Introduction to Linear Modeling

Professor Jeff Saltz
Professor Jeff Stanton

School of Information Studies
SYRACUSE UNIVERSITY
ischool.syr.edu

Summary of Previous Learning

What you should know and be able to do at this point :

1. List major skills needed by data scientists and describe the development of a DS project with domain analysis, SMEs, data and modeling
2. Explain and use a data frame in R and various diagnostics for variables and data structures; describe and use multiple strategies for accessing external data from R; automate with functions
3. Define and calculate the most common descriptive statistics; describe the effects of randomness on sampling; create and interpret a sampling distribution including defining the law of large numbers and the central limit theorem
4. Use plot and ggplot to visualize data and create maps

Data Science Day

In person and virtually (Nov 4 & 5th)

Registration is required:
https://graduate.admissions.go.syr.edu/register/?id=fe51b3db-7d2f-4df3-a8c0-16c29ea08964

Thursday talks (starting at ~4pm):

- Exploring What is Data Science & Analytics
- What is Visual Analytics?
- Metadata & Analytics
- Learning Analytics
- Graduate Student Panel

Friday talks (starting at ~10am):

- The Use of Data Science to Better Address COVID-19 - Thibaut Jombart (WHO / UK researcher)
- Applied Deep Learning, Examples in Industry
- Ethics in Data Science: Problems, Approaches, and Future

Break for lunch

- Managing Risk in Data Science
- Data Science in the Real World – a discussion of what our adjunct faculty do outside the iSchool
- Data Science Job Market Update

Schedule: https://calendar.syracuse.edu/events/2021-nov-04/data-science-day/

Exam Review

School of Information Studies
SYRACUSE UNIVERSITY

Covers Weeks 1-8

1. List major skills needed by data scientists
2. Describe a DS project with domain analysis and SMEs
3. Explain basic concepts of data modeling
4. Explain and use a data frame in R
5. Define the most common descriptive statistics and calculate them with R using appropriate functions
6. Demonstrate the development of a simple function in R
7. Describe the effects of randomness on sampling
8. Create and interpret a sampling distribution including defining the law of large numbers and the central limit theorem
9. Visualize a distribution and interpret a histogram
10. Describe multiple strategies for accessing external data from R
11. Use ggplot to visualize data including maps
12. Create, use and understand linear models
13. Be able to discuss how to apply data science "in the real world"

Homeworks

- Read in a CSV data set; summarize a data frame with summary(), str(), dim()
- Locate and repair missing data
- Run and interpret basic descriptive statistics: mean(), median(), sd(), var(), range()
- Create and interpret a histogram and scatterplot
- Randomly sample rows from a dataset
- Write a function to perform some repetitive work
- Use ggplot to create visualizations
- Understand pipes ('%>%') and how to use them
- Create and understand a linear model

Example Topics

Basic data foundations:

- Understand measure of dispersion, law of large numbers, central tendency
- What is an independent and dependent variable (within a specific dataset)

Vectors:

- How to create a vector with a specified list of elements (numbers, etc).
- Add a value to all the elements within the vector.
- Apply basic functions on a vector of numbers (ex. sum the numbers).

Data frames:

- Understand data frames and be able to write R code that outputs a column, a row, or a specific element in the data frame.
- Add or remove a columns or rows from a data frame.
- Understand how to access the data frame

Example Topics

Functions:

- Understand the quantile function - what is it, why to use it, how to use it
- Understand the sample function - what is it, why to use it, how to use it
- Understand the replicate function - what is it, why to use it, how to use it
- Understand the histogram function - what is it, why to use it, how to use it
- How to use ggplot (and create maps)
- Be able to create and use a function

Linear models:

- How to create a basic model and how to measure quality of the model

Study Strategy

- General principle 1: **Use spaced practice, not massed practice**; make a list of six big things you want to study and do one each day, leaving a sleep cycle in between; put the most important elements first and reinforce those on later days.
- General principle 2: spend at least **an hour without interruptions** in each of the six study sessions.
- Write / Run at least one piece of **code from each chapter** to check your understanding.
- In the **textbook, reread/review assigned chapters**
- For each chapter, **make written notes** on paper of the five-to-seven most pertinent concepts or techniques shown in the chapter.
- Review homeworks; open each file and **run the code line by line**; make sure you correct your code and understand what it does (if needed, "redo" the homework from scratch – it's the best way to learn)

Midterm Format

- This will be a open-book, open-notes, laptop-based exam.
- There will be different versions of the exam.
- The exam is designed to be completed in 45 minutes, but you may take up to 80 minutes.
- The exam will contain multiple choice questions
- The Exam will also contain R coding questions - You will be asked to read R code and **write R code**

Objectives for the Week

- Explain the difference between supervised and unsupervised machine learning
- Be able to create a prediction "model" for a given data set
- Use the model to interpret/understand the data
- Identify relationships between variables
- Visually and statistically determine the influence of one or more variables over another variable
- Interpret output of a linear regression model
- Evaluate how least-squares formula can be used to predict dependent variable
- Utilize R code to create linear models

Machine Learning Overview

Trad Statistical Analysis vs. Machine Learning / Data Science

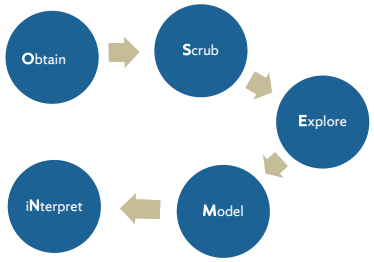
Statistical Analysis

- Often focused on confirming a hypothesis in light of some theory
- Data arrives with provenance, screening & cleaning are trivial
- Measurement strategy planned in advance, thoughtful choice of constructs and scales
- Metric variables often follow the normal distribution

Data Mining

- More exploratory, looking for novel patterns and needles in haystacks
- Data may arrive with little provenance, may require extensive diagnosis and restructuring
- No planned measurement strategy: work with what you have available
- Mixture of textual, categorical, & metric, no presumption of distribution

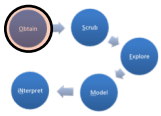
The Machine Learning Process OSEMN (Rhymes With Possum)



Hilary Mason and Chris Wiggins, 2010 - http://www.dataists.com/2010/09/a-taxonomy-of-data-science

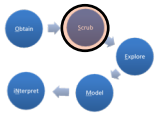
OSEMN Phase: Obtain

- Collect data
  - Deal with file formats and how to read data
  - Query databases or data repositories
  - Extract data from other sources
  - Generate data (e.g., surveys, sensors)



OSEMN Phase: Scrub

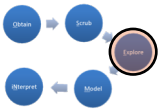
- Get data into a useable format structure
  - Filter/subset
  - Extract attributes
  - Replace/handle missing, illegal, and anomalous values
  - Transform/bin/code data attributes



Typically 50% or more of the work is in this phase

OSEMN Phase: Explore

- Explore patterns and trends
  - Start to "understand" the data, detect outliers
  - Visualize attributes (e.g., scatter plots, histograms)
  - Calculate/visualize descriptive statistics (e.g., distributions)
  - Feature selection (what attributes are most interesting)

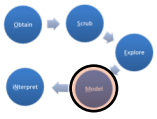


OSEMN Phase: Model

- Build predictive models (machine learning)
  - Type of modeling:
    - Supervised (classification, regression)
    - Unsupervised (e.g., clustering)

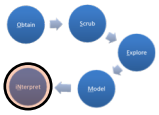
  → *Will discuss shortly*

  - Model tuning and comparing candidate models



OSEMN Phase: Interpret

- Understand/explain the results
  - Draw conclusions from and data models
  - Evaluate the meaning of results
  - Communicate the results
  - Ensure actionable insight



Question

How is the OSEMN process different from or the same as the overall data science process?

Machine Learning Techniques

Unsupervised learning includes a variety of machine learning techniques that do not use a criterion or dependent variable, but rather look for patterns solely among "independent" variables.

Supervised learning is parallel in concept to the predictive statistical techniques used by many social science researchers, such as linear regression, but without the restriction of only exploring linear relationships.

Another form of learning is known as "**reinforcement learning**." Evolution of these models depends on success/failure cues from real or simulated environments.

Machine Learning in a Nutshell

- "Supervised" refers to the idea that there is a **criterion** used during an algorithm training phase
- A supervised algorithm uses a set of input variables to optimize the prediction of an outcome variable
- Supervised data mining is closely connected with machine learning techniques: What is the difference?

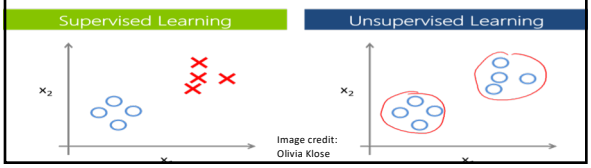


Image credit: Olivia Klose

Supervised Data Mining

Question

What is the meaning of the word "supervised" in the context of Machine Learning?

25

Two types of prediction problems

Classification problems

- **Predicting membership in two or more categories** based on a set of predictors (model features)
  - **Examples of criteria to predict**: Medical diagnosis, Employment outcomes (e.g., hiring), Financial outcome (e.g., customer default on a loan)

Regression problems

- **Predicting a continuous numeric outcome** based on on a set of predictors
  Examples of criteria to predict: sales volume, employee engagement, customer satisfaction

Traditional Approaches:

Classification problems
Logistic regression
Discriminant analysis

Regression problems
Ordinary least squares regression
linear models
Generalized linear models
Lasso regression

Supervised Learning Example

Train a machine learning algorithm to predict the weather

- Collect weather data over a period of time
  - Sunny, cloudy
  - Temperature
  - Barometer
  - Wind speed and direction
- Train a machine learning algorithm with these collected variables
- Collect more weather data and predict the weather via our trained algorithm
  - Classification would be predicting good weather or bad weather
  - Regression would be predicting the temperature
- Then validate the prediction

The Modeling Process

- Use a substantial number of training cases
  - The machine learning algorithm can use that data to build a model
- Use the results of this process (i.e., the model) on test data set to determine how well algorithm performed
  - Validate the model on new data
- The result is a model that can be used for prediction
  - Predict data that was not used during training
  - Predict future instances of data

*Note: The model is **not always useful** for explaining results to managers.*

Some algorithms produce results that are not easy to interpret or visualize or explain how they work; for some algorithms there is no output that is like a regression coefficient.

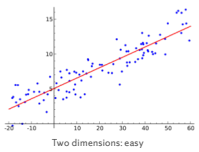
Linear Modeling Overview

School of Information Studies
SYRACUSE UNIVERSITY
ischool.syr.edu

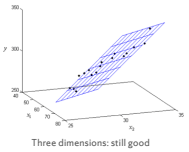
Linear Modeling

"Does X influence Y?" .

Linear Relationships in 2+ Dimensions



Two dimensions: easy



Three dimensions: still good

?

Four dimensions?

Image credit: Sewaqu - Public Domain, https://commons.wikimedia.org/w/index.php?curid=11967659

4D Cube: Tesseract

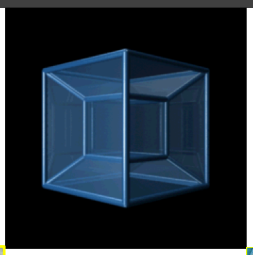
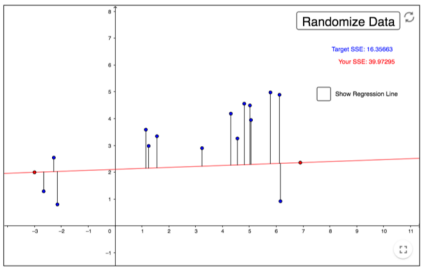


Image credits: Wikipedia, Warner Bros., Disney

AN EXAMPLE



https://www.geogebra.org/m/xC6zq7Zv

A Simple Model to Predict GPA

Predicting a student's semester GPA using three pieces of information?

Data collection:

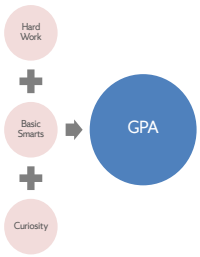
- Measure three predictors, using tests or surveys with multiple students (e.g., n=120)
- Then at the end of the semester we would have four pieces of information:
  - The criterion (GPA)
  - The three predictors (hard work, basic smarts, & curiosity).

Model creation:

- Using linear regression, calculate coefficients for each predictor to make an equation:

GPA = (B1 * HardWork) +
(B2 * BasicSmarts) + (B3 * Curiosity)

[Y = MX + b] or Y = BX + E

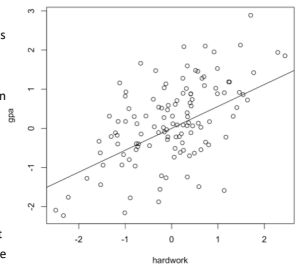


Regression Terminology

- Criterion/**dependent variable**: what we are trying to predict
- Predictor/**independent variable**: one of the variables we use to predict the criterion; there are usually multiple predictors
- Coefficients/weights: the strength of prediction for each predictor; sometimes also called B-weights (or the standardized version is called a beta-weight)
- Regression equation: the result of the regression analysis in the form of an algebraic equation
  - $\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + \ldots$
  - Y-hat is the predicted Y, subscripts on Bs and Xs refer to predictor number

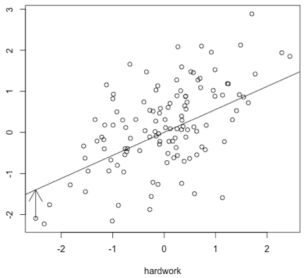
Finding the Best Fitting Line

- The diagram at left visualizes a cloud of points representing scores on our "hardwork" predictor and our GPA criterion/outcome.
- The other two predictors are still in the data, we're just ignoring them for the moment.
- How can we decide the best slope and intercept for the line? What does it mean to have a good fit to these points?
- The lm() procedure uses the "least squares criterion" to select the one best value for the slope.

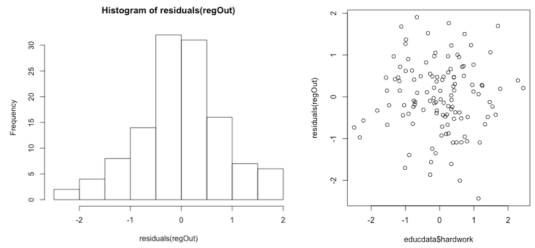


Prediction Errors: The Least Squares Criterion

- Each point **on the line** represents our prediction of Y given a certain value of X.
- To the extent that an observed value does not fall on the line, we have a prediction error: the vertical (Y-axis) distance between the point and the line. See the arrow at the lower left.
- For the best fitting line, all of the positive errors (too high) and all of the negative errors (too low) will sum to zero.
- When we square all prediction errors and sum them, we get a measure of the overall error in the model. The lm() function chooses a slope that makes this value as small as possible.



Residuals Are Errors of Prediction



A Worked Example with R

School of Information Studies
SYRACUSE UNIVERSITY
ischool.syr.edu

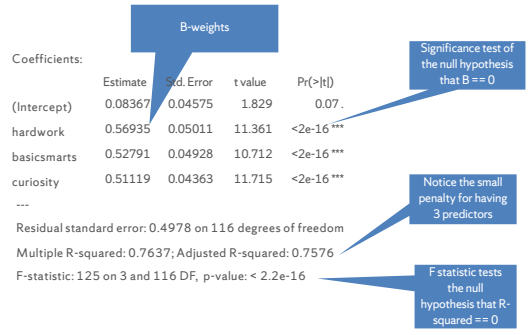
GPA Model With All Three Predictors

```
regOut3 <- lm(gpa ~ hardwork + basicsmarts + curiosity, data=educdata)

summary(regOut3)
Call:
lm(formula = gpa ~ hardwork + basicsmarts + curiosity, data = educdata)
Residuals:
     Min       1Q   Median       3Q      Max
-1.02063 -0.37301  0.00361  0.31639  1.32679
```

Residuals look good: median should be near zero and symmetric

Significance Tests on Predictors



Interpreting the Model

- Adjusted R-squared value 0.7576
- Known as the coefficient of determination
- The proportion of the variation that is accounted for in the dependent variable by the whole set of independent variables.
- The closer to 1.0 , the greater the influence the independent variable has on predicting the value of the dependent variable.
- The R-squared value of 0.7576 indicates that the three factors account for 75.76% of GPA.

Evaluating Regression Results

1. Examine the F-test on R-squared, **if the p-value is < 0.05, it is *significant* and OK to proceed**
2. Examine the **adjusted R-squared value**: This translates as, "the proportion of variance in Y that is accounted for by all Xs working together."
3. Look for ***significant* predictors** (but usually ignore the intercept) by seeing *, **, or ***
4. Look in the "Estimate" column for the value of the slope on a significant predictor: One unit change in X causes that much change in Y

Another Example

Car Maintenance

- We manage a "fleet" of cars
  - Cars get replaced every three years
  - Have information on:
    - Past repairs
    - Miles driven
    - # of oil changes during past three years

→ How often to change the oil?

→ Can we build a model to predict repair costs?

Question

In this new example, which are independent and which are dependent variables? Why?

	oilChanges	repairs	miles
1	3	300	100500
2	5	300	116000
3	2	500	136000
4	3	400	110500
5	2	700	150500
6	6	420	117000
7	6	100	89500
8	4	290	99500
9	3	475	100500
10	2	620	120500
11	0	600	106000
12	8	150	115000
13	7	200	104000
14	8	50	98500

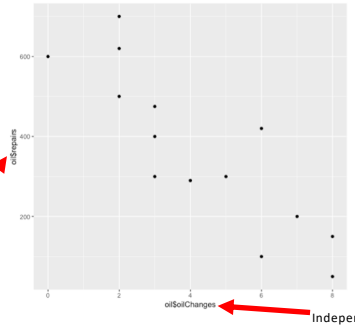
Small Data Set, n=14

```
oilChanges <- c(3,  5,  2,  3,  2,  6,  6,  4,  3,  2,  0,  8,  7,  8)
repairs <-    c(300,300,500,400,700,420,100,290,475,620,600,150,200,50)
miles <-      c(100500,116000,136000,110500,150500,117000,89500,99500,100500,
                120500,106000,115000,104000,98500)
oil <- data.frame(oilChanges, repairs, miles)
View(oil)
```

	oilChanges	repairs	miles
1	3	300	100500
2	5	300	116000
3	2	500	136000
4	3	400	110500
5	2	700	150500
6	6	420	117000
7	6	100	89500
8	4	290	99500
9	3	475	100500
10	2	620	120500
11	0	600	106000
12	8	150	115000
13	7	200	104000
14	8	50	98500

qplot(oil$oilChanges, oil$repairs)

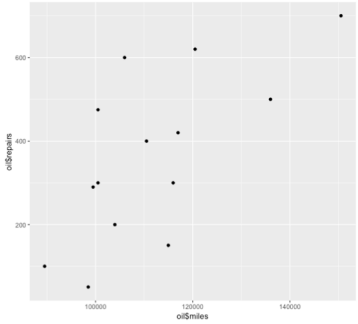
Note: qplot is in the ggplot package



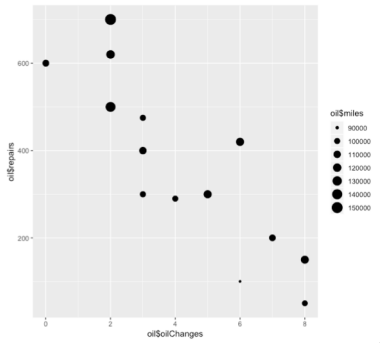
Dependent Variable

Independent Variable

qplot(oil$miles, oil$repairs)



qplot(oil$oilChanges, oil$repairs, size=oil$miles)



lmOut <- lm(repairs ~ oilChanges + miles, data=oil) summary(lmOut)

```
Call:
lm(formula = repairs ~ oilChanges + miles, data = oil)

Residuals:
    Min      1Q  Median      3Q     Max
-115.39  -38.76  -19.61   31.48  130.62

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.716647 172.343530   0.207  0.83961
oilChanges  -57.256143   8.945717  -6.400 5.08e-05 ***
miles         0.005104   0.001377   3.706  0.00346 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.71 on 11 degrees of freedom
Multiple R-squared:  0.8828,  Adjusted R-squared:  0.8615
F-statistic: 41.44 on 2 and 11 DF,  p-value: 7.56e-06
```

Interpreting the Model

- R-squared value 0.8615
- Known as the coefficient of determination
- The proportion of the variation that is accounted for in the dependent variable by the whole set of independent variables.
- The closer to 1.0 , the greater the influence the independent variable has on predicting the value of the dependent variable.
- The R-squared value of 0.8615 indicates that the oil changes accounts for 86.15% of the cost of repairs.

What's an Oil Change Worth?

- The initial overall model was significant (F-test) and the oilChange predictor was significant with a slope of -57
- In the initial model, every oil change reduces the dollar amount of repairs by $57

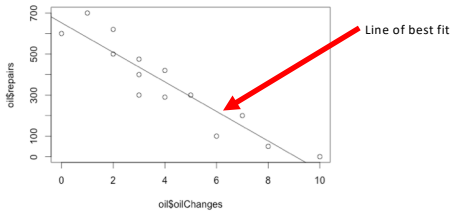
```
Call:
lm(formula = repairs ~ oilChanges + miles, data = oil)

Residuals:
    Min      1Q  Median      3Q     Max
-115.39  -38.76  -19.61   31.48  130.62

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.716647 172.343530   0.207  0.83961
oilChanges  -57.256143   8.945717  -6.400 5.08e-05 ***
miles         0.005104   0.001377   3.706  0.00346 **
```

Looking at the "abline"

abline(lmOut)



The model suggests that we should do as many oil changes as possible.
→ it predicts very low (almost 0) repairs if we do 9 or more oil changes, but about $680 if we do no oil changes.

Predict New Values

```
> oilChanges <- c(1,      1,      2,      2,      3,      3)
> miles <-    c(10000, 20000, 20000, 40000, 40000, 80000)
> oilPred <- data.frame(oilChanges, miles)
> predict(lmOut1, oilPred)
       1        2        3        4        5        6
29.50321 80.54592 23.28977 125.37519 68.11905 272.28987
```

Questions

Is this an accurate model?

Working Through a Refined Example

Cost of Oil Change

→ How "model" the cost?

→ What might be some ranges of the cost?

Include the Cost of an Oil Change

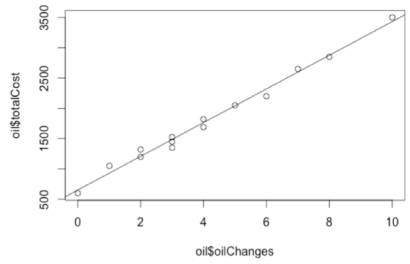
```
# What if oil changes cost $350 each?

oil$oilChangeCost <- oil$oilChanges * 350
oil$totalCost <- oil$oilChangeCost +
                                    oil$repairs

m <- lm(formula=totalCost ~ oilChanges,
        data=oil)
```

What is oil changes cost $350 each?

```
plot(oil$oilChanges, oil$totalCost)
abline(m)
```



Predict New Values

```
oilChanges <- c(1,      1,      2,      2,      3,      3)
miles <-    c(10000, 20000, 20000, 40000, 40000, 80000)
oilPred <- data.frame(oilChanges, miles)
predict(m, oilPred)
```

1	2	3	4	5	6
379	430	723	825	1118	1322

Question

How accurate is the model?

- Did we have all the facts?
- Did we have all the data?