Trabajo práctico 3

Fundamentos de análisis de datos

Luciana Gattuso

En el siguiente informe se analizaron los datos *data multiple regression exercice.cs*v que corresponden las medidas de circunferencias y diámetros corporales, edad y altura con el objetivo de estudiar cómo depende la variable weight (peso) en función de las otras variables. Los datos muestran 24 covariables que van a ser traducidas de la siguiente manera:

biacromial	biacromial
pelvic.breadth	ancho pélvico
bitrochanteric	bitrocantéreo
chest.depth	profundidad del pecho
chest.diam	diámetro del pecho
elbow.diam	diámetro del codo
wrist.diam	diámetro de la muñeca
knee.diam	diámetro de la rodilla
ankle.diam	diámetro del tobillo
shoulder.girth	circunferencia del hombro
chest.girth	circunferencia del pecho
waist.girth	circunferencia de la cintura
navel.girth	circunferencia del ombligo
hip.girth	circunferencia de cadera
thigh.girth	circunferencia de muslo
bicep.girth	circunferencia de bíceps
forearm.girth	circunferencia de antebrazo
knee.girth	circunferencia de rodilla
calf.girth	circunferencia de pantorrilla
ankle.girth	circunferencia de tobillo
wrist.girth	altura de la muñeca
age	edad
height	altura

En primer lugar, se separaró el conjunto de registros dos. El primero se utilizará para calcular los parámetros de cada modelo y el segundo para realizar predicciones y evaluar los resultados. Los datos de prueba se utilizaron para comprobar si el modelo generado a partir de los datos de entrenamiento es correcto. Asimismo, con el fin de generar resultados estadísticamente significativos y representativos del conjunto de datos, se seleccionó el 80% del volumen de datos a

entrenamiento y el 20% restante de prueba, conformando un conjunto de entrenamiento consta de 406 observaciones.

En segundo lugar, se realizó un modelo de regresión múltiple con todas las variables involucradas y la validación del modelo. Se utilizó una matriz de correlaciones para validar la relación de dependencia entre las variables para el conjunto de entrenamiento.

Figura 1

En la matriz de correlaciones no hay variables con relación inversa o negativa. Todas las variables tienen algún tipo de correlación positiva entre ellas, aunque sólo pocas una correlación significativa, por encima del 0.8.

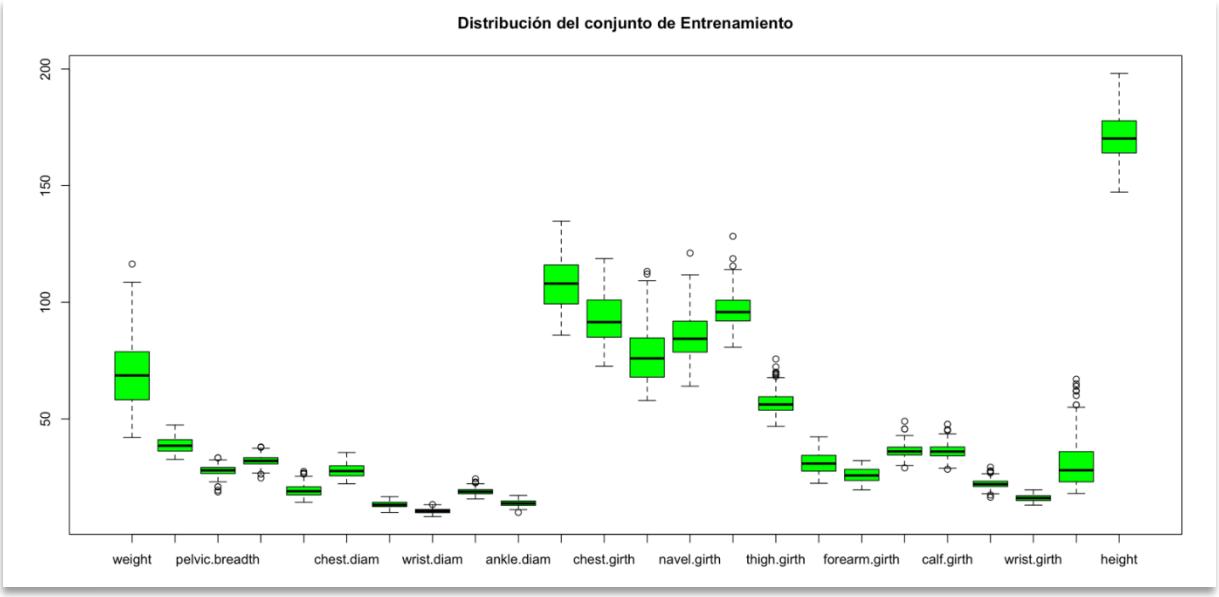
Por otro lado, se evaluó la correlación de cada variable con la variable de interés (Y).

Figura 2

	Variables	Corrrelación.con.Y
1	1.00	weight
2	0.90	chest.girth
3	0.90	waist.girth
4	0.88	shoulder.girth
5	0.88	forearm.girth
6	0.87	bicep.girth
7	0.83	chest.diam
8	0.83	wrist.girth
9	0.81	elbow.diam
10	0.81	knee.girth
11	0.80	chest.depth
12	0.78	knee.diam

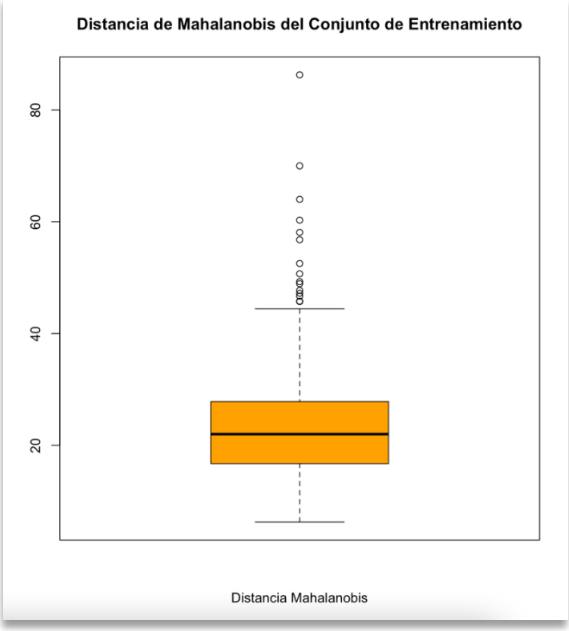
De las 24 variables, todas con correlación positiva, sólo 11 superan el 0.80. En la figura 2 se visualizan las correlaciones ordenadas de mayor a menor demostrando las circunferencias de pecho, cintura, hombro, antebrazo, bíceps, muñeca y rodilla y los diámetros de rodilla y pecho podrían afectar positivamente al peso de una persona, es decir a más presencia de la variable más peso. Asimismo, se graficó la distribución del conjunto de entrenamiento en un boxplot.

Figura 3



El gráfico demuestra una alta variabilidad entre las variables y la presencia de outliers. Esta conclusión resulta útil para el cálculo de componentes principales. Además, se utilizó la Distancia de Mahalanobis sobre el conjunto de entrenamiento.

Figura 4



En la figura 4 se visualizan los outliers del conjunto. De 406 registros 57 son outliers, es decir que representan un 14%. Es necesario tener este dato en cuenta para futuras conclusiones.
En segundo lugar, para implementar el modelo de regresión lineal múltiple con todas las variables se aplicó la función lm() de Rstudio (modelo 1).

Figura 5

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-119.70705	2.86159	-41.832	< 2e-16	***
biacromial	-0.03820	0.07600	-0.503	0.615522	
pelvic.breadth	0.06395	0.07631	0.838	0.402531	
bitrochanteric	-0.08353	0.10430	-0.801	0.423719	
chest.depth	0.29178	0.07864	3.710	0.000238	***
chest.diam	0.17759	0.09345	1.900	0.058126	.
elbow.diam	0.28267	0.21129	1.338	0.181743	
wrist.diam	0.25328	0.25656	0.987	0.324154	
knee.diam	0.39434	0.15583	2.531	0.011789	*
ankle.diam	-0.03058	0.17776	-0.172	0.863515	
shoulder.girth	0.07657	0.03496	2.190	0.029107	*
chest.girth	0.12663	0.04348	2.912	0.003798	**
waist.girth	0.32265	0.02859	11.283	< 2e-16	***
navel.girth	0.02596	0.02734	0.950	0.342919	
hip.girth	0.28371	0.05364	5.289	2.07e-07	***
thigh.girth	0.20748	0.06232	3.330	0.000955	***
bicep.girth	0.07495	0.09535	0.786	0.432289	
forearm.girth	0.33038	0.15742	2.099	0.036498	*
knee.girth	0.26958	0.08900	3.029	0.002620	**
calf.girth	0.38263	0.07917	4.833	1.95e-06	***
ankle.girth	0.06442	0.11307	0.570	0.569176	
wrist.girth	-0.23805	0.23740	-1.003	0.316615	
age	-0.06177	0.01359	-4.544	7.42e-06	***
height	0.29585	0.02026	14.603	< 2e-16	***

El resultado del modelo se evaluó la columna Pr (>t) que indica el p valor. Un p valor menor a 0.05 refleja una baja probabilidad de observar una relación entre las variables de predicción (peso) y respuesta. Sólo 12 de las 24 variables cuentan con p valores menores que 0.05: la profundidad y

diámetro del pecho; diámetro de rodilla, hombro, pecho, cadera, cintura, muslo, antebrazo y pantorrilla; edad y altura. Este grupo de variables podrían explicar la variable de interés, peso. Bajo el supuesto de estas 12 variables no se rechazaría la hipótesis alternativa concluyendo una relación entre ellas y el peso. Sin embargo, como las otras 12 variables dan un resultado opuesto, es decir que rechazarían la hipótesis nula se podría pensar en un modelo que las descarte para obtener predicciones más certeras.

El modelo calculó una desviación estándar residual de 2.12, un R cuadrado de 0.97 y un R cuadrado ajustado de 0.97. Ambos valores cercanos a uno indican que el modelo de regresión explica la varianza observada en la variable de respuesta. Si bien se podría R cuadrado se ve influencia por la cantidad de variables utilizadas en el modelo (24) es necesario resaltar que el R cuadrado ajustado da exactamente el mismo valor.

El p valor del estadístico F (2.2e-16) es menor a 0.05 reforzando la idea de que se podría rechazar la hipótesis nula. En base a este indicador, se podría concluir que utilizando todas las variables del modelo la relación con la variable de interés (peso) es ausente.

En segundo lugar, se utilizó el método de selección de variables Mixed Selection a partir del paquete Olsrr de Rstud El método seleccionó 12 variables de las 24 iniciales del conjunto de entrenamiento. Es importante resaltar estas coinciden con aquellas variables de p valores menores a 0.05 del modelo anterior. Este segundo modelo utiliza la profundidad de pecho, diámetro de rodilla, circunferencia del hombro, del pecho, cadera, cintura, muslo, antebrazo, rodilla, pantorrilla, altura y edad.

El método de selección de variables calculó un R cuadrado de 0.97 y un R cuadrado ajustado de 0.97, el mismo valor que el modelo anterior. Un valor alto en ambos modelos podría inferir un alto grado de efectividad de las variables independientes para explicar la variable dependiente. El coeficiente de variación de 3.07 y un error cuadrático medio de 4.50 indicador que se utilizará para la comparación de modelos.

El error estándar residual para este modelo es de 2.1 mientras que en el modelo 1 es 706.1. Ya que este indicador proporciona la desviación estándar de los residuales y refleja la predicción en los datos de entrenamiento, se podría decir que un valor menor de error estándar residual refleja un modelo más acertado. Más adelante se volverá a analizar este tema en gráficos de dispersión e histogramas.

Figura 6

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -121.64003    2.64662 -45.961  < 2e-16 ***
waist.girth     0.32943    0.02687  12.262  < 2e-16 ***
knee.girth      0.27072    0.08424   3.214 0.001418 **
height          0.30447    0.01702  17.891  < 2e-16 ***
thigh.girth     0.20547    0.05487   3.745 0.000207 ***
chest.girth     0.17607    0.03867   4.554 7.04e-06 ***
calf.girth      0.39345    0.06964   5.650 3.09e-08 ***
hip.girth       0.30225    0.04133   7.314 1.47e-12 ***
forearm.girth   0.40037    0.10448   3.832 0.000148 ***
age            -0.05571    0.01277  -4.362 1.65e-05 ***
knee.diam       0.47134    0.13486   3.495 0.000528 ***
chest.depth     0.27987    0.07673   3.648 0.000300 ***
shoulder.girth  0.08358    0.03096   2.700 0.007237 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.123 on 393 degrees of freedom
Multiple R-squared:  0.9764,    Adjusted R-squared:  0.9757
F-statistic:  1354 on 12 and 393 DF,  p-value: < 2.2e-16
```

Según la figura 6 todas las variables del Modelo 2 presentan un p valor menor a 0.05 indicando significancia. Si el valor p es menor que 0.05, entonces esa variable es significativa con al menos el 95% de confianza y se podría no rechazar la hipótesis alternativa de que existe una relación lineal entre las 12 variables y el peso de una persona. Respecto de los t valores sirven para juzgar las hipótesis nulas. En este caso, los valores son significativamente diferentes a cero y por eso, se podría pensar en una relación entre las variables.

En tercer lugar, debido a la gran variabilidad de escala entre las variables reflejadas en la figura 3 resulta recomendable utilizar un modelo de regresión múltiple tomando las componentes principales del conjunto de variables explicativas. Por un lado, se reducen la cantidad de variables a analizar sin perder información y, se estandarizan entre ellas, evitando el sobredimensionamiento de una variable. Para esto se utilizó el comando pcr() del paquete pls.

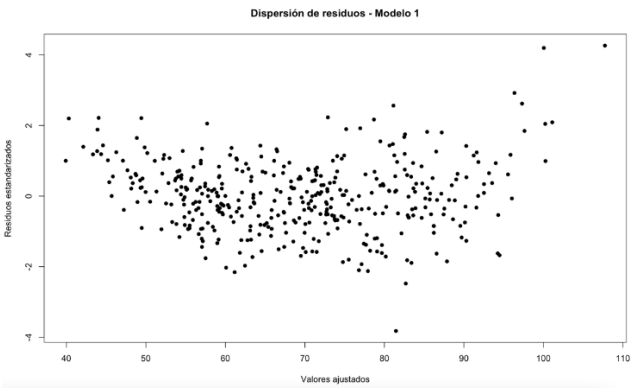
Este modelo lineal (modelo 3) construido en función de las componentes principales da como resultado que con 4 componentes es suficiente para explicar el 95% de los datos según la carga de la figura 7.

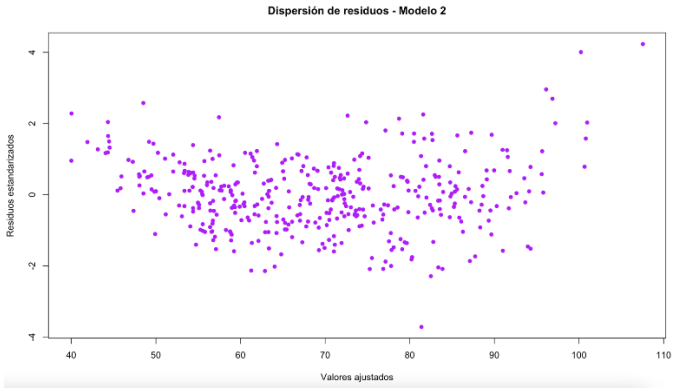
Figura 7

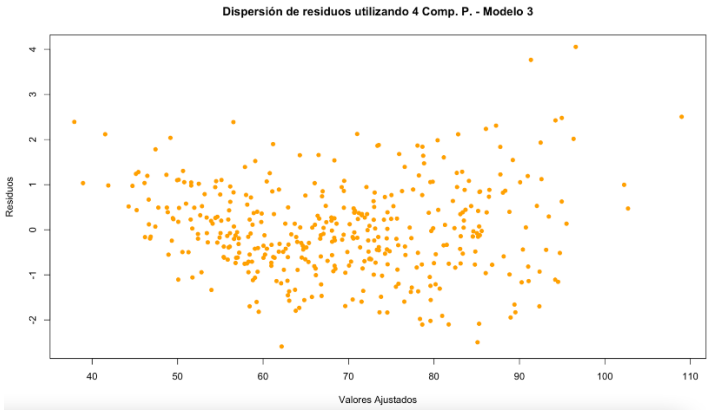
```
Data:   X dimension: 406 23
        Y dimension: 406 1
Fit method: svdpc
Number of components considered: 23
TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X         62.13    72.57    78.01    82.72    85.62    87.38    89.08    90.49    91.78
weight    94.05    94.61    94.86    95.26    96.11    96.15    96.32    96.49    96.72
        10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps
X          92.92     94.03     94.97     95.79     96.52     97.19     97.78     98.28
weight     97.01     97.01     97.03     97.11     97.13     97.22     97.45     97.49
        18 comps  19 comps  20 comps  21 comps  22 comps  23 comps
X          98.70     99.07     99.41     99.65     99.85     100.0
weight     97.56     97.63     97.70     97.70     97.70      97.7
>
```

Por último, utilizando todos los modelos se realizó una comparación entre ellos para seleccionar el modelo más acertado. La dispersión de los residuos de los modelos se grafica en función de los residuos estandarizados y los valores ajustados.

Figuras 8, 9 y 10

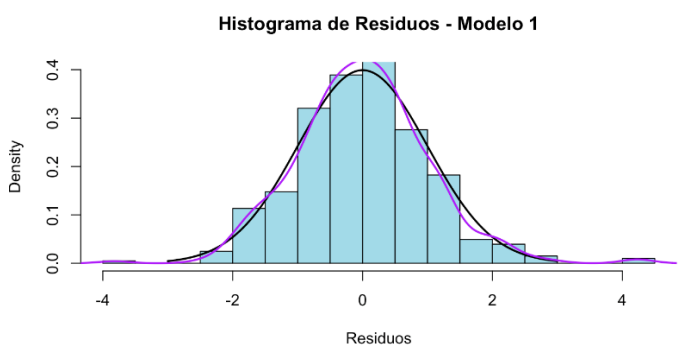


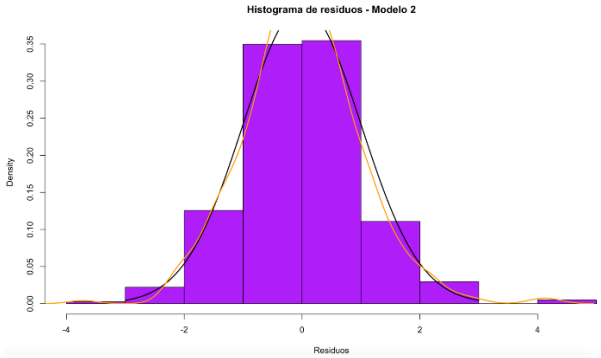


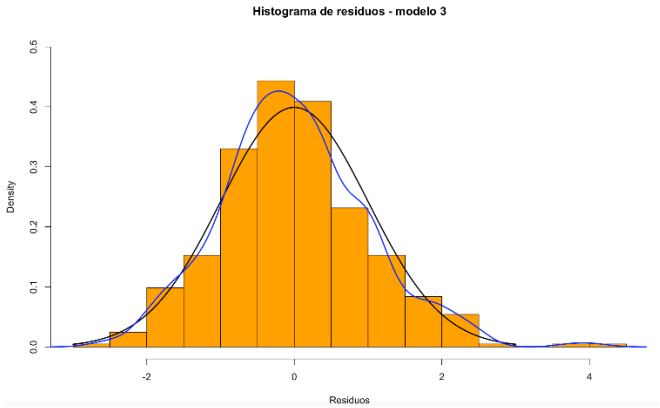


Los modelos 1 y 2 están más agrupados al centro o cero que el modelo 3, este cuenta con registros más dispersos. El modelo 3 no sigue un patrón de distribución de datos, a simple vista no refleja una asociación entre las variables, sino que están esparcidos de manera independiente y se visualizan algunos outliers. Para conclusiones más significativas en este tipo de gráficos se podría plantear a futuro eliminar los valores outliers del conjunto de entrenamiento.

Figuras 11, 12 y 13

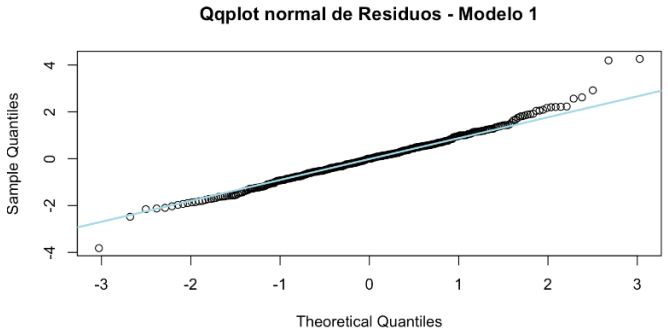


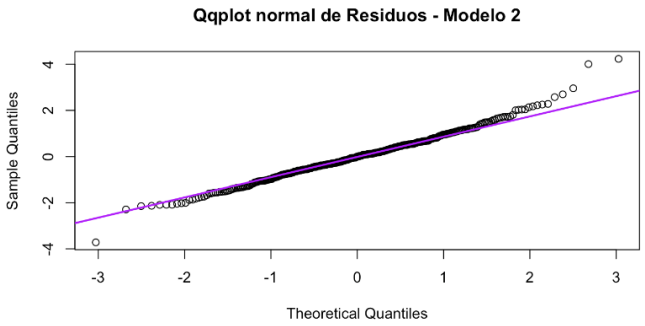


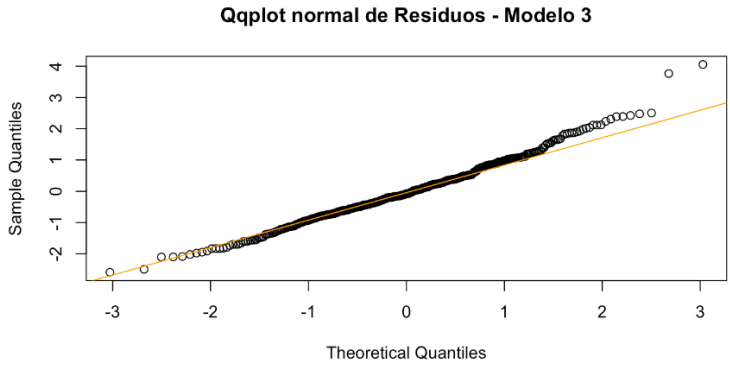


Teniendo en cuenta que uno de los supuestos para la prueba de hipótesis es que los errores y los residuos sigan una distribución gaussiana, todos los histogramas 3 están centrados al cero, pero el modelo 1 y 2 tienen una distribución más uniforme de los datos. En el modelo 3 se visualizan valores aislados en el punto 4, es decir que el mínimo y el máximo no tienen una magnitud similar.

Figuras 14, 15 y 16







En el gráfico de Q-Q (quantile-quantile), utilizado para comparar las distribuciones de probabilidad en bases a sus cuartiles, se visualizan una distribución prácticamente igual en los 3 modelos. Si bien los registros están en su mayoría alineados a la recta, cuentan con curvaturas en las colas que podría indicar una asimetría.

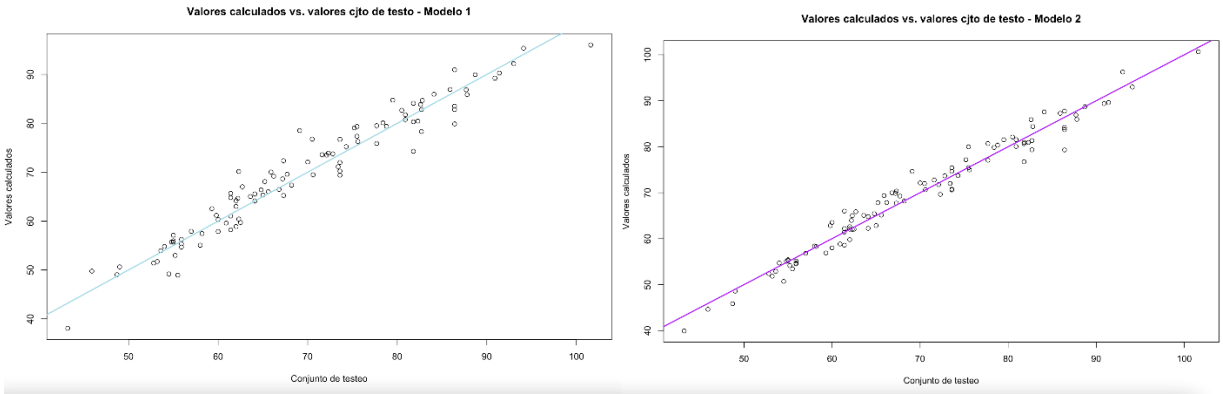
Por último, también se compararon los tres modelos utilizando el comando predict() sobre el segundo conjunto de datos data.body2 y se calcularon los errores cuadráticos medios y absolutos como se visualizan en la siguiente tabla.

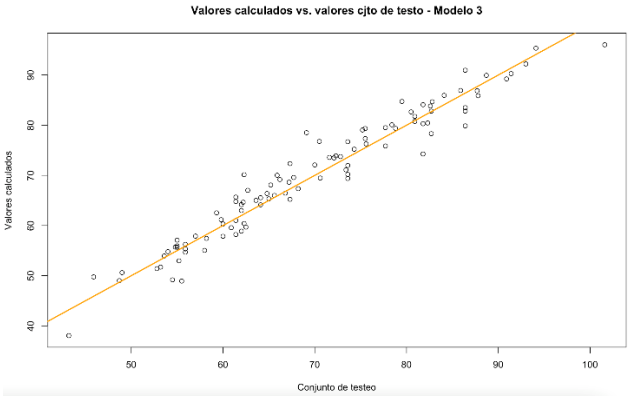
Tabla 1

Modelo 1 – R.L.M.		Modelo 2 – S. M.		Modelo 3 – C.P.	
EC Absoluto	EC Medio	EC Absoluto	EC Medio	EC Absoluto	EC Medio
465.5569	4.609475	474.37	4.696762	883.1689	8.744247

Para ambos indicadores, el modelo 1 es el que tiene menor error con el valor más cercano al cero. El segundo modelo cuenta con un error muy cercano al primero.

Figuras 17, 18 y 19





Se graficaron los valores calculados por la función predict () en función de los valores reales del conjunto de testeo y la línea de regresión estimada. Se puede visualizar que el modelo 2 cuenta con datos más cerca de la línea de regresión, esto indica que el modelo de regresión ajusta mejor los datos que para los modelos 1 y 3.

A modo de conclusión, para inferir que existe una relación de tipo lineal de la variable Y en función de las 24 variables, se puede utilizar el alto valor de R cuadrado calculado en el modelo 1 y 2. Si bien ambos modelos presentan el mismo R cuadrado, el modelo 1 es el más indicado para este conjunto de datos por tener el error cuadrático medio más bajo. Respecto de la distribución de los residuos estos siguen una distribución cercana a la normal.

Para futuros informes resulta interesante explorar la idea de trabajar sólo con el conjunto de variables, seleccionadas por el método Mixed Selection, que impactan positivamente sobre la variable de interés, peso. Así como probar un modelo recortando los outliers mencionados en la distancia de Mahalanobis.