Multiple Regression 1

AU STAT-615

Emil Hvitfeldt

2021-03-03

Example 1

First order model with two predictor variables

When there are two variables X_1 and X_2, the regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Assuming $E\{\varepsilon_i\} = 0$ we have

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

The response function is a plane

Example 1

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

- β_0, y-intercept. If $X_1 = X_2 = 0$ then β_0 represents the mean response $E\{Y\}$
- β_1 Indicates change in mean per unit increase in X_1 when X_2 is constant
- β_2 Indicates change in mean per unit increase in X_2 when X_1 is constant

Example 2

First order model with more than two predictor variables

For $p - 1$ predictor variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Or

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{l1} + \varepsilon_i$$

Example 2

Assuming $E\{\varepsilon_i\} = 0$ we obtain

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

Here the response function is a hyperplane

Qualitative predictor variables

For the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

This model encompasses not only quantitative predictor variables but also qualitative ones such as sex or disability status

For example, let

$$\begin{aligned} X_1 &= \text{Age of patients} \\ X_2 &= \begin{cases} 1, & \text{patient female} \\ 0, & \text{patient male} \end{cases} \\ Y &= \text{Length of hospital stay} \end{aligned}$$

Qualitative predictor variables

We have

$$E\{Y\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_2$$

and for male patients

$$E\{Y\} = \beta_0 + \beta_1 X_1$$

and for female patients

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 X_1$$

These two response functions are straight lines that are parallel with each other

Polynomial Regression

Special case of general linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

More on Chapter 8

interaction Effects

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

The effect of one predictor variable depends on the levels of the other predictor variables

Meaning of Linear in General Linear Regression Model

We say that a regression model is linear in the parameters when it can be written in the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

The term linear model refers to the fact that the equation is linear in parameters, it does not refer to the shape of the response variable

An example of a non-linear regression model

$$Y_i = \beta_0 \cdot e^{\beta_1 X_i} + \varepsilon_i$$

General Linear Regression model in matrix form

The model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Can be written using matrices as

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p} \cdot \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

General Linear Regression model in matrix form

Where

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X}_{n\times P} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

General Linear Regression model in matrix form

For

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p} \cdot \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

- $\mathbf{Y}_{n\times 1}$, vector of responses
- $\mathbf{X}_{n\times p}$, Matrix of constants
- $\boldsymbol{\beta}_{p\times 1}$, vector of parameters
- $\boldsymbol{\varepsilon}_{n\times 1}$, vector of independent normal random variables

Properties

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p} \cdot \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

We have that

$$E\{\boldsymbol{\varepsilon}\} = \mathbf{0}$$

and

$$V\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} \sigma_2 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_2 \end{bmatrix} = \sigma_2 \mathbf{I}$$

Thus $E\{\mathbf{Y}\}_{n\times 1} = \mathbf{X}\boldsymbol{\beta}$ and $V\{\mathbf{Y}\}_{n\times n} = \sigma_2 \mathbf{I}_{n\times n}$

Estimation of Regression Coefficients

In the general linear case, we have the following criterion

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

The vector of least squares estimated coefficients $b_0, b_1, \ldots, b_{p-1}$ is denoted by

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

Estimation of Regression Coefficients

The least squares normal equations for the general linear regression model are given by

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

and the least square estimators are

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

Fitted values & Residuals

let $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$ and $e_i = Y_i - \hat{Y}_i$ is written as $\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$

The fitted values are represented by

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

and

$$\mathbf{e}_{n\times 1} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

Fitted values & Residuals

We know that $\mathbf{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y}$ so get get that

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$
$$\hat{\mathbf{Y}} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y}$$
$$\hat{\mathbf{Y}} = \mathbf{H} \cdot \mathbf{Y}$$

where we substitute $\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$

Fitted values & Residuals

Therefore $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H}) \cdot \mathbf{Y}$

and the variance-covariance is

$$V\{\boldsymbol{\varepsilon}\} = \sigma^2 \cdot (\mathbf{I} - \mathbf{H})$$

and

$$s^2\{\boldsymbol{\varepsilon}\} = MSE(\mathbf{I} - \mathbf{H})$$

Analysis of Variancee

		df	MS
regression	SSR	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	SSE	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	SSTO	$n-1$	

Analysis of Variancee

Where

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

We have that $\mathbf{b}^T \cdot \mathbf{X}^T = \mathbf{Y}^T$ so we get

$$SSR = \mathbf{Y}^T \cdot \mathbf{Y} - \frac{1}{n}\mathbf{Y}^T \cdot \mathbf{J} \cdot \mathbf{Y}$$

where $\mathbf{J}_{n\times n}$ of all 1s

Analysis of Variancee

And we have that

$$SSE = \mathbf{e}^T \cdot \mathbf{e} = \cdots = \mathbf{Y}^T \cdot \mathbf{Y} - \mathbf{b}^T \mathbf{X}^T \cdot \mathbf{Y}$$

The expectation of MSE is σ^2 as for simple linear regression

The expectation of MSR is σ^2 plus a quantity that is non-negative

F Test for Regressioon Relation

To test whether there is a regression relation between Y and a set of variables $X_1, \ldots, X_{p-1}$ we have

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$
$$h_1 : \text{not all } \beta_k (k = 1, \ldots, p-1) \text{ equal zero}$$

We have $F^* = \dfrac{MSR}{MSE}$

The decision rule to control type 1 error at α is

$$\text{If } F^* \leq F(1-\alpha; p-1, n-p) \text{ conclude } H_0$$

$$\text{If } F^* > F(1-\alpha; p-1, n-p) \text{ conclude } H_1$$

Coefficients of multiple determination

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Measures the proportionate reduction of total variation in Y associated with the use of the set of X variables $X_1, \ldots, X_{p-1}$

Coefficients of multiple determination

Since adding more variables to the regression model can only increase R^2 and never reduce it because SSE can never become larger with more X variables and SSTO is always the same for a given set of responses

So we can use another metric, adjusted coefficient of multiple determination

$$R_\alpha^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right) \cdot \frac{SSE}{SSTO}$$

Note: A larger value of R^2 does not necessarily imply that the fitted model is a useful one.