CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W6: 9/22-24/15

Administrative

- Updated Weekly Schedule (on Moodle)
- HW-2: To be posted (9/22-23/15)
  - DUE: 10/4/15
- 1st Midterm: 10/6/15
- Feedback (HWs)
  - What worked best for enhancing your understanding and what didn't?
  - Anything (change) that you would like to see in HW2?

NC STATE UNIVERSITY

So far

- 1st Module is completed
  - Exploratory Data Analysis
  - Summary Statistics, Histograms, etc.
  - Basic Probability (including set operations)
  - Basic Linear/Matrix Algebra, Calculus
  - Probability distributions
  - Parameter estimation
  - Sampling distribution, CLT, C.I.
  - Hypothesis testing
- 1st Midterm will be based on 1st module
  - All these topics will be reviewed on 10/5

NC STATE UNIVERSITY

Today

- Regression Analysis (covered in 3 lectures)

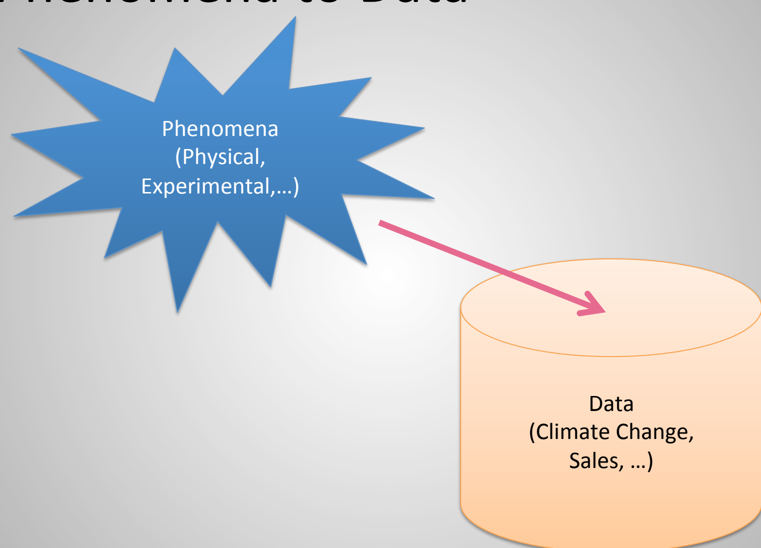
NC STATE UNIVERSITY

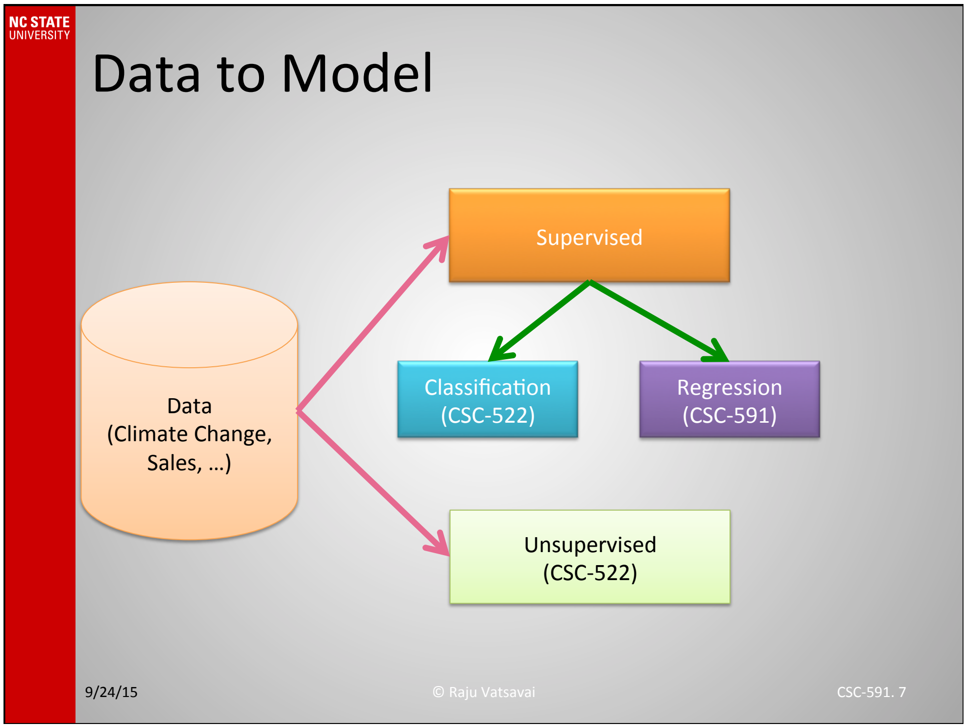
Learning

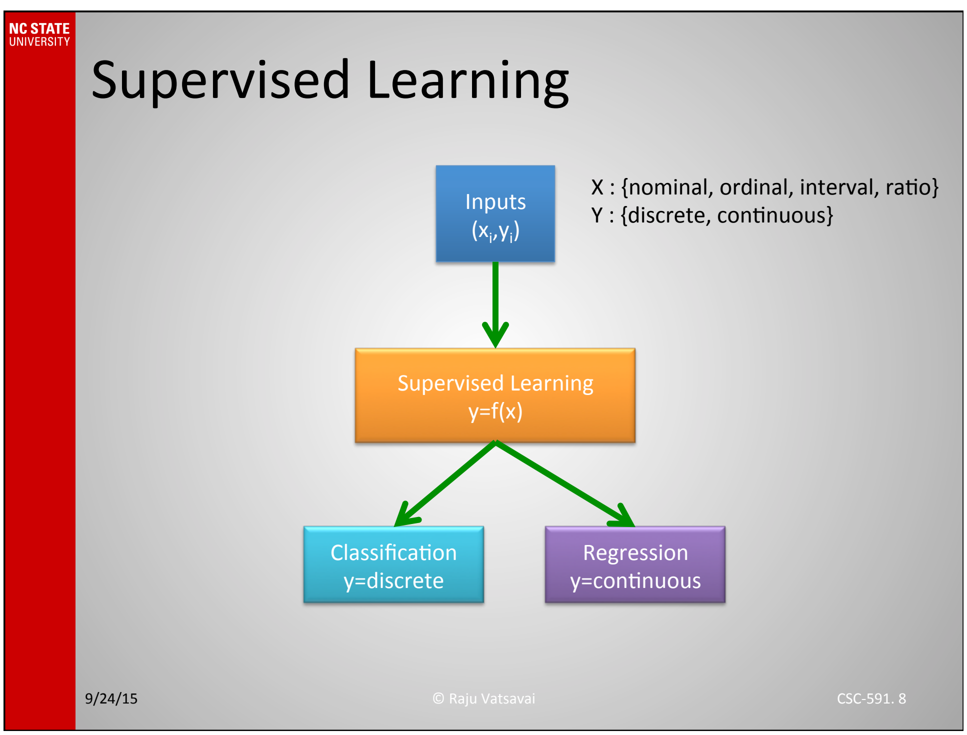
- Observe a phenomena
- Construct a model of that phenomena
- Make predictions using the model

NC STATE UNIVERSITY

Phenomena to Data



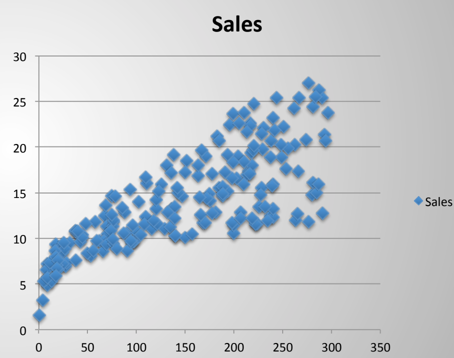

NC STATE UNIVERSITY
Data to Model
Supervised
Data
(Climate Change,
Sales, …)
Classification
(CSC-522)
Regression
(CSC-591)
Unsupervised
(CSC-522)
9/24/15
© Raju Vatsavai
CSC-591. 7


NC STATE UNIVERSITY
Supervised Learning
Inputs
(x_i, y_i)
X : {nominal, ordinal, interval, ratio}
Y : {discrete, continuous}
Supervised Learning
y=f(x)
Classification
y=discrete
Regression
y=continuous
9/24/15
© Raju Vatsavai
CSC-591. 8

NC STATE UNIVERSITY

Ex: Advertising Data

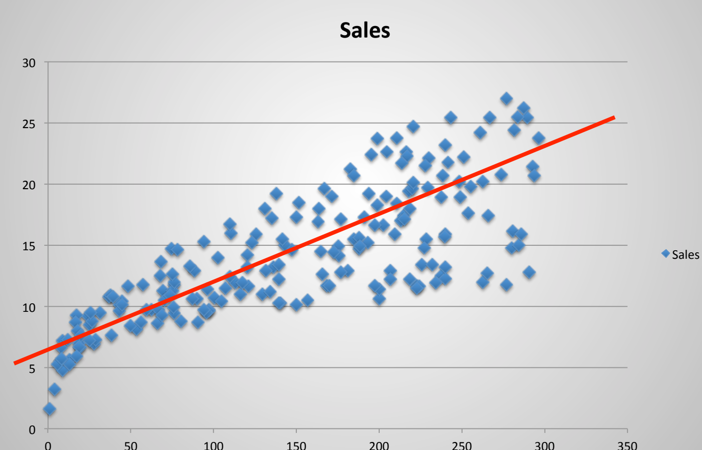
TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24	4	17.4
23.8	35.1	65.9	9.2



Is there a relationship between advertising budget (TV) and sales?

NC STATE UNIVERSITY

Can We Establish y = f(x)



NC STATE UNIVERSITY

Regression

- Regression is a method for studying the relationship between response variable Y and covariates X.
  - The covariate is also called predictor variable or explanatory variable or a feature
- The term "regression" is due to Sir Francis Galton (1822-1911) who noticed that tall and short men tend to have their sons with heights closer to the mean. He called this "regression towards the mean"

9/24/15 © Raju Vatsavai CSC-591. 11

NC STATE UNIVERSITY

Simple Linear Regression

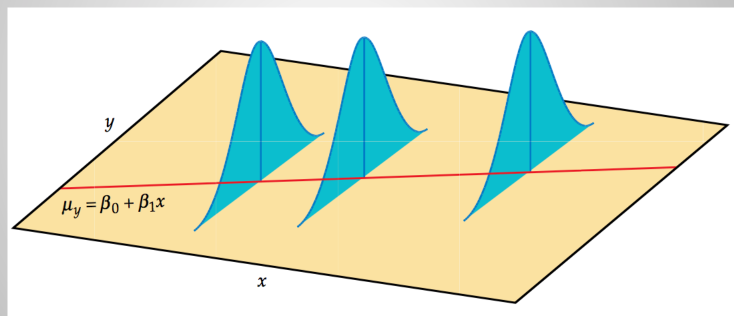
- SLR studies the relationship between a response variable *y* and a single explanatory variable *x*.
- SLR assumes that for each value of x the observed values of the response variable y are Normally distributed with a mean (μ_y) that depends on *x*.
  - The mean of response variable, μ_y changes as *x* changes. The means all lie on a straight line. That is, $\mu_y = \beta_0 + \beta_1 x$.
  - Individual responses of *y* with the same *x* vary according to a Normal distribution. These Normal distributions all have the same standard deviation.

9/24/15 © Raju Vatsavai CSC-591. 12

NC STATE UNIVERSITY

Simple Linear Regression

- The SLR model ($\mu_y = \beta_0 + \beta_1 x$), with intercept β_0 and slope β_1, assumes that all means (μ_y) lie on a line when plotted against *x*. This is the population regression line.



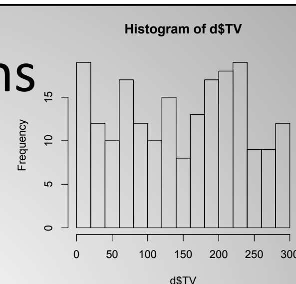
The statistical model for linear regression; the mean response is a straight-line function of the explanatory variable.

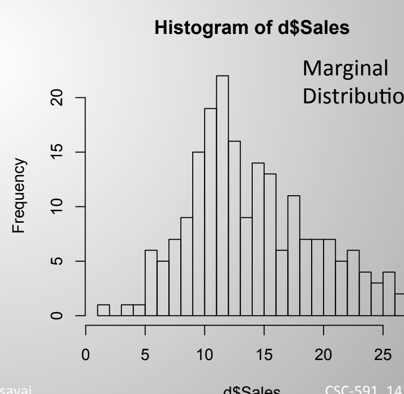
NC STATE UNIVERSITY

Marginal Distributions

- Sales Data

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24	4	17.4
23.8	35.1	65.9	9.2





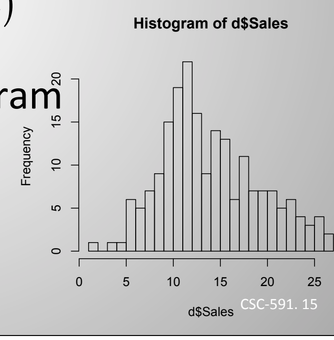
Marginal Distributions

NC STATE UNIVERSITY

Least Squares

- How to find "middle" via least squares?
- Let Y_i be sales, i = 1..200, then the middle μ is the one that minimizes

$$\sum_i^n \left(Y_i - \mu\right)^2$$



- This is the center of histogram
- Then, $\mu = \bar{Y}$

NC STATE UNIVERSITY

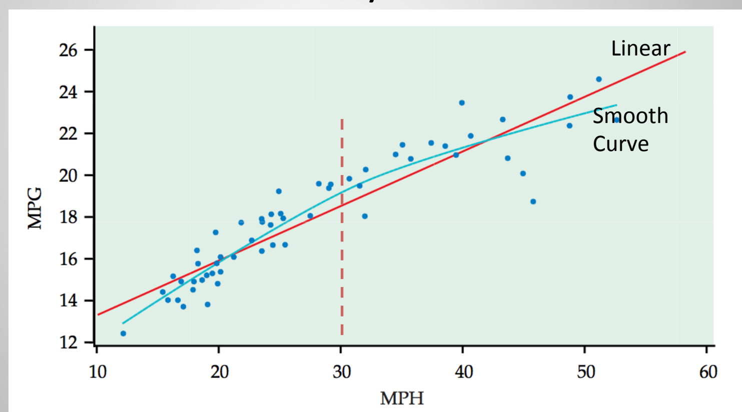
Show that $\mu = \bar{Y}$

Proof is demonstrated in the class.
You should try it as practice question.

NC STATE UNIVERSITY

Relationship is Approximate Linear

- Consider the relationship between car driven speed and fuel efficiency

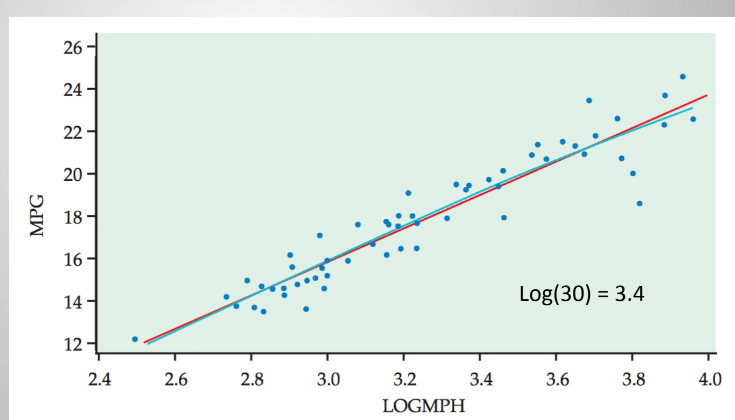


9/24/15 © Raju Vatsavai CSC-591. 17

NC STATE UNIVERSITY

Options

- Confine SLR to a region (e.g., up to 30MPH)
- Transformation (e.g., log)

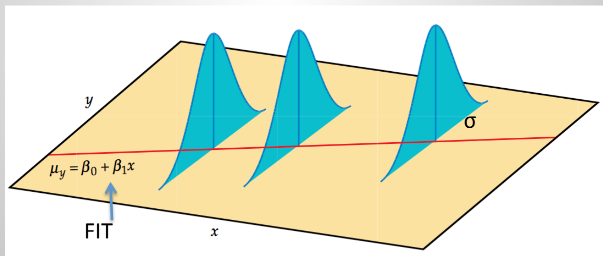


9/24/15

NC STATE UNIVERSITY

SLR Model

- Population regression gives mean value. However, we can't observe this line on sample data. The statistical model for linear regression consists of population regression line and a description of the variation of *y* about the line. That is, data = fit + residual. The residual part represents deviations of the data from the line of population means. We assume that these deviations are Normally distributed with standard deviation σ.



NC STATE UNIVERSITY

SLR Model

SIMPLE LINEAR REGRESSION MODEL

Given n observations of the explanatory variable x and the response variable y,

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

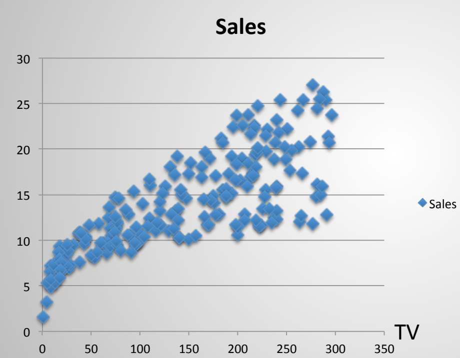
the **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Here $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviations ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ.

The parameters of the model are β_0, β_1, and σ.

Correlation



The correlation measures the direction and strength of the relationship between two quantitative variables, written as *r.*

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

9/24/15 © Raju Vatsavai CSC-591. 21

Regression Parameters

- Least-squares line:

$$\hat{y} = b_0 + b_1 x$$

- Intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

- Slope

$$b_1 = r\frac{s_y}{s_x}$$

- Residual, e_i = observed response – predicted response

$$e_i = y_i - b_0 - b_1 x_i$$

9/24/15 © Raju Vatsavai CSC-591. 22

Regression Parameters

- For SLR, the estimate of σ^2 is the average of squared residual

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

- The quantity (n – 2) is called the degrees of freedom for s^2
- The estimate of σ is given by $s = \sqrt{s^2}$

9/24/15 © Raju Vatsavai CSC-591. 23

Properties

- Sum of residuals is 0
- Sum of squared residuals is minimum (this is the constraint to be satisfied in deriving least squares estimators of the regression parameters)
- Sum of the observed values of Y_i equals sum of the fitted values $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$
- Sum of the weighted residuals is zero when residual in the i^{th} trail is weighted by x_i

$$\sum_{i=1}^{n} X_i e_i = 0 \qquad \text{Likewise,} \qquad \sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

9/24/15 © Raju Vatsavai CSC-591. 24

NC STATE UNIVERSITY

Example

- Consider linear regression model with μ_y = 40.5 – 2.5*x* and σ = 2.0
  - What is the slope of the population regression line
  - What is y when x = 10?
  - Between what two values would approximately 95% of the observed responses, *y*, fall when x = 10?

NC STATE UNIVERSITY

Acknowledgements

- G. James, et. al., Moore, et. al. Caffo, et. al.