CWRU DSCI351-451: Predictive Analytics

Roger H. French, JiQi Liu

25 October, 2018

Contents

9.2.0.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - OIS 6.1-2 for today 10/25/2018
  - OIS 7 for next Tuesday 10/30/2018
- Homeworks
  - HW5 Inference due Tuesday 10/30/2018
- Data Science Projects:
  - Project 2 Time Series Analysis due Thursday 11/1/2017
- 451 SemProjects:
  - SemProjects Report Out #2, in class, Tuesday 10/30/2018
- Friday Comm. Hour
  -

9.2.0.2 Textbooks

- Peng: R Programming for Data Science

- Peng: Exploratory Data Analysis with R
- Open Intro Stats, v3
- Wickham: R for Data Science
- Hastie: Intro to Statistical Learning with R

9.2.0.3 Syllabus

Day:Date	Foundation	Practicum	Reading	Due
w1a:Tu:8/28/18	ODS Tool Chain	R, Rstudio, Git		
w1b:Th:8/30/18	Setup ODS Tool Chain	Bash, Git, Twitter	PRP4-33	HW1
w2a:Tu:9/4/18	What is Data Science	OIS:Intro2R	PRP35-64	**HW1 Due**
w2b:Th:9/6/18	Data Analytic Style, Git	451SempProj, Git	PRP65-93, OI1-1.9	HW2
w3a:Tu:9/11/18*	Struct. of Data Analysis	ISLR:Intro2R, Loops	PRP94-116, OIS3	**HW2 Due**
w3b:Th:9/13/18*	OIS3 Intro to Data	GapMinder, Dplyr, Magrittr		
w4a:Tu:9/18/18	OIS3, Intro2Data part 2, Data	EDA: PET Degr.	EDA1-31	Proj1
w4b:Th:9/20/18	Hypothesis Testing	GGPlot2 Tutorial	EDA32-58	HW3
w5a:Tu:9/25/18	Distributions	SemProj RepOut1	R4DS1-3	**HW3 Due**
w5b:Th:9/27/18	Wickham DSCI in Tidyverse	SemProj RepOut1	R4DS4-6	**SemProj1,**
w6a:Tu:10/2/18	OIS Found. of Inference	Inference	R4DS7-8	**Proj1 Due**
w6b:Th:10/4/18		Midterm Review	R4DS9-16 Wrangle	
w7a:Tu:10/9/18*	Summ. Stats & Vis.	Data Wrangling		
w7b:Th:10/11/18*	**MIDTERM EXAM**			HW4
w8a:Tu:10/16/18	Numerical Inference	Tidy Check Explore	OIS4	**HW4 Due**
w8b:Th:10/18/18	Algorithms, Models	Pairwise Corr. Plots	OIS5.1-4	Proj 2, HW5
Tu:10/23	**CWRU FALL BREAK**		R4DS17-21 Program	
w9b:Th:10/25/18	Categorical Infer	Predictive Analytics	OIS6.1,2	
w10a:Tu:10/30/18	SemProj	SemProj	OIS7	**SemProj2 HW5 Du**
w10b:Th:11/1/18	Lin. Regr.	Lin. Regr.	OIS8	**Proj.2 due**
w11a:Tu:11/6/18	Inf. for Regression	Curse of Dim.	OIS8	Proj 3
w11b:Th:11/8/18	Model Accuracy	Training Testing	ISLR3	HW6
w12a:Tu:11/13/18	Multiple Regr.	Mul. Regr. & Pred.	ISLR4	**HW6 due**
w12b:Th:11/15/18	Classification		ISLR6	
w13a:Tu:11/20/18	Classification	Clustering	ISLR5	**Proj 3 due**
Th:11/22/18	**THANKSGIVING**			Proj 4
w14a:Tu:11/27/18	Big Data	Hadoop		
w14b:Th:11/29/18	InfoSec	VerisDB		**SemProj3**
w15a:Tu:12/4/18	SemProj ReportOut3			
w15b:Th:12/6/18	SemProj ReportOut3			**Proj4**
	FINAL EXAM	**Monday12/17, 12:00-3:00pm**	Olin 313	**SemProj4 due**

MidTerm Results

In Canvas Grades, the "Total" column is in %

- DSCI 351 graded out of 100 points
  - 35 points done through the MidTerm

- DSCI 451 graded out of 140 points
  - 45 points done through the MidTerm
  - SemProj is 40 points total

9.2.0.3.1 MidTerm DSCI 351, 351M Grades

Out of 40 points

- A > 91 %, 32 points
- B > 80%, 28 points
- C < 80%, 28 points

9.2.0.3.2 MidTerm DSCI 451 Grades

Out of 50 points

- A > 89%, 40 points
- B > 80%, 36 points
- C < 80%, 36 points

9.2.0.4 SemProj Report Out 2, next Tuesday 10/31/2017 in class

- Submit your report Rmd, pdf and/or Rpres to the Assignment page on Canvas
- 8 minutes presentation
- We'll have DSCI 352/452 also present

9.2.0.4.1 Basic steps we use to construct a data analysis.

Modified from Jeff Leek's slides

- (available in your repo in 17f-dsci351-451-prof/3-readings/)

9.2.0.4.2 SemProj. Part a) Define Question

- Background on the research area and critical issues
- Define the question
- Define the ideal data set
- Determine what data you can access
- Define critical capabilities and identify packages you will draw upon
- Obtain the data, define you target data structure
- Clean and tidy the data

9.2.0.4.3 SemProj Part b) Cleaning and EDA

- Write you databook, defining variables, units and data structures
- Data visualization and exploratory data analysis
- Observations of trends and functional forms
- Power transformations
- Validate with reference to domain knowledge
- Evaluate the types of Modeling Approaches to take

9.2.0.4.4 SemProj Part c) Modeling and Statistical Learning

- Types of modeling to try
- Statistical prediction/modeling
- Model selection
- Cross-validation, Predictive R2
- Interpret results
- Challenge results

9.2.0.4.5 SemProj Part d) Present your final models and learnings

- Present your results
- Present reproducible code
- Comparison to other modeling approaches in the literature

9.2.0.5 Supervised and Unsupervised Learning

Two broad families of algorithms will be covered:

- Unsupervised learning algorithms
- Supervised learning algorithms

9.2.0.5.1 Unsupervised learning

In unsupervised learning,

- the algorithm will seek to find the structure that organizes unlabelled data.

9.2.0.5.2 Supervised learning

In supervised learning,

- we know the class or the level of some observations of a given target attribute.

9.2.0.6 Classification and Regression Problems

There are basically two types of problems that predictive modeling deals with:

- Classification problems
- Regression problems

9.2.0.6.1 Classification

In some cases, we want to predict which group an observation is part of.

Here, we are dealing with a quality of the observation.

9.2.0.6.2 Regression

In other cases, we want to predict an observation's level on an attribute.

Here, we are dealing with a quantity, and this is a regression problem.

9.2.0.7 The critical role of domain knowledge

- in modeling and prediction

Domain knowledge informs and is informed by data understanding.

- The understanding of the data
  - then informs how the data has to be prepared.

The next step is data modeling,

- which can also lead to further data preparation.

Data models have to be evaluated,

- and this evaluation can be informed by field knowledge,
  - which is also updated through the data mining process.

Finally,

- if the evaluation is satisfactory,
  - the models are deployed for prediction.

9.2.0.8 Caveat: For Predictive Analytics

Of course, predictions are not always accurate,

- and some have written about the caveats of data science.

What do you think about the relationship between

- the attributes titled Predictor and Outcome on the following plot?

It seems like there is a relationship between the two.

- For the statistically inclined,
  - I tested its significance:
    - r = 0.4195, p = .0024.
- The value p is the probability of obtaining a relationship of this strength or stronger
  - if there is actually no relationship between the attributes.
- (This is the p-value of hypothesis testing, if $p<0.05$
  - typically we assert we can reject the null hypothesis)
- We could conclude that the relationship between these variables
  - in the population they come from is quite reliable,
  - **right?**

9.2.0.8.1 No: Lets think about this

Believe it or not,

- the population these observations come from
  - is that of randomly generated numbers.
- We generated a data frame of 50 columns
  - of 50 randomly generated numbers.
- We then examined all the correlations (manually)
  - and generated a scatterplot of the two attributes
  - with the largest correlation we found.

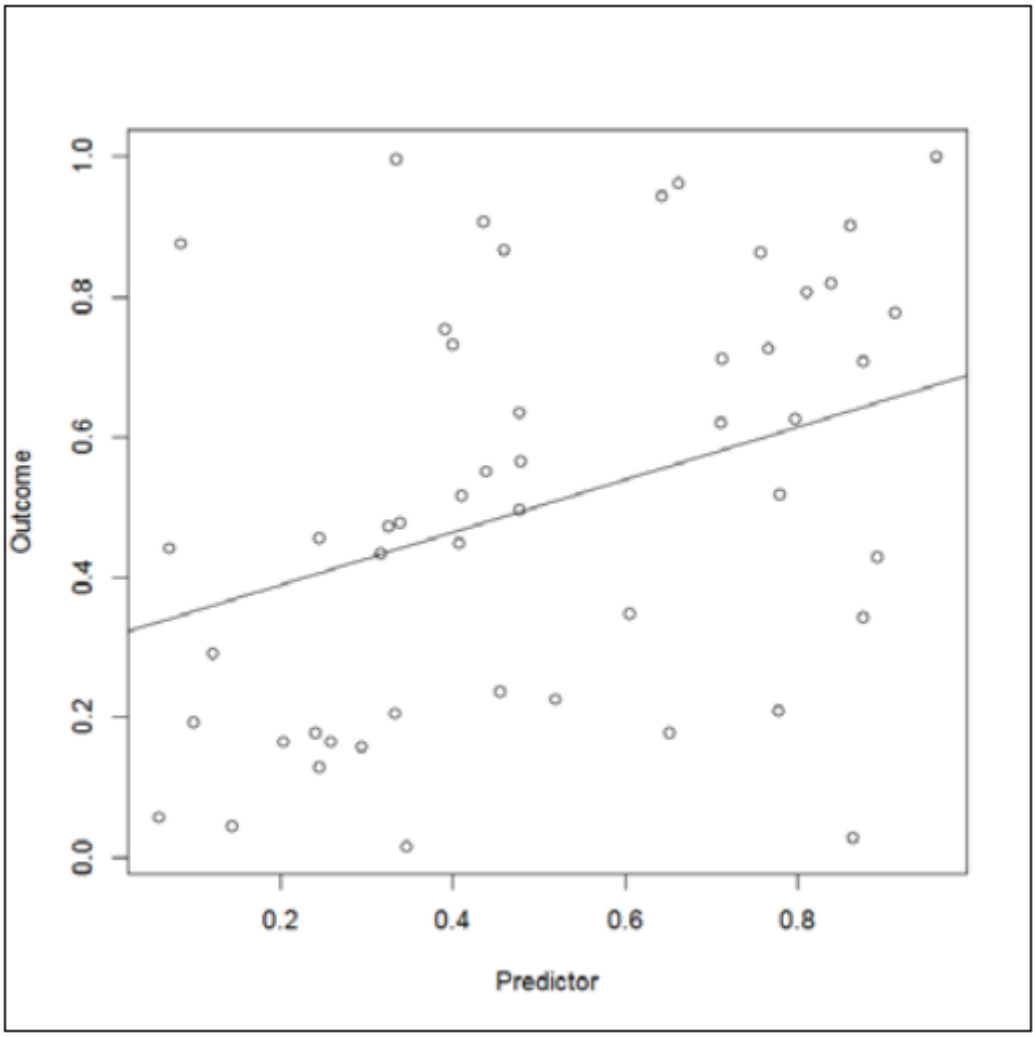


Figure 1: Relationship between Predictor & Outcome

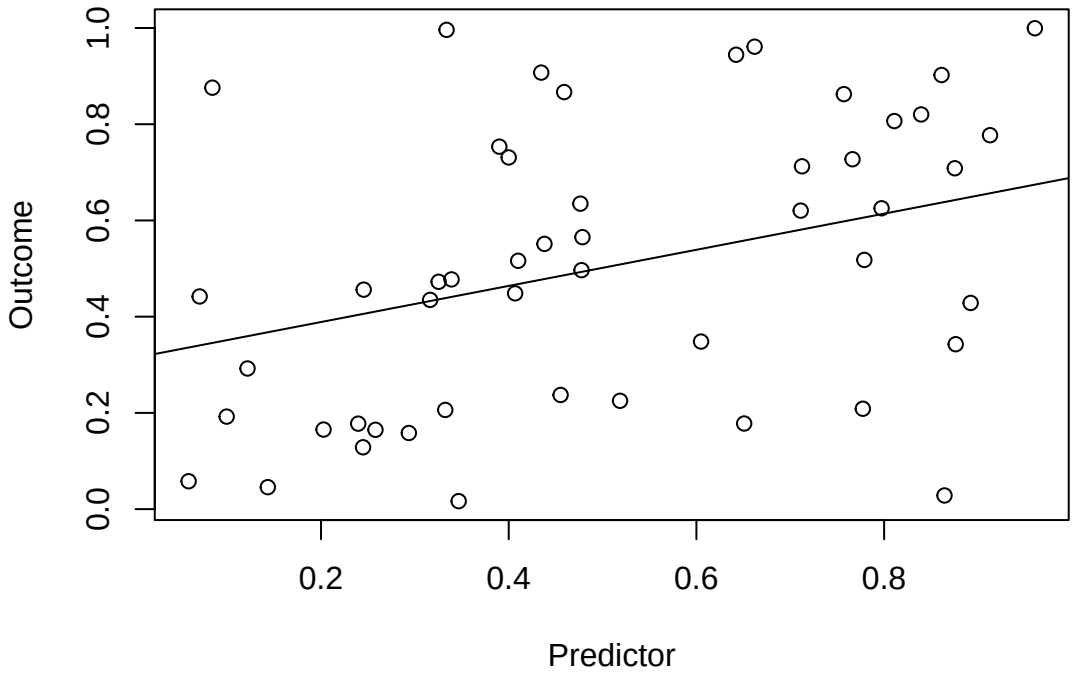
9.2.0.8.2 The code is provided here,

We'll use runif()

- help(runif)
  - The Uniform Distribution
  - Description
  - These functions provide information about the uniform distribution on the interval from min to max.
    - dunif gives the density,
    - punif gives the distribution function
    - qunif gives the quantile function and
    - runif generates random deviates.

```
set.seed(1)
DF = data.frame(matrix(nrow = 50,ncol = 50))
for (i in 1:50) {
  DF[,i] = runif(50)
}

plot(DF[[2]],DF[[16]], xlab = "Predictor", ylab = "Outcome")
abline(lm(DF[[2]]~DF[[16]]))
```




```
cor.test(DF[[2]], DF[[16]])
```

```
##
##  Pearson's product-moment correlation
##
## data:  DF[[2]] and DF[[16]]
## t = 3.2023, df = 48, p-value = 0.002421
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1598919 0.6249314
## sample estimates:
##       cor
```

```
## 0.4195666
```

- in case you want to check it yourself
- line 1 sets the seed so that you find the same results as we did,
- line 2 generates to the data frame,
- line 3-5 the for loop fills it with random numbers, column by column,
- line 7 generates the scatterplot,
- line 8 fits the regression line, and
- line 9 tests the significance of the correlation:

Normally we reject the null with a p-value of <0.05

- i.e. we'll be wrong 5% of the time
  - in a set of 20 trials

Here we did 50 trials

- And cherry picked the best correlation
  - But its all randomly generated numbers
  - There is no predictive or causal relationship
- And we'd only recognize this if we consider
  - That our p-value is reported for 1 trial
  - But we run many trials

9.2.0.8.3 Bonferroni Correction for multiple comparisons

How could this relationship happen given that the odds were 2.4 in 1000 ?

- Well, think of it;
  - we correlated all 50 attributes 2 x 2,
  - which resulted in 2,450 tests
    - (not considering the correlation of each attribute with itself).
- Such spurious correlation was quite expectable.

The usual threshold below which we consider a relationship significant is p = 0.05.

- This means that we expect to be wrong once in 20 times.
- You would be right to suspect that there are other significant correlations
  - in the generated data frame (there should be approximately 125 of them in total).
- This is the reason why we should always correct the number of tests.
- In our example,
  - as we performed 2,450 tests,
  - our threshold for significance
  - should be 0.0000204 (0.05 / 2450).
- This is called the Bonferroni correction.

9.2.0.9 Overfitting: The need for Training and Testing Datasets

Spurious correlations are always a possibility in data analysis

- and this should be kept in mind at all times.

A related concept is that of overfitting.

- Overfitting happens, for instance,
  - when a weak classifier bases its prediction on the noise in data.
- We will discuss overfitting when discussing
  - Training datasets for fit a model to
  - Testing datasets for evaluating the goodness of fit

    - when using various types of cross-validation
  - And when evaluating $Predictive, Adjusted, R^2$

9.2.0.10 Lets get some basic ideas for background

- Standard Error
- Margin of Error
- Z score
  - Used in Z-test
  - The analog of Students t-test
- $Adjusted R^2$
- PRESS, predicted residual error sum of squares statistic

9.2.0.11 Dice statistics

- Weldon's Dice
  - Weldon's Dice Automated
  - Weldon's Dice Revisited, In readings Kemp and Kemp - 1991 - Weldon's Dice Data Revisited.pdf
- Reproducing Weldon's Dice Experiment
- Fair Dice (Part 2) - Numberphile

9.2.0.12 Citations

- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2014. http://www.R-project.org/.
- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. 1st ed. 2013, Corr. 5th printing 2015 edition. Springer Texts in Statistics. New York: Springer, 2013.
- Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. OpenIntro Statistics: Third Edition. 3 edition. S.l.: OpenIntro, Inc., 2015.
- Mayor, Eric. Learning Predictive Analytics with R. Packt Publishing - ebooks Account, 2015.