Linear Regression Models

Segment 2 – Multiple Linear Regression Model

Topic 1 – Matrix Approach for Multiple Linear Regression Model

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. Matrix Notations for Data: Design Matrix
2. Dealing with Categorical Covariates
3. Multiple Linear Regression Models (MLRM) and assumptions
4. Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof
5. Residual Vector and its Properties

Matrix Notations for Data: Design Matrix

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example:

Matrix Notations for Data: Design Matrix

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*):

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$,

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f:

Matrix Notations for Data: Design Matrix

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \dots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \dots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
- To that end, we collect *sample* data from the *population* and use the following notation:

Matrix Notations for Data: Design Matrix

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
- To that end, we collect *sample* data from the *population* and use the following notation:

$y^{(i)} =$ ith sample's response value,

Matrix Notations for Data: Design Matrix

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Suppose Y is the response variable and we are interested in studying its relationship with multiple predictors $X_1, X_2, \ldots, X_p$.
- Example: Let fuel efficiency (mpg) be the response variable and horse power, weight, and transmission type (automatic or manual) be the predictors.
- In a multiple linear regression model (MLRM), the true relationship (*real population model*): $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$, for an unknown nonlinear function f is modeled using a *linear approximation* of the function f: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$.
- To that end, we collect *sample* data from the *population* and use the following notation:

$y^{(i)}$ = ith sample's response value, $x_j^{(i)}$ = ith sample's jth predictor value.

Matrix Notations for Data: Design Matrix

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix}$$

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(n)} - \hat{y}^{(n)} \end{bmatrix}$$

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(n)} - \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 x_1^{(1)} + \cdots + \beta_p x_p^{(1)} \\ \vdots \\ \beta_0 + \beta_1 x_1^{(n)} + \cdots + \beta_p x_p^{(n)} \end{bmatrix}$$

Matrix Notations for Data: Design Matrix

- The MLRM model predicts Y as an approximation $\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
- Residual $R = Y - \hat{Y} = Y - (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$.
- After sampling data, we have

$$\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ \vdots \\ y^{(n)} - \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 x_1^{(1)} + \cdots + \beta_p x_p^{(1)} \\ \vdots \\ \beta_0 + \beta_1 x_1^{(n)} + \cdots + \beta_p x_p^{(n)} \end{bmatrix}$$

$$= \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} - \begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_p^{(1)} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_1^{(n)} & \ldots & x_p^{(n)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Matrix Notations for Data: Design Matrix

Matrix Notations for Data: Design Matrix

$$\underbrace{\begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(n)} \end{bmatrix}}_{\text{residual vector } \mathbf{r}} = \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\text{true response vector } \mathbf{y}} - \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}}_{\text{design matrix } \mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}}_{\text{unknown coefficients vector } \boldsymbol{\beta}}$$

Matrix Notations for Data: Design Matrix

$$\underbrace{\begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(n)} \end{bmatrix}}_{\text{residual vector } \mathbf{r}} = \underbrace{\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\text{true response vector } \mathbf{y}} - \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{bmatrix}}_{\text{design matrix } \mathbf{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}}_{\text{unknown coefficients vector } \boldsymbol{\beta}}$$

$$\Rightarrow \mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

Dealing with Categorical Covariates

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values:

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced:

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced: (1) `heatinghot air`, (2) `heatinghot water/steam`.

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced: (1) `heatinghot air`, (2) `heatinghot water/steam`.
- The dummy encoding for building the model is as follows:

Dealing with Categorical Covariates

- Suppose we have a categorical predictor `heating` which can take three possible values: (1) `electric` (2) `hot water/steam` (3) `hot air`.
- Based on alphabetical order, the categorical level `electric` is chosen as the *reference*.
- Two new dummy predictors are introduced: (1) `heatinghot air`, (2) `heatinghot water/steam`.
- The dummy encoding for building the model is as follows:

	heatinghot air	heatinghot water/steam
electric	0	0
hot air	1	0
hot water/steam	0	1

Multiple Linear Regression Models (MLRM) and assumptions

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

Multiple Linear Regression Models (MLRM) and assumptions

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The random errors for yet to be decided samples $i = 1, 2, \dots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:

Multiple Linear Regression Models (MLRM) and assumptions

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean:

Multiple Linear Regression Models (MLRM) and assumptions

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM
$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$
can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.
- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: Cov$(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$

Multiple Linear Regression Models (MLRM) and assumptions

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
    1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
    2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$
    3. the design matrix has full rank:

Multiple Linear Regression Models (MLRM) and assumptions

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$.

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$
  3. the design matrix has full rank: $\text{rank}(\mathbf{X}) = p + 1$

Multiple Linear Regression Models (MLRM) and assumptions

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}.$

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$
  3. the design matrix has full rank: $\text{rank}(\mathbf{X}) = p + 1$
  4. the random error vector is (multivariate) normally distributed:

Multiple Linear Regression Models (MLRM) and assumptions

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The random errors for yet to be decided samples $i = 1, 2, \ldots, n$ in the MLRM

$$Y^{(i)} = \beta_0 + \beta_1 X_1^{(i)} + \beta_2 X_2^{(i)} + \cdots + \beta_p X_p^{(i)} + \epsilon^{(i)}$$

can be put into a random vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}.$

- For drawing statistical inferences about the coefficients estimates, we will assume that:
  1. random error has zero mean: $E[\boldsymbol{\epsilon}] = \mathbf{0}$
  2. random errors across samples are uncorrelated with constant variance: $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$
  3. the design matrix has full rank: $\text{rank}(\mathbf{X}) = p + 1$
  4. the random error vector is (multivariate) normally distributed: $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates,

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset:

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\|\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix}\right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to minimizing $\|\mathbf{r}\|^2$ = minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2.$
- Note that $\sum_{i=1}^{n} \left(r^{(i)}\right)^2 = \underbrace{\left\| \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix} \right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2.$
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to minimizing $\|\mathbf{r}\|^2$ = minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$.
- The resulting solution is the OLS solution: $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

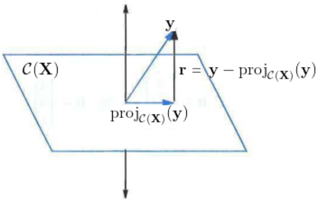
MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- To find the coefficient estimates, just as in SLRM, we minimize the the sum of the squares of the residuals (RSS) for all samples in the dataset: $\min \sum_{i=1}^{n}\left(r^{(i)}\right)^2 = \sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2$.
- Note that $\sum_{i=1}^{n}\left(r^{(i)}\right)^2 = \underbrace{\left\|\begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(n)} \end{bmatrix}\right\|^2}_{\text{norm of vector squared}} = \|\mathbf{r}\|^2$.
- Using the equation $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$, minimizing the RSS corresponds to minimizing $\|\mathbf{r}\|^2$ = minimizing $\|\mathbf{y} - \mathbf{X}\beta\|^2$.
- The resulting solution is the OLS solution: $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Full rank of the design matrix $\mathbf{X}$ ensures the existence of $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Minimizing $\|\mathbf{r}\|^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2 = \left\| \mathbf{y} - \left(\underbrace{\beta_0 \mathbf{x}_1 + \beta_1 \mathbf{x}_2 + \cdots + \beta_p \mathbf{x}_{p+1}}_{\text{linear combination of columns of } \mathbf{X}} \right) \right\|^2$ is equivalent to solving the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ which represents the *orthogonal projection* of $\mathbf{y}$ on to the column space of the design matrix $\mathcal{C}(\mathbf{X})$ (set of all possible linear combinations of the columns of $\mathbf{X}$):



Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}.$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{z}) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{z}) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{z}) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as
  $\mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{z}\right) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as $\mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{X}\left(\hat{\beta} - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right) = \mathbf{0}$.

Ordinary Least Squares (OLS) Solution: Intuition, Geometry, & Algebraic Proof

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- Let $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} \Rightarrow \mathbf{r} = \mathbf{y} - \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{y} - \mathbf{X}\mathbf{z}$.
- Residual vector orthogonal to the column space of $\mathbf{X} \Rightarrow \mathbf{r} \perp \mathcal{C}(\mathbf{X})$ $\Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{r} = \mathbf{0}$.
- This implies $\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{z}) = \mathbf{0} \Rightarrow \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{z} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{z} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- This leads to $\text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y}) = \mathbf{X}\mathbf{z} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.
- Therefore, the equation $\mathbf{X}\hat{\beta} = \text{proj}_{\mathcal{C}(\mathbf{X})}(\mathbf{y})$ can be written as $\mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \Rightarrow \mathbf{X}\left(\hat{\beta} - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right) = \mathbf{0}$.
- Using the fact that the design matrix $\mathbf{X}$ has full rank (that is, its columns are linearly independent), we arrive at the unique OLS solution $\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Residual Vector and its Properties

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector $\mathbf{1}$.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n} \left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$

Residual Vector and its Properties

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n}\left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}^{(i)}$.

Residual Vector and its Properties

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n}\left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}^{(i)}$.
- This means the true and fitted response values always have the same sample mean.

Residual Vector and its Properties

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- By construction, the residual vector $\mathbf{r}$ is orthogonal to the columns of the design matrix $\mathbf{X}$.
- In particular, the residual vector is orthogonal to the first column of $\mathbf{X}$ which is the column full of ones or the ones-vector **1**.
- This implies that $\mathbf{1}^{\mathrm{T}}\mathbf{r} = 0$ which leads to the fact that sum of the residuals is always equal to 0.
- This further implies that $\sum_{i=1}^{n}\left[\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)}\right] = 0$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{y}}^{(i)}$.
- This means the true and fitted response values always have the same sample mean.
- This is a reiteration of the fact that linear regression works best on an average.