Análisis Estadístico sobre el IMC

Luis Eduardo Cisneros Valdez

15 de mayo de 2018

Introducción

En este proyecto se evaluaron las bases de datos de la Encuesta Nacional de Salud y Nutrición para modelar el IMC cual es un indicador del peso de una persona en relación con su altura, sin embargo, se consideraron factores que atañen directamente la exposición de la población a la obesidad lo cual afecta directamente el valor del IMC y por tanto la calidad de la información que éste arroja. Para el estudio se consideraron las siguientes variables:

- IMC: El índice de masa corporal (kg/m2).
- Sexo: 1=mujer, 0=hombre.
- Edad: Edad en años.
- Horas_suenio: Horas de sueño promedio.
- Cintura: Cirfunferencia de la cintura.
- Tam_pob: Densidad poblacional
- Act_fis_vig: Minutos por semana de actividad física vigorosa.
- Act_fis_mod: Minutos por semana de actividad física moderada
- Act_fis_cam: Minutos por semana de caminata.
- Act_fis: Minutos de actividad física total.
- Tiempo_pantalla: Tiempo en pantalla en minutos.

La causa de la obesidad se remonta a que la cantidad de energía ingerida es menor que la que se consume, sin embargo, se pretende probar qué variables influyen en mayor proporción al aumento de la masa corporal.

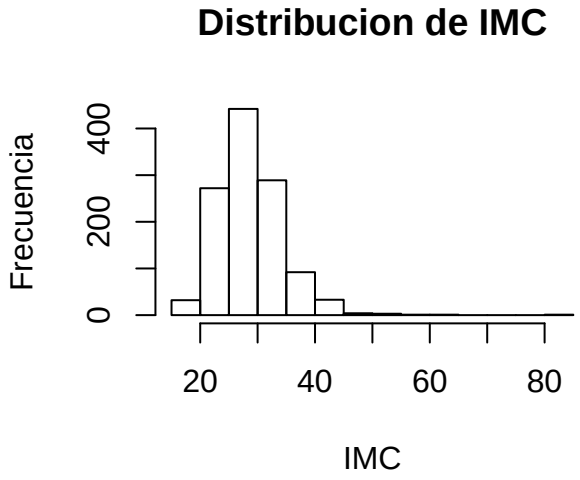
I: Análisis exploratorio

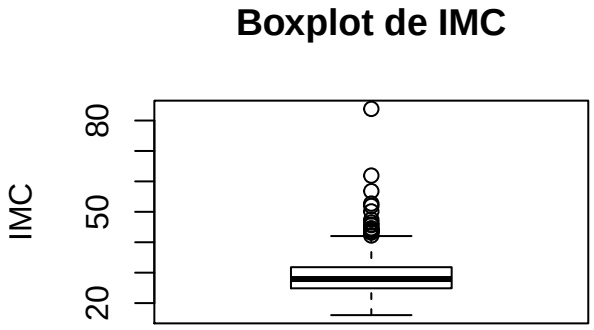
Comenzamos cargando los datos de la base de datos. Incialmente se cargan los datos por separado, luego se mezclan, se cambia el nombre de columnas y utilizamos 'datos' como nuestra nueva fuente de los datos a usar. Los datos de fumador no son suficientes y no serán considerados en el proyecto. La variable cintura es puesta como valores numéricos para que R no lo tome como muchos niveles distintas. Por último cambiamos la escala de los datos de actividad física, para tener en promedio cuántos minutos al día se tienen para cada actividad.

Visualizamos los datos limpiados con ayuda de summary, un boxplot y un histograma. Iniciamos con una muestra de tamaño N = 1170.

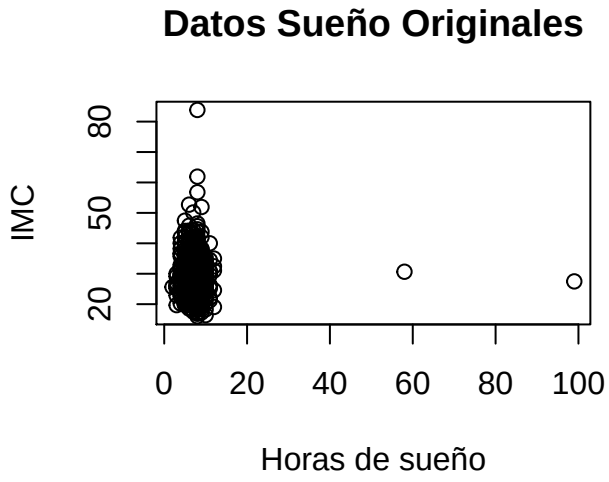
```
##       imc             sexo             edad        horas_suenio
##  Min.   :16.05   Min.   :0.0000   Min.   :20.00   Min.   : 2.000
##  1st Qu.:24.87   1st Qu.:0.0000   1st Qu.:34.00   1st Qu.: 6.000
##  Median :27.92   Median :1.0000   Median :44.00   Median : 8.000
##  Mean   :28.72   Mean   :0.6017   Mean   :45.84   Mean   : 7.508
##  3rd Qu.:31.81   3rd Qu.:1.0000   3rd Qu.:57.00   3rd Qu.: 8.000
##  Max.   :83.83   Max.   :1.0000   Max.   :95.00   Max.   :99.000
##     cintura         tam_pob     act_fis_vig      act_fis_mod
##  Min.   : 26.30   Rural :561   Min.   :  0.00   Min.   :  0.0
##  1st Qu.: 86.00   Urbano:609   1st Qu.: 42.89   1st Qu.:  0.0
##  Median : 94.00                Median :128.57   Median : 40.0
##  Mean   : 95.62                Mean   :142.50   Mean   : 66.1
```

```
##  3rd Qu.:102.10                3rd Qu.:207.86   3rd Qu.:120.0
##  Max.   :222.20                Max.   :540.00   Max.   :180.0
##   act_fis_cam         act_fis       tiempo_pantalla
##  Min.   :  0.000   Min.   :  0.00   <=840min:679
##  1st Qu.:  8.571   1st Qu.: 42.89   840+min :491
##  Median : 21.429   Median :128.57
##  Mean   : 45.719   Mean   :142.50
##  3rd Qu.: 60.000   3rd Qu.:207.86
##  Max.   :180.000   Max.   :540.00
```

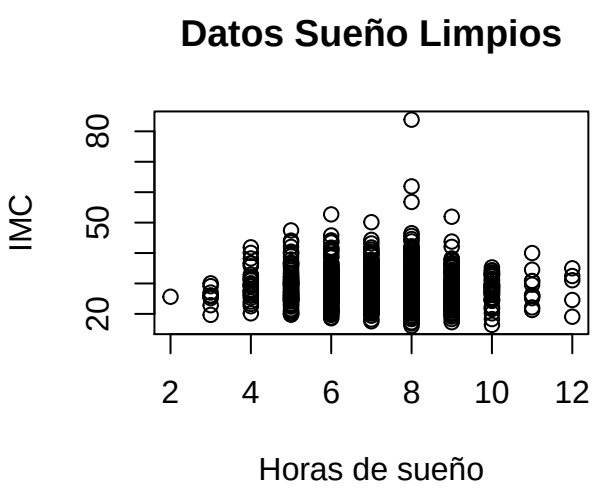




Al graficar los datos de hora de sueño notamos que dos valores (58 y 99) no son posibles.



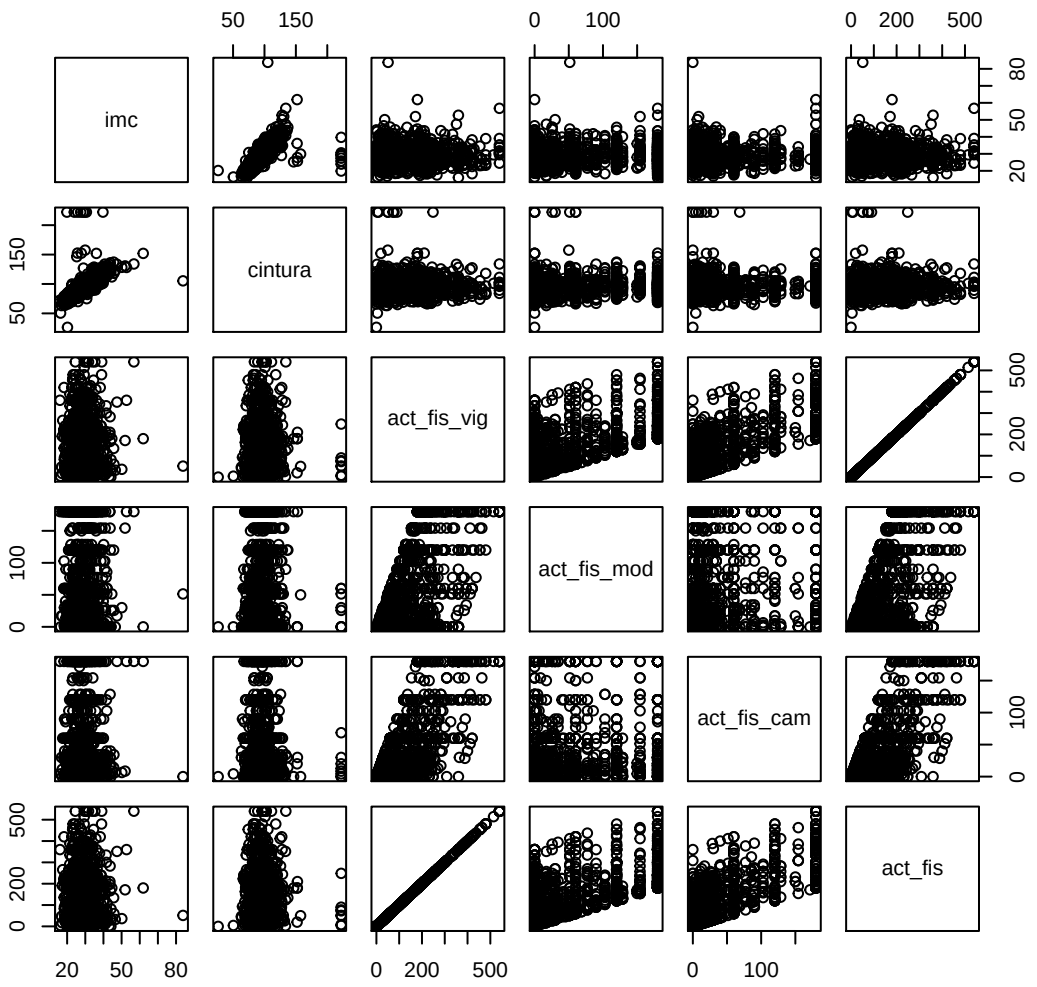
Eliminamos los dos valores que no son válidos(58 y 99)



Al eliminar estos dos datos tenemos una muestra de tamaño N = 1168.

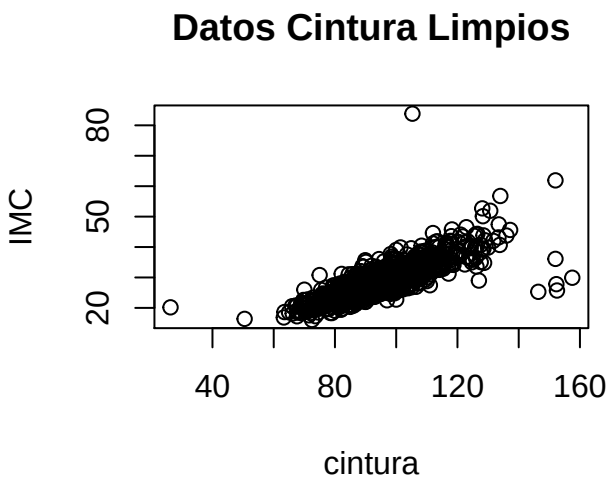
Obtenemos las graficas de dispersion para otras variables continuas para evaluar si todos los datos son posibles.

Variables Continuas



Se tienen registrados individuos con cintura de más de 200 cm y un IMC bajo, se quitaran estos valores de la base de datos.

```
##  [1] 222.2 222.2 222.2 222.2 222.2 222.2 222.2 222.2 157.5 152.5
```



Después de eliminar estos datos, nuestro tamaño de muestra final es N = 1160.

II: Selección de modelo.

Ya que tenemos los datos limpios (N = 1160) buscamos las variables significativas.

Mostramos los modelos

```
fwd.model
```

```
##
## Call:
## lm(formula = imc ~ cintura + sexo + edad + tam_pob, data = datos)
##
## Coefficients:
##   (Intercept)        cintura           sexo           edad  tam_pobUrbano
##      -3.56359        0.34381        1.70299       -0.03561        0.60997
```

```
bwd.model
```

```
##
## Call:
## lm(formula = imc ~ sexo + edad + cintura + tam_pob, data = datos)
##
## Coefficients:
##   (Intercept)           sexo           edad        cintura  tam_pobUrbano
##      -3.56359        1.70299       -0.03561        0.34381        0.60997
```

```
both.model
```

```
##
## Call:
## lm(formula = imc ~ sexo + edad + cintura + tam_pob, data = datos)
##
## Coefficients:
##   (Intercept)           sexo           edad        cintura  tam_pobUrbano
##      -3.56359        1.70299       -0.03561        0.34381        0.60997
```

Como los métodos, backward, forward y both dan el mismo modelo, seleccionamos el modelo con variables: cintura, sexo, edad y tamaño de población.

```
anova(fwd.model)
```

```
## Analysis of Variance Table
##
## Response: imc
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## cintura      1 24152.7 24152.7 2293.276 < 2.2e-16 ***
## sexo         1   890.5   890.5   84.548 < 2.2e-16 ***
## edad         1   360.0   360.0   34.180 6.531e-09 ***
## tam_pob      1   106.7   106.7   10.127    0.0015 **
## Residuals 1155 12164.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fwd.model)$adj.r.squared
```

```
## [1] 0.6759973
```

III: Ajustes de modelos de regresión.

Nota: Los valores TRUE indican que para un nivel de confianza al (1 - 0.05)% el coefieciente asociado a la variable es distinto de cero. FALSE indica que el coeficiente es cero.

Ahora hacemos modelos de regresión lineal simple para todas las variables a considerar:

```
## Analysis of Variance Table
##
## Response: imc
##             Df Sum Sq Mean Sq F value    Pr(>F)
## cintura      1  24153 24152.7  2068.5 < 2.2e-16 ***
## Residuals 1158  13522    11.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.6410938
```

```
## [1] TRUE
```

En este modelo podemos ver que se explica el **64.10938%** de la variabilidad del modelo. Fijandonos en el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable cintura **sí** es significante y podemos suponer que el coefieciente asociado a la variable cintura es distinto de 0. Entonces tenemos un modelo que explica una buena parte de la variabilidad y es significante. Este modelo resulta ser la **mejor** regresión lineal simple, explica más la variabilidad por mucho y en la gráfica notamos que ajusta adecuadamente a los datos.

```
## Analysis of Variance Table
##
## Response: imc
##                Df Sum Sq Mean Sq F value  Pr(>F)
## horas_suenio    1    185 184.525  5.6997 0.01713 *
## Residuals    1158  37490  32.375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.004897903
```

```
## [1] TRUE
```

Se observa que el 0.04% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable horas_suenio sí resulta significante y podemos suponer que el coefieciente asociado a la variable horas de sueño es distinto de 0. Entonces el modelo es significante. Pero no es muy bueno ya que la recta no ajusta adecuadamente y explica muy poca variabilidad.

```
## Analysis of Variance Table
##
## Response: imc
##             Df Sum Sq Mean Sq F value Pr(>F)
## edad         1     12  12.461  0.3831 0.5361
## Residuals 1158  37662  32.523
```

```
## [1] 0.0003307478
```

```
## [1] FALSE
```

Se observa que el 0.03% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que el coefieciente asociado a la variable edad NO resulta significante, podemos suponer que el coeficiente es IGUAL a 0. El modelo no explica la variabilidad y no es significante, por lo tanto no es un modelo a considerar.

```
## Analysis of Variance Table
##
## Response: imc
##             Df Sum Sq Mean Sq F value    Pr(>F)
## sexo         1   1188 1188.11  37.709 1.127e-09 ***
## Residuals 1158  36486   31.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.03153652
```

```
## [1] TRUE
```

Se observa que el 3% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable sexo sí resulta significante y podemos suponer que el coefieciente asociado a la variable sexo es distinto de 0. El modelo es significante, pero no es muy bueno ya que explica poca variabilidad.

```
## Analysis of Variance Table
##
## Response: imc
##             Df Sum Sq Mean Sq F value    Pr(>F)
## tam_pob      1    486  486.28  15.142 0.0001054 ***
## Residuals 1158  37188   32.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.01290747
```

```
## [1] TRUE
```

Se observa que el 1.2% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable tam_pob sí resulta significante y podemos suponer que el coefieciente asociado a la variable tam_pob es distinto de 0. El modelo es significante, sigue sin explicar mucha variabilidad, sin embargo, explica más que las otras variables discretas.

```
## Analysis of Variance Table
##
## Response: imc
##               Df Sum Sq Mean Sq F value Pr(>F)
## act_fis_vig    1     45  45.021  1.3855 0.2394
## Residuals   1158  37629  32.495
```

```
## [1] 0.001195021
```

```
## [1] FALSE
```

Se observa que el .01% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable act_fis_vig NO resulta significante, podemos suponer que el coefieciente asociado a la variable act_fis_vig es IGUAL a 0. Este modelo no es significante, lo que parece sorprendente, ya que se suele asociar el ejercicio vigoroso con la pérdida de peso.

```
## Analysis of Variance Table
##
## Response: imc
##               Df Sum Sq Mean Sq F value  Pr(>F)
## act_fis_mod    1    167 167.261  5.1641 0.02324 *
## Residuals   1158  37507  32.389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.004439674
```

```
## [1] TRUE
```

Se observa que el 0.04% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable act_fis_mod sí resulta significante y podemos suponer que el coefieciente asociado a la variable act_fis_mod es distinto de 0. El modelo es significante, pero no muy bueno, lo que de nuevo es sorprendente, ya que sugiere que la actividad física no es la mejor manera de modelar el imc.

```
## Analysis of Variance Table
##
## Response: imc
##               Df Sum Sq Mean Sq F value Pr(>F)
## act_fis_cam    1      2   2.173  0.0668 0.7961
## Residuals   1158  37672  32.532
```

```
## [1] 5.766944e-05
```

```
## [1] FALSE
```

Se observa que el modelo explica casi 0% de la variabilidad. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable Actividad Fisica Cam NO resulta significante, podemos suponer que el coefieciente asociado a la variable act_fis_cam IGUAL a 0. El modelo es no significante, para ninguna de las variables de actividad física se tiene un buen ajuste.

```
## Analysis of Variance Table
##
## Response: imc
##              Df Sum Sq Mean Sq F value Pr(>F)
## act_fis       1     45  45.021  1.3855 0.2394
## Residuals  1158  37629  32.495
```

```
## [1] 0.001195021
```

```
## [1] FALSE
```

Se observa que el .01% de la variabilidad del modelo está explicada. considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable Actividad Fisica NO resulta significante, podemos suponer que el coefieciente asociado a la variable act_fis IGUAL a 0. El modelo es no significante, lo que mantiene que la actividad física no es una variable adecuada para explicar la variable respuesta de imc.

```
## Analysis of Variance Table
##
## Response: imc
##                   Df Sum Sq Mean Sq F value Pr(>F)
## tiempo_pantalla    1      7   6.921  0.2128 0.6447
## Residuals       1158  37667  32.528
```

```
## [1] 0.000183705
```

```
## [1] FALSE
```

Se observa que el .001% de la variabilidad del modelo está explicada. Considerando el p-value de la prueba F, para un nivel de significancia de 0.05, vemos que la variable tiempo_pantalla NO resulta significante, podemos suponer que el coefieciente asociado a la variable tiempo_pantalla IGUAL a 0 El modelo es no significante, siguiendo de que es una variable discreta y no han demostrado mucha efectividad en explicar el imc.

Por lo tanto concluimos que el mejor modelo de regresión lineal simple para nuestros datos de tamaño N = 1160 es el que utiliza a la variable cintura.

Regresión Lineal Múltiple

Ahora creamos los modelos de regresión lineal múltiple, enfocandonos en las variables que ya vimos tienen significancia en el modelo, como lo son: cintura, sexo, tam_pob y edad

Dado que cintura es la variable que mejor ajusta el modelo, empezamos buscando términos de interacción entre esta y las otras 3 variables elegidas.

Primero vemos la interacción cintura ~ edad

```
#Establecemos un nivel de confianza de 0.05
alpha = 0.05

modelo1 <- lm(imc ~ cintura + edad + cintura:edad, data=datos)
anova(modelo1)
```

```
## Analysis of Variance Table
##
## Response: imc
##                 Df  Sum Sq Mean Sq  F value    Pr(>F)
## cintura          1 24152.7 24152.7 2157.054 < 2.2e-16 ***
## edad             1   398.3   398.3   35.574  3.26e-09 ***
## cintura:edad     1   179.4   179.4   16.019  6.67e-05 ***
## Residuals     1156 12943.8    11.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Obtenemos el coeficiente de determinación:
summary(modelo1)$r.squared
```

```
## [1] 0.6564276
```

```
summary(modelo1)$adj.r.squared   #R^2 adj
```

```
## [1] 0.6555359
```

Después la interacción cintura ~ sexo

```
#Establecemos un nivel de confianza de 0.05
alpha = 0.05

modelo2 <- lm(imc ~ cintura + sexo + cintura:sexo, data=datos)
anova(modelo2)
```

```
## Analysis of Variance Table
##
## Response: imc
##                Df  Sum Sq Mean Sq   F value Pr(>F)
## cintura         1 24152.7 24152.7 2212.9596 <2e-16 ***
## sexo            1   890.5   890.5   81.5867 <2e-16 ***
## cintura:sexo    1    14.2    14.2    1.3039 0.2537
## Residuals    1156 12616.8    10.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Obtenemos el coeficiente de determinación
summary(modelo2)$r.squared
```

```
## [1] 0.6651071
```

```
summary(modelo2)$adj.r.squared   #R^2 adj
```

```
## [1] 0.664238
```

La interacción tam_pob ~ cintura

```
modelo3 <- lm(imc ~ tam_pob + cintura + cintura:tam_pob, data=datos)
anova(modelo3)
```

```
## Analysis of Variance Table
##
## Response: imc
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## tam_pob            1   486.3   486.3   42.560 1.023e-10 ***
## cintura            1 23832.5 23832.5 2085.852 < 2.2e-16 ***
## tam_pob:cintura    1   147.2   147.2   12.883 0.0003454 ***
## Residuals       1156 13208.2    11.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Obtenemos el coeficiente de determinación
summary(modelo3)$r.squared
```

```
## [1] 0.6494097
```

```
summary(modelo3)$adj.r.squared   #R^2 adj
```

```
## [1] 0.6484998
```

Ahora probamos el modelo con las 4 variables más significativas y lo usaremos para contrastar con los términos de interacción

```
modelo4 <- lm(imc ~ cintura + sexo + edad + tam_pob, data=datos)
anova(modelo4)
```

```
## Analysis of Variance Table
##
## Response: imc
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## cintura      1 24152.7 24152.7 2293.276 < 2.2e-16 ***
## sexo         1   890.5   890.5   84.548 < 2.2e-16 ***
## edad         1   360.0   360.0   34.180 6.531e-09 ***
## tam_pob      1   106.7   106.7   10.127    0.0015 **
## Residuals 1155 12164.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(modelo4)
```

```
##
## Call:
## lm(formula = imc ~ cintura + sexo + edad + tam_pob, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.609  -1.555  -0.216   1.325  51.754
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.563586   0.713320  -4.996 6.76e-07 ***
## cintura        0.343810   0.007213  47.665  < 2e-16 ***
## sexo           1.702993   0.195458   8.713  < 2e-16 ***
## edad          -0.035614   0.006128  -5.811 8.00e-09 ***
## tam_pobUrbano  0.609973   0.191681   3.182   0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.245 on 1155 degrees of freedom
## Multiple R-squared:  0.6771, Adjusted R-squared:  0.676
## F-statistic: 605.5 on 4 and 1155 DF,  p-value: < 2.2e-16
```

Nuestro coeficiente de correlación ajustado es de 67.59973%, en los modelos siguientes veremos si podemos explicar más variabilidad.

Eliminamos el intercepto del modelo.

```
modelo4 <- lm(imc ~ cintura + sexo + edad + tam_pob - 1, data=datos)
anova(modelo5)
```

```
## Analysis of Variance Table
##
## Response: imc
##               Df Sum Sq Mean Sq F value Pr(>F)
## act_fis_vig    1     45  45.021  1.3855 0.2394
## Residuals   1158  37629  32.495
```

```
summary(modelo5)
```

```
##
## Call:
## lm(formula = imc ~ act_fis_vig, data = datos)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.036  -3.810  -0.757   3.028  55.267
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.473695   0.266663 106.778   <2e-16 ***
## act_fis_vig  0.001707   0.001450   1.177    0.239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.7 on 1158 degrees of freedom
## Multiple R-squared:  0.001195,   Adjusted R-squared:  0.0003325
## F-statistic: 1.385 on 1 and 1158 DF,  p-value: 0.2394
```

Al eliminar el intercepto se explica mucha más variabilidad, ahora explica el 98.79% de esta, sin embargo la variable tam_pobRural no es significante en este modelo.

Agregamos términos de interacción a este nuevo modelo sin intercepto

```
modelo6 <- lm(imc ~ cintura + sexo + edad + tam_pob + cintura:tam_pob - 1, data=datos)
anova(modelo6)
```

```
## Analysis of Variance Table
##
## Response: imc
##                   Df Sum Sq Mean Sq   F value    Pr(>F)
## cintura            1 980547  980547 94648.197 < 2.2e-16 ***
## sexo               1    737     737    71.128 < 2.2e-16 ***
## edad               1    562     562    54.292 3.284e-13 ***
## tam_pob            2    346     173    16.710 7.015e-08 ***
## cintura:tam_pob    1    209     209    20.182 7.747e-06 ***
## Residuals       1154  11955      10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(modelo6)
```

```
##
## Call:
## lm(formula = imc ~ cintura + sexo + edad + tam_pob + cintura:tam_pob -
##     1, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.550  -1.555  -0.213   1.431  51.440
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## cintura                 0.311424   0.010156  30.664  < 2e-16 ***
## sexo                    1.730291   0.193950   8.921  < 2e-16 ***
## edad                   -0.037523   0.006093  -6.159 1.01e-09 ***
## tam_pobRural           -0.449871   0.990401  -0.454     0.65
## tam_pobUrbano          -5.877199   0.969474  -6.062 1.82e-09 ***
## cintura:tam_pobUrbano   0.063732   0.014186   4.492 7.75e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.219 on 1154 degrees of freedom
## Multiple R-squared:  0.988,  Adjusted R-squared:  0.9879
## F-statistic: 1.58e+04 on 6 and 1154 DF,  p-value: < 2.2e-16
```

Al agregar el término de interacción entre cintura y tam_pob, se obtiene una pequeña mejora en el coefieciente de correlación, sin embargo tam_pobRural no es significativo.

Cambiamos el término de interacción por cintura:sexo.

```
modelo7 <- lm(imc ~ cintura + sexo + edad + tam_pob + cintura:sexo - 1, data=datos)
anova(modelo7)
```

```
## Analysis of Variance Table
##
## Response: imc
##                Df Sum Sq Mean Sq    F value    Pr(>F)
## cintura         1 980547  980547 93117.9423 < 2.2e-16 ***
## sexo            1    737     737    69.9777 < 2.2e-16 ***
## edad            1    562     562    53.4139 5.033e-13 ***
## tam_pob         2    346     173    16.4399 9.122e-08 ***
## cintura:sexo    1     13      13     1.1984    0.2739
## Residuals    1154  12152      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(modelo7)
```

```
##
## Call:
## lm(formula = imc ~ cintura + sexo + edad + tam_pob + cintura:sexo -
##     1, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.917  -1.532  -0.212   1.364  51.875
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## cintura        0.332794   0.012381  26.880  < 2e-16 ***
## sexo           0.141605   1.439633   0.098   0.9217
## edad          -0.035376   0.006132  -5.770 1.02e-08 ***
## tam_pobRural  -2.540571   1.175604  -2.161   0.0309 *
## tam_pobUrbano -1.921719   1.188862  -1.616   0.1063
## cintura:sexo   0.016514   0.015086   1.095   0.2739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.245 on 1154 degrees of freedom
## Multiple R-squared:  0.9878, Adjusted R-squared:  0.9877
## F-statistic: 1.555e+04 on 6 and 1154 DF,  p-value: < 2.2e-16
```

Notamos que al cambiar el termino de interacción cintura:tam_pob por cintura:sexo, el coeficiente de correlación (ajustado y normal) disminuye y las variables cintura:sexo, tam_pobUrbano y sexo no son significativas.

```r
modelo8 <- lm(imc ~ cintura + sexo + edad + tam_pob + cintura:edad - 1, data=datos)
anova(modelo8)
```

```
## Analysis of Variance Table
##
## Response: imc
##                Df Sum Sq Mean Sq   F value    Pr(>F)
## cintura         1 980547  980547 94425.747 < 2.2e-16 ***
## sexo            1    737     737    70.960 < 2.2e-16 ***
## edad            1    562     562    54.164 3.494e-13 ***
## tam_pob         2    346     173    16.671 7.288e-08 ***
## cintura:edad    1    181     181    17.423 3.217e-05 ***
## Residuals    1154  11984      10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(modelo8)
```

```
##
## Call:
## lm(formula = imc ~ cintura + sexo + edad + tam_pob + cintura:edad -
##     1, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.966  -1.548  -0.165   1.414  51.369
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## cintura          0.422769   0.020227  20.901  < 2e-16 ***
## sexo             1.683797   0.194137   8.673  < 2e-16 ***
## edad             0.119410   0.037635   3.173  0.00155 **
## tam_pobRural   -10.823449   1.877979  -5.763 1.06e-08 ***
## tam_pobUrbano  -10.167833   1.872107  -5.431 6.82e-08 ***
## cintura:edad    -0.001678   0.000402  -4.174 3.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.222 on 1154 degrees of freedom
## Multiple R-squared:  0.9879, Adjusted R-squared:  0.9879
## F-statistic: 1.577e+04 on 6 and 1154 DF,  p-value: < 2.2e-16
```

En este modelo todas nuestras variables son significativas, además al tener explicar 98.79% de la variabilidad, explica la misma cantidad que el modelo6 con término de interacción cintura:tam_pob, que tambien explica el 98.79% de la variabilidad, pero el modelo6 tiene variables no significativas.

Utilizamos el criterio de la información de Akaike para encontrar el modelo más parsimonioso.

```r
AIC(modelo1, modelo2, modelo3, modelo4, modelo5, modelo6, modelo7, modelo8)
```

```
##         df      AIC
## modelo1  5 6100.088
## modelo2  5 6070.407
## modelo3  5 6123.544
## modelo4  6 6030.048
## modelo5  3 7333.996
```

```
## modelo6  7 6011.936
## modelo7  7 6030.844
## modelo8  7 6014.666
```

Vemos que el modelo6(cintura + sexo + edad + tam_pob + cintura:tam_pob - 1) tiene la menor Información de Akaike, sin embargo el modelo8(imc ~ cintura + sexo + edad + tam_pob + cintura:edad - 1) tiene casi la misma información de Akaike y todas sus variables son significativas.

Como el modelo8(cintura + sexo + edad + tam_pob + cintura:edad - 1) tiene una información de Akaike(6014.666) muy pequeña, explica la mayor cantidad de la variabilidad (Coeficiente de Determinación ajustado = 0.9879) y todas sus variables son significativas se elige a esto como el mejor modelo para estos datos con tamaño de muestra N = 1160.

Prueba de significancia del modelo completo con términos de interacción

Probamos la significancia de la regresión del mejor modelo de regresión lineal múltiple(modelo6):

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \quad vs \quad H_1 : \beta_j \neq 0$$

- β_1 es el coeficiente asociado a la variable cintura,
- β_2 es el coeficiente asociado a la variable sexo.
- β_3 es el coeficiente asociado a la variable edad.
- β_4 es el coeficiente asociado a la variable tam_pobRural.
- β_5 es el coeficiente asociado a la variable tam_pobUrbana.
- β_6 es el coeficiente asociado a la variable cintura:edad.

```
#Establecemos un nivel de confianza de 0.05
alpha = 0.05

SCE <- sum((datos$imc - fitted(modelo8))^2)
SCR <- sum((fitted(modelo8) - mean(datos$imc))^2)

F <- (SCR/(6))/(SCE/1154)
F > qf(1 - alpha, 1154, 6)
```

```
## [1] TRUE
```

Como el estadístico F es mayor al percentil 0.95 de una distribución F, para un nivel de confianza de 0.05 decimos que la regresión es significante.

Detalles del modelo seleccionado

```
anova(modelo8)
```

```
## Analysis of Variance Table
##
## Response: imc
##                Df Sum Sq Mean Sq   F value    Pr(>F)
## cintura         1 980547  980547 94425.747 < 2.2e-16 ***
## sexo            1    737     737    70.960 < 2.2e-16 ***
## edad            1    562     562    54.164 3.494e-13 ***
## tam_pob         2    346     173    16.671 7.288e-08 ***
## cintura:edad    1    181     181    17.423 3.217e-05 ***
## Residuals    1154  11984      10
## ---
```

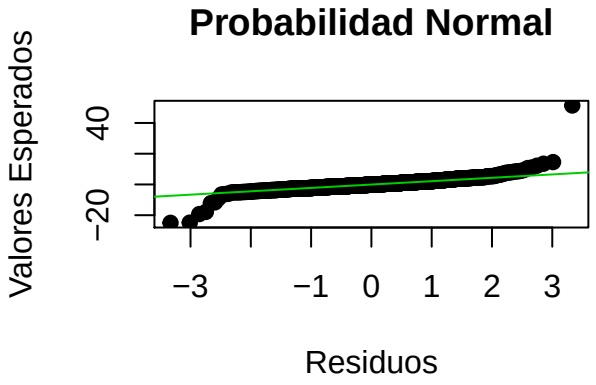
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

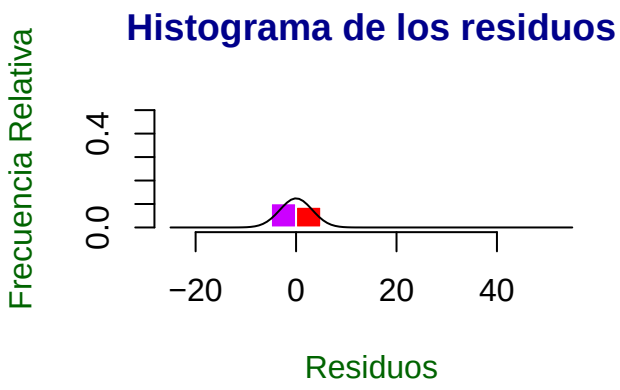
```
summary(modelo8)
```

```
##
## Call:
## lm(formula = imc ~ cintura + sexo + edad + tam_pob + cintura:edad -
##     1, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.966  -1.548  -0.165   1.414  51.369
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## cintura          0.422769   0.020227  20.901  < 2e-16 ***
## sexo             1.683797   0.194137   8.673  < 2e-16 ***
## edad             0.119410   0.037635   3.173  0.00155 **
## tam_pobRural  -10.823449   1.877979  -5.763 1.06e-08 ***
## tam_pobUrbano -10.167833   1.872107  -5.431 6.82e-08 ***
## cintura:edad   -0.001678   0.000402  -4.174 3.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.222 on 1154 degrees of freedom
## Multiple R-squared:  0.9879, Adjusted R-squared:  0.9879
## F-statistic: 1.577e+04 on 6 and 1154 DF,  p-value: < 2.2e-16
```

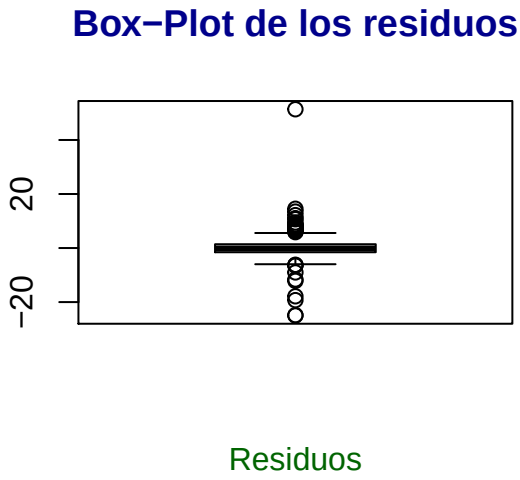
Diagnóstico del modelo seleccionado.

Graficamos los residuales.





Realizamos la prueba de Anderson-Darling para cerificar la bondad de ajuste




```
##
##  Anderson-Darling normality test
##
## data:  rnorm(rstandard(modelo), mean = 0, sd = 1)
## A = 0.53156, p-value = 0.174
```

Como el p-value de la prueba de Anderson-Darling es mayor que un nivel de confianza 0.05, decimos que nuestros errores se distribuyen normal(0, 1).

Buscamos datos influyentes mediante la distancia de Cook

```
modelo8 <- lm(imc ~ sexo + edad + tam_pob + cintura:tam_pob, data=datos)
n <- dim(datos)[1]
cooksd <- cooks.distance(modelo6)
as.numeric(names(cooksd)[(cooksd > (4/n))])
```

```
##  [1]  113  132  160  161  171  215  351  480  481  584  586  668  679  729
## [15]  831  836  856  875  901  949  951  987 1030 1033 1039 1075 1091 1104
## [29] 1108 1150 1166 1197 1264
```

Resulta haber varios datos influyentes, sin embargo todos parecen ser posibles.

```
n <- dim(datos)[1]
influential <- which(cooksd > (4/n))
datos[influential,]$edad
```

```
##  [1] 58 65 34 63 27 41 34 55 33 41 25 38 57 25 89 57 35 47 49 43 49 47 95
## [24] 80 62 40 23 25 39 39 34 72 90
```

```
datos[influential,]$act_fis
```

```
##  [1] 180.000000 117.142857 360.000000  17.142857 411.428571 244.285714
##  [7] 171.428571 180.000000  51.428571 107.142857 180.000000  40.000000
## [13] 150.000000   0.000000   4.285714  50.000000 210.000000 360.000000
## [19]  30.000000  20.000000  90.000000  60.000000   0.000000  30.000000
## [25] 180.000000 540.000000 197.142857 351.428571  40.000000  32.142857
## [31]  36.428571 180.000000 150.000000
```

Son datos de personas que son de edades mayores o personas que hacen muy poco o mucho ejercicio.

Revisamos la multicolinealidad usando el Vif

```
Vif <- 1/(1 - summary(modelo8)$r.squared)
Vif
```

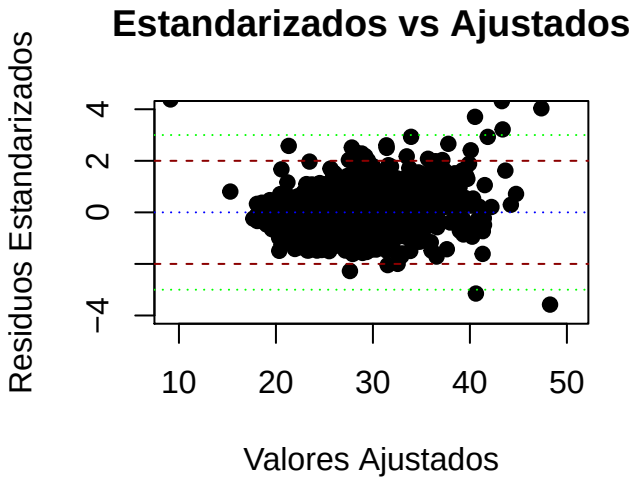
```
## [1] 3.151248
```

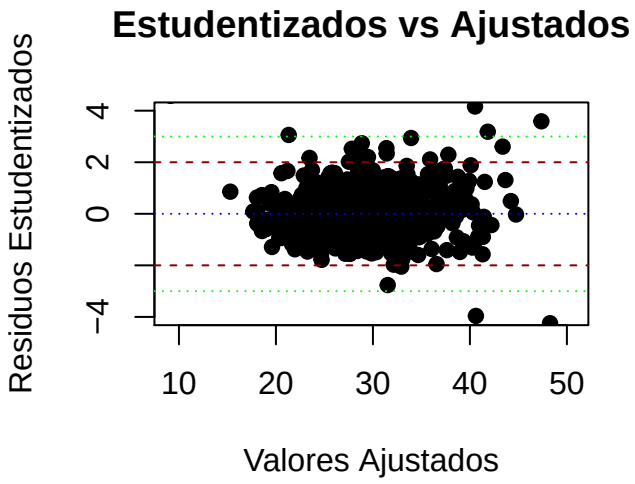
```
Vif < 10
```

```
## [1] TRUE
```

Como el Vif < 10 no existe ninguna correlación significante.

Probamos la Homocedasticidad.





Conclusión

El mejor modelo, es el modelo8 de la sección de regresión lineal múltiple de este proyecto, este modelo, que lleva las variables, cintura, sexo, edad, tam_pob y la variable de interacción tam_pob:edad resulto ser la mejor forma de explicar la variable respuesta de imc. Nos pareció importante notar que con estos datos la actividad física no parece tener un rol importante en la predicción del imc. En nuestra experiencia tambien nos parecio sorprendente que al cambiar la variable cintura por un término de interacción con edad obtuviesemos un mejor modelo, esperabamos que la interacción con sexo tuviera un peso más importante, pero no fue así.

Nota Técnica

A lo largo del código hay varios TRUE que aparecen, estos indican si las regresiones son significantes o no. Usamos las librerías: foreign, MASS, MPV y lmtest. Decidimos entregar un trabajo elaborado en Markdown para poder mostrar con agilidad los resultados junto con nuestras interpretaciones y comentarios.

Bibliografía

University of Sheffield,Ellen Marshall and Sofia Maria Karadimitriou, https://www.sheffield.ac.uk/polopoly_fs/1.536483!/file/MASH_multiple_regression_R.pdf, 16/05/2018