows and columns

$$\begin{bmatrix} 4 & 7 \\ 3 & 9 \end{bmatrix} \qquad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

- A column vector is a matrix with a single column

$$\begin{bmatrix} 4 \\ 7 \\ 10 \end{bmatrix} \qquad \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}$$

- All vectors (row or column) are matrices, all scalars are 1×1 matrices.

Transpose

- The transpose of a matrix is another matrix in which the rows and columns have been interchanged

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 7 & 10 \\ 3 & 4 \end{bmatrix}$$

$$\mathbf{A}^T = \begin{bmatrix} 2 & 7 & 3 \\ 5 & 10 & 4 \end{bmatrix}$$

Equality of Matrices

- Two matrices are the same if they have the same dimension and all the elements are equal

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 4 \\ 7 \\ 3 \end{bmatrix}$$

$\mathbf{A} = \mathbf{B}$ implies $a_1 = 4,\ a_2 = 7,\ a_3 = 3$

Matrix Addition and Substraction

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}$$

Then

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 6 \\ 4 & 8 \\ 6 & 10 \end{bmatrix}$$

Multiplication of a Matrix by a Scalar

$$\mathbf{A} = \begin{bmatrix} 2 & 7 \\ 9 & 3 \end{bmatrix}$$

$$k\mathbf{A} = k\begin{bmatrix} 2 & 7 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} k2 & k7 \\ k9 & k3 \end{bmatrix}$$

Multiplication of two Matrices

$$\underset{2\times 2}{\mathbf{A}} = \begin{bmatrix} 2 & 5 \\ 4 & 1 \end{bmatrix} \qquad \underset{2\times 2}{\mathbf{B}} = \begin{bmatrix} 4 & 6 \\ 5 & 8 \end{bmatrix}$$

$$\begin{array}{cc} & \mathbf{A} \\ \begin{array}{r} \text{Row 1} \\ \text{Row 2} \end{array} & \begin{bmatrix} \boxed{2 \quad 5} \\ 4 \quad 1 \end{bmatrix} \end{array} \qquad \begin{array}{cc} \multicolumn{2}{c}{\mathbf{B}} \\ \multicolumn{2}{c}{\begin{bmatrix} 4 & \boxed{\begin{matrix} 6 \\ 8 \end{matrix}} \\ 5 & \end{bmatrix}} \\ \text{Col. 1} & \text{Col. 2} \end{array} \qquad \begin{array}{ccc} & \multicolumn{2}{c}{\mathbf{AB}} \\ \text{Row 1} & \multicolumn{2}{c}{\begin{bmatrix} 33 & 52 \\ & \end{bmatrix}} \\ & \text{Col. 1} & \text{Col. 2} \end{array}$$

$$\mathbf{A}_{l\times m}\mathbf{B}_{m\times N} = \mathbf{AB}_{l\times N},$$

where

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{m} A_{ik} B_{jk},$$

for $i = 1, ..., l$ and $j = 1, ..., N$.

Another Matrix Multiplication Example

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 1 & 3 & 4 \\ 0 & 5 & 8 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 26 \\ 41 \end{bmatrix}$$

Special Matrices

- If $\mathbf{A} = \mathbf{A}^T$, then $\mathbf{A}$ is a symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix} \qquad \mathbf{A}^T = \begin{bmatrix} 1 & 4 & 6 \\ 4 & 2 & 5 \\ 6 & 5 & 3 \end{bmatrix}$$

- If the off-diagonal elements of a matrix are all zeros it is then called a diagonal matrix

$$\mathbf{A} = \begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

Identity Matrix

A diagonal matrix whose diagonal entries are all ones is an identity matrix. Multiplication by an identity matrix leaves the pre or post multiplied matrix unchanged.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

Vector and matrix with all elements equal to one

$$\mathbf{1}_N = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}_{N\times 1} \qquad \mathbf{J}_N = \begin{bmatrix} 1 & \dots & 1 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & \dots & 1 \end{bmatrix}_{N\times N}$$

$$\mathbf{1}_N \mathbf{1}_N^T = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}_{N\times 1} \begin{bmatrix} 1 & 1 & . & . & . & 1 \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & \dots & 1 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & \dots & 1 \end{bmatrix}_{N\times N} = \mathbf{J}_N$$

$$\mathbf{1}_N^T \mathbf{1}_N = N$$

Linear Dependence and Rank of Matrix

Consider

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \end{bmatrix}$$

and think of this as a matrix of a collection of column vectors.
Note that the third column vector is a multiple of the first column vector.

Linear Dependence

When m scalars$k_1, ..., k_m$ not all zero, can be found such that:

$$k_1 A_1 + ... + k_m A_m = 0$$

where 0 denotes the zero column vector and A_i is the i^{th} column of matrix **A**, the m column vectors are called linearly dependent. If the only set of scalars for which the equality holds is $k_1 = 0, ..., k_m = 0$, the set of m column vectors is linearly independent.

In the previous example matrix the columns are linearly dependent.

$$5 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 0 \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} - 1 \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix} + 0 \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Rank of Matrix

The rank of a matrix is defined to be the maximum number of linearly independent columns in the matrix.

Rank properties include:

- The rank of a matrix is unique
- The rank of a matrix can equivalently be defined as the maximum number of linearly independent rows
- The rank of an $r \times c$ matrix cannot exceed $min(r, c)$
- The row and column rank of a matrix are equal
- The rank of a matrix is preserved under nonsingular transformations., i.e. Let $\mathbf{A}(N \times N)$ and $\mathbf{C}(k \times k)$ be nonsingular matrices. Then for any $N \times k$ matrix $\mathbf{B}$ we have

$$rank(\mathbf{B}) = rank(\mathbf{AB}) = rank(\mathbf{BC})$$

Inverse of Matrix

- Like a reciprocal

$$6 * 1/6 = 1/6 * 6 = 1$$

$$x\frac{1}{x} = 1$$

- But for matrices

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Uses of inverse matrix

- Suppose that we have an equation:

$$\mathbf{A}\mathbf{W} = \mathbf{C},$$

  where both $\mathbf{A}$ and $\mathbf{C}$ are known.

- Solve for $\mathbf{W}$ by multiplying both sides by $\mathbf{A}^{-1}$, i.e.,

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{W} = \mathbf{A}^{-1}\mathbf{C} \Rightarrow \mathbf{W} = \mathbf{A}^{-1}\mathbf{C}.$$

Example

$$\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} -.1 & .4 \\ .3 & -.2 \end{bmatrix}$$

$$\mathbf{A}^{-1}\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

More generally,

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{D}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

where
$D = ad - bc$

Inverses of Diagonal Matrices are Easy

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

then

$$\mathbf{A}^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$$

Finding the inverse

- Finding an inverse takes (for general matrices with no special structure)

$$O(N^3)$$

  operations (when N is the number of rows in the matrix)
- We will assume that numerical packages can do this for us in R: *solve*($\mathbf{A}$) gives the inverse of matrix $\mathbf{A}$

Example

Solving a system of simultaneous equations

$$2y_1 + 4y_2 = 20$$

$$3y_1 + y_2 = 10$$

$$\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

List of Useful Matrix Properties

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$$

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$$

$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$$

$$k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$$

$$(\mathbf{A}^T)^T = \mathbf{A}$$

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}, (\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$$

Random Vectors and Matrices

Let's say we have a vector consisting of three random variables

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

The expectation of a random vector is defined as

$$\mathbb{E}(\mathbf{Y}) = \begin{pmatrix} \mathbb{E}(Y_1) \\ \mathbb{E}(Y_2) \\ \mathbb{E}(Y_3) \end{pmatrix}$$

The expectation of a random matrix is defined similarly

$$\mathbb{E}(\mathbf{Y}) = [\mathbb{E}(Y_{ij})]$$

for $i = 1, \dots, N$; and $j = 1; \dots; p$.

Variance-covariance Matrix of a Random Vector

The variances of three random variables $\sigma^2(Y_i)$ and the covariances between any two of the three random variables $\text{Cov}(Y_i; Y_j)$, are assembled in the variance-covariance matrix of $\mathbf{Y}$

$$\text{Cov}(\mathbf{Y}) = \sigma^2\{\mathbf{Y}\} = \begin{pmatrix} \sigma^2(Y_1) & \sigma(Y_1, Y_2) & \sigma(Y_1, Y_3) \\ \sigma(Y_2, Y_1) & \sigma^2(Y_2) & \sigma(Y_2, Y_3) \\ \sigma(Y_3, Y_1) & \sigma(Y_3, Y_2) & \sigma^2(Y_3) \end{pmatrix}$$

remember $\sigma(Y_1, Y_2) = \sigma(Y_2, Y_1)$ so the covariance matrix is symmetric.

Important result

- Let $\mathbf{Y}$ be a random vector and let $\mathbf{A}$ be a constant matrix.
- Then $\mathbf{W} = \mathbf{AY}$ is also a random vector with

$$E(\mathbf{W}) = E(\mathbf{AY}) = \mathbf{A}E(\mathbf{Y}),$$

and

$$\sigma^2(\mathbf{W}) = \sigma^2(\mathbf{AY}) = \mathbf{A}\sigma^2(\mathbf{Y})\mathbf{A}^T.$$

Matrix Approach to Regression

- Matrix algebra is commonly used in statistical analysis.
- While it is not required for simple linear regression, it is extremely useful in the *multiple linear regression* setting.
- The simple linear regression we can write our model as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_N \end{bmatrix}}_{=\mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}}_{=\boldsymbol{\varepsilon}}.$$

- It reduces to

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}.$$

Simple Linear Regression Model in Matrix Terms

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

$$\mathbb{E}[\mathbf{Y}] = \mathbf{X}\beta, \quad \mathbb{E}[\varepsilon] = \mathbf{0}, \quad \text{and} \quad \boldsymbol{\sigma}^2\{\varepsilon\} = \sigma^2 \mathbb{I}_{N\times N}.$$

Normal Equations

- Some matrix products: $\mathbf{Y}^T\mathbf{Y} = \sum_{i=1}^N Y_i^2$,

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} N & \sum_{i=1}^N X_i \\ \sum_{i=1}^N X_i & \sum_{i=1}^N X_i^2 \end{bmatrix}, \qquad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^N Y_i \\ \sum_{i=1}^N X_i Y_i \end{bmatrix}.$$

- Therefore, the **normal equations**:

$$Nb_0 + b_1 \sum_{i=1}^N X_i = \sum_{i=1}^N Y_i \quad \text{and} \quad b_0 \sum_{i=1}^N X_i + b_1 \sum_{i=1}^N X_i^2 = \sum_{i=1}^N X_i Y_i,$$

in the matrix terms are $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$, where $\mathbf{b} = [b_0, b_1]^T$.

Simple Linear Regression Model in Matrix Terms

Solving Normal Equations

- Normal Equations:

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}.$$

- Multiply both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$ (we assume that this inverse exists!)

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

- Hence, the least square estimates are given by

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Fitted Values and Hat Matrix

Fitted values in matrix form are given by

$$\widehat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

Since $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, we can write them as

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

or equivalently

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}, \qquad \text{where } \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

- The fitted values $\widehat{\mathbf{Y}}$ can be expressed as linear combinations of the response variable observations $\mathbf{Y}$, with the coefficients being elements of the matrix $\mathbf{H}$.
- The matrix $\mathbf{H}$ involves only the observations on the predictor variable $\mathbf{X}$.
- $\mathbf{H}$ is a square matrix and it is called the *hat matrix*. It is a projection matrix (we will revisit this later) because it is symmetric and idempotent, i.e.,

$$\mathbf{H}\mathbf{H} = \mathbf{H}$$

In matrix notation the vector of residuals is given by

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$$

The variance-covariance matrix of the residuals is given by

$$\sigma^2\{\mathbf{e}\} = \sigma^2 (\mathbf{I} - \mathbf{H})$$

and is estimated by

$$s^2\{\mathbf{e}\} = MSE\,(\mathbf{I} - \mathbf{H})$$

Proof of the variance of the error term

$$\begin{aligned}
\sigma^2\{\mathbf{e}\} &= (\mathbf{I} - \mathbf{H})\,\sigma^2\{\mathbf{Y}\}\,(\mathbf{I} - \mathbf{H})^T \\
&= \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T \\
&= \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) \\
&= \sigma^2 (\mathbf{I} - \mathbf{H}).
\end{aligned}$$

Analysis of Variance

$$SSTO = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N} Y_i^2 - \frac{(\sum_{i=1}^{N} Y_i)^2}{N} = \mathbf{Y}^T\mathbf{Y} - \frac{1}{N}\mathbf{Y}^T\mathbf{J}\mathbf{Y},$$

where $\mathbf{J}$ is a square matrix with all elements 1.

$$\begin{aligned} SSE &= \mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T(\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{b} - \mathbf{b}^T\mathbf{X}^T\mathbf{Y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{b} - \mathbf{b}^T\mathbf{X}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{b} - \mathbf{b}^T\mathbf{X}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{X}\mathbf{b} \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{b}^T\mathbf{X}^T\mathbf{Y} \end{aligned}$$

Finally, since $SSR = SSTO - SSE$, we get

$$SSR = \mathbf{b}^T\mathbf{X}^T\mathbf{Y} - \frac{1}{N}\mathbf{Y}^T\mathbf{J}\mathbf{Y}.$$

In short

$$SSTO = \mathbf{Y}^T\left[\mathbf{I} - \frac{1}{N}\mathbf{J}\right]\mathbf{Y}, \quad SSE = \mathbf{Y}^T[\mathbf{I} - \mathbf{H}]\mathbf{Y}, \quad SSR = \mathbf{Y}^T\left[\mathbf{H} - \frac{1}{N}\mathbf{J}\right]\mathbf{Y}$$

Variance-Covariance Matrix of $\mathbf{b}$

The variance-covariance matrix of $\mathbf{b}$

$$\sigma^2\{\mathbf{b}\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} \end{bmatrix} = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1},$$

where

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} \frac{1}{N} + \frac{\bar{X}^2}{\sum_{i=1}^N (X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum_{i=1}^N (X_i - \bar{X})^2} & \frac{1}{\sum_{i=1}^N (X_i - \bar{X})^2} \end{bmatrix}.$$

When MSE is substituted for σ^2, we obtain the estimated variance-covariance matrix of $\mathbf{b}$.

Mean Response vs. Prediction of New Observation

Fitted value in matrix form is given by

$$\widehat{Y}_h = \mathbf{X}_h^T \mathbf{b}, \text{ where } \mathbf{X}_h = [1 \quad X_h]^T$$

The variance of $\widehat{Y}_h$ in matrix notation is

$$\sigma^2 \left\{ \widehat{Y}_h \right\} = \sigma^2 \mathbf{X}_h^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}_h.$$

The corresponding estimator is given by

$$s^2 \left\{ \widehat{Y}_h \right\} = MSE(\mathbf{X}_h^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}_h).$$

For the prediction of new observation the variance in matrix notation is given by

$$\sigma^2 \{\text{pred}\} = \sigma^2 (1 + \mathbf{X}_h^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}_h).$$

The corresponding estimator is given by

$$s^2 \{\text{pred}\} = MSE(1 + \mathbf{X}_h^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}_h).$$

Multiple regression

- A regression with *two* or *more* explanatory variables is called a *multiple* regression.
- Multiple regression analysis is one of the most widely used of all statistical methods.
- In matrix notation regression models for multiple regression will appear exactly as those for simple linear regression.
- Only the degrees of freedom, constants related to the number of explanatory variables and the dimensions of some variables will be different.

General linear regression

- Suppose we have N observations on $(p-1)$ explanatory variables $X_1, X_2, \ldots, X_{p-1}$ and one response variable Y.

- We can write this as follows:

$$\begin{array}{rl} \text{1st observation:} & (X_{11}, X_{12}, \ldots, X_{1,p-1}, Y_1) \\ \ldots\ldots & \\ i\text{'th observation:} & (X_{i1}, X_{i2}, \ldots, X_{i,p-1}, Y_i) \\ N\text{'th observation:} & (X_{N1}, X_{N2}, \ldots, X_{N,p-1}, Y_N) \end{array}$$

- The *general linear regression* model is:

$$Y_i = \beta_0 + \beta X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

  for $i = 1, \ldots, N$.

- The errors ε_i are independent and follow a $N(0, \sigma^2)$ distribution.

- The model parameters are $\beta_0, \beta_1, \ldots, \beta_{p-1}$ and σ^2.

- The regression function is given by

$$E(Y|X_1, \ldots, X_{p-1}) = \beta_0 + \beta X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}.$$

- This makes up a $(p-1)$-dimensional *regression surface*.

- For more than two variables the surface consists of a *hyper-plane*, and is not possible to visualize.

- The interpretation of the coefficients β_i differ from SLR.

- The multiple linear regression model states that each explanatory variable has a *straight-line relationship* with the mean of Y, given that the other explanatory variables are *fixed*.

- The estimate of β_i represents the effect of the explanatory variable X_i, while *controlling* (fixing the value) effects of all other explanatory variables in the model.