Big Data Analysis with Apache Spark

[Logos: Apache Spark, Databricks, Berkeley X]
This Lecture: Relation between Variables

An association

A trend
  » Positive association or Negative association

A pattern
  » Could be any discernible “shape”
  » Could be Linear or Non-linear

Visualize, then quantify, but be cautious!
Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950’s

- Experiment: subjects guess whether 10 hidden cards were red or blue

He found that about 1 person in 1,000 had *Extra Sensory Perception*

- They could correctly guess the color of all 10 cards

*Example from Jeff Ullman/Anand Rajaraman*
Rhine Paradox

Called back "psychic" subjects and had them repeat test
» They all failed

Concluded that act of telling psychics that they have psychic abilities causes them to lose it…(!)

Q: What’s wrong with his conclusion?
Rhine’s Error

Q: What’s wrong with his conclusion?

$2^{10} = 1,024$ combinations of red and blue of length 10

0.98 probability at least 1 subject in 1,000 will guess correctly
The Correlation Coefficient $\rho$

**Pearson product-moment correlation coefficient $\rho$**

» Measures *linear association* between $X$ and $Y$

» Based on standard units (Standard Deviation)

Ranges from $-1 \leq \rho \leq 1$

» $\rho = 1$: scatter is perfect straight line sloping up

» $\rho = -1$: scatter is perfect straight line sloping down

» $\rho = 0$: No linear association; uncorrelated
Total positive correlation

$\rho = +1$

https://commons.wikimedia.org/wiki/User:Kiatdd
- Correlation

Total *negative* correlation

\[ \rho = -1 \]

https://commons.wikimedia.org/wiki/User:Kiatdd
0 Correlation

No correlation

\[ \rho = 0 \]

https://commons.wikimedia.org/wiki/User:Kiatdd
$-1 < \rho < 1$ Correlation

$-1 < \rho < 1$ correlation

https://commons.wikimedia.org/wiki/User:Kiatdd
Definition of $\rho$

**Correlation Coefficient $\rho$:**

» Average of product of (x in standard units) and (y in standard units)
» Measures how clustered the scatter is around a straight line

Further Properties of $\rho$

» $\rho$ is a pure number with no units
» $\rho$ is not affected by changing units of measurement
» $\rho$ is not affected by switching the x and y axes

Remember: *Correlation is not causation*
Graph of Averages

A *visualization* of $x$ and $y$ pairs

» Group each $x$ with a representative $x$ value (e.g., rounding)
» Average the corresponding $y$ values for each group

One point per representative $x$ value and average $y$ value

If the association between $x$ and $y$ is linear, then points in the graph of averages tend to fall on the regression line
Regression to the Mean

A statement about $x$ and $y$ pairs

- Measured in *standard units*
- Describing the deviation of $x$ from 0 (the average of $x$’s)
- And the deviation of $y$ from 0 (the average of $y$’s)

*On average, $y$ deviates from 0 less than $x$ deviates from 0*
Regression to the Mean

A statement about $x$ and $y$ pairs

» Measured in *standard units*

» Describing the deviation of $x$ from 0 (the average of $x$’s)

» And the deviation of $y$ from 0 (the average of $y$’s)

On average, $y$ deviates from 0 less than $x$ deviates from 0

$$y_{(su)} = \rho \times x_{(su)}$$

Not true for all points — a statement about averages
Slope & Intercept

In original units, the regression line has this equation:

$$\frac{\text{estimate of } y - \text{average of } y}{\text{SD of } y} = \rho \times \frac{\text{the given } x - \text{average of } x}{\text{SD of } x}$$
Slope & Intercept

In original units, the regression line has this equation:

\[
\frac{\text{estimate of } y - \text{average of } y}{\text{SD of } y} = \rho \times \frac{\text{the given } x - \text{average of } x}{\text{SD of } x}
\]

- \( y \) in standard units
- \( x \) in standard units
Slope & Intercept

In original units, the regression line has this equation:

\[
\frac{\text{estimate of } y - \text{average of } y}{\text{SD of } y} = \rho \times \frac{\text{the given } x - \text{average of } x}{\text{SD of } x}
\]

Lines can be expressed by slope & intercept

\[y = \text{slope} \times x + \text{intercept}\]
Regression Line

Original Units

SD $\times$

$\rho \ast$ SD $\gamma$
Regression Line

Original Units

(Average x, Average y)

ρ * SD y

SD x
Regression Line

Standard Units

Original Units

(Average x, Average y)

\[
(0, 0) \quad \rho \quad SD \times \quad \rho \ast SD \gamma
\]
The Regression Model

A model is a set of *assumptions* about the data.

Regression model *justifies* using regression line as a predictor.

For each point:
- Sample an $x$
- Find its $y$ on regression line (*signal*)
- Add a sample of deviation (*noise*)
Predicting with Regression

Justified by assuming that $x$ and $y$ are linearly related

Only reasonable within the range of observed data

Varies by sample with more variability at the extremes

Predictions are average values, not perfect guesses
Errors: Evaluating Prediction Accuracy

*Error*: the difference between estimated and actual values
» Errors can be positive or negative
» Depend on data and the line chosen

Common metric:
» Root Mean Squared Error (or Root Mean Squared Deviation)

**Root Mean Squared Error (RMSE)** =

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$
Regression by Minimizing Errors

The *regression line* is the one that minimizes MSE of a collection of paired values.

Minimizing any of these quantities yields equivalent results:

» Root Mean Squared Error
» Mean Squared Error
» Total Squared Error
Regression by Minimizing Errors

The *regression line* is the one that minimizes MSE of a collection of paired values.

The *slope* and *intercept* are unique for regression.

Numerical minimization is approximate but effective.

Lots of machine learning involves minimizing error.
Multiple Linear Regression

*Simple regression*: one input → one output

*Multiple regression*: many inputs → one output

\[
GPA = a_{days} \times days + a_{contributions} \times contributions + b
\]

Find \( a \)'s and \( b \) by minimizing RMSE
Statistics Terminology

**Inference:** Making conclusions from random samples

**Population:** The entire set that is the subject of interest

**Parameter:** A quantity computed for the entire population

**Sample:** A subset of the population

» **Random Sample:** we know chance any subset of population will enter the sample, in advance

**Statistic:** A quantity computed for a particular sample
Estimating a Parameter

1. Describe the population and a parameter of interest
Estimating a Parameter

1. Describe the population and a parameter of interest
2. Acquire a random sample
Estimating a Parameter

1. Describe the population and a parameter of interest
2. Acquire a random sample
Estimating a Parameter

1. Describe the population and a parameter of interest
2. Acquire a random sample
3. Compute statistics
Estimating a Parameter

1. Describe the population and a parameter of interest
2. Acquire a random sample
3. Compute statistics
4. Pick an estimate
Estimating a Parameter

1. Describe the population and a parameter of interest
2. Acquire a random sample
3. Compute statistics
4. **Pick an estimate**
   - Draw conclusions
Empirical Distributions, Statistics & Parameters

A reasonable way to estimate a parameter (e.g., population average, max, median, ...) is to compute the corresponding statistic for a sample
Empirical Distributions, Statistics & Parameters

A reasonable way to estimate a parameter (e.g., population average, max, median, …) is to compute the corresponding statistic for a sample.

Different samples will lead to different estimates.

Population (fixed) $\rightarrow$ Sample (random) $\rightarrow$ Statistic (random)
Empirical Distributions, Statistics & Parameters

A reasonable way to estimate a parameter (e.g., population average, max, median, …) is to compute the corresponding statistic for a sample.

Different samples will lead to different estimates.

Goal: Infer the variability of a statistic, using only a sample.
Sample Variability

Anatomy of a sample:

» A sample contains not just a statistic, but a whole data set!

The same sample can be used for multiple purposes:

» Compute a statistic that is an \textit{estimate} of a parameter

» Approximate the shape of the \textit{population distribution}
Confidence Intervals: A Margin of Error

Estimation is a process with a random outcome

Population (fixed) $\rightarrow$ Sample (random) $\rightarrow$ Statistic (random)

Instead of picking a single estimate of the parameter, we can pick a whole interval: lower bound to upper bound
Confidence Intervals: A Margin of Error

Estimation is a process with a random outcome

Population (fixed) $\rightarrow$ Sample (random) $\rightarrow$ Statistic (random)

Instead of picking a single estimate of the parameter, we can pick a whole interval: lower bound to upper bound

A 95% *Confidence Interval* is an interval constructed so that it will contain the parameter for 95% of samples
Confidence Intervals: A Margin of Error

Estimation is a process with a random outcome

Population (fixed) $\rightarrow$ Sample (random) $\rightarrow$ Statistic (random)

A 95% Confidence Interval is an interval constructed so that it will contain the parameter for 95% of samples

For a particular sample, the interval either contains the parameter or it doesn’t: the process works 95% of the time
Resampling

**Inferential idea:** When we wish we could sample again from the population, we instead sample from the sample.
Intervals

If an interval *around the parameter* contains the estimate, then a (reflected) interval of the same width *around the estimate* contains the parameter (and vice versa)
Resampled Confidence Interval

Inferential idea: The variability of the sampled distribution is a useful proxy for the variability of the original distribution

Collect a random sample

Compute your estimate (e.g., sample average)

Resample $K$ samples from the sample, with replacement

- Compute the same statistic for each resampled sample
- Take percentiles of the deviations from the estimate
Verifying Intervals

When all you have is a sample, it is impossible to verify empirically whether the interval you compute is correct.

If you have the whole population, then you can check how often intervals are correct.
Simple Linear Regression

**Simple:** One explanatory or predictor variable $x$

Can be used to estimate *response variable* $y$ based on $x$

Strength of linear relation between $x$ and $y$ is measured by *correlation* $\rho$
Regression Line

Estimate of $y = \text{slope} \times x + \text{intercept}$

slope $= \rho \times \frac{(SD \text{ of } y)}{(SD \text{ of } x)}$

intercept $= (\text{average of } y) - \text{slope} \times (\text{average of } x)$

“Best” among all straight lines for estimating $y$ based on $x$

“Best”: minimizes RMSE of estimation
Tyche, the Goddess of Chance
A “Model”: What Tyche does

Distance drawn at random from normal distribution with mean 0

Another distance drawn independently from the same normal distribution
Plan for Prediction

If our model is good:

» Regression line is close to Tyche’s true line
» Given a new value of x, predict y by finding the point on the regression line at that x
» Bootstrap the scatter plot
» Get a new prediction using the regression line that goes through the resampled plot
» Repeat the two steps above many times
» Get an interval of predictions of y for the given x
Predictions at Different Values of $x$

Since $y$ is correlated with $x$, the predicted values of $y$ depend on the value of $x$.

The width of the prediction interval also depends on $x$.

» Typically, intervals are wider for values of $x$ that are further away from the mean of $x$. 


Rain on the Prediction Parade

We observed a positive slope and used it to make our predictions.

But what if the scatter plot got its positive slope just by chance?

What if the true line is actually FLAT?
Confidence Interval for True Slope

Steps:
» Bootstrap the scatter plot
» Find the slope of the regression line through bootstrapped plot
» Repeat
» Draw the empirical histogram of all the generated slopes
» Get the “middle 95%” interval

That’s an approximate 95% confidence interval for the slope of the true line
Inference for the True Slope

**Null hypothesis:** The slope of the true line is 0  
**Alternative hypothesis:** No, it’s not

**Method:**
- Construct a bootstrap confidence interval for the true slope
- If the interval doesn't contain 0, reject the null hypothesis
- If the interval does contain 0, there isn't enough evidence to reject the null hypothesis
Confidence Intervals for Testing

**Null hypothesis:** A parameter is equal to a specified value

**Alternative hypothesis:** No, it’s not

**Method:**
- Construct a confidence interval for the parameter
- If the specified value isn’t in the interval, reject the null hypothesis
- If the interval does contain 0, there isn’t enough evidence to reject the null hypothesis