G-PAC 5: Anàlisi de variància, regressió lineal simple i múltiple. Màrqueting

PID_00141419

Universitat Oberta de Catalunya
www.uoc.edu

Índex

1. Introducció

En aquesta darrera prova d'avaluació continuada de l'assignatura treballarem diverses tècniques estadístiques per a l'anàlisi de les relacions entre diferents variables. Després de les quatre primeres PAC, centrades més en l'anàlisi univariable, ara ens fixem principalment en l'anàlisi bivariable. Addicionalment, i de manera introductòria, també tractarem alguna tècnica d'anàlisi multivariable, com és la regressió múltiple.

Quan hem de treballar amb diverses variables, un element que hem de tenir en compte és si hi ha o no relació de causalitat entre aquestes variables. Aquest també serà un dels punts a desenvolupar a través de les activitats d'aquesta PAC.

L'aplicació de les tècniques estadístiques d'aquesta activitat es durà a terme en l'àmbit del màrqueting i, més concretament, en l'àmbit del comerç electrònic.

2. Objectius i competències

Els **objectius** que es volen assolir amb aquesta quinta PAC són els següents:

1) Saber comparar l'homogeneïtat de mitjanes de dues o més poblacions a través de l'anàlisi de la variancia (ANOVA).

2) Conèixer la utilitat de les taules de contingència a l'hora de valorar relacions de dependència entre variables.

3) Saber valorar si hi pot haver o no relació entre dues o més variables a partir del coeficient de correlació.

4) Saber estimar un model de regressió, simple o múltiple per, *a posteriori*, interpretar estadísticament els resultats obtinguts.

Addicionalment, en aquesta PAC 5 es treballaran les **competències** següents:

- Capacitat per a generar coneixement econòmic rellevant a partir de dades, aplicant els instruments tècnics pertinents.
- Capacitat per a aplicar els coneixements teòrics i les eines d'investigació dels mercats en la definició de solucions de negoci.
- Capacitat per a valorar críticament situacions empresarials concretes i establir possibles evolucions d'empreses i mercats.
- Capacitat per a utilitzar i aplicar les tecnologies de la informació i la comunicació en els àmbits acadèmic i professional.

3. Guia de continguts

3.1. Anàlisi de la variància

L'anàlisi de la variància és una tècnica d'inferència estadística per a contrastar si un nombre determinat de poblacions (K) tenen la mateixa mitjana. Per tal d'aplicar correctament aquest mètode, és necessari que les poblacions siguin normals i que tinguin la mateixa variància.

$$X_j \approx N(\mu_j;\sigma),\ j = 1, 2, \ldots, K$$

La hipòtesi nul·la del contrast és:

$$H_0 = \mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_K$$

La hipòtesi alternativa H_1 representa el cas que almenys una mitjana és diferent de la resta.

Per a fer el contrast, s'agafa una mostra (no necessàriament de la mateixa mida) de cadascuna de les K poblacions. La taula següent representa la informació de la qual es disposarà i la que es calcularà posteriorment (mitjanes i variàncies):

	Mostres						
Obs.	1	2	....	j	....	K	Total
1	X_{11}	X_{21}	....	X_{j1}	....	X_{k1}	
2	X_{12}	X_{22}	....	X_{j2}	....	X_{k2}	
....	....	....	....	....	....	....	
i	X_{1i}	X_{2i}	....	X_{ji}	....	X_{ki}	
....	....	....	....	....	....	....	
n_j	X_{1n_1}	X_{2n_2}	....	X_{jn_j}	....	X_{kn_k}	
Mides	n_1	n_2	....	n_j	....	n_k	N
Mitjanes	$\overline{X}_1$	$\overline{X}_2$	....	$\overline{X}_j$	....	$\overline{X}_k$	$\overline{X}$
Variàncies	S_1^2	S_2^2	....	S_j^2	....	S_k^2	S^2

En què n_j és la grandària de la mostra extreta de la població j, i X_j és l'observació i-èsima de la població j-èsima.

Per a fer el contrast ANOVA s'han de calcular les quantitats següents:

- **Suma de quadrats total (SQT)**. Reflecteix la dispersió de totes les dades respecte a la mitjana global:

$$SQT = \sum_{j=1}^{k} \sum_{i=1}^{nj} (X_{ji} - \overline{X})^2$$

- **Suma de quadrats entre grups (SQE)**. La variació entre grups reflecteix com són de diferents (o no) les *K* mitjanes mostrals:

$$SQE = \sum_{j=1}^{k} (\overline{X}_j - \overline{X})^2 n_j$$

- **Suma de quadrats dins dels grups (SQD)**. La variació dins dels grups reflecteix com són de disperses les mostres:

$$SQD = \sum_{j=1}^{k} \sum_{i=1}^{nj} (X_{ji} - \overline{X}_j)^2$$

La relació que hi ha entre aquestes tres expressions és la següent:

$$SQT = SQE + SQD$$

L'estadístic per a fer el contrast ANOVA és el següent:

$$F^* = \frac{\frac{SQE}{k-1}}{\frac{SQD}{n-k}}$$

Si H_0 és certa, aquest estadístic es comportarà segons una distribució *F* amb (*K* – 1, *N* – *K*) graus de llibertat. Si l'estadístic de prova pren un valor relativament alt (superior a un valor crític amb un nivell de significació α fixat per al test):

$$F^* > F_{K-1;N-K;\alpha}$$

es rebutjarà la hipòtesi nul·la i conclourem que les poblacions no tenen la mateixa mitjana; que les poblacions són diferents. El contrast ANOVA és un contrast unilateral, a una cua superior.

3.2. Taules de contingència

Les taules de contingència s'empren a l'hora de fer una anàlisi preliminar de la distribució conjunta entre dues variables qualitatives. Una de les seves principals utilitats és saber si hi ha o no dependència estadística entre les dues variables.

La hipòtesi nul·la en el contrast estadístic de dependència és la següent:

H_0: No hi ha relació entre les dues variables.

I l'estadístic de contrast, que segueix una distribució khi-quadrat, compara la freqüència observada O_{ij}, dels casos que pertanyen a la categoria *i*, si ens fixem en la primera variable, i a la categoria *j*, si ens fixem en la segona variable, amb la freqüència teòrica T_{ij}, calculada suposant que les dues variables són independents:

$$\chi^{2*} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}$$

Si aquest valor és molt "petit", voldrà dir que no hi ha diferència entre les freqüències observades i les teòriques i, per tant, ens indicarà que hi ha independència entre les dues variables. Si aquest valor és "gran" ens indicarà el contrari, i haurem de rebutjar la hipòtesi nul·la.

- ***p-valor***. Una forma alternativa per a poder decidir si rebutgem o no la hipòtesi nul·la és fixar-nos en el *p*-valor de l'estadístic de prova:

$$Valor - P\,(\chi^{2*}) = Pr(\chi^2_{(I-1).(J-1)} > \chi^{2*})$$

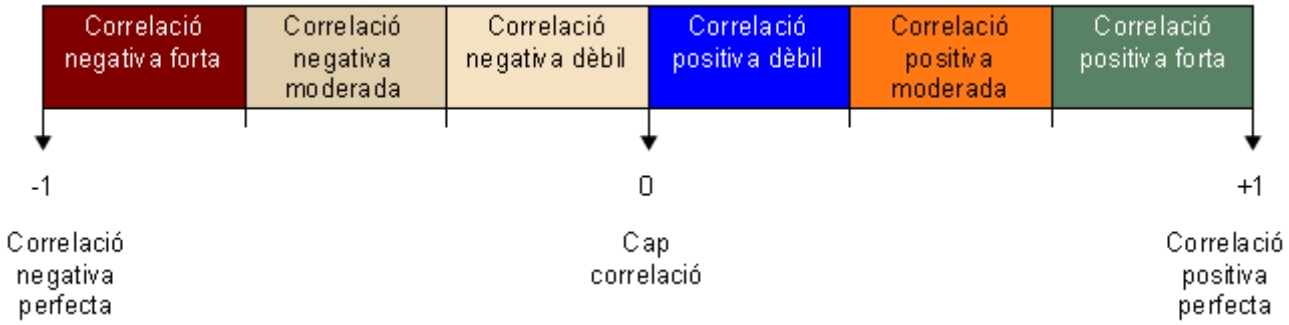
Si aquest valor és superior al nivell de significació α que fixem, no rebutjarem la hipòtesi nul·la. Si és més petit, la rebutjarem, és a dir, afirmarem estadísticament que hi ha relació entre les dues variables.

3.3. Correlació lineal

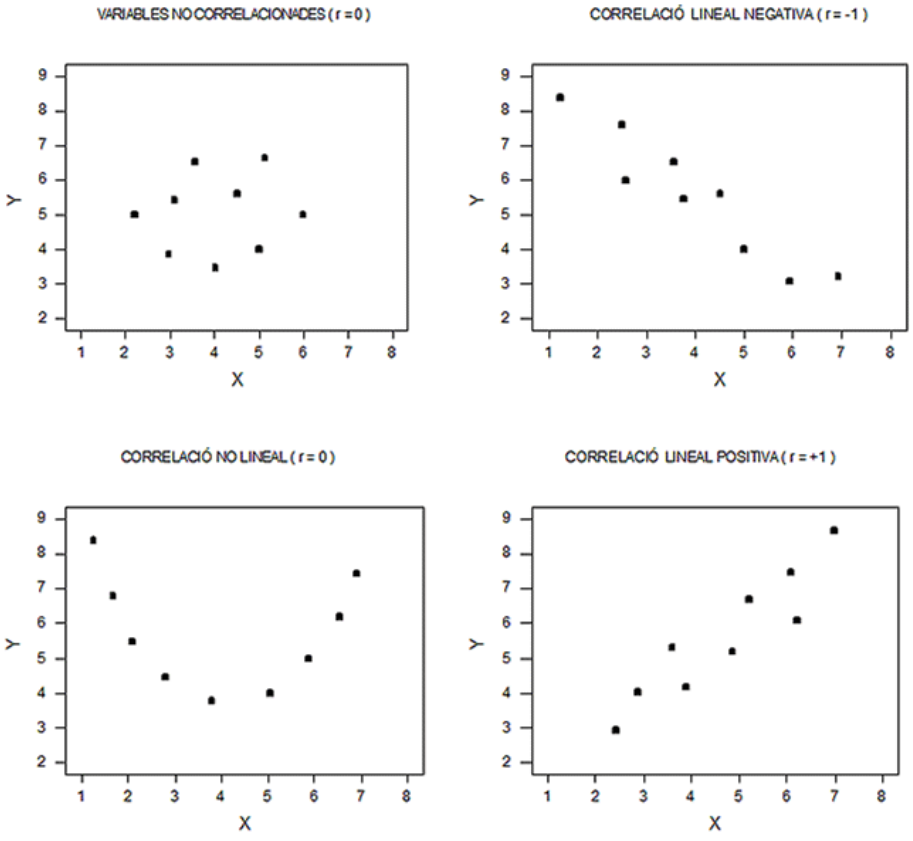
Usarem aquesta tècnica quan ens interessi estudiar si hi ha o no algun tipus de relació entre dues variables aleatòries, sense que *a priori* hi hagi d'haver cap relació de causalitat entre elles. L'existència de correlació entre variables no implica causalitat. En particular, quan vulguem quantificar la intensitat de la relació lineal entre aquestes dues variables. El paràmetre que ens mesura la correlació lineal és el **coeficient de Pearson *r***:

$$r = \frac{Cov(X, Y)}{s_X \, s_Y}$$

El seu valor oscil·la entre –1 i +1. Si aquest està pròxim a +1 direm que la correlació tendeix a ser lineal directa (majors valors d'una variable s'associen a majors valors de l'altra). Si s'aproxima a –1 ens indicarà que la relació entre les variables tendeix a ser lineal inversa (majors valors d'una variable estan relacionats amb menors valors de l'altra).



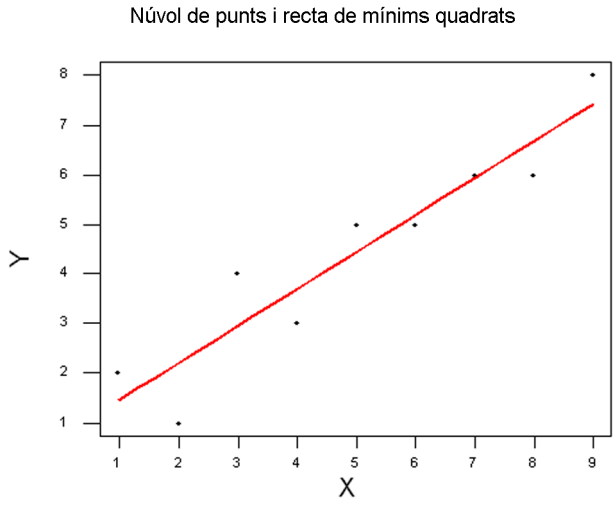
Els següents gràfics mostren exemples de núvols de punts corresponents a diferents tipologies de correlacions lineals:



3.4. Regressió lineal simple

Quan el coeficient de correlació lineal està proper a +1 o a –1 (correlació forta), té sentit considerar si entre les variables hi ha una relació de causalitat que pot ser modelitzada a través de l'equació d'una recta. Un cop determinada quina és la variable explicativa (independent) i quina la variable explicada (dependent), l'equació de la recta que representa la relació entre les variables es calcula pel procés de mínims quadrats ordinaris. Un dels principals usos de l'esmentada recta serà el de predir o estimar els valors de y que obtindríem per a diferents valors de x.

Aquests conceptes quedaran representats en el que anomenem *diagrama de dispersió*:



L'equació de la recta de regressió (en forma punt-pendent) és la següent:

$$y - \overline{y} = \frac{Cov(X, Y)}{s_X^2}(X - \overline{X})$$

Una altra expressió per a la recta de regressió és:

$$y = \hat{\alpha} + \hat{\beta} \cdot X$$

en què $\hat{\alpha}$ i $\hat{\beta}$ són les estimacions de l'ordenada a l'origen i del pendent, respectivament.

Per a poder fer una inferència estadística a partir dels resultats obtinguts, cal que es donin els supòsits següents:

- En la població, la relació entre les variables x i y ha de ser aproximadament lineal, és a dir: $y = \alpha_1 + \alpha_2 x + \epsilon$, essent ϵ la variable aleatòria que representa

els **residus**, diferències entre els punts reals de les dades i les estimacions del model.

- Els residus s'han de distribuir segons una distribució normal de mitjana 0 i variància σ^2 constant, això és, $\epsilon \approx N(0,\sigma^2)$.

- Els residus són independents els uns dels altres.

Afortunadament, el model de regressió lineal és bastant robust, cosa que significa que no és necessari que les condicions anteriors es compleixin amb exactitud per tal de poder treure conclusions de les dades.

El **coeficient de determinació** o la **bondat de l'ajust**, R^2, és el coeficient que ens indica el percentatge de l'ajust que s'ha aconseguit amb el model lineal, és a dir, el percentatge de la variació de y que s'explica a través del model lineal pel comportament de x. Concretament, el coeficient de determinació és una mesura de la proximitat o de l'ajust de la recta de regressió al núvol de punts. A major percentatge, el nostre model prediu millor el comportament de la variable dependent.

El coeficient de determinació és igual al quadrat del coeficient de correlació r:

$$R^2 = r^2$$

Quan la bondat d'ajust sigui prou elevada, estarem interessats a saber si realment variacions en la variable independent afecten la variable dependent. Per a saber això, haurem de saber si, estadísticament, el coeficient β que acompanya la variable explicativa x (i que és el pendent de la recta de regressió) és significativament diferent de 0. El contrast d'hipòtesis que cal plantejar és el següent:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

L'estadístic de contrast que s'utilitza per a fer el test és el següent:

$$t^* = \frac{\hat{\beta}}{s_{\hat{\beta}}}$$

que segueix una distribució t d'Student amb $n - 2$ graus de llibertat, i en què $s_{\hat{\beta}}$ és l'error estàndard del pendent:

$$s_{\hat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} \cdot x_i)^2}{(n-2)\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

- ***p-valor***. Per a decidir si rebutgem o no la hipòtesi nul·la ens podem fixar en el *p*-valor de l'estadístic de prova:

$$Valor - P(t^*) = Pr(t_{n-2} > |t^*|) + Pr(t_{n-2} < -|t^*|)$$

Si aquest valor és superior al nivell de significació α que fixem, no rebutjarem la hipòtesi nul·la. Si és més petit, la rebutjarem, és a dir, afirmarem estadísticament que el coeficient que acompanya la variable explicativa és diferent de 0.

Finalment, també podem obtenir l'interval de confiança per a β a un nivell de confiança 1 – α utilitzant l'expressió:

$$\hat{\beta} \pm t(n-2, \alpha/2).s_{\hat{\beta}}$$

3.5. Regressió lineal múltiple

La regressió lineal múltiple parteix de la mateixa idea que la simple, però les coses es compliquen pel fet que en lloc de tenir només una variable explicativa o independent, ara s'incorporen en el model tants regressors o variables explicatives com vulguem.

$$Y = \beta_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + ... + \beta_k \cdot X_k$$

Això fa que els càlculs necessaris per a estimar el model es facin habitualment amb un paquet estadístic, ja que cal fer servir expressions matricials:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

La bondat de l'ajust es mesura pel coeficient de determinació R^2, que és, com en el model de regressió simple:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

Aquest coeficient està delimitat entre 0 i 1. Com més proper a 1, millor és l'ajust i més forta és la relació que hi ha entre les variables que el model vol captar. El paquet estadístic que usem ens presentarà la taula ANOVA, en la qual aquestes sumes de quadrats apareixeran sota la capçalera SS (*sums of squares*, és a dir, sumes de quadrats).

Com en el cas de la regressió simple, per veure si una variable independent pot explicar la variable dependent, hem d'estudiar si el coeficient β_j que l'acompanya és significativament diferent de 0 o no:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Aquest contrast requereix el càlcul de l'estadístic de prova següent:

$$t_j^* = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}$$

Aquest estadístic segueix una distribució *t* d'Student, i si el seu valor absolut (contrast a dues cues) és superior al valor crític $t_{n-k,\alpha/2}$ rebutjarem la hipòtesi nul·la, i direm que el regressor X_j és estadísticament significatiu; que X_j explica *Y*.

- ***p-valor***. Una manera alternativa per a decidir si rebutgem o no la hipòtesi nul·la és fixar-nos en el *p*-valor de l'estadístic de prova:

$$Valor - P(t_j^*) = Pr(t_{n-k} > |t_j^*|) + Pr(t_{n-k} < -|t_j^*|)$$

Si aquest valor és superior al nivell de significació α que fixem, no rebutjarem la hipòtesi nul·la. Si és més petit, la rebutjarem, és a dir, afirmarem estadísticament que el coeficient que acompanya la variable explicativa és diferent de 0.

Per tal de validar el model de manera conjunta farem un contrast d'hipòtesi en què la hipòtesi nul·la és:

H_0: El model no és significatiu en conjunt

i l'estadístic de contrast serà el següent:

$$F^* = \frac{\frac{SQR}{k-1}}{\frac{SQE}{N-k}}$$

Aquest estadístic segueix una distribució *F*. Es tracta d'un contrast a una cua superior (a diferència del contrast *t* de significació individual que és bilateral, o a dues cues). Si l'estadístic supera el valor crític $F_{k-1;n-k,\alpha}$, direm que el model és globalment significatiu, és a dir, que rebutjarem la hipòtesi nul·la.

- ***p-valor***. Una forma alternativa per poder decidir si rebutgem o no la hipòtesi nul·la és fixar-nos en el *p*-valor de l'estadístic de prova:

$$Valor - P(F^*) = Pr(F_{k-1;n-k} > F^*)$$

Si aquest valor és superior al nivell de significació α que fixem, no rebutjarem la hipòtesi nul·la. Si és més petit, la rebutjarem, és a dir, afirmarem estadísticament que el model és significatiu de manera conjunta.

4. Fonts d'informació

Per a resoldre la PAC 5 haureu de fer servir dues bases de dades, una de caràcter temporal i una altra de caràcter transversal, i llegir la lectura proposada, que trobareu a la unitat didàctica corresponent. Així mateix, podeu consultar tot un seguit de fonts d'informació que us proposem a continuació, i que han servit per a confeccionar les bases de dades amb què treballareu. Trobareu arxius amb dades actualitzades a la unitat didàctica corresponent.

1) Base de dades de caràcter temporal

N. sèrie	Nom de la variable (unitats de mesura)	Font
Sèrie 01	Volum de negoci del comerç electrònic (milions d'euros)	www.aecem.org
Sèrie 02	Població (% total)	www.ine.es
Sèrie 03	Població urbana (% total)	www.ine.es
Sèrie 04	Llars amb ordinador (% total)	www.ine.es
Sèrie 05	Servidors web segurs (per milió habitants)	www.netcraft.com
Sèrie 06	Preus d'ordinadors a Europa (€)	www.eito.com
Sèrie 07	Línies digitals (milers)	www.eito.com
Sèrie 08	Dominis d'Internet (total mundial)	www.isc.org
Sèrie 09	Dominis.es (total)	www.isc.org
Sèrie 10	Usuaris d'Internet (milers)	www.aimc.es
Sèrie 11	Usuaris d'Internet (% població > 14)	www.aimc.es
Sèrie 12	Població > 14 (milers)	Calculat
Sèrie 13	Despesa d'alimentació (% total)	www.ine.es
Sèrie 14	Solter/divorciat/separat/vidu (% població > 15 anys)	www.ine.es
Sèrie 15	Formació universitària (% població > 15 anys)	www.ine.es

2) Base de dades de caràcter transversal

Les fonts d'informació usades per a elaborar la base de dades corresponen a tres projectes:

- e-Business W@tch Project.
- Projecte Regional-IST.
- Projecte Internet Catalunya–PIC.