Práctica 01

Peredo Leonel

Correlación

Ejercicio 1.

El conjunto de datos `bdims` del paquete `openintro` que se habilita en el workspace del `R` con `data(bdims, package = "openintro")` consiste en medidas del diámetro y circunferencia de distintas partes del cuerpo (21 variables), así como edad, peso, altura y género de 507 personas físicamente activas. (Para más detalle, tipear `help(bdims, package = "openintro")`)

a. Calcular las correlaciones muestrales entre las 21 variables que miden el diámetro o circunferencia de las distintas partes del cuerpo. ¿Cuántas correlaciones debe calcular? ¿Cuál sería la mejor manera de exhibir esta información? ¿Están positiva o negativamente correlacionadas estas variables?

```r
data(bdims, package = "openintro")
correlations <- cor(bdims[1:21])
knitr::kable(correlations[,1:7])
```

	bia_di	bii_di	bit_di	che_de	che_di	elb_di	wri_di
bia_di	1.0000000	0.3090358	0.4862726	0.5832585	0.7691406	0.7658212	0.7228388
bii_di	0.3090358	1.0000000	0.6734567	0.3567852	0.3311695	0.3228573	0.2792363
bit_di	0.4862726	0.6734567	1.0000000	0.4725560	0.5241288	0.5257579	0.4681583
che_de	0.5832585	0.3567852	0.4725560	1.0000000	0.6650702	0.6652377	0.6081147
che_di	0.7691406	0.3311695	0.5241288	0.6650702	1.0000000	0.7588682	0.7308643
elb_di	0.7658212	0.3228573	0.5257579	0.6652377	0.7588682	1.0000000	0.8399305
wri_di	0.7228388	0.2792363	0.4681583	0.6081147	0.7308643	0.8399305	1.0000000
kne_di	0.6359621	0.4377883	0.6083021	0.5502889	0.6590648	0.7315042	0.7124844
ank_di	0.6614162	0.3683128	0.4954057	0.5978540	0.6685389	0.8210977	0.7724489
sho_gi	0.7925957	0.2772388	0.4787637	0.7376115	0.8706480	0.8194698	0.7783992
che_gi	0.7218401	0.3256838	0.4880845	0.8065033	0.8703062	0.8031396	0.7665426
wai_gi	0.6416072	0.4347003	0.5702148	0.8037549	0.7880334	0.6946192	0.6807824
nav_gi	0.3057128	0.5805152	0.6175048	0.6212365	0.5012123	0.4387605	0.3992720
hip_gi	0.3400615	0.5641529	0.7482328	0.5563131	0.5212073	0.4393353	0.4223687
thi_gi	0.1219279	0.4141551	0.5317738	0.3576541	0.3147735	0.2069166	0.1940200
bic_gi	0.6950618	0.2991071	0.4801457	0.7328977	0.7923345	0.8047840	0.7621594
for_gi	0.7526421	0.2896823	0.4780849	0.7175490	0.8071175	0.8582063	0.8147088
kne_gi	0.5079070	0.4724691	0.6233547	0.5636517	0.5928721	0.5909794	0.5818739
cal_gi	0.5108144	0.4070641	0.5929802	0.5535016	0.5969089	0.5799083	0.5814377
ank_gi	0.6034678	0.3358175	0.5390628	0.5873425	0.6350210	0.6641619	0.6546945
wri_gi	0.7715976	0.2632546	0.4795170	0.6802408	0.7608931	0.8457563	0.8625527

```
knitr::kable(correlations[,8:14])
```

	kne_di	ank_di	sho_gi	che_gi	wai_gi	nav_gi	hip_gi
bia_di	0.6359621	0.6614162	0.7925957	0.7218401	0.6416072	0.3057128	0.3400615
bii_di	0.4377883	0.3683128	0.2772388	0.3256838	0.4347003	0.5805152	0.5641529
bit_di	0.6083021	0.4954057	0.4787637	0.4880845	0.5702148	0.6175048	0.7482328
che_de	0.5502889	0.5978540	0.7376115	0.8065033	0.8037549	0.6212365	0.5563131
che_di	0.6590648	0.6685389	0.8706480	0.8703062	0.7880334	0.5012123	0.5212073
elb_di	0.7315042	0.8210977	0.8194698	0.8031396	0.6946192	0.4387605	0.4393353
wri_di	0.7124844	0.7724489	0.7783992	0.7665426	0.6807824	0.3992720	0.4223687
kne_di	1.0000000	0.7232729	0.6818019	0.6522224	0.6239675	0.4712506	0.5795936
ank_di	0.7232729	1.0000000	0.6921115	0.7058718	0.6369715	0.4365745	0.4077358
sho_gi	0.6818019	0.6921115	1.0000000	0.9271923	0.8234546	0.5154661	0.5336717
che_gi	0.6522224	0.7058718	0.9271923	1.0000000	0.8837994	0.6229823	0.5834991
wai_gi	0.6239675	0.6369715	0.8234546	0.8837994	1.0000000	0.7547704	0.6923506
nav_gi	0.4712506	0.4365745	0.5154661	0.6229823	0.7547704	1.0000000	0.8258924
hip_gi	0.5795936	0.4077358	0.5336717	0.5834991	0.6923506	0.8258924	1.0000000
thi_gi	0.4315276	0.1926277	0.3234272	0.3630508	0.4210849	0.6026428	0.8289411
bic_gi	0.6814055	0.6862886	0.8951884	0.9081845	0.8047044	0.5578071	0.5598848
for_gi	0.7206519	0.7352504	0.8949838	0.8875909	0.7807924	0.4862181	0.5143585
kne_gi	0.7338176	0.5423538	0.6247826	0.6140547	0.6582072	0.6120932	0.7349017
cal_gi	0.6860935	0.5436159	0.6270538	0.6088643	0.6313445	0.5247789	0.6745805
ank_gi	0.6547070	0.6772298	0.6797568	0.6691396	0.6558891	0.5194785	0.5770429
wri_gi	0.7311803	0.7627486	0.8407085	0.8246754	0.7289813	0.4354197	0.4588567

```
knitr::kable(correlations[,15:21])
```

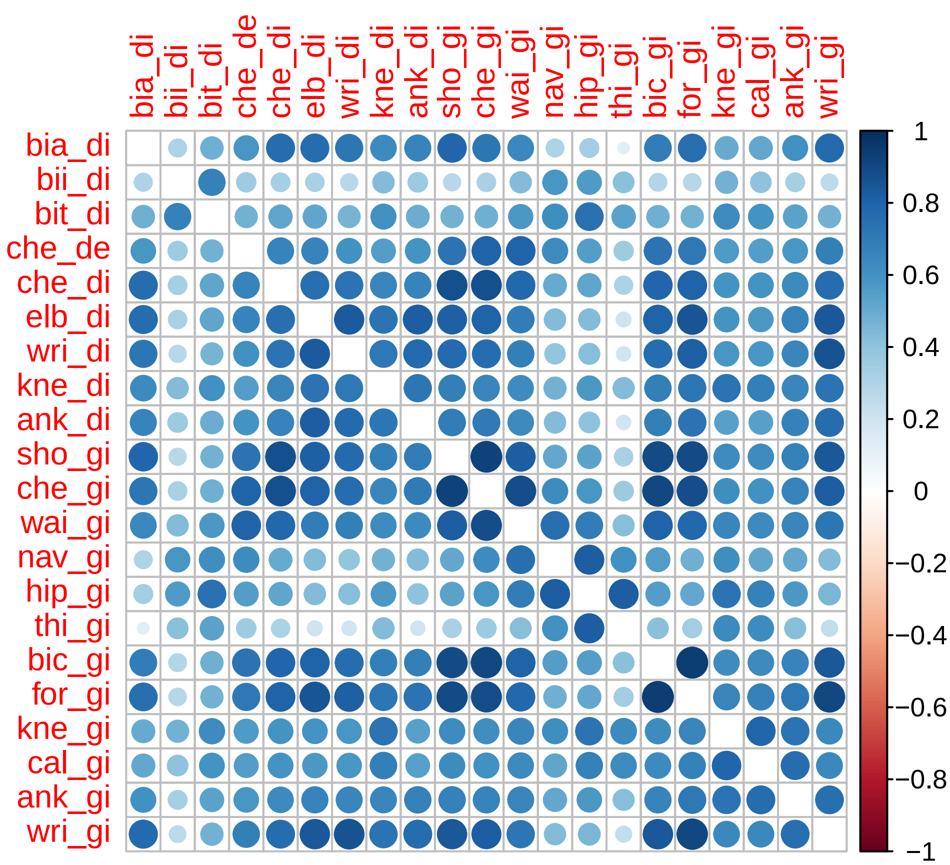
	thi_gi	bic_gi	for_gi	kne_gi	cal_gi	ank_gi	wri_gi
bia_di	0.1219279	0.6950618	0.7526421	0.5079070	0.5108144	0.6034678	0.7715976
bii_di	0.4141551	0.2991071	0.2896823	0.4724691	0.4070641	0.3358175	0.2632546
bit_di	0.5317738	0.4801457	0.4780849	0.6233547	0.5929802	0.5390628	0.4795170
che_de	0.3576541	0.7328977	0.7175490	0.5636517	0.5535016	0.5873425	0.6802408
che_di	0.3147735	0.7923345	0.8071175	0.5928721	0.5969089	0.6350210	0.7608931
elb_di	0.2069166	0.8047840	0.8582063	0.5909794	0.5799083	0.6641619	0.8457563
wri_di	0.1940200	0.7621594	0.8147088	0.5818739	0.5814377	0.6546945	0.8625527
kne_di	0.4315276	0.6814055	0.7206519	0.7338176	0.6860935	0.6547070	0.7311803
ank_di	0.1926277	0.6862886	0.7352504	0.5423538	0.5436159	0.6772298	0.7627486
sho_gi	0.3234272	0.8951884	0.8949838	0.6247826	0.6270538	0.6797568	0.8407085
che_gi	0.3630508	0.9081845	0.8875909	0.6140547	0.6088643	0.6691396	0.8246754
wai_gi	0.4210849	0.8047044	0.7807924	0.6582072	0.6313445	0.6558891	0.7289813
nav_gi	0.6026428	0.5578071	0.4862181	0.6120932	0.5247789	0.5194785	0.4354197
hip_gi	0.8289411	0.5598848	0.5143585	0.7349017	0.6745805	0.5770429	0.4588567
thi_gi	1.0000000	0.4114580	0.3452848	0.6384400	0.6288901	0.4217687	0.2416102
bic_gi	0.4114580	1.0000000	0.9423755	0.6207299	0.6374041	0.6693240	0.8479443
for_gi	0.3452848	0.9423755	1.0000000	0.6575450	0.6701918	0.7125539	0.9047086
kne_gi	0.6384400	0.6207299	0.6575450	1.0000000	0.7958277	0.7377154	0.6409596
cal_gi	0.6288901	0.6374041	0.6701918	0.7958277	1.0000000	0.7622219	0.6476269
ank_gi	0.4217687	0.6693240	0.7125539	0.7377154	0.7622219	1.0000000	0.7536365
wri_gi	0.2416102	0.8479443	0.9047086	0.6409596	0.6476269	0.7536365	1.0000000

Se calcularon $21^2 = 441$ correlaciones. La mejor forma de exhibir esta información es con una tabla. Todos los pares de variables tienen $r > 0$ lo cual indica que están positivamente correlacionadas. Otra forma más gráfica y menos ruidosa es con un gráfico de este estilo:

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```
corrplot.mixed(
  correlations,
  lower="circle",
  upper="circle",
  tl.pos = c("lt"),
  diag = c("n", "l", "u"),
  bg = "white",
  addgrid.col = "grey",
  lower.col = NULL,
  upper.col = NULL,
  plotCI = c("n", "square", "circle", "rect"),
  mar = c(0, 0, 0, 0),
  )
```



b. Encontrar las dos variables con mayor correlación entre sí. Realizar un scatter plot. ¿Le parece que este número resume adecuadamente el vínculo entre ambas variables?

```r
# Máximo de matriz correlations con la diagonal anulada
maxcorrelation <- max(`diag<-`(correlations,0))
colnames(correlations)[which(correlations == maxcorrelation)]
```

```
## [1] NA NA
```

```r
correlations[332]
```

```
## [1] 0.9423755
```

c. Repetir con las de menor correlación.
d. Hacer un scatter plot de peso en el eje y y altura en el eje x y calcular la correlación muestral o de Pearson. ¿Le parece que este número resume adecuadamente el vínculo entre ambas variables?
e. Hacer scatter plots de la variable `bia_di`, que es la distancia biacromial (informalmente, la distancia emtre los hombros) con las siguientes cuatro variables y calcular las correlaciones de a pares para ambas. Observar cómo se comportan los scatterplots para distintos valores de la correlación.

- `age`, la edad
- `bii_di`, el ancho de la pelvis
- `che_de`, la profundidad del pecho
- `wri_di`, la circunferencia de la muñeca

Ejercicio 2

Sean $(X_i, Y_i)_{1 \leq i \leq n}$ observaciones bivariadas, la covarianza muestral entre X e Y, basada en las observaciones se define por

$$\widehat{\text{cov}}((X_1, Y_1), ..., (X_n, Y_n)) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Por simplicidad en vez de escribir $\widehat{\text{cov}}((X_1, Y_1), ..., (X_n, Y_n))$ a veces escribiremos $\widehat{\text{cov}}(X_i, Y_i)$

a. Sean $a, b \in \mathbb{R}$ constantes.
b. Definimos $X_i^* = X_i + a$, con $i = 1, ..., n$. Probar que $\widehat{\text{cov}}(X_i^*, Y_i) = \widehat{\text{cov}}(X_i, Y_i)$

Por definición, tenemos que

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)(Y_i - \bar{Y})$$

Reemplazando según $X_i^* = X_i + a$, primero calculamos $\bar{X}_i^*$:

$$\bar{X}_i^* = \frac{1}{n} \sum_{i=1}^{n} X_i^* = \frac{1}{n} \sum_{i=1}^{n} (X_i + a) = \frac{1}{n} \sum_{i=1}^{n} X_i + \frac{1}{n} \sum_{i=1}^{n} a = \bar{X} + \frac{na}{n} = \bar{X} + a$$

Por lo tanto,

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i + a - (\bar{X} + a))(Y_i - \bar{Y}) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i + a - \bar{X} - a)(Y_i - \bar{Y})$$

Es decir que

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \widehat{\text{cov}}(X_i, Y_i)$$

como se quería probar.

ii. Definimos $X_i^* = bX_i + a$, con $i = 1, ..., n$. Probar que $\widehat{\text{cov}}(X_i^*, Y_i) = b \cdot \widehat{\text{cov}}(X_i, Y_i)$

Por definición, tenemos que

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)(Y_i - \bar{Y})$$

Reemplazando según $X_i^* = bX_i + a$, primero calculamos $\bar{X}_i^*$:

$$\bar{X}_i^* = \frac{1}{n} \sum_{i=1}^{n} X_i^* = \frac{1}{n} \sum_{i=1}^{n} (bX_i + a) = \frac{b}{n} \sum_{i=1}^{n} X_i + \frac{1}{n} \sum_{i=1}^{n} a = b\bar{X} + \frac{na}{n} = b\bar{X} + a$$

Por lo tanto,

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (bX_i + a - (b\bar{X} + a))(Y_i - \bar{Y}) = \frac{1}{n-1} \sum_{i=1}^{n} (bX_i + a - b\bar{X} - a)(Y_i - \bar{Y})$$

Es decir que

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} b(X_i - \bar{X})(Y_i - \bar{Y}) = b\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = b \cdot \widehat{\text{cov}}(X_i, Y_i)$$

como se quería probar.

b. Sean $X_i^* = X_i - \bar{X}$, y $Y_i^* = Y_i - \bar{Y}$, con $i = 1, ..., n$. Probar que $\widehat{\text{cov}}(X_i^*, Y_i^*) = \widehat{\text{cov}}(X_i^*, Y_i) = \widehat{\text{cov}}(X_i, Y_i)$

Por definición,

$$\widehat{\text{cov}}(X_i^*, Y_i^*) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)(Y_i^* - \bar{Y}^*)$$

Primero calculamos $\bar{Y}^*$:

$$\bar{Y}^* = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{n} \sum_{i=1}^{n} \bar{Y} = \bar{Y} - \frac{n\bar{Y}}{n} = 0$$

Luego, reemplazando,

$$\widehat{\text{cov}}(X_i^*, Y_i^*) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)(Y_i - \bar{Y} - 0) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i^* - \bar{X}^*)(Y_i - \bar{Y}) = \widehat{\text{cov}}(X_i^*, Y_i)$$

Ahora calculamos $\bar{X}^*$:

$$\bar{X}^* = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} \bar{X} = \bar{X} - \frac{n\bar{X}}{n} = 0$$

Luego, reemplazando,

$$\widehat{\text{cov}}(X_i^*, Y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X} - 0)(Y_i - \bar{Y}) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \widehat{\text{cov}}(X_i, Y_i)$$

Por lo tanto, por propiedad transitiva, $\widehat{\text{cov}}(X_i^*, Y_i^*) = \widehat{\text{cov}}(X_i^*, Y_i) = \widehat{\text{cov}}(X_i, Y_i)$, como se quería probar.