Statistics & Linear Regression

Zvi Rosen
Department of Mathematics

October 31, 2016

Plan

How do we analyze a set of data using MATLAB?

In particular, how do we obtain representative statistics for a set of real-number data?

When our data points consist of two real-numbers, how do we analyze the data set, and the relationship between the two coordinates?

Mean, Median, Mode

Let X be a data set, equivalently, a point in $\mathbb{R}^n$.

- mean(X) is the average of all values: $\frac{1}{n}\sum_{i=1}^{n} x_i$.
- median(X), when n is odd, is the value of index $(n-1)/2$ when the data are ordered from smallest to largest. When n is even, it is the mean of the values with indices $n/2$ and $n/2+1$.
- mode(X) is the most commonly occurring value in the dataset. When there is a tie, MATLAB gives the tie to the smallest value.

The MATLAB commands are `mean`, `median`, and `mode`.

Representative Statistics from Metrics

Suppose we use $d(p, q)$ to denote the distance between points p and q. Let X be our data set, and let Y be some point with the same number in each coordinate (i.e. $(y, y, \ldots, y)$).

1. $d(X, Y) = \sum \mathbf{1}[X_i \neq Y_i]$. (Hamming)
2. $d(X, Y) = \sum |X_i - Y_i|$. (Taxicab)
3. $d(X, Y) = \max |X_i - Y_i|$. ($\infty$-norm)
4. $d(X, Y) = \sqrt{\sum (X_i - Y_i)^2}$ (2-norm)

What value of y minimizes the distance between X and Y?

Standard Deviation & Variance

Mean, median, and mode give us a single number to estimate the whole data set. We often want an estimate for how spread out the values are from that single estimate.

- Range: $\max(X) - \min(X)$.
- Variance: $S_t = \sum (X_i - \bar{X})^2$, where $\bar{X}$ is the mean.
- Standard Deviation: $s_y = \sqrt{S_t/(n-1)}$.

Normal Distribution

Many data distributions are distributed normally. The normal distribution is defined as

$$f(x) = -\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

This is the well-known bell curve with mean μ and standard deviation σ. The probability of any data point falling between values a and b is equal to the integral of f from a to b.

Generating random numbers in MATLAB

To generate random numbers in the uniform distribution, i.e. any range between 0 and 1 is equally likely, use the command `rand(m,n)`. This returns an $m \times n$ matrix of random numbers.

To generate random numbers in the normal distribution with mean 0 and standard deviation 1, use the command `randn(m,n)`. This returns an $m \times n$ matrix of random normal numbers.

1. How do we generate random numbers in the uniform distribution between 0 and 100?
2. How do we generate random numbers in the normal distribution with mean 1 and standard deviation 0.5?

Regression

Now, we consider data points with data points with two coordinates (x, y). We want to assess the hypothesis that these two variables are related. In particular, can we produce a line that is close to predicting their relationship?

History: the term "regression" comes from a biological phenomenon of children of tall parents regressing to average population height. It was eventually adopted to describe the analysis of relationships between two data sets.

Least-squares Error

In order to find a line "close" to the data, we need to know what "close" means. (pick your metric!)

The taxicab metric does not specify a unique line.
The ∞-metric gives too much weight to outliers.
The Euclidean metric (corresponding to the 2-norm) is our best option. In particular, we take the set of x-data points, and compare the corresponding line values to the actual y-values. This tells us how "close" to the data our line is.

Least-squares Line

To find the least-squares error, i.e. minimum Euclidean distance, we use our optimization tool box.
Let S_r be the sum of the residuals,
and let $a_0 + a_1 \cdot x$ be the linear estimator.

We take dS_r/da_0 and dS_r/da_1 to find the optimal values of these parameters for the least-squares line.

Quantifying Error

Just like we measured spread from a central estimator in one coordinate, we want to measure spread from our regression line in two coordinates.

Our approach is to use the residuals:

$$s_{y/x} = \sqrt{S_r} n - 2.$$

(standard error of the estimate)

The goodness of the fit by our line can be described by:

$$r^2 = \frac{S_t - S_r}{S_t}$$

(r is the correlation coefficient. This describes the proportion of the total variance in y-values that can be described by the line as opposed to residual effects)