Probabilitat i Estadística 2, GCED

Prof. Marta Pérez-Casany i Jordi Valero, 2019-2020

Exercici1 1. Un enginyer químic vol estudiar la duresa de quatre barreges de pintura. A tal efecte, es van agafar sis mostres de cada barreja de pintura i amb cadascuna d'elles es va pintar un troç de metall (tots els troços tenien la mateixa dimensió). Passats uns dies, i un cop la pintura ja estava seca, se'n va mesurar la duresa. Les variables que es van considerar van ser: *Pintura* (hi ha 4 barreges), *Duresa*, *Operador* (persona que feia la mesura) i *Temperatura* (mitja del dia en que es va pintar). Basant-nos amb això, contesteu les preguntes següents:

1) Quina és la variable resposta i quines són les explicatives? Quina distribució té sentit assumir per la resposta? De quin tipus són les explicatives?

   La variable resposta és la Duresa de la pintura. Atès que es mesura clavant-hi una mena de punxó i veient quan s'enfosa, té sentit assumir que segueix una distribució Normal. Les variables explicatives són Tipus de pintura que és categòrica amb quantre nivells, Operador, que també és categòrica amb tants nivells com operaris hagin participat en la recollida de dades i finalment la Temperatura que és numèrica contínua ja que prové de promitjar diverses temperatures del dia en que es va pintar el metall.

2) Formula tres preguntes que tingui sentit contestar en aquest entorn. Formula-les en termes de les variables que intervenen. És a dir, han de ser preguntes que entengui el públic en general i en particular el propietari de l'emrpesa de pintures que us ha encarregat les anàlisis.

   En aquest entorn té sentit preguntar-se entre d'altres coses el següent:

   - Té alguna influència el tipus de pintura en la duresa final?
   - Té alguna influència en la dures final la temperatura mitja del dia en que es realitza la pintada?
   - Es possible que l'efecte del tipus de pintura en la duresa final canvii dependent de la temperatura del dia en que es realitza la pintada?

3) Defineix què vol dir *interacció* entre dues variables. Té sentit considerar algún tipus d'interacció entre les variables d'aquest experiment? Si la resposta és que sí, explica quina i perquè. Si la resposta és que no, justifica perquè no.

   La interacció és un fenòmen que té lloc entre dues variables explicatives categòriques o entre una variables explicativa categòrica i una numèrica. Entenem que dues variables interaccionen quan l'efecte que produeix una d'elles (pot ser categòrica o numèrica) en la variable resposta depèn del nivell de la variable categòrica amb la que apareix combinada quan es realitza l'experiment. Important remarcar que NO és possible definir interacció sense esmentar la variable resposta.

   En l'experiment d'aquest enunciat, té sentit preguntar-se si l'efecte del tipus de pintura en la duresa depèn de la temperatura que hi havia el dia

que es va pintar. Per tant, en aquest cas té sentit parlar d'interacció entre Pintura i Temperatura. Aquesta pregunta correspon a la pregunta tres de l'apartat anterior.

4) Consideraries un ML o un MLG per ajustar les dades de l'experiment anterior?

   Atès que té sentit assumir que la duresa segueix una distribució Normal, començariem per un ML, concretament seria un model ANCOVA oerquè tenim una explicativa numèrica i dues categòriques. El motiu és que són més senzills d'ajustar i d'interpretar. En el cas que amb l'anàlisi dels residus veiessim que les hipòtesis de Normalitat, independència i igualtat de variàncies que requereix el ML no es complissin, recorreriem a un MLG.

Exercici2 2. Es vol saber si un aparell de mesura està ben calibrat. Concretament, si mesura correctament una distància. A tal efecte, s'han mesurat un total de 25 distàncies entre dos punts amb l'aparell, i aquestes es denoten per y_i. Atès que es coneix el valor exacta de les distàncies, el *dataset* conté punts de la forma (x_i, y_i), $i = 1 \cdots 25$, on x_i és la distància exacta. En aquesta situació, contesteu les preguntes següents:

1) Quin és el model de regressió que heu d'ajustar per tal de saber si l'aparell funciona correctament? Raoneu perquè és aquest i escriviu-lo en foma matricial. Quin nom rep aquest model?

   En aquest cas volem comparar la nostra mesura de la distància amb la distància real, per tant, el model que té sentit ajustar és un model de REGRESSIÓ LINEAL SIMPLE.

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

   assumint que $e_i \sim N(0, \sigma^2)$ i que dos errors diferents són independents, és a dir: e_i indepentdent de e_j si $i \neq e_j$. Matricialment s'escriu com:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{25} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{25} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{25} \end{pmatrix}$$

2) Quins són els dos test d'hipòtesis que té sentit portar a terme en aquest entorn?, Expliciteu-los tot donant la hipòtesi nul.la i l'alternativa de cadascún d'ells. Sabent que la sortida obtinguda a l'aplicar un ML és la que apareix a la taula de sota, porteu a terme els test d'hipòtesis esmentats. Què concluiu?

	$\hat{\beta}$	$s.d.(\hat{\beta})$
intercept	14.06	0.07
slope	1.76	0.4

   Els dos test d'hipòtesis que té sentit realitzar són: Test 1: $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$, Test 2: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. En el cas de NO rebutjar cap de les dues hipòtesis nul.les conclouriem que l'aparell de

mesura mesura correctament les distàncies ja que podriem assumir que $y_i = x_i + e_i$, és a dir que la recta que ajusta el núvvol de punts és la recta identitat. Tot seguit realitzem els test:

Test 1: L'estadístic de prova és $\frac{\hat{\beta}_0}{s.d.(\hat{\beta}_0)} = \frac{14.06}{0.07} = 200$, si H_0 és certa, aquesta observació ha de provenir d'una distribució t-d'Student amb $n - p = 25 - 2 = 23$ graus de llibertat. Prenent $\alpha = 0.05$, es té que $t_{0.975,23} = 2.06$. Atès que 200 no cau a la regió de no rebuig que és $(-2.6, 2.6)$ concluim que β_0 és estadísticament diferent de zero.

Test 2: L'estadístic de prova és $\frac{\hat{\beta}_1}{s.d.(\hat{\beta}_1)} = \frac{1.76-1}{0.4} = 1.9$, si H_0 és certa, aquesta observació ha de provenir d'una distribució t-d'Student amb $n - p = 25 - 2 = 23$ graus de llibertat. Prenent $\alpha = 0.05$, es té que $t_{0.975,23} = 2.06$. Atès que 1.9 cau a la regió de no rebuig que és $(-2.6, 2.6)$ concluim que β_0 és estadísticament diferent de zero.

3) Com interpreteu les estimacions dels paràmetres que apareixen en la sortida anterior?

Atès que $\hat{\beta}_0 = 14.06$ i aquest és estadísticament diferent de zero, entenem que l'aparell de mesura fa un error sispemàtic positiu de 14 unitats. Es a dir l'aparell sobreestima totes les mesures en 14 unitats.

Atès que $\hat{\beta}_1 = 1.9$, entenem que per a cada increment d'una unitat de la distància real, la distància mesurada amb l'aparell incrementarà en 1.9 unitats. Ara bé, hem vist que aquest valor no és estadísticament diferent de 1, per tant concluim que l'error que fa l'aparell no depèn de la mesura en questió, és un error sistemàtic positiu independent de la distància real entre els dos punts.

4) Definiu test *Omnibus*. A partir de la taula ANOVA que figura tot seguit, porteu a terme l'Omnibus test per al model anterior. Què concluiu? (Prèviament completeu la taula posant un valor allà on hi ha una x)

El test Omnibus és el que es fa servir en els ML per a comparar el nostre model amb el model nul. Si es rebutja la hipòtesi nul.la, és conclou que el nostre model explica una part significativa de la variabilitat de les dades. Ve definit a través les les següents hipòtesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \;\; vs \;\; H_1 : \exists \; i \text{ tal que} \beta_i \neq 0$$

Per a fer el test Omnibus cal completar la taula. El resultat de la completació és el següent:

	SQ	G. ll.	MSQ	F_0
Regressió	150.6	1	150.6	$\frac{150.6}{4.72} = 31.9$
Error	108.7	23	$\frac{108.7}{23} = 4.72$	
Total	259.3	24		

Si la hipòtesi nul.la és certa, F_0 ha de ser una observació que prové d'una distribució de Fisher amb $n_1 = 1$ i $n_2 = 23$ graus de llibertat. Atès que l'Omnibus test és unilateral, es rebutjarà H_0 quan $F_0 > F_{1-\alpha,n_1,n_2}$, en el nostre cas, amb $\alpha = 0.05$ es té que $F_{0.95,1,23} = 4.27$ Atès que $31.0 > 4.23$, rebutgem H_0 i concluim que el model és millor que el nul.

Exercici 3.

1) Defineix MLG. Què ha de passar perquè un MLG sigui un ML?

   Un Model Lineal Generalitzat és una fórmula matemàtica que ens permet explicar una variable anomenada variable resposta (Y) en fució d'unes altres anomenades variables explicatives. Es defineix a través de tres components que són:

   1) Component aleatòria: És el vector d'observacions de la variable Y. Les observacions han de ser independents unes de les altres.S'assumeix que cada observació Y_i segueix una distribució de la forma:

      $$Y_i \sim \exp\Big(\frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y,\phi)\Big),$$

      on θ_i canvia com a funció de les covariants i ϕ és constant.

   2) Component determinista: Es una combinació lineal de les variables explicatives. Es denota per $X\beta$ on X és la matriu del disseny i, per tant, conté totes les condicions experimentals associades a les observacions.

   3) Una component enllaç, que és la que enllaça les dues componenta anteriors. Concretament, s'assumeix el model:

      $$g(\mu) = X\beta,$$

      On $\mu = E(Y)$ i g és una funció monòtona diferenciable que s'anomena funció enllaç.

   Si es considera que $Y_i \sim N(\mu_i, \sigma^2)$ i que la funció g és la identitat, i.e. $g(\mu) = \mu$, el GLM es redueix a un LM.

2) Demostra que la distribució Binomial pot ser utilitzada com a distribució de la variable resposta en un MLG (assumeix n conegut). Dedueix quin és el paràmetre canònic, el paràmetre de dispersió, la funció variància i el link canònic.

   Aquest apartat es va fer a classe, per tant el teniu als apunts.

3) Es possible utilitzar la distribució de probabilitat discreta que figura a continuació, com a distribució de la variable resposta en un MLG?.

   $$p(y;p) = \frac{-1}{\log(1-p)}\frac{p^y}{y}, \quad y = 1,2,3,\cdots \quad p \in (0,1)$$

   En cas de resposta afirmativa, expliciteu quin valor prenen el paràmetre canònic i el de dispersió.

   Sí que és possible, perquè

   $$p(y;p) = \exp(\log(\frac{-1}{\log(1-p)}\frac{p^y}{y})) = \exp(-\log(-\log(1-p)) + y\log(p) - \log(y))$$

   Com parant aquesta expressió amb la que apareix a la segona component d'un MLG s'obté que elparàmetre canònic és igual a: $\theta = \log(p)$ i el de dispersió igual a $\phi = 1$. A més, $C(y;\phi) = -\log(y)$ i $b(\theta) = \log(-\log(1-p))) = \log(-\log(1-e^{\theta}))$

4) En el cas d'un MLG, la variància de la variable resposta $Var(Y_i)$ canvia en funció de les covariables o és constant? Acompanya la teva resposta d'alguns exemples que ho clarifiquin.

En els MLG es permet que la variància canvii en funció de les variables explicatives perquè es té que $Var(Y_i) = a(\phi)V(\mu_i)$. Atès que en la component tercera d'una MLG es veu que el valor esperat canvia en funció de les covariants, també ho farà la variància. L´'únic cas en que no canvia és quan s'assumeix normalitat per la resposta, que en aquest cas $a(\phi) = \phi = \sigma^2$ i $V(\mu_i) = 1$. En el cas Poisson per exemple $Var(Y_i) = \mu_i$ i, per tant canvia a l'igual que ho fa l'esperança perquè són iguals.

Exercici 4. Contesta les preguntes següents:

1) Com es defineix el X^2 de Pearson generalitzat en un GLM? Calculeu a què és igual en el cas que la variable resposta sigui Normal i en el cas que la variable resposta sigui Poisson.

El χ^2 de Pearson generalitzat és una mesura de bondat d'ajust del nostre model. S'obté a partir dels residus crus (diferència entre observat i predit) estandaritzats apropiadament. Concretament és igual a:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Important tenir en compte que és un estadístic fàcil d'entendre (com més petit sigui millor perquè vol dir que observats i predits estan molt aprop), però que no ens permet comparar models aniuats.

Pel cas particular de la distribució Normal, com que tal com hem dit abans $V(\mu_i) = 1$, el $X^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ que és la suma de quadrats residuals.

En el cas Poisson, com que $V(\mu_i) = \mu_i$, $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$.

2) Imagineu-vos que heu ajustat un model de Poisson amb un total de 35 observacions i 6 paràmetres. El valor de l'estadístic X^2 de Pearson generalitzat després de portar a terme l'ajust és igual a 25.78. Porteu a terme el test que us permet decidir entre les hipòtesis: H_0: el model és apropiat vs H_1: el model no és apropiat.

Si el model de Poisson és apropiat, el valor de X^2 ha de provenir d'una distribució χ^2_{n-p}. Prenent $\alpha = 0.05$ i tenint en compte que $n-p = 35-5 = 29$ i que aquest test és unilateral perquè només rebutjarem quan X^2 sigui gran, es rebutja quan $X^2 > \chi^2_{0.95,29} = 42.55$. Atès que en el nostre cas això no es compleix perquè $25 < 42.55$, es conclou que no hi ha motius per a rebutjar el model de Poisson per a les nostres dades.

3) En l'entorn dels ML, definiu R^2 i R^2_{adj}? Quina és la seva utilitat?

Tant R^2 com R^2_{adj} són mesures de bondat d'ajust d'un ML. La definició del primer és:

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{ESS}{TSS},$$

i, en conseqüència és la proporció de variabilitat que hi ha en la variable resposta que és explicada pel nostre model. O dit d'una altre manera és u menys la proporció de variabilitat en la resposta no explicada pel nostre model. L'R^2 sempre creix quan al model s'hi afegeixen variables, encara que aquestes tinguin poc sentit. Per aquest motiu, i pel fet que no té en compte quants paràmetres hi ha al model, per a comparar models amb un nombre diferent de paràmetres, cal utilitzar una altre mesura que en aquest cas és R^2_{adj}. Aquesta darrera es calcula mitjançant

$$R^2_{adj} = 1 - \frac{ESS/(n-p)}{TSS/(p-1)},$$

en aquest cas les sumes de quadratas apareixen dividides pels seus graus de llibertat. Si es tenen dos models amb diferent nombre de paràmetres, el que té més paràmetres segur que donarà lloc a un R^2 més gran, però no necessàriament a un $R^2 - adj$ més gran.

4) Per a un ML, explica quina és la diferència entre un PI i un CI.

Els PI s'anomenen intervals de predicció, i són intervals que contenen el valor de la Y_i en unes certes condicions experimentals, amb probabilitat $1-\alpha$. El CI s'anomenen intervals de confiança i són intervals que contenen el valor esperat de Y_i, és a dir μ_i, amb una probabilitat igual a $1-\alpha$. Els intervals PI són sempre més amples que el CI, com a conseqüència de que la variabilitat de la variable resposta és sempre superior a lavariabilitat del valor esperat de la resposta. Ambdós intervals compelixen que a mida que les condicions experimentals s'allunyen del centre de gravetat de les condicions experimentals, els intervals es fan més amples, com a conseqüència de que estem intentant predir en unes condicions diferents de les utilitzades en l'experiment.