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

CURS 2015-2016 Q1 – EXAMEN PARCIAL : MODEL LINEAL GENERALITZAT

(Data: 16 d'Octubre del 2015 a les 15:00 Aula S04-FME)

Professors:	Lídia Montero – Josep Anton Sanchez
Localització:	ETSEIB 6a Planta 6-67
Normativa de l'examen:	ÉS POT DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 00 min
Sortida de notes:	Abans del 29 d'Octubre al Web Docent de MLGz
Revisió de l'examen:	29 d'octubre a 16 h a Sala Professors FME– Campus Sud

El conjunto de datos contiene información de deportistas de alto nivel del "Australian Institute of Sport". (Referencia: Cook, R. D., and Weisberg, S. (1994). An Introduction to Regression Graphics. Wiley, New York.). En concreto se han seleccionado los datos de 113 deportistas de 4 disciplinas: Voley playa (BBall), Remo (Row), Natación (Swim) y atletismo en 400 m lisos (T400m). Los campos son los siguientes:

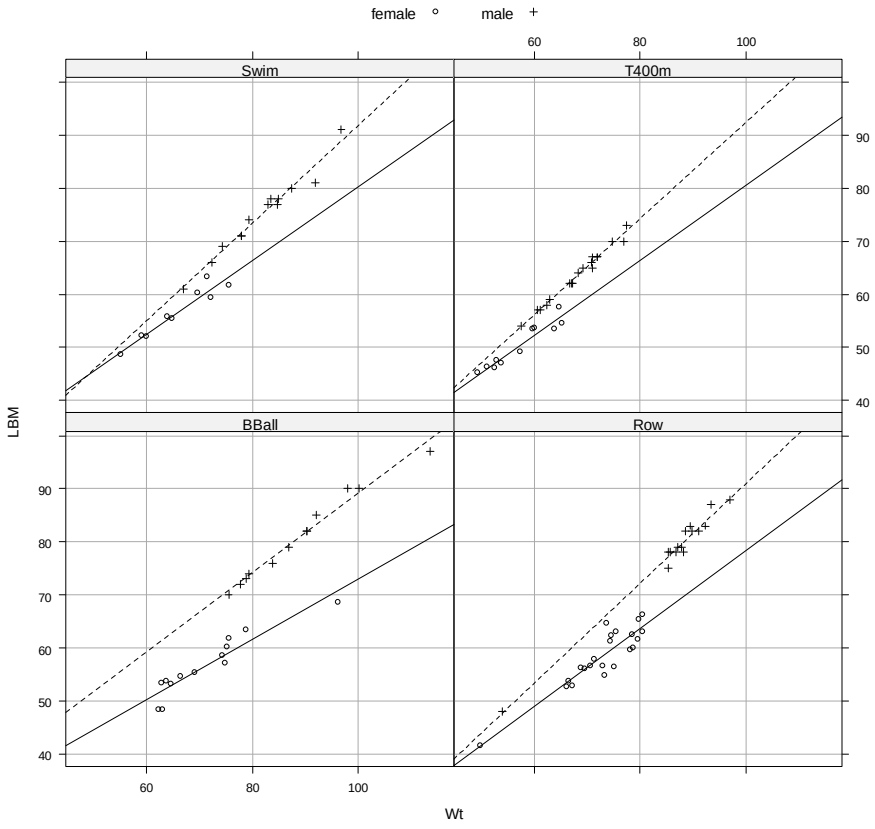
```
Sex:  Género (female/male)
Sport:Deporte que practica (BBall, Row, Swim, T400m)
RCC:  Recuento de Hematíes en sangre (millones por microlitro of sangre)
WCC:  Recuento de Leucocitos (Miles por microlitro de sangre)
Hc:   % Hematocrito (fracción de hematíes del total del volumen de sangre)
Hg:   Hemoglobina (g/dL)
Ferr: Concentración de Ferritina en plasma (ng/mL)
BMI:  Índice de masa corboral - body mass index (weight/height2)
SSF:  Suma de los siete plieges principales (mm)
Bfat: % grasa corporal
LBM:  Masa muscular – Lean Body Mass (kg)
Ht:   Altura(cm)
Wt:   Peso (kg)
```

En primer lugar se trata de analizar la relación lineal que hay entre el Peso (Wt) y la Masa Muscular (LBM) que actúa como respuesta, teniendo en cuenta el Género y el Deporte que practica. La descriptiva numérica y la gráfica es la siguiente:

```
              LBM        Wt
Min.     41.54000  49.20000
1st Qu.  55.73000  66.00000
Median   62.46000  74.40000
Mean     65.07000  74.43000
3rd Qu.  75.00000  83.50000
Max.     97.00000 113.70000
sd       12.14763  12.36673
```

Sex	Sport	Wt.Min.	Wt.Q1	Wt.Median
female	BBall	62.3	63.7	69.1
female	Row	49.8	69.7	73.9
female	Swim	55.1	60.0	64.8
female	T400m	49.2	52.6	57.3
male	BBall	75.5	79.2	88.6
male	Row	53.8	86.2	88.2
male	Swim	67.0	78.0	83.0
male	T400m	57.4	63.8	68.7

Sex	Sport	Wt.Mean	Wt.Q3	Wt.Max.
female	BBall	71.3	75.2	96.3
female	Row	72.9	78.4	80.5
female	Swim	65.7	71.4	75.6
female	T400m	57.2	61.8	65.2
male	BBall	88.9	93.6	113.7
male	Row	86.8	90.4	97.0
male	Swim	81.6	85.0	96.9
male	T400m	68.2	71.6	77.5



El modelo ajustado que considera la variable explicativa Peso y categóricas Sexo y Deporte, incluyendo las interacciones de orden 2 de la numérica con las categóricas es:

```
m1 <- lm(formula = LBM ~ Wt * (Sex + Sport), data = dades)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2266 -0.9024  0.0279  0.9209  5.8850

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.008507   2.847227   4.920 3.29e-06 ***
Wt              0.610053   --(1)--   15.931  < 2e-16 ***
Sexmale       -10.596595   2.660872  -3.982 0.000127 ***
SportRow       -5.789240   3.085363  -1.876 --(2)--
SportSwim       --(3)--    3.446450   0.637 0.525844
SportT400m     -0.862153   3.777918  -0.228 0.819937
Wt:Sexmale      0.250048   0.035708   7.003 2.65e-10 ***
Wt:SportRow     0.076302   0.038523   1.981 0.050296 .
Wt:SportSwim   -0.006173   0.044419  -0.139 0.889743
Wt:SportT400m   0.036916   0.055468   0.666 0.507194
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.762 on 103 degrees of freedom
Multiple R-squared: --(4)--, Adjusted R-squared: --(5)--
F-statistic: 579.9 on --(6)-- and --(7)-- DF,  p-value: < 2.2e-16
```

1) Completa la tabla para los valores 1 al 6, indicando como realizas el cálculo.

1- Error estándar de $\hat{\beta}_{Wt}$: $\frac{0.610053}{s_{\hat{\beta}}} = 15.931 \rightarrow S_{\hat{\beta}} = 0.03829$

2- P-valor de $\hat{t}_{SportRow}$: $P(|t_{103}| < |-1.876|) = 0.06$

3- Estimación $\hat{\beta}_{SportSwim}$: $\frac{\hat{\beta}_{SportSwim}}{3.44645} = 0.637 \rightarrow \hat{\beta}_{SportSwim} = 2.19$

4- Coeficiente de Determinación R^2 : $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{103*1.762^2}{112*12.1476^2} = 0.9806$

5- Coeficiente de Determinación Ajustado: $R^2_{adj} = 1 - \frac{S^2_R}{S^2_{LBM}} = 1 - \frac{1.762^2}{12.1476^2} = 0.979$

6- Grados de libertad del numerador: p-1 = 10-1 = 9

7- Grados de libertad del denominador: N-p = 113-10 = 103

Para decidir las variables que incluye el modelo teórico se dispone de las siguientes salidas:

```
> anova(m1)
Analysis of Variance Table

Response: LBM
           Df  Sum Sq Mean Sq   F value    Pr(>F)
Wt          1 14391.0 14391.0 4634.0227 < 2.2e-16 ***
Sex         1  1540.7  1540.7  496.1230 < 2.2e-16 ***
Sport       3    69.6    23.2    7.4700 0.0001421 ***
Wt:Sex      1   190.2   190.2   61.2571 4.609e-12 ***
Wt:Sport    3    15.9     5.3    1.7072 0.1701734
Residuals 103   319.9     3.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Anova(m1,type=2)
Anova Table (Type II tests)

Response: LBM
          Sum Sq  Df   F value    Pr(>F)
Wt        4427.0   1 1425.5487 < 2.2e-16 ***
Sex        878.5   1  282.8776 < 2.2e-16 ***
Sport       54.8   3    5.8797  0.000957 ***
Wt:Sex     152.3   1   49.0357 2.647e-10 ***
Wt:Sport    15.9   3    1.7072  0.170173
Residuals  319.9 103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Anova(m1,type=3)
Anova Table (Type III tests)

Response: LBM
            Sum Sq  Df  F value    Pr(>F)
(Intercept)  75.17   1  24.2069 3.292e-06 ***
Wt          788.18   1 253.8024 < 2.2e-16 ***
```

```
Sex          49.25   1  15.8593 0.0001274 ***
Sport        19.06   3   2.0459 0.1120732
Wt:Sex      152.28   1  49.0357 2.647e-10 ***
Wt:Sport     15.90   3   1.7072 0.1701734
Residuals   319.87 103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2) **Explica a que es debido que el término Sport presente tres p-valores diferentes. Especifica las hipótesis de los tests asociados a cada una de las tres opciones para el término Sport. ¿Consideras que este predictor ha de estar incluido en el modelo? Justifica la respuesta.**

El primer método corresponde al test de varianzas incrementales, que considera la inclusión secuencial de los términos. Para una línea de esta tabla, el test compara el modelo que incluye los términos anteriores con el modelo que incorpora el nuevo término.
El segundo método corresponde el test del ANOVA con suma de cuadrados tipo 2, lo cual quiere decir que se compara el modelo completo con el que resulta de suprimir el término de la línea y el resto de términos (interacciones de orden superior) donde aparezca. Con este método se preserva el principio de marginalidad.
Finalmente, el último método considera únicamente la eliminación del término de la línea, sin eliminar interacciones donde pueda aparecer. Este último caso, para variables explicativas numéricas y binarias, que poseen un grado de libertad, equivale a la aplicación del test de Wald para el coeficiente asociado y reproduce los p-valores de la tabla de coeficientes del método summary.

Para el primer método la línea Sport, en notación de R, hace el siguiente test:
H0: LBM~Wt+Sex
H1: LBM~Wt+Sex+Sport

F-Value=7.47, p-value=0.0001421 ➔ Sport es un término significativo en el modelo

Para el segundo método la línea Sport, en notación de R, hace el siguiente test:
H0: LBM~Wt+Sex+Wt:Sex
H1: LBM~Wt+Sex+Wt:Sex+Sport

F-Value=5.8797, p-value=0.000957 ➔ Sport es un término significativo en el modelo

En el tercer método, la línea Sport hace el siguiente test:
H0: LBM~Wt+Sex+Wt:Sex+ Wt:Sport
H1: LBM~Wt+Sex+Wt:Sex +Sport+ Wt:Sport

F-Value=2.0459, p-value=0.11207 ➔ En este caso, con un 5% de nivel de significación, aceptaríamos que el término no es significativo, pero el modelo que resulta incorpora una interacción en la que interviene el término Sport. No es adecuado tomar la decisión en base a un test que compara dos modelos con uno de ellos que posee interacciones y uno de los términos que la conforman no está incluido.
Como conclusión, el término Sport debe estar en el modelo.

Independientemente de la respuesta anterior, se decide trabajar a continuación con el siguiente modelo:

```
m2 <- lm(formula = LBM ~ Wt + Sex + Sport + Wt:Sex, data = dades)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1693 -0.8034  0.0913  0.8249  5.6650

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.96250    2.12646   5.626 1.52e-07 ***
Wt            0.63518    0.02976  21.345  < 2e-16 ***
Sexmale     -10.74935    2.40708  -4.466 2.00e-05 ***
SportRow      0.24318    0.46343   0.525 0.600868
SportSwim     1.84821    0.53849   3.432 0.000855 ***
SportT400m    1.91985    0.61770   3.108 0.002418 **
Wt:Sexmale    0.25274    0.03261   7.750 5.89e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 106 degrees of freedom
Multiple R-squared:  0.9797, Adjusted R-squared:  0.9785
F-statistic: 851.9 on 6 and 106 DF,  p-value: < 2.2e-16
```

3) **En base a la tabla de coeficientes y a las tablas del ANOVA anteriores, ¿se puede simplificar el modelo eliminando términos no significativos? Indica los tests, estadísticos de contraste y p-valores en que basas tu respuesta.**

A pesar de que uno de los p-valores de la tabla de coeficientes es no-significativo (el correspondiente al parámetro SportRow) para decidir la eliminación de este término se debe considerar la variable categórica al completo, Sport, que tiene 3 grados de libertad. De acuerdo a lo argumentado en el punto precedente y puesto que el modelo está anidado en el considerado anteriormente, la variable globalmente es significativa y por lo tanto no es adecuado eliminarla del modelo:

Sport: test de la F con el método Anova y suma de cuadrados tipo 2: F-value=5.87, p-value=0.00095

El resto de variables son cuantitativas o binarias o interacción entre ambas y todas poseen un grado de libertad y un p-valor que establece la significación de todos los términos:

Wt: Test de wald de la tabla summary: t-value=21.345, p-value<2e-16
Sex: Test de wald de la tabla summary: t-value=-4.466, p-value=2e-5
Wt:Sex: Test de wald de la tabla summary: t-value=7.75, p-value=5.89e-12

4) **Suponiendo que se ha validado el modelo, interpreta los coeficientes estimados. ¿Hay diferencias en la relación entre el peso y la masa corporal según sexo? ¿Y según el deporte que se práctica? Si existen diferencias, indica de qué tipo son. Justifica la respuesta indicando los tests, estadísticos y p-valores correspondientes.**

El contraste activo es el que se aplica por defecto: tipo baseline con la primera categoría como referencia. Hay dos variables categóricas, Sex y Sport, por lo tanto el Intercept se asocia al valor esperado de la variable respuesta para Wt=0 cuando el individuo es una mujer (Sex=female) que practica Voley Playa (Sport=BBall).
El coeficiente estimado para la variable Wt se puede interpretar como el incremento en el valor esperado de la respuesta por cada unidad (kg) de peso (Wt) que aumente. En este caso, cada kilo de aumento supone 635g de incremento en la masa muscular, para las mujeres que practican Voley Playa.
Sobre este modelo de referencia, se pueden calcular los modelos lineales para cada combinación de niveles de los factores:

Cuando cambiamos de deporte, las diferencias se centran en un cambio de nivel sin que afecte a la relación lineal (pendiente).

Sex=female & Sport=BBall: $LBM = 11.96 + 0.635Wt$
Sex=female & Sport=Row: $LBM = (11.96 + 0.243) + 0.635Wt$
Sex=female & Sport=Swim: $LBM = (11.96 + 1.848) + 0.635Wt$
Sex=female & Sport=T400m: $LBM = (11.96 + 1.919) + 0.635Wt$

Entre BBall y Row, no se han encontrado diferencias significativas en los modelos (test de Wald para el coeficiente SportRow: t-value=0.525, p-value=0.6). En cambio, Swim y T400 presentan un intercept significativamente superior que el de BBall. (test de Wald para el coeficiente SportSwim: t-value=3.432, p-value=0.00085 y test de Wald para el coeficiente SportT400m: t-value=3.108, p-value=0.002)

Cuando comparamos los modelos para ambos sexos, hay cambio en el nivel del intercept (test de Wald para el coeficiente Sexmale: t-value=-4.466, p-value=2e-5) pero también en la pendiente del modelo (test de Wald para el coeficiente Wt:Sexmale: t-value=7.75, p-value=5.89e-12): en el caso de los hombres deportistas, por cada kg de aumento supone un incremento de masa muscular del 887g:

Sex=male & Sport=BBall: $LBM = (11.96 - 10.75) + (0.635 + 0.252)Wt$
Sex=male & Sport=Row: $LBM = (11.96 - 10.75 + 0.243) + (0.635 + 0.252)Wt$
Sex=male & Sport=Swim: $LBM = (11.96 - 10.75 + 1.848) + (0.635 + 0.252)Wt$
Sex=male & Sport=T400m: $LBM = (11.96 - 10.75 + 1.919) + (0.635 + 0.252)Wt$

Para este modelo, se realizan los siguientes tests:

1) Linear hypothesis test

```
Hypothesis:
Wt:Sexmale = 0

Model 1: restricted model
Model 2: LBM ~ Wt + Sex + Sport + Wt:Sex

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    107 526.00
2    106 335.77  1    190.23 60.055 5.89e-12 ***
```

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

    2)  Linear hypothesis test

Hypothesis:
Sexmale  + Wt:Sexmale = 0

Model 1: restricted model
Model 2: LBM ~ Wt + Sex + Sport + Wt:Sex

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    107 397.64
2    106 335.77  1     61.87 19.532 2.398e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    3)  Linear hypothesis test

Hypothesis:
Wt  + Wt:Sexmale = 0

Model 1: restricted model
Model 2: LBM ~ Wt + Sex + Sport + Wt:Sex

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    107 4542.9
2    106  335.8  1    4207.1 1328.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

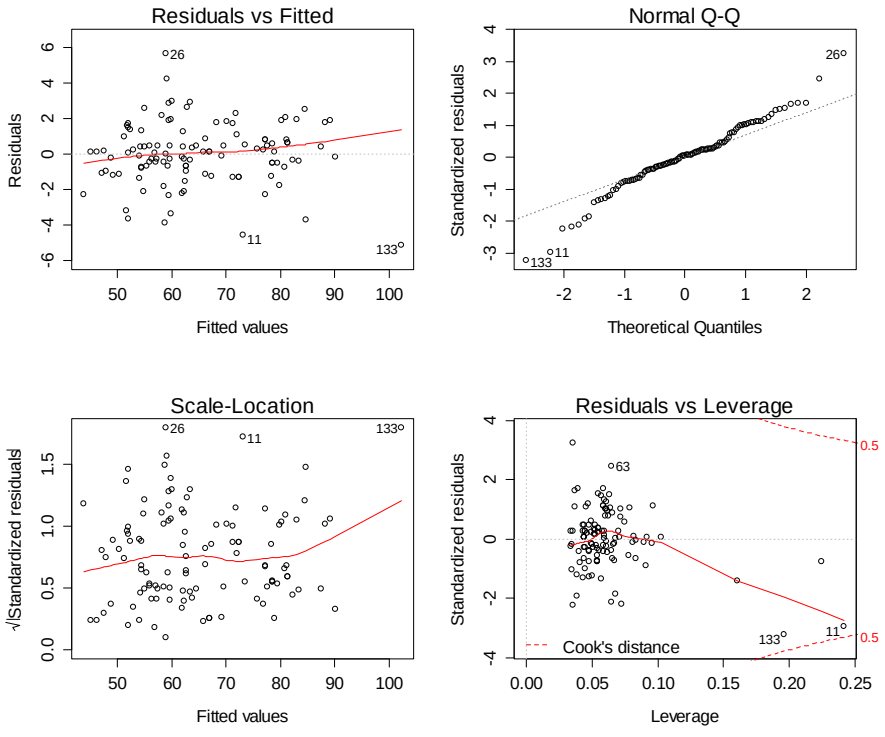
5) Si queremos establecer que <u>en el caso de los hombres</u>, el peso (Wt) está relacionado linealmente de forma significativa con la Masa Corporal (LBM), ¿Cuál de los anteriores es el estadístico de contraste y el p-valor que permite justificar tal afirmación?

La relación lineal se establece a partir de la significación de la pendiente del modelo lineal. Puesto que el contraste activo es de tipo baseline con la primera categoría de referencia, la estimación de la pendiente del modelo para los hombres se obtiene sumando al término linel (Wt) el valor del coeficiente estimado de la interacción de la variable Sex con la Covariable lineal (Wt:Sexmale). Para el modelo anterior, la estimación puntual de la pendiente de la recta para el caso de los hombres, en cualquier deporte ya que no incluye la interacción Wt:Sport, es:

Wt+Wt:Sexmale = 0.635 + 0.252 = 0.887

El test que establece la significación de este coeficiente es el test 3, y por lo tanto el estadístico de contraste del test F es 1328.2 y su p-valor es <2.2e-16

A continuación se incluyen los gráficos para validar el modelo anterior:



6) **Realiza la validación del modelo, comentando si se cumplen las premisas. Indica si son atípicos y/o datos influyentes las siguientes observaciones: 11, 26 y 133. Caracterízalos en base a su predicción, residuo, leverage y distancia de Cook. En particular, para la observación 133 deduce el sexo y el deporte que practica.**

Las premisas del modelo son : linealidad , homocedasticidad , normalidad e independencia .

El primer plot es el de los residuos frente las predicciones , permite ver si la disposición de los residuos es aleatoria alrededor del cero , sin que se observe ningún patrón que indicas desviaciones de la relación lineal. El ajuste local (línea roja) es prácticamente horizontal , confirmando en este caso no parece haber patrones de no linealidad . En este plot también se puede verificar descriptivamente si la varianza puede considerarse constante, frente a las predicciones. En este caso, no se observa incremento de la variabilidad de los residuos a medida que aumenta la predicción, indicando que se puede asumir homocedasticidad (el ligero incremento observado al final viene determinado por unos pocos puntos y puede suponer una pobre estimación de la varianza). También en este plot , aparecen etiquetadas las observaciones con residuos estandarizados superior a 2 (aprox) en valor absoluto (valores atípicos). En este caso no se obsevan indicios que cuestionen la linealidad y homocedasticidad de los residuos.

El segundo plot es el plot de normalidad, que permite determinar si podemos considerar que la distribución Normal es adecuada para los residuos. Si los puntos están alineados podemos asumir Normalidad de los residuos. Este plot permitiría ver patrones de asimetría o colas pesadas en los residuos que irían en contra de la hipótesis de normalidad. También se etiquetan los atípicos. En este caso, la disposición de los puntos no está del todo alineada y existen valores que se separan en las colas, sugiriendo que la normalidad de los residuos es dudosa, seguramente por la presencia de atípicos.

El tercer plot representa la raíz cuadrada de los valores absolutos de los residuos frente a las predicciones. Es un plot que permite determinar de forma más clara la presencia de heteroscedasticidad . El ajuste local mediante la recta no indica un claro incremento de los valores que constituyen una estimación de la varianza de los residuos, excepto al final debido a la presencia de un valor extremo. No es concluyente para confirmar la presencia de varianza no constante y además se ve influido por la presencia de atípicos que están relacionados con valores altos de las predicciones. La presencia de éstos últimos podría explicar el ligero incremento del ajuste local.

El cuarto plot permite identificar y caracterizar los datos influyentes. Representa los residuos estandarizados frente al factor de anclaje/apalancamiento (leverage). Además incluye curvas de nivel para indicar la distancia de Cook de las observaciones. Valores con una distancia de Cook alta pueden ser valores influyentes y se debe analizar su efecto en el ajuste del modelo. La distancia de Cook es una función creciente de los residuos al cuadrado y del leverage. Las observaciones que tienen un valor alto de la distancia de Cook aparecen etiquetadas (pueden ser por tener muy leverage , o tener un residuos alto en valor absoluto o una combinación de ambas situaciones no tan extremas). Las observaciones etiquetadas como influyentes parece que tienen un leverage alto ya la vez tienen un residuo de magnitud elevada. Habría que analizar qué efecto tienen en la estimación del modelo. Claramente, existen dos valores altamente influyentes (distancia de Cook próxima a 0.5) que poseen factores de apalancamiento de los más altos.

Caso 26: Su predicción está próxima a 60 pero el modelo explica mal esta observación ya que su residuo constituye un valor atípico (residuo estandarizado aprox. 3). Siendo positivo se interpreta que se observa una masa muscular (observado) muy superior a la que predice el modelo (esperado). Su influencia a priori es baja (factor de apalancamiento de los más bajos) y su distancia de Cook, siendo alta no es de las observaciones más influyentes a posteriori.

Caso 11: Predicción de unos 75 Kg, pero residuo atípico (próximo a -3) que indica que la masa muscular es menor de lo esperado de acuerdo a sus características. Es el valor con un leverage mayor y la combinación de residuo atípico y leverage alto da lugar a una distancia de Cook de las mayores, indicando que el dato es altamente influyente a posteriori.

Caso 133: Su predicción es la mayor, de más de 100kg de masa muscular. El residuo es atípico y su valor estandarizado es menor que -3. Sin ser el que posee el leverage más alto (es el tercer valor en leverage) vuelve a tener una distancia de Cook de las mayores y por lo tanto, también es influyente a posteriori. Teniendo en cuenta que su predicción es de unos 102kg de masa muscular y su residuo es de -6 aproximadamente, la masa muscular de este individuo es de unos 97kg. Observando los plots descriptivos de las rectas de regresión y la descriptiva por grupos, este valor corresponde al individuo de mayor peso, que es un hombre que juega al Voley Playa.

Ajustamos el siguiente modelo lineal general sin interacciones:

```
m3 <- lm(formula = LBM ~ Wt + Sex + SSF + Hg + Sport, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73873 -0.61041  0.06346  0.57223  2.16399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

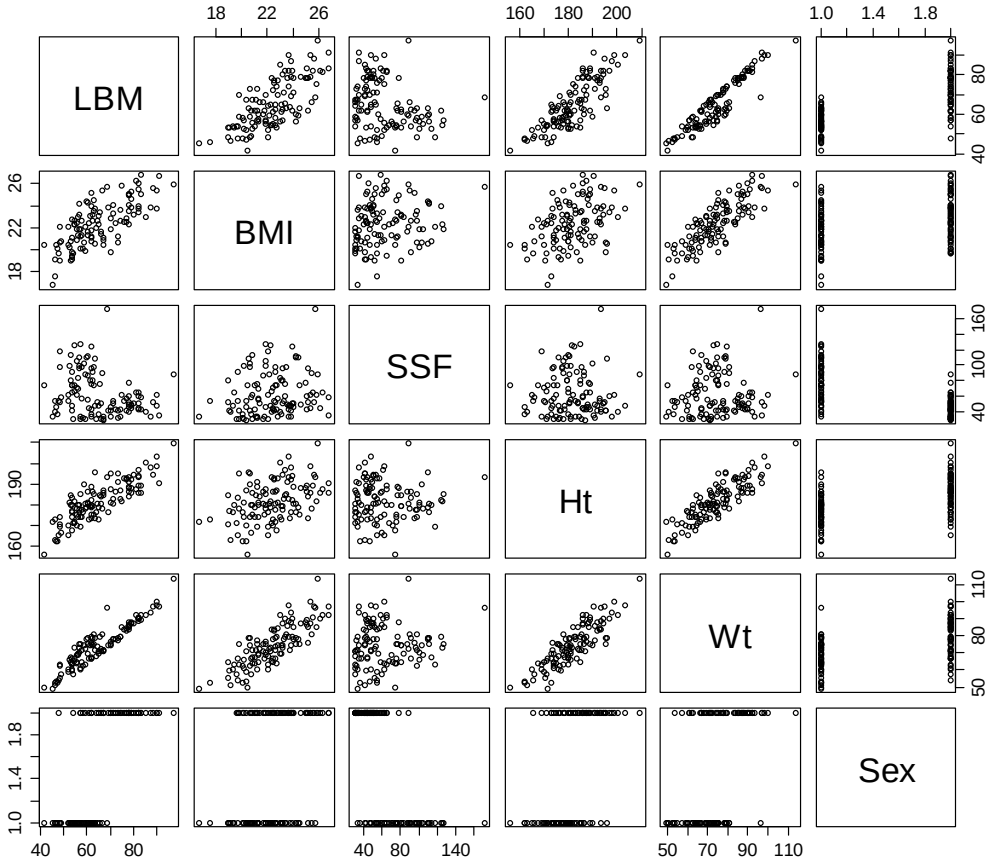
```
(Intercept)  7.96703    1.53864   5.178 1.09e-06 ***
Wt           0.89093    0.01093  81.497  < 2e-16 ***
Sexmale      2.37960    0.34822   6.834 5.60e-10 ***
SSF         -0.12880    0.00520 -24.771  < 2e-16 ***
Hg          -0.14575    0.10286  -1.417    0.159
SportRow    -0.28138    0.23061  -1.220    0.225
SportSwim    0.30394    0.27215   1.117    0.267
SportT400m  -0.45911    0.31539  -1.456    0.148
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8548 on 105 degrees of freedom
Multiple R-squared:  0.9954, Adjusted R-squared:  0.995
F-statistic:  3216 on 7 and 105 DF,  p-value: < 2.2e-16

> anova(m3)
Analysis of Variance Table

Response: LBM
           Df  Sum Sq Mean Sq    F value  Pr(>F)
Wt          1 14391.0 14391.0 19696.5026 < 2e-16 ***
Sex         1  1540.7  1540.7  2108.7269 < 2e-16 ***
SSF         1   508.8   508.8   696.4143 < 2e-16 ***
Hg          1     2.4     2.4     3.2401 0.07473 .
Sport       3     7.7     2.6     3.5085 0.01791 *
Residuals 105    76.7     0.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7) **Según el summary del modelo, los coeficientes asociados a la variable Sport son todos no significativos. Sin embargo, el test de variancias incrementales del modelo indica que la variable Sport es claramente significativa. Explica a qué se debe esta situación en este caso.**

La interpretación de los coeficientes para una variable categórica se hace en base al contraste activo, que en este caso es de tipo baseline y la categoría de referencia es la primera. En este caso, cada p-valor de la tabla summary establece que no hay diferencias entre los deportes considerados y el de referencia (BBall). Ahora bien, con estos valores no es posible determinar que existan otras dos categorías entre las que haya diferencias significativas. En este caso, el valor de la categoría de referencia es bastante central respecto al resto de categorías por lo que es razonable que no hayan diferencias entre ésta y el resto. Hay que tener en cuenta que la significación estadística no es transitiva y el hecho que entre BBall y Swim (coeficiente=0.30) no haya diferencias y que entre BBall y T400m (coeficiente=-0.46) tampoco las haya, no implica que entre Swim y T400m (coeficiente=0.30-(-0.46)=0.76) no haya diferencias significativas. El p-valor de la tabla ANOVA realiza los contrastes simultáneos y establece que existe diferencias entre algún par de categorías. Probablemente, las categorías citadas, que son las más extremas tengan diferencias significativas. Se podría comprobar con el método linearHypothesis.

A continuación ajustamos un modelo lineal general usando el mecanismo stepwise incluyendo interacciones de las categóricas con las numéricas y utilizando el criterio BIC. El modelo resultante es el siguiente:

```
m4 <- lm(formula = LBM ~ BMI + SSF + Ht + Wt + Sex + Wt:Sex, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92834 -0.45245  0.05298  0.54226  1.59871

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.741802  14.319146  -2.706  0.00795 **
BMI           1.081681   0.334741   3.231  0.00164 **
SSF          -0.112271   0.004931 -22.767  < 2e-16 ***
Ht            0.276194   0.081640   3.383  0.00101 **
Wt            0.463746   0.109644   4.230 4.99e-05 ***
Sexmale      -6.568938   1.317832  -4.985 2.43e-06 ***
Wt:Sexmale    0.130442   0.019460   6.703 1.02e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7539 on 106 degrees of freedom
Multiple R-squared:  0.9964, Adjusted R-squared:  0.9961
F-statistic:  4829 on 6 and 106 DF,  p-value: < 2.2e-16

> vif(m4)
       BMI        SSF         Ht         Wt        Sex     Wt:Sex
 92.552272   3.615238 120.349505 362.350245  86.270221 126.821486
```



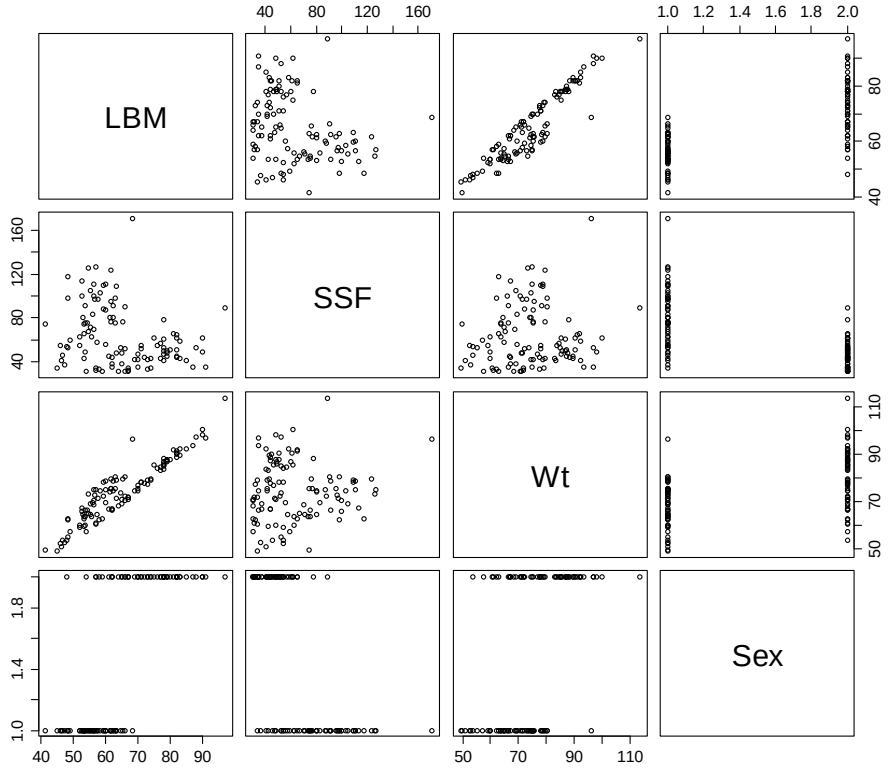
8) Suponiendo validado el modelo, ¿tiene sentido realizar la interpretación de los coeficientes obtenidos? Justifica la respuesta. En caso afirmativo, realiza la interpretación de los coeficientes.

Para realizar la interpretación de los coeficientes a partir únicamente del modelo, es preciso que los predictores sean lo más independientes posible, y por lo tanto, no debe existir multicolinealidad. Si interpretamos la variación en el valor esperado de la respuesta por cada aumento unitario de un predictor hay que tener en cuenta que si este predictor está correlacionado con otro, este otro también afecta a la respuesta.
En este caso hay claramente predictores altamente correlacionados positivamente (BMI, Altura y Peso). Todos ellos tienen un VIF muy alto indicando la presencia de multicolinealidad. Por ello, los coeficientes no son directamente interpretables, porque si se considera el efecto del valor esperado de la respuesta por cada kg unitario de incremento en el peso, también supondrá que aumente su altura y su BMI y se han de combinar los efectos de estos otros predictores. La solución es eliminar variables explicativas correlacionadas para obtener predictores lo más ortogonal posible y poder interpretar los coeficientes directamente.

En el modelo anterior, eliminamos las variables Indice de Masa Corporal (BMI) y altura (Ht). El modelo obtenido es el siguiente:

```
m5 <- lm(formula = LBM ~ SSF + Wt + Sex + Wt:Sex, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22732 -0.43298  0.01151  0.60754  1.62124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.229758   0.840559  10.980  < 2e-16 ***
SSF         -0.115502   0.005052 -22.861  < 2e-16 ***
Wt           0.824985   0.015511  53.186  < 2e-16 ***
Sexmale     -3.908192   1.091429  -3.581 0.000515 ***
Wt:Sexmale   0.091232   0.016099   5.667 1.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.787 on 108 degrees of freedom
Multiple R-squared:  0.996,  Adjusted R-squared:  0.9958
F-statistic:  6645 on 4 and 108 DF,  p-value: < 2.2e-16



> vif(m5)
     SSF       Wt      Sex   Wt:Sex
```

```
3.482399  6.654560 54.299177 79.645681
```



9) Suponiendo validado el modelo, ¿tiene sentido realizar la interpretación de los coeficientes obtenidos? Justifica la respuesta. En caso afirmativo, realiza la interpretación de los coeficientes.

En este caso el valor grande de los VIF corresponde a las variables Wt, Sex y su interacción. En realidad, la correlación introducida por la interacción es artificial y viene provocada por el diseño del modelo. Puesto que es habitual que a la hora de interpretar modelos con variables categóricas se realice la interpretación estratificando por las categoría de la variable, la presencia de VIFs altos asociados a interacciones no se asimilan a multicolinealidad y es posible interpretar el modelo.
En este caso, el modelo para una mujer sería:

$$\boldsymbol{LBM = 9.23 - 0.11SSF + 0.82 * Wt}$$

Por cada cm de incremento (=10mm) en la suma de los siete pliegues principales supone un decremento de 1.1kg de masa muscular y por cada kilo de peso más, la masa muscular aumenta en 820 gramos.

En el caso de un hombre, el modelo sería:

$$\boldsymbol{LBM = (9.23 - 3.91) - 0.11SSF + (0.82 + 0.09) * Wt}$$

La interpretación respecto a la variable SSF es la misma que en el anterior caso. Debido a la interacción, en el caso de un hombre, cada kg de aumento en el peso está relacionado con un incremento de 910g en la masa muscular.

10) El Índice de Masa Corporal (BMI) se calcula dividiendo el Peso en kg por la altura al cuadrado medida en metros. Para verificar usando un modelo lineal que el cálculo en esta base de datos se ha realizado de forma correcta ¿qué variables crearías, cómo definirías el modelo y que valores de los estimadores y del coeficiente de determinación esperarías obtener del modelo lineal ajustado?

La fórmula indica:

$$\boldsymbol{BMI = \frac{Wt}{(Ht/100)^2}}$$

Tomando logaritmos en la expresión para obtener una relación lineal exacta entre las variables:

$$\log(BMI) = \log(Wt) - 2\log(Ht) + 2\log(100)$$

$$\log(BMI) = 9.21 + \log(Wt) - 2\log(Ht)$$

Quiere decir, que si hacemos una regresión lineal usando como variable respuesta el logaritmo de BMI y predictores log(Wt) y log(Ht), el coeficiente de determinación ha de ser del 100%, el intercept ha de valer 9.21 y los coeficientes del modelo han de ser 1 y -2 respectivamente, siempre que el cálculo se haya realizado correctamente.

Como verificación, con estos datos:

```
> summary(lm(log(BMI)~log(Wt)+log(Ht),dades))

Call:
lm(formula = log(BMI) ~ log(Wt) + log(Ht), data = dades)

Residuals:
       Min         1Q     Median         3Q        Max
-2.655e-04 -1.077e-04  3.733e-06  1.072e-04  2.154e-04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.2095021  0.0019816    4648   <2e-16 ***
log(Wt)      0.9998451  0.0001527    6546   <2e-16 ***
log(Ht)     -1.9997086  0.0004854   -4120   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0001345 on 110 degrees of freedom
Multiple R-squared:      1,  Adjusted R-squared:      1
F-statistic: 2.64e+07 on 2 and 110 DF,  p-value: < 2.2e-16
```