Regresión Lineal Múltiple Taller # 2

Julián Camilo Riaño Moreno

22/3/2020

Actividad: desarrollo de taller # 2 (Regresión lineal múltiple)

En el archivo adjunto denominado atletas.xls se encuentran los datos de 98 mujeres atletas de élite que fueron entrenadas en el Instituto Australiano del Deporte. El objetivo del análisis de este conjunto de datos es explicar el comportamiento de la masa corporal magra (lbm) a partir de la estatura en centímetros (ht), el peso en kilogramos (wt) y el conteo de glóbulos rojos (rcc).

1. Haga un análisis descriptivo del conjunto de datos, es decir, construya diagramas de dispersión de cada variable explicativa con la variable respuesta. las correlaciones simples y parciales entre las variables.Comente.

Table 1: Coeficiente de correlación de Pearson para las variables dadas

	masa_c_magra_%	Estatura_cm	Peso_kg	ContGR_M_mm3
masa_c_magra_%	1.000	0.711	0.939	0.085
Estatura_cm	0.711	1.000	0.715	0.015
Peso_kg	0.939	0.715	1.000	0.021
ContGR_M_mm3	0.085	0.015	0.021	1.000

Se aplicó un coeficiente de correlación de Pearson dado que las variables dadas ('masa_c_magra_%', 'Estatura_cm', 'Peso_kg', 'ContGR_M_mm3') son cuantitativas continuas. Los resultados son mostrados en la Tabla 1. Desde estos resultados se puede afirmar que la variable regresora 'Peso_kg' es la que mayor correlación (0.939) tiene con la variable respuesta 'masa_c_magra_%', seguida de la variable 'Estatura_cm' (0.711). Incluso, es posible una mediana correlación entre ellas ('Peso_kg'y 'Estatura_cm' = 0.715). Por otra parte, la variable regresora 'ContGR_M_mm3', tiene escasa correlación tanto con, las variable respuesta como las otras variables regresoras.

Diagrama de dispersión y correlaciones simples y parciales

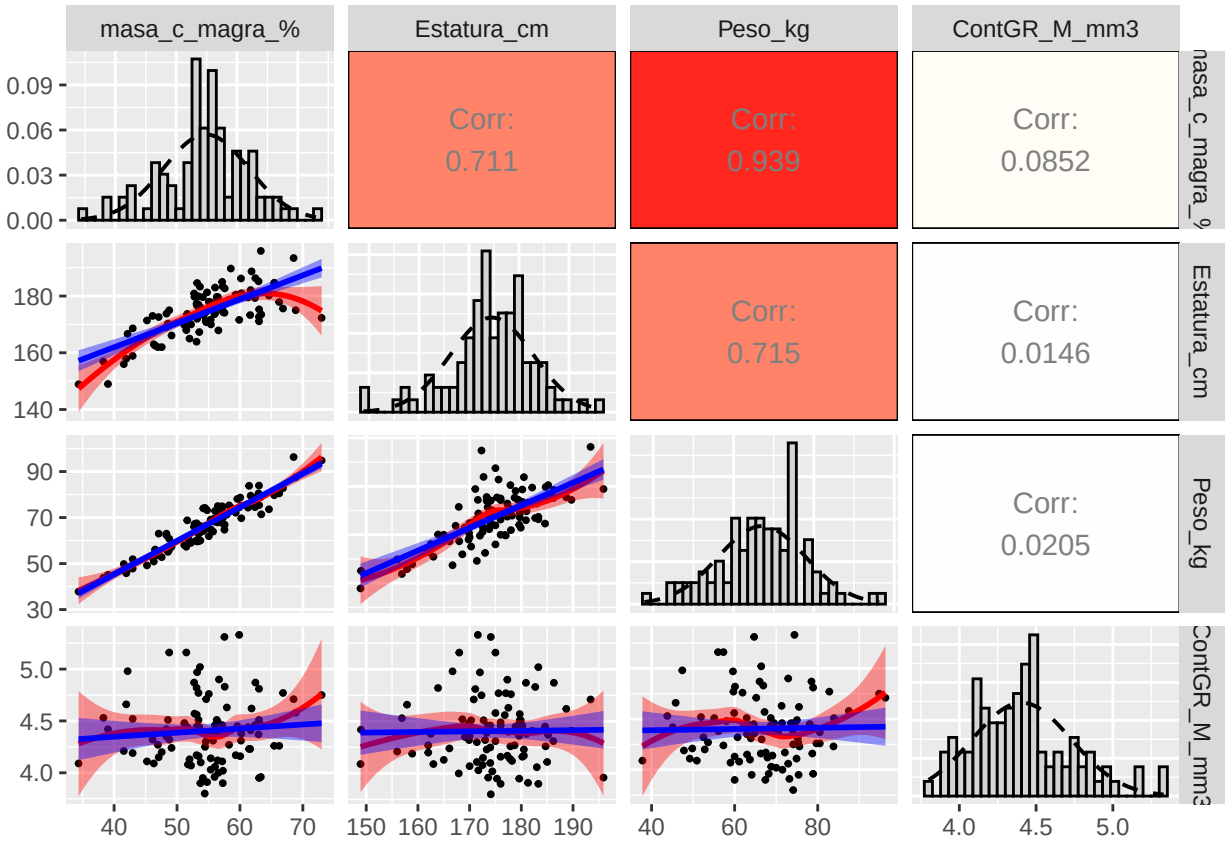

Figure 1: Gráfica de dispersión y correlaciones. En las cuadrículas inferiores a la izquierda muestran los gráficos de dispersion. La lineas azules corresponde al modelo de regresión lineal (método: `lm`). La linea roja corresponde a la correlación local método: `loess`. Las cuadriculas superiores a la derecha muestran los coeficientes de correlación entre las variables. Los colores corresponden a un mapa de calor que muestra en rojo la máxima correlación y en blanco la mínima correlación. En la cuadricual diagonal se muestran la distribución de las variables en gráfico de histograma y densidad

Se elaboró una gráfica con los diagramas de dispersion, distribuciones y correlaciones de y entre las variables dadas (Figura 1). Como se puede observar todas las variables tiene una distribución normal, sin embargo, la variable 'ContGR_M_mm3' presenta una distribución sesgada levemente a la derecha. Los diagramas dispersión mostrados muestran dos tipos de correlaciones, por una parte, las lineas rojas corresponde a una distribución local entre las variables y la linea azul corresponde a la linea regresora del modelo de regresión aplicado. Las lineas de correlación parcial y la correlación del modelo se sobre ponen entre la variable 'masa_c_magra_%' (respuesta) y la variable 'Peso_kg' (regresora), lo que soporta que la mayor correlación entre variables se da entre estas dos. Por otra parte, las gráficas muestras que la variable 'ContGR_M_mm3' es la que mayor dispersión presenta con respecto a las otros variables y las lineas correlación y regresión son menos consistentes y muestran mayores desviaciones (sobra roja).

2. Ajuste un modelo de regresión lineal múltiple que relacione la masa corpo- ral magra (lbm) con la estatura en centímetros (ht), el peso en kilogramos (wt) y el conteo de glóbulos rojos (rcc).

Análisis del modelo de regresión múltiple

Table 2: Estimaciones de parametros de la regresión

	Estimado	ErrorStand
β_0	-1.308	6.711
β_1	0.068	0.041
β_2	0.566	0.032
β_3	1.425	0.740

Los parametros de las tablas mostradas para el análisis del modelo serán mostradas en forma de β_i donde i corresponde a cada una de las variables dadas.

- β_0: corresponde a la variable respuesta ('masa_c_magra_%').
- β_1: corresponde a la variable regresora ('Estatura_cm').
- β_2: corresponde a la variable regresora ('Peso_kg').
- β_3: corresponde a la variable regresora ('ContGR_M_mm3').

Los resultados mostrados en la tabla 2 para las estimaciones del modelo, muestran un intercepto (β_0) con valor < 0, de manera que, *no tiene sentido práctico* realizar una interpretación de los resultados ya que solo son interpretables los resultados con $\beta_0 > 0$.

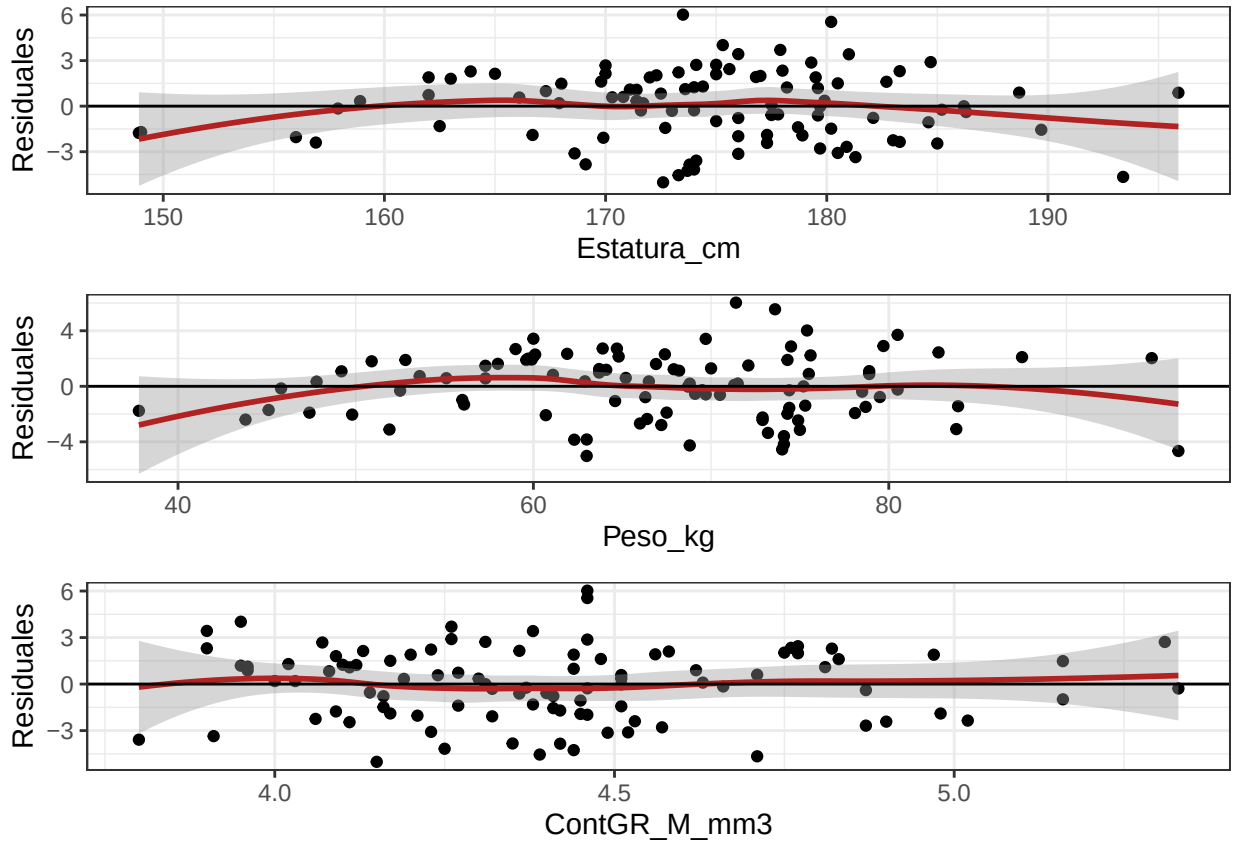

Figure 2: Gráfica de distribución de los residuos para cada variable regresora

Se realizó un diagrama de dispersion (figura 2) para los *"residuales"* del modelo con el objetivo de ver su distribución, en este caso los residuos deben de distribuirse aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje X. Como se observa la dispersión de los residuos es mayor para el β_3 ('ContGR_M_mm3') con lo cual se puede considerar que puede aportar menos en la regresión.

Table 3: $t - values$ para los parametros dados

	$t - value$	$p - value$
β_0	-0.195	0.84590
β_1	1.660	0.10026
β_2	17.945	0.00000
β_3	1.925	0.05723

A partir de la tabla 3 es posible deducir que los estimadores de variables regresoras β_1 ('Estatura_cm') y β_3 ('ContGR_M_mm3'), no son significativos ($p - value \geq 0.05$) en el modelo de regresión. No obstante, se reconoce la variable β_2 ('Peso_kg') como significativa ($p - value < 0.05$). Por lo tanto, en caso que se requiera una interpretación del modelo, sería posible asumir que existe una correlación lineal entre esta variable y la variable respuesta, donde se podría afirmar que un incremento en una unidad de el Peso (en Kg) se puede incrementar en 0.566 el % de masa corporal, siempre que las otras dos variables no cambien.

3. La masa corporal magra depende del peso del atleta?

Según el análisis realizado antes acerca de los $t-values$ y sus respectivos $p-values$ y en caso que se requiera una interpretación entendiendo que $\beta_0 < 0$, entonces, es posible afirmar que existe dependencia de la masa corporal magra del peso del atleta, en cuanto que β_2 (`Peso_kg`) estimado es 0.566 con un $t-value$ de 17.945 ($p - value < 0.05$).

4. La masa corporal magra depende de la estatura del atleta?

Según el análisis realizado antes acerca de los $t-values$ y sus respectivos $p-values$ y en caso que se requiera una interpretación entendiendo que $\beta_0 < 0$, entonces, es posible afirmar que NO existe dependencia de la masa corporal magra conteo dde la estatura del atleta, en cuanto que β_1 ('ContGR_M_mm3') estimado es 0.068 con un $t - value$ de 1.660 ($p - value \geq 0.05$) no significativo.

5. La masa corporal magra depende del conteo de glóbulos rojos del atleta?

Según el análisis realizado antes acerca de los $t - values$ y sus respectivos $p - values$ y en caso que se requiera una interpretación entendiendo que $\beta_0 < 0$, entonces, es posible afirmar que NO existe dependencia de la masa corporal magra del conteo de glóbulos rojos del atleta, en cuanto que β_3 ('ContGR_M_mm3') estimado es 1.425 con un $t - value$ de 1.925 ($p - value \geq 0.05$) no significativo.

6. Construya la tabla de análisis de varianza, y pruebe la significancia de la regresión.

Análisis de varianzas y significancia de la regresión (F_{global}: F de Fisher)

Table 4: Tabla ANOVA para suma de cuadrados de referencia

	Suma_cuadrados	Grados_de_libertad	Cuadrados_medios	F_{global}
Regresión	$SC_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$	k	$\frac{SC_{reg}}{k}$	...
Residuales	$SC_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k-1$	$\frac{SC_{res}}{n-k-1}$	...
Total	$Sc_t = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$	$n-1$	...	F_{fisher}

Table 5: Tabla ANOVA para suma de cuadrados obtenidos por el modelo

	Suma_cuadrados	Grados_de_libertad	Cuadrados_medios	F_{global}
Regresión	4218.151	3	1406.05	...
Residuales	523.244	94	5.566	...
Total	4741.395	97	...	252.595

Comparación F de Fisher y prueba de hipotesis

Table 6: Comparación entre estadistico F de Fisher manual y obtenido con la función `summary`

$F_{summary}$	F_{global}	Grados_libertad	$p-value$
252.59494028313	252.59494028313	94	$<$ 2.2e-16

Se realizo una tabla de varianza (ANOVA) de manera manual a través de las indicaciones dadas en clase (tabla 5) para la suma de cuadrados obtenidos de la regresión (SC_{reg}), los residuales (SC_{res}) y la suma de cuadrados total (SC_t) (los parametros pueden ser revisados en la tabla 6). A partir de estos resultados se puede encontrar que la ($SC_{reg} \approx SC_t$). Lo que sugiere que el modelo tiene *buen ajuste* y puede explicar la variabilidad de la variable respuesta β_0 identificada, sin tener en cuenta las variables regresoras.

Además se plantea una prueba de hipotesis global de la siguiente manera:

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \; para \; cualquier \; j \; dado$$

Para hacer la prueba de hipotesis se utilizó el estadístico F de Fisher con 94 grados de libertad. Los calculos se obtuvieron manualmente F_{global} y a través de la función `summary` ($F_{summary}$) Se compararon los resultados de f obtenidos donde $F_{global} = F_{summary}$ (ver tabla 6). Al obtener el $p-value$ se encuentra alta significancia estadística, con lo que se rechaza la hipotesis nula. De tal forma, se puede afirmar que almenos una de la variables regresoras (β_i) son diferentes de 0. Con los resultados provenientes del análisis de correlación y el modelo de regresión podemos afirmar que la variable regresora más probable es 'Peso_kg' (β_2).

7. Calcular R^2 y R^2_{adj} para este modelo. Interprete los resultados.

Table 7: R^2 y R^2_{adj} para el modelo

R^2	R^2_{adj}
0.8896435	0.8861215

La tabla 6. muestra los R^2 y R^2_{adj} para el modelo de regresión aplicado. Estos resultados ($R^2 \approx R^2_{adj}$), permiten proponer que aproximadamente el 88% de la variabilidad de la variable respuesta ('masa_c_magra_%') puede ser explicado por las variables regresoras en el modelo; de manera que, la variabilidad restante puede ser consecuencia a otras variables no tenidas encuenta en el modelo o por azar.

8. Determinar un intervalo de confíanza de 95% para los paráametros del modelo.

Table 8: Intervalos de confianza para los parametros de 95%

	IC 2.5%	IC 97.5%
β_0	-14.633	12.018
β_1	-0.013	0.150
β_2	0.504	0.629
β_3	-0.045	2.894

9. Un intervalo de confíanza de 95% para la masa corporal magra promedio de una atleta cuya estatura es 180 cms, con un peso de 78 kilogramos y un conteo de globulos rojos de 4.50 (millones de glóobulos rojos por microlitro de sangre).

Table 9: Predicción de intervalos de confianza para el caso dado

Ajuste	IC 2.5%	IC 97.5%
61.607	60.921	62.294

Anexo

Ejercicio realizado con dos variables excluyendo variable `ContGR_M_mm3`

Se realizó un análisis independiente del modelo excluyendo la variable que menor correlación (figura 1) presentó en el caso anterior (`ContGR_M_mm3`). Se planteó un modelo con dos variables regresoras y una variable respuesta ('masa_c_magra_%' = Variable respuesta; 'Estatura_cm' y 'Peso_kg' = Variables regresora) con la finalidad de probar el mejor modelo para el caso dado.

Table 10: Estimaciones de parametros de la regresión para dos variables regresoras

	Estimado	ErrorStand
β_0	4.913	5.965
β_1	0.068	0.042
β_2	0.567	0.032

Table 11: $t - values$ para los parametros dados

	$t - value$	$p - value$
β_0	0.824	0.41225
β_1	1.637	0.10504
β_2	17.724	0.00000

Al aplicar el modelo y como se evidencia en la tabla 10. el valor de β_0 en este caso es positivo > 0, de manera que se puede realizar una interpretación práctica de los resultados de las pendiente en este caso. No obstante, en la tabla 11. se evidencia que la variable 1 (`Estatura_cm` (β_1)) no es presenta signifancia estadistica para el modelo con un $p-value \geq 0$ (0.10504). De manera que la interpretación que se podría obtener proviene de la variables regresora 2 (`Peso_kg` (β_2)). El resultado de la pendiente es 0.567 con significancia $p-value < 0$, por lo tanto, es posible afirmar que el incremento de una unidad de peso en kg en los atletas incrementaría la masa muscular magra en 0.567 veces en el caso que las demás variables no cambiaran.

7. Calcular R^2 y R^2_{adj} para este modelo. Interprete los resultados. 0.8861215 0.8852922 0.8828773

Table 12: R^2 y R^2_{adj} para el modelo

R^2	R^2_{adj}
0.8852922	0.8828773

Al estimar de este modelo de dos variables regresoras, el R^2_{adj} (tabla 12), es menor que el R^2_{adj} obtenido en el modelo del caso propuesto con tres variables regresoras. De manera, que el primer modelamiento realizado tiene mayor probabilidad de explicar la variabilidad de la β_0 que el modelo de dos variables anexo. En este caso el R^2 no ajustado no puede servir como una medida para explicar la variabilidad dado que esté se estimador se puede ver afectado por el número de variables introducidas.

Para reconocer evaluar el mejor modelo de regresión para las variables dadas, se realizó una estratégia *stepwise* a través de la función `step` que utiliza un modelo matemático Akaike (AIC) para valorar los diferentes modelos. El resultado obtenido (ver abajo) evidencia que el mejor modelo en este caso es el que corresponde a uno de 3 variables regresoras que fue el primero realizado en esta actividad.

Modelo stepwise mixto para evaluación de modelo de regresión lineal múltiple.

```
## Start:  AIC=172.16
## `masa_c_magra_%` ~ Estatura_cm + Peso_kg + ContGR_M_mm3
##
##                Df Sum of Sq     RSS    AIC
## <none>                       523.24 172.16
## - Estatura_cm   1     15.34  538.58 172.99
## - ContGR_M_mm3  1     20.63  543.87 173.95
## - Peso_kg       1   1792.50 2315.74 315.93

##
## Call:
## lm(formula = `masa_c_magra_%` ~ Estatura_cm + Peso_kg + ContGR_M_mm3,
##     data = Athl)
##
```

```
## Coefficients:
##  (Intercept)   Estatura_cm       Peso_kg  ContGR_M_mm3
##     -1.30797       0.06848       0.56637       1.42480
```