Data Driven Computer Security

Executive Education Course

Data Analysis

Spring 2015

Introductions

About me: John Gasper, PhD

- Time at CMU-Qatar: Aug-2010 – present.
- Courses:
  - Regression and Forecasting
  - Stochastic Modeling and Simulation
  - Decision Analysis
  - Game Theory for Business (strategic decision making)
  - Behavioral Decision Making (Psychology of decision making)

- Full disclosure: I’m not a computer security expert.

Outline

Day 1: Why do we care / what is the data

Day 2:

- Correlation and summary statistics
- Regression Modeling
- Logistic Regression and prediction
- Potential uses of machine learning and classification.

Understanding Data and Analysis

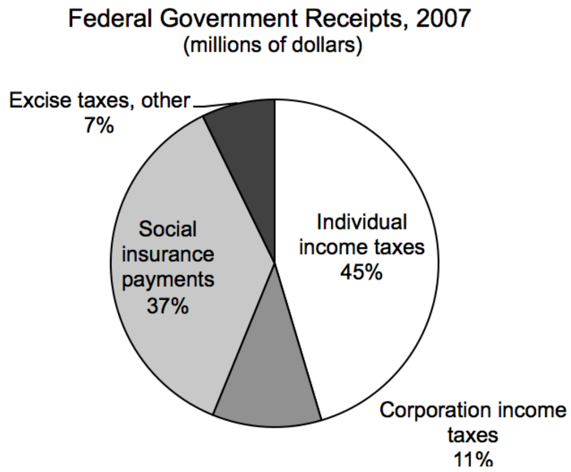
- Everyone – from upper level executives to analysts – will make better decisions with a better understanding of data and data analysis.

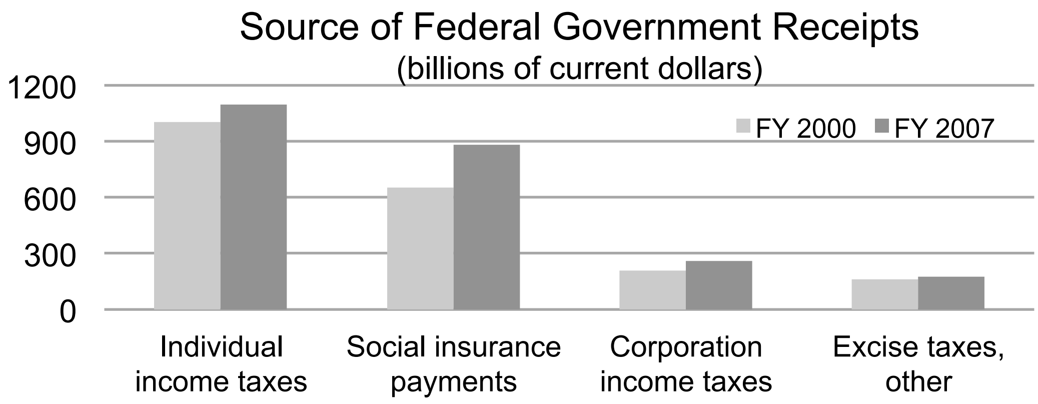
- What kinds of data do you deal with or collect in your organizations?

Describing Data

Graphical displays of data.

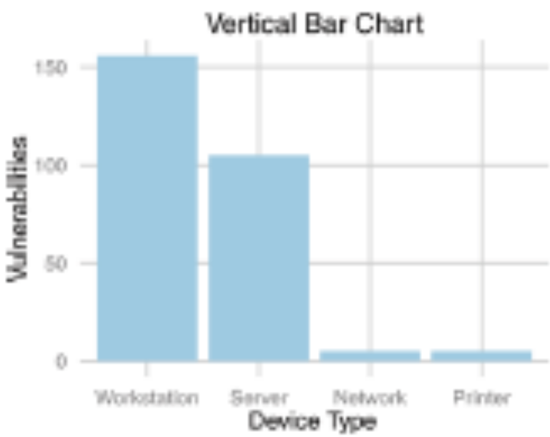
- Present meaningful data
- Define data unambiguously
- Do not distort the data – no 3D effects, please.
- Present the data efficiently.

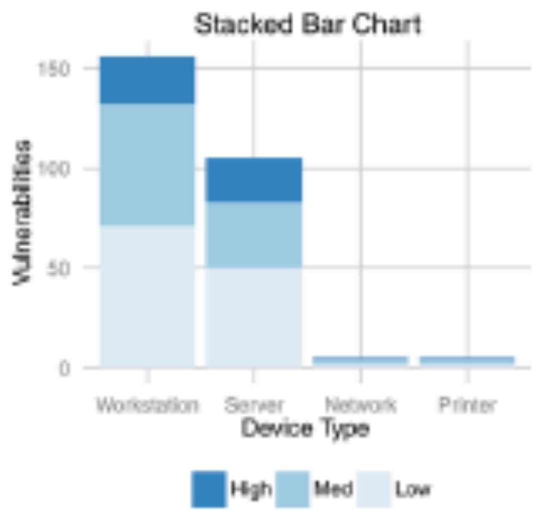


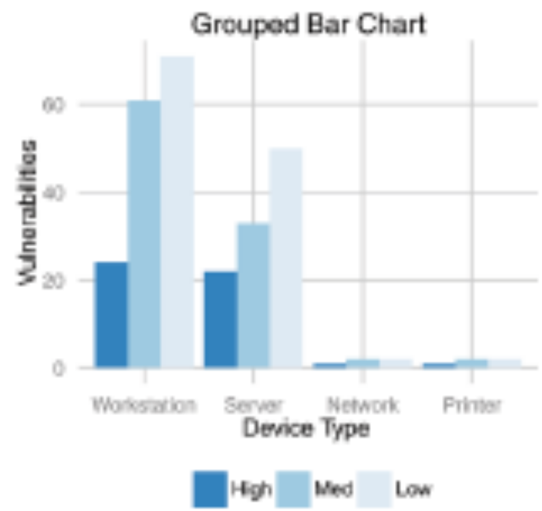


Describing Data

Convey the appropriate information

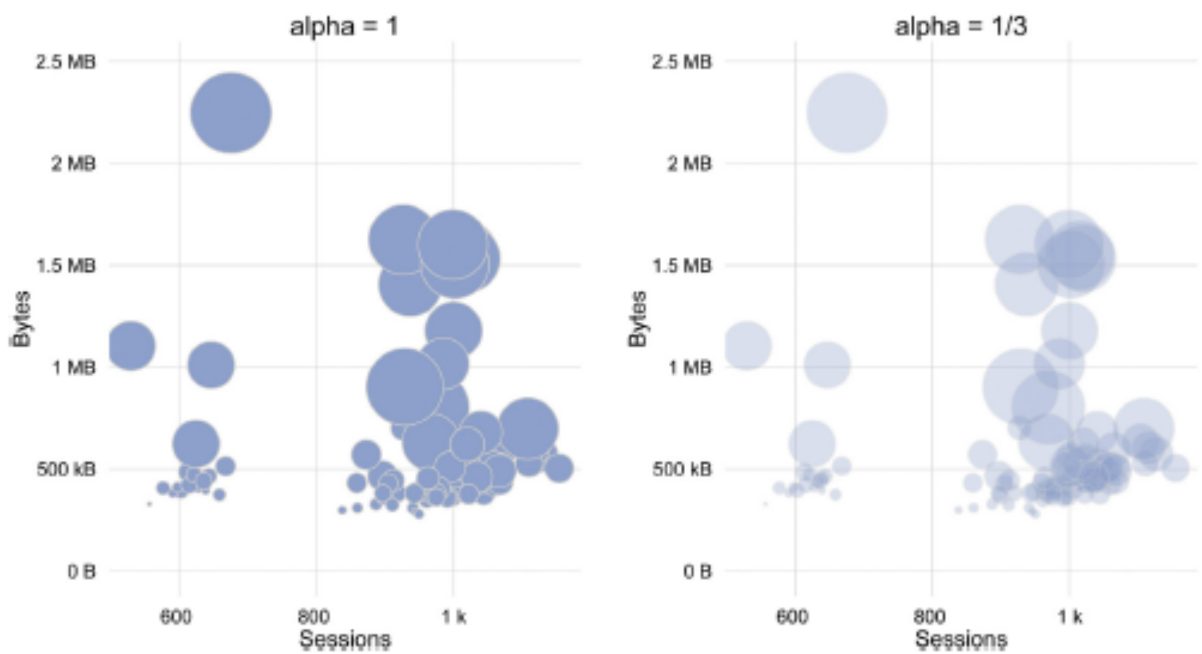






Describing Data

Leverage opacity or colors to highlight intensity



- 8 hours of firewall data for networking devices split into 5-minute totals.
- x-axis = number of network sessions & y-axis= number of bytes
- Size of each bubble" is proportional to the packet count.

Describing Data

Strive to move beyond summarizing a single variable.

- Think about relationships ***between*** variables

Describing Data

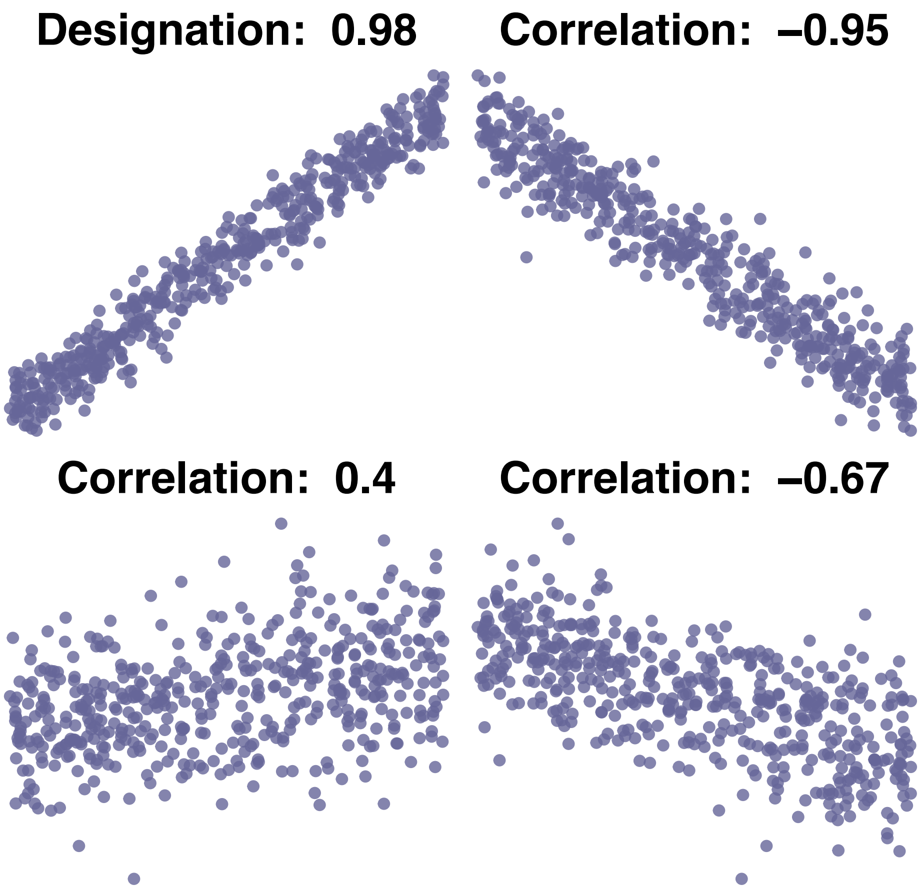
<u>Graphical summaries of data can be incredibly useful</u>

- Benefits
  - Understanding how data are dispersed. Not just the average amount, but what are possible outcomes?
  - A lot of intuition about relationships between variables

- Dangers
  - Sometimes oversimplifies the relationships
  - Can be misleading

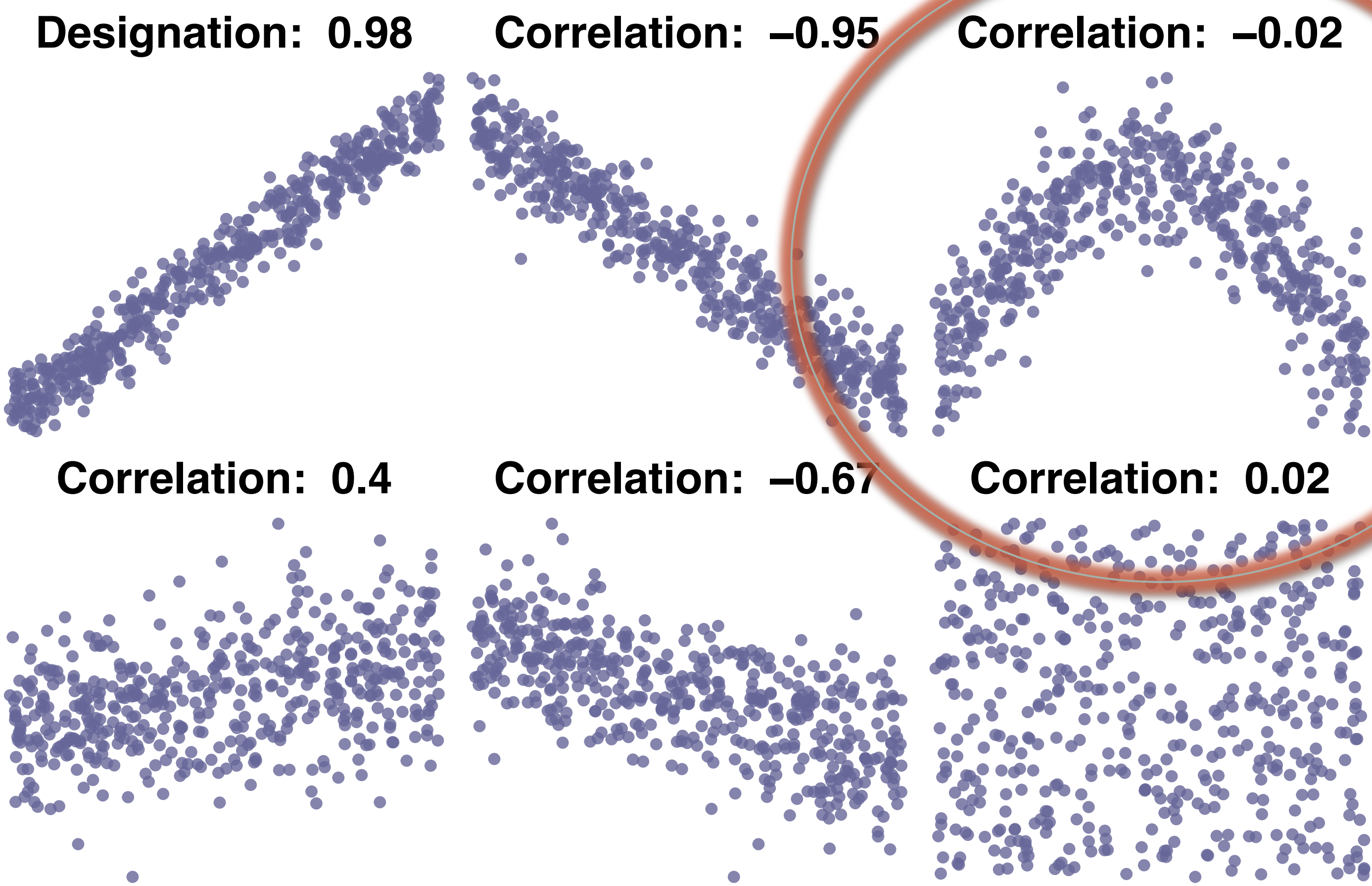
Describing Data: Correlation

One of the most used and fundamental ways to describe the relationships between variables is correlation: $-1 < \rho < 1$



Describing Data: Correlation

- One of the most used and fundamental ways to describe the relationships between variables is correlation: $-1 < \rho < 1$



Describing Data: Correlation

But don’t be fooled by significant correlations.

- Bivariate scatterplots are good starting places but can be misleading

Also remember that Correlation is NOT Causation.

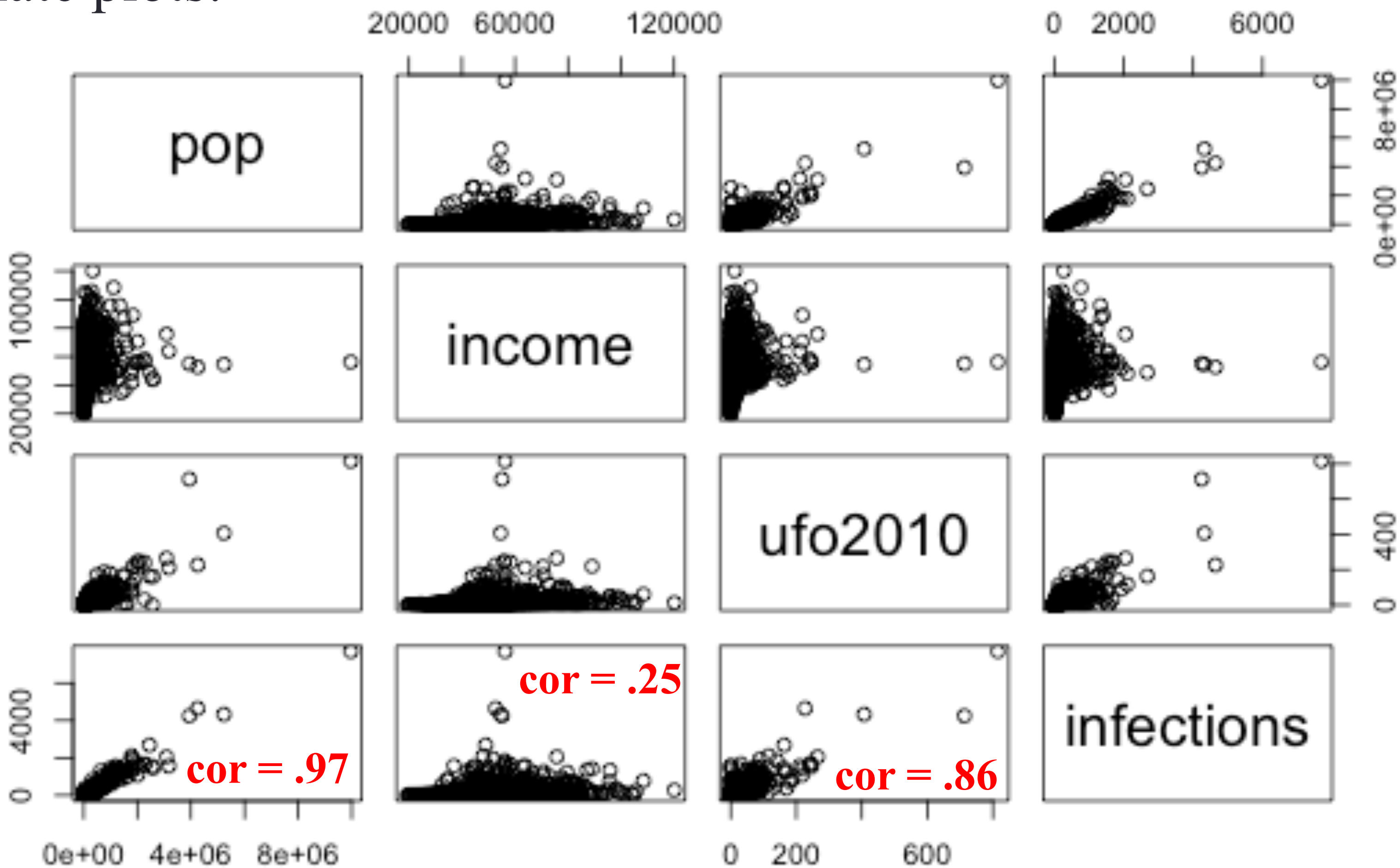
- Just because two things are correlated doesn’t mean that one causes the other (as much as we might like it to)

Example: ZA infections

- USA Zero Access infection data

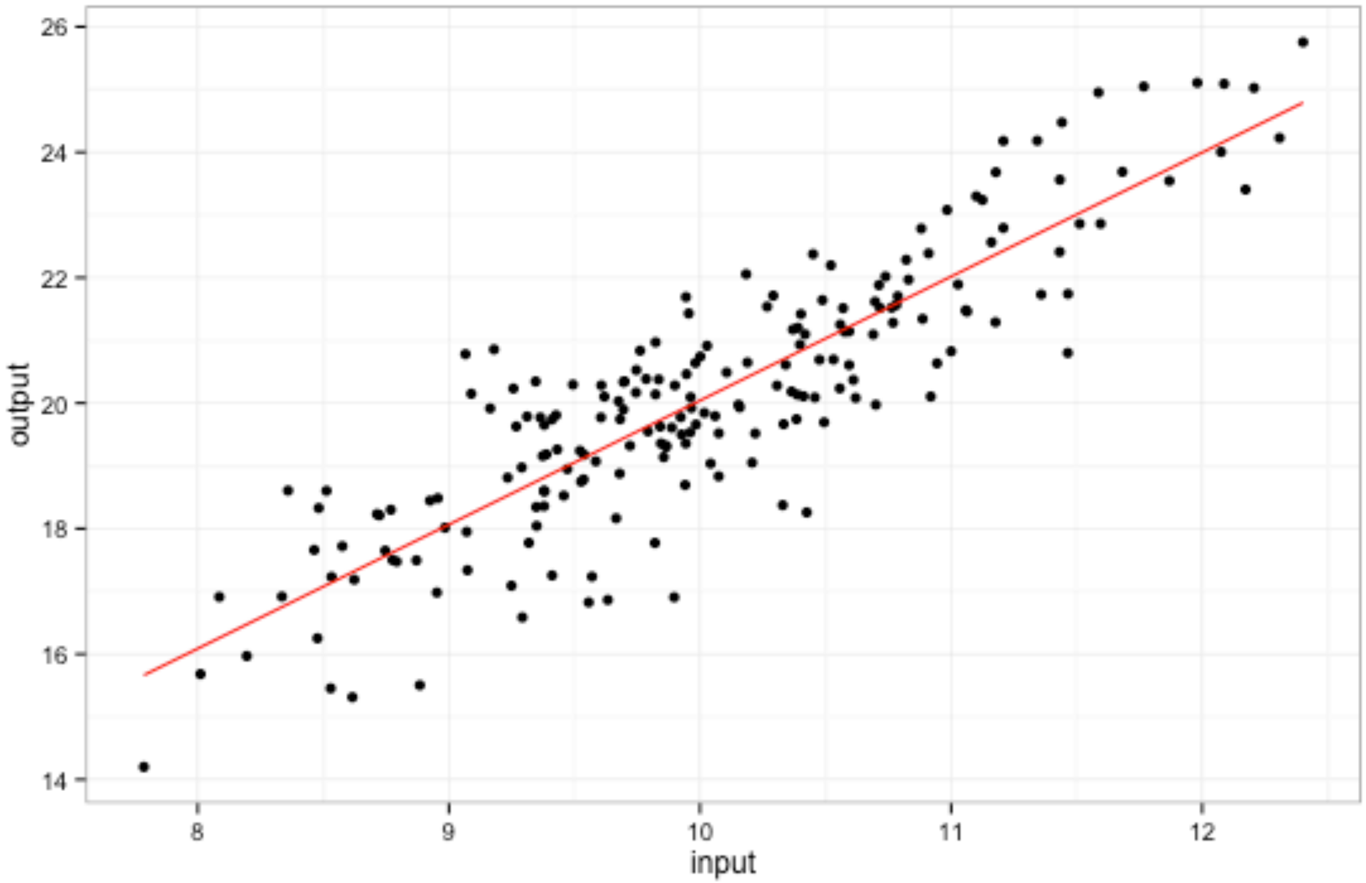
Example: ZA infections

- Bivariate plots:



Statistical Modeling: Regression

One of the most commonly used statistical modeling techniques used is linear regression

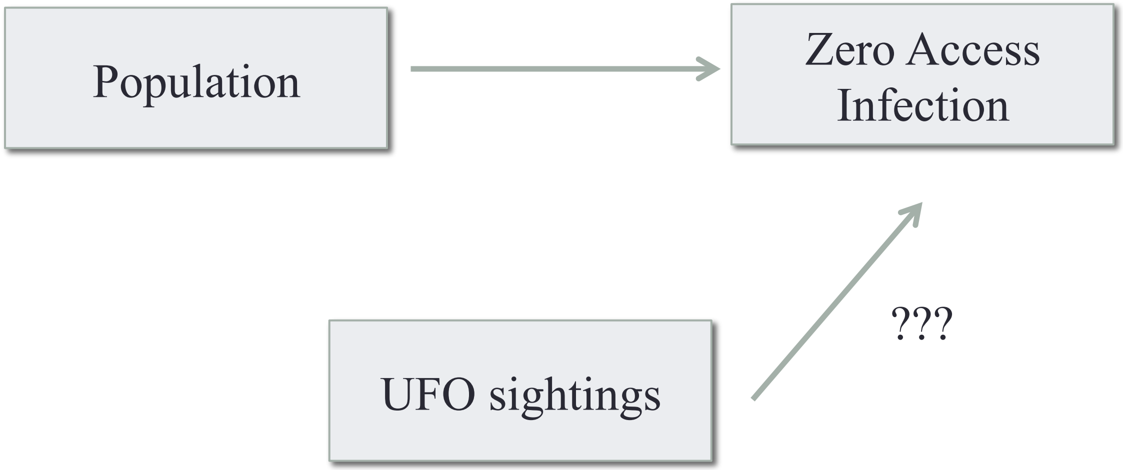


Statistical Modeling: Regression

What does it do?

- Still a correlational analysis, but allows you to partial out other effects

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$



Simple Regression

Results?

```
Call:
lm(formula = infections ~ pop10000, data = za.county)

Residuals:
     Min       1Q   Median       3Q      Max
-1076.23    -9.17    -2.78     2.46  1106.31

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.21413    1.25794   -0.17    0.865
pop10000     8.31725    0.03722  223.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.54 on 3070 degrees of freedom
Multiple R-squared:  0.9421,    Adjusted R-squared:  0.9421
F-statistic: 4.994e+04 on 1 and 3070 DF,  p-value: < 2.2e-16
```

Says that with every increase of 10,000 people we see, on average, about 8.3 more computers infected.

Multiple Regression

- What about UFOs? Are they infecting computers?

```
Call:
lm(formula = infections ~ ufo2010, data = za.county)

Residuals:
     Min       1Q   Median       3Q      Max 
-1699.72   -25.07   -14.76    -0.76  2738.94 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.75794    2.59828   6.834 9.89e-12 ***
ufo2010      8.31411    0.08711  95.445  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.8 on 3070 degrees of freedom
Multiple R-squared:  0.7479,    Adjusted R-squared:  0.7479 
F-statistic:  9110 on 1 and 3070 DF,  p-value: < 2.2e-16
```

Multiple Regression

- Multiple regression with UFO sighting and Population:

```
Call:
lm(formula = infections ~ pop10000 + ufo2010, data = za.county)

Residuals:
     Min       1Q   Median       3Q      Max 
-1060.19    -9.10    -2.88     2.42  1150.89 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.21857    1.25357  -0.174    0.862    
pop10000     7.99024    0.07840 101.914  < 2e-16 ***
ufo2010      0.41640    0.08796   4.734  2.3e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.3 on 3069 degrees of freedom
Multiple R-squared:  0.9425,    Adjusted R-squared:  0.9425 
F-statistic: 2.516e+04 on 2 and 3069 DF,  p-value: < 2.2e-16
```

Multiple Regression

UFOs really are infect computers???

- Again, probably not..
- There could be lots of reasons. Ideas?

- Reported UFO sightings could be related to education level. I don’t have education in the data so we can’t check, but maybe.
- UFO sightings are highly correlated with population
  - A problem known as “collinearity” in Regression-speak.
- Outliers; there might be a few
- Could just be random

→ **Good modeling isn’t just looking for statistical significance.**

Prediction

1. Regression analysis is useful for looking at relationships between variables (e.g., population and ZA infections)
2. Also useful for prediction
   - Basic linear regression is useful for predicting a quantitative variable: how many computers will be infected?

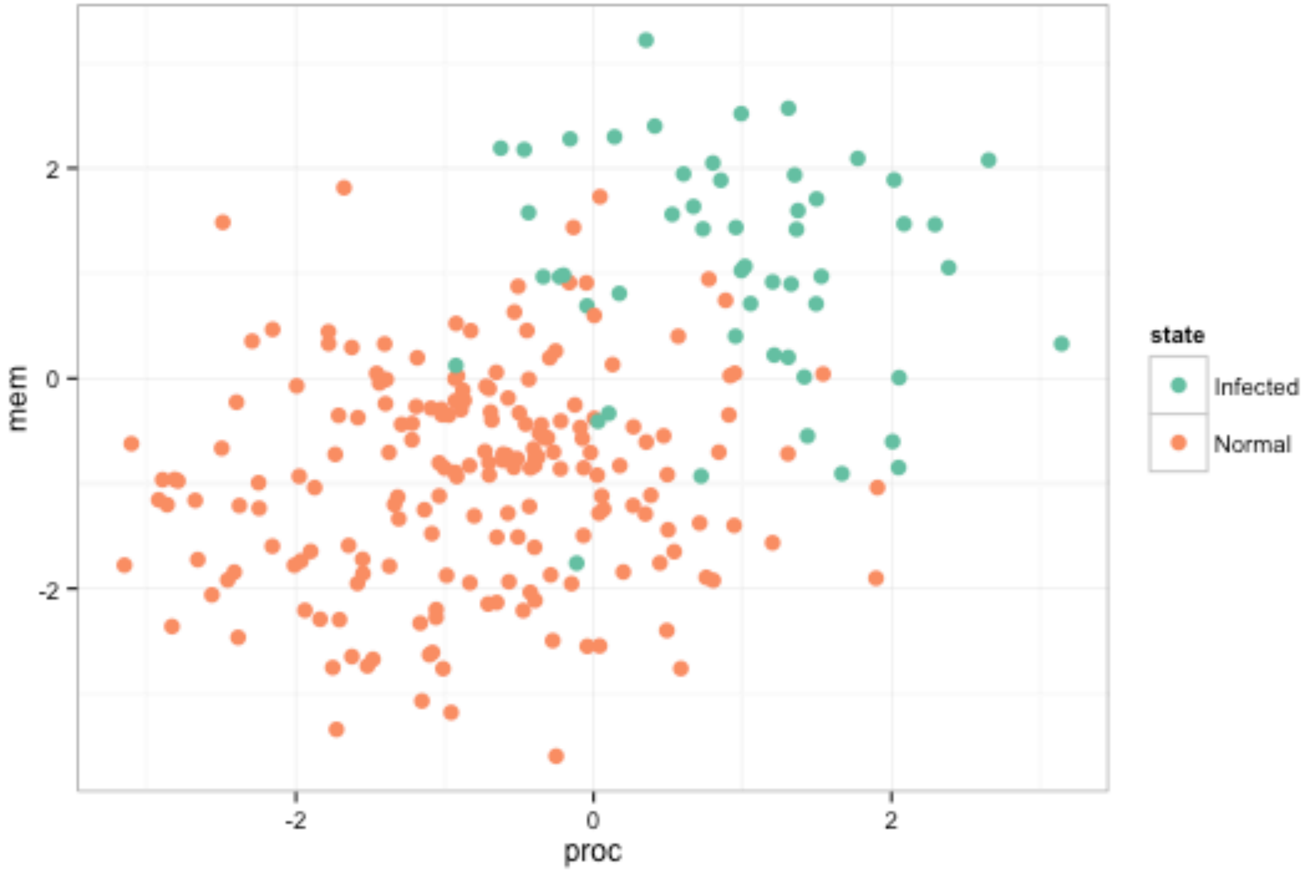
When the prediction task is qualitative, basic regression (OLS) isn't the best choice.

- Suppose we wanted to detect if a system was infected?

Prediction

Simple example that's easy to see:

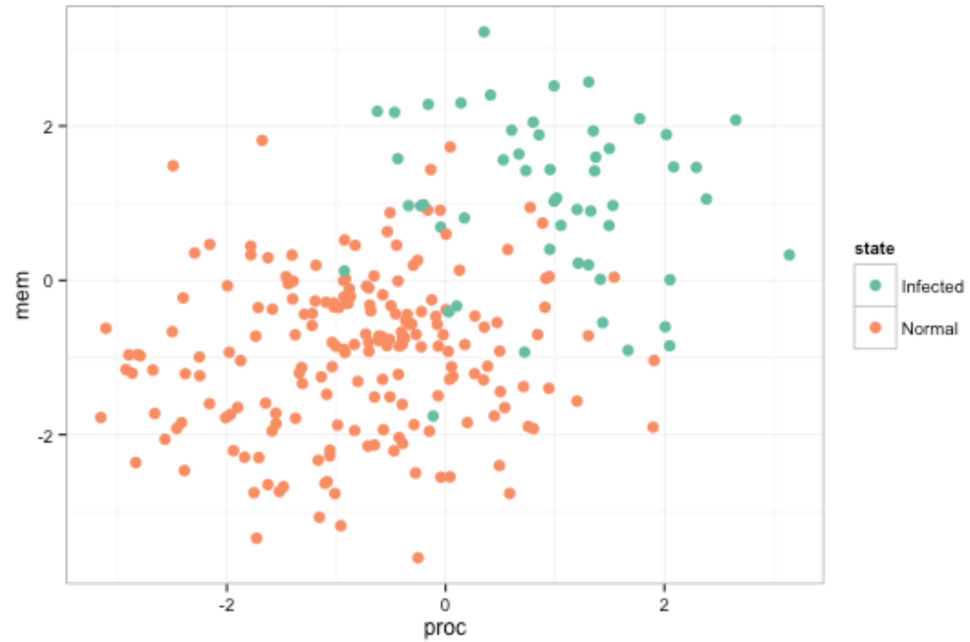
- Suppose we have memory and processor data



Prediction

With two variables it's easy to see but with more, nearly impossible.

- Need a way to classify, given memory processor data, the **probability of being infected**.



Logistic Regression

A common method to predict a 0-1 event is Logistic Regression.

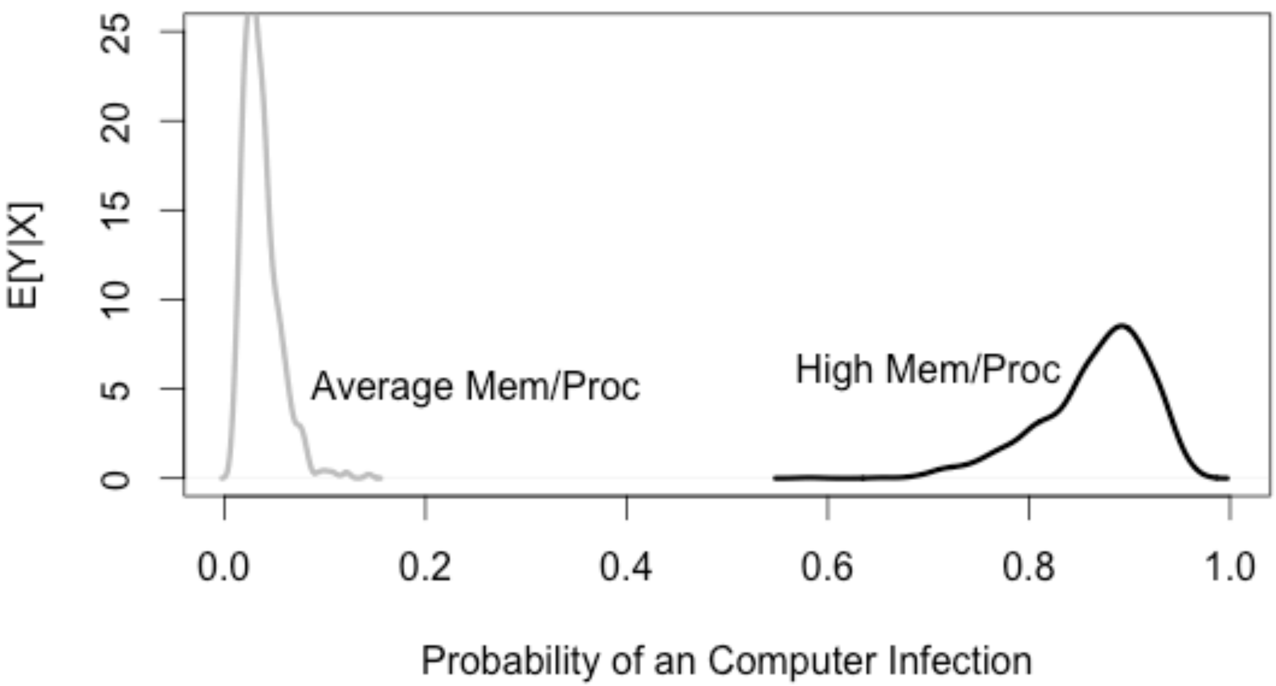
- Details of Logistic Regression go beyond a 2-day mini course.

```
Call:
glm(formula = formula, weights = w, family = binomial(link = "logit"),
    model = F, data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.82651  -0.24100  -0.09186  -0.01296   3.15601

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6641     0.3248  -5.123 3.00e-07 ***
proc          1.8660     0.3354   5.563 2.65e-08 ***
mem           1.7612     0.3038   5.797 6.77e-09 ***
---
```

Not Easily Interpretable

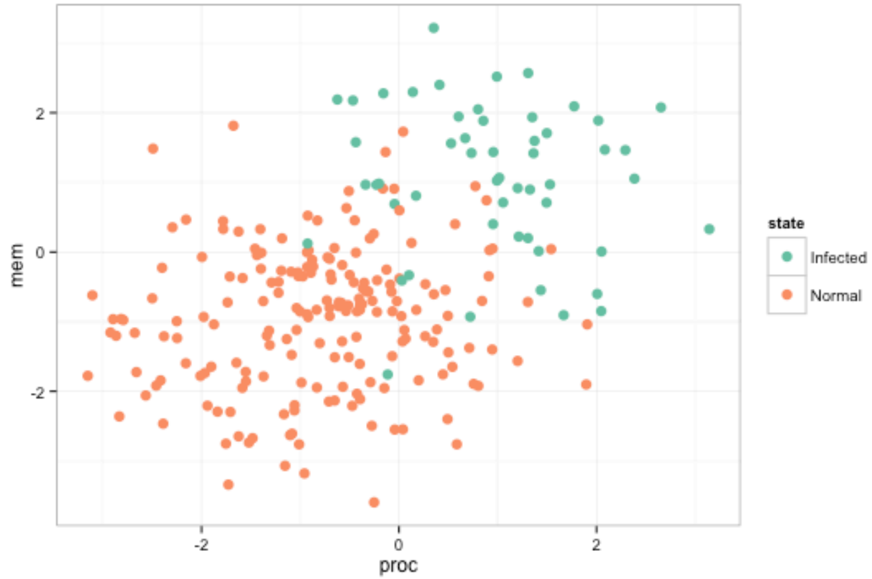
Logistic Regression

- However we can use simulation to make sense of these outputs
  - The previous model predicts that if Memory and Processor usage are both 1 standard deviation above the average, then there is an 87% chance the computer is infected, versus a 3% if at average levels!



Logistic Regression

- Obviously we could see some relationship from the initial scatterplot:



- What happens when we have lots of different factors that could contribute?
  - That's the power of the statistical prediction, it works the same.

Prediction tasks

- What are some other prediction tasks that use in your organizations? (or would like to use)

- There are all kinds of useful prediction tasks that statistical models can help do
  - Higher risk or more vulnerable employees
  - High risk phishing links
  - Non computer security predictions?

Machine Learning

- Building appropriate statistical models takes time

- The faster we can develop, implement, and respond to these models/results the better we'll do
  - Time till Compromise vs Time till Detection

- Can we automate some of these models so that they optimally change and "learn" over time?
  - Yes: machine learning.

Machine Learning

- Machine learning techniques go well beyond the scope of a two day overview course.

- Many **many** successful applications
  - Classic example: spam filtering
  - Potential applications? Automated threat detections, automated sentiment analysis, many more

It's not a magic bullet

- Classic problem with Machine Learning: over-fitting the data.
- Can be computationally intensive
- It's a growing field that is honestly still relatively young.