Stat 4 MR Hi !! 2023. W02

Tests for independence + Permutation tests

Ex.

X \ Y	Online	offline
A	25 ?	15
B	10	10
C	15	10

2D cont. table

← obs-d freq.

Q. Are X and Y dependent?

H_0: X and Y are indep.
H_A: X and Y are dep.

Any ideas?

Classic Pearson's Independence χ^2 test:

$$P.S. = \sum_{i,j} \frac{(N_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

$$\sum_i \frac{(N_i - E(N_i))^2}{E(N_i)}$$

$\hat{\mu}_{ij}$ - is the expected value of N_{ij} estimated under H_0 (independency)

idea: estimate $\hat{p}_A, \hat{p}_B, \hat{p}_C$, $\hat{p}_{online}, \hat{p}_{offline}$ under H_0

$$\hat{p}_A = \frac{N_{A,on} + N_{A,off}}{N}$$

$$\hat{p}_{on} = \frac{N_{A,on} + N_{B,on} + N_{C,on}}{N}$$

$$\hat{\mu}_{A,on} = N \cdot \hat{p}_A \cdot \hat{p}_{on}$$

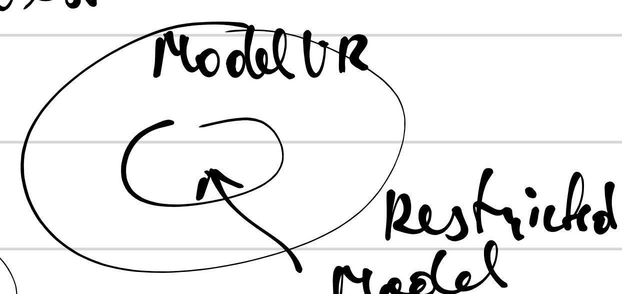
$$\hat{p}_{i\bullet} = \frac{\sum_j N_{ij}}{N} \qquad \hat{p}_{\bullet j} = \frac{\sum_i N_{ij}}{N}$$

$$\hat{\mu}_{ij} = N \times \hat{p}_{i\bullet} \times \hat{p}_{\bullet j} = \frac{\left(\sum_i N_{ij}\right) \cdot \left(\sum_j N_{ij}\right)}{N}$$

$$PS. = \sum_{i,j} \frac{(N_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \xrightarrow[\text{dist}]{N \to \infty,\ H_0} \chi^2_{df}$$

$$df = (c_x - 1) \cdot (c_y - 1)$$ where c_x - number of possible values of X, c_y — // — of Y

nested models:



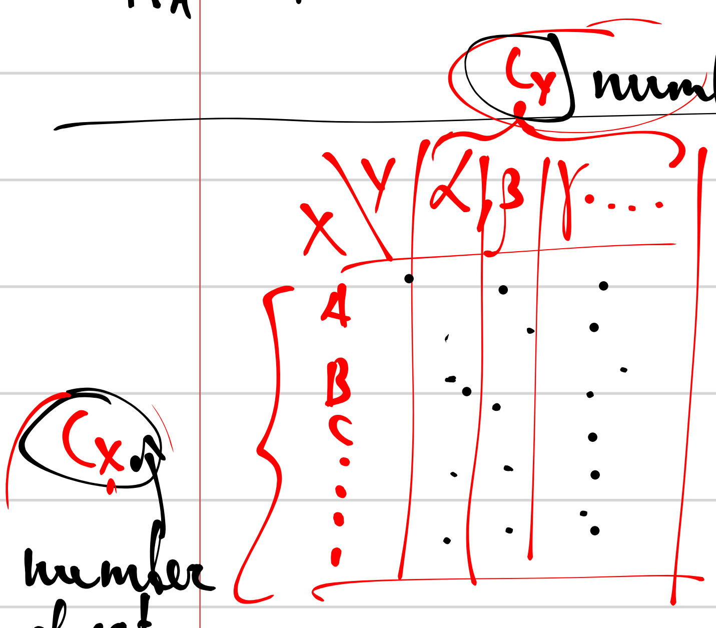
Likelihood Ratio test.

$$LR = 2 \cdot (\ln L_{UR} - \ln L_R) \xrightarrow[\text{dist}]{H_0,\ N \to \infty} \chi^2_{p_{UR} - p_R}$$

p_R, p_{UR} - number of par-s.

H_0: Restricted model is TRUE

H_A: R-model is FALSE, but UR-model is TRUE



R-model: X and Y are indep.

UR-model : maybe X and Y are depend., any probab-s are OK.

$$p_{UR} \stackrel{?}{=} c_x \cdot c_y - 1$$

one constraint: $\sum_{ij} p_{ij} = 1$

free parameters.

$c_x \cdot c_y$ — number of cells in the table

$$p_{UR} - p_R = c_x \cdot c_y - 1 - c_x - c_y + 2 = (c_x - 1) \cdot (c_y - 1)$$

$$p_R ? = \underbrace{c_x - 1}_{\text{prob-s for } X} + \underbrace{c_y - 1}_{\text{probs for } Y} = c_x + c_y - 2$$

UR - model:

X \ Y	Online	offline
A	25 ?	15
B	10	10
C	15	10

$N = 85$

X \ Y	On	Off
A	$\hat{p}_{11}$	$\hat{p}_{12}$
B	$\hat{p}_{21}$	$\hat{p}_{22}$
C	$\hat{p}_{31}$	$\hat{p}_{32}$

it is not a free parameter!

$\hat{p}_{11}^{UR} = \frac{25}{85}$ $\qquad \hat{p}_{12}^{UR} = \frac{15}{85}$

$\hat{p}_{21}^{UR} = \frac{10}{85}$ $\qquad \vdots$

$p_{UR} = 2 \cdot 3 - 1 = 5$

$$LR = 2 \cdot \left(\ln L_{UR} - \ln L_R \right) = \text{[use past lecture]}$$

$$= 2 \cdot \sum_{i,j} N_{ij} \cdot \left(\ln \hat{p}_{ij}^{UR} - \ln \hat{p}_{ij}^{R} \right) \xrightarrow[\text{dist}]{n\to\infty,\ H_0} \chi^2_{df}$$

$df = p_{UR} - p_R = (c_X - 1) \cdot (c_Y - 1)$

R - model: [H_0 of independency]

X \ Y	Online	offline
A	25 ?	15
B	10	10
C	15	10

$$\left.\begin{array}{l} \hat{p}_{1\cdot} = \frac{25+15}{85} \\ \hat{p}_{2\cdot} = \frac{20}{85} \end{array}\right\} \text{2 free params}$$

$\hat{p}_{3\cdot} = \frac{25}{85}$

it is not a free param!

$$\left.\hat{p}_{\cdot 1} = \frac{25+10+15}{85}\right\} \text{1 free param}$$

$\hat{p}_{\cdot 2} = \frac{15+10+10}{85}$

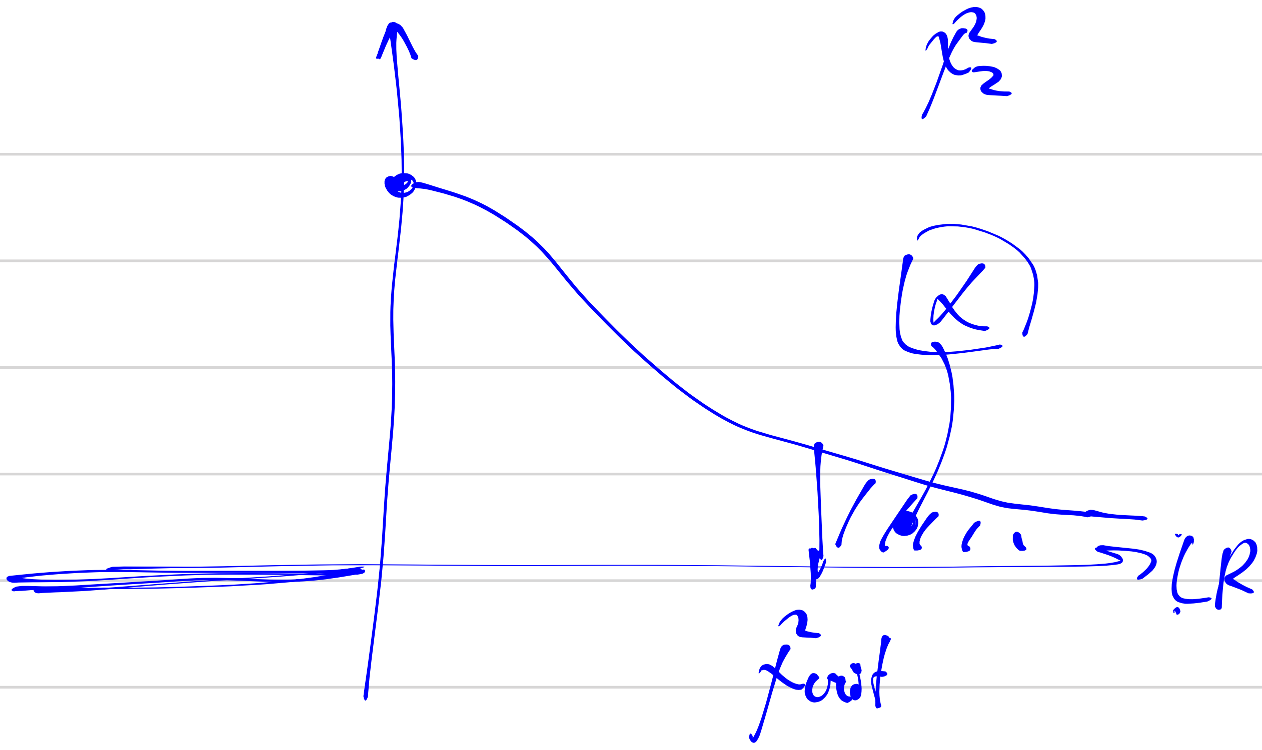
$p_R = 1 + 2 = 3$ param.

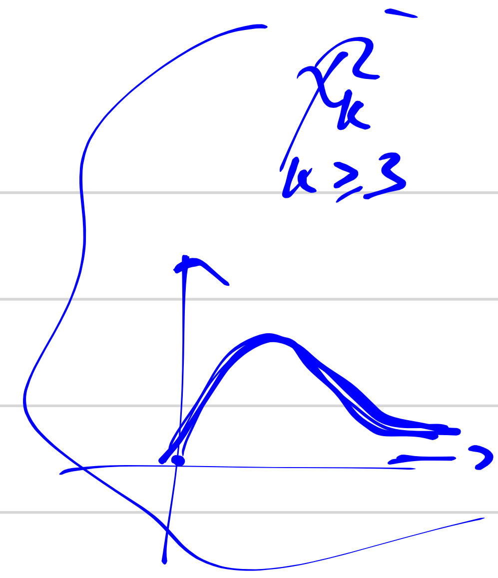
H_0: X and Y are indep.

$$\hat{p}_{ij}^{R} = \hat{p}_{i\cdot} \times \hat{p}_{\cdot j}$$

$$\hat{p}_{11}^{R} = \frac{25+15}{85} \cdot \frac{25+10+15}{85}$$

in my ex: $\chi^2_{(3-1)\cdot(2-1)} = \chi^2_2$





if $LR > \chi^2_{crit}$ then reject H_0.

Permutation test. / Bootstrap.

idea: write a long table, row = = one observ.

X	Y
A	On
B	On
C	Off
A	Off
⋮	⋮
C	On

Calculate any [!] reasonable measure of similarity (correla-n)

example: [of a reas. measure for this case]

Y_{on}	X_A	X_B	X_C
1	1	0	0
1	0	1	0
0	.		
0			
⋮			
1			

Step 1 regress: $\hat{Y}^{on}_i = \hat{\beta}_1 \cdot X^A_i + \hat{\beta}_2 \cdot X^B_i + \hat{\beta}_3 X^C_i$

Step 2: calculate R^2_{orig}

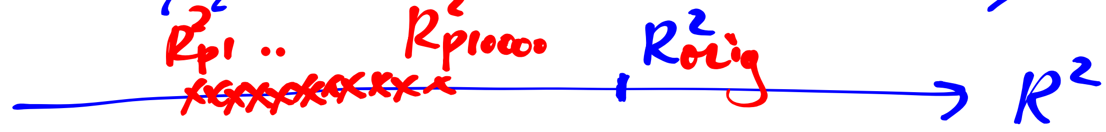
---

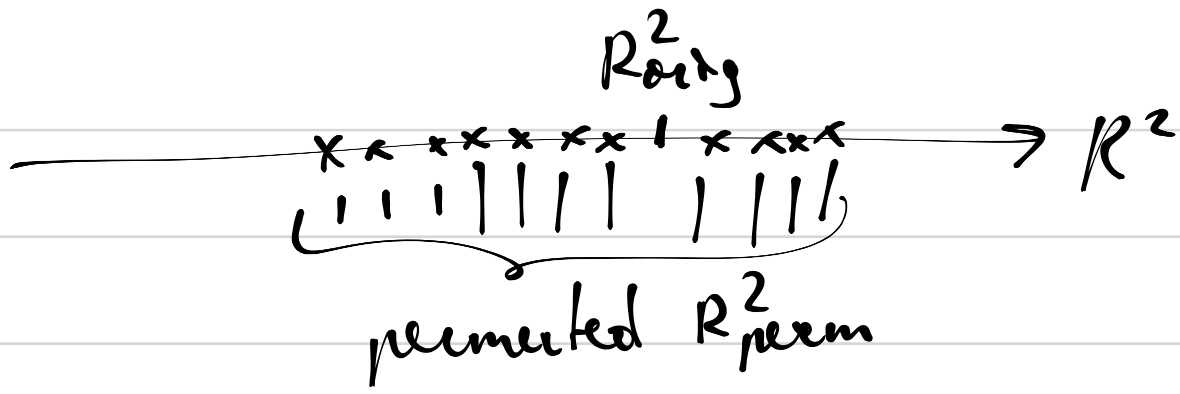
① Randomly permute values of Y => obtain $R^2_{perm\,1}$

② — // — => obtain $R^2_{perm\,2}$

⋮

10000 . — // — => obtain $R^2_{perm\,10000}$

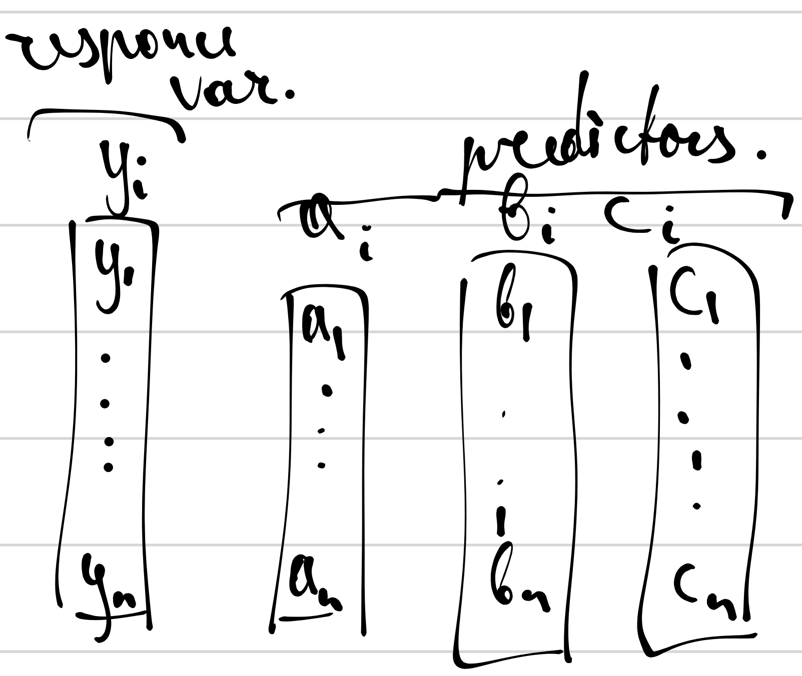




P-value = fraction of R^2_{perm} higher than R^2_{orig}.

If p-value $< \alpha = 0.05$ then reject H_0
If p-value $\geq \alpha = 0.05$ then do not reject H_0.

Permutation tests in machine learning.



→ Classification problem
$y_i \in \{A, B, C\}$
→ Gradient boosting

Q. Which predictors are really important?

Step 1. Split into train and test

Train → Step 2. estimate the param.s of the algorithm on the training set. only one model est-n!

Test

A_{orig} · Step 3. calculate accuracy on test sample (or any other reasonable quality measure)

Step 4 Run permut-n test for variable a.
→ permute values of var-le a. obtain $A_{perm\,1}$

→ — // — obtain $A_{perm\,2}$

Step 5 Calculate : $A_{orig} - A_{perm}$