# CISC 271 Class 9

## Design Matrix and Standardized Data

Text Correspondence: $\sim$

*Main Concepts:*

- *Design matrix: gathering data vectors*
- *Zero-mean data: difference from mean*
- *Standard score: multiple of standard deviation*
- *Standardized data: standard score of zero-mean data*

**Sample Problem, Data Analysis:** What is an effective way to prepare a set of data for linear analysis?

One guiding principle of this course can be caricatured as "see a set of vectors, gather them into a matrix". A caution is that it is often unclear whether an information source, such as a scientific article or a fragment of code, is using the data as column vectors or as rows in a matrix. We must always read a source carefully to determine how the data is represented. For example, the textbook in this course – which is highly regarded – uses one convention whereas `MATLAB` – which is also highly regarded – uses a different convention.

Here, we will adopt the `MATLAB` convention. We will interpret the values of independent variables or dependent variables as *columns*, so each column has the same physical dimensions or other semantics. We will interpret a set of such variables as *observations*, so an observation or a "reading" is a matrix row that can have different semantics for each entry.

For example, historical data from Statistics Canada are available that describe basic body measurements such as: gender, weight, height, the ratio of the waist circumference to the hip circumference, and the sum of the measurements of flaps of skin. These measurements are available for multiple years and the age group of the subjects is also known. We might want to interpret these *anthropomorphic* measurements as independent variables, and interpret the age groups and the survey year as dependent variables. To reduce the problem to a single dependent variable, we might try to analyze each year separately. The dependent data could be given numerical labels such as those in Table 9.1.

For the first dependent value, which is ages $6 - 11$, the actual mean values are provided in Table 9.2.

**Table 9.1:** Numerical codes for the dependent variable of age group.

| Ages | Code |
|---|---|
| $6-11$ | 1 |
| $12-19$ | 2 |
| $20-39$ | 3 |
| $40-59$ | 4 |
| $60-79$ | 5 |

**Table 9.2:** Values of independent variables for Age Group #1. These are statistical means of a sample population that was measured in 2009.

| Variable | Value |
|---|---|
| Gender | Female |
| Mass | 33.15 kg |
| Height | 135.28 cm |
| Hip circ. | 73.3 cm |
| Waist/hip | 0.82 |
| Net skinfold | 52.8 mm |

In statistics, data analytics, and machine learning, the values of the independent variables are gathered into a data vector that we will write as $\vec{x}_i$. For example, we might write the values in Table 9.2 as the data vector

$$\vec{x}_1 = \begin{bmatrix} 33.15 \\ 135.28 \\ 73.3 \\ 0.82 \\ 52.8 \end{bmatrix}$$

There are two immediate problems that we have: how do we gather multiple data vectors into a matrix, and how can we manage the disparate ranges of the values of the data.

## 9.1 Data Matrix and Design Matrix

One common method of linear data analysis of a set of vectors is to use *regression*. Regression is a term from statistics for estimating relationship between a number of independent variables and, for this course, a single dependent variable. One such might be to try to find trends in a population.

The approximation problem for linear regression is: given a set of $m$ independent observations $t_k$, and $m$ dependent data $c_i$, to find a weight vector $\vec{w}$ that approximates the dependent data values as

$$t_1 w_1 + t_2 w_2 + \cdots + t_n w_n \approx c_i$$

We can "vectorize" this approximation by gathering the independent observations $t_k$ into a *data matrix* $A$, and gathering the dependent data $c_i$ into a data vector $\vec{c}$. When we create the data matrix $A$, we have two choices: a data vector $\vec{a}_i$ can be the $i^{\text{th}}$ column of $A$, or a transpose as $\vec{t}_i^{\,T}$ can be the $i^{\text{th}}$ row of $A$. We will follow the convention in statistics and in `MATLAB` that uses the second version, so we will define the design matrix as

$$A \stackrel{\text{def}}{=} \begin{bmatrix} \vec{t}_1^{\,T} \\ \vec{t}_2^{\,T} \\ \vdots \\ \vec{t}_m^{\,T} \end{bmatrix} \tag{9.1}$$

Using Equation 9.1, our regression problem is to find a weight vector $\vec{w}$ so that

$$A\vec{w} \approx \vec{c} \tag{9.2}$$

We will assume that the observations $\vec{t}_i^{\,T}$ either exactly determine or over-determine the weight vector $\vec{w}$. Mathematically, we will assume that the data matrix $A$ is full rank, so $\text{rank}(A) = n$.

This matrix $A$ is the data matrix, which addresses the first of our two problems. The second problem – how to manage the disparate ranges of values in data vectors – is addressed by using another concept from statistics.

A *design matrix* for observations $\vec{t}_i^{\,T}$, usually written as $X$, has two important statistical properties. The first property is that each column of the matrix $X$ has an average value of 0, which is called a *zero-mean* column. The second property is that the standard deviation of the values of a column is 1. These two properties are used to *standardize* data for subsequent processing.

## 9.2 Standardized Data or *Z* Score

In regression – especially for methods that are developed from a statistical point of view – it is usual practice to use *standard* or standardized data. These are data that have a mean of zero and a variance of one. Historically, the concept seems to have been suggested, at around 1900, by Karl Pearson when he introduced the concept of a probability "ellipsoid", in which data were scored by their number of standard deviations from the mean in each relevant direction.

We can derive a standardization transformation for a single vector $\vec{a} \in \mathbb{R}^m$ and then generalize to a matrix of data. Previously, we defined the mean of a vector $\vec{a}$ as

$$\bar{a} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} a_i \tag{9.3}$$

For a matrix $A$, we will define the mean as a "row" matrix, which is called a *1-form*. The entries of the mean 1-form are the means of the columns. We will define the mean of a matrix $A$, which is a generalization of Definition 9.3, as

$$\bar{A} \stackrel{\text{def}}{=} \begin{bmatrix} \bar{a}_1 & \bar{a}_2 & \cdots & \bar{a}_n \end{bmatrix} \tag{9.4}$$

Statistically, the variance of a set of data has two definitions: the *sample* variance and the *population* variance. Here, we will use the sample variance; we will be cautious because other writings and code may explicitly or implicitly use the population variance.

For a vector $\vec{a}$, the sample variance is the sum of the squares of the entries of the vector divided by one less that the size of the vector. Using the conventional statistical symbol $\sigma^2$, we will define the sample variance of a vector as

$$\sigma_{\vec{a}}^2 \stackrel{\text{def}}{=} \frac{1}{m-1} \sum_{i=1}^{m} (a_i - \bar{a})^2 = \frac{\|\vec{a} - \vec{1}\,\bar{a}\|^2}{m-1} \tag{9.5}$$

We will define the standard-deviation of a matrix $A$ as a diagonal matrix $D$. The $j^{\text{th}}$ diagonal entry of $D$ is the sample standard deviation of the $j^{\text{th}}$ column of $A$. The definition of the sample variance of a matrix $A$, using Definition 9.5, is

$$D_A \stackrel{\text{def}}{=} \begin{bmatrix} \sigma_{\vec{a}_1} & & & \\ & \sigma_{\vec{a}_2} & & \\ & & \ddots & \\ & & & \sigma_{\vec{a}_n} \end{bmatrix} \tag{9.6}$$

We can use these definitions to concisely state how we will standardize data. Suppose that we are given a regression problem with a matrix of independent data $A$ and a vector of dependent data $\vec{c}$, that is to be approximated as

$$A\vec{w} \approx \vec{c} \tag{9.7}$$

We want to convert Equation 9.7 to a standardized form, which implies that we seek:

- A design matrix $X$ with columns that are zero-mean and unit variance
- A vector of dependent data $\vec{y}$ with zero mean and unit variance
- A linear model of the data that is

$$X\vec{u} = \vec{c} \tag{9.8}$$

For the independent data in $A$, we must subtract the mean of $A$ from $A$ and scale the difference to have a sample variance of one. We must likewise modify the dependent data in $C$. These operations can be expressed as

$$\begin{aligned} X &= [A - [\vec{1}\bar{A}]]D_A^{-1} \\ \vec{y} &= [\vec{c} - [\vec{1}\bar{c}]]D_{\vec{c}}^{-1} \end{aligned} \tag{9.9}$$

Observation: **Equation 9.7 and Equation 9.8 solve different problems**

We can explore some effects of using zero-mean data by considering a problem where the sample variances are one, that is, where $D_A = I$ and $D_{\vec{c}} = 1$. For any such unit-variance problem, we would estimate a solution of Equation 9.8 to be the vector $\hat{u}$. We would then estimate the dependent data, which are in the vector $\vec{c}$ of Equation 9.7, from the zero-mean independent data in the matrix $X$ as

$$\hat{c} = X\hat{u} + [\vec{1}\bar{y}] \tag{9.10}$$

In general, the solution $\hat{u}$ of Equation 9.8 is *not* the same as a solution $\hat{w}$ of Equation 9.7.

Methods that are derived from the statistical point of view generally try to find solutions that are based on standard data, as stated in Equation 9.8. In particular, we generally want to avoid problems where a constant offset term is stated. We can see that the transformation in Equation 9.9 will produce a division by zero when any column of $A$ is a constant value, because we would be trying to invert a scaling matrix that refers to the variance of the zero vector $\vec{0}$.

## Extra Notes: Example of a Standardized Data Vector

Consider the vector $\vec{a} = \begin{bmatrix} 15 \\ 17 \\ 31 \\ 19 \\ 3 \end{bmatrix}$

The mean value of $\vec{a}$ is 17, so the zero-mean vector is $\vec{m} = \begin{bmatrix} -2 \\ 0 \\ 14 \\ 2 \\ -14 \end{bmatrix}$

The variance of $\vec{a}$ is the sum of the squares of the entries of $\vec{m}$, divided by $(5 - 1) =$, so $\sigma^2 = 100$

The standardization of $\vec{a}$ is $\vec{m}$ divided by $\sigma = 10$, which is $\vec{x} = \begin{bmatrix} -0.2 \\ 0.0 \\ 1.4 \\ 0.2 \\ -1.40 \end{bmatrix}$



 