Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada II
Cali, febrero de 2018

Tarea 2:El modelo de regresión lineal simple

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

El objetivo del estudio es estimar el peso de una persona a partir de las otras variables (Altura, edad y sexo), para ello se proponen varios modelos lineales y se evaluaran con respecto a algunos estadísticos.

Para este trabajo, contamos con una base de datos de 99 personas, de las cuales, 12 eran mujeres y el resto (87) eran hombres. Nos pedian trabajar con una muestra de 24 personas, fijando las 12 mujeres, es decir, tenemos 87 hombres de los cuales debemos seleccionar 12 para completar la muestra. Con la ayuda del software R Core Team (2017) generamos 12 números aleatorios (todos los números tienen la misma probabilidad de salir) entre 1 y 87, donde cada número representa la posición en la base de datos del hombre seleccionado. Ya con nuestra base de datos conformada, procedimos a responder cada uno de los literales dados.

2. Punto 1:Modelo simple

El modelo ajustado para la variable 'Peso' con la variable de predicción 'Altura' fue el siguiente:

$$Peso = -99.0330 + 0.9778 Estatura$$

Para darnos una idea de que tan bueno es nuestro modelo, sin necesidad de evaluar a fondo cada uno de sus coeficientes, generamos una gráfica de dispersión entre las dos variables involucradas donde se observa la recta de regresión ajustada.

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

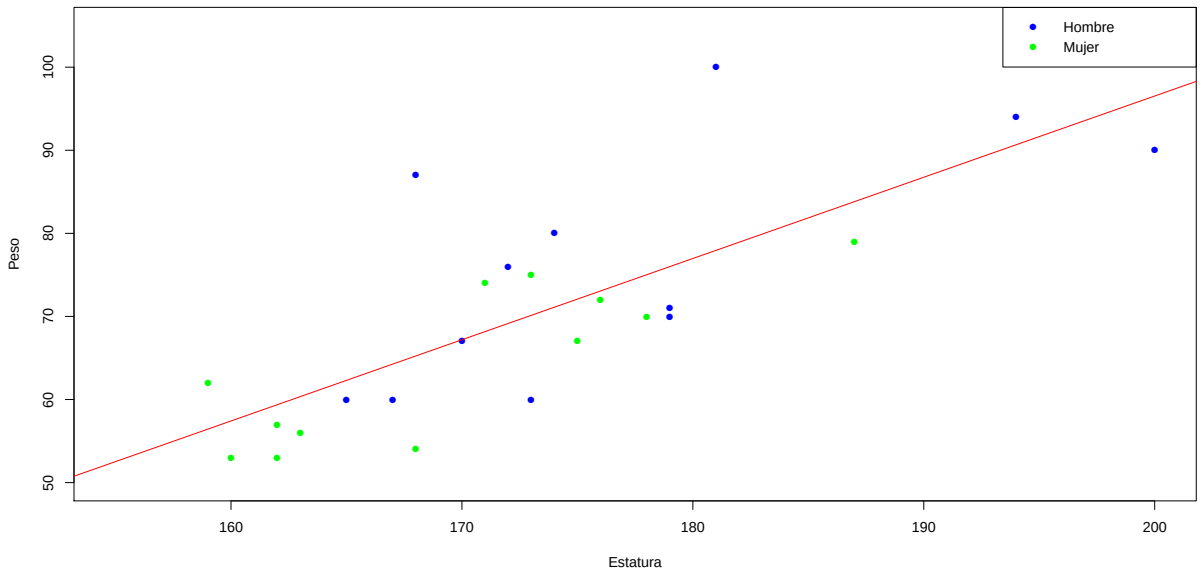

FIGURA 1: Gráfica de dispersión con la recta de regresión ajustada

En este gráfico podemos ver que nuestro modelo no es tan bueno a la hora de predecir el peso dando una estatura, ya que en algunos casos, el peso obtenido va a estar muy por encima o muy por debajo del peso real, es decir, los errores de nuestro modelo son aparentemente altos. Sin embargo, el modelo explica el comportamiento creciente de los puntos, por lo cuál observando solo esta gráfica, podríamos decir que el modelo ajustado, más específicamente la 'Altura' podría explicar aproximadamente entre el 40 y el 60 % de la variabilidad total de la variable 'Peso'.

3. Punto 2:Bondad del modelo e interpretaciones

Con respecto a las interpretaciones del modelo, se obtuvieron los valores p, de cada una de las pruebas de significancia para los coeficientes del modelo (β_0 y β_1), y se obtuvo el R^2.

- $\beta_0 = -99.0330$: Este valor del intercepto es para un mejor ajuste del modelo teniendo en cuenta que no tenemos alturas negativas nos ayuda a ajustar el peso en función de la altura, en pocas palabras, sirve de balanceo o de ajuste para una predicción mas acertada.
- $\beta_1 = 0.9778$: Cuando la variable 'Estatura', aumenta en una unidad (1 centimetro), se espera que el 'Peso' de la persona aumente en 0.9778 kg.
- $R^2 = 0.5754$: El 57.54 % de la variabilidad total de la variable Y:'Peso' es explicada por la variable X:'Estatura'
- p-valor $\beta_0 = 0.00425$: Como mi p-valor es menor que 0.05 rechazo mi hipótesis nula y digo que β_0 si es significante para el modelo
- p-valor $\beta_1 = 0.0000174$: Como mi p-valor es menor que 0.05 rechazo mi hipótesis nula y digo que β_1 si es significante para el modelo

Según las medidas calculadas para el modelo tomando como primer punto el R^2, podríamos decir que no

se ajusta muy bien ya que el valor de 0.5754 es aún muy bajo para las variables que queremos explicar. Algo curioso fue que el intercepto da un valor negativo, esto se puede deber a que como el modelo trata de ajustarse a los datos y tenemos datos un poco dispersos el intercepto trata de balancear esto, además no hay alturas menores que 150 cm ni pesos (Kgs) muy grandes por ende este ubica mi modelo en cierto punto para poder ajustarlo de manera correcta. Con respecto a los p-valores obtenidos podemos ver que en los dos casos nuestro p-valor es bastante bajo lo que es bueno, ya que basados en la hipótesis nula de que nuestros coeficientes son iguales a 0, con un valor tan bajo podemos rechazar con confianza esta hipótesis y decir que nuestros coeficientes son significativos para el modelo.

4. Punto 3:Intervalos de confianza para β_0 y β_1

Para generar los intervalos de confianza, utilizamos directamente la función confint en R, los intervalos generados fueron los siguientes:

- $\beta_{0\,95\,\%} = (-163.4559651; -34.60997)$: El verdadero valor de β_0 está entre -163.456 y -34.610 con una confianza del 95 %
- $\beta_{1\,95\,\%} = (0.6064059; 1.34922)$: El verdadero valor de β_1 está entre 0.6064 y 1.3492 con una confianza del 95 %

Podemos notar que el valor del intercepto real siempre viene siendo negativo, sin embargo el intervalo obtenido es demasiado amplio por lo que aproximarse al verdadero valor del parámetro es difícil. Ahora respecto a la pendiente podemos ver que es mayor que 0 y menor que 2, y este si tiene una amplitud mucho más reducida por lo que es más certera una estimación puntual de este.

5. Punto 4:Inclusión de la variable 'Sexo' en el modelo

Para incluir la variable sexo al modelo, generamos una variable indicadora (variable 'Dummy') que básicamente codifica la variable categórica (en nuestro caso es la variable 'Sexo') en términos binarios, la cual tomaba el valor 0 cuando es mujer y 1 cuando es hombre,además, es claro que la variable 'Altura', también depende del sexo de la persona (normalmente la media de la estatura de los hombres es mayor a la media de las mujeres), por lo cuál se tuvo en cuenta este cambio en la altura dependiendo del sexo, en pocas palabras, se tuvo en cuenta la interacción entre estas dos variables ('Altura' y 'Sexo'), el modelo ajustado incluyendo la variable 'Sexo' y teniendo en cuenta la interacción entre las variables 'Altura' y 'Sexo' fue el siguiente:

$$Peso = -92.8094 + 0.9271 Altura + 19.4939 Sexo - 0.0813(Altura \cdot Sexo)$$

6. Punto 5:Interpretación y comparación de modelos

Como lo mencionamos, se hizo una codificación de la variable sexo, donde vale 1 si es hombre y 0 si es mujer, por lo cual, podemos obtener dos modelos o mejor, dos interpretaciones, una para cada sexo:
Cuando el empleado es mujer (sexo=0), el modelo quedaría de la siguiente forma:

$$Peso = -92.8094 + 0.9271 Altura$$

donde β_0 es el intercepto, el cuál balancea o sirve para acercar más los valores ajustados a los reales, y $\beta_1 = 0.9271$ nos dice que por cada centímetro adicional de la altura de una mujer, su peso aumenta en 0.9271 kilogramos aproximadamente.
Ahora, cuando el empleado es hombre (sexo=1), el modelo quedaría de la siguiente forma:

$$Peso = -73.3155 + 0.8459 Altura$$

Podemos notar que el intercepto aumenta, y el coeficiente correspondiente a la variable 'Altura' (β_1) disminuye.
Este modelo nos dice que por cada centímetro adicional en la altura de un hombre, su peso aumentará aproximadamente en 0.8459 kilogramos.

Note que los dos modelos se desprenden del modelo general, solo se reemplazan los dos posibles valores en la variable sexo. interpretando los valores en el modelo general tendríamos lo siguiente:

- $\beta_1 = 0.9271$: por cada centímetro que aumente un empleado, sea mujer o hombre, su peso aumentara en 0.9271 kilogramos siempre y cuando todas las otras variables sean 0.
- $\beta_2 = 19.4939$: Si el empleado es hombre (Sexo=1), su peso aumentará en promedio 19.4939 kilogramos, suponiendo que todas las otras variables son 0. Si el empleado es mujer (Sexo=0), su peso no aumentará.
- $\beta_3 = -0.0813$: Con este coeficiente hay que tener cuidado en su interpretación. Primero, si es hombre (Sexo=1), al coeficiente original de la altura (0.9271) se le restara 0.0813, lo cuál daría 0.8458, es decir, si el sexo es masculino, la influencia de la altura en el Peso, disminuiría a 0.8458 kgms por centímetro adicional de altura.Si es mujer (Sexo=0), no generaría ningún cambio en el coeficiente de la altura.
  Resumiendo, el coeficiente β_3 mide el cambio en la influencia de la altura en el peso, por sexo.

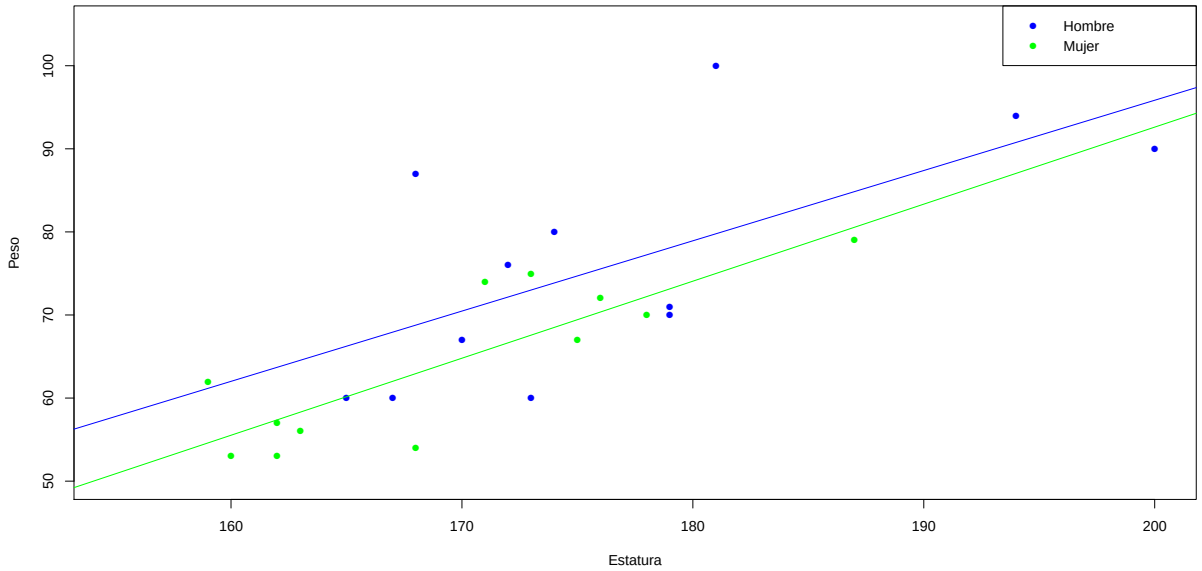

FIGURA 2: Gráfica de dispersión con la recta de regresión ajustada por sexo

Para comparar los modelos y decidir si uno es mejor que el otro, generamos la siguiente tabla, donde se comparan con respecto al $R^2_{ajustado}$ y al $CME = \sigma^2$

Tabla 1: Tabla comparativa entre los modelos ajustados

	Peso-Altura	Peso-Altura,Sexo
$R^2_{ajustado}$	0.5561	0.5574
$CME = \sigma^2$	77.33809	77.10571

Podemos apreciar que el R^2 y el $CME = \sigma^2$ no son muy diferentes entre los dos modelos sin embargo podemos decir que el modelo donde incluimos la variable sexo es un poco mejor que el otro, ya que se ve, aunque no en gran medida que el R^2 es mayor y el CME menor, lo cual indica que este explica un poco mejor el comportamiento de la variable.

7. Punto 6:Inclusión de la variable 'Edad' en el modelo

El modelo ajustado incluyendo la variable 'Edad' es el siguiente:

$$Peso = -108.63737 + 0.91223 Altura + 14.77551 Sexo + 0.97851 Edad - 0.05999(Altura \cdot Sexo)$$

Con respecto a la interpretación de los coeficientes de este modelo, son exactamente igual al modelo del punto 5 (teniendo en cuenta el cambio en los valores de los coeficientes, por el ingreso de la nueva variable), solo añadimos la variable edad, este coeficiente nos dice que mientras todas las otras variables estén constantes, por un año adicional en la edad del empleado, su peso aumentará en aproximadamente 0.97851 kilogramos.

Para comparar los tres modelos obtenidos, simplemente adicionamos este ultimo a la tabla comparativa, la cual quedo de la siguiente manera:

Tabla 2: Tabla comparativa entre los modelos ajustados

	Peso-Altura	Peso-Altura,Sexo	Peso-Altura,Sexo,Edad
$R^2_{ajustado}$	0.5561	0.5574	0.5572
$CME = \sigma^2$	77.33809	77.10571	77.14896

En la tabla anterior, tuvimos en cuenta el $R^2_{ajustado}$ y el $CME = \sigma^2$ para comparar los tres modelos generados. Podemos ver que el modelo con la Altura, el Sexo y la interacción entre ambas sigue siendo el mejor, aunque la diferencia no es muy grande. Esto nos dice que la variable Edad, no ayuda a explicar la variable de respuesta 'Peso'.
También, el R^2 es muy bajo, por lo cuál se deberían pensar en posible nuevas variables para explicar el peso de una persona, se nos ocurren variables como por ejemplo, horas dedicadas al ejercicio, cantidad de comida ingerida diaria,inclusive se podría pensar en la contextura de sus antepasados como variable predictiva. En pocas palabras, se deberían buscar nuevas variables para armar un buen modelo que permita predecir de manera adecuada el peso de los empleados.

Referencias

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/