Total Deviation = Explained Deviation + Unexplained Deviation/Residual/Error

$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$

$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (b_0 + b_1 x_i))^2$

$SST = SSE + SSR$

$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\bar{y} - \hat{y}_i)^2$

$R^2 = 1 - (SSE/SST) = SSR/SST$

Adjusted $R^2 = 1 - \{SSE/(n-p-1)\} / \{SST/(n-1)\}$

p = no. of independent variables

n = no. of observations

ANOVA table the SSE is Residuals

$F = \frac{R^2/p}{(1-R^2)/(n-p-1)}$

Adjusted R^2: $1 - \frac{(1-R^2)(n-1)}{(n-p-1)}$

Compare two models: $F = \frac{(R_2^2 - R_1^2)/(p_2 - p_1)}{(1-R_2^2)/(n-p_2-1)}$

$0 \le R^2 \le 1$

$-1 \le R \le 1$

Assumptions of Linear Regression

① Linearity Assumption ② Error terms are i.i.d each w mean zero and constant variance σ^2 ③ Predictor variables in multiple regression are assumed to be linearly independent

Common Problems: ① Nonlinear relationship between response & predictor. Check linearity using residuals vs fitted values ② Correlation of error terms. An important assumption we have is that error terms are uncorrelated. If they are correlated we have auto correlation ③ Heteroskedasticity (Non constant error variance) We need to transform the Y variable ④ Outliers ⑤ High-leverage points → Residuals vs Fitted (check linearity) → Normal QQ (check if normally distributed) → Scale Location → Residuals vs Leverage (find influential point, if any, with $C_i > 1$) ⑥ Multicollinearity: check using VIF or corr matrix

regress predictor X_j against other predictor (X). The resulting $R^2 = R_j^2$. $VIF_j = 1/(1-R_j^2)$. VIF > 5 mean there is multicollinearity

R^2 will increase with new variables or stay same if interaction term does not improve model. Adjusted R^2 may go up or down

OLS to reduce SSE. Indicator = Dummy

For n factors, n-1 are not encoded

Purpose of log transformation – make distribution more normal and get better fit like a higher R^2

Eg. Age Mid = $\sum 1_{0/w}$, If Age = middle

Amt Spent = $b_0 + b_1$ * Age Mid + b_2 * Age Old

Amt Spent = −6.12 + 0.002 Salary −4.81 Age Mid + 23.28 Age Old

At the same salary level, old customers spent more than young customers

Categorical variable w M values, you will need M−1 indicator (dummy) variables

Amt Spent = $b_0 + b_1$ Salary + b_2 Fav

If Fav = 1, Relevant slope = $b_1 + b_3$ + b_3 Salary Fav

Price = $b_0 + b_1$ Capacity + b_2 Private + b_3 House + b_4 Private*Cap + b_5 House*Cap. $b_4 + b_5$ is slope for private

$b_1 + b_5$ is slope for house

Interaction Term = Synergy

linear-linear. $Y = b_0 + b_1 X$. If X increase by 1 unit, Y changes by b_1 units

linear-log $Y = b_0 + b_1 \ln X$. If X increase by 1%, Y changes by $(b_1/100)$ units or $0.01 \times b_1$ units. Because $\ln(1.01x) = \ln 1.01 + \ln x = 0.01 + \ln x$

log-linear $\ln Y = b_0 + b_1 X$

Y changes by $(100 \times b_1)$ % for 1 unit increase in X. Proof $y = e^{(b_0 + b_1 x)}$

$\frac{dy}{dx} = b_1 e^{b_0 + b_1 x} = b_1 y \Rightarrow \frac{dy}{y} = b_1 dx$

$100\, dy/y = 100\, b_1\, dx$, let $dx = 1$.

Accurate % change in Y = $(e^{b_1} - 1) * 100$ for one unit change in X

log-log $\log Y = b_0 + b_1 \log X$

Y changes by b_1% for 1% increase in x

Proof: $b_0 + b_1 \ln(1.01x) = b_0 + b_1 \ln 1.01 + b_1 \ln x$

$= b_0 + b_1 \ln x + b_1 (0.01) = \log y + b_1 (0.01)$

Elasticity: % change in Y when X ↑ by 1%

b_1 captures elasticity in log-log.

Reason for log transformation: ① Achieve nonlinear r/s ② Make distribution more normal ③ Make variance more constant ④ Get better fit ↑ R^2

Quadratic Model

price = $b_0 + b_1$ lotsize + b_2 lotsize^2

$\frac{d(\text{price})}{d(\text{lotsize})} = b_1 + 2 b_2 \text{lotsize}$

Clear signal of non linearity ① residuals vs. fitted values indicate heteroscedasticity ② QQ plot shows non linearity ③ Residuals not normally distributed.

$\ln n + \ln m = \ln mn$

$\ln n - \ln m = \ln(\frac{n}{m})$

$\ln n^r = r \ln n$

Polynomial regression does not allow for an isolated interpretation of coefficients of non-linear variables.

Poly regression provides better approx. of relationship between dependent & independent variables. Outliers can greatly affect. Fitted model more reliable when built on larger n.

QQ plot is standardised residuals against theoretical quantities

Odds For: P(will happen) > P(won't happen)

Odds For = $\frac{p}{1-p}$, $p = \frac{\text{odds for}}{1 + \text{odds for}}$

$p(x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$, $1 - p = \frac{1}{1 + e^{b_0 + b_1 x}}$

$\frac{p}{1-p} = e^{b_0 + b_1 x}$, $\ln \frac{p}{1-p} = b_0 + b_1 x$

logit(p)

x ↑ by 1 unit, log odds ↑ by b_1, i.e. odds ↑ by factor $e^{b_1} \approx 100 \times b_1$ %

↑ cutoff value True-ve ↑ & False+ve ↓

↑ cutoff value False-ve ↑ & True+ve ↓

Sensitivity (recall) = $\frac{\text{true +ve}}{\text{true +ve + false −ve}}$

Specificity = $\frac{\text{true −ve}}{\text{true −ve + false +ve}}$

Roc Curve Plot Sensitivity against 1−specificity

Type 1 Error: null hyp rejected when it should not be rejected (i.e. true) (false +ve)

Odds against = $\frac{1-p}{p}$

Type 2 Error: Failing to reject the null hypothesis when it is false (false −ve)

null hypothesis represents the default position or assumption of no effect or no relationship between variables.

Precision = $\frac{TP}{TP + FP}$ ↑ cutoff ⇒ precision ↑ & specificity ↑ recall/sensitivity ↓

Orthogonality Assumption: $Cor(X, \varepsilon) = 0$ where $Y = a + bX$ is model & ε error term

① Voluntary Response Bias ② Non Response

$Y = b_0 + b_1 X + \varepsilon$

$b_{OLS} = \frac{cov[X,Y]}{cov[X,X]} = b_1 + \frac{cov[e,X]}{cov[X,X]}$

Orthogonality Assumption: $Cov[e,X] = 0$

When X is a dummy variable

$b_{OLS} = b_1 + \frac{cov[e,x]}{cov[x,x]} = b_1 + (\bar{e}_1 - \bar{e}_0)$

b_1 = treatment effect, $(\bar{e}_1 - \bar{e}_0)$ = selection bias

Natural Experiments: A treatment/event that just happened, not intentionally designed as an experiment

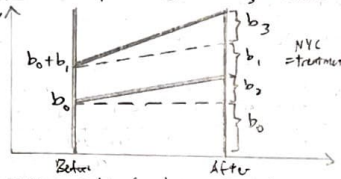
Difference In Difference

	Before	After	Difference
Control	A	C	C−A
Treatment	B	D	D−B

Diff-in-Diff = (D−B) − (C−A)

= (D−C) − (B−A) = (D−C) − (E−C) = D−E

Sales = $b_0 + b_1$ NYC + b_2 After + b_3 NYC·After



Diff in Diff is b_3 coeff of treatment X after

Analyse treatment effect using DID Estimator

DID is coefficient of interaction term

Diff estimator = $\frac{cov(y_1, y_0)}{cov(y_0, y_0)} = \frac{\text{mean}(y_1) - \text{mean}(y_0)}{}$

y_0 is set of obs in control group & y_1 is set of obs in treatment group

Natural Experiment = no intervention or manipulation by researchers

~~DID compares absolute values for not average~~

We regress the indicator variable on the other control variables & check if there are any significant coefficients. Random assignment is present if coefficients of the regression model are insignificant

Estimate causal impact of treatment, compare treatment outcome to the counterfactual. Avoid selection bias

① Randomised Control Experiments ② Natural Experiment ③ A/B Control

DID is comparison of the average change over time of dependent var for treatment vs control

$t = \frac{\text{price}_t}{\text{price}_{t-1}} - 1$

factor = stock split adjustment factor

Eg. 3 for 2 split ⇒ factor = 3/2 = 1.5

Compound Return = $\left[\prod_{i=1}^{t} (r_i + 1)\right] - 1$

$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ ↑ sd means ↑ risk

sd = total risk = firm specific + market wide

↑ stocks in portfolio ⇒ firm specific ↓ which means SD falls

$r_i = \alpha + \beta r_m + \varepsilon$, r_m = market return

β = stock sensitivity to overall market movements

R^2 = measure % of fund's performance that occurs as a result of the market

High beta ⇒ high market risk

$PD_t = \frac{HWM_t - P_t}{HWM_t}$

Sharpe Ratio = $\frac{\bar{x} - R^f}{\sigma(x - R^f)}$ $R - R^f$ = excess of portfolio above risk free rate

$\sigma(R - R^f)$ = sd of excess return

↑ sharpe ratio = ↑ reward / unit risk

Treynor Ratio = $\frac{\bar{x} - R^f}{\beta}$ ↑ treynor ⇒ ↑ reward / unit risk

Jensen's Alpha:

$r_{i,t} - r_f = \alpha_i + \beta_i (R_{m,t} - r_f) + \varepsilon_{i,t}$

return on asset i at time t; return on market index at time t

$\alpha_i > 0$ ⇒ fund outperformed

$\alpha_i < 0$ ⇒ fund underperformed

1 basis point = 0.01% = 0.0001

Types of fees: ① Commission ② Bid Ask Spread ③ Delay cost

Efficient Capital Market: i.e. where stock prices fully reflect available info

① Weak Form: security prices reflect all info found in past prices & volume

② Semi-strong: security prices reflect all publicly available info

③ Strong: prices reflect all info public & private

Behavioural Biases: ① Overconfidence ② Loss Aversion ③ Recency ④ Anchoring

Ask price: what seller is willing to sell at i.e. a buy order will be executed at the ask

Bid price: what buyer is willing to buy at.

Stock split does not change market value of the firm

5 for 3 stock split means 3 stocks becomes 5 stocks

① Size Effect ② Value Effect ③ Momentum Effect ④ Profitability Effect ⑤ Ritle Effect

① Size effect → smaller firms have higher returns than larger firms on average size measured by market cap.

② Value effect → inexpensive stocks tend to outperform expensive stocks expensiveness ⇒ book/market ratio

③ Momentum Effect → stocks that performed well in the past year tend to continue to perform well.

④ Profitable stocks tend to outperform unprofitable stocks

⑤ Risk Effect → low beta assets outperform high beta assets (volatility effect)

Factor Regression

Market Beta $r^m - r^f$, Size SMB, Value HML, Momentum WML, Risk BAB, Quality QMJ

$r_t^{fund} - r_t^f = \alpha + \beta(r_t^m - r_t^f) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 WML_t + \beta_5 BAB_t + \beta_6 QMJ_t + E_t$

+ve & significant coefficient indicates that the fund manager has outperformed

Low Book/Market ratio ⇒ stock inexpensive

Low B/M ratio ⇒ Value stock

−ve coefficient on HML ⇒ fund tilted towards growth stocks

Treasury bonds have historically been the safest asset.

Value factor = book value / price

Momentum = % change in price

Coeff of Size Factor −ve & significant ⇒ Large-Cap Fund.

Value Factor −ve & significant ⇒ Growth Fund.

Marketing and Advertising

Traditional methods ① outdoor display & promotions ② Print: newspaper & Magazines ③ mailings: door to door ④ Radio ⑤ Television

CPM = Cost per Mille or Thousand

Rating = Households watching a program / Total no. of households with TV

Share = Households watching a program / Total no of households watching TV

HUT = Households watching TV / Households with TV

GRP = Gross Rating Points

Reach = Households watching a program / Total TV households

Frequency = GRP / Reach

Digital Advertising: Email, Direct Messaging, Display Advertising

Digital Marketers: build campaign, buy media, optimize media campaign.

Ad Placement systems - audience relevancy & budget allocation.

Ad networks not restricted to work w only one media buying agency

2017 & 2018, search ads had highest revenue

2014 to 2018 mobile ads had more growth than desktop ads

2012 to 2018: Social media ads boom

2007 to 2018 Performance Pricing > CPM > hybrid model

Exponential Smoothing A_t: actual demand, F_t: forecast demand

$F_{t+1} = F_t + \alpha(A_t - F_t)$ Where $0 \le \alpha \le 1$

Forecast Including Trend $FIT_t = F_t + T_t$

$F_t = FIT_{t-1} + \alpha(A_{t-1} - FIT_{t-1})$

$T_t = T_{t-1} + \delta(F_t - FIT_{t-1})$

Seasonality Index = Period average demand / avg demand for all periods

① realized seasonal demand in previous period ($\sim A_t$), seasonal forecast for previous period ($\sim F_t$) & smoothing constant α

② $A_t = \sim A_t / SI_t$ and $F_t = \sim F_t / SI_t$

③ $F_{t+1} = \alpha A_t + (1-\alpha) F_t$ *Mod 4 Lesson 4

④ $\sim F_{t+1} = F_{t+1} \times SI_{t+1}$

error = actual − forecast

Mean forecast Error (Bias) = $\frac{\Sigma(A_t - F_t)}{n}$ } Running Sum Forecast Error

Mean Absolute Deviation = $\frac{\Sigma|A_t - F_t|}{n}$

Tracking Signal TS = RSFE ÷ MAD (|TS| < 4 ideal) for evaluating models

5 Methods of Digital Marketing

① Display Ads ② Search Engine Marketing ③ Social Media Marketing ④ Mobile Ads ⑤ Direct Messaging & Email Marketing

① Display Ads – in websites, 1st page, good for branding, low click through rates, works better with retargeting

Tracking Data – Impressions, clicks, time of day, device/OS, region

② Search Engine Marketing

PPC (pay per click) – pay only for each click of an ad

SERP (search engine results page) – Click Through Rate – total clicks on link divided by number of times an ad link has been shown

Landing Page – first page user reaches

Vickrey Auction Model

Quality Score Factors ① keywords ie relevance of keywords ② Ad Copy – ie relevance of ad copy to search term ③ Landing Page – relevance of the landing page to the search term ④ CTR of Ad based on history.

CPC Bid × Quality Score = Ad Rank

(max bid) (Keyword relevance + user exp) ↓ Ad position on page

max bid for keyword; how relevant & useful your ad is to user. Components are CTR, relevance & landing page

Profitability of ad-campaign: ① Cost Per Click ② Conversion Rate ③ Sale Value ④ Profit Margin ⑤ Repeat Behaviour/loyalty/lifetime Value of a customer

Conversion Rate = rate at which ppl who land on your site will go to your page, register or buy smth etc.

Awareness → interest → desire → research/comparison → purchase decision → payment → thanks → repeat

LTV (lifetime Value) of customer

Break Even Point (BEP) Analysis

* PPC Conversion Cost = Avg CPC (cost per click) / Conversion Rate in Decimals

Profit margin per sale = Avg sale Value × Profit Margin in decimals

BEP CPC per avg sale = Profit Margin per sale × Conversion rate in decimals

BEP CPC for LTV (Life time Value) = n × BEP CPC per avg sale

n = avg no. of sales per cust. over lifetime

SMART: Specific, Measurable, Attainable, Relevant, Time-Bound

Recall Exploration-Exploitation

Payment Models ① CPM – Cost per Mille ② Cost Per View (CPV) ③ Cost per Click (CPC) ④ CPA/CPL cost per acquisition or lead ⑤ CPS – Cost per sale

Display ads → size, position, duration & air time play a role in pricing

Only Search Engine Marketing considers a quality score

Digital Marketers: planning & building AD campaign → buying media or ad placements → optimizing media campaigns

CPS (cost per sale) payment guarantees tangible return (profit) for any money spent on ads.

Bounce Rate: gives an indication of the proportion of visitors who did not interact with the website

Conversion Rate Optimisation

Gather → Analyse → Correct → Design → Implement → Assess & Repeat

Goals: Increase Traffic, Enhance Engagement, Increase New Leads, Grow Sales, Improve Conversion to Leads & To Sale

Bounce Rate – proportion of visitors who did not interact with the website

Metrics for Google Analytics ① Users ② Bounce Rate ③ Sessions ④ Average Session Duration ⑤ % of new sessions ⑥ Pages per sessions ⑦ Page views ÷ total pageviews ÷ total no. of sessions

Advertising Steps: ① Campaign ② Ad Set ③ Ad

Ad = Audience + Creative a) Decide on an objective 1) Select an Target Audience 2) Upload the creative (of the ad)

Impt consideration for landing pages ① Heading ② offer ③ Images ④ Call to action

FB insight reporting tool Reports ① message to page ② people engaged with content ③ Time of a page ④ Daily & hourly activity

FB network → FB, Insta, Audience Network

Google behaviour report → shows people engaged, pages viewed, landing & exit

Risks of running ads for advertisers from riskiest to least: cost per view, cost per click, cost per sale

Hybrid Advertisements: cost per view + cost per click model

Landing page is the first page a user reaches when clicking on a link in an online marketing campaign

Google AdWords system quality score: keywords, Ad Copy, Landing Page and CTR

An organisation that attains a higher quality score relative to their competitors can realise a lower CPC & still rank higher

Planning & Setting up a paid search campaign: establish budget, conduct keyword search, define business goals supported by market research

Reasons for supply interruptions ① Labor shortages ② Supplier failures ③ Inaccurate inventory ④ Poor Forecasting.

Large impacts on supply chain operations ① Autonomous Vehicles ② Robotics ③ AI ④ 3D printing

Process Capability (CP) Ratio Week 10 Module 3 Lesson 1

$C_p = (\text{upper specification} - \text{Lower specification})/6\sigma$

$C_p \ge 1.0$ indicates a process is capable

Process Capability Index (Cpk)

$C_{pk} = \min[\hat{z}(\text{upper specification} - \bar{x})/3\sigma\}, \hat{z}(\bar{x} - \text{lower specification})/3\sigma\}]$

$C_{pk} = 1 \Rightarrow$ process meets specifications

$C_{pk} < 1 \Rightarrow$ does not meet specifications

$C_{pk} > 1 \Rightarrow$ outperform specifications

Operation Management

Improve efficiency, effectiveness, value

Queuing: Strategic, Tactical, Operational

Customer Pop'n → Arrival → Queue → Service → Exit

Arrival Rate λ – mean arrival rate

Exponentially distributed interarrival time $f(t) = \lambda e^{-\lambda t}$ OR

Poisson Distributed arrivals per unit time $P_T(n) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}$

Service Rate (μ) → how many channels & how many phases?

Exponentially distributed service time $f(t) = \mu e^{-\mu t}$

λ = arrival rate, μ = service rate

Utilization = $\rho = \frac{\lambda}{\mu}$ (M/M/1)

Avg # of customers in system = $\frac{\lambda}{\mu - \lambda} = L_s$

Avg # of customer in queue = $\frac{\lambda^2}{\mu(\mu-\lambda)} = L_s \times \rho$

Avg time a customer spends in the system = $W_s = \frac{1}{\mu - \lambda} = \frac{L_s}{\lambda}$ in hours

Avg time a customer spends in the queue = $W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{L_q}{\lambda}$ in hours

Prob. of n units in the system $P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)(\rho)^n$

Eg. customer arrival 15 per hour, staff can service customers 1 every 3 mins

$\lambda = 15/h$, $\mu = 20/h$

Utilization: $\rho = \frac{\lambda}{\mu} = \frac{15}{20} = 75\%$

Avg # of customers in the system: $L_s = \frac{\lambda}{\mu - \lambda} = \frac{15}{20-15} = 3$ cust

$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = L_s \times \rho = 3 \times 0.75 = 2.25$ cust

$W_s = \frac{1}{\mu - \lambda} = \frac{L_s}{\lambda} = 3/15 = 0.2h = 12$ mins

$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{L_q}{\lambda} = \frac{2.25}{15} = 0.15h = 9$ mins

Eg. Ensure with 95% confidence (or service level) that no more than 3 cust will be in the system at any time. What confidence exists currently for no more than 3 cust?

$P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)(\rho)^n$, $\rho = \frac{15}{20}$

$P_3 = (1-0.75)(0.75)^3 = 0.105$

$P_2 = (1-0.75)(0.75)^2 = 0.141$

$\vdots$

$P_0 + P_1 + P_2 + P_3 = 0.25 + 0.186 + 0.141 + 0.105 = 0.684 \rightarrow 68.4\%$

For 95% service level we want: $P_0 + P_1 + P_2 + P_3 = 0.95$

$1 - \rho^4 = 0.95 \Rightarrow \rho = 0.47$

$\rho = \frac{\lambda}{\mu} \Rightarrow 0.47 = \frac{15}{\mu} \Rightarrow \mu = 32$ ppl per hour

Quality-price Trade-off | Queuing System

① Line Length ② Queue Discipline ③ Service rate ④ Arrival Rate | M/M/1 model

① Random Arrivals ② Single Line ③ DO customer pop'n ④ FCFS

$S_n = \frac{a(1-r^n)}{1-r}$ *SA 9

Garvin 8 Dimensions of Product Quality: ① Performance ② Functionality ③ Durability ④ Reliability ⑤ Conformance to Specifications ⑥ Serviceability ⑦ Aesthetics ⑧ Perceived Quality

Dimensions of Service Quality ① Consistency ② Courtesy ③ Convenience ④ Communication ⑤ Accuracy / Reliability ⑥ Timeliness / Responsiveness ⑦ Credibility / Trustworthiness ⑧ Security

Cost of Quality ① Appraisal Cost ② Prevention Cost ③ Internal Failure Cost ④ External Failure Cost

Cost of Good Quality → ① & ②

Cost of Bad Quality → ③ & ④

Statistical Process Control monitoring a variable ① Collect Data ② Calculate $\bar{R}$ ③ Calculate UCL_R and LCL_R ④ Plot R-chart ⑤ Calculate $\bar{\bar{x}}$ ⑥ Calculate $UCL_{\bar{x}}$ and $LCL_{\bar{x}}$ ⑦ Plot $\bar{x}$-chart

* Week 10: Lesson 6 + Lesson 7 Module 3

$UCL_R = D_4 * \bar{R}$, $LCL_R = D_3 * \bar{R}$

$UCL_{\bar{x}} = \bar{\bar{x}} + A_2 * \bar{R}$, $LCL_{\bar{x}} = \bar{\bar{x}} - A_2 * \bar{R}$

Variation: random/common causes, assignable/special causes

With only common causes of variation variable will be Normally Distributed.

UCL & LCL are based on common causes of variation

Investigate causes of variation when ① A point outside of UCL & LCL bound ② 2 points near upper control ③ Erratic