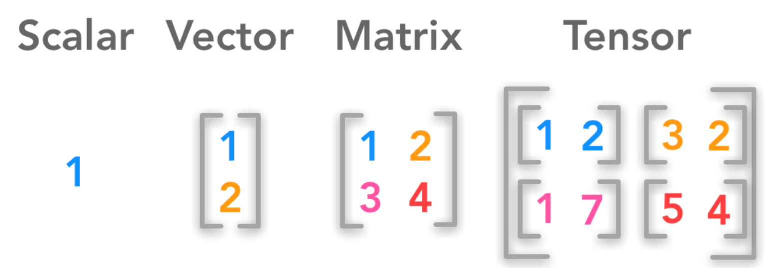
Linear Regression in Matrix Form



The SLR Model in Scalar Form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim^{iid} N(0, \sigma^2)$$

Consider now writing an equation for each observation:

$$\begin{array}{rcl} Y_1 & = & \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 & = & \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \vdots & \vdots \\ Y_n & = & \beta_0 + \beta_1 X_n + \epsilon_n \end{array}$$

The SLR Model in Matrix Form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- $\mathbf{X}$ is called the design matrix.
- β is the vector of parameters
- ϵ is the error vector
- $\mathbf{Y}$ is the response vector

The Design Matrix

$$\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

Vector of Parameters

$$\beta_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vector of Error Terms

$$\epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vector of Responses

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Thus,

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

Least squares estimates in matrix notation

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

where:

- $(X'X)^{-1}$ is the **inverse** of the $X'X$ matrix, and
- X' is the **transpose** of the X matrix.

An example

Let's consider the data in soapsuds.txt, in which the height of suds (y = *suds*) in a standard dishpan was recorded for various amounts of soap (x = *soap*, in grams) (Draper and Smith, 1998, p. 108). Using statistical software to fit the simple linear regression model to these data, we obtain:

```
Regression Equation

suds = -2.68 + 9.500 soap
```

Let's see if we can obtain the same answer using the above matrix formula. We previously showed that:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

We can easily calculate some parts of this formula:

x_i soap	y_i suds	$x_i \times y_i$ so*su	x_i^2 soap²
4.0	33	132.0	16.00
4.5	42	189.0	20.25
5.0	45	225.0	25.00
5.5	51	280.5	30.25
6.0	53	318.0	36.00
6.5	61	396.5	42.25
7.0	62	434.0	49.00
---	---	-----	-----
38.5	**347**	**1975.0**	**218.75**

That is, the 2×2 matrix ***X'X*** is:

$$X'X = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.75 \end{bmatrix}$$

And, the 2×1 column vector ***X'Y*** is:

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} 347 \\ 1975 \end{bmatrix}$$

So, we've determined ***X'X*** and ***X'Y***. Now, all we need to do is to find the inverse $(X'X)^{-1}$. As mentioned before, it is very messy to determine inverses by hand. Letting computer software do the dirty work for us, it can be shown that the inverse of ***X'X*** is:

$$(X'X)^{-1} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix}$$

And so, putting all of our work together, we obtain the least squares estimates:

$$b = (X'X)^{-1}X'Y = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix} \begin{bmatrix} 347 \\ 1975 \end{bmatrix} = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

That is, the estimated intercept is b_0 = -2.67 and the estimated slope is b_1 = 9.51.
estimates are the same as those reported above (within rounding error)!

Multiple Regress

Data for Multiple Regression

- Y_i is the response variable (as usual)
- $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ are the $p-1$ explanatory variables for cases $i = 1$ to n.
- Example – In Homework #1 you considered modeling GPA as a function of entrance exam score. But we could also consider intelligence test scores and high school GPA as potential predictors. This would be 3 variables, so $p = 4$.
- *Potential problem to remember!!!* These predictor variables are likely to be themselves correlated. We always want to be careful of using variables that are themselves strongly correlated as predictors together in the same model.

The Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i \text{ for } i = 1, 2, \ldots, n$$

where

- Y_i is the value of the response variable for the ith case.
- $\epsilon_i \sim^{iid} N(0, \sigma^2)$ (exactly as before!)
- β_0 is the intercept (think multidimensionally).
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables.
- $X_{i,k}$ is the value of the kth explanatory variable for the ith case.
- Parameters as usual include all of the β's as well as σ^2. These need to be estimated from the data.

Model in Matrix Form

$$\begin{aligned}\mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \\ \epsilon &\sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n\times n}) \\ \mathbf{Y} &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})\end{aligned}$$

Design Matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Coefficient matrix β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Interesting Special Cases

- Polynomial model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_{p-1} X_i^{p-1} + \epsilon_i$$