GRAU INTERUNIVERSITARI D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA (UB- UPC)

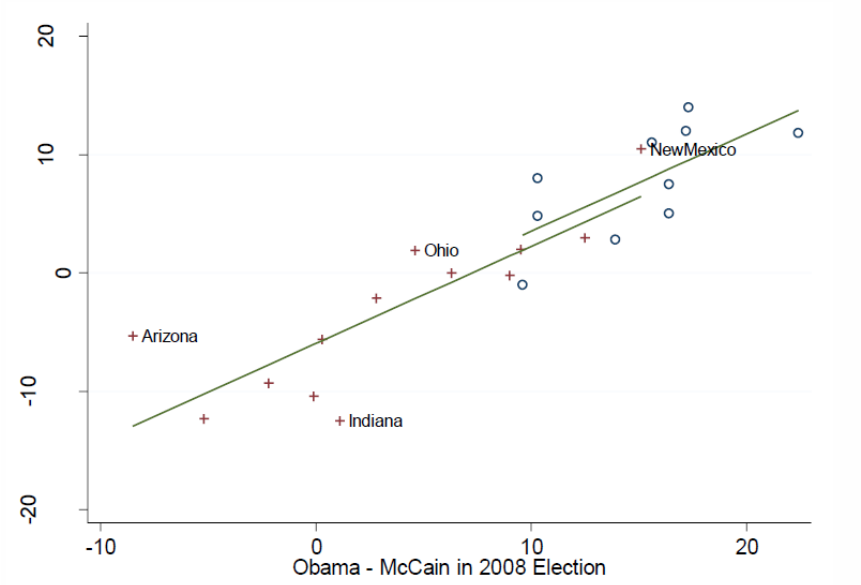
CURS 2014-2015 Q1 – EXAMEN REAVALUACIÓ :MODEL LINEAL GENERALITZAT

(Data: 7/07/2014 a les 18:00h Aula 002-FME)

Nom de l´alumne:	**DNI:**
Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	Edifici C5 D217 o H6-67
Normativa:	SÓN PERMESOS APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 00 min
Sortida de notes:	Abans del 13 Juliol a la Intranet Docent de MLGz
Revisió de l'examen:	13 Juliol a 13h a C5-217-C Nord (Problema 1) o H- P6-67 a 15h (Problemes 2)

Problema 1 (5 punts): Resposta normal

El gráfico muestra los datos de una encuesta de Intención de Voto por internet desarrollada Octubre de 2012 en 23 estados de EEUU de resultado dudoso. En 20 de los 23 estados ganó Obama en 2008. Se excluyen el resto de estados al tener una tendencia política demasiado clara. El eje Y indica el margen de Obama sobre Romney en las encuestas y en el eje X se muestra el margen de Obama sobre McCain en las elecciones del 2008. Los estados con resultado Demócrata en 2004 (Kerry se indican con un circulo) y los estados con resultado Republicano (Bush) con un signo +. Se indican adicionalmente los nombres de 4 estados de interés donde ganó Bush en 2004.



El margen de Obama en promedio en los 23 estados es 1.548 puntos porcentuales y su desviación estándar es 8.020. La correlación lineal entre el margen de Obama en la encuesta y su margen sobre McCain en 2008 es 0.8913. La regresión múltiple entre el margen de Obama en la encuesta y el margen en las elecciones del 2008 con el indicador de resultados demócratas en 2004 tiene un coeficiente de determinación de 0.7976 y la relación resultante es:

$$(\text{rcp2012}) = -5.955 + 0.8211 \times \text{margin2008} + 1.2777 \times \text{dem2004}$$

Si se añade la interacción entre el margen de Obama en las elecciones del 2008 y el factor binario para los demócratas en 2004, el estimador de la interacción muestra en el contraste de hipótesis de coeficiente igual a 0 un estadístico de Student de valor −0.0299.

1. Indicar cuál de los dos modelos de análisis de la covarianza es más adecuado: el aditivo o el completo.

La RSS del model nul és $8.020^2 \times 22 = 1415.05$. El marge del 2008 explicada un $0.8913^2 = 0.7944$ d'aquesta suma de quadrats total i per tant deixa per explicar 290.91 unitats. El model additiu explica 0.7976 de la TSS of 1128.64 deixant per explicar una RSS de 286.41, aleshores l'indicador del 2004 explica 4.50 unitats de la suma de quadrats total. Ara bé, per valorar la interacció s'ha d'usar la relació que el quadrat de l'estadístic d'Student és la F del contrast de variança incremental per la interacció: $t^2 = F$ i F=SSI/[(286.41−SSI)/19] on SSI és la suma de quadrats explicada per la interacció, ara ja només cal resoldre l'equació d'on surt que SSI és 0.01 i per tant RSS(Mcomplert) és 286.40. El model additiu és equivalent al model complert segons el contrast de variança incremental i per tant aquest model és el preferit.

2. Contrastar la hipótesis que el margen de Obama en la encuesta del 2012 no depende de su margen de elección en 2008, sin considerar por ahora el factor binario de resultado en 2004.

Comparar el model nul amb el model que té només l'efecte brut del marge de les eleccions al 2008.

$$F = \frac{(1415.05 - 290.91)/1}{(290.91)/(23-2)} = 81.15 \text{ amb } 1 \text{ i } 21 \text{ g. ll. Equivalentment, } t = \sqrt{F} = 9.01 \text{ amb } 21 \text{ g. ll.}$$

Per tant, els dos models no són equivalent i es pot concloure que una persistència

en els marges als llarg del temps

3. Considerar el modelo aditivo con el margen en el 2008 y el factor binario Demócrata en 2004. Interpretad el coeficiente del estimador de la *dummy* y contrastar si es estadísticamente significativo.

El coeficient 1.2777 sobre la dummy indica que el marge d'Obama és 1.28 punts percentuals superior en els estats on va guanyar Kerry al 2004, en front dels estats on a 2004 va sortir Bush amb el mateix marge per Obama al 2008.
A qualsevol nivell de significació l'estimador de la *dummy* de victòria de Kerry (demócrata) al 2004 no és estadísticament significatiu. : F=4.50/(286.4120)=0.31 amb 1 i 20 g.ll. Equivalentment, t=0.56 amb 20 g.ll. i per tant, el pvalor ($P(|t_{20}|>0.56)$) >0.05

4. Cuál es el coeficiente de correlación múltiple entre las predicciones con el modelo aditivo y el margen de Obama obtenido en las encuestas del 2012?

El quadrat del coeficient de correlació múltiple entre les prediccions del model additiu i el marge d'Obama al 2012 (les dades de la resposta) és el coeficient de determinació, que s'indica a l'enunciat que és del 0.7976, per tant la correlació múltiple és l'arrel quadrada $\sqrt{7976} = 0.8931$.

5. Contrastar la hipótesis que la asociación entre el margen de Obama en la Encuesta del 2012 y el margen obtenido en el 2008 no depende de si se obtuvo un resultado Demócrata en el estado en 2004.

Se está demanant un contrast sobre la interacció entre l'indicador de la dummy dem2004 i el marge d'Obama al 2008 (covariant). Tal com diu l'enunciat el contrast de l'estimador igual a zero mitjançant el test d'Student (el que surt per defecte a la taula que proporciona un summary per defecte) : és -0.0299 amb 19 g.ll i per tant, el pvalor ($P(|t_{19}|>0.0299)$) >0.05, acceptant-se la hipótesi nul.la d'on es conclou que la interacció no és estadísticament significativa.

6. Hay dos estados que tienen unos residuos estudentizados elevados en valor absoluto 2.73 y −2.26, siendo el siguiente residuo en magnitud 1.27. Indicad cuales son esos estados (están detallados en el diagrama bivariante).

Arizona es el positivo e Indiana el negativo.

7. Hay dos estados que tienen factores de anclaje elevados, pero no alarmantes (0.277 y 0.266, siendo el siguiente estado en magnitud según el factor de anclaje 0.191) y están indicados en el gráfico incluido. Detallad cuales son y justificad por qué tienen un anclaje elevado.

Arizona és l'estat amb el factor d'anclatge més elevat i alhora és l'estat més republicà dels 23 estats al 2008. L'altre observació amb factor d'anclatge és New Mexico que va donar el marge més elevat a Obama al 2008, malgrat ser un estat on va guanyar Bush al 2004.

8. Uno de los estados tiene una distancia de Cook de 0.720 (siendo el siguiente valor en magnitud 0.189) y está indicado en el diagrama bivariante. Indicad cual es y que pensaís que pasaría a la pendiente del margen del 2008 si se omitiera ese estado del análisis.

L'estat més influent és Arizona al tenir un factor d'anclatge elevat i un residu elevat (aquest estat mostra més suport a Obama del que caldria esperar pels resultats del 2008). Si s'ometés s'incrementaria la pendent apropant-la a 1.

Problema 2 (5 punts)

Los datos corresponden a 935 estudiantes del Grado en Ingeniería en Tecnologías Industriales que entraron por acceso de Selectividad entre los años 2010 y 2012 en la UPC. Se recogen las siguientes variables:

Sexe:	H: Hombre, D: Mujer
AnyEntrada:	Año de entrada en el grado (2010, 2011, 2012)
Barcelona:	Vive fuera de la provincia de Barcelona (S/N)
Informatica:	Nota superior a 7 en la asignatura Informática en la fase selectiva-Q1 (S/N)
NotaEstadistica:	Nota superior a 7 en la asignatura Estadística en tercero-Q5 (S/N)

Se quiere analizar cuáles de estos factores están asociados con el hecho de sacar más de un 7 (notable o excelente) en la asignatura de Estadística del Grado.

				Nota Estadistica >7	
Gènere	**AnyEntrada**	**Barcelona**	**Informatica>7**	**No**	**Si**
D	**2010**	N	N	9	1
D	**2010**	N	S	6	1
D	**2010**	S	N	35	4
D	**2010**	S	S	9	4
D	**2011**	N	N	13	0
D	**2011**	N	S	6	3
D	**2011**	S	N	24	5
D	**2011**	S	S	13	15
D	**2012**	N	N	6	0
D	**2012**	N	S	3	2
D	**2012**	S	N	38	9
D	**2012**	S	S	16	6
H	**2010**	N	N	15	4
H	**2010**	N	S	8	1
H	**2010**	S	N	114	6
H	**2010**	S	S	52	11
H	**2011**	N	N	19	5
H	**2011**	N	S	19	14
H	**2011**	S	N	110	15
H	**2011**	S	S	53	31
H	**2012**	N	N	15	6
H	**2012**	N	S	4	16
H	**2012**	S	N	79	12
H	**2012**	S	S	60	38

1. Determina la tabla de datos agregados necesaria para la estimación del modelo de respuesta binaria para la probabilidad de obtener una nota en la asignatura de estadística superior a 7 con el único efecto del año de entrada. ¿Cuál es la probabilidad de obtener una nota en Estadística superior a 7 que marginalmente corresponde a cada alumno?

Año de Entrada	Nota EST>7 (respuesta positiva)	Número Alumnos	Probabilidad
2010	32	280	0.114
2011	88	345	0.255
2012	89	310	0.287
	209	935	0.223

P('NotaEST>7')=209/935= 0,223

2. Estima manualmente a partir de la tabla del punto anterior y empleando la transformación ***logit*** cuál es el estimador del término constante en el modelo nulo.

$$logit(\pi_i)=log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta \quad \rightarrow \hat{\eta}=ln\left(\frac{209/935}{1-209/935}\right)=ln\left(\frac{0.223}{1-0.223}\right)=ln(0.2878)=-1.245$$

3. Estima manualmente a partir de la tabla del punto anterior y empleando la transformación ***probit*** cuáles son los estimadores de la constante y de los coeficiente de las *dummies* para el efecto bruto del **año de entrada** que incluye exclusivamente el factor AnyEntrada (nivel de referencia '2010').

$$\Phi^{-1}(\pi_i)=\eta+\alpha_i \quad i=2010,2011,2012 \quad \alpha_{2010}=0 \quad \rightarrow$$

$$\hat{\eta}=\Phi^{-1}(\pi_{2010})=-1.204$$
$$\hat{\alpha}_{2011}=\Phi^{-1}(\pi_{2011})-\Phi^{-1}(\pi_{2010})=0.545$$
$$\hat{\alpha}_{2012}=\Phi^{-1}(\pi_{2012})-\Phi^{-1}(\pi_{2010})=0.642$$

```
> summary(mod<-glm(cbind(Si,No)~AnyEntrada,notes,family=binomial(probit)))

Call:
glm(formula = cbind(Si, No) ~ AnyEntrada, family = binomial(probit),
    data = notes)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.747  -1.176  -0.022   1.525   4.757

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.20405    0.09839  -12.24  < 2e-16 ***
AnyEntrada2011  0.54543    0.12256    4.45 8.57e-06 ***
AnyEntrada2012  0.64216    0.12397    5.18 2.22e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 133.59  on 23  degrees of freedom
Residual deviance: 102.52  on 21  degrees of freedom
AIC: 181.89

Number of Fisher Scoring iterations: 4
```

4. Calcula el odds-ratio del **género** sobre la incidencia de **notas superiores a 7 en Estadística**. Interpreta el efecto del género en la escala de los odds de la probabilidad de obtener más de un 7 en EST. ¿Es significativo este efecto?

```
> p1=50/(50+178)
> p2=159/(159+548)
> log(p1/(1-p1))
```

```
[1] -1.269761
> log(p2/(1-p2))-log(p1/(1-p1))
[1] 0.03238946
> summary(mod<-glm(cbind(Si,No)~Sexe,notes,family=binomial))

Call:
glm(formula = cbind(Si, No) ~ Sexe, family = binomial, data = notes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3241  -1.1768  -0.3322   0.8163   5.4562

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.26976    0.16006  -7.933 2.14e-15 ***
SexeH        0.03239    0.18366   0.176     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 133.59  on 23  degrees of freedom
Residual deviance: 133.56  on 22  degrees of freedom
AIC: 210.93

Number of Fisher Scoring iterations: 4

> coef(mod)
(Intercept)       SexeH
-1.26976054  0.03238946
> exp(coef(mod))
(Intercept)       SexeH
  0.2808989   1.0329197
> 100*(1-exp(coef(mod)))
(Intercept)       SexeH
  71.910112   -3.291971
```

Els odds de treure una nota superior a 7 en Estadística augmenta en 3,29% en el grup dels homes respecte els odds del grup de referencia de les dones.

Aún así, este efecto es no significativo:

```
> anova(moda,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    23     133.59
Sexe  1 0.031198        22     133.56   0.8598
```

5. Calcula i interpreta en la escala del predictor lineal i del odds-ratio del efecto bruto de haber sacado **más de un 7 en informática** sobre el hecho de sacar más de un 7 en Estadística.

```
> p1=67/(67+477)
> p2=142/(142+249)
> log(p2/(1-p2))-log(p1/(1-p1))
[1] 1.401198
```

```
> summary(modd<-glm(cbind(Si,No)~Informatica,notes,family=binomial))

Call:
glm(formula = cbind(Si, No) ~ Informatica, family = binomial,
    data = notes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2957  -0.9901   0.0019   0.8426   4.0008

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.9628     0.1305 -15.045   <2e-16 ***
InformaticaS   1.4012     0.1676   8.362   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 133.589  on 23  degrees of freedom
Residual deviance:  58.386  on 22  degrees of freedom
AIC: 135.76

Number of Fisher Scoring iterations: 4
```

En l'escala del predictor lineal, aquest puja 1.4 unitats si s'ha tret més d'un 7 en Informàtica. En l'escala de l'odds-ratio, el factor d'augment és exp(1.4)=4.05 vegades respecte a la categoría base de no haver tret més d'un 7 en informática.

6. Determina si el efecto NETO de la **Pertenencia a Barcelona** es estadísticamente significativo, cuando el haber sacado **más de un 7 en Informática** y el **año de entrada** ya se encuentran en el modelo. Indica el valor del estadístico del test y la distribución de referencia que utilizas para justificar la decisión.

```
summary(modb<-
glm(cbind(Si,No)~AnyEntrada+Informatica+Barcelona,notes,family=binomial))>
anova(modb,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         23    133.589
AnyEntrada  2   31.070        21    102.519 1.791e-07 ***
Informatica 1   66.753        20     35.766 3.078e-16 ***
Barcelona   1    4.875        19     30.891   0.02724 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Si que ho és. Estadístico: 4.875, distribución de referencia: Chi-sq amb 1 grau de llibertat

7. El modelo que incluye el **año de entrada**, una **nota superior a 7 en Informática** y pertenecer a **Barcelona** ¿explica de forma adecuada los datos observados? Indica el estadístico y la distribución de referencia en que basas el test para justificar la respuesta.

Test sobre la deviància residual. Estadístic: 30.891, distribució de referència: chi-sq amb 19 graus de llibertat. El p-valor és 0.041, inferior al nivel de significació habitual del 5%. Per tant, el model no és satisfactori per explicar les dades observades.

```
> 1-pchisq(30.891,df=19)
[1] 0.04149844
```

8. ¿Existe evidencia para afirmar que los efectos de haber sacado **más de un 7 en Informática** no son los mismos según los distintos **años de entrada** en la incidencia notas superiores a 7 en Estadística? Indica el valor del estadístico y la distribución de referencia del test.

```
> mod1<-glm(cbind(Si,No)~AnyEntrada+Informatica,notes,family=binomial)
> mod2<-glm(cbind(Si,No)~AnyEntrada*Informatica,notes,family=binomial)
> anova(mod2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev
NULL                                    23    133.589
AnyEntrada            2   31.070        21    102.519
Informatica           1   66.753        20     35.766
AnyEntrada:Informatica 2    1.462        18     34.304

> anova(mod1,mod2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Si, No) ~ AnyEntrada + Informatica
Model 2: cbind(Si, No) ~ AnyEntrada * Informatica
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        20     35.766
2        18     34.304  2   1.4622   0.4814
```

Demana si la relació entre el factor "nota>7 en Informática" i la incidència de "notes>7 en Estadística" ha canviat per als diferents anys d'entrada, per tant, és un contrast entre el model complert i el model additiu, el qual té un pvalor superior al 5% habitual i per tant, no es pot rebutjar la hipòtesi nul.la: NO, la relació entre el factor NotaInformatica>7 i la incidència de NotaEstad>7 no depen de l'any d'entrada al grau. L'estadístic de referència és una Chi quadrat amb 2 graus de llibertat i l'estadístic val 1.462.

RESULTATS R

```
> summary(notes)
 Sexe   AnyEntrada Barcelona Informatica       No              Si
 D:12   2010:8     N:12      N:12        Min.   :  3.00   Min.   : 0.000
 H:12   2011:8     S:12      S:12        1st Qu.:  8.75   1st Qu.: 2.750
        2012:8                           Median : 15.50   Median : 5.500
                                         Mean   : 30.25   Mean   : 8.708
                                         3rd Qu.: 41.50   3rd Qu.:12.500
                                         Max.   :114.00   Max.   :38.000
> by(notes$No, notes$Sexe,sum)
notes$Sexe: D
[1] 178
------------------------------------------------------------------------
```

```
notes$Sexe: H
[1] 548
> by(notes$Si, notes$Sexe,sum)
notes$Sexe: D
[1] 50
------------------------------------------------------------------------
notes$Sexe: H
[1] 159

> by(notes$No, notes$Informatica,sum)
notes$Informatica: N
[1] 477
------------------------------------------------------------------------
notes$Informatica: S
[1] 249
> by(notes$Si, notes$Informatica,sum)
notes$Informatica: N
[1] 67
------------------------------------------------------------------------
notes$Informatica: S
[1] 142


> anova(moda,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    23     133.59
Sexe  1 0.031198        22     133.56   0.8598

> anova(modb,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                         23     133.59
Barcelona  1   7.1301        22     126.46  0.00758 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(modc,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: probit

Response: cbind(Si, No)

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          23     133.59
AnyEntrada  2    31.07        21     102.52 1.791e-07 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(modd,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           23    133.589
Informatica  1   75.204        22     58.386 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> anova(modb)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev
NULL                           23    133.589
AnyEntrada   2   31.070        21    102.519
Informatica  1   66.753        20     35.766
Barcelona    1    4.875        19     30.891
> anova(mod2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


                      Df Deviance Resid. Df Resid. Dev
NULL                                     23    133.589
AnyEntrada             2   31.070        21    102.519
Informatica            1   66.753        20     35.766
AnyEntrada:Informatica 2    1.462        18     34.304
```