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

CURS 2012-2013 Q1 – EXAMEN REAVALUACIÓ: MODEL LINEAL GENERALITZAT

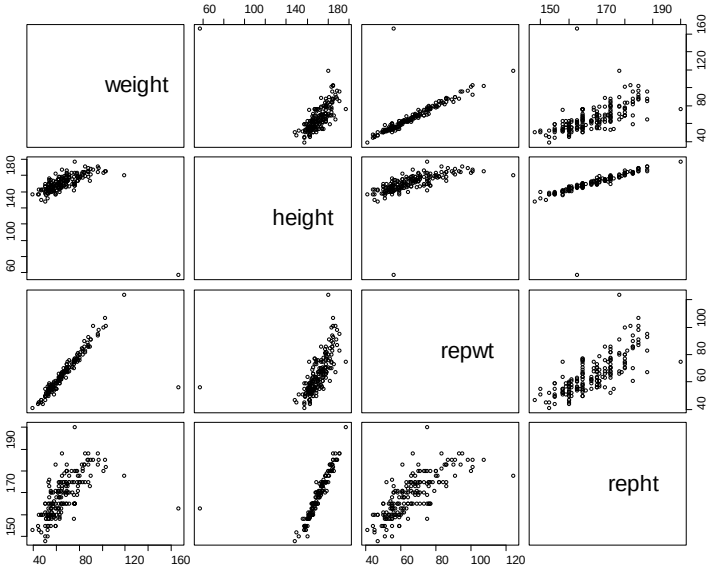
(Data: 1/07/2013 a les 18:00h Aula 004-FME)

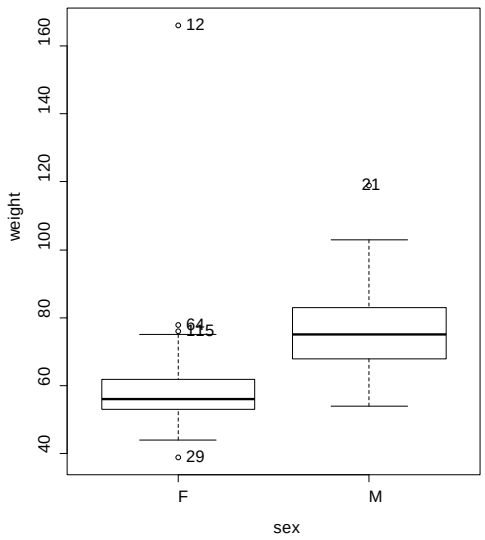
Nom de l´alumne:	
DNI:	
Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	Edifici C5 D217 o H6-67
Normativa de l'examen:	ÉS PERMÉS DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 00 min
Sortida de notes:	Abans del 5 de Juliol al Web Docent de MLGz
Revisió de l'examen:	5 de Juliol a 10:00 h a C5-217-C Nord o H- P6-67

Problema 1 (5 punts): Peso real y percibido

El conjunto de datos Davis contiene 181 registros de individuos que realizan ejercicio de forma regular. La información contenida incluye: el sexo del individuo, su peso (kg) y altura (cm) y el peso y altura reportados por los propios individuos.

```
sex        weight          height          repwt           repht
 F:99   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0
 M:82   1st Qu.: 56.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:161.0
        Median : 63.0   Median :169.0   Median : 63.00   Median :168.0
        Mean   : 66.3   Mean   :170.2   Mean   : 65.68   Mean   :168.7
        3rd Qu.: 75.0   3rd Qu.:178.0   3rd Qu.: 74.00   3rd Qu.:175.0
        Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
```





Se plantea ver la correspondencia entre el peso real y el peso que el individuo reporta. Para ello, ajustamos un modelo ANCOVA para ver si hay diferencias entre la relación de interés dependiendo del sexo del individuo.

```
Call:
lm(formula = weight ~ repwt * sex, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max
 -7.655  -1.859  -0.777   0.661 108.346
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.950399   7.243291   0.545    0.586
repwt        0.958986   0.126933   7.555 2.16e-12 ***
sexM        -2.156119   9.373734  -0.230    0.818
repwt:sexM   0.009932   0.148330   0.067    0.947
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.487 on 177 degrees of freedom
Multiple R-squared: 0.6991,      Adjusted R-squared: 0.694
F-statistic: 137.1 on 3 and 177 DF,  p-value: < 2.2e-16
```

1. A partir del ajuste del modelo ¿cómo se interpretan los coeficientes asociados a sexM y repwt:sexM? ¿Hay alguna influencia del sexo del individuo en la relación entre su peso reportado y su peso real? Justifica la respuesta.

El coeficiente asociado a la variable sexM corresponde al efecto sobre el nivel global de la relación entre el peso y el peso reportado para un hombre si pasa a considerarse esta relación en una mujer. Sería el efecto sbre la ordenada en el origen del modelo de regresión lineal. En cambio, la interacción repwt:sexM correspondería al efecto en la pendiente del modelo lineal del hombre respecto a la mujer. En este caso, ambas variables son nmo significativas, indicando que la relación entre el peso y el peso reportado es la misma para hombres y mujeres tanto en el nivel como en la pendiente de la recta lineal.

2. En base al ajuste anterior, realiza la prueba de hipótesis para determinar si el coeficiente asociado a la variable repwt en el modelo teórico puede valer 1.

El ajuste del modelo indica que la estimación del coeficiente asociado a la variable numérica repwt (pendiente de la recta de regresión lineal) vale 0.959 y su error estándar 0.127. Por lo tanto, planteando el test de Wald para contrastar el valor de referencia 1 tenemos:

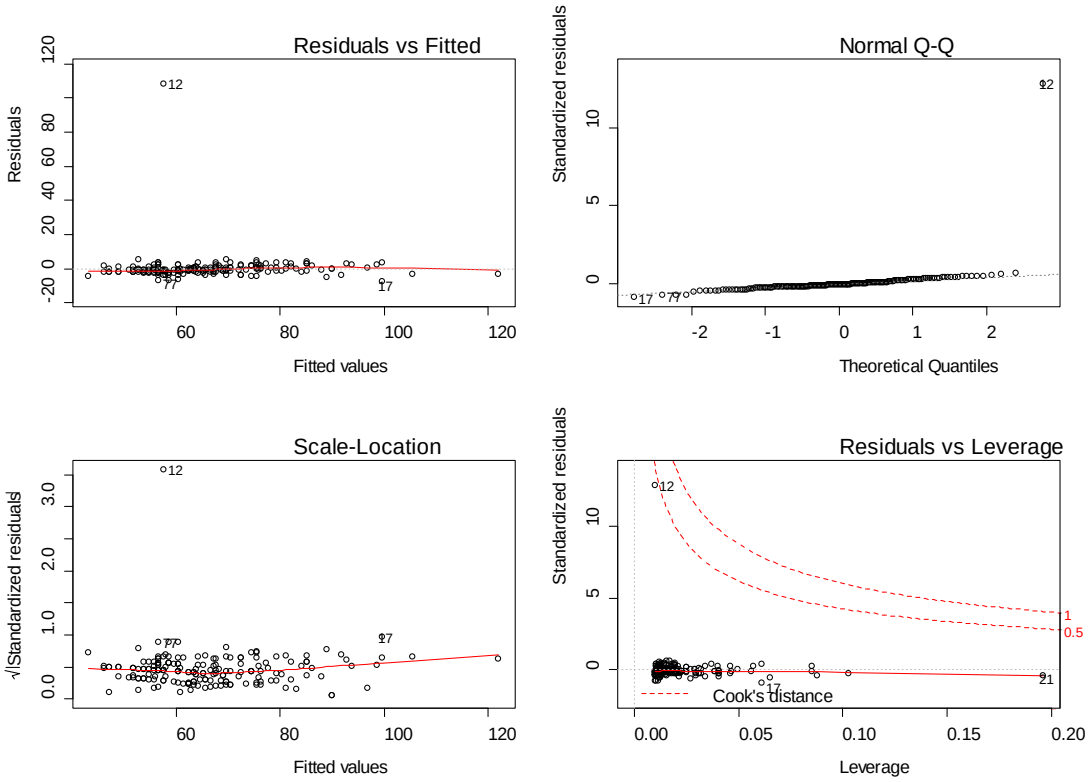
$$H_0: \beta = 1 \qquad H_1: \beta \neq 1$$

$$\hat{t} = \frac{\hat{\beta} - 1}{S_{\hat{\beta}}} \sim_{H_0} t_{n-4} \approx N(0,1)$$

$$\hat{t} = \frac{0.959 - 1}{0.127} = -0.322$$

I com $|-0.322|<1.96=Z_{0.975}$ la decisión es por la hipòtesis nula que indica que no hay evidencias estadísticas significativas para rechazar el valor de 1 del modelo teórico.

Los plots de validación del modelo aparecen a continuación:



3. Indica en cada plot que premisas del modelo lineal se puede analizar y que información aporta para la validación del modelo. Caracteriza las observaciones 12 y 21 en términos de error estandarizado, factor de apalancamiento/anclaje y distancia de Cook.

El primer plot corresponde a los residuos frente a las predicciones del modelo. Este plot nos permite evaluar la linealidad de los datos y la varianza constante de los residuos, si observamos una disposición aleatoria de los puntos en el gráfico y sin cambios en la variabilidad. Este plot también permite detectar valores atípicos correspondientes a observaciones mal explicadas por el modelo.
El segundo plot es el plot de normalidad, para verificar el supuesto de que los residuos provienen de una distribución normal.
El tercer plot representa la raíz cuadrada de los valores absolutos de los residuos estandarizados frente a las predicciones. Es similar al primero pero permite incidir en el análisis de la variabilidad para comprobar la hipótesis de homocedasticidad (varianza constante). Tanto este plot como el primero incluyen un ajuste suave para facilitar la interpretación.
El último plot refleja las componentes de la medida de influencia (distancia de Cook). En el eje de abscisas se refleja el factor de apalancamiento (leverage) y en el eje de ordenadas, el residuo estandarizado. Ademàs se incluyen curvas de nivel para indicar la posición relativa de cada observación según su distancia de Cook.
La observación 12 es un valor atípico que no es correctamente explicado por el modelo. La magnitud de su residuo estandarizado hace difícil considerarlo un dato extremo y seguramente corresponde a un error de entrada de datos (hay una altura de 57 cm y la altura mínima percibida es de 148cm!). Los plots descriptivos ya indicaban la presencia de esta dato anómalo. Por otro lado, no corresponde a un valor extremo del espacio de los predictores, ya que su leverage no es excesivo. Sin embargo, la distancia de Cook resultante es considerable respecto al resto de observaciones (próxima a 0.5), indicando que es un dato altamente influyente, que afecta a la estimación y que se debería eliminar del ajuste.
La observación 21 no presenta un residuo excesivo pero presenta un leverage muy alto. Corresponde a un valor extremo del espacio de los predictores pero que no es influyente ya que su distancia de Cook no es excesiva.

Se ajusta un modelo lineal general con los datos reportados por el individuo y su sexo, sin interacciones y habiendo eliminado la observación 12:

```
> summary(davis.mod)

Call:
lm(formula = weight ~ repht + repwt + sex, data = Davis, subset = -12)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7006 -1.0920 -0.2427  1.3194  6.3200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.513669   4.488014   0.337    0.736
repht        0.004664   0.030380   0.154    0.878
repwt        0.969114   0.019991  48.478   <2e-16 ***
sexM        -0.556341   0.532440  -1.045    0.298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.246 on 176 degrees of freedom
Multiple R-squared: 0.9726,      Adjusted R-squared: 0.9721
F-statistic:  2080 on 3 and 176 DF,  p-value: < 2.2e-16

> anova(davis.mod)
Analysis of Variance Table

Response: weight
           Df  Sum Sq Mean Sq   F value Pr(>F)
repht       1 18116.8 18116.8 3591.3748 <2e-16 ***
repwt       1 13357.6 13357.6 2647.9272 <2e-16 ***
sex         1     5.5     5.5    1.0918 0.2975
Residuals 176   887.8     5.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. ¿A que son debidas las discrepancias en cuanto a la significación de la variable repht según se considere la tabla del summary o la tabla del anova? ¿Se debe eliminar la variable del modelo por ser no significativa o se ha de mantener?

El p-valor de la tabla summary corresponde al test que comprueba la significación del parámetro en el modelo, comparando el modelo que tiene todos los predictores con el modelo que contiene todos los parámetros menos el del test. Según el resultado, un p-valor de 0.878 indica que el parámetro no se debe incluir en el modelo ya que la variabilidad explicada por este predictor no es significativa.
En cambio, el método anova incluye la inferencia de tipo secuencial, es decir, permite establecer la significación de los coeficientes a medida que son introducidos en el modelo. Por ello, el p-valor reportado para la variable repht corresponde al test que compara el modelo que no contiene ningún predictor con el que contiene sólo esta variable. En ausencia del predictor que mejor explica la respuesta (repwt), la covariable testada posee alguna significación (la que se extrae de la regresión lineal simple entre esta variable y la respuesta).

5. ¿Qué diferencias hay entre la R-squared y la Adjusted R-squared (no hace falta poner las expresiones exactas)? ¿Para qué se utilizan?

La R-squared corresponde al coeficiente de determinación y representa la proporción de variabilidad de la respuesta que es explicada por el modelo. Como porcentaje, es un valor entre 0 y 100, de forma que cuando más proxima a 100 es, menor es la variabilidad residual que es la parte no explicada por el modelo.

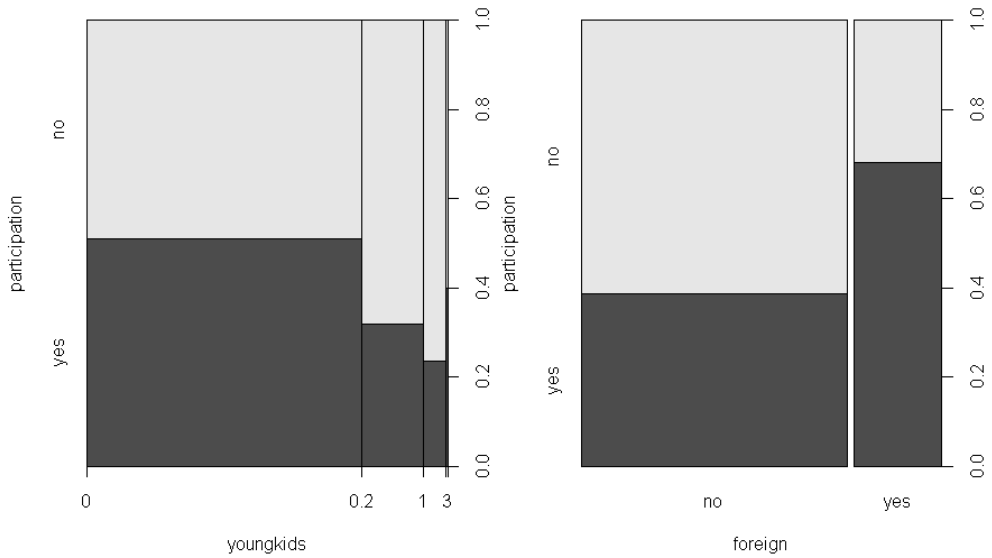
La Adjusted R-squared es el cálculo de la R-squared corregida por los grados de libertad. También es una medida de ajuste expresada como porcentaje con una interpretación similar a la del coeficiente de determinación.
En caso de añadir nuevos predictores al modelo, la R-squared siempre aumenta, en cambio la Adjusted R-squared decrece en caso de que los predictores que se incluyen en el modelo no sean significativos. Esta característica hace que a la hora de construir el modelo sea más adecuado controlar la Adjusted R-squared porque cuando empieza a decrecer es síntoma de que no debemos incluir en el modelo las últimas variables introducidas.

Problema 1. Participación Laboral femenina en Suiza (5 puntos)

Los datos SwissLabor del paquete AER de R son un juego de datos trabajado originariamente por Gerfin (1996) que contienen indicación de las características de la participación laboral femenina en una muestra de 872 mujeres de Suiza. Hay muchas variables pero de entrada se va a trabajar con la respuesta partipación laboral y la presencia de hijos pequeños en el hogar o extranjería. La siguiente tabla muestra los datos agregados correspondientes.

Factor Younkids	Participación - NO		Participación - SI		m
	Foreign-NO	Foreign-YES	Foreign-NO	Foreign-YES	
0	288	38	225	114	**665**
1	80	20	23	24	**147**
2	32	12	6	7	**55**
3	2	1	0	2	**5**
	471		***401***		**872**

```
> summary(dfex)
 participation     income           age          education      youngkids oldkids foreign
 no :471       Min.   : 7.187   Min.   :2.000   Min.   : 1.000   0:665     0:393   no :656
 yes:401       1st Qu.:10.472   1st Qu.:3.200   1st Qu.: 8.000   1:147     1:198   yes:216
               Median :10.643   Median :3.900   Median : 9.000   2: 55     2:208
               Mean   :10.686   Mean   :3.996   Mean   : 9.307   3:  5     3: 55
               3rd Qu.:10.887   3rd Qu.:4.800   3rd Qu.:12.000             4: 14
               Max.   :12.376   Max.   :6.200   Max.   :21.000             5:  2
                                                                           6:  2
```




```
>
```

```
> anova(m0);anova(m1);anova(m2)
Analysis of Deviance Table
Response: participation

     Df Deviance Resid. Df Resid. Dev
NULL                   871     1203.2

Analysis of Deviance Table
Response: participation
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                       871     1203.2
youngkids  3   30.469       868     1172.8

Analysis of Deviance Table
Response: participation
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev
NULL                     871     1203.2
foreign  1   56.867       870     1146.4
> anova(m1,m3,test="Chis")
Analysis of Deviance Table

Model 1: participation ~ youngkids
Model 2: participation ~ youngkids + foreign
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       868     1172.8
2       867     1105.7  1    67.08 2.607e-16 ***

> anova(m2,m3,test="Chis")
Analysis of Deviance Table

Model 1: participation ~ foreign
Model 2: participation ~ youngkids + foreign
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       870     1146.4
2       867     1105.7  3   40.681 7.641e-09 ***

>
```

1. Calcular el modelo nulo según la transformación logit

```
> log(401/471)
[1] -0.1608967
> summary(m0)

Call:
glm(formula = participation ~ 1, family = binomial, data = dfex)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.16090    0.06795  -2.368   0.0179 *

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1203.2  on 871  degrees of freedom
AIC: 1205.2
```

2. Estimar manualmente el modelo de regresión logística para modelar la probabilidad de trabajar (participación laboral SI) según la presencia de hijos en el hogar.

```
>> summary(m1)

Call:
glm(formula = participation ~ youngkids, family = binomial, data = dfex)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03910    0.07757   0.504 0.614201
youngkids1  -0.79413    0.19312  -4.112 3.92e-05 ***
youngkids2  -1.21182    0.32673  -3.709 0.000208 ***
youngkids3  -0.44457    0.91616  -0.485 0.627498


    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1172.8  on 868  degrees of freedom
AIC: 1180.8

Number of Fisher Scoring iterations: 4

> log(339/326)
[1] 0.03910273
>log(339/326)-log(47/100);log(339/326)-log(13/42);log(339/326)-log(2/3)
[1] 0.7941253
[1] 1.211823
[1] 0.4445678
>
```

3. La probabilidad de participación laboral depende del número de hijos pequeños? Formular la hipótesis nula e indicar algún estadístico adecuado de bondad del ajuste y cálculo del pvalor de la hipótesis nula.

```
> anova(m0,m1,test="Chis")
Analysis of Deviance Table

Model 1: participation ~ 1
Model 2: participation ~ youngkids
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       871     1203.2
2       868     1172.8  3   30.469 1.100e-06 ***

H0: Rebutjada – La participación laboral depèn del nombre de fills de la llar
```

4. La probabilidad de participación laboral positiva depende de la procedencia de extranjería? Formular la hipótesis nula e indicar algún estadístico adecuado de bondad del ajuste y cálculo del pvalor de la hipótesis nula.

```
> anova(m0,m2,test="Chis")
Analysis of Deviance Table

Model 1: participation ~ 1
Model 2: participation ~ foreign
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       871     1203.2
2       870     1146.4  1   56.867 4.662e-14 ***
H0: Rebutjada – La participación laboral depèn de la situación d'estrangeria
```

5. Calcular el modelo probit nulo

```
> qnorm(401/(401+471))
[1] -0.1007804
> m0<-glm(participation~1,family=binomial(link=probit),data=dfex)
> summary(m0)

Call:
glm(formula = participation ~ 1, family = binomial(link = probit),
    data = dfex)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.10078    0.04252   -2.37   0.0178 *
```

6. Calcular el modelo probit para modelar la probabilidad de trabajar (participación laboral positiva) según la presencia de hijos en el hogar.

```
> summary(m1)

Call:
glm(formula = participation ~ youngkids, family = binomial(link = probit),
    data = dfex)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.02450    0.04861   0.504 0.614181
youngkids1  -0.49296    0.11807  -4.175 2.98e-05 ***
youngkids2  -0.74255    0.19208  -3.866 0.000111 ***
youngkids3  -0.27785    0.56916  -0.488 0.625428

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1172.8  on 868  degrees of freedom
AIC: 1180.8

Number of Fisher Scoring iterations: 4

> qnorm(339/(339+326))
[1] 0.02450333
> qnorm(47/(47+100))-qnorm(339/(339+326))
[1] -0.4929632
> qnorm(13/(13+42))-qnorm(339/(339+326))
[1] -0.742552
> qnorm(2/(5))-qnorm(339/(339+326))
[1] -0.2778504
```

7. Pensaís que el efecto neto/bruto del número de hijos es más importante que el efecto neto/bruto de la condición de extranjería?

```
Es disposa dels contrastos per deviança amb l'enllaç logit. L'efecte brut i
net de foreign sempre suposa una reducció superior de la deviança del model,
ja sigui respecte el model nul (brut) o el model amb el factor fills petits
(net).
```