Data Analytics

in Accounting

Introduction

What is Accounting?

- Accountants: Work with economic constructs and distill them into measures
- Data Analysts: Only ever work with measures

Accounting ⇒ Measurement ⇒ Statistics ⇒ Data Analytics

Descriptive Stats In Accounting

- Bring meaning to numbers/data
- Telling a story based on data
- Create comparisons and relationships
- Understand the known unknown with known metrics
- Risk assessment and probability of risk
- Dispersion: How far away from the mean (Find the Normal, Probability, and Predictive Analysis)
- Normal Distribution: Predict what is common/normal/expected
- Validity of what is being counted

Validity and Causality

- Constructs: What counts, and measures are what get counted
- Measures: Comparable, consistent, and **Validity**

Validity

- Internal Validity: does X cause Y?
- External Validity: would an observed effect generalize to other settings?
- Construct Validity: does the measure sufficiently capture the construct?
- Statistical Conclusion Validity: have proper analytical procedures been followed?

Causality

Causality is <u>difficult</u> to infer

What are three prerequisites for causal inference?

- Temporal precedence: (x must occur before y)
- Significant correlation: (x must be related to y)
- Alternative explanations must be eliminated (The difficult part)

Five Groups of Alternative Explanations

Weakness in any of these links can cast doubt on a causal relation between X and Y.

- Correlated Omitted Variables: Assuming that X causes Y when really Z causes both X and Y.
- Reverse Causality: Assuming that X causes Y when really Y causes X.
- Selection Bias: When the sample of individuals analyzed systematically differs from the population.
- Measurement Error: When values of measured observations randomly/systematically differ from true values.
- Spurious Correlation: When the correlation between X and Y is really just due to chance.

Data Models

- The process of creating a visual representation of information systems to show how data is stored and connected within a system
- Data models serve as a blueprint for designing and communicating the information structure of a system

Relational Data Models

Relational databases dominate the landscape of database management systems.

- Class: A collection of things about which an organization wants to collect and store data (entire table, columns and rows)
- Attributes: The specific facts or dimensions of a class for which we will collect and store data (columns of the table)
- Associations: A formally stated or acknowledged relationship between two classes (records)

Data Keys

Primary Key (PK)

An attribute that uniquely identifies every instance in a class (i.e., a row of a table)

- Each record must have a primary key
- Values for a table's primary key must NOT repeat

Foreign Key (FK)

A field in one table that is the primary key of another table in the database

- Links tables together
- The table that is the "many" side of a one-to-many relationship gets the foreign key.

Advanced Topics in MS Excel

Dynamic Array Functions

- Unique: Returns a list of unique values from a range, eliminating duplicates
- Sort: Sorts data in ascending or descending order automatically
- Sequence: Generates a sequence of numbers, with an optional starting value and step increment
- XLookUp: Searches for a value in a range and returns the corresponding value from another range, with additional options for matching and handling missing data
- TextSplit: Splits text into multiple cells based on a delimiter, with an option to ignore empty results
- Filter: Filters data based on a specified condition and returns the matching results

Pivot Tables

- Data summarization tool found in spreadsheet and data analysis programs that quickly reorganizes and summarizes large datasets to highlight key patterns, trends, and relationships
- A powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data

Power Pivot

COM add-in for Excel that enables relational data modeling and analysis

Features

- Building data models in Excel using internal/external data sources
- This enables Excel to analyze larger datasets
- Multi-table calculations using DAX formulas

Function

$= RELATED(< column >)$: Returns a related value from another table.

Lambda Functions

$$= LAMBDA(parameter1, parameter2, \ldots, calculation)$$

- The Excel LAMBDA function allows users to create custom, reusable functions using Excel's own formula language, without needing VBA or macros
- Define a LAMBDA by specifying the function's parameters and the calculation logic, then save it as a named function in the Name Manager to use it throughout your workbook

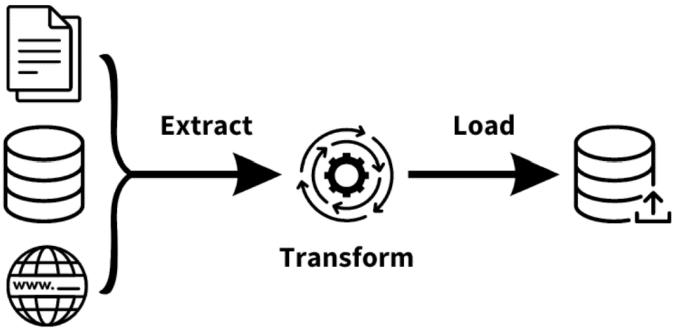
Power Query

Process

- Connect to Data
- Intro to UI
- Data Profiling
- Adding Columns & Groupby
- Appending Queries
- Joins (Merging) Tables

ETL Process Overview

Consists of all the activities needed to prepare the data for analysis. Often, the most time-consuming part of the whole data analytics process.



Includes:

- Collection
- Cleaning
- Combining
- Structuring
- Transformation
- Profiling
- Formatting

High Quality Information

Characteristic	Description
Accurate	Correct; free from error; accurately represents underlying facts
Available	Accessible to users when needed
Complete	Does not omit needed data; sufficient in depth and breadth
Consistent	Presented in the same format every time
Current	Includes data that is up-to-date to the present
Objective	Unbiased; impartial
Relevant	Appropriately pertains to the requested situation
Timely	Provided in time for users to make decisions
Understandable	Easily comprehended and communicated
Verifiable	Can be checked/confirmed by others; supported with documentation

Data

Facts that are:

- Raw
- Unorganized
- Meaningless

Information

Data that has been:

- Processed
- Structured
- Made useful

Workflow Automation

Data Structuring

- **Transposing:** Switching the rows and columns of the data
- **Pivoting:** Reshaping the data from tall to wide format. The unique values of one column are converted to columns, and values are specified for aggregation.
- **Unpivoting:** Reshaping the data from wide to tall format. Takes multiple related columns and transforming them into a single column of values.

Combining Data

- **Appending:** Stacking datasets vertically (adding rows).
- **Connecting to a Folder:** One of the options to "Get Data." When importing a folder containing multiple files, combining the files is equivalent to "appending."
- **Merging (Joining):** Combines datasets horizontally based on a common primary/foreign key match (adding columns).
    - Left: Retains all records from 1st & only matches from 2nd
    - Right: Retains all records from 2nd & only matches from 1st
    - Inner: Retains only matches between the two tables
    - Full (Outer): Retains all records of both tables, with matching records being joined where possible

Process Science

Business Process Modeling

The What? The Why? The How?

Segregation of Duties

Requiring certain tasks be performed by separate individuals. Within a business process, individuals should not perform duties across more than one of the following areas:

- Custody of Assets
- Authorization of Procedures
- Recording of Information

Process Mining

- The intersection of Business Process Modeling and Business Intelligence.
- An approach to automate business process modeling and analysis.
- Used for: Process discovery; Locating inefficiencies; Assessing compliance

Event Logs

Process mining platforms retrieve data from event logs to produce insights. Three required data inputs:

Key	Description	Meaning
Case ID	Unique reference to identify each instance of a cycle flow	WHICH
Activity	Description of the process that the instance has undergone	WHAT
Timestamp	Record of when the case went through an activity	WHEN

Variant Analysis

- A variant is a unique sequence of activities taken by at least one case.
- The number of times that an activity is performed is listed within each bubble.
- The number of times a case proceeds from one activity to the next is listed within each arrow.
- The most common variant is called the “happy path.”

Limitations of Process Mining

1. The entire organization needs to sufficiently support providing the necessary data to create event logs.
2. Not all processes will be recorded (and hence cannot be analyzed).
3. Analysis of complicated processes (i.e., with many variants) can be unwieldy.

Statistics

Descriptive Stats

Descriptive statistics summarize the distributional properties of a dataset

Measures of Dispersion

- Max/Min: The largest and smallest values.
- Range: The difference between the maximum and minimum values.
- Standard Deviation: Measures the dispersion or spread of data points around the mean.

Measures of Central Tendency

- Mean: The average value of the dataset.
- Median: The middle value when the data points are arranged in order.
- Mode: The most frequently occurring value in the dataset.

Correlation

Correlation measures the strength and direction of a linear relationship between two variables.

Pearson Correlation Coefficient r: Ranges from -1 to 1

- Positive Correlation: As one variable increases, the other variable also increases.
- Negative Correlation: As one variable increases, the other variable decreases.
- No Correlation: No predictable linear relationship between variables.

Distribution

Histograms: A histogram is a visual representation of the distribution of numerical data. It groups data into bins (intervals) and shows the frequency of data points within each bin.

Linear Regression (OLS)

A statistical technique that models the relationship between one or more independent variables and a dependent variable. Useful for understanding associations, making predictions, and assessing potential causality.

Adjusted R^2: How well X explains Y

Common Mistakes

- Nonlinear Relationship
- Multicollinearity (corr between X's) >.5
- Data Mining (Overfitting)
- Omitted Variables
- Reverse Causality
- Extrapolating Beyond the Data

Regression Output

- Sign
- Size
- Significance

Simple Regression Model

$$y = \alpha + \beta x + \varepsilon$$

- Dependent Variable (y): The variable we are trying to predict or explain.
- Independent Variable (x): A variable we think may explain y.
- Intercept (α): The predicted value of y for when x equals zero. If the p-value is greater than .05, there can be no conclusion that the value is not 0.
- Slope (β): Estimate of how y will change with a one-unit increase in x.
- Error (ε): The part of y that remains unexplained by x.

Multiple Regression Model

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \varepsilon$$

Multiple linear regression allows for more than one independent variable.

What is the benefit of including multiple X variables in one regression? Each slope coefficient reflects the effect of the individual Xi on Y, holding all other independent variables constant. This is sometimes referred to as "controlling for" other X variables.

Multiple Linear Regression Assumptions

1. Linearity: The relationship between Y and X must be linear.
2. Independence: The Y value for one observation should not correlate with the next.
3. No Multicollinearity: The X variables should not be too highly correlated with each other (ρ>0.6)
4. Normality: The Y variable should be follow a normal distribution.

Backwards Elimination

One method to set up a prediction model is using 'backwards elimination.' This is a process of starting with a full model of independent variables and iteratively removing insignificant variables until all remaining are statistically significant.

1. Run a regression with a series of X variables that plausibly can help predict Y.
2. Identify which X variables were statistically significant predictors.
3. Run another regression with only the X variables from Step 2.
4. Repeat until the model contains only significant X variables.

Predictive Analytics

Cost Behavior

- Fixed costs: do not vary with production/sales volume
- Variable costs: vary entirely based on production/sales volume
- Mixed costs: vary somewhat with production volume but also have a fixed cost component

Earnings Persistence

- A measure of how well current earnings predict future earnings.
- Higher earnings persistence indicates stable and predictable earnings over time.
- Investors prefer firms with high earnings persistence.

Formulas

- Earnings Persistence: $\text{Net Income}_{t+1} = \alpha + \beta \text{Net Income}_t + \epsilon$ (β is typically a number between 0 and 1. The higher the β, the more persistent the earnings.)
- Differential Persistence of Accruals vs Cash: $\text{Net Income}_{t+1} = \alpha + \beta_1 \text{CFO}_t + \beta_2 \text{Accruals}_t + \epsilon$
- Discretionary Accruals: $\frac{\text{Total Accruals}_t}{\text{Assets}_{t-1}} = \beta_0 + \beta_1 \frac{1}{\text{Assets}_{t-1}} + \beta_2 \frac{\Delta \text{Sales}_t}{\text{Assets}_{t-1}} + \beta_3 \frac{\text{PP\&E}_t}{\text{Assets}_{t-1}} + \epsilon$

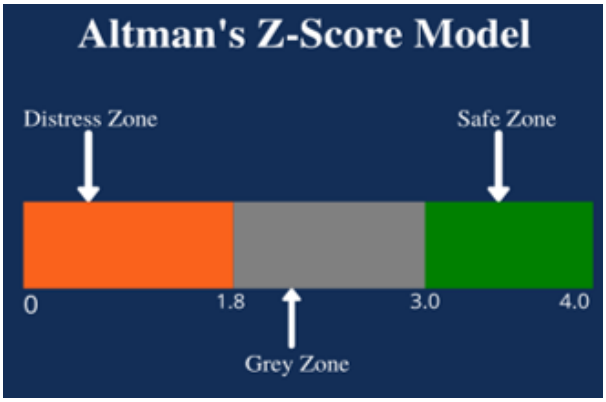
Altman Z-Score $Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$

- A model invented in 1968 by Edward Altman to predict bankruptcy risk.
- Designed to assess the likelihood of a firm going bankrupt within the next two years.
- While others have improved upon the original model, it is still a widely used model.

- $X_1 = Working\ Capital / Total\ Assets$
- $X_2 = Retained\ Earnings / Total\ Assets$
- $X_3 = EBIT / Total\ Assets$
- $X_4 = Market\ Cap / Total\ Liabilities$
- $X_5 = Sales / Total\ Assets$

Z > 2.99 → Safe Zone

- The company is financially stable with a low probability of bankruptcy
- Typically applies to well-established firms with solid financials



1.81 ≤ Z ≤ 2.99 → Gray Zone

- The company is in a financial risk zone where distress is possible
- Requires closer analysis of financial trends and industry conditions

Z < 1.81 → Distress Zone

- High likelihood of financial distress or bankruptcy within two years
- Often seen in struggling firms or industries facing downturns

Monte Carlo

- A computational technique that uses repeated random sampling to estimate outcomes.
- Helps to model uncertainty across multiple dimensions.

Process

1. Define the problem and identify key variables
2. Assign probability distributions to the variables
3. Run simulations
4. Analyze the results

Data Storytelling

Data Visualization

Exploratory

- Allows the user to explore data as a form of analysis
- Can be conducted for a problem that has not been clearly defined

Explanatory

- Explains to an audience what it needs to know
- Effectively communicates critical takeaways

Types of Data

Type	Meaning	Chart/Plot
Categorical	Presents an analysis of different groups	Bar Chart
Univariate	Map summary statistics for a single variable	Histogram
Multivariate	Analysis between 2+ variables	Scatter
Time Series	Captures differences in observations over time	Line
Proportional	Tracks response variables that express percentages/fractions	Stacked Bar

Intro to Power BI

1. Free now, free later
2. Synergy with Power Query
3. Currently, the most relevant software

Large Language Models

AI Use Cases

- Classification
- Summary
- Coding
- ETL
- Financial Analysis
- Deep Research
- Statistical Analysis
- Visualization