EDA → What kind of variables
→ What distributions
→ anomalies ?
→ interesting associations ?

center/location → Mean
Median (outliers x)
Mode (categ)

Quant
numeric, Ratio interval

categorical
Nominal
ordinal
logical order

Location
↳ 25th, 50th, 75th percentile

categor quant
reference values
or
dummy
optimal values

spread/variation
↳ diff btw Min & Max Range
third & first quartile IQR
variance & s.d (dispersion/data around mean)

symm/normal
more than 3 S.D mean

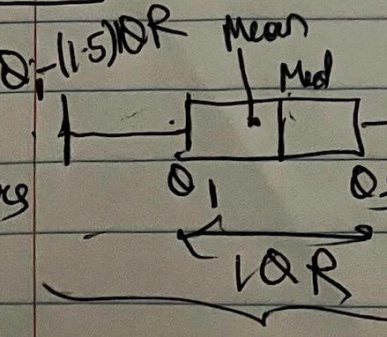
shape uni / bi / complicated model
left skew / right skew (long tail)
Kuertosis (thin or thick tails normal)

Boxplot
1.5 IQR 3Q
less than
1.5 IQR 1Q

Normal dist assumptions → symmetric, defined mean & sd, bellshaped, m=md=mod

distributions → hist, normprob/QQplots, boxplots

assymptotic to X axis (boundaries -∞ to ∞)



deviation from normality

small s.d data close to mean
large s.d → data away from mean or spread out

Why EDA → make necessary assumptions about analysis & apply
↳ understanding variables, relations
↳ which stat analysis might be appropriate

Nominal – no logical ordering
↳ while modeling converted dummy variables
↓
indicate presence/absence

Linear reg → reference/one hot effect
ML models → one hot / other variating
↳ dummy variable for each level
↳ similar to one hot but reference level is dropped

Interval variables → arithmetic operations ✓

Continuous → quantitative can take any value within range
Categorical → qualitative, represent types or categories

ordinal variables → treated as interval variables / categorical variables

Histogram – continuous
boxplot → o nominal

Mean – avg
Median – middle
mode – most common

IQR = 25th perc – 75th per.

s.d = $\sqrt{\text{variance}}$

$V = \sqrt{\frac{\sum (x - \text{mean})^2}{n-1}}$

→ s.d preferred same unit of variance ✓

caddy

Anomalies → normal dist 3 SD from mean
BOXplot → > 1.5 IQR Q3
< 1.5 IQR Q1

Point Estimates → sample statistic → estimate of pop parameters
↳ xp of populatn
see. to know more about

Variability among Samples

CLT → dist of sample mean is normal regardless of pop dist's shape, if sample size is large enough.

Standerro. (Sx) ↳ variability four estimate

Stadevia → variability in data

$S_x = \frac{S}{\sqrt{n}}$ → provide info variability around statistic

CI → contains pop parameter with some degree of confidence

CL → % indicats strength four confidence that interval created actually captures true paramdr.

95% crated intervals capture pop paramdr

Hypothesis → if true pop value' sig diff from assumed value

statemnt we wish to investigate scientifically through process of stat inference

① Initial assump, Null hypo

② analyze data to determine other observations are likely given null hypo

③ if given data observed data unlikely reject initial assumption favour of alternate

α → sig level how much evidence req to reject null hypo

p → probabilty that we observed something as extreme or more extreme than we did if $p < \alpha$ reject null

Type I → reject null

Type II → fail to reject null

t test

mean.

One sample t test

mean against hypovalue

2 side t test

t can be +ve or -ve

hypovale → inside CI
- tail to reject null

→ outside CI
- reject null

one side test

saleprice > 170k$

shape, var, symmetric

* stat significance → observed data is not due to random chance

2 sample t test

↳ testing diff btw 2 means

assump: ind observations

normality each group

equal var

→ QQ, shapiro

F test (normality ✓)

2 sample t test → with equal
→ without equal

when normality fails
↳ wilcoxon rank test

Manwhitney w/ coxen test
diff in means/medians/dist dominance

ftest, shapiro, ttest

null → equal mean
- equal var
- normality

Response \ predictor	catg	cont	cont/catg
cont	ANOVA	OLS	OLS
catg	Log	Log	Log

Honest Modelass

Linear Model

↳ ~~explanatory~~ response

→ explore dist, outliers, missing

→ then split before relationship

↓

train / valid / test

Exp

→ x related y

→ descriptions

→ few var & simple model

→ focus on pvalue & confid intervals

pred

→ x_i can predict y

→ predictions focus

→ many variable & complex

overfit *

↳ model captures nuances on train

→ pattern not hold in validation or test

→ performance suffers

doesn't generalize

model overfit

50-40-10 Lots

70-20-10 not so much not enough

cross

Bivariate EDA

expected value of one variable changes at diff levels of another

(1) linear association scatterplot straight line

(2) grouped boxplots overlaid hist cont vs catego

One way Anova

categ input & quant response

H_0 → means of each level equal

H_a → atleast one diff

Null → means equal

Alt → means not equal

→ we do reference level coding

$y = \beta_0 + \beta_A x_A + \beta_B x_B + \epsilon_i$

$x_i \rightarrow y = \beta_0$

β_A diff means for A vs first c

① EDA

② verify Ass

③ pvalue of overall F test in Anova

$P < \alpha$ reject null

Assump → Ind observations → good data collect

→ groups normally (residuals) → QQ, hist / formal test

→ equal var → classic / welch → formal test (homoskedasticity)

when normality fails → kruskal wallis

tails → Levene's → Fligners

requires norm → no normality

post hoc Anova

→ which groups diff, multiple 2 sample t tests

multiple hypothesis

compounds error

Exp wise Error Rate

Comparisons each having error rate

poop that 1 make atleast one error

$1-(1-\alpha)^n$

to control → Tukey (all pairwise)

Dunnett (comparisons to control group)

cont Resp & continuous pred.

Corr (-ve to +ve)

→ Pearsoncoeff → measures linear relation btw 2 continuous attributes

→ if we don't explore data sometimes anamalous observations declare relationships are significant & outliers skew

→ raw correlation misleading so scatterplot

corr value, test, plot

→ can mislead when both variables affected by other variable

SLR

→ how much one variable changes as other changes by one unit

→ creates Linear eq by Minsquared btw observed data & model predictions

OLS → Min SS residuals

Loss function

→ predict expected value of y for each value of x

→ y is expected to change for unit change in x

→ purpose

$Y = \beta_0 + \beta_1 x + \varepsilon$

determ → random

H0 $\beta_1 = 0$ $\beta_1 \neq 0$

global F test

para t test } equal

pearson

Assump

↳ linearity of mean

errors are normal

hist, QQ normality

Res Vs pred nice band

→ equal Var

→ Independent

data collection

or explore for auto correlation

no pattern in residual plot

2 way
n way Anova

2 ways → 2 variables
K_1 & K_2
categories

Model → relation btw exp var & Resp var

Effect → expected chg in rspns with value of explanatory

2 ways 1] EDA
2] 2 way anova → each variable has F test
post Hoc test

3] Main effect → effect of single explanatory var

Inter actions → effect of one variable changes as levels of another

mask effects of variables

Slicing → ftest mean of one
within effects testing

ANOVA essentially a linear reg

assump → Indep observations
→ Eq of variance (errors)
→ Norm of categories (errors)

Multiple Linear Reg → linear comb of variables

→ more than one predictor

β → predicted change in Y with one unit increase in X given all other variables held constant *

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ ✓

$\beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \varepsilon$ ✓

$\beta_1 = \beta_2 = \ldots = 0$

Global → None of var useful
Ftest → atleast one variable useful } ①

If Global Ftest sig → Local test → individual ttests ②

Multicollinearity

→ pred var are correlated with each other

→ Linear reg breaks only when collinear perfect

Adv

↔ investigate relation among Y & several indp variables

Disadv

↑ complexity makes difficult

→ which model best

→ interpret model

Assump Linear Reg

① mean of Y accurately modeled by linear fn of X

② random error term assumed to have normal dist mean '0'

③ → const var

④ errors independent

Application

Predict

focus on predictions

value & coeff & stat sig less important

Explain to develop underst pred var & Resp var

R^2 → prop of variances in dependent var explained by independent var

R^2_{adj} → consider no of variables & penalizes excessive variables

* with addition of variables R^2 ↑ in MLR so R^2_{adj} → penalizes model for adding variable that doesn't provide any useful information

Categorical variable in MLR → dummy encoded

1,0 = X

Avg diff between catg Y & N a $\beta_0 + \beta_1 X$

effect coding

average diff between catg Y & overall avg of cat Y & N.

Model selection

Information → used to select variables for model

Selection Algorithm → auto tchq to evaluate variables on some selection criteria

- stepwise → for, back, stepwise
- Allreg selection (R^2, Ra^2, Mallow(p))

Model selection based on training data (to avoid overfitting)

AIC → crude, large sample approx of leave one out CV

BIC → favors small models / penalizes model complexity more.

→ low value better than higher, no amount of lowering better.

forward: start with intercept, build the model one variable at a time

backward: start with full model & remove variables

stepwise: start with empty, can add & remove

dir = forward, direction = both

K = 2 AIC

K = log(nrow(train.sel))

BIC selection

K = qchisq(0.05, 1, lower.tail = false)

↳ p value selection with alpha 0.05

Issues with AutoSelAlgo

→ bias in parameter estimates, pred & standard errors

→ Incorrect cal of DOF

→ pvalue tend to err on side of overstating significance. (Type I error)

→ load best not global

→ don't use result from ASalgorithm blindly

→ Adjust pvalue for large samples.

Conservative pvalues

sample size ↑ pvalue ↓

→ Automated stepwise can provide subset of potential variables

→ no model blindly final model

explore other models & investigate assumptions

Diagnostics

4 Reg assumptions

Violated

1. mean of Y accurately modeld by linear func of X → misspecified results not trusted
2. constant var random error → S.E compromised → coeff ✓ parameter estimates
3. random error → normal dist with mean zero → parameter estimates ✓ effects test results
4. errors are independent → not effect parameter estimates standard error compromised
5. no perfect collinearity

Residual plot → randomly scattered

Misspecified → pattern detected (poly, interaction, splines...)

* **Model linearly** when adding higher order, all lower terms included in model

when straight line not appro

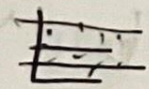
→ fit poly / complex reg model

→ transfor dep / ind var to obtain linearity

→ fit non linear reg model;

→ fit non param reg model (LOESS)

Const Variance (homoscedacity)



→ examine residual plots or spearman Rank correlation

→ if violated (Hypothesis, & CI based on t, f, x^2 not valid)

0 → Variance is potent homo

+ve V ↑ mean ↑

-ve V ↓ mean ↑

if relation is not linear test will not detect it

corr coeff btwen absolute value of residuals & predicted values

→ accounting hetero [use WLS
transform data
use diff distribution

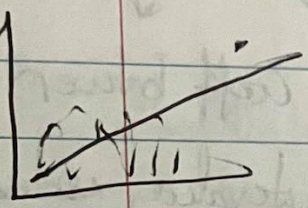
Normality → histogram of residuals
→ Normal probplot (QQplot)
→ formal tests for normality

accounting for lack of normality
→ outliers (robust reg)
→ nonnormal (transf needed)
can by BoxCox transformation

Outlier → large standardized residuals, far away fitted known Y direction

Influential

Leverage points → points fall outside normal range X space
large influence on regression line.



Residual only focus on outliers

Diagnostics → standard, studentred } good for outliers

CookSD
DFFITS
DFBETAS
Hat values } good for influential observations

How to handle Influential observations

↳ Recheck data to ensure no transcription or data entry

→ data is valid model adequate
[high order terms poly & interaction
nonlinear model

↕

run analysis with & without influential observations

→ Robust Regression
→ WLS

Collinearity → look at corr matrix
→ use VIF (> 10 potential colli
> 5 GVIF ...)

dealing
→ exclude redundant independent variables
→ Redefine variables
→ use biased Reg techniques
→ center independ var in poly reg models
or model with interaction term.

Model building & Scoring for prediction

Linear reg — BLUE
best linear unbiased estimate

→ if assumptions hold best linear min variance of all unbiased estimators.

if assumptions don't hold?
if biased estimators had small variance?

no. of var ↑ → assump ×
multicollinearity
↳ coeff vary
leads to overfitting
high variance

→ More var than observations

Regular Reg — puts constraints on estimated coeff & shrink estimate to 0
coeff biased but improve variance

→ Reg Reg add penalty term to min(SSE + penalty)
loss function

↳ penalty alpha=0 ridge → approach to 0 — $\lambda = 0$ OLS, $\lambda \to \infty$ coeff → 0
↳ penalty alpha=1 — Lasso → coeff = 0, variables removed

Lasso → does variable selection

Ridge → Keeps all variables

ElasticNet → Lasso + Ridge alpha between 0 & 1

ok

Fear of Overfitting → need to select λ for any regularised regression approach

→ don't want to minimise variance to point of overfitting to training

Crossvalidation prevent overfitting

Modelmetrics RMSE not easily interpretable

MAE not scale invariant

MAPE not symmetric

Categorical Data Analysis

→ when samples are smaller diff can't compare

→ data freq of data
→ possible association among variables

Crosstable function in R → only associations exists

Chisquare test

Null
Alternate

Pearson
likelihood
→ any type of categorical variable.

Sample size Req
80% or more expected cell count > 5

when we don't meet assumptions
↳ fishers exact test

ordinal more info → Mantel Haenz chisquare test

---

strength of association [odds ratio, Cramer's V
Spearman's Correlation]

odds = p/1-p

odds ratio → how much more likely wrt to odds a certain events occurs in one group relative to its occurence in another

Cramer's V — more than 2 categories in one or both variables → use Cramer's V

odd's ratio — B vs B

Spearman corr → strength & association btw 2 ordinal variables

↳ Pearson on rank instead of value of observations

Continuous Target → Linear Reg
Anova
~~Lo~~ Reg Reg

Categorical target → Logistic Reg

Intro to logistic Reg

↳ pred probability between 0 & 1

→ parameter estimates don't enter model equation linearly

→ rate of change of probs vary's X vary's

→ logit applied to probabilities $\log\left(\frac{p_i}{1-p_i}\right)$

→ relation btw logits linears log applied to probability

→ logits unbounded

assumptions ↳ Indp observations

↳ logit is linearly related to variables

→ $100 \times (e^{\beta} - 1)\%$ change in odds

→ β is change in logit

we want high # concordant

- Concordant (0,1) ↳ high pred prob
- Discordant (0,1) ↳ high pred prob
- Tied (0,1) bonus eligible has same pred prob as non bonus eligible

(glm(↳function)

Variable selection

↳ for, back, stepwise, Lasso

→ Reg Reg can use same link function to obtain logistic Regression