JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLGY

DEPARTMENT OF MATHEMATICS AND ACTUARIAL SCIENCE

BACHELOR OF SCIENCE: BOR 4 AND BBS 4

STA 2408: REGRESSION MODELLING II

BY

THEOPHILUS ASAMOAH

SEPTEMBER, 2024 TO DECEMBER, 2024

SECTION I: REVIEW OF LINEAR MODELS

1.1 The General Form of the Linear Model

- The linear model is the foundation for many predictive modeling techniques.
- For simple linear regression, we have a single independent variable, and the model is of the form: $Y_i = \beta_0 + X_1\beta_{1i} + \varepsilon_i$ (1)

Where:

- Y_i is the i^{th} observation of the dependent variable.
- X_{i1} is the independent variable.
- β_0 is the intercept term or constant
- ε_i the random error for the i^{th} observation

1.1 The General Form of the Linear Model Cont'd

- For the general form (multiple linear regression), where we have multiple independent variables, the model is of the form:

$$Y_i = \beta_0 + X_1\beta_{1i} + X_2\beta_{2i} + \cdots + X_{ki}\beta_{ki} + \varepsilon_i \qquad (2)$$

Where:

- Y_i is the i^{th} observation of the dependent variable.
- $X_{i1}, X_{i2}, \cdots X_{ik}$ are the independent variables.
- β_0 is the intercept term or constant
- ε_i the random error for the i^{th} observation.

1.1 The General Form of the Linear Model Cont'd

- In matrix form, the general form of the regression model is;

$$Y = X\beta + \varepsilon \tag{3}$$

Where:

- Y is the response vector (dependent variable).
- X is the matrix of predictor variables (independent variables).
- β is the vector of unknown parameters (coefficients).
- ε is the error term, assumed to follow a normal distribution with mean, 0 and constant variance, σ^2.

1.2 Assumptions of the Linear Models

- The following assumptions are critical for the validity of a linear model;
  - ✓ ***Linearity:*** The relationship between the dependent variable and independent variables is linear. That is, $E(Y \mid X) = X\beta$ (4)
  - ✓ ***Independence:*** Observations are independent of each other.
  - ✓ ***Homoscedasticity:*** The variance of the errors is constant across all levels of the independent variables. That is, $Var(\varepsilon_i) = \sigma^2$ for all i. (5)

1.2 Assumptions of the Linear Models Cont’d

- ✓ ***Normality:*** The residuals (errors) are normally distributed with mean zero. That is,

$$\varepsilon \sim N\left(0,\ \sigma^2\right) \tag{6}$$

- ✓ ***No Multicollinearity:*** Independent variables are not highly correlated with each other. This means the matrix $X'X$ is invertible.

- ✓ X ***is of full rank:*** Full rank means that the columns of X (the independent variables) are linearly independent, and thus the rank of p is equal to the number of predictors.

1.2 Assumptions of the Linear Models Cont’d

- If these assumptions are violated, the estimates obtained from the model might be biased or inefficient.

- NB: What does it means for X to be of **Full Rank**, and why?

1.3. Estimation Procedures of the Linear Model

(The Matrix Approach)

- To estimate the parameters of the linear model, we use the Ordinary Least Squares (OLS) method.

- The goal of OLS is to minimize the sum of squared residuals (differences between the observed and predicted values).

- The matrix form of the linear model is:

$$Y = X\beta + \varepsilon \tag{7}$$

1.3. Estimation Procedures of the Linear Model (The Matrix Approach) Cont'd

- The OLS estimator for β is given by:

$$\hat{\beta} = \left(X'X\right)^{-1} + X'Y \qquad (8)$$

- Where:
  - ✓ X' is the transpose of matrix X
  - ✓ $\left(X'X\right)^{-1}$ is the inverse of the matrix $\left(X'X\right)$
  - ✓ $\hat{\beta}$ is the vector of estimated coefficients.

1.3. Estimation Procedures of the Linear Model (The Matrix Approach) Cont'd

- This OLS leads to the following optimization problem;

$$Q(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \qquad (9)$$

- The least squares estimate, $\hat{\beta}$ of β, is the solution of β in the equation,

$$\frac{dQ}{d\beta} = 0 \qquad (10)$$

Now, since $(\beta' X' Y)$ is a 1×1 matrix or scalar, and it transpose,

$$(\beta' X' Y)' = Y' X \beta \qquad (11)$$

is the same scalar, $Q(\beta)$ can be expressed as;

1.3. Estimation Procedures of the Linear Model (The Matrix Approach) Cont'd

$$\begin{aligned} Q(\beta) &= Y'Y - \beta' X'Y - Y'X\beta + \beta' X'X\beta \\ &= Y'Y - 2Y'X\beta + \beta' X'X\beta \end{aligned} \qquad (12)$$

- The least squares estimator of β must satisfy $\frac{dQ}{d\beta} = 0$ (13)
- To solve, we differentiate the objective function with respect to β and set the derivative to zero, and we have;

$$\frac{dQ}{d\beta}\Big|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} \qquad (14)$$

1.3. Estimation Procedures of the Linear Model (The Matrix Approach) Cont'd

- Now, setting this equal to zero gives the **Normal Equations**:

$$X'X\hat{\beta} = X'Y \qquad (15)$$

- Solving for $\hat{\beta}$, we have: $\hat{\beta} = \left[(X'X)^{-1}X'Y\right]$ (16)

Where:

- $\hat{\beta}$ is the p×1 vector of estimated coefficients.
- $(X'X)^{-1}$ is the inverse of the matrix $X'X$, assuming $X'X$ is non-singular.

1.4. Estimation of the Parameters with a Given Dataset

(The Matrix Approach)

- **Example 1:** The following data relate to the prices *(Y)* of five randomly chosen houses in a certain neighborhood, the corresponding ages of the houses *(x1)*, and square footage *(x2)*.

Price (*Y*) in thousands in of dollars	Age (*X*1) years	Square footage *X*2 in thousands of square feet
100	1	1
80	5	1
104	5	2
94	10	2
130	20	3

- Fit a multiple linear regression model, $Y_i = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon_i$ to the foregoing data

1.4.1 Matrix Representation of the Given Data

- The response vector ***Y***, the design matrix ***X***, and the coefficient vector ***β*** are represented as:

$$Y_i = \begin{bmatrix} 100 \\ 80 \\ 104 \\ 94 \\ 130 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 10 & 2 \\ 1 & 20 & 3 \end{bmatrix}, \text{ and } \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

- Here, the first column of ***X*** corresponds to the intercept term.

1.4.2 Estimate the Coefficients

- The coefficients $\boldsymbol{\beta}$ are estimated using the formula in (16):
- First, we compute $X'X$ and $X'Y$ given as:

$$X'X = \begin{bmatrix} 5 & 41 & 9 \\ 41 & 551 & 96 \\ 9 & 96 & 19 \end{bmatrix} \text{ and } X'Y = \begin{bmatrix} 508 \\ 4560 \\ 966 \end{bmatrix}$$

- Next, we find the inverse of $X'X$ given as:

1.4.2 Estimate the Coefficients Cont’d

$$(X'X)^{-1} = \begin{bmatrix} 2.3076 & 0.1565 & -1.8840 \\ 0.1565 & 0.0258 & -0.2044 \\ -1.8840 & -0.2044 & 1.9779 \end{bmatrix}$$

- Finally, we calculate $\hat{\beta}$ and have;

$$(X'X)^{-1}(X'Y) = \begin{bmatrix} 66.1252 \\ -0.3794 \\ 21.4365 \end{bmatrix}$$

1.4.2 Estimate the Coefficients Cont’d

- Thus, the estimated coefficients are;

$$\beta_0 = 66.1252;$$

$$\beta_1 = -0.3794$$

$$\beta_2 = 21.4365.$$

- The regression model is: $y = 66.1252 - 0.3794x_1 + 21.4365x_2$

1.5 Goodness-of-Fit of Linear Regression Models

- **1.5.1 ANOVA Output**

Source	DF	SS	MS	F-Value	P-Value
Regression	2	956.50	478.248	2.50	0.286
Error	2	382.70	191.352		
Total	**4**	**1339.20**			

- **1.5.2 Model Summary (R^2)**

S	R-squared	R-squared(adj)
13.8330	71.42%	42.85%

1.6 Practical Example in R

- **# Define the response vector Y**
  - ✓ Y <- matrix(c(100, 80, 104, 94, 130), nrow = 5, byrow = TRUE)

- **# Define the predictor vectors X1 and X2**
  - ✓ X1 <- matrix(c(1, 5, 5, 0, 20), nrow = 5, byrow = TRUE)
  - ✓ X2 <- matrix(c(1, 1, 2, 2, 3), nrow = 5, byrow = TRUE)

- **# Define the design matrix X**
  - ✓ X <- matrix(c(1, 1, 1, 1, 1, 1, 5, 5, 0, 20, 1, 1, 2, 2, 3), nrow = 5, byrow = TRUE).

1.6 Practical Example in R Cont’d

- **# Define the parameters vector *B***
  - ✓ β <- matrix(c(=β_0, β_1, β_2), nrow = 3, byrow = TRUE)

- **# Estimate the coefficients using the matrix formula**
  - ✓ beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y

- **# Output the estimated coefficients**
  - ✓ print(beta_hat)

1.7 Conclusions

- This review of linear models sets the foundation for moving into non-linear regression.
- We started by understanding the general form and assumptions of linear models, followed by estimation techniques using matrix algebra, and a practical example in R.
- In non-linear regression, we will relax the linearity assumption and explore models where the relationship between dependent and independent variables is non-linear, often using iterative estimation procedures instead of closed-form solutions like OLS.

1.8. Assignment I

- Given the data:

x_1	x_2	y
3	1	4
2	5	3
3	3	6
1	2	4

1.8. Assignment I Cont’d

- (a) Write the multiple regression model in matrix form.
- (b) Find: $X'X$, $(X'X)^{-1}$, and $X'Y$
- (c) Estimate $\hat{\beta}$.
- (d) Test the statistical significance of the model.
- (e) Check if the linear model assumptions discussed in class hold for the given data.

Questions/Comments/Suggestions

THANK YOU