Second Midterm Exam: November 18 (Tuesday) 3:30pm – 5:30pm.
Lectures 6 – 20, Open-note, JH 245

Lecture 21

Statistics

- **Statistics**

- ***Arithmetic mean***: the sum of the individual data points (y_i) divided by he number of points n:

$$\bar{y} = \frac{\sum y_i}{n}$$

In MATLAB, mean(Y) returns the mean value of the elements in Y if Y is a vector. For matrices, it returns a row vector containing the mean value of each column.

```
>> Y = [0, 2, 5, 1];
>> m = mean(Y)
>> ans =
      2
>> Y = [0, 2, 5, 1; 0, 2, 5, 1];
>> m = mean(Y)
ans =
      0    2    5    1
```

How to calculate the mean value of each row?

How to calculate the mean value of the matrix?

Statistics

- **Statistics**

- ***Median***: returns the midpoint of a group of data.

  In MATLAB, For vectors, median(Y) returns the median value of the elements in Y. For matrices, it returns a row vector containing the median value of each column. The median value is the middle number or the mean of the middle two numbers **in sorted order (denpends on the number of values)**.

```
>> Y = [5  2  3  6  9];

>> n = median(Y)

n =

     5

>> Y = [5  2  3  6  9  10];

>> n = median(Y)

n =

     5.5000
```

- ***Mode***: returns the value that occurs most frequently in a group of data. mode(Y)
- `>> Y = [1 4 2 8 2];`
- `>> p = mode(Y)`

What if all values appear once?

Statistics

➢ Statistics

- ***Standard deviation***. the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

$$s_y = \sqrt{\frac{S_t}{n-1}}$$

where *St* is the sum of the squares of the data residuals:

$$S_t = \sum (y_i - \bar{y})^2 \ , \qquad \bar{y} = \frac{\sum y_i}{n}$$

and n-1 is referred to as the degree of freedom.

For vectors, std(Y) returns the standard deviation. For matrices, it returns a row vector containing the standard deviation of each column.

Statistics

➢ **Statistics**

- ***Variance***, measures how far a set of numbers are spread out from their average value. It is calculated as the average squared deviation of each number from the mean of a data set. Variance is the square of the standard deviation.

$$\text{variance} = s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

  In MATLAB, var(Y) returns the variance of the values in vector Y. For matrices, it returns a row vector containing the variance of each column of Y

- Standard deviation and variance are the most commonly used measures of spread.
- Coefficient of variation:

$$c.v. = \frac{s_y}{\bar{y}} \times 100\%$$

- **In MATLAB, if a matrix is given, the statistics will be returned for each column**

How to do the statistics on the rows?

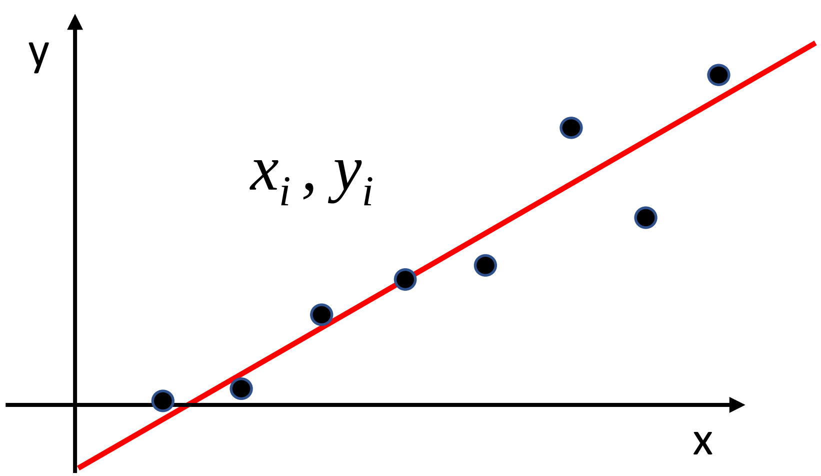
Linear Least-Squares Regression

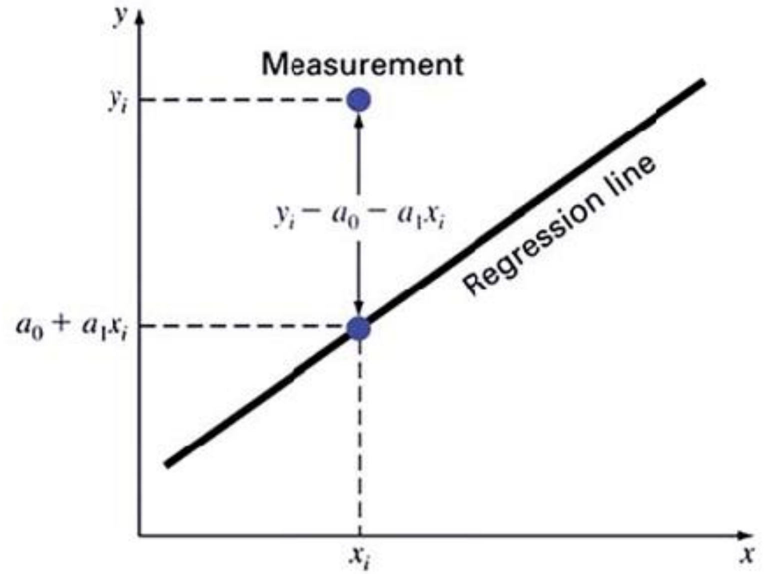
- Linear least-squares regression is a method to determine the "best" coefficients in a linear model $y = a_0 + a_1 x$ for a given data set in (x, y) space.
- "Best" for least-squares regression means minimizing the sum of the squares of the estimate residuals. For a straight line model $y = a_0 + a_1 x$, this gives

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - y(x_i)]^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

Basically, we will find the values of a0 and a1 to minimize Sr.

How to minimize Sr? $\frac{\partial S_r}{\partial a_0} = 0, \quad \frac{\partial S_r}{\partial a_0} = 0$





Linear Least-Squares Regression

- This method will yield a unique line for a given set of data.
- For the model:

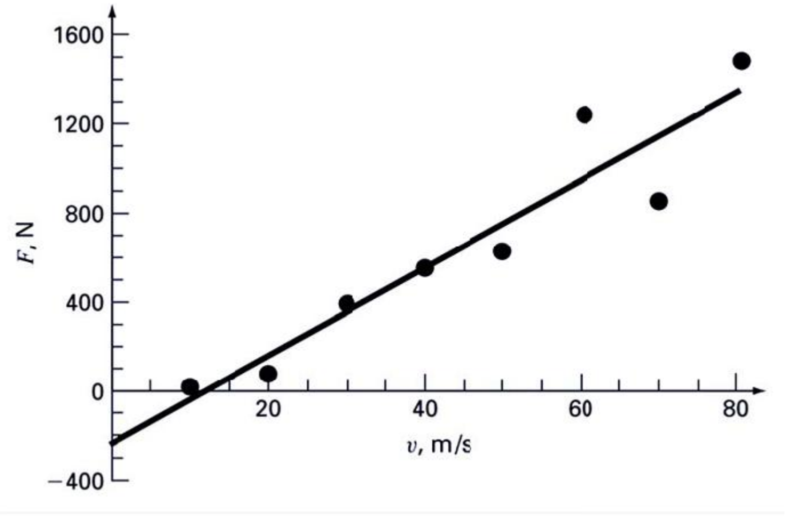
$$y = a_0 + a_1 x$$

the slope and intercept producing the best fit can be found by making:

$$\begin{cases} \dfrac{\partial S_r}{\partial a_0} = \displaystyle\sum_{i=1}^{n} (-2)(y_i - a_0 - a_1 x_i) = 0 \\ \dfrac{\partial S_r}{\partial a_1} = \displaystyle\sum_{i=1}^{n} (-2x_i)(y_i - a_0 - a_1 x_i) = 0 \end{cases}$$

Then a_1 and a_0 can be obtained as

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \qquad a_0 = \bar{y} - a_1 \bar{x}$$



$\sum$ Can be calculated using function sum() or using a for loop in MATLAB.

Linear Least-Squares Regression

Method 1

```
function [a0,a1] = mylinfit(x,y)
% least squares regression for a straight line

n = length(x);

numerator = n*sum(x.*y)-sum(x)*sum(y);
denom = n * sum(x .^2) – (sum(x))^2;

a1 = numerator / denom;
a0 = mean(y) – a1*mean(x);
```

Method 2

p = polyfit(x,y,N) finds the coefficients of an polynomial p(x) of degree N that fits the y-data best in a least-squares sense.

$$y = a_0 + a_1 x$$

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

```
>> f = @(x) x.^2 + 4*x + 2;
>> x = linspace(-6,5,10);
>> y = f(x);
>> plot(x,y,'bo')
>> [a0,a1] = mylinfit(x,y)

>> p = polyfit(x,y,1)
```

a0 is the constant, a1 is the coefficient of x;
p(1) is the coefficient of x, and
p(2) is the constant

Nonlinear Relationships

- Linear regression is predicated on the fact that the relationship between the dependent and independent variables is linear - that is not always the case.

$$\text{exponential:} \qquad y = \alpha_1 e^{\beta_1 x}$$

$$\text{power:} \qquad y = \alpha_2 x^{\beta_2}$$

$$\text{saturation - growth - rate:} \qquad y = \alpha_3 \frac{x}{\beta_3 + x}$$

Nonlinear Relationships

- One option for finding the coefficients for a nonlinear fit is to linearize it. For the three common models, this way involve taking logarithms or inversion:

Model	Nonlinear	Linearized
exponential :	$y = \alpha_1 e^{\beta_1 x}$	$\ln y = \ln \alpha_1 + \beta_1 x$
power :	$y = \alpha_2 x^{\beta_2}$	$\log y = \log \alpha_2 + \beta_2 \log x$
saturation - growth - rate :	$y = \alpha_3 \dfrac{x}{\beta_3 + x}$	$\dfrac{1}{y} = \dfrac{1}{\alpha_3} + \dfrac{\beta_3}{\alpha_3} \dfrac{1}{x}$

$y' \equiv \ln y, \quad x' \equiv x$
$a' \equiv \ln\alpha_1, \; b' \equiv \beta_1$

$y' \equiv \log y, \quad x' \equiv \log x$
$a' \equiv \log\alpha_2, \; b' \equiv \beta_2$

$y' \equiv 1/y, \quad x' \equiv 1/x$
$a' \equiv 1/\alpha_3, \; b' \equiv \beta_3/\alpha_3$

Unified linear form, $y' = a' + b' x'$

- Define new variables (x', y') and coefficients (a', b').
- Convert x_i and y_i to x'_i and y'_i, and then do linear LS regression in space (x', y'), and find a' and b'
- Convert x'_i, y'_i, a' and b' back to x_i, y_i, α and β

- **Homework on Canvas**