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).

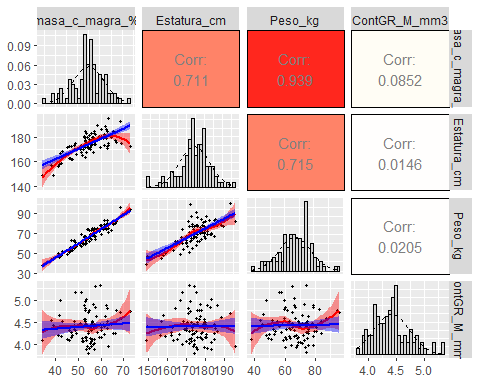
1. Haga un análisis descriptivo del conjunto de datos, es decir, construya diagramas de dispersión de cada variable explicativa con la variable respuesta. las correlaciones simples y parciales entre las variables.Comente.

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



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).

1. 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

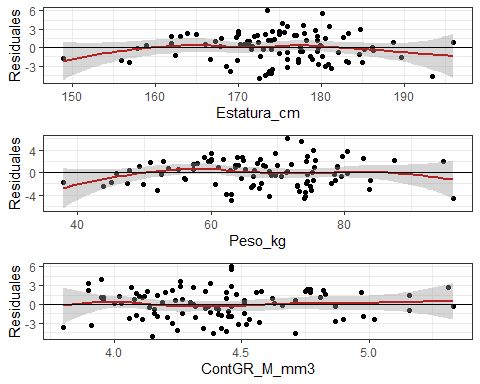
Estimaciones de parametros de la regresión

|  |  |  |
| --- | --- | --- |
|  | Estimado | ErrorStand |
|  | -1.308 | 6.711 |
|  | 0.068 | 0.041 |
|  | 0.566 | 0.032 |
|  | 1.425 | 0.740 |

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

* : corresponde a la variable respuesta (‘masa\_c\_magra\_%’).
* : corresponde a la variable regresora (‘Estatura\_cm’).
* : corresponde a la variable regresora (‘Peso\_kg’).
* : corresponde a la variable regresora (‘ContGR\_M\_mm3’).

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



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 (‘ContGR\_M\_mm3’) con lo cual se puede considerar que puede aportar menos en la regresión.

para los parametros dados

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | -0.195 | 0.84590 |
|  | 1.660 | 0.10026 |
|  | 17.945 | 0.00000 |
|  | 1.925 | 0.05723 |

A partir de la tabla 3 es posible deducir que los estimadores de variables regresoras (‘Estatura\_cm’) y (‘ContGR\_M\_mm3’), no son significativos () en el modelo de regresión. No obstante, se reconoce la variable (‘Peso\_kg’) como significativa (). 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.

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

Según el análisis realizado antes acerca de los y sus respectivos y en caso que se requiera una interpretación entendiendo que , entonces, es posible afirmar que existe dependencia de la masa corporal magra del peso del atleta, en cuanto que (‘Peso\_kg’) estimado es 0.566 con un de 17.945 ().

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

Según el análisis realizado antes acerca de los y sus respectivos y en caso que se requiera una interpretación entendiendo que , entonces, es posible afirmar que NO existe dependencia de la masa corporal magra conteo dde la estatura del atleta, en cuanto que (‘ContGR\_M\_mm3’) estimado es 0.068 con un de 1.660 () no significativo.

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

Según el análisis realizado antes acerca de los y sus respectivos y en caso que se requiera una interpretación entendiendo que , entonces, es posible afirmar que NO existe dependencia de la masa corporal magra del conteo de glóbulos rojos del atleta, en cuanto que (‘ContGR\_M\_mm3’) estimado es 1.425 con un de 1.925 () no significativo.

1. Construya la tabla de análisis de varianza, y pruebe la signifícancia de la regresión.

### Análisis de varianzas y significancia de la regresión (: de Fisher)

Tabla ANOVA para suma de cuadrados de referencia

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Suma\_cuadrados | Grados\_de\_libertad | Cuadrados\_medios |  |
| Regresión |  |  |  | … |
| Residuales |  |  |  | … |
| Total |  |  | … |  |

Tabla ANOVA para suma de cuadrados obtenidos por el modelo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Suma\_cuadrados | Grados\_de\_libertad | Cuadrados\_medios |  |
| Regresión | 4218.151 | 98 | 43.042 | … |
| Residuales | 523.244 | 94 | 5.566 | … |
| Total | 4741.395 | 97 | … | 252.595 |

### Comparación de Fisher y prueba de hipotesis

Comparación entre estadistico de Fisher manual y obtenido con la función summary

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Grados\_libertad |  |
| 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 (), los residuales () y la suma de cuadrados total () (los parametros pueden ser revisados en la tabla 6). A partir de estos resultados se puede encontrar que la (). Lo que sugiere que el modelo tiene *buen ajuste* y puede explicar la variabilidad de la variable respuesta identificada, sin tener en cuenta las variables regresoras.

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

$$
H\_o: \beta\_1 = \beta\_2 = ... =\beta\_k = 0 \\
$$

Para hacer la prueba de hipotesis se utilizó el estadístico de Fisher con 94 grados de libertad. Los calculos se obtuvieron manualmente $F\_{global$ y a través de la función summary () Se compararon los resultados de obtenidos donde (ver tabla 6). Al obtener el 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 () 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’ ().

1. Calcular y para este modelo. Interprete los resultados.

y para el modelo

|  |  |
| --- | --- |
|  |  |
| 0.8896435 | 0.8861215 |

La tabla 6. muestra los y para el modelo de regresión aplicado. Estos resultados (), 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.

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

Intervalos de confianza para los parametros de 95%

|  |  |  |
| --- | --- | --- |
|  | IC 2.5% | IC 97.5% |
|  | -14.633 | 12.018 |
|  | -0.013 | 0.150 |
|  | 0.504 | 0.629 |
|  | -0.045 | 2.894 |

1. 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).

Predicción de intervalos de confianza para el caso dado

|  |  |  |
| --- | --- | --- |
| Ajuste | IC 2.5% | IC 97.5% |
| 61.607 | 60.921 | 62.294 |