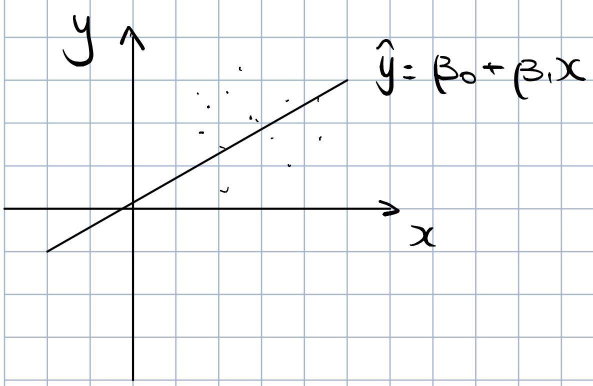
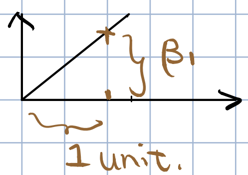
Intuition on Stats.

Start from a simple linear regression, which has 1 predictor.



β_1 : slope.



$\Delta x \to \Delta y : \beta_1$

model : $\hat{y} = \beta_0 + \beta_1 x$.

model assumption : we expect / model predicts

$\hat{y}$: on average

slope : $\Delta x \to \Delta y$: β_1

"1" Δx : - For numeric var, 1 unit.

- For categorical var with k levels.

"1" can represent 1 2 3 ... k

any gap between two neighbors

"distance" between neighbors may vary!

$\boxed{k-1}$ different representation of "1" Δx

need $\boxed{k-1}$ different β's to estimate Δy resulted from different "1" in Δx.

method 1. 1 2 3 4 ... k $\Rightarrow$ baseline change.

method 2. 1 2 3 4 ... k $\Rightarrow$ baseline fix : level 1.

1 $\Delta x \to \Delta y$:

baseline		increment	
β_0			level 1
β_0	+	$\beta_{1.1}$	level 2
β_0	+	$\beta_{1.2}$	level 3
	$\vdots$		$\vdots$
β_0	+	$\beta_{1.k-1}$	level k

Interpret $\beta_{1.1}, \cdots, \beta_{1.k-1}$. as " increment " :

- " slope " : $\Delta x \to \Delta y$

Compare to level 1 (baseline), level 2 has Δy : $\beta_{1.1}$

model : $\hat{y} = \beta_0 + \beta_{1.1} I(\text{level } 2) + \beta_{1.2} I(\text{level } 3) + \cdots + \beta_{1.k-1} I(\text{level } k)$

model assumption : we expect / model predicts

$\hat{y}$: on average

slope : $\Delta x \to \Delta y$: β_1

compared to level 1, level 2 will have Δy : β_1

For model with 1 predictor. (categorical)

$$\hat{y} = \beta_0 + \beta_{1.1} I(\text{level}2) + \beta_{1.2} I(\text{level } 3) + \cdots + \beta_{1.k-1} I(\text{level } k)$$

	I(level 2)	I(L3)	I(L4)	...	I(Lk)
level 1	0	0	0		0
2	1	0	0		0
3	0	1	0		0
$\vdots$					
k	0	0	0		1

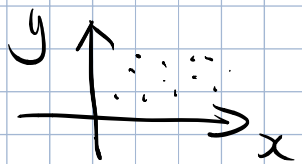
Already multivariate Linear regression !

Special case ,

one observation can only have 1 level ,

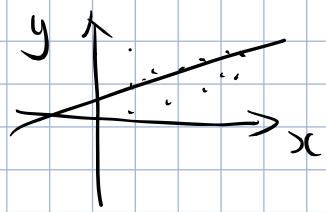
SLR.

Y . x



EDA $\to$ linear trend.

$\downarrow$



linear regression

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

model fitting: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Interprention of $\hat{\beta}_0$. $\hat{\beta}_1$

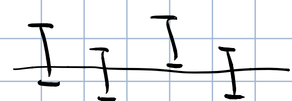
prediction : $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

extrapolation.

evaluation : R^2

RMSE.

Inference :

CI  true.

HT $\beta_0 = 0$ vs $\beta_0 \neq 0$

method

simulation
- CI : bootstrap
- HT : permutation

math. ($\varepsilon \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$)

CI $\hat{\beta} \pm t^* \times se(\hat{\beta})$ $\to$ degree of freedom

HT : $t^* = \dfrac{\hat{\beta}}{se(\hat{\beta})}$ p-val $= Pr(|t| > t^*)$

Interpretation.

Predictive Interval.

Diagnosis:

Condition (4) ⇐ residual plot

Influential pts:

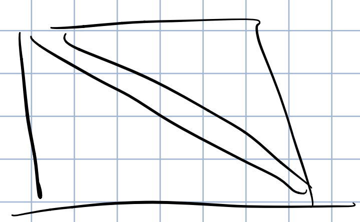
leverage

standardized residuals

Cook's Distance

MLR.

EDA : ggpair



Model assumption : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$.

Model fitting : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

Interpretation : - Hold constant - on average

- expect - $\Delta x \to \Delta y$

prediction : $\hat{y}^* = f(x^*)$

Type of predictors:

- mean-center
- categorical : dummy variable
- Interaction : $\Delta\beta$
- Transformation : $\log(\cdot)$; $x^2 + x$

Model comparison :

ANOVA

Adj R^2

AIC

BIC

Feature engineering -

split data

recipe .

workflow

fit model

prediction .

CV

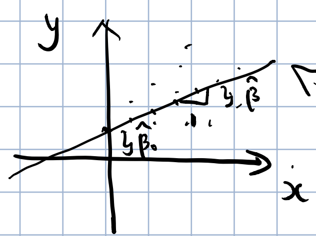
conditions : residual plot

+ multicollinearity .

Simple Linear Regression.

$y \sim x$

EDA.



Model Assumption : $y = \beta_0 + \beta_1 x + \varepsilon$

Fit : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \rightarrow$

tidy( ) : point. se. t-stat. p-val.

meaningful? $\rightarrow$ $\hat{\beta}_0$: $x = 0 \Rightarrow \hat{y} = \hat{\beta}_0$

center x.

interpret.

$\hat{\beta}_1$: $\Delta x \rightarrow \Delta \hat{y}$: $\hat{\beta}_1$

* we expect / predict.

* on average

* 1 unit. $\rightarrow \hat{\beta}$..

prediction : $y^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

$x^* \rightarrow$ extrapolation.

evaluation : R^2 ↑

RMSE. ↓

Inference. { CI : ~~I I I I I~~
HT

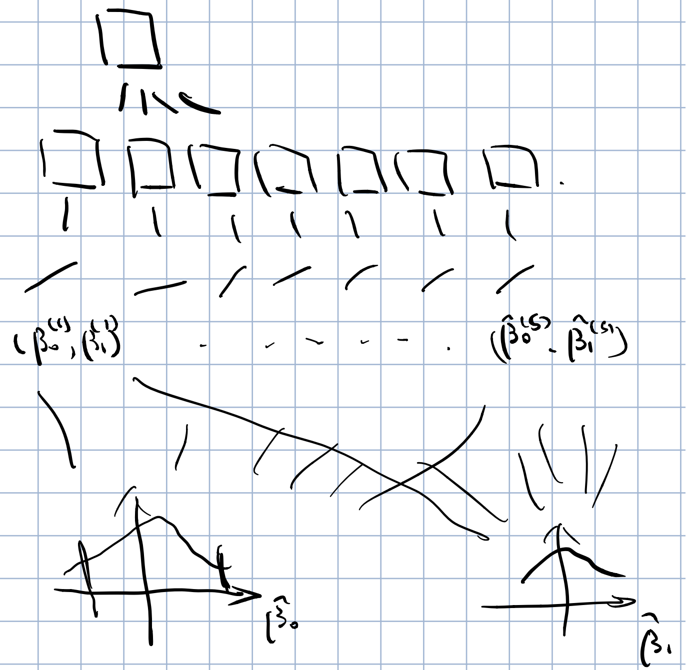
Interprete. CI : - 95%.

- model predict / expect.
- on average.
- Δx ↑. [,].

- CI for β
- CI for $\bar{y}$ individual.
- prediction interval. > CI. $\bar{y}$. average mean.

Simulation. Bootstrap: sample w/ replacement

math.



Quantile. 0.025 0.975.

math: $y = \beta_0 + \beta_1 x + \varepsilon$

$\varepsilon \sim N(0, \sigma_\varepsilon^2)$

$\hat\beta \pm t^* \times se(\hat\beta)$

t_{df}

$df = n - p - 1 = n-2.$

#predictor

HT:

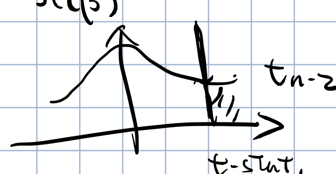
$H_0: \beta_1 = 0$ vs: $H_1: \beta_1 \neq 0$

Assume H_0 is true.

check obs. is reasonable.

Math: t-stat. $\dfrac{\hat\beta}{se(\hat\beta)} \sim t_{n-2}$

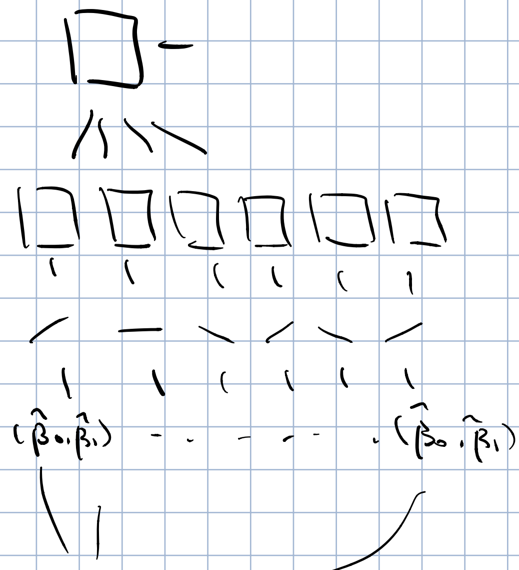
p-val

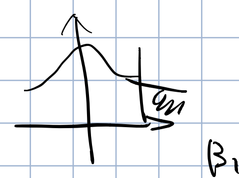


$Pr(|t^*| > t)$ p-val. small.

reject H_0.

Simulation : permutation. (sample w/o replacement)

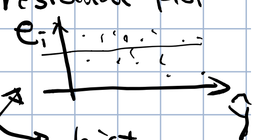




Calculate p-val.

Diagnosis:

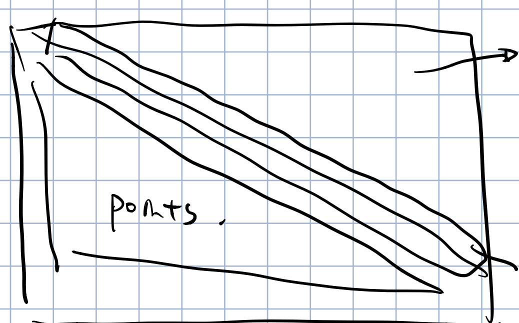
- check conditions.
  - linearity → residual plot
  - ind → context
  - normality → hist
  - constant var.

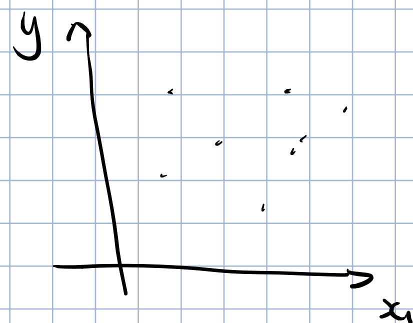


- influential points.
  - outlier. ← standardize res. > 3
  - high leverage. ←
  - → Cook dist. : 0.5, 1.

MLR.

EDA : ggpair. (·)





size.
color.
⋮

Model : $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$.

Fit : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p + \varepsilon$.

hold other constant.

Interpretation. $\hat{\beta}_1$ (4) - on avg - we expect. - 1 unit $\to \cdots \Delta y : \beta_1$

- hold other constant.

Δx : numeric var $\quad \Delta x \to \Delta y : \hat{\beta}_1$

Δx : 1 unit increase.

Δx : categorical var. k level. Δ

1 2 3 4 .... k

Δy β_1 β_1 β_1 β_1

Δy 1 2 3 4 ... k $\quad$ fix reference level. 1.

k levels. $\quad \beta_{1,1}$ $\beta_{1,2}$ - - ... $\beta_{1,k-1}$.

k-1 dummy.	$I(2)$	$I(3)$	----	$I(k)$.
1	0	0	....	0
2	1	0	...	0
⋮				
k	0	0	. . --	1.

$y = \beta_0 + \beta_1 \underline{x_1} + \cdots$

x_1 is k-level cate...

$y = \underline{\beta_0} + \underline{\beta_{1,1}} I(x_1 = 2) + \cdots + \underline{\beta_{1,k-1}} I(x_1 = k)$

β_0 increment. $\quad$ - $\Delta x \to \Delta y$: compare to level 1.

$\beta_0 + \boxed{\beta_{1,1}}$ $\quad$ we expect the y of obs w/ level 2. will increase by $\beta_{1,1}$, on avg. hold other constant.

⋮ + ...

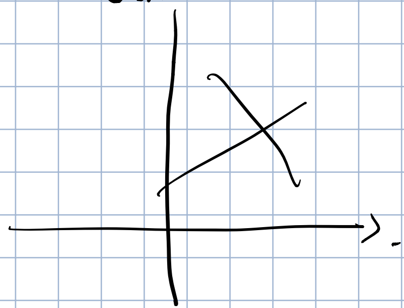
$\beta_0 + \beta_{1,k-1}$

baseline. $\quad \hat{\beta}_0$: obs. with level .....

Interpretation : 1. numeric vs. categorical.

2. interaction.

EDA



w/o int. same slope.
diff intercept

w

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

x_1 : numeric x_2 = categorical.

w/o

$$y = (\beta_0 + \beta_2 x_2) + \beta_1 x_1$$

intercept.

$$y = \underbrace{(\beta_0 + \beta_2 x_2)}_{\text{inter}} + (\beta_1 + \underbrace{\beta_3 x_2) x_1}_{\text{slope.}}$$

β_0

$\beta_0 + \boxed{\beta_{2.1} \ \vdots \ \beta_{2.k-1}}$

baseline + increment.

β_1

$\beta_1 + \boxed{\beta_{3.1} \ \vdots \ \beta_{3.k-1}}$

Δy

$\Delta \beta_1$:

$\hat{\beta}_{3,1}$

- $\Delta x_2 \Rightarrow \Delta \beta$: $\hat{\beta}_{3,1}$
- expect.
- on avg
- hold other const.

the potential influence of

Compare to level 1, we expect obs w/ level 2, on outcome, increase by $\hat{\beta}_{3,1}$ on avg, hold other constant.

- mean center.

$\Delta x \to \Delta y$

- log. transformation.

$$y = \beta_0 + \beta_1 \log x_1$$

$\log(x_1 + \Delta x)$

$\log(x_1 \cdot c) = \log x_1 + \log c$

$x + 0.1x$.

$\Delta x \to \Delta y$.

$$y = \beta_0 + \beta_1 \log(1.1x) = \beta_0 + \beta_1 \log x + \beta_1 \log 1.1$$

x increase by 10%.

y. $\beta_1 \log 1.1$

$y = \beta_0 + \beta_1 \log x + \varepsilon$. ? linear regression.

Yes,

linearity w.r.t. β.

Type of predictors.

Model comparison.

- anova. $H_0: \beta_1 = \cdots = \beta_k = 0$. p-val. small. reject H_0.
- Adj. R^2 : ↑
- AIC, BIC : ↓.

Feature engineer: extract feature from predict.

represent predictor.

spend data: split (training) + test

prediction.

model fit.

create recipe:

- recipe(...) } recipe.
- step_...

f... } model.

recipe + model. } workflow.

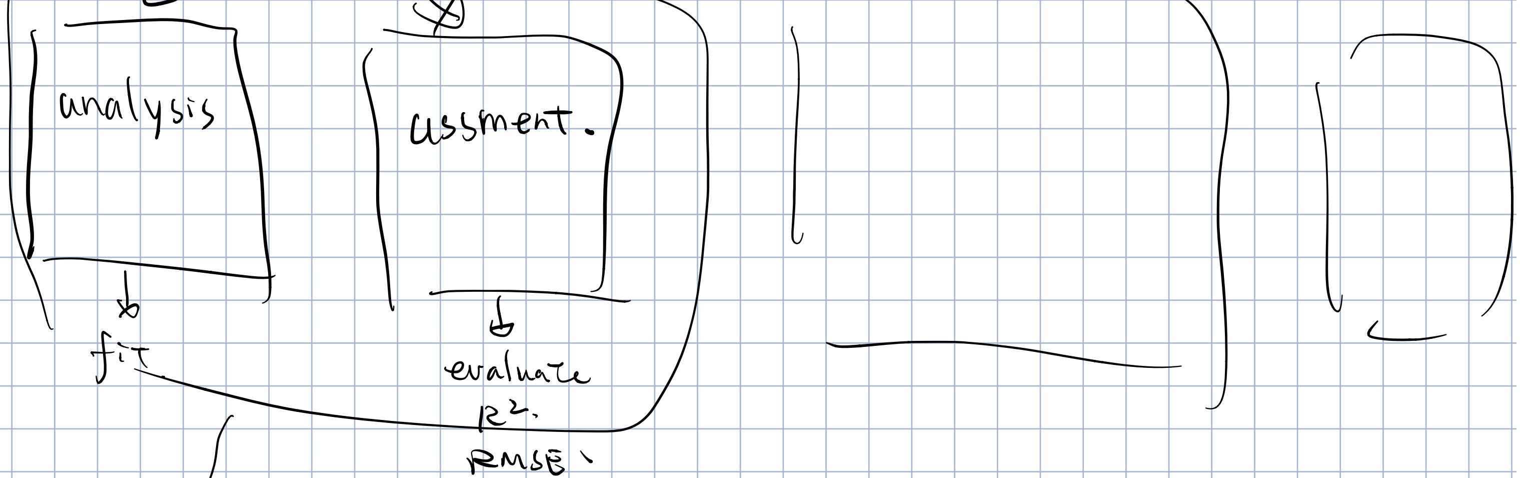
...... } fit. model.

... } prediction.

? → test.

CV.

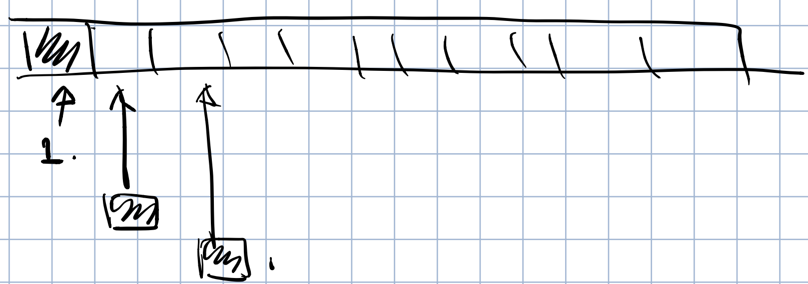
training + test.



$\boxed{\text{CV.}}$

RMSE					
R^2.					

10-fold.



CV. { cv_fold.( ).
fit_samples( )

Inference. { CI.
HT. : $H_0: \beta_1 = 0$ $\boxed{\text{hold other constant}}$

Diagnosis : { Conditions : (4). residual plot. - SLR.
multicollinearity. → non-identifiable.
↓
se (·) ↑.
↓
inference imprecise.

detect: { corr. > 0.9.
abnormal. coeff.
VIF. : > 10.

Select one to remove.

model comparison