Multiple Linear Regression

- features/predictors: $X_1, \ldots, X_p$
- response/outcome variable: Y

The linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

where

β_0 is the intercept

β_j is the regression coefficient associated with X_j

e is the error term often assumed to have mean zero and variance σ^2.

Housing Data

Y: sale price of a house

X_1: # of bedrooms

X_2: # of bathrooms

X_3: square feet

......

Multiple Linear Regression

- features/predictors: $X_1, \ldots, X_p$
- response/outcome variable: Y

The linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

where

β_0 is the intercept

β_j is the regression coefficient associated with X_j

e is the error term often assumed to have mean zero

and variance σ^2.

Housing Data

Y: sale price of a house

X_1: # of bedrooms

X_2: # of bathrooms

X_3: square feet

......

Multiple Linear Regression

- features/predictors: $X_1, \ldots, X_p$
- response/outcome variable: Y

The linear regression model assumes

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

where

β_0 is the intercept

β_j is the regression coefficient associated with X_j

e is the error term often assumed to have mean zero and variance σ^2.

Housing Data

Y: sale price of a house

X_1: # of bedrooms

X_2: # of bathrooms

X_3: square feet

$\ldots\ldots$

Training Data $(x_{i1}, \ldots, x_{ip}, y_i)_{i=1}^n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i$$

$$i = 1, \ldots, n$$

Matrix Representation

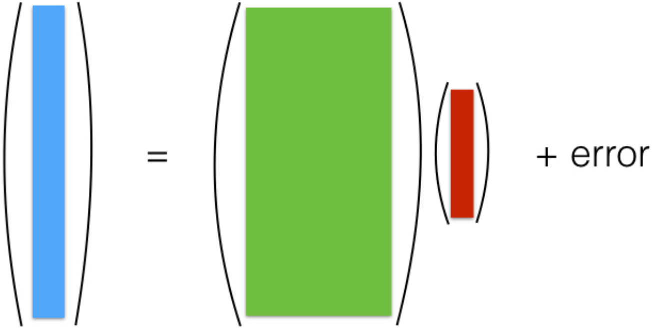
Express the regression model on $(x_{i1}, \ldots, x_{ip}, y_i)_{i=1}^n$ in the following matrix form

$$\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + e_1 \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + e_2 \\ \ldots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + e_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ 1 & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}$$

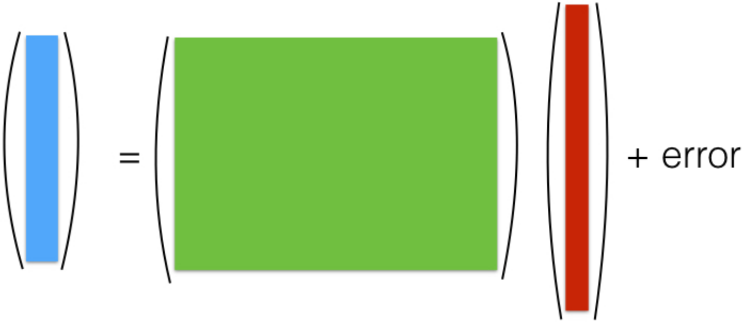
$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \mathbf{e}_{n\times 1}$$

The classical large n small p regression model:



Focus of **this** week

The modern large p small n regression model:

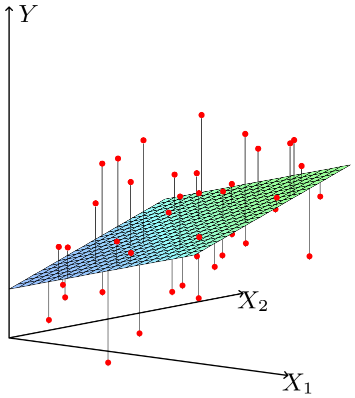


Focus of **next** week

Least Squares Estimation

Given a set of training data $(x_{i1}, \ldots, x_{ip}, y_i)_{i=1}^n$, we estimate the regression coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ by minimizing the residual sum of squares (RSS)

$$\begin{aligned} & \text{RSS}(\beta_0, \beta_1, \cdots, \beta_p) \\ = \; & \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2. \end{aligned}$$



Least Squares Estimation: Continued I

Using matrix representation, we can express the regression model as

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \mathbf{e}_{n\times 1}.$$

The least squares method estimates $\boldsymbol{\beta}$ by minimizing

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p\right)^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2. \end{aligned}$$

Least Squares Estimation: Continued II

Differentiating RSS($\boldsymbol{\beta}$) with respect to $\boldsymbol{\beta}$ and setting to zero, we have

$$\begin{aligned}
\frac{\partial \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{\partial \boldsymbol{\beta}} &= \mathbf{0}_{(p+1)\times 1} = -2\mathbf{X}^t_{(p+1)\times n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})_{n\times 1} \\
&\Longrightarrow \mathbf{X}^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \text{normal equation} \\
&\Longrightarrow (\mathbf{X}^t\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y} \\
&\Longrightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}
\end{aligned}$$

Here we assume the rank of $\mathbf{X}$ is $(p+1)$ and then the inverse of the $(p+1)\times(p+1)$ matrix $(\mathbf{X}^t\mathbf{X})$ exists.

Least Squares Estimation: Continued II

Differentiating RSS($\boldsymbol{\beta}$) with respect to $\boldsymbol{\beta}$ and setting to zero, we have

$$
\begin{aligned}
\frac{\partial \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{\partial \boldsymbol{\beta}} &= \mathbf{0}_{(p+1)\times 1} = -2\mathbf{X}^t_{(p+1)\times n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})_{n\times 1} \\
&\Longrightarrow \mathbf{X}^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \text{normal equation} \\
&\Longrightarrow (\mathbf{X}^t\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y} \\
&\Longrightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}
\end{aligned}
$$

Here we assume the rank of $\mathbf{X}$ is $(p+1)$ and then the inverse of the $(p+1) \times (p+1)$ matrix $(\mathbf{X}^t\mathbf{X})$ exists. What if rank$(X) < (p+1)$? Not a serious issue.

Some LS Outputs

Prediction at a new point $\mathbf{x}^*$

$$\hat{y}^* = \hat{\beta}_0 + x^*_{i1}\hat{\beta}_1 + \cdots + x^*_{ip}\hat{\beta}_p.$$

Fitted value at $\mathbf{x}_i$:

$$\hat{y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \cdots + x_{ip}\hat{\beta}_p.$$

Residual at $\mathbf{x}_i$: $r_i = y_i - \hat{y}_i$.

RSS $= \sum_{i=1}^n r_i^2$.

The error variance is estimated by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1} = \frac{\sum_{i=1}^n r_i^2}{n-p-1}$$

The degree of freedom (df) of the residuals is $n-(p+1)$. In general

$$\begin{aligned} df(\text{residuals}) &= (\text{sample-size}) \\ &\quad -(\text{number-of-linear-coefs}) \end{aligned}$$

The Residual Vector

$\mathbf{X}^t\mathbf{r} = \mathbf{0}_{(p+1)\times 1}$ implies that the residual vector $\mathbf{r}$ is subject to $(p+1)$ equality constraints, therefore it loses $(p+1)$ degrees of freedom.