NYU

NYU Summer Machine Learning Program

Presenter Name Here
Date Here

NYU

Linear Regression

Day 2

NYU

Learning Objectives

- ❑How to load data from a text file
- ❑How to visualize data via a scatter plot
- ❑Describe a linear model for data
  - ❑ Identify the target variable and predictor
- ❑Compute optimal parameters for the model using the regression formula
- ❑Fit parameters for related models by minimizing the residual sum of squares
- ❑Compute the measure of the fit

Outline

- ❑Motivating Example: Virus inactivation in wetland waters due to sunlight
- ❑Linear Model
- ❑Least Squares Fit Problem
- ❑Sample Mean and Variance
- ❑LS Fit Solution
- ❑Assessing Goodness of Fit

Example: Wetland Virus Inactivation from Sunlight Exposure

- Getting the data:
- Data can be found here.

10 lines (9 sloc) | 623 Bytes

Raw Blame History

Search this file...

Time (h)	Ct1 Clear	Ct1 5cm	Ct1 20cm	Ct2 Clear	Ct2 5cm	Ct2 20cm
0	1000000	1000000	1000000	1000000	1000000	1000000
0.75		547323.127587941	680310.384129047		427843.265171708	724747.957348342
1	124572.849876906			124242.476640091		
1.5		191599.336145909	434275.771005791		163016.250572439	491827.603977672
2	23632.9534751967			28479.8042537738		
2.25		66699.2575995024	267168.775509554		60215.5974896074	224766.301820444
3	2364.27505448768	21358.0697267647	123072.776977687	3703.73580494049	15289.2787968066	112632.54568239
4	472.99222292394	4481.98088500142	69507.0729364707	569.997564484227	3151.41791835128	33667.6652164115

Example: Wetland Virus Inactivation from Sunlight Exposure

- Reading & Visualizing the Data:
- Using Python packages - Pandas, Numpy, Matplotlib.
- Pandas:
  - Used for reading and writing data files
  - Loads data into dataframes

```
import pandas as pd
import io
```

- Numpy:
  - Used for numerical operations, including linear algebra
  - Data is stored in ndarray structure
  - We convert from dataframes to ndarray

```
import numpy as np
```

- Matplotlib:
  - Used for MATLAB-like plotting and visualization

```
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
```

Example: Wetland Virus Inactivation from Sunlight Exposure

- Reading the data using python's pandas library:
- *pd.read_csv* converts a comma-separated values file into a 2D data structure with labeled axes.

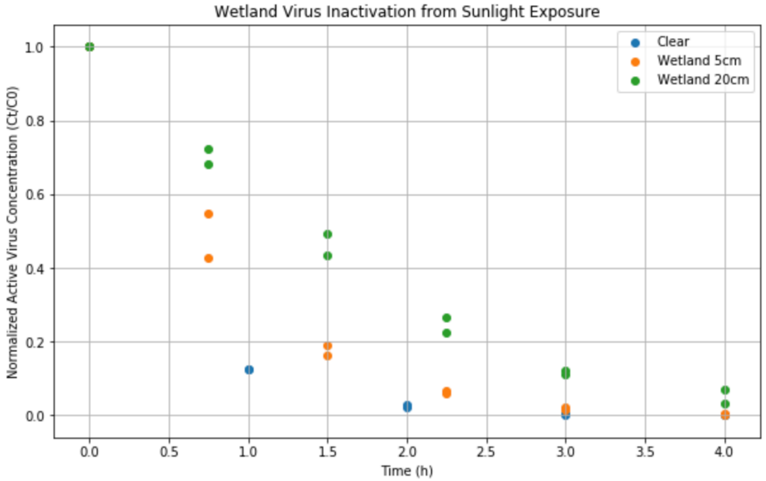
```
import pandas as pd
import io

df = pd.read_csv(io.BytesIO(uploaded[filename]))
df
```

	Time (h)	Ct1 Clear	Ct1 5cm	Ct1 20cm	Ct2 Clear	Ct2 5cm	Ct2 20cm
0	0.00	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000
1	0.75	NaN	547323.127588	680310.384129	NaN	427843.265172	724747.957348
2	1.00	124572.849877	NaN	NaN	124242.476640	NaN	NaN
3	1.50	NaN	191599.336146	434275.771006	NaN	163016.250572	491827.603978
4	2.00	23632.953475	NaN	NaN	28479.804254	NaN	NaN
5	2.25	NaN	66699.257600	267168.775510	NaN	60215.597490	224766.301820
6	3.00	2364.275054	21358.069727	123072.776978	3703.735805	15289.278797	112632.545682
7	4.00	472.992223	4481.980885	69507.072936	569.997564	3151.417918	33667.665216

Example: Wetland Virus Inactivation from Sunlight Exposure

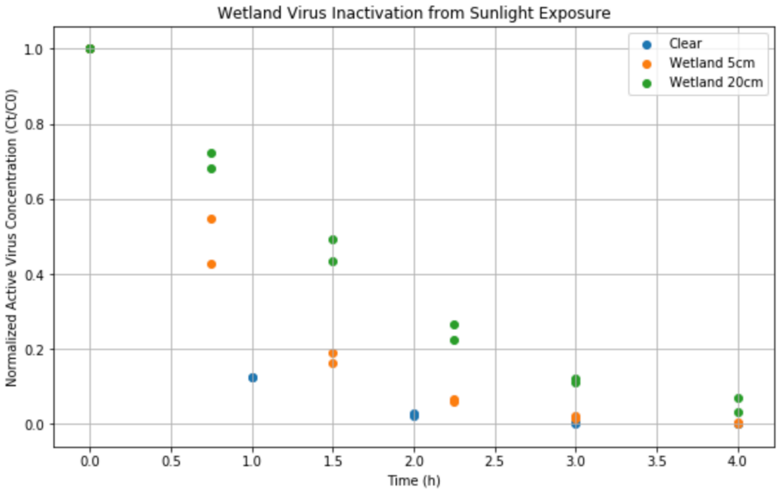
- ❑ Visualizing the Data:
- ❑ It's always a good idea to visualize the data before performing any operations on it.
- ❑ Using python's Matplotlib:
- ❑ *plt.scatter(x,y)* plots a scatter plot of y vs x.

```
plt.scatter(cat_time[~nanloc_clear], ct_clear[~nanloc_clear])
plt.scatter(cat_time[~nanloc_wetla], ct_5cm[~nanloc_wetla])
plt.scatter(cat_time[~nanloc_wetla], ct_20cm[~nanloc_wetla])
plt.grid(True)
plt.legend(['Clear','Wetland 5cm','Wetland 20cm'])
plt.title('Wetland Virus Inactivation from Sunlight Exposure')
plt.xlabel('Time (h)')
plt.ylabel('Normalized Active Virus Concentration (Ct/C0)');
```



Exercise: Postulate a Model

- ❑ Try to find a mathematical model to predict the active virus concentration from time:
  - ❑ Try to make a reasonable/eyeball guess, without using a program.



Outline

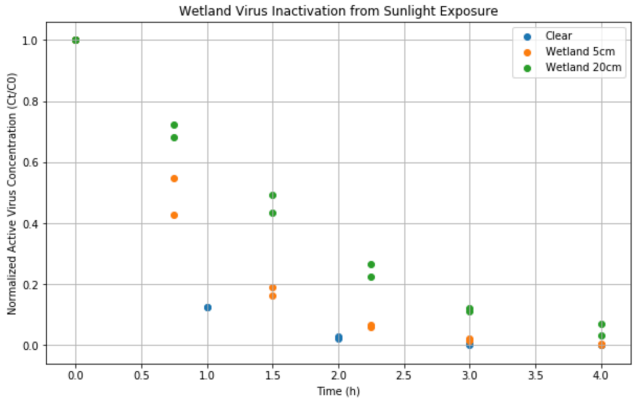
- Motivating Example: Virus inactivation in wetland waters due to sunlight
- Linear Model
- Least Squares Fit Problem
- Sample Mean and Variance
- LS Fit Solution
- Assessing Goodness of Fit

Understanding the Data:

❑ Our y axis: The variable we are trying to predict. Can be called: Dependent variable, response variable, target, regressand, …

❑ Our x axis: The variable we are using to predict. Can be called: Predictor, attribute, covariate, regressor …

❑ Each data point is called a sample. In this example, we are using a scatter plot to view the samples.



Linear Model

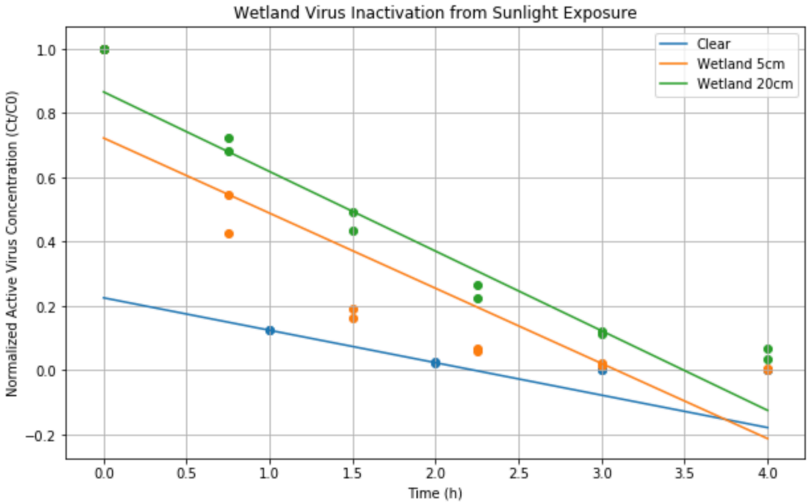
- Assume a linear relationship among the samples - using the intercept (beta0) and slope (beta1).

$$\beta_0 = c_a - \beta_1 t_a \qquad \beta_1 = \frac{c_b - c_a}{t_b - t_a}$$

- Why do we use a Linear Model?
  - Generally, most natural phenomena have a linear relationship
  - Simple computation, and easy to interpret.

Linear Model

- ❑ Plotting the Linear Model using python's Matplotlib:
- ❑ *[plt.plot]()* is used to plot y versus x as lines or markers.



Outline

- ❑Motivating Example: Virus inactivation in wetland waters due to sunlight
- ❑Linear Model
- ❑Least Squares Fit Problem
- ❑Sample Mean and Variance
- ❑LS Fit Solution
- ❑Assessing Goodness of Fit

NYU

Linear Model Residual

- As we can see, a linear model does not fit all the sample, which means it is not a good fit for our data.
- We add a residual term to our lineal model (e):

$$y = \beta_0 + \beta_1 x + \epsilon$$

- For a $\beta = (\beta_0, \beta_1)$
- We define a residual sum of squares (RSS): $\mathrm{RSS}(\beta_0, \beta_1) := \sum_{I=1}^{n} (y_i - \hat{y}_i)^2$
- A Least Squares Solution is to find (β_0, β_1) to minimise RSS. $\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2 = \sum_{i=0}^{N-1} (y_i - \beta_0 - \beta_1 x_i)^2$

NYU

Outline

- Motivating Example: Virus inactivation in wetland waters due to sunlight
- Linear Model
- Least Squares Fit Problem
- Sample Mean and Variance
- LS Fit Solution
- Assessing Goodness of Fit

NYU

Least-Squares Fit Solution: Sample Mean and Standard Deviations

- Given data: $(x_i, y_i), i = 1, \dots, N$
- Sample mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Sample variances

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 \,, \qquad s_y^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2$$

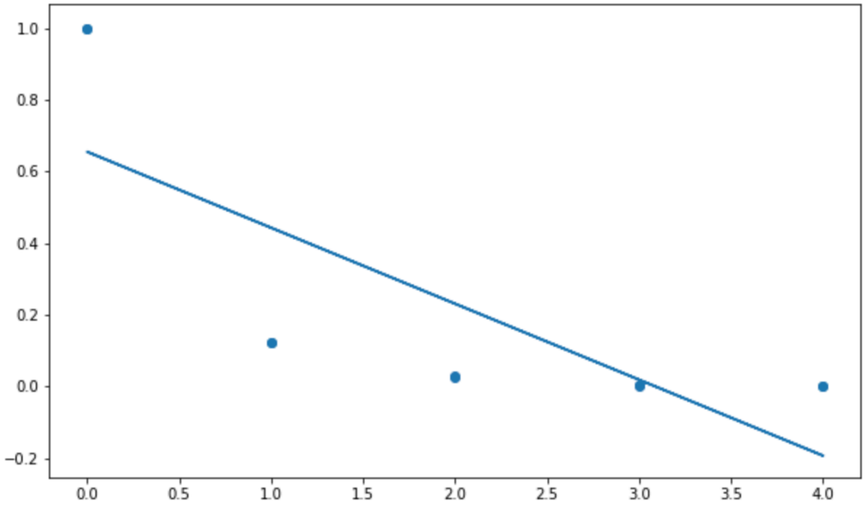
Outline

- Motivating Example: Virus inactivation in wetland waters due to sunlight
- Linear Model
- Least Squares Fit Problem
- Sample Mean and Variance
- LS Fit Solution
- Assessing Goodness of Fit

NYU

Least-Squares Fit Solution

- ❑ Using python to find the Least Squares solutions
- ❑ *np.mean* uses python's numpy library to compute the arithmetic mean along the specified axis.




```
# Calculate the mean of x and y
xm = np.mean(x)
ym = np.mean(y)

syy = np.mean((y-ym)**2)          # Variance of y
syx = np.mean((y-ym)*(x-xm))      # Covariance of x and y
sxx = np.mean((x-xm)**2)          # Variance of x

beta1 = syx/sxx
beta0 = ym - beta1*xm
print('beta0 = {:.2f}, beta1 = {:.2f}'.format(beta0,beta1))
```

```
beta0 = 0.65, beta1 = -0.21
```

Outline

- ❑Motivating Example: Virus inactivation in wetland waters due to sunlight
- ❑Linear Model
- ❑Least Squares Fit Problem
- ❑Sample Mean and Variance
- ❑LS Fit Solution
- ❑Assessing Goodness of Fit

NYU

Assessing the goodness of the fit

- We can use the R2 score to estimate the goodness of our fit.
- The best (and maximum) R2 score is 1. It can be negative if the fit is really bad.
- It can be calculated on python by:

```
RSS = np.sum((y - y_hat)**2)
N = y.size    # Number of samples in the data set
R2 = 1 - (RSS/N)/syy
print('R^2 = {:.2f}'.format(R2))
```

```
R^2 = 0.60
```