PEC 2 Regresión modelos y métodos

Marc Bañuls Tornero

21/6/2020

Contents

Ejercicio 1

Se cargan los datos en R:

```
peru1 <- read.delim("peru.txt")
```

Debido a que se el ejercicio pide unas variables concretas para el análisis de los datos encontrados en "peru.txt", adaptamos estos datos en una nueva tabla:

```
Fraction <- peru1$Years / peru1$Age
peru <- data.frame(peru1$Age, peru1$Years, Fraction, peru1$Weight, peru1$Height, peru1$Chin, peru1$Forea
colnames(peru) <- c("Age", "Years", "Fraction", "Weight", "Height", "Chin", "Forearm", "Calf", "Pulse",
```

Posteriormente realizamos el modelo de regresión múltiple deseado:

```
lm_peru <- lm(Systol ~. , data = peru)
```

(a) Estudiar la posible multicolinealidad de este modelo

La multicolinealidad es la presencia de una correlación alta entre más de dos variables en un modelo de regresión múltiple, por lo que se han de buscar en el modelo ajustado más de dos variables con una alta correlación.

Para saber si existe multicolinealidad en el modelo, se puede primero observar el resumen de éste:

```
summary(lm_peru)
```

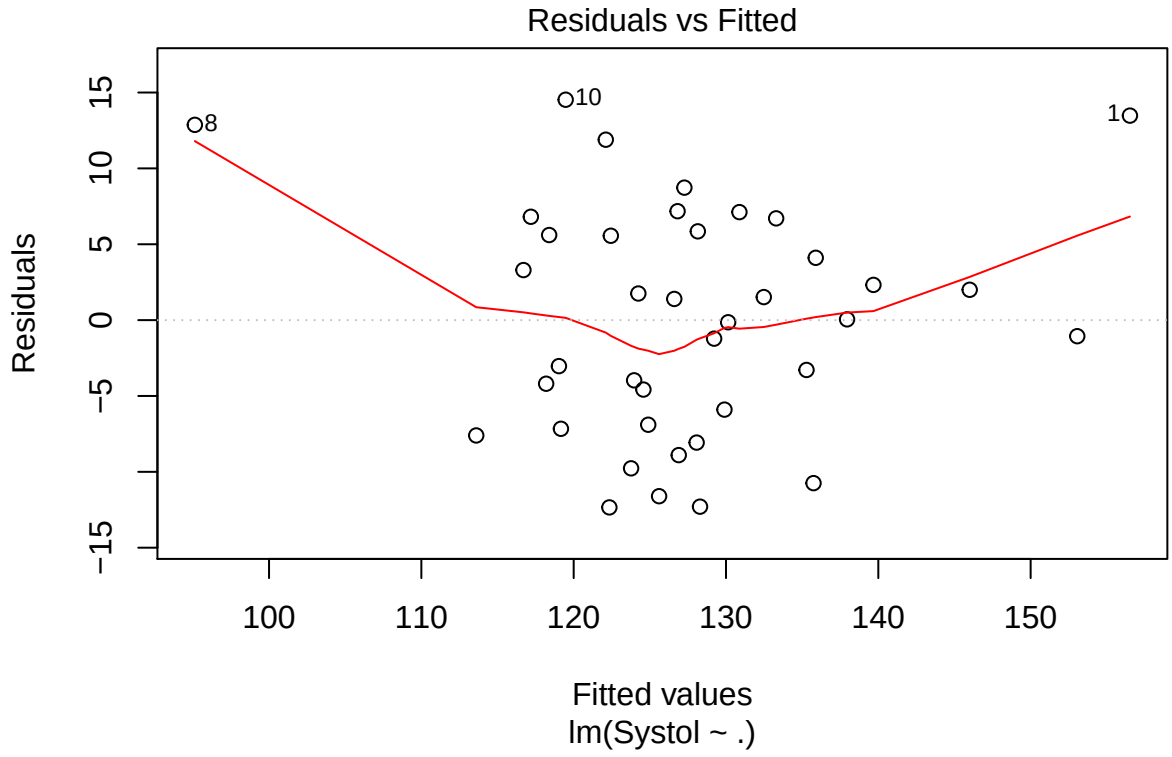
```
##
## Call:
## lm(formula = Systol ~ ., data = peru)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3442  -6.3972   0.0507   5.7292  14.5257
##
## Coefficients:
```

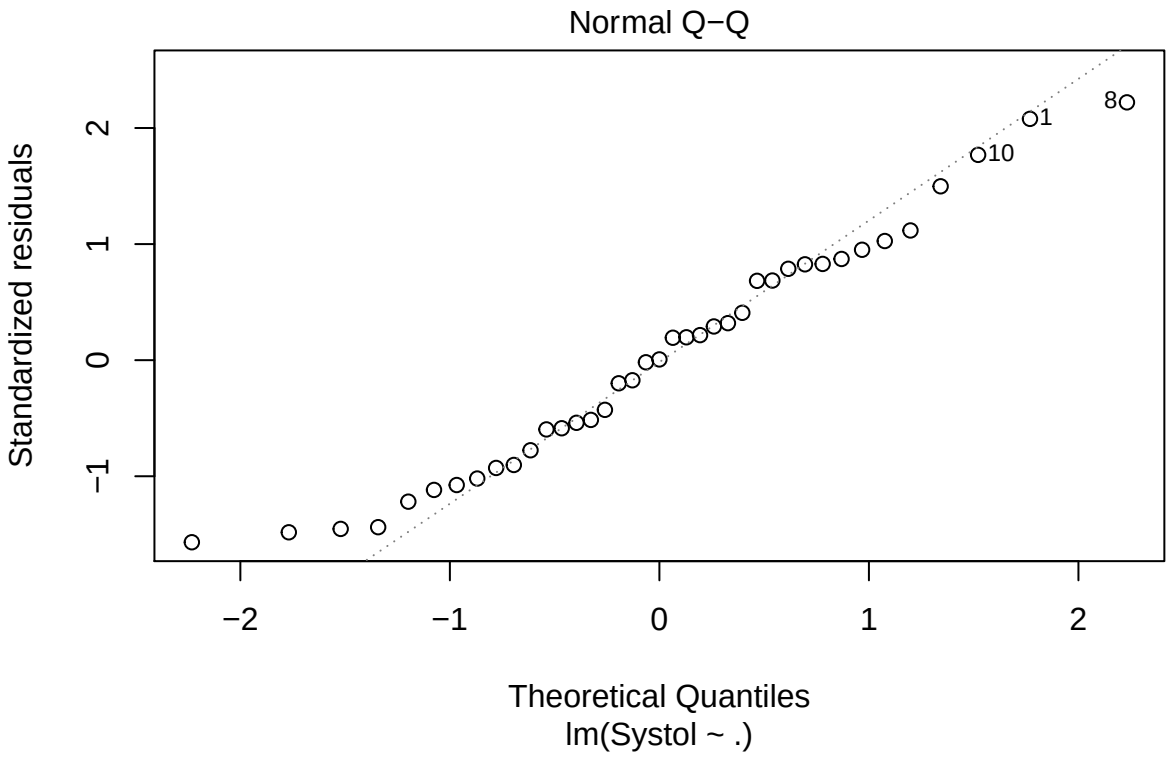
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.81907   48.97096   2.998 0.005526 **
## Age           -1.12144    0.32741  -3.425 0.001855 **
## Years          2.45538    0.81458   3.014 0.005306 **
## Fraction    -115.29395   30.16900  -3.822 0.000648 ***
## Weight         1.41393    0.43097   3.281 0.002697 **
## Height        -0.03464    0.03686  -0.940 0.355194
## Chin          -0.94369    0.74097  -1.274 0.212923
## Forearm       -1.17085    1.19329  -0.981 0.334612
## Calf          -0.15867    0.53716  -0.295 0.769810
## Pulse          0.11455    0.17043   0.672 0.506818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.655 on 29 degrees of freedom
## Multiple R-squared:  0.6674, Adjusted R-squared:  0.5641
## F-statistic: 6.465 on 9 and 29 DF,  p-value: 5.241e-05
```

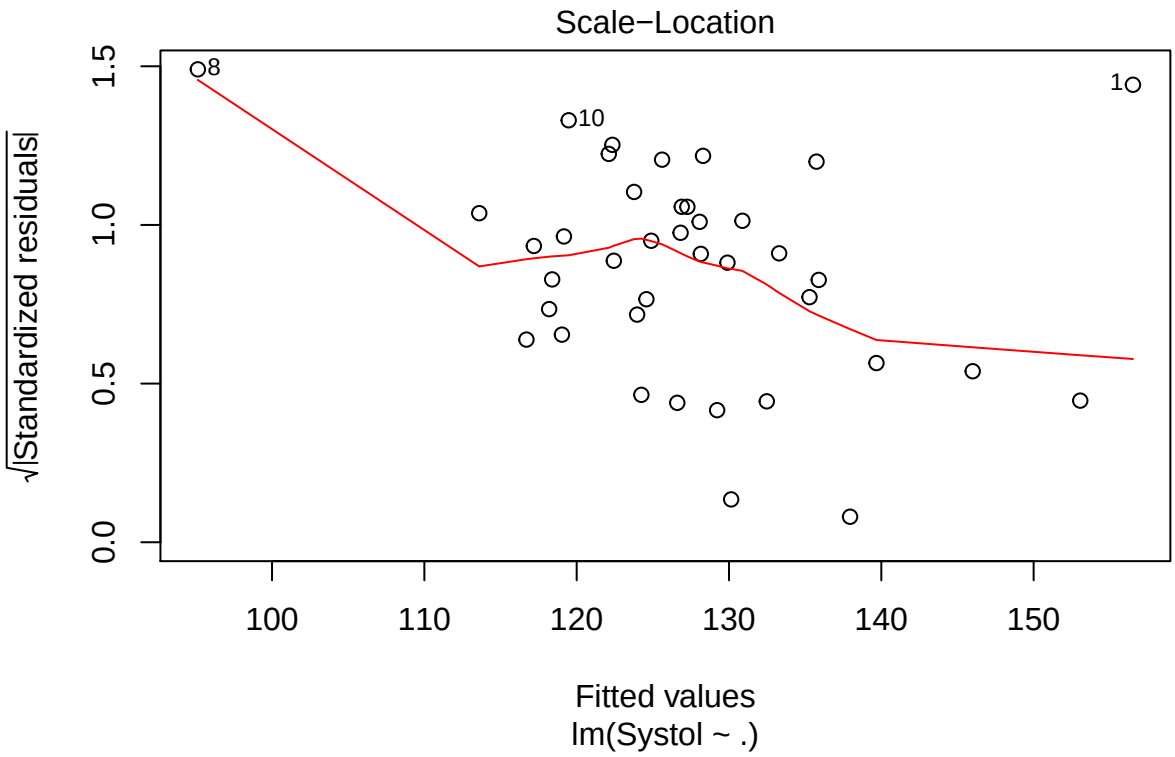
Tan solo parece que las variables *Age*, *Years*, *Fraction*, y *Weight* y su intercepto son significantes al rechazar la hipótesis nula del t-test (p-valor menor a 0.05) a la hora de ajustar el modelo. Además el coeficiente de determinación tan solo explica el 56% de la variación del modelo. observándose que existen pocas variables realmente significativas, sabiendo además que la variable *Fraction* procede de las variables *Age* y *Years*.

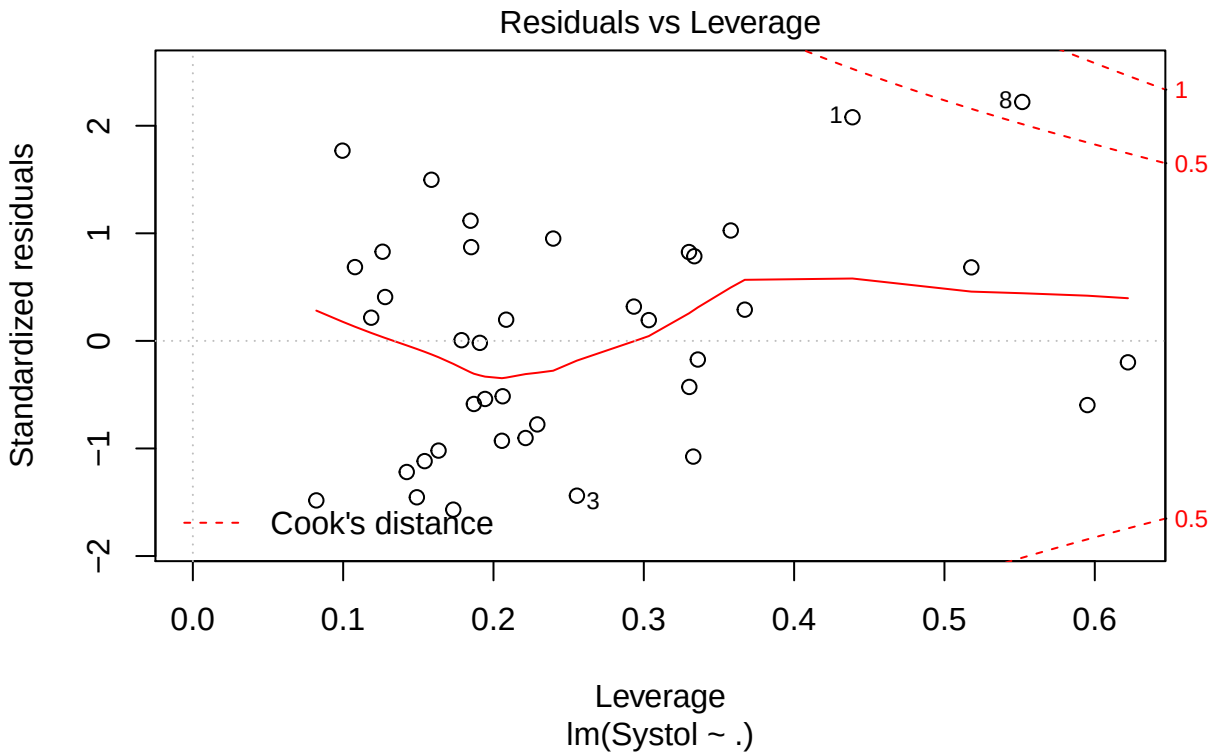
Para observar el ajuste del modelo se puede visualizar el resumen de gráficos del modelo:

```
plot(lm_peru)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Systol ~ .)
8
1
10
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
8
10
1
100
110
120
130
140
150
Fitted values
lm(Systol ~ .)



El gráfico de residuos contra valores ajustados se observa que aparte de algunos outliers los residuos se encuentran bien distribuidos. Observando la gráfica de normalidad se observa que los residuos siguen principalmente una distribución normal. La gráfica de los valores ajustados contra los residuos estandarizados indica que éstos no siguen un patrón o tendencia concreta, indicando que el modelo tiene homocedasticidad de la varianza. La última gráfica permite detectar los outliers que más afectan al ajuste del modelo, indicado con la distancia de Cook. En este caso la observación 8 es el outlier que más significativamente afecta al ajuste del modelo.

Para obtener pruebas de una posible multicolinealidad se realiza una tabla de correlaciones:

```
round(cor(peru),2)
```

```
##             Age Years Fraction Weight Height  Chin Forearm  Calf Pulse Systol
## Age        1.00  0.59     0.36   0.43   0.06  0.16    0.06 -0.01  0.09   0.01
## Years      0.59  1.00     0.94   0.48   0.07  0.22    0.14  0.00  0.24  -0.09
## Fraction   0.36  0.94     1.00   0.29   0.05  0.12    0.03 -0.11  0.21  -0.28
## Weight     0.43  0.48     0.29   1.00   0.45  0.56    0.54  0.39  0.31   0.52
## Height     0.06  0.07     0.05   0.45   1.00 -0.01   -0.07  0.00  0.01   0.22
## Chin       0.16  0.22     0.12   0.56  -0.01  1.00    0.64  0.52  0.22   0.17
## Forearm    0.06  0.14     0.03   0.54  -0.07  0.64    1.00  0.74  0.42   0.27
## Calf      -0.01  0.00    -0.11   0.39   0.00  0.52    0.74  1.00  0.21   0.25
## Pulse      0.09  0.24     0.21   0.31   0.01  0.22    0.42  0.21  1.00   0.14
## Systol     0.01 -0.09    -0.28   0.52   0.22  0.17    0.27  0.25  0.14   1.00
```

Se observa que hay una elevada cantidad de correlaciones. Por ejemplo, la variable *Years* está altamente correlacionada con *Fraction* y en cierta medida con *Age*. También se observan correlaciones positivas entre *Weight* con *Chin* y *Forearm*, entre los propios *Chin* y *Forearm*, y entre *Calf* y *Forearm*. Estas correlaciones tienen cierto sentido en la interpretación de los datos (a mayor peso, generalmente aumenta la cantidad de

piel en las distintas zonas donde se acumula grasa, como en el antebrazo o el mentón).

Ahora se realiza la decomposición de eigen de $X^T X$:

```
x <- model.matrix(lm_peru)
e <- eigen(t(x) %*% x)
e$val
```

```
##  [1] 9.774654e+07 6.059674e+03 3.264509e+03 1.358492e+03 1.140882e+03
##  [6] 3.974082e+02 1.410343e+02 4.889686e+01 8.368043e-02 3.103563e-02
```

```
sqrt(e$val[1]/e$val)
```

```
##  [1]     1.0000   127.0065   173.0381   268.2391   292.7051   495.9436
##  [7]   832.5080  1413.8724 34177.3789 56120.3576
```

Se observa que existen en el modelo unos valores elevados en distintos "eigenvalues." Esto implica que existe más de una combinación linear.

Ahora se comprueba el factor de inflación de la varianza (FIV):

```
car::vif(lm_peru)
```

```
##       Age     Years  Fraction    Weight    Height      Chin   Forearm      Calf
##  3.213372 34.289194 24.387468  4.747711  1.913991  2.063866  3.802313  2.414602
##     Pulse
##  1.329233
```

Se encuentra inflación elevada de la varianza, principalmente en las variables *Years* y *Fraction*, indicando así una muy elevada colinealidad.Además se encuentran las variables *Weight* y *Age*, que también tienen valor suficiente para indicar una posible colinealidad.

Por lo mencionado en los análisis de las variables hechos parece que sí que existe multicolinealidad.

(b) Eliminar una única observación de la muestra de forma que el modelo mejore apreciablemente. Razonar la elección.

La observación que más sentido tiene eliminar del modelo es la que se aleje más significativamente de otras observaciones (desviando la media y modificando el ajuste del modelo). Por lo tanto se busca el outlier más significativo:

```
outlierTest(lm_peru)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 8 2.395715           0.023515      0.91709
```

Se indica que el outlier significativo está producido por la observación 8. Por ello se ajusta un nuevo modelo sin esta observación para saber si se realiza un mejor ajuste.

```
peru_n <- peru[-8,]
lm_peru_n <- lm(Systol ~. , data = peru_n)
summary(lm_peru_n)
```

```
##
## Call:
## lm(formula = Systol ~ ., data = peru_n)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -12.1497  -4.1489  -0.2525   5.2688  16.7433
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  166.35727   46.12801   3.606 0.001194 **
## Age           -1.18974    0.30488  -3.902 0.000546 ***
## Years          3.02918    0.79227   3.823 0.000673 ***
## Fraction    -145.53620   30.68660  -4.743  5.6e-05 ***
## Weight         1.65662    0.41220   4.019 0.000399 ***
## Height        -0.05052    0.03481  -1.451 0.157876
## Chin          -0.94857    0.68696  -1.381 0.178257
## Forearm       -2.38282    1.21649  -1.959 0.060168 .
## Calf           0.38367    0.54705   0.701 0.488874
## Pulse          0.07024    0.15909   0.442 0.662228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.025 on 28 degrees of freedom
## Multiple R-squared:  0.7066, Adjusted R-squared:  0.6123
## F-statistic: 7.492 on 9 and 28 DF,  p-value: 1.696e-05
```

El coeficiente de determinación ha aumentado de 0.56 a 0.61 tan solo eliminando el outlier más significativo, además de aumentar el nivel de significación de las variables ya significativas previamente, por lo que parece una buena idea descartar esta observación para mejorar el ajuste del modelo.

(c) Con los 38 datos restantes, hallar el "mejor" modelo consensuado por dos métodos diferentes de selección de variables como, por ejemplo R^2_{adj} y C_p de Mallows

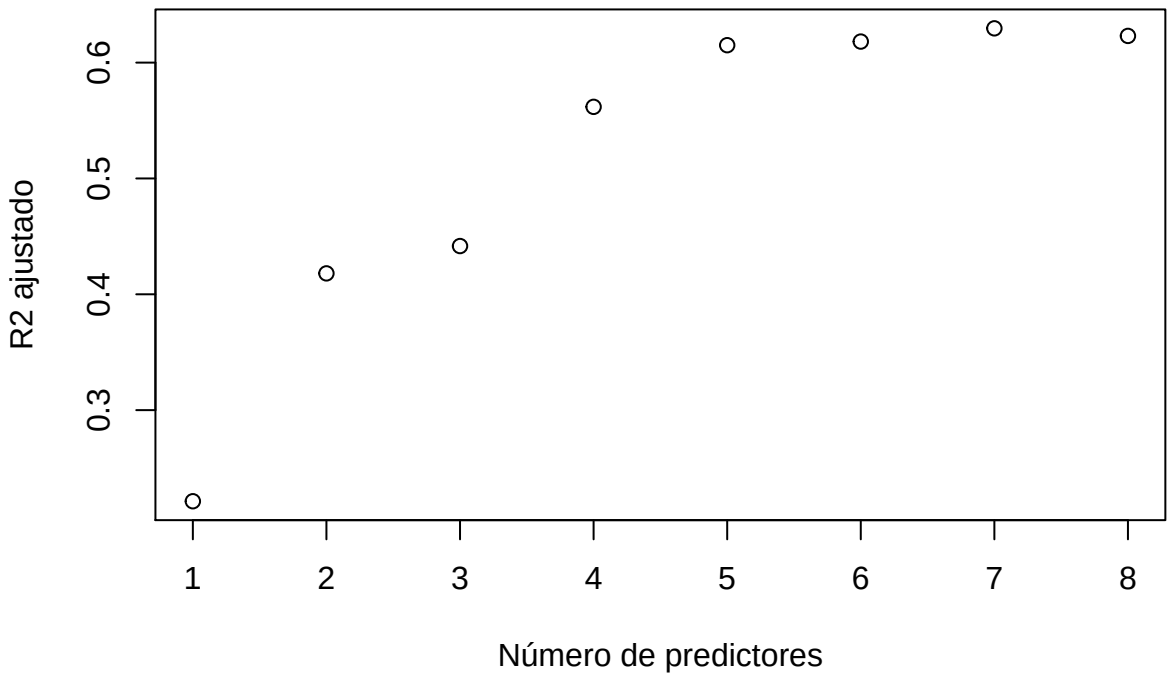
Se realiza el método de selección de variables de R^2_{adj}:

```
b <- regsubsets(Systol~., data=peru_n)
rs <- summary(b)
rs$adjr2
```

```
## [1] 0.2213449 0.4180425 0.4416393 0.5618138 0.6150372 0.6181600 0.6295699
## [8] 0.6230287
```

Se puede observar además gráficamente qué número de predictores es el que tiene un R^2_{adj} óptimo:

```
k <- length(rs$rss)
p <- k + 1
plot(1:k,rs$adjr2, xlab="Número de predictores",
     ylab="R2 ajustado")
```



Se puede observar que el coeficiente de determinación ajustado más elevado es el que contiene 7 predictores, aunque su valor entre 5 y 7 predictores no es muy distinto (de 0.61 a 0.63) disminuyendo su valor en los 8 predictores.

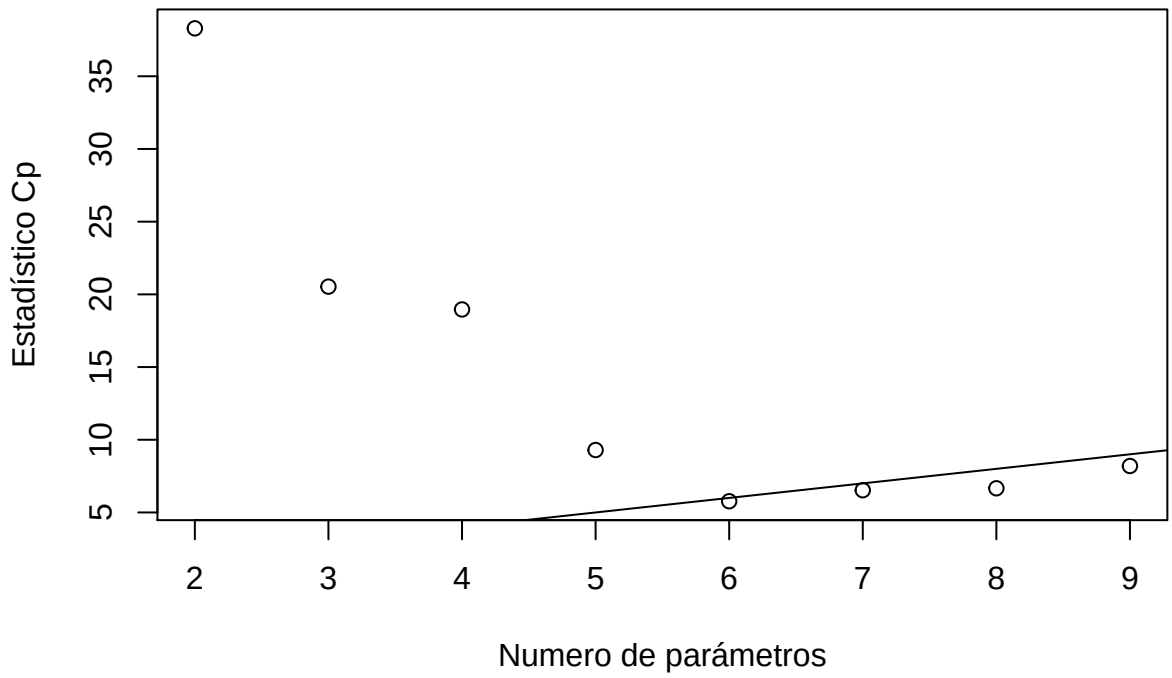
Ahora se realiza el método de selección de variables de C_p de Mallows:

```
rs$cp
```

```
## [1] 38.295709 20.532028 18.961941  9.293879  5.771204  6.528675  6.661067
## [8]  8.194940
```

También se puede observar con otro gráfico para facilitar la interpretación:

```
plot(2:p,rs$cp, xlab="Numero de parámetros",
     ylab="Estadístico Cp")
abline(a=0,b=1)
```



El mejor valor del estadístico C_p se encuentra en los 6 parámetros (7 predictores), concordando con el R^2_{adj} óptimo.

(i) ¿Cuáles son las variables seleccionadas?

Como en ambos modelos se prefiere el uso de los 7 mejores predictores, se descartan las variables menos significativas para el ajuste del modelo (los que tengan su p-valor más elevado). Para ello también podemos observar el RSS mínimo:

```
rs$which
```

```
##   (Intercept)   Age Years Fraction Weight Height  Chin Forearm  Calf Pulse
## 1        TRUE FALSE FALSE    FALSE   TRUE  FALSE FALSE   FALSE FALSE FALSE
## 2        TRUE FALSE FALSE     TRUE   TRUE  FALSE FALSE   FALSE FALSE FALSE
## 3        TRUE  TRUE  TRUE     TRUE  FALSE  FALSE FALSE   FALSE FALSE FALSE
## 4        TRUE  TRUE  TRUE     TRUE   TRUE  FALSE FALSE   FALSE FALSE FALSE
## 5        TRUE  TRUE  TRUE     TRUE   TRUE  FALSE FALSE    TRUE FALSE FALSE
## 6        TRUE  TRUE  TRUE     TRUE   TRUE   TRUE FALSE    TRUE FALSE FALSE
## 7        TRUE  TRUE  TRUE     TRUE   TRUE   TRUE  TRUE    TRUE FALSE FALSE
## 8        TRUE  TRUE  TRUE     TRUE   TRUE   TRUE  TRUE    TRUE  TRUE FALSE
```

Los predictores menos significativos entonces son *Pulse* y *Calf*, por lo que los descartamos para ajustar un nuevo modelo, y seleccionamos las restantes.

(ii) ¿Cual es el coeficiente de determinación ajustado de este modelo? Compararlo con el del modelo completo.

Para saber el coeficiente de determinación ajustado del nuevo modelo, se ajusta primero dicho modelo y se observa en el resumen el coeficiente de determinación ajustado:

```
lm_peru_adj <- lm(Systol~ Age + Years + Fraction + Weight + Height + Chin + Forearm,
                  data = peru_n)
summary(lm_peru_adj)
```

```
##
## Call:
## lm(formula = Systol ~ Age + Years + Fraction + Weight + Height +
##     Chin + Forearm, data = peru_n)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2689  -4.3720  -0.8238   4.8757  15.6880
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  167.48929   43.90880   3.814 0.000634 ***
## Age           -1.20029    0.29712  -4.040 0.000342 ***
## Years          3.02338    0.77342   3.909 0.000490 ***
## Fraction    -144.56746   29.93081  -4.830 3.76e-05 ***
## Weight         1.62411    0.39182   4.145 0.000256 ***
## Height        -0.04689    0.03337  -1.405 0.170248
## Chin          -0.92559    0.66201  -1.398 0.172317
## Forearm       -1.72543    0.87339  -1.976 0.057468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.843 on 30 degrees of freedom
## Multiple R-squared:  0.6997, Adjusted R-squared:  0.6296
## F-statistic: 9.983 on 7 and 30 DF,  p-value: 2.205e-06
```

El coeficiente de determinación ajustado en el modelo reducido es de 0.63 mientras que el del modelo completo es de 0.61, por lo que sí que explica un mayor porcentaje de la variación del modelo. Por ello, se consideraría mejor este modelo reducido (explica el mismo o mayor porcentaje de la variación del modelo con menos variables, siguiendo entonces el principio de la navaja de Occam).

(iii) ¿Se gana en eficiencia con el modelo reducido? Comparar los intervalos de confianza de la estimación del coeficiente de la variable *Age*.

Al tener el modelo ajustado un coeficiente de determinación similar o mejor que el original con un menor número de predictores, se supone que es por lo tanto más eficiente. Para saber si se gana eficiencia con el modelo reducido se observan los intervalos de confianza del coeficiente de la variable *Age* al 95%. De esta manera se puede observar la precisión de ambos modelos.

```
confint(lm_peru_n)
```

```
##                   2.5 %       97.5 %
## (Intercept)   71.8683281 260.84620905
## Age           -1.8142681  -0.56522175
## Years          1.4062872   4.65207036
## Fraction    -208.3948497 -82.67754163
## Weight         0.8122712   2.50097753
```

```
## Height        -0.1218283   0.02079582
## Chin          -2.3557476   0.45860437
## Forearm       -4.8746857   0.10904896
## Calf          -0.7369032   1.50424248
## Pulse         -0.2556352   0.39611568
```

```
confint(lm_peru_adj)
```

```
##                    2.5 %       97.5 %
## (Intercept)   77.8155617 257.16301928
## Age           -1.8071033  -0.59348604
## Years          1.4438452   4.60291451
## Fraction    -205.6943200 -83.44059009
## Weight         0.8238988   2.42432045
## Height        -0.1150347   0.02125967
## Chin          -2.2775941   0.42640652
## Forearm       -3.5091256   0.05827442
```

El intervalo de confianza de la variable *Age* en el modelo completo es de (-1.81,-0.56) y un coeficiente de -1.19 mientras que en el modelo reducido es de (-1.81,-0.59) con un coeficiente de -1.20. Esto indica que la precisión en la estimación del coeficiente de la variable *Age* es similar, por lo que el modelo reducido es más eficiente, ya que con una menor cantidad de variables se obtiene el mismo ajuste.

(d) Los investigadores sugieren adoptar el modelo reducido que contenga únicamente las variables significativas ($\alpha = 0.1$) con el test t en sustitución del modelo completo con las 9 variables explicativas. ¿Es ese un buen criterio de selección? Realizar un test adecuado que resuelva su sugerencia. Discutir el resultado en consonancia con los resultados obtenidos en el apartado anterior.

El criterio de selección comentado en el apartado en principio está bien fundamentado, ya que así tan solo se escogerían las variables más significativas para el ajuste del modelo. Algún problema que puede dar tal reducción es que disminuya el coeficiente de determinación al eliminar algunas variables que puedan explicar algún porcentaje de la varianza en la variable respuesta.

Para saber si la selección ha sido una buena idea, se puede realizar un análisis de la varianza o ANOVA para saber si ha habido diferencias significativas entre el modelo completo y el anidado. En este ANOVA la hipótesis nula es que no hay diferencias significativas enre los dos modelos, y la hipótesis alternativa que sí hay diferencias significativas entre los modelos. En este análisis consideramos $\alpha = 0.05$. Para realizar el ANOVA primero se requiere del ajuste un modelo con las variables con $\alpha = 0.1$. Cabe mencionar que se considera como modelo completo el que contiene 38 variables (al haber eliminado el outlier). Las variables del modelo con $\alpha < 0.1$ son *Age*, *Years*, *Fraction*, *Weight* y *Forearm*.

```
lm_peru_sig <- lm(Systol~Age + Years + Fraction + Weight + Forearm,
                  data = peru_n)
```

```
anova(lm_peru_n,lm_peru_sig)
```

```
## Analysis of Variance Table
##
## Model 1: Systol ~ Age + Years + Fraction + Weight + Height + Chin + Forearm +
##     Calf + Pulse
## Model 2: Systol ~ Age + Years + Fraction + Weight + Forearm
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 1803.0
## 2     32 2045.8 -4   -242.84 0.9428 0.4539
```

El análisis de varianza indica que no hay diferencias significativas entre los dos modelos (p-valor > 0.05). Esto

indica que la sugerencia de los investigadores es acertada, ya que siguiendo el principio de la navaja de Occam es preferible usar el modelo más simple, siendo en este caso el modelo reducido. Además se puede observar que las variables utilizadas son las 5 variables explicativas seleccionadas en el apartado (d) de este ejercicio.

(e) Comprobar si hemos solucionado el problema de multicolinealidad en el modelo reducido del apartado anterior. Como los investigadores no quieren prescindir de más variables, se plantea una regresión Partial Least Squares (PLS). ¿Cuantas componentes se necesitan para minimizar el RMSEP? Calcular los coeficientes de las variables originales, también para β_0, que proporciona este método con el número de componentes necesario. ¿Es adecuado este método de regresión con estas variables? ¿Es útil?

Para observar el problema de multicolinealidad se realizan los mismos análisis que en el apartado (a) pero con el modelo reducido:

Para saber si existe multicolinealidad en el modelo, se puede observar el resumen de éste:

```
summary(lm_peru_sig)
```

```
##
## Call:
## lm(formula = Systol ~ Age + Years + Fraction + Weight + Forearm,
##     data = peru_n)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.162  -5.498   0.333   5.539  15.894
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  116.6799    20.0133   5.830 1.78e-06 ***
## Age           -1.2035     0.3024  -3.980 0.000371 ***
## Years          3.2951     0.7724   4.266 0.000165 ***
## Fraction    -153.2570    30.1142  -5.089 1.53e-05 ***
## Weight         1.1624     0.2817   4.126 0.000245 ***
## Forearm       -1.7233     0.7307  -2.358 0.024623 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.996 on 32 degrees of freedom
## Multiple R-squared:  0.6671, Adjusted R-squared:  0.615
## F-statistic: 12.82 on 5 and 32 DF,  p-value: 6.991e-07
```

En este caso todas las variables (como cabe esperar debido a lo hecho en el anterior apartado) son significativas para el ajuste del modelo, siendo *Forearm* la menos significativa de éstas.

Para obtener pruebas de una posible multicolinealidad se realiza una tabla de correlaciones con las variables del modelo:

```
round(cor(peru_n[,c(1,2,3,4,7,10)]),2)
```

```
##             Age Years Fraction Weight Forearm Systol
## Age        1.00  0.64     0.47   0.41    0.05  -0.04
## Years      0.64  1.00     0.96   0.54    0.15  -0.05
## Fraction   0.47  0.96     1.00   0.42    0.05  -0.21
## Weight     0.41  0.54     0.42   1.00    0.55   0.49
```

```
## Forearm   0.05  0.15     0.05   0.55    1.00   0.27
## Systol   -0.04 -0.05    -0.21   0.49    0.27   1.00
```

Se sigue observando la misma correlación positiva entre *Age* y *Years* y *Fraction* con las dos variables anteriores. En comparación con el modelo del apartado (a) ya no se encuentran las correlaciones que aparecían entre las otras variables (debido a que ya no están las variables en sí) por lo que la multicolinealidad se ha visto considerablemente reducida.

Ahora se realiza la decomposición de eigen de $X^T X$:

```
x <- model.matrix(lm_peru_sig)
e <- eigen(t(x) %*% x)
e$val
```

```
## [1] 2.167253e+05 3.191569e+03 1.100315e+03 1.423289e+02 3.495301e-01
## [6] 5.681601e-02
```

```
sqrt(e$val[1]/e$val)
```

```
## [1]    1.000000    8.240483   14.034478   39.021891  787.431012 1953.077392
```

Se observa que siguen presentes valores elevados en distintos "eigenvalues." Esto implica que sigue existiendo más de una combinación linear.

Ahora comprobamos el factor de inflación de la varianza (FIV):

```
car::vif(lm_peru_sig)
```

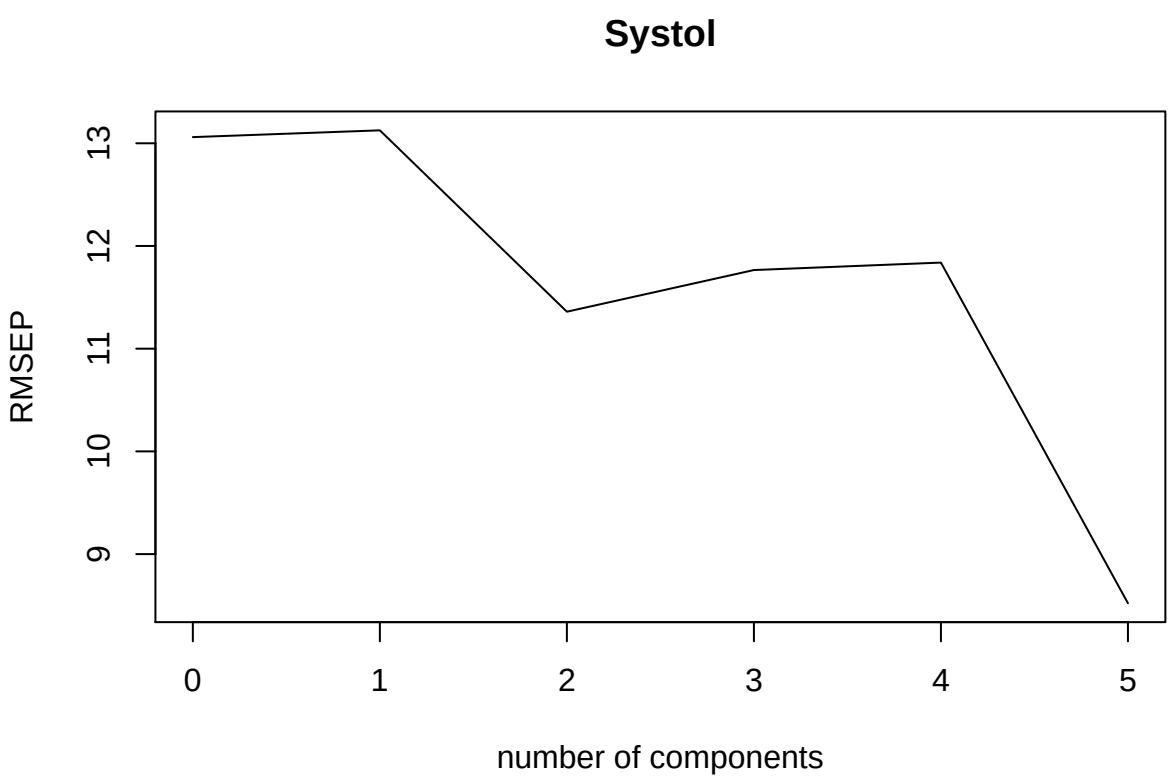
```
##       Age     Years  Fraction    Weight   Forearm
##  3.105001 35.118905 24.738540  2.245658  1.665539
```

Se sigue encontrando una inflación elevada de la varianza en las variables *Years* y *Fraction* indicando la colinealidad por lo que se ha podido observar en el análisis de colinealidad del modelo reducido, que sigue existiendo multicolinealidad.

Como se indica en el apartado, se pretende realizar una regresión PLS para evitar prescindir de más variables. Primero se ajusta el modelo y se utiliza la validación cruzada (crossvalidation) para determinar el número de componentes en la predicción:

```
set.seed(111)
pls_peru <- plsr(Systol~Age + Years + Fraction + Weight + Forearm,
                 data = peru_n, validation = "CV")
plsCV <- RMSEP(pls_peru, estimate = "CV")
```

```
plot(plsCV)
```




```
which.min(plsCV$val)
```

```
## [1] 6
```

Según los resultados obtenidos en la regresión PLS, se necesitan de 6 componentes (el intercepto y los 5 predictores) para obtener el el RMSEP mínimo posible.

Para obtener los coeficientes de las variables originales utilizando los 6 componentes se extraen los valores del modelo realizado con los componentes necesarios:

```
coef(pls_peru, intercept = TRUE)
```

```
## , , 5 comps
##
##                  Systol
## (Intercept)  116.679857
## Age           -1.203490
## Years          3.295147
## Fraction    -153.256987
## Weight         1.162357
## Forearm       -1.723303
```

Se observa que los coeficientes de regresión utilizando el método PLS son los mismos (o similares) que en el modelo reducido, por lo que el método PLS no parece que haya sido de utilidad a la hora de reducir el modelo. Al seguir utilizando todos los componentes del modelo equivalentes al intercepto y los predictores, no ha habido cambios significativos, por lo que no se considera útil este método de regresión. Aun así, el método PLS se suele utilizar en modelos que tienen una alta correlación, como es el caso del modelo utilizado, así que el planteamiento de su uso en este modelo es adecuado.

(f) Siguiendo con el modelo reducido, otra posibilidad es utilizar la Ridge Regression. ¿Cuales son los coeficientes obtenidos? Explicar brevemente las ventajas e inconvenientes de este método frente a la selección de variables. Calcular el RMSE de la regresión OLS, PLS (con 5, 4, 3 y 2 componentes) y Ridge (con λ óptima por GCV) para el modelo reducido.

Se ajusta el modelo de Ridge Regression con las variables utilizadas en el modelo reducido:

```
rg_peru <- lm.ridge(Systol~Age + Years + Fraction + Weight + Forearm,
                    data = peru_n, lambda=(seq(0,50,0.001)))
```

De esta manera se pueden observar los coeficientes del modelo de la misma manera que en el modelo PLS:

```
head(coef(rg_peru))
```

```
##                       Age    Years  Fraction   Weight   Forearm
##  0.000 116.6799 -1.203490 3.295147 -153.2570 1.162357 -1.723303
##  0.001 116.5783 -1.201845 3.289482 -153.0367 1.162866 -1.721601
##  0.002 116.4770 -1.200204 3.283835 -152.8171 1.163373 -1.719904
##  0.003 116.3761 -1.198569 3.278206 -152.5982 1.163878 -1.718213
##  0.004 116.2755 -1.196939 3.272595 -152.3800 1.164381 -1.716526
##  0.005 116.1752 -1.195314 3.267002 -152.1625 1.164882 -1.714844
```

Los coeficientes para cada variable son los mismos (o similares) a los del modelo reducido, pero van variando conforme se modifica λ. Por ello es recomendable encontrar el valor óptimo de λ.

```
nGCV <- which.min(rg_peru$GCV)
lGCV <- rg_peru$lambda[nGCV]
lGCV
```

```
## [1] 0.033
```

Por lo tanto con este nuevo valor se realiza de nuevo el modelo con el valor de λ óptimo:

```
rg_peru_opt <- lm.ridge(Systol~Age + Years + Fraction + Weight + Forearm,
                        data = peru_n, lambda= lGCV)
coef(rg_peru_opt, intercept = TRUE)
```

```
##                     Age       Years    Fraction      Weight     Forearm
##  113.497230   -1.151803    3.117327 -146.340388    1.178175   -1.669631
```

La ventaja de este método es que permite reducir los modelos teniendo en cuenta las multicolinealidad entre las distintas variables, para así ajustarlos de manera más eficiente y obtener una reducción de la varianza. Sin embargo, esto va ligado a la principal desventaja del método. Ésta es que los modelos ajustados con este método tienen coeficientes sesgados. La presencia de coeficientes sesgados en el modelo implica que en algunos casos las predicciones realizadas con un modelo ajustado mediante éste método se encuentre alejada de la realidad.

Ahora se van a utilizar los distintos modelos obtenidos para calcular su respectivo RMSE:

Para calcular el RMSE de la regresión OLS se utiliza la misma ecuación que la que ya se utilizó en la PEC anterior:

```
sqrt(mean(lm_peru_sig$residuals^2))
```

```
## [1] 7.337436
```

Para calcular el RMSE de la regresión utilizando el método PLS (con 5, 4, 3 y 2 componentes respectivamente) se puede utilizar la función `RMSEP()`:

```
RMSEP(pls_peru, ncomp = 2:5, estimate = "CV", intercept = FALSE)
```

```
## 2 comps  3 comps  4 comps  5 comps
##  11.360   11.766   11.838    8.522
```

En el caso de Ridge Regression al utilizar el paquete MASS no existe una función de predicción, por lo que se realizará un ajuste a partir de otra función del paquete caret:

```
rg_peru_opt_fit <- train(Systol~Age + Years + Fraction + Weight + Forearm,
                         data = peru_n, method = "ridge")
rg_peru_opt_fit$results
```

```
##   lambda      RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
## 1  0e+00  9.291219 0.5837182 7.608259 2.132085  0.1666872 1.845144
## 2  1e-04  9.289245 0.5832246 7.605752 2.117720  0.1669625 1.831983
## 3  1e-01 11.227632 0.3584770 8.993151 2.789252  0.2001160 2.072087
```

```
rg_peru_opt_fit$bestTune
```

```
##   lambda
## 2  1e-04
```

El RMSE con el mejor valor (lambda óptimo) usando ridge regression es 9.3.

(g) Sabemos que el RMSE calculado en un modelo para todos los datos observados es muy optimista. Es mejor un cálculo por validación cruzada. Con el modelo reducido de los apartados anteriores y para comparar los métodos estudiados OLS, PLS (con 4 componentes) y Ridge (con λ óptimo por GCV) haremos lo siguiente:

Se ajusta cada modelo con los parámetros requeridos a partir del train y test que se piden, obteniendo finalmente un conjunto de 1000 valores de RMSE para cada método:

```
repeticiones_lm <- rep(0,1000)
repeticiones_pls <- rep(0,1000)
repeticiones_ridge <- rep(0,1000)
for (i in seq(1:1000)) {
  sample <- sample.int(nrow(peru_n), size = 8, replace = F)
  test_peru <- peru_n[sample, ]
  train_peru <- peru_n[-sample,]


  lm_peru_sig_t <- lm(Systol~Age + Years + Fraction + Weight + Forearm,
                      data = train_peru)
  pls_peru_t <- plsr(Systol~Age + Years + Fraction + Weight + Forearm,
                     data = train_peru, validation = "CV",ncomp = 4)
  rg_peru_t <- lm.ridge(Systol~Age + Years + Fraction + Weight + Forearm,
                        data = train_peru, lambda=(seq(0,50,0.001)))
  nGCV_t <- which.min(rg_peru_t$GCV)
  lGCV_t <- rg_peru_t$lambda[nGCV]
  rg_peru_t_lgcv <-train(Systol~Age + Years + Fraction + Weight + Forearm,
                         data = train_peru, method = "ridge")


  prediccion_lm <- predict(lm_peru_sig_t,
                           newdata = test_peru, type = "response")
```

```
  prediccion_pls <- predict(pls_peru_t,
                            newdata = test_peru, type = "response")
  prediccion_ridge <- predict(rg_peru_t_lgcv,
                              newdata = test_peru)

  rmse_lm_peru <- RMSE(prediccion_lm,test_peru$Systol)
  rmse_pls_peru <- RMSE(prediccion_pls,test_peru$Systol)
  rmse_ridge_peru <- RMSE(prediccion_ridge, test_peru$Systol)
  repeticiones_lm[i] <- rmse_lm_peru
  repeticiones_pls[i] <- rmse_pls_peru
  repeticiones_ridge[i] <- rmse_ridge_peru
}
```

Con los valores de todos los rmse de cada método se puede obtener una media para obtener así un rmse estimado:

```
mean(repeticiones_lm)
```

```
## [1] 8.597375
```

```
mean(repeticiones_pls)
```

```
## [1] 11.90945
```

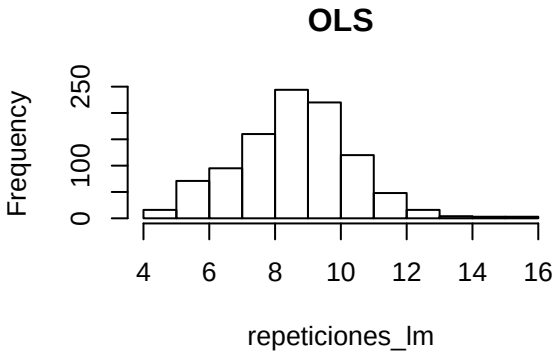
```
mean(repeticiones_ridge)
```

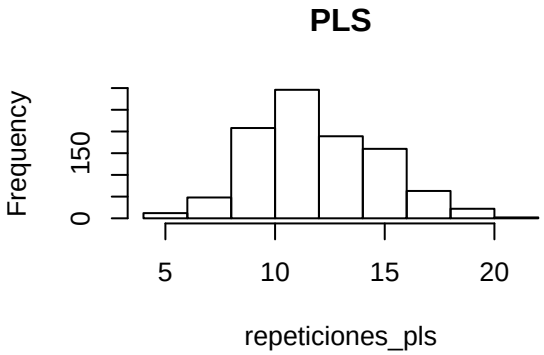
```
## [1] 8.640714
```

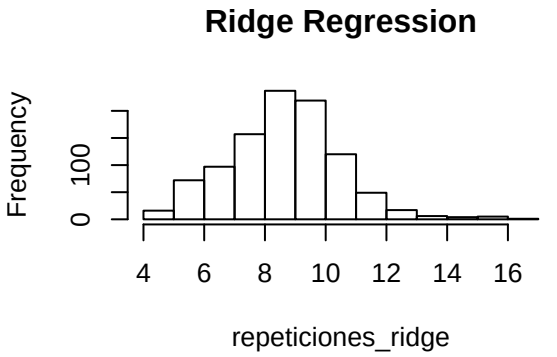
Por lo tanto el valor medio del rmse para OLS es 8.46, para el método PLS 11.91 y para ridge 8.49 pareciendo ser el mejor método (el que tiene menor RMSE) es el OLS.

Se puede observar en un histograma la distribución de los resultados en cada método:

```
par(mfrow=c(2,2))
hist(repeticiones_lm, main = "OLS")
hist(repeticiones_pls, main = "PLS")
hist(repeticiones_ridge, main = "Ridge Regression")
```



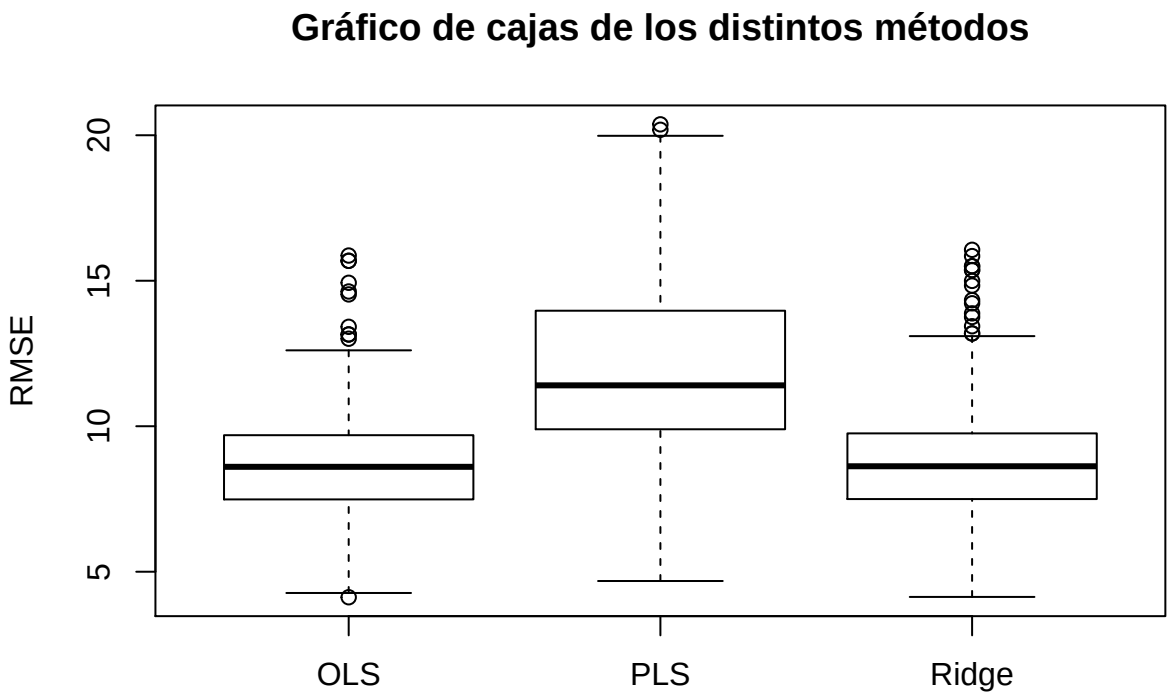




En los tres histogramas se observa que los tres métodos siguen una distribución normal.

Para comparar gráficamente cúal de los métodos es más preciso a la hora de realizar las predicciones, un buen gráfico es el de cajas:

```
boxplot(repeticiones_lm, repeticiones_pls, repeticiones_ridge, main = "Gráfico de cajas de los distintos
        names = c("OLS", "PLS", "Ridge"))
```



Se observa que el método que tiene un menor RMSE y una media con una menor desviación estándar y por lo tanto más precisión en las 1000 repeticiones está igualado entre el método OLS y Ridge. En cambio el método PLS tiene una distribución más amplia de resultados con una media sobrepasando 10 de RMSE, implicando que este método es peor para ajustar el modelo del ejercicio.

(h) Calcular los grados de libertad de la Ridge regression para el λ óptimo del apartado (e)

No he encontrado la manera de solucionar este apartado.

Ejercicio 2

Se lee la tabla adjunta en el ejercicio:

```
cancer <- read.delim("T33.1", sep = "", header = FALSE)
```

Posteriormente se eliminan las columnas que no aportan información (las tres primeras) y se nombran las columnas según los nombres que se introdujeron en el artículo:

```
cancer <- cancer[,c(-1:-3)]
colnames(cancer) <- c("Case", "Sex", "Age", "A", "B", "C", "D")
```

También se eliminan los símbolos + que se encuentran en las columnas A y C:

```
# Se introduce as.integer para que no se interpreten estas columnas como "character", complicando poste
cancer$A <- as.integer(gsub('\\+', '', cancer$A))
cancer$C <- as.integer(gsub('\\+', '', cancer$C))
```

Se crea un vector con los tipos de cáncer ordenados respectivamente conforme se encuentran en la tabla y se añaden como una nueva columna:

```
type <- c(rep("Stomach", 13), rep("Bronchus", 17), rep("Colon", 17), rep("Rectum", 7),
          rep("Ovary", 6), rep("Breast", 11), rep("Bladder", 7), rep("Kidney", 8),
          rep("Gallbladder", 2), rep("Esophagus", 2), rep("Reticulum cell sarcoma", 2),
          rep("Prostate", 2), "Uterus", "Brain", rep("Pancreas",3),
          "Chronic lymphatic leukemia")

cancer$Type <- type
```

(a) Estudiar la transformación que mejora la distribución de los datos C y los datos D (100 observaciones en cada caso). Se puede utilizar el método de Box-Cox. Una vez transformados, comparar si el tiempo de supervivencia C es superior al de los controles D con todas las observaciones.

Se utiliza el método de transformación Box-Cox para saber la mejor transformación tanto para los datos de C como de D:

```
BoxCoxTrans(cancer$C)
```

```
## Box-Cox Transformation
##
## 100 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.0    57.5   136.5   293.3   338.2  2270.0
##
## Largest/Smallest: 284
## Sample Skewness: 2.78
##
## Estimated Lambda: 0
## With fudge factor, Lambda = 0 will be used for transformations
```

```
BoxCoxTrans(cancer$D)
```

```
## Box-Cox Transformation
##
## 100 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   19.75   32.00   37.79   52.75  129.00
##
## Largest/Smallest: 12.9
## Sample Skewness: 1.15
##
## Estimated Lambda: 0
## With fudge factor, Lambda = 0 will be used for transformations
```

Ambas variables tienen un valor estimado de $\lambda = 0$ por lo que lo recomendable es realizar una transformación logarítmica de estas dos variables.

```
cancer$C <- log(cancer$C)
cancer$D <- log(cancer$D)
```

Ahora se realiza una comparación en todas las observaciones que indicará como TRUE si la observación de C es mayor a la de D y viceversa. Se puede observar en una tabla:

```
compare <- ifelse(cancer$C > cancer$D, TRUE, FALSE)
table(compare)
```

```
## compare
## FALSE  TRUE
##    13    87
```

Se ha encontrado que hay 13 observaciones donde el tiempo de supervivencia de los controles D es superior a la de C.

(b) Ahora estamos interesados en comparar la mejora en función del tipo de cáncer. Nos centraremos exclusivamente en los tres tipos de cáncer de la tabla 1 de más arriba y no tendremos en cuenta el sexo... Calcular los elementos de dicha tabla con la matriz de diseño X de este modelo y resolver con ellos el contraste $H_0 : \mu_1 = \mu_2 = \mu_3$ cuando la variable respuesta Y es el logaritmo de la razón entre la supervivencia de los tratados y la supervivencia de los controles. ¿Cual es la conclusión?

Como sólo se van a utilizar los 3 tipos de cáncer de la tabla 1, se crea una nueva tabla con tan solo esos tipos de cáncer:

```
cancer_tabla <- cancer[cancer$Type == "Stomach" | cancer$Type == "Bronchus" | cancer$Type == "Colon",]
```

La matriz de diseño del modelo mostrado es el mismo que el de la "one-way anova" o anova de un factor, por lo que pasamos la variable *Type* a categórica (factor) para así poder utilizarla en el formato correcto.

```
cancer_tabla$Type <- as.factor(cancer_tabla$Type)
```

```
cancer_tabla$C <- exp(cancer_tabla$C)
cancer_tabla$D <- exp(cancer_tabla$D)
```

Con los datos facilitados se encuentra la matriz del modelo sin intercepto:

```
head(model.matrix(log(C / D)~ Type -1, data = cancer_tabla))
```

```
##   TypeBronchus TypeColon TypeStomach
## 1            0         0           1
## 2            0         0           1
## 3            0         0           1
## 4            0         0           1
## 5            0         0           1
## 6            0         0           1
```

A partir de esta matriz se puede realizar el "one-way anova" o la anova de un factor. De esta manera se contrasta la hipótesis nula en que la media de los factores (tipos de cáncer) son iguales entre ellas (con $\alpha = 0.05$). Además cabe tener en cuenta que se ha de eliminar el intercepto (como indica el apartado):

```
lm_cancer_tabla <- lm(log(C / D)~ Type -1,
                      data = cancer_tabla)
summary(lm_cancer_tabla)
```

```
##
## Call:
## lm(formula = log(C/D) ~ Type - 1, data = cancer_tabla)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51899 -0.98435 -0.03999  0.74344  2.57376
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## TypeBronchus   1.3080     0.2993   4.370 7.48e-05 ***
## TypeColon      2.0182     0.2993   6.742 2.73e-08 ***
## TypeStomach    1.2572     0.3423   3.673 0.000647 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.234 on 44 degrees of freedom
## Multiple R-squared:  0.6395, Adjusted R-squared:  0.6149
## F-statistic: 26.01 on 3 and 44 DF,  p-value: 7.793e-10
```

Según el resumen del modelo todos los tipos de cáncer son significantes (p-valor <0.05) a la hora de ajustar el modelo, siendo el modelo significativo (p-valor < 0.05).

Ahora se comprueba la normalidad y homogeneidad del modelo. Para la normalidad se realiza un test de Shapiro-Wilk, donde la hipótesis nula es que los residuos del modelo siguen una distribución normal (con $\alpha = 0.05$):

```
shapiro.test(lm_cancer_tabla$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm_cancer_tabla$residuals
## W = 0.98128, p-value = 0.6459
```

Se confirma entonces (con un p-valor > 0.05) la hipótesis nula, es decir, que el modelo sigue una distribución normal.
Para estudiar la homogenidad del modelo se realiza un test de Bartlett, donde la hipótesis nula (con $\alpha = 0.05$) es que el modelo tiene homogeneidad de varianza:

```
bartlett.test(log(C / D)~ Type -1,
              data = cancer_tabla)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  log(C/D) by Type
## Bartlett's K-squared = 0.21532, df = 2, p-value = 0.8979
```

Se confirma (con un p-valor > 0.05) la hipótesis nula, por lo que el modelo tiene homogeneidad de varianza.

Ahora se realiza un análisis de varianza del modelo, donde se contrasta la hipótesis nula (con $\alpha = 0.05$) en que no hay diferencias significativas entre las medias de supervivencia entre distintos tipos de cáncer:

```
anova_cancer <- anova(lm_cancer_tabla)
anova_cancer
```

```
## Analysis of Variance Table
##
## Response: log(C/D)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Type       3 118.876  39.625  26.015 7.793e-10 ***
## Residuals 44  67.021   1.523
```

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

El p-valor del contraste de hipótesis (p-valor < 0.05) rechaza la hipótesis nula, indicando que la media de supervivencia es diferente en almenos un tipo de cáncer. Se encuentra que el factor tiene 3 grados de libertad y 44 los grados de libertad del error (residuales), la media de cuadrados del grupo es de 39.625 y la del error (residuales) 1.523, la suma de cuadrados del grupo es 118.88, y la de los residuos 67.02, y la suma de cuadrados total (SStotal) no se encuentra en la tabla pero se puede calcular, ya que es la suma de cuadrados del grupo con el del error (SSE):

```
sstotal <- 118.876 + 67.021
sstotal
```

```
## [1] 185.897
```

Por lo que el SStotal es de 185.897. Si se quiere calcular el valor de F manualmente se divide la media de la suma de cuadrados del grupo entre la media de la suma de cuadrados del error:

```
F <- 39.625 / 1.523
F
```

```
## [1] 26.01773
```

En resumen, la conclusión que se puede sacar de este apartado es que existen diferencias significativas entre las medias de los distintos grupos de cáncer.

(c) La edad de los pacientes presenta una cierta variabilidad y puede influir en su supervivencia. Añadir a la matriz X del apartado anterior el vector columna con las edades centradas. Utilizar las sumas de cuadrados de los residuos de este modelo y del anterior para contrastar la importancia de ajustar con la edad. ¿Se puede utilizar un test t de Student?

Para saber si la edad de los pacientes puede influir a la supervivencia, se ajusta un nuevo modelo añadiendo las edades mostradas en la tabla centradas. Se compara el modelo obtenido con el del anterior apartado para saber si el modelo cambia significativamente al añadir la edad. Se indicará como hipótesis nula (con $\alpha = 0.05$) que ambos modelos no tienen diferencias significativas entre ellos.

```
lm_cancer_tabla_edad <- lm(log(C / D)~ Type + scale(Age, center = TRUE, scale = FALSE) -1,
                           data = cancer_tabla)
anova(lm_cancer_tabla, lm_cancer_tabla_edad)
```

```
## Analysis of Variance Table
##
## Model 1: log(C/D) ~ Type - 1
## Model 2: log(C/D) ~ Type + scale(Age, center = TRUE, scale = FALSE) -
##     1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     44 67.021
## 2     43 66.716  1   0.30501 0.1966 0.6597
```

Se observa que el nuevo modelo sigue indicando que se acepta la hipótesis nula (p-valor > 0.05), por lo que se acepta que los dos modelos no tienen diferencias significativas entre ellos. Esto implica que la variable *Age* no parece tener un efecto significativo en el ajuste del modelo.
Además se puede obtener la suma de cuadrados del nuevo modelo mediante un análisis de varianza de éste:

```
anova(lm_cancer_tabla_edad)
```

```
## Analysis of Variance Table
##
```

```
## Response: log(C/D)
##                                          Df  Sum Sq Mean Sq F value    Pr(>F)
## Type                                      3 118.876  39.625 25.5395 1.207e-09
## scale(Age, center = TRUE, scale = FALSE)  1   0.305   0.305  0.1966    0.6597
## Residuals                                43  66.716   1.552
##
## Type                                     ***
## scale(Age, center = TRUE, scale = FALSE)
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede observar que este modelo tiene una suma de cuadrados de los residuos más baja que el modelo ajustado en el apartado (b) respecto a los residuos, (siendo 66.71 mientras que en el modelo del anterior apartado era de 67.02), mientras que respecto a la la suma de cuadrados del grupo es similar entre los dos modelos. Esto puede indicar que la variable *Age* mejora en cierta medida el ajuste del modelo (ya que su introducción al modelo disminuye su RMSE) aunque ésta variable no sea significativa.

Se puede utilizar el test t de student para observar si la edad es significativa para el ajuste del modelo (con $\alpha = 0.05$), observable en el resumen de éste:

```
summary(lm_cancer_tabla_edad)
```

```
##
## Call:
## lm(formula = log(C/D) ~ Type + scale(Age, center = TRUE, scale = FALSE) -
##     1, data = cancer_tabla)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6250 -0.9079 -0.0989  0.7922  2.5282
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## TypeBronchus                             1.297381   0.303055   4.281 0.000102
## TypeColon                                2.022655   0.302269   6.692  3.6e-08
## TypeStomach                              1.265273   0.345951   3.657 0.000690
## scale(Age, center = TRUE, scale = FALSE) 0.008012   0.018070   0.443 0.659713
##
## TypeBronchus                             ***
## TypeColon                                ***
## TypeStomach                              ***
## scale(Age, center = TRUE, scale = FALSE)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.246 on 43 degrees of freedom
## Multiple R-squared:  0.6411, Adjusted R-squared:  0.6077
## F-statistic:  19.2 on 4 and 43 DF,  p-value: 3.994e-09
```

Se indica mediante el p-valor del test t (p-valor > 0.05) que la variable *Age* no afecta significativamente al ajuste del modelo, por lo que la variable no parece ser importante en el modelo.

(d) Aunque la regresión de la edad en el modelo anterior pudiera no ser importante, se decidió que cada grupo debería tener su propia regresión sobre la edad para verificar si la edad no es importante en niguno de los grupos. Modificar adecuadamente la matriz de diseño para acomodar esta nueva situación y completar el test para la hipótesis nula de que la regresión sobre la edad es la misma en los tres grupos de cáncer. ¿Cual es la conclusión?

Se ajustan los nuevos modelos con lo comentado en el enunciado:

```
cancer_tabla_stomach <- subset(cancer_tabla, Type == "Stomach")
cancer_tabla_bronchus <- subset(cancer_tabla, Type == "Bronchus")
cancer_tabla_colon <- subset(cancer_tabla, Type == "Colon")

lm_cancer_tabla_stomach <- lm(log(C / D) ~ scale(Age, center = TRUE, scale = FALSE) -1,
                        data = cancer_tabla_stomach)
lm_cancer_tabla_bronchus <- lm(log(C / D) ~ scale(Age, center = TRUE, scale = FALSE) -1,
                        data = cancer_tabla_bronchus)
lm_cancer_tabla_colon <- lm(log(C / D) ~ scale(Age, center = TRUE, scale = FALSE) -1,
                        data = cancer_tabla_colon)
```

```
summary(lm_cancer_tabla_stomach)
```

```
##
## Call:
## lm(formula = log(C/D) ~ scale(Age, center = TRUE, scale = FALSE) -
##     1, data = cancer_tabla_stomach)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6224  0.0155  1.2874  2.1252  3.7583
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## scale(Age, center = TRUE, scale = FALSE)  0.01275    0.06001   0.213    0.835
##
## Residual standard error: 1.833 on 12 degrees of freedom
## Multiple R-squared:  0.00375,    Adjusted R-squared:  -0.07927
## F-statistic: 0.04517 on 1 and 12 DF,  p-value: 0.8353
```

```
summary(lm_cancer_tabla_bronchus)
```

```
##
## Call:
## lm(formula = log(C/D) ~ scale(Age, center = TRUE, scale = FALSE) -
##     1, data = cancer_tabla_bronchus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2260  0.3265  1.2086  2.0190  3.8279
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## scale(Age, center = TRUE, scale = FALSE)  0.00808    0.04138   0.195    0.848
##
## Residual standard error: 1.771 on 16 degrees of freedom
```

```
## Multiple R-squared:  0.002377,   Adjusted R-squared:  -0.05997
## F-statistic: 0.03813 on 1 and 16 DF,  p-value: 0.8476
```

```
summary(lm_cancer_tabla_colon)
```

```
##
## Call:
## lm(formula = log(C/D) ~ scale(Age, center = TRUE, scale = FALSE) -
##     1, data = cancer_tabla_colon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5765  1.2845  1.7332  3.0785  4.2812
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## scale(Age, center = TRUE, scale = FALSE) 0.005723   0.054684   0.105    0.918
##
## Residual standard error: 2.438 on 16 degrees of freedom
## Multiple R-squared:  0.0006841,  Adjusted R-squared:  -0.06177
## F-statistic: 0.01095 on 1 and 16 DF,  p-value: 0.9179
```

Se observa en los tres modelos que la edad no mejora el ajuste del modelo significativamente, ya que se acepta la hipótesis nula del test t realizado en cada modelo (p-valor > 0.05).

Ahora se realiza un test anova, donde se indica como hipótesis nula en el modelo ajustado del apartado (c) (con $\alpha = 0.05$) que las medias de los tres grupos de cáncer sobre la edad son iguales. Teniendo en cuenta esto, se realiza el análisis de varianza:

```
anova(lm_cancer_tabla_edad)
```

```
## Analysis of Variance Table
##
## Response: log(C/D)
##                                         Df  Sum Sq Mean Sq F value    Pr(>F)
## Type                                     3 118.876  39.625 25.5395 1.207e-09
## scale(Age, center = TRUE, scale = FALSE)  1   0.305   0.305  0.1966    0.6597
## Residuals                                43  66.716   1.552
##
## Type                                     ***
## scale(Age, center = TRUE, scale = FALSE)
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El test F indica que se acepta la hipótesis nula (p-valor > 0.05) para la edad, demostrando que la regresión sobre la edad es la misma en los tres grupos de cáncer.

Ejercicio 3

Se leen los datos adjuntos al ejercicio para poder utilizarlos en R:

```
diabetes <- read.csv("diabetes.txt", header = TRUE)
```

(a) Ajustar un modelo de regresión logística para predecir la diabetes utilizando todas las otras variables como predictoras. Dar la ecuación del modelo obtenido y clasificar las variables según sean factores protectores o de riesgo para la diabetes.

Se ajusta el modelo de regresión logística como indica el apartado:

```
logit_diabetes <- glm(relevel(diabetes, ref = "neg") ~ .,
                      data = diabetes, family = "binomial")
summary(logit_diabetes)
```

```
##
## Call:
## glm(formula = relevel(diabetes, ref = "neg") ~ ., family = "binomial",
##     data = diabetes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -0.6603  -0.3642   0.6409   2.5612
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## pressure    -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## mass         7.054e-02  2.734e-02   2.580  0.00989 **
## pedigree     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

Se ha de tener en cuenta que la variable *diabetes* se considera como un valor binomial . Se ha ordenado dicho factor indicando el primer nivel como "neg" significando que no se ha desarrollado diabetes y el segundo nivel como "pos", indicando que sí se ha desarrollado diabetes.

El resumen del modelo indica mediante la diferencia entre la desviación nula y la residual que este modelo tiene un buen ajuste.

Para observar más fácilmente los coeficientes obtenidos, se redondean:

```
round(logit_diabetes$coefficients, 3)
```

```
## (Intercept)    pregnant     glucose    pressure     triceps     insulin
##     -10.041       0.082       0.038      -0.001       0.011      -0.001
##        mass    pedigree         age
##       0.071       1.141       0.034
```

La ecuación del modelo obtenido (sin tener en cuenta que parte de las variables del modelo no son significativas para el ajuste del modelo) al basarse en una regresión logística, se basa en la probabilidad de aparición de diabetes (por lo que la respuesta obtenida es categórica, es decir, diabetes o no diabetes) entendiendo entonces que la ecuación cambia respecto a los modelos de regresión lineal realizados con anterioridad en este trabajo. La ecuación obtenida en este caso es:

$$p = \frac{e^{-10.041+0.082*pregnant+0.038*glucose-0.001*pressure+0.011*triceps-0.001*insulin+0.071*mass+1.141*pedigree+0.034*age}}{1+e^{-10.041+0.082*pregnant+0.038*glucose-0.001*pressure+0.011*triceps-0.001*insulin+0.071*mass+1.141*pedigree+0.034*age}}$$

Como se ha comentado anteriormente, al ser el cálculo de una probabilidad de obtener diabetes y al ser 0 el no tener diabetes y 1 tener diabetes, todos los predictores que disminuyan esa probabilidad se considerarán factores protectores para la diabetes. Consecuentemente todos los predictores que aumenten la probabilidad en la ecuación se considerarán factores de riesgo para la diabetes.
Con lo observado previamente, se consideran factores protectores para la diabetes las variables *pressure* y *insulin*, mientras que los factores de riesgo para la diabetes son *pregnant*, *glucose*, *triceps*, *mass*, *pedigree* y *age*. Cabe recalcar que algunas de las variables del modelo no son significativas para su ajuste (concretamente *pregnant*, *pressure*, *triceps*, *insulin* y *age*)

(b) Calcular el odds ratio de la variable *pedigree*, así como su intervalo de confianza.

Los odds ratios se pueden obtener fácilmente a partir de los coeficientes calculados del modelo. Estos coeficientes son los log odds, por lo que realizando su exponente se obtiene el odds ratio:

```
round(exp(logit_diabetes$coefficients),2)
```

```
## (Intercept)    pregnant     glucose    pressure     triceps     insulin
##        0.00        1.09        1.04        1.00        1.01        1.00
##        mass    pedigree         age
##        1.07        3.13        1.03
```

Por lo tanto el odds ratio de *pedigree* es de 3.13.

Para calcular el intervalo de confianza:

```
exp(cbind(coef(logit_diabetes), confint(logit_diabetes)))
```

```
## Waiting for profiling to be done...
```

```
##                                2.5 %       97.5 %
## (Intercept) 4.358754e-05 3.548295e-06 0.0004258459
## pregnant    1.085629e+00 9.743237e-01 1.2116311454
## glucose     1.039011e+00 1.027717e+00 1.0513035403
## pressure    9.985807e-01 9.757909e-01 1.0223068780
## triceps     1.011285e+00 9.778466e-01 1.0457799522
## insulin     9.991750e-01 9.966180e-01 1.0017675218
## mass        1.073085e+00 1.017827e+00 1.1335373213
## pedigree    3.129611e+00 1.378380e+00 7.3682727463
## age         1.034535e+00 9.985446e-01 1.0735228530
```

Se observa que el intervalo de confianza del odd ratio en la variable *pedigree* es (1.38,7.37).

(c) Calcular el odds ratio y la probabilidad de tener diabetes para el individuo de la observación 9

Se obtienen los valores de la observación 9:

```r
diabetes_c <- diabetes[9,]
```

Se realiza la predicción del individuo:

```r
predict_c <- predict(logit_diabetes,
                     newdata = diabetes_c, type = "response")
predict_c
```

```
##         9
## 0.2194284
```

Por lo tanto la probabilidad de que el individuo 9 tenga diabetes es del 22% (redondeado). Para obtener el odds ratio se divide la probabilidad por uno menos la probabilidad:

```r
odds_d <- (predict_c)/(1 - predict_c)
odds_d
```

```
##         9
## 0.2811125
```

Por lo tanto el ratio de la probabilidad (o odds ratio) de que el individuo de la observación 9 sea diabético es de 0.28.

(d) ¿Como valoras la bondad de ajuste del modelo? Realizar los contrastes o cálculos que se consideren necesarios.

Para valorar la bondad del ajuste se utiliza el test de bondad de Hosmer-Lemeshow. En este caso el test tiene como hipótesis nula que no hay evidencia de una mala bondad de ajuste en el modelo (con $\alpha = 0.05$):

```r
# Se obtienen los valores necesarios a partir de los valores ajustados. Si la probabilidad es menor a 0
# se indica entonces como "no diabetes" y si es mayor a 0.5 como "sí diabetes".
predicciones <- ifelse(test = logit_diabetes$fitted.values > 0.5, yes = 1, no = 0)
hoslem.test(logit_diabetes$y, predicciones)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  logit_diabetes$y, predicciones
## X-squared = 9.6002, df = 8, p-value = 0.2942
```

Se observa que se acepta la hipótesis nula (p-valor > 0.05) indicando que no hay evidencias significativas que indiquen que este modelo no esté correctamente ajustado.

(e) Considerar ahora el modelo reducido con las variables *pregnant*, *glucose*, *mass*, *pedigree* y *age*. ¿Es significativa la variable *pregnant*? Comparar los dos modelos.

Se ajusta un nuevo modelo de regresión logístico con las variables que se piden en el apartado:

```r
logit_diabetes_e <- glm(relevel(diabetes, ref = "neg") ~ pregnant + glucose + mass + pedigree + age,
                        data = diabetes, family = "binomial")
summary(logit_diabetes_e)
```

```
##
## Call:
## glm(formula = relevel(diabetes, ref = "neg") ~ pregnant + glucose +
##     mass + pedigree + age, family = "binomial", data = diabetes)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8827  -0.6535  -0.3694   0.6521   2.5814
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## pregnant     0.083953   0.055031   1.526 0.127117
## glucose      0.036458   0.004978   7.324 2.41e-13 ***
## mass         0.078139   0.020605   3.792 0.000149 ***
## pedigree     1.150913   0.424242   2.713 0.006670 **
## age          0.034360   0.017810   1.929 0.053692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
##
## Number of Fisher Scoring iterations: 5
```

La variable *pregnant* según el resumen del modelo indica que se acepta la hipótesis nula (p-valor > 0.05), por lo que implica que esta variable no afecta significativamente al ajuste del modelo.

Para comparar los dos modelos, al ser modelos anidados, se puede realizar un análisis de la varianza entre los dos modelos. El contraste de la hipótesis consiste en que la hipótesis nula (con $\alpha = 0.05$) es que no existen diferencias significativas entre estos dos modelos.

```
anova(logit_diabetes_e,logit_diabetes,test='LRT')
```

```
## Analysis of Deviance Table
##
## Model 1: relevel(diabetes, ref = "neg") ~ pregnant + glucose + mass +
##     pedigree + age
## Model 2: relevel(diabetes, ref = "neg") ~ pregnant + glucose + pressure +
##     triceps + insulin + mass + pedigree + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       386     344.89
## 2       383     344.02  3   0.8639   0.8341
```

Se observa que se acepta la hipótesis nula (p-valor > 0.05), por lo que se considera que estos dos modelos tienen un ajuste similar. Siguiendo el principio de la navaja de Occam se utilizaría el modelo más simple, siendo el modelo ajustado en este apartado.