19-B-MA5821-ONL-EXT-SP85 Advanced Statistical Methods for Data Scientists

Week-2

Presented by
Ban (JCU)
Banmali.Pradhan@jcu.edu.au

Week 1 - Topics

- Misleading statistics
- Supervised learning
  - Linear regression, Logistic regression, Bayes classifier, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Decision tree, Neural Network, SVM, kNN, Random Forest, AdaBoost, Gradient Boosting
- Unsupervised learning
  - Clustering: partitional, model selection, hierarchical clustering, density based clustering
  - PCA, dimension reduction
  - Outlier detection
  - Recommenders: Association rules, “basket analysis”
- Theory of Probability
- Frequentist vs Bayesian

Key dates

Key dates	Date
Census date	12 September, 2019
Last date to withdraw without academic penalty	19 September,2019
Assessment 1 – Weekly quizzes. Total: 30% • Week 1 quiz: 10% • Week 3 quiz: 5% • Week 4 quiz: 5% • Week 5 quiz: 5% • Week 6 quiz: 5%	Week 1 quiz : A1A Due Sunday Week 1 Week 3 quiz : A1B Due Sunday Week 3 Week 4 quiz : A1C Due Sunday Week 4 Week 5 quiz : A1D Due Sunday Week 5 Week 6 quiz : A1E Due Sunday Week 6 See LearnJCU for details on date and time.
Assessment 2 – Weekly workbook exercise submissions (including short answers). Total: 30% • Week 2 submission (Regression) : 7.5% • Week 3 submission (General Linear Modelling) : 7.5% • Week 4 submission (Logistic regression) : 7.5% • Week 5 submission (Decision Trees and Cluster Analysis) : 7.5%	• Week 2 submission: A2A Due Sunday of Week 2. • Week 3 submission: A2B Due Sunday of Week 3. • Week 4 submission: A2C Due Sunday of Week 4. • Week 5 submission: A2D Due Sunday of Week 5. See LearnJCU for details on date and time
Assessment 3 – Capstone project. Total :40%	Due Wednesday of Week 7. See LearnJCU for details on date and time

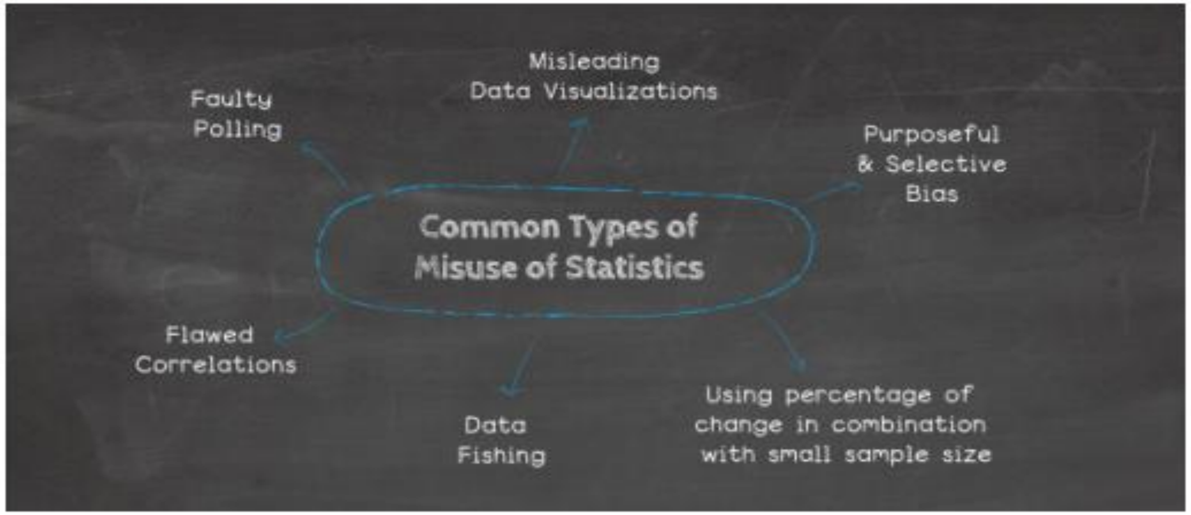
Misleading Statistics

Misleading statistics are simply the misusage – purposeful or not – of a numerical data. The results provide a misleading information to the receiver, who then believes something wrong if he or she does not notice the error or does not have the full data picture.

73.6% Of All Statistics Are Made Up

33.7% of scientists surveyed admitted to questionable research practices, including modifying results to improve outcomes, subjective data interpretation, withholding analytical details and dropping observations because of gut feelings…. Scientists!

How Statistics Can Be Misleading



20 COGNITIVE BIASES THAT SCREW UP YOUR DECISIONS

1. Anchoring bias.

People are **over-reliant** on the first piece of information they hear. In a salary negotiation, whoever makes the first offer establishes a range of reasonable possibilities in each person's mind.

2. Availability heuristic.

People **overestimate the importance** of information that is available to them. A person might argue that smoking is not unhealthy because they know someone who lived to 100 and smoked three packs a day.

3. Bandwagon effect.

The probability of one person adopting a belief increases based on the number of people who hold that belief. This is a powerful form of **groupthink** and is reason why meetings are often unproductive.

4. Blind-spot bias.

Failing to recognize your own cognitive biases is a bias in itself. People notice cognitive and motivational biases much more in others than in themselves.

5. Choice-supportive bias.

When you choose something, you tend to feel positive about it, even if that **choice has flaws**. Like how you think your dog is awesome — even if it bites people every once in a while.

6. Clustering illusion.

This is the tendency to **see patterns in random events**. It is key to various gambling fallacies, like the idea that red is more or less likely to turn up on a roulette table after a string of reds.

7. Confirmation bias.

We tend to listen only to information that confirms our **preconceptions** — one of the many reasons it's so hard to have an intelligent conversation about climate change.

8. Conservatism bias.

Where people favor prior evidence over new evidence or information that has emerged. People were **slow to accept** that the Earth was round because they maintained their earlier understanding that the planet was flat.

9. Information bias.

The tendency to **seek information when it does not affect action**. More information is not always better. With less information, people can often make more accurate predictions.

10. Ostrich effect.

The decision to **ignore dangerous or negative information** by "burying" one's head in the sand, like an ostrich. Research suggests that investors check the value of their holdings significantly less often during bad markets.

11. Outcome bias.

Judging a decision based on the **outcome** — rather than how exactly the decision was made in the moment. Just because you won a lot in Vegas doesn't mean gambling your money was a smart decision.

12. Overconfidence.

Some of us are **too confident about our abilities**, and this causes us to take greater risks in our daily lives. Experts are more prone to this bias than laypeople, since they are more convinced that they are right.

17. Selective perception.

Allowing our expectations to **influence how we perceive** the world. An experiment involving a football game between students from two universities showed that one team saw the opposing team commit more infractions.

18. Stereotyping.

Expecting a group or person to have certain qualities without having real information about the person. It allows us to quickly identify strangers as friends or enemies, but people tend to **overuse and abuse** it.

13. Placebo effect.

When **simply believing** that something will have a certain effect on you causes it to have that effect. In medicine, people given fake pills often experience the same physiological effects as people given the real thing.

14. Pro-innovation bias.

When a proponent of an innovation tends to **overvalue its usefulness** and undervalue its limitations. Sound familiar, Silicon Valley?

15. Recency.

The tendency to weigh the **latest information** more heavily than older data. Investors often think the market will always look the way it looks today and make unwise decisions.

16. Salience.

Our tendency to focus on the **most easily recognizable features** of a person or concept. When you think about dying, you might worry about being mauled by a lion, as opposed to what is statistically more likely, like dying in a car accident.

19. Survivorship bias.

An error that comes from focusing only on surviving examples, causing us to **misjudge a situation**. For instance, we might think that being an entrepreneur is easy because we haven't heard of all those who failed.

20. Zero-risk bias.

Sociologists have found that **we love certainty** — even if it's counterproductive. Eliminating risk entirely means there is no chance of harm being caused.

Machine learning synopsis

Purpose	Problem Space	ML Technique
Anomaly Detection	more features, aggressive boundary	One-class SVM
	less features, fast training	PCA-based anomaly detection
Prediction	Linear model, fast training	Linear regression
	Linear model, small dataset	Bayesian linear regression
	Accuracy, long training time	Neural network regression
	Accuracy, fast training	Decision forest regression
	Predict event counts	Poisson regression
	Accuracy, fast training, large memory	Boosted decision tree regression
Discovering structure	Clustering	K-means
Classification (two class, multi-class)	Fast training, linear model	Logistic regression
	Accuracy, long training time	Neural network
	Accuracy, fast training	Decision forest, Decision jungle
	More features	Deep SVM
Recommendation	What you may also like	Association rules, matchbox
Text Analytics	NER, Sentiment Analysis	Rule based, SVM
Computer Vision	Image recognition	CNN, OpenCV Library

JAMES COOK UNIVERSITY AUSTRALIA

Theory of probability

Key idea: $P(A) = \dfrac{\text{number times event A occurs}}{\text{number of all events}}$

Rules: $0 \leq P(A) \leq 1$, $P(\bar{A}) = 1 - P(A)$,

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Independent events: $P(A \text{ and } B) = P(A) \times P(B)$

Counting with Combinations and Permutations

$$_nP_r = \frac{n!}{r!}, \qquad _nC_r = \frac{n!}{(n-r)!r!}$$

Theory of probability

Odds against is given by number of unfavourable outcomes to number of favourable outcomes.

At betting scenario, bookmakers quote odds as odds against winning.
P(A) = Number of unfavourable outcomes/Number of favourable outcomes
e.g., odds in against of throwing a die to get 6 dots is 5:1 or 5/1

Probability of the event = Number of favourable outcomes / (Number of favourable outcomes + Number of unfavourable outcomes)

Dividend = 1 + odds against

If dividend for $1 stake on a win in a game for a team is $3.50, then:
Odds against = $3.50 - $1 = 2.50 to 1(25 to 10)
It means that, out of 35 games 25 times that particular team is expected to lose based on the above odds
The probability of wining for that team = 10 / 35 = 0.2857

Frequentist vs Bayesian

I have misplaced my phone somewhere in the home. I can use the phone locator on the base of the instrument to locate the phone and when I press the phone locator the phone starts beeping.

Problem: Which area of my home should I search?

Frequentist Reasoning

I can hear the phone beeping. I also have a mental model which helps me identify the area from which the sound is coming. Therefore, upon hearing the beep, I infer the area of my home I must search to locate the phone.

Bayesian Reasoning

I can hear the phone beeping. Now, apart from a mental model which helps me identify the area from which the sound is coming from, I also know the locations where I have misplaced the phone in the past. So, I combine my inferences using the beeps and my prior information about the locations I have misplaced the phone in the past to identify an area I must search to locate the phone.

Bayes' theorem (intuition)

Machine 1

Machine 2

Bayes' theorem (intuition)

What is the probability?

M2

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Bayes' theorem (intuition)

- Machine 1 (M1): 40 units/hour
- Machine 2 (M2): 30 units/hour
- Of all parts produced in a batch, there are 2% defective
- Of all defective product 50% from M1 and 50% from M2

Q1: What is the probability that a wrench produced by M2 is defective

Q2: What is the probability that a wrench produced by M1 is not defective

Bayes' theorem (intuition)

- Machine 1 (M1): 40 units/hour — P(M1) = 40/70 = 0.572
- Machine 2 (M2): 30 units/hour — P(M2) = 30/70 = 0.428
- There are 2% defective products — P(defect) = 2% = 0.02
- Of all defective product 50% from M1 and 50% from M2

P(M1|defect) = P(M2|defect) = 50% = 0.50

Q1: P(defect|M2)

Q2: 1 – P(defect|M1)

$$P(\text{defect}|M2) = \frac{P(M2|defect) * P(defect)}{P(M2)} \qquad \frac{0.50 * 0.02}{0.428} = 0.0233 = 2.33\%$$

Lets verify this with frequentist theory

- 8400 produced in a batch
- M1 produced: 4800
- M2 produced: 3600
- There are 168 defective products (which is 2% of the production)
- M1 produced 84 and M2 produced 84 defective products

$$P(\text{defect}|\text{M2}) = \frac{Total\ defective\ by\ M2}{Total\ Production\ by\ M2} \qquad \frac{84}{3600} = 0.0233 = 2.33\%$$

Why learning SAS VA

- Read the attached file: [sas-visual-analytics-105682.pdf]
- Improve your CV for better jobs

What does SAS® Visual Analytics do?

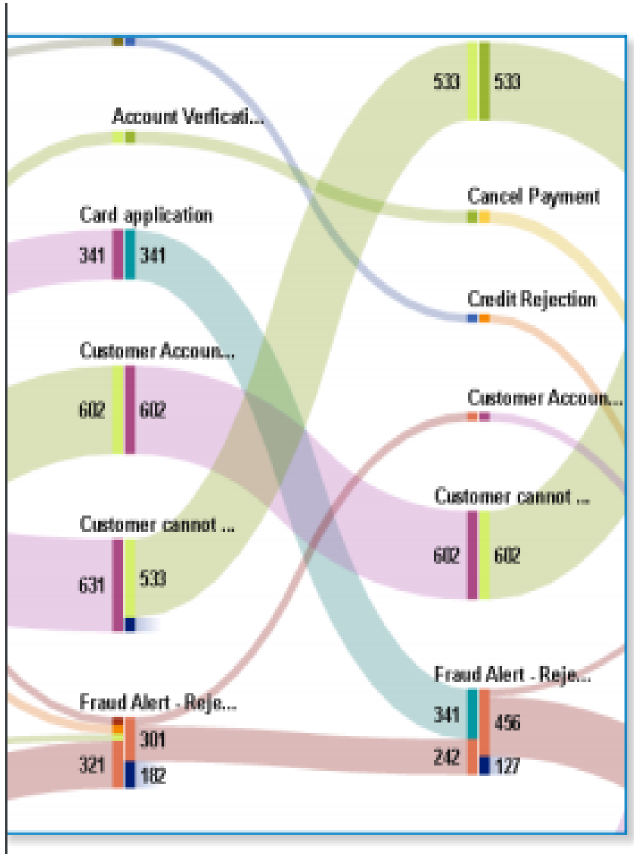
SAS Visual Analytics provides a complete platform for analytics visualization, enabling you to identify patterns and relationships in data that weren't initially evident. Interactive, self-service BI and reporting capabilities are combined with out-of-the-box advanced analytics so everyone can discover insights from any size and type of data, including text.

Why is SAS® Visual Analytics important?

Users of all skill levels can visually explore data on their own while tapping into powerful in-memory technologies for faster analytic computations and discoveries. It's an easy-to-use, self-service environment that can scale on an enterprisewide level.

For whom is SAS® Visual Analytics designed?

It's designed for anyone in your organization who wants to use and derive insights from data – from influencers, decision makers and analysts to statisticians and data scientists. It also offers IT an easy way to protect and manage data integrity and security.

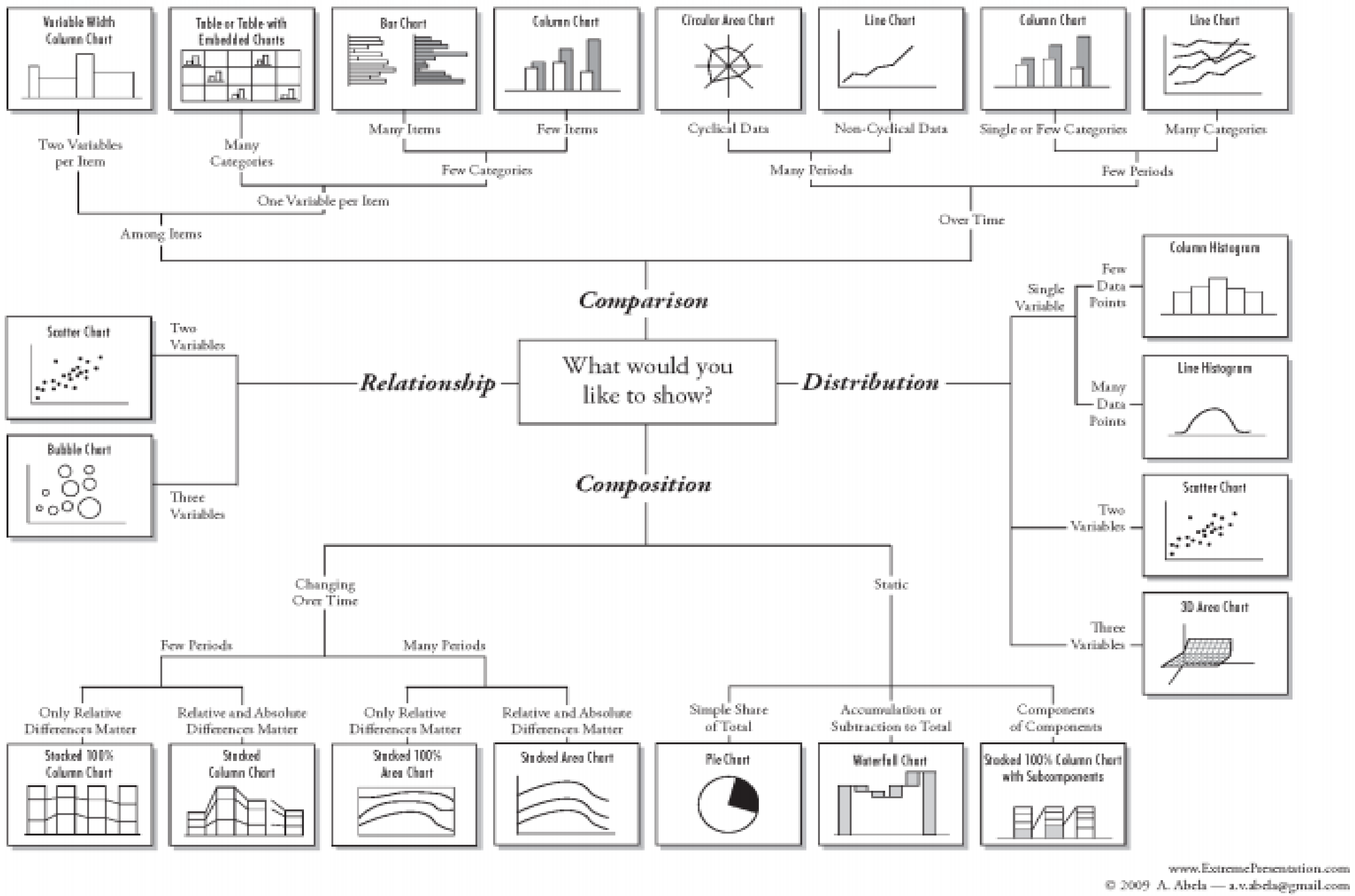


Week 2 - Topics

- Warm-up exercise
- Regression: Simple Linear Regression, Logistic Regression
- Multiple Linear Regression
- Assumptions for multiple linear regression
  - How to adjust for multicollinearity in multiple regression
- Different types of regression
- Terminology related to regression
- Summary flow chart for multiple linear regression
- Modeling Linear Regression with SAS VA

JAMES COOK UNIVERSITY AUSTRALIA

Chart Suggestions—A Thought-Starter

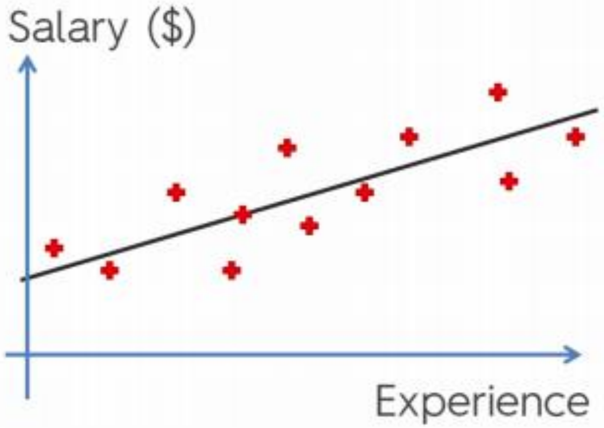


Simple Linear Regression - Intuition

salary

WorkExpYears	CurrentSalary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445
3.7	57189
3.9	63218
4	55794
4	56957
4.1	57081
4.5	61111
4.9	67938
5.1	66029
5.3	83088
5.9	81363
6	93940

Simple Linear Regression:



$y = b_0 + b_1 {}^* x$

Salary = b_0 + b_1 *Experience

Multiple Linear Regression

Simple Linear Regression

$$y = b_0 + b_1{*}x_1$$

Dependent variable (DV)

Independent variables (IVs)

Multiple Linear Regression

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + \ldots + b_n{*}x_n$$

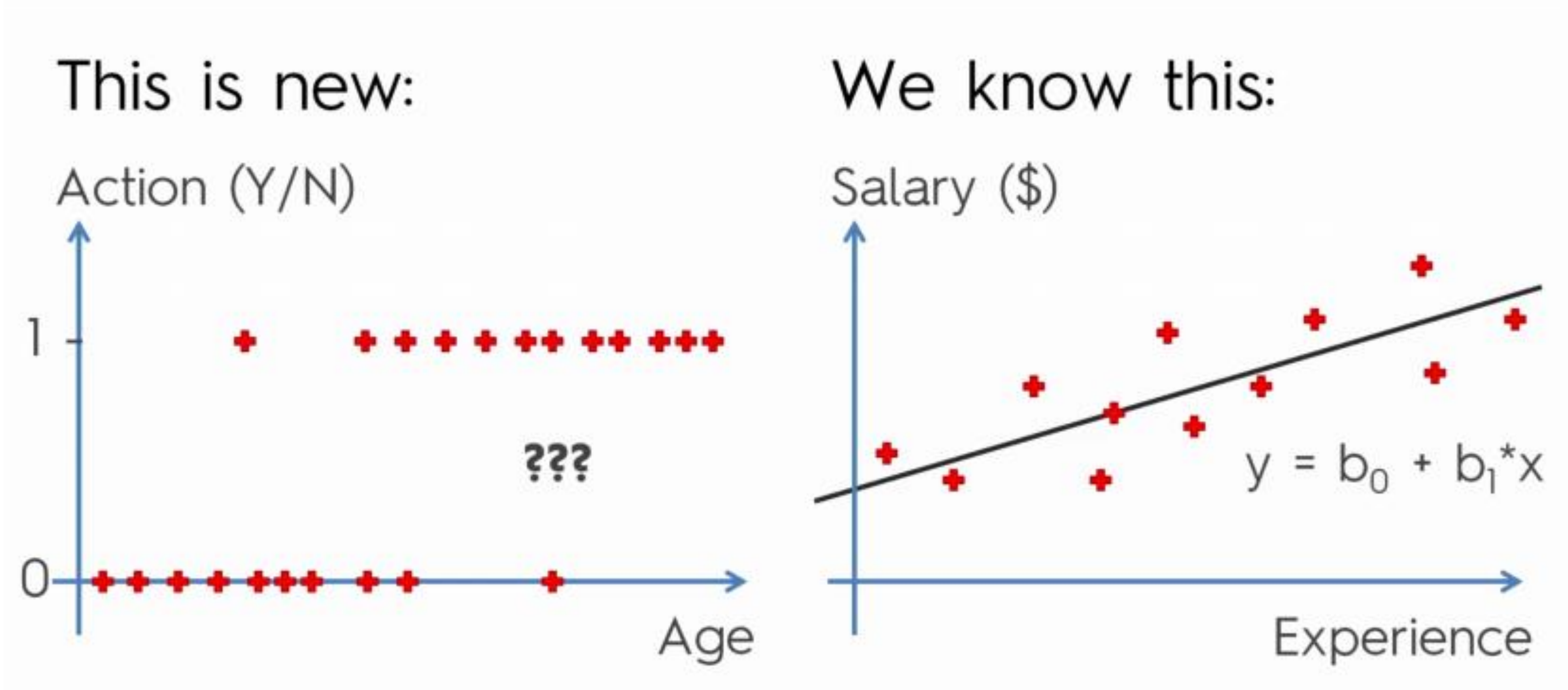
Constant

Coefficients

How many linear regression?

Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

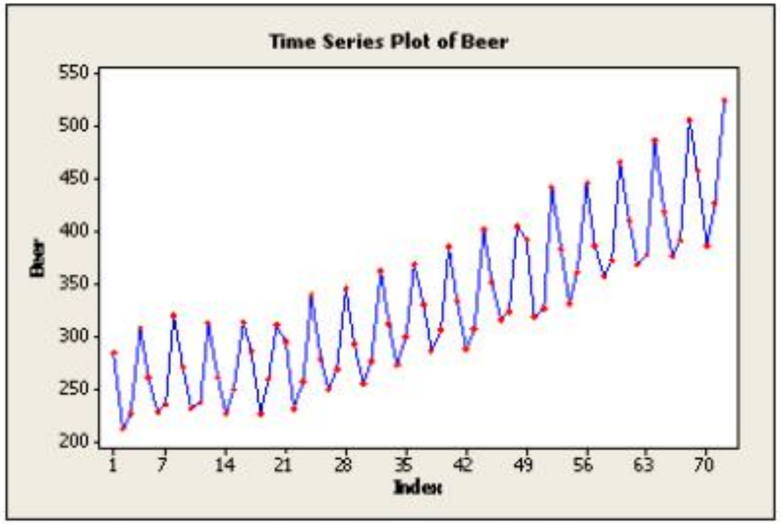
Logistic regression

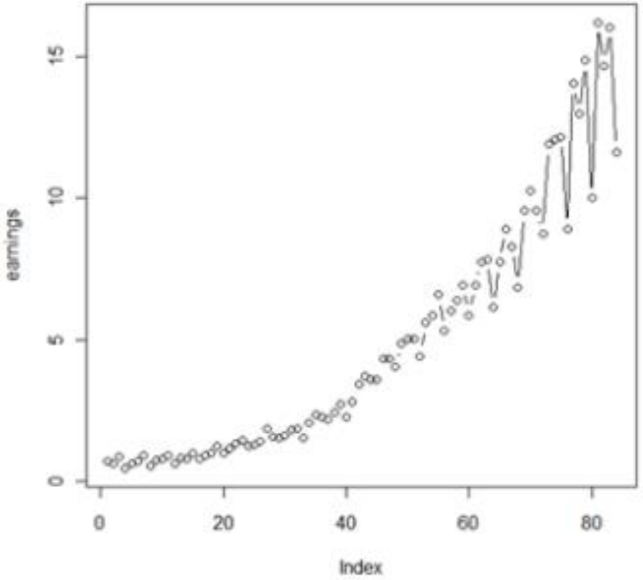


Linear Models in time series

The following two structures are considered for basic models:
1. Additive: Data = Seasonal effect + Trend + Cyclical + Residual
2. Multiplicative: Data = Seasonal effect * Trend * Cyclical * Residual

How to Choose Between Additive and Multiplicative
- The additive model is useful when the seasonal variation is relatively constant over time
- The multiplicative model is useful when the seasonal variation increases over time





Assumptions for Linear Regression

- Linearity
  - relationship between the independent and dependent variables to be linear.
  - check for outliers since linear regression is sensitive to outlier effects.
  - linearity assumption can best be tested with scatter plots

- Multivariate normality
  - This assumption can best be checked with a histogram or a Q-Q-Plot.
  - Normality can be checked with a goodness of fit test, e.g., the Kolmogorov-Smirnov test.
  - If data is not normally distributed a non-linear transformation (e.g., log-transformation) might work

Assumptions for Linear Regression

- No or little multicollinearity
  - Multicollinearity occurs when the independent variables are too highly correlated with each other

Testing for multicollinearity:

- Correlation matrix – computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1
- Tolerance – tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis. Tolerance is defined as $T = 1 – R^2$ for these first step regression analysis. With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly is.
- Variance Inflation Factor (VIF) – variance inflation factor of the linear regression is defined as $VIF = 1/T$. With $VIF > 10$ there is an indication that multicollinearity may be present; with $VIF > 100$ there is certainly multicollinearity among the variables

Assumptions for Linear Regression

Multicollinearity resolution steps

Might be able to ignore the multicollinearity if:

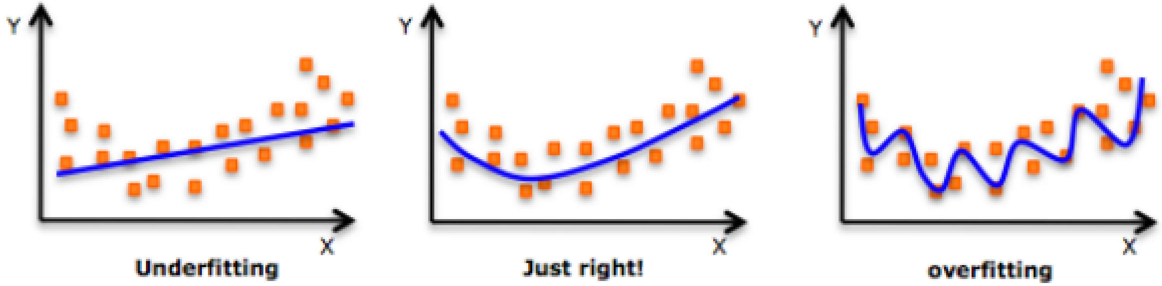
- The variables with high VIFs are control variables, and the variables of interest do not have high VIFs
- The high VIFs are caused by the inclusion of powers or products of other variables.
- The variables with high VIFs are indicator (dummy) variables that represent a categorical variable with three or more categories.

If cannot ignore them:

- conducting a factor analysis and rotating the factors to insure independence of the factors in the linear regression analysis
- remove the highly correlated predictor variable(s), starting with the least interesting variable(s).

Terminology related to regression

- The value of R-square is always between 0 and 1, where 0 means that the model does not explain any variability in the target variable (Y) and 1 meaning it explains full variability in the target variable.
- Parsimonious models are simple models with great explanatory predictive power. They explain data with a minimum number of parameters.
- Overfitting: an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably



- In shrinkage methods we don't actually select variables explicitly but rather we fit a model containing **all p** predictors using a technique that constrains or regularizes the coefficient estimates that shrinks the coefficient estimates towards zero relative to the least squares estimates. These methods do not use full least squares to fit but rather different criterion that has a penalty that:
  - penalize the model for having a big number of coefficients or a big size of coefficients
  - will shrink the coefficients towards, typically, 0.

P-value

A p-value is:
Provided H_0 is true then p is the probability a test-statistic will be more extreme than what was observed.

If p is small, e.g. $p < 0.05$ (usually 0.05, but not always), we reject the null hypothesis

If p is not small, we fail to reject the null hypothesis (never accept it).

Caution: It is not the probability of Ho being true.

Flow chart for Multiple Regression