A3: Modelización predictiva

Estadística Avanzada, Universitat Oberta de Catalunya

Paula Muñoz Lago

10 diciembre 2019

Contents

1. Modelo de regresión lineal

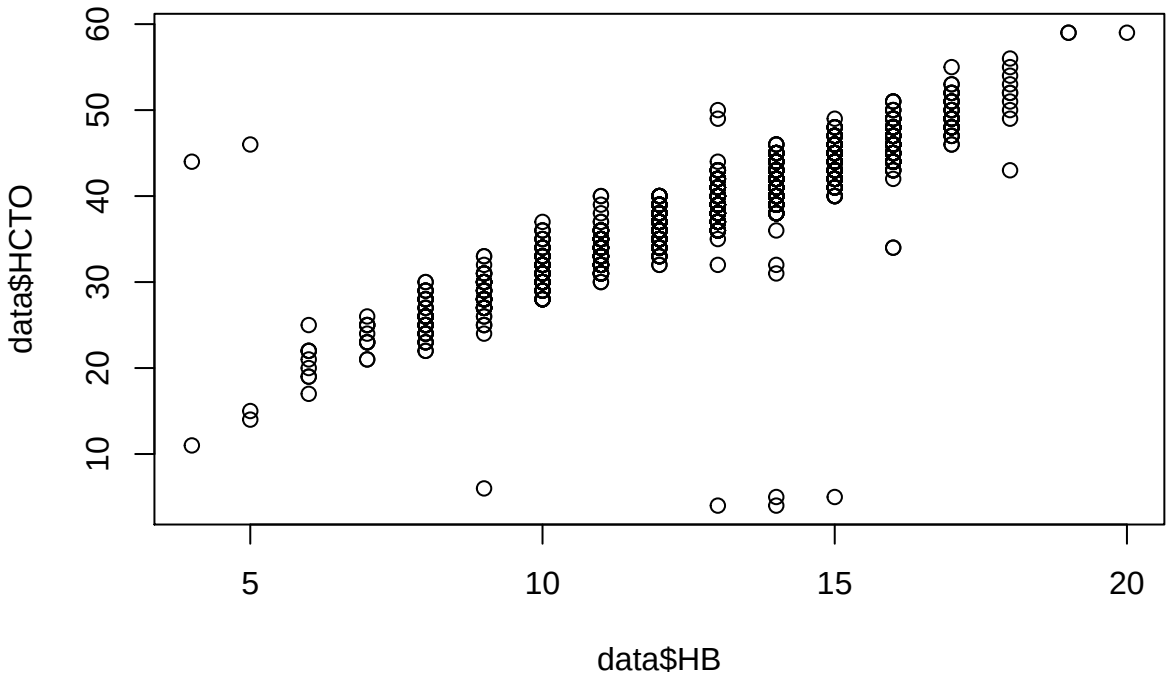
1.1. Modelo de regresión lineal simple

a) Toda la muestra

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable hematocrito en función de la hemoglobina. Evaluar la bondad de ajuste a través del coeficiente de determinación (R2). Podéis usar la instrucción de R lm.

Antes de comenzar, observamos el diagrama de dispersión para las variables Hematocrito (x) y Hemoglobina (y). Como se puede ver, ambas variables se encuentran sobre una recta, aunque no se ajustan perfectamente, con pendiente positiva, que india que a medida que los Hematocritos aumentan, también lo hace la hemoglobina. Además, destacamos la presencia de algunos outliers.

```r
plot(data$HB, data$HCTO)
```



Una vez estudiado el diagrama de dispersión y comprobado que existe una relación, procedemos a encontrar la ecuación de la recta que mejor se ajuste a la nube de puntos. La llamaremos la recta de regresión y la buscaremos con el método de regresión lineal de mínimos cuadrados ordinarios, que consiste en encontrar una relación entre dos variables en un plano y minimizar sus residuos, que son la diferencia entre el valor real y el valor predicho.

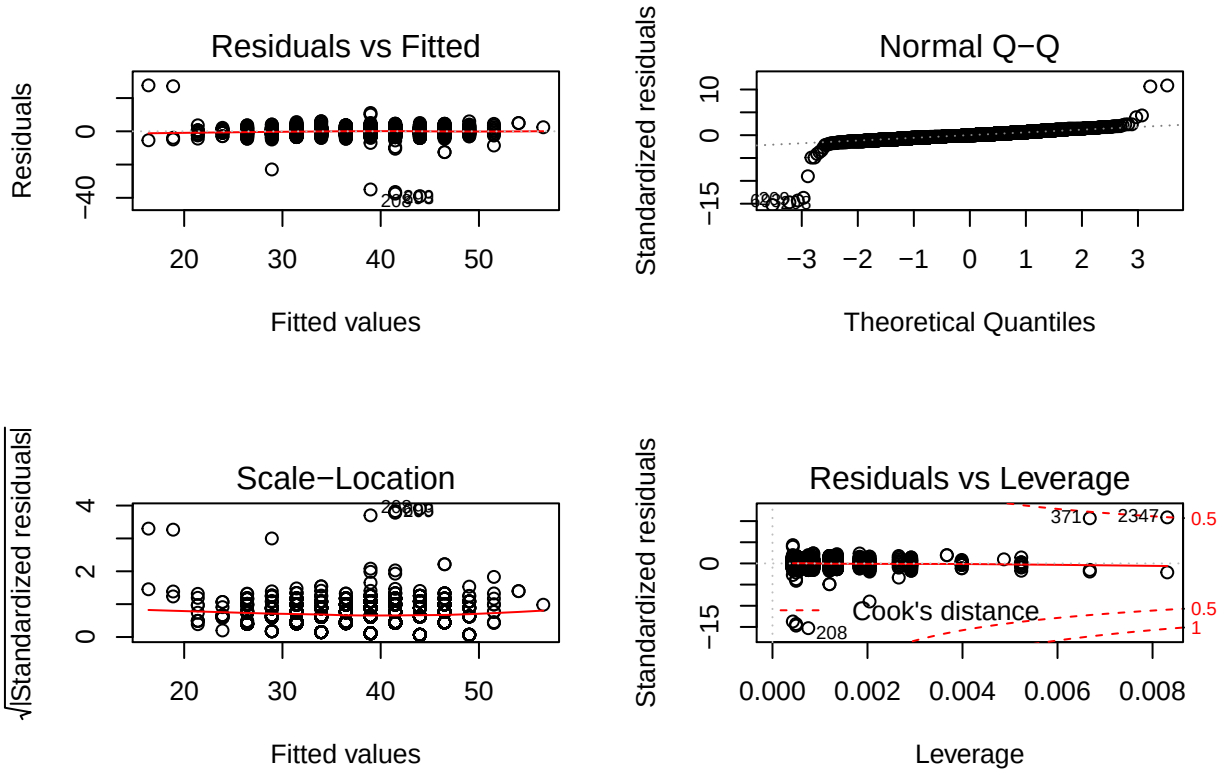
```r
model <- lm(formula=data$HCTO ~ data$HB)
summary(model)
```

```
##
## Call:
## lm(formula = data$HCTO ~ data$HB)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.989  -0.989   0.034   1.056  27.636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.31870    0.32991   19.15   <2e-16 ***
## data$HB      2.51136    0.02479  101.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.551 on 2341 degrees of freedom
```

```
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.8143, Adjusted R-squared:  0.8142
## F-statistic: 1.026e+04 on 1 and 2341 DF,  p-value: < 2.2e-16
```

Una vez construido el modelo, procedemos a visualizar los residuos generados.

```
par(mfrow = c(2, 2))
plot(model)
```



A continuación, evaluaremos la bondad del ajuste basándonos en el coeficiente de determinación (R2). Éste nos indica el grado de ajuste de la recta de regresión a los valores de la muestra.

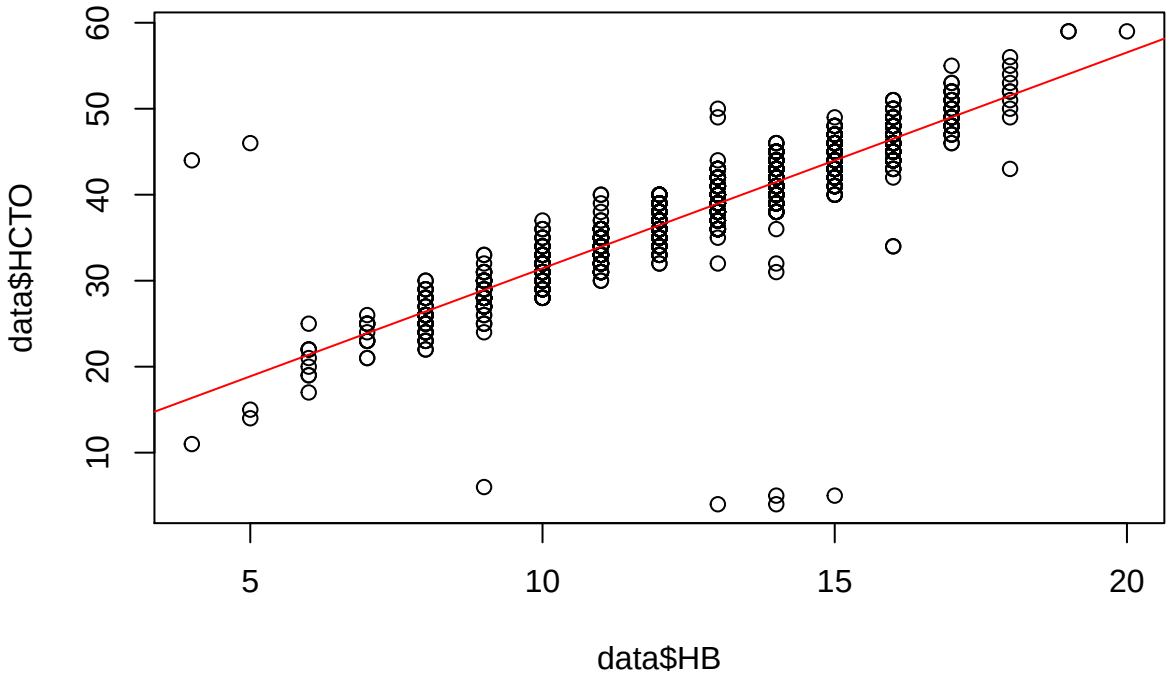
```
r2 <- summary(model)$r.squared
r2
```

```
## [1] 0.8142837
```

En este caso, el coeficiente de determinación es 0.8142. Dado que r2 = 1 indica un ajuste perfecto, es decrir, todos los puntos se encuentran sobre la recta de regresión, mientras que r2 = 0 denota la falta de relación entre los Hematocritos y la Hemoglobina (X e Y respectivamente). En este caso, dado que r2 se encuentra más cerca de 1, concluimos que existe una fuerte relación entre ambas variables, sin llegar a ser un ajuste perfecto.

La recta de regresión lineal obtenida es la siguiente.

```
plot(data$HB, data$HCTO)
abline(model, col = "red")
```



b) División de la muestra en dos

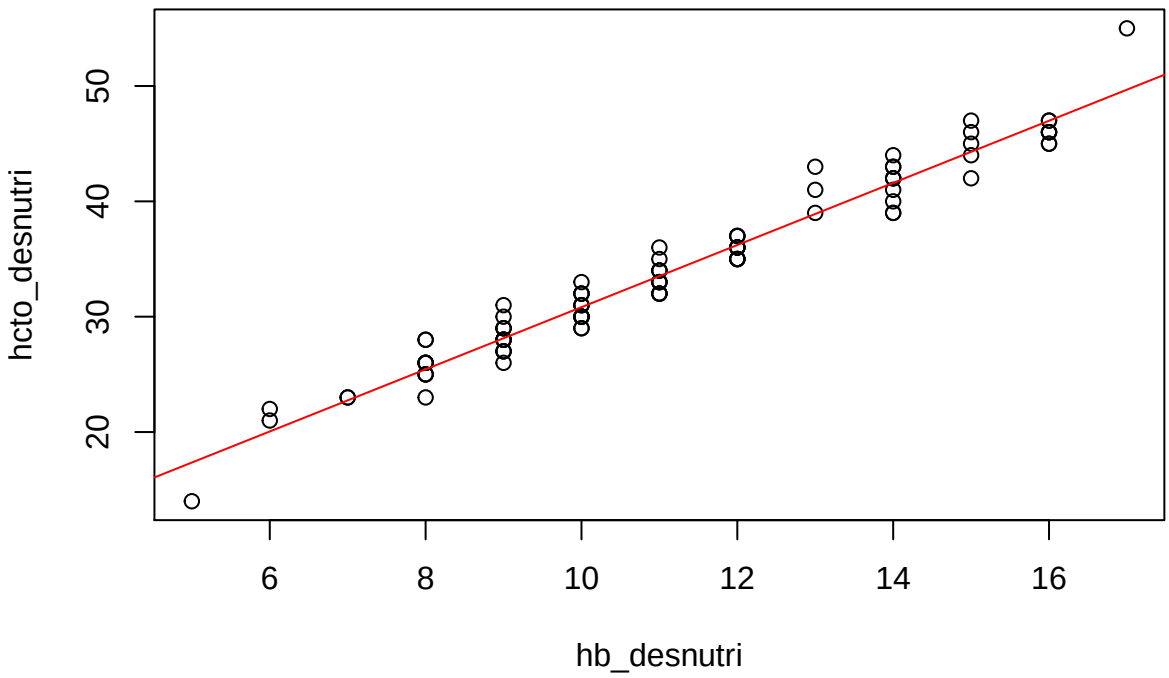
Algunos estudios afirman que la relación calculada anteriormente varía según la persona esté en condiciones óptimas de salud o no. Para contestar a esta pregunta, se dividirá la muestra en dos, según si la persona presenta desnutrición o no. Posteriormente se repetirá el estudio para cada muestra por separado. A partir de los resultados del modelo lineal en cada una de las muestras, ¿se puede tomar como cierta dicha conclusión? Justificar la respuesta.

En primer lugar, estudiaremos la relación entre os Hematocritos y la Hemoglobina en personas que presentan desnutrición. Repetiremos el proceso anterior: tras la división de la muestra, observaremos el diagrama de dispersión de ambas variables, que, como anteriormente, se muestran relaionadas en el eje positivo. A continuación, ejecutaremos el modelo y extraeremos la variable R2.

```
hcto_desnutri <- data$HCTO[which(data$DESNUTR == "si")]
hb_desnutri <- data$HB[which(data$DESNUTR == "si")]
model_desnutri <- lm(hcto_desnutri ~ hb_desnutri)
r2_desnutri <-summary(model_desnutri)$r.squared
r2_desnutri
```

```
## [1] 0.9603207
```

```
plot(hb_desnutri, hcto_desnutri)
abline(model_desnutri, col = "red")
```



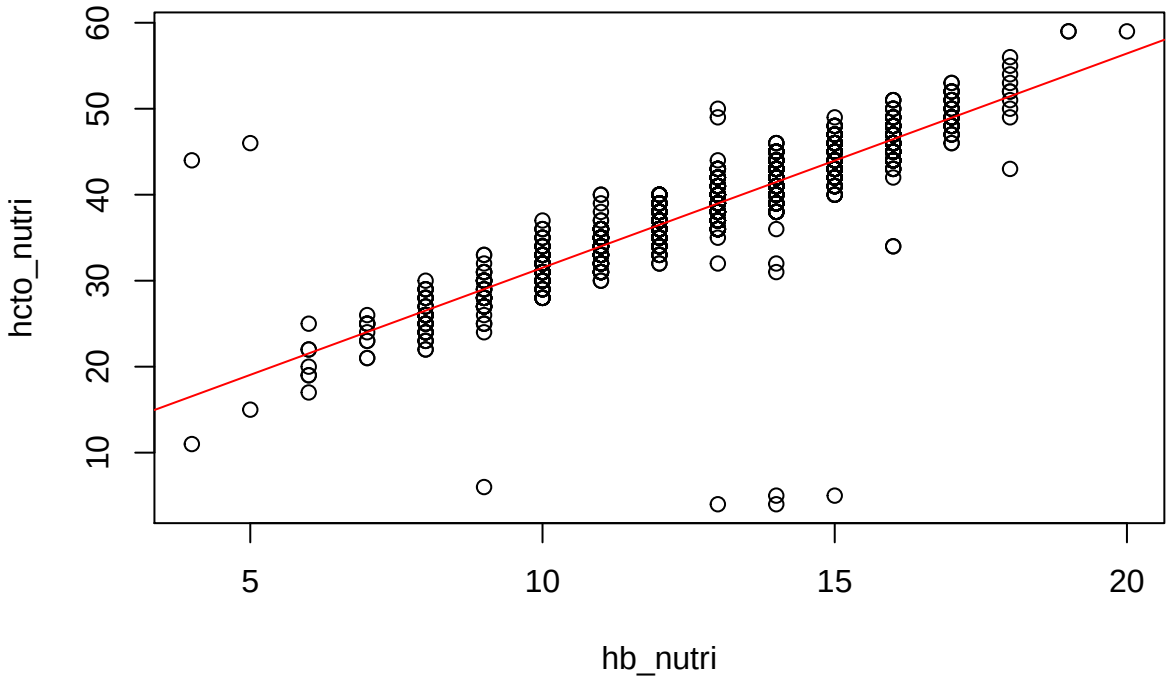
Dado que la variable R2 es muy cercana a 1, podemos concluir que ambas variables están fuertemente relacionadas en casos de desnutrición.

A continuación, extraeremos del conjunto de datos únicamente a los individuos que no presentan desnutrición en el momento del estudio. En este caso, la gráfica de dispersión muestra los valores algo más distanciados de la recta central, aunque sigue podiendose distinguir claramente la tendencia a que cuando aumentan los Hematocritos, aumenta también la Hemoglobina.

```
hcto_nutri <- data$HCTO[which(data$DESNUTR == "no")]
hb_nutri <- data$HB[which(data$DESNUTR == "no")]
model_nutri <- lm(hcto_nutri ~ hb_nutri)
r2_nutri <-summary(model_nutri)$r.squared
r2_nutri
```

```
## [1] 0.7966207
```

```
plot(hb_nutri, hcto_nutri)
abline(model_nutri, col = "red")
```



Como pudimos preveer en la gráfica de dispersión, los valores X e Y en el caso de individuos que no presentan desnutrición están menos relacionados, hemos podido comprobar esta sospecha gracias al valor r2_nutri, que es inferior a r2_desnutri pero igualmente está lo suficientemente cerca de 1 como para concluir que también existe una relación entre ambas variables. Es por ello que concluimos que independientemente de la condición nutricional del individuo, el nivel de Hematocritos y Hemoblobina estarán relacionados en mayor o menor medida.

1.2. Regresión linear múltiple (regresores cuantitativos)

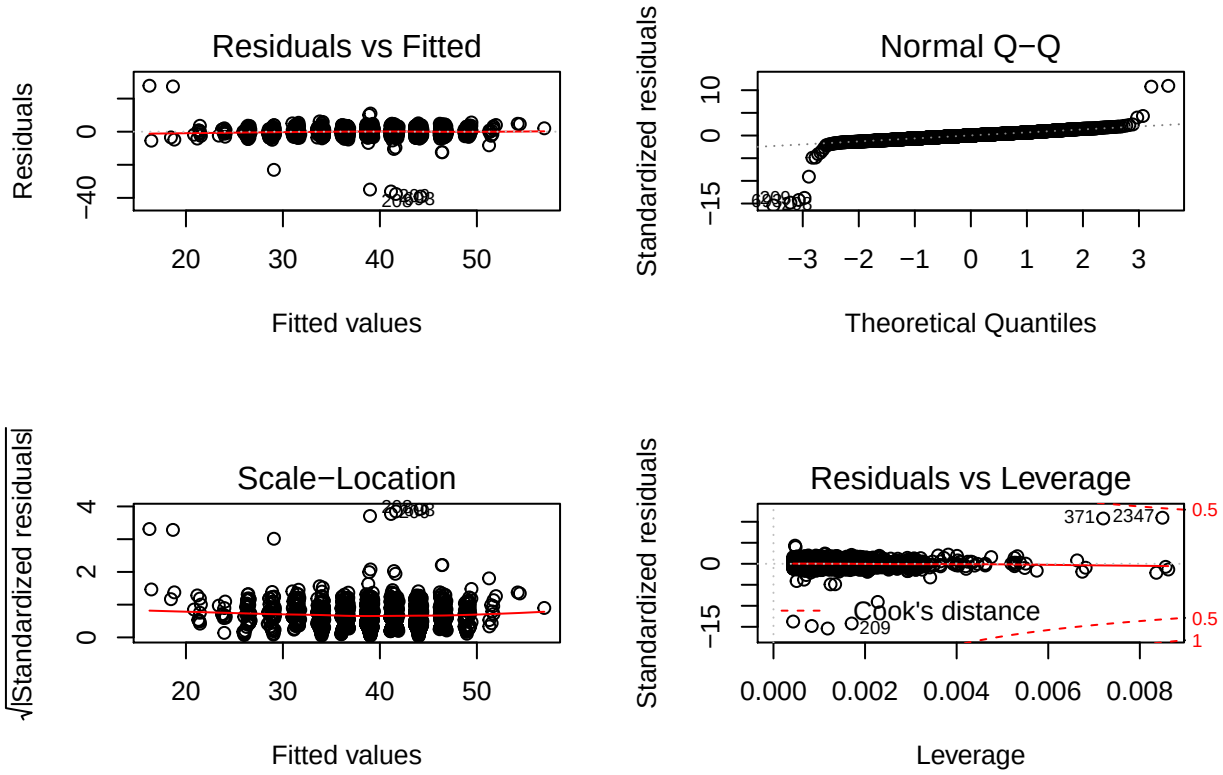
Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable hematocrito en función de la hemoglobina y la edad. Evaluar la bondad del ajuste y comparar el resultado con el obtenido en el apartado 1.1.a). Podéis usar la instrucción de R lm y usar el coeficiente R-cuadrado ajustado en la comparación. Interpretar también el significado de los coeficientes obtenidos y su significación estadística.

```
model_multiple <- lm(HCTO ~ HB + EDAD, data=data)
summary(model_multiple)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.188  -1.075   0.006   1.196  27.762
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.469021   0.399018  13.706  < 2e-16 ***
## HB          2.533452   0.025430  99.624  < 2e-16 ***
## EDAD        0.010250   0.002706   3.787 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.545 on 2338 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8153, Adjusted R-squared:  0.8152
## F-statistic:  5162 on 2 and 2338 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(model_multiple)
```



Al visualizar estos gráficos se recalca que, dado el primero de ellos, tenemos una relación linear entre ambas variables, al distribuirse los residuos sobre la línea horizontal sin patrones aparentes. Según la segunda gráfica, llamada *Normal Q-Q*, los residuos se distribuyen siguiente una línea recta, por lo que concluimos que tienen una distribución normal, pese a los outliers. Dada la gráfica *Scale-Location*, se asume que tenemos la misma varianza. Finalmente, la gráfica *Residuals vs Leverage* muestra la influencia de los outliers en el resultado del análisis dada la distancia de Cook. Los outliers que encontrasemos fuera de las líneas punteadas serían influyentes a la hora de obtener la línea de regresión, pero, dado que en este caso no existen tales valores, concluimos que los outliers de los que disponemos no alteran el resultado.

```
r2_multiple <- summary(model_multiple)$r.squared
r2_multiple
```

```
## [1] 0.8153395
```

El valor R2 es ligeramente supertior al estudiado en el apartado 1.1.a), por lo que concluimos que la edad incrementa levemente la relación entre estas dos variables.

Según los resultados obtenidos en la ejecución de la función *summary*, la tabla de coeficientes contiene diferente información. En primer lugar, la columna *Estimate* nos da los valores a, b y c para completar la fórmula que determina la recta, siendo x_1 el valor correspondiente a la Hemoglobina y x_2 el valor correspondiente a la edad.

$$y = a + bx_1 + cx_2$$

$$y = 5.46 + 2.53x_1 + 0.01x_2$$

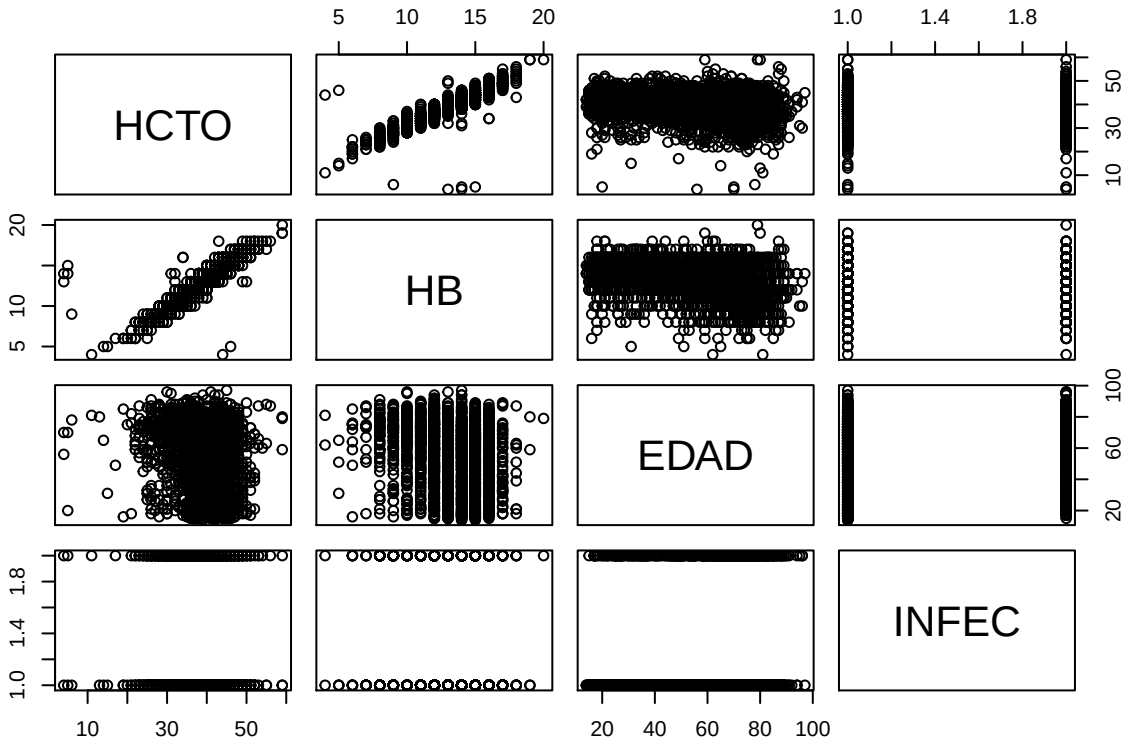
A continuación, encontramos una columna dedicada al error estandar, que indica la diferencia entre el coeficiente estimado y el valor real. También encontramos el valor t (e en t-Student) y al p-value, o la probabilidad de encontrar un valor más alto que t, que en los tres casos es mínimo.

1.3. Regresión linear múltiple (regresores cuantitativos y cualitativos)

a) Todo el conjunto

Queremos conocer en qué medida se relacionan los hematocritos, con la hemoglobina y la edad, dependiendo de si los pacientes tienen o no infección postquirúrgica. Aplicar un modelo de regresión lineal múltiple y explicar el resultado. En primer lugar, observaremos las relaciones entre las variables.

```
d_3 <- data.frame(data$HCTO, data$HB, data$EDAD, data$INFEC)
names(d_3) <- c("HCTO", "HB", "EDAD", "INFEC")
pairs(d_3)
```



```r
model_multiple_2 <- lm(HCTO ~ HB + EDAD + INFEC, data=data)
summary(model_multiple_2)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD + INFEC, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.208  -1.072   0.001   1.186  27.718
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.504500   0.401798  13.700  < 2e-16 ***
## HB           2.531145   0.025615  98.816  < 2e-16 ***
## EDAD         0.010525   0.002731   3.854 0.000119 ***
## INFECsi     -0.102044   0.134815  -0.757 0.449173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.545 on 2337 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8154, Adjusted R-squared:  0.8151
## F-statistic:  3441 on 3 and 2337 DF,  p-value: < 2.2e-16
```

```r
r2_multiple_2 <- summary(model_multiple_2)$r.squared
r2_multiple_2
```

```
## [1] 0.8153848
```

Como se ha explicado anteriormente, en el resultado de la función *summary*, en la tabla *Coefficients* encontramos la estimación de cuanto aumenta o disminuye la variable estudiada en función de la variable indicada en cada fila. Es decir, por cada unidad de hemoglobina que aumenta, la cantidad de Hematocritos se incrementa en un promedio de 2.53. El mismo caso aplica a la edad, que por cada año que aumenta, la cantidad de hematocritos se incrementa en 0.01. Sin embargo, al usar variables cualtivativas, se establece que un valor de referencia, que en este caso es que no haya infección, al que se le asigna el valor 0, mientras que al otro valor, que sí haya infección, se le asigna el valor 1. En este caso, indica que el valor de hematocritos en personas que no presentan infección post-operatoria es 0.1 veces mayor.

$$NivelHematocritos = 5.5 + 2.53 * HB + 0.01 * EDAD - 0,1 * INFEC_{sí}$$

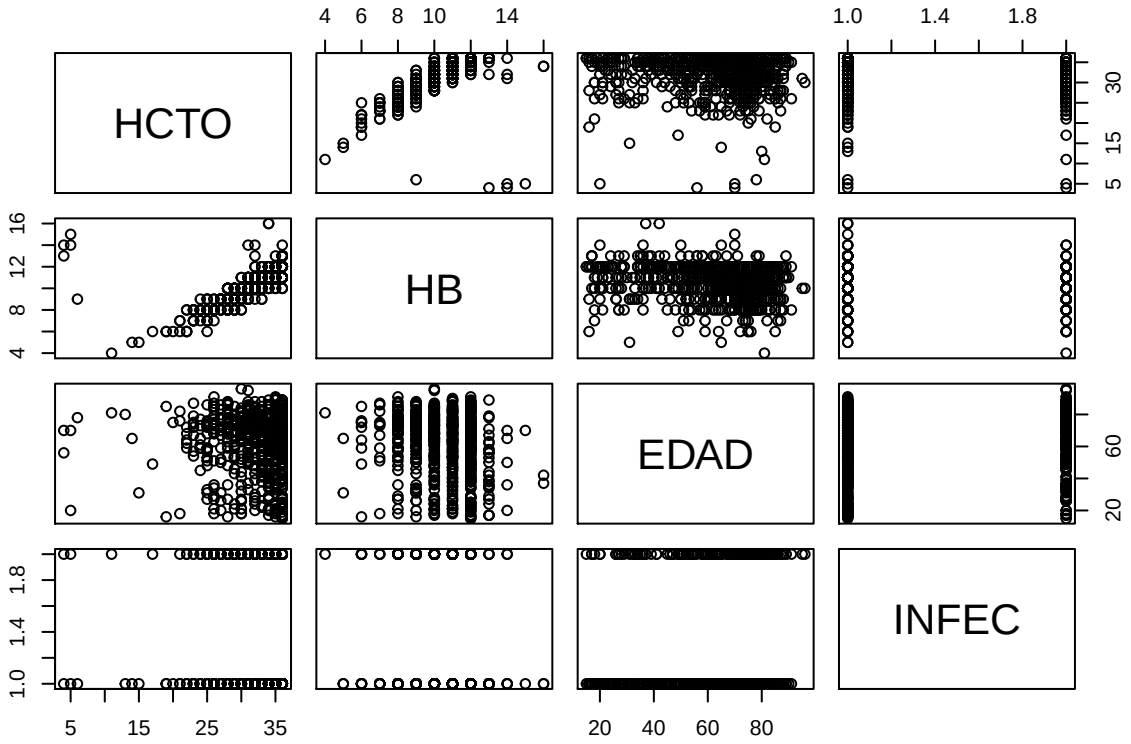
Según el valor R2 = 0.815, concluimos que las variables están relacionadas.

b) Candidad de Hematocritos < 37

Se hará el mismo estudio, pero tomando sólo aquéllos pacientes, cuya cantidad de hematocritos sea < 37. Comparar con el modelo anterior y extraer conclusiones.

En primer lugar, observaremos las relaciones entre las variables después de filtrar según el valor de los hematocritos.

```r
hcto_indexes_less_37 <- which(data$HCTO < 37)
d_3 <- data.frame(data$HCTO[hcto_indexes_less_37], data$HB[hcto_indexes_less_37], data$EDAD[hcto_indexes
names(d_3) <- c("HCTO", "HB", "EDAD", "INFEC")
pairs(d_3)
```



En esta segunda tabla, además de observar una menor densidad en los gráficos, vemos que los valores de hematocritos se distrubuyen desde 0 hasta el 37, y se puede apreciar cómo, por ejemplo, el gráfico de la edad, queda *partido* a menos de la mitad que el anterior. Procedemos a ejecutar el modelo lineal para estudiar la presencia de relación entre estas variables dada la condición anterior.

```
hcto_indexes_less_37 <- which(data$HCTO < 37)
model_multiple_2_less37 <- lm(HCTO ~ HB + EDAD + INFEC, data=data[hcto_indexes_less_37,])
summary(model_multiple_2_less37)
```

```
##
## Call:
## lm(formula = HCTO ~ HB + EDAD + INFEC, data = data[hcto_indexes_less_37,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.616  -0.806   0.350   1.450   5.874
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.784778   1.103507   9.773   <2e-16 ***
## HB           1.942643   0.086347  22.498   <2e-16 ***
## EDAD         0.009873   0.007430   1.329    0.184
## INFECsi     -0.252656   0.306318  -0.825    0.410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.465 on 612 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.4662, Adjusted R-squared:  0.4635
## F-statistic: 178.1 on 3 and 612 DF,  p-value: < 2.2e-16
```

```
r2_multiple_2 <- summary(model_multiple_2_less37)$r.squared
r2_multiple_2
```

```
## [1] 0.4661596
```

En este caso, no podemos concluir que las variables estén relacionadas. Podemos concluir que dada la media del nivel de Hematocritos, al extraer únicamente casos menores a la media, no disponemos de información suficiente para hacer la regresión lineal, dado que esta condición nos deja con 619 observaciones frente a las 2353 anteriores.

1.4. Predicción de la concentración de hematocritos

Suponer un paciente de 60 años, con infección postquirúrgica y con un valor de hemoglobina de 10. Realizar la predicción del valor de hematocritos, con los dos modelos del apartado 1.3. Interpretar los resultados.

Según el primer modelo estudiado en el punto anterior, el nivel de Hematocritos predicho para una persona de 60 años on infección postquirúrjuca y un nivel de Hemoglobina de 10.0 es 31.34, mientras que, según el segundo modelo, elaborado estudiando únicamente los casos en los que el nivel de hematocritos es menor que 37, la predición resulta 30.55.

```
new_pacient <- data.frame(EDAD = 60, INFEC = "si", HB = 10.0)
p <- predict(model_multiple_2, new_pacient)
p
```

```
##        1
## 31.34539
```

```
p2 <- predict(model_multiple_2_less37, new_pacient)
p2
```

```
##        1
## 30.55095
```

Dado que el primer modelo dispone de todos los valores, y el segundo únicamente los individuos con un nivel de hematocritos menor a 37, el valor predicho será menor que el primero, cuya recta de regresión es más precisa al disponer de más valores.

2. Modelo de regresión logística

2.1. Análisis crudo. Estimación de OR

Se desea identificar cuáles son los factores de riesgo en la infección postquirúrgica. Por tanto, se evaluará la probabilidad de que un paciente pueda o no tener una infección, dependiendo si presenta o no unas determinadas características. Para evaluar esta probabilidad, primero se realizará un análisis crudo de los posibles factores (características). Es decir, un análisis univariante de posibles factores de riesgo asociados a la infección postquirúrgica.

a) Relación entre infección postquirúrgica con el resto de variables. Estimar e interpretar las OR.

Estudiar la relación entre la infección postquirúrgica y cada una de las variables siguientes: diabetes, desnutrición, obesidad, edad y hematocrito. Estimar e interpretar las OR en cada caso. Dicha estimación será efectuada a partir de las tablas de contingencia. Antes de calcular los valores de las odds ratio, se recomienda aplicar el test chi-cuadrado, para valorar la relación entre las variables.

Para estudiar la relación entre la infección postquirúrgica y el resto de variables, realizaremos el test chi-cuadrado, que pertenece a las llamadas pruebas de bondad de ajuste o contraste, y se utiliza para analizar la relación entre variables nominales o cualitativas. El estadístico chi-cuadrado tomará el valor 0 si existe una concordancia perfecta entre las frecuencias observadas de los valores en las tablas de contingencia, con las frecuencias esperadas, mientras que si existe una discrepancia entre dichas frecuencias, el estadístico tomará un valor grande. Se establecerá como hipótesis nula que las variables son independientes, mientras que la alternativa será que las variables presentan una dependencia entre ellas. Si el p-valor obtenido en el test es menor que 0.05, concluiremos que la hipótesis nula es falsa y aceptaremos la alternativa y viceversa.

Para realizar el estudio, antes de ejecutar el test, estudiaremos la tabla de contingencias entre ambas variables e imprimiremos un gráfico que muestra las relaciones entre ambas,para todas las variables categórigas, y únicamente para una contínua, y a continuación realizar el test en base a dicha tabla de contingencias.

```
conting_diab <- table(data$INFEC, data$DIABETES)
conting_diab
```

```
##
##        no   si
##   no 1792   97
##   si  419   45
```

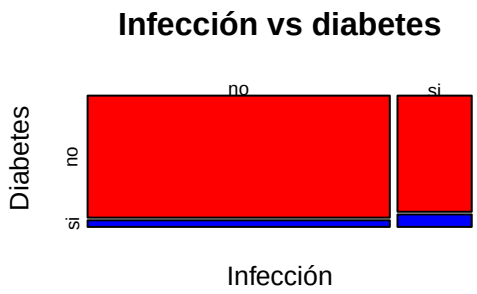
```
conting_desnutri <- table(data$INFEC, data$DESNUTR)
conting_desnutri
```

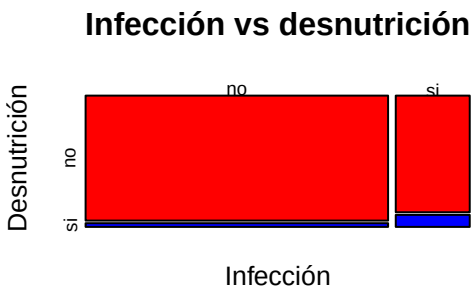
```
##
##        no   si
##   no 1831   54
##   si  418   43
```

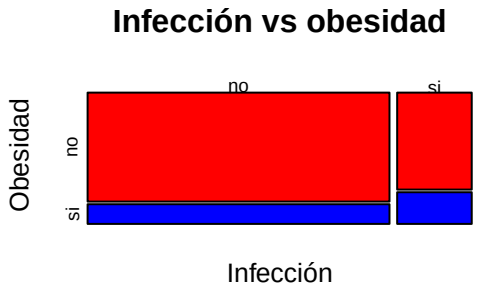
```
conting_obes <- table(data$INFEC, data$OBES)
conting_obes
```

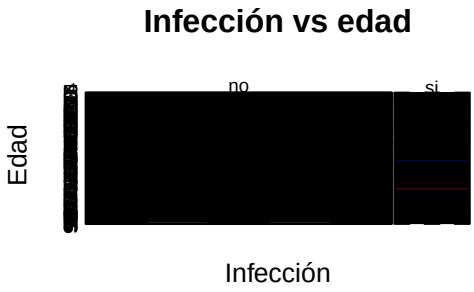
```
##
##        no   si
##   no 1379  249
##   si  303   99
```

```
conting_edad <- table(data$INFEC, data$EDAD)
conting_hcto <- table(data$INFEC, data$HCTO)
```










```
chisq.test(conting_diab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conting_diab
## X-squared = 12.886, df = 1, p-value = 0.0003311
```

```
chisq.test(conting_desnutri)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conting_desnutri
## X-squared = 37.419, df = 1, p-value = 9.53e-10
```

```
chisq.test(conting_obes)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  conting_obes
## X-squared = 19.115, df = 1, p-value = 1.231e-05
```

```
chisq.test(conting_edad)
```

```
## Warning in chisq.test(conting_edad): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  conting_edad
## X-squared = 144.46, df = 82, p-value = 2.536e-05
```

```
chisq.test(conting_hcto)
```

```
## Warning in chisq.test(conting_hcto): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  conting_hcto
## X-squared = 161.68, df = 46, p-value = 8.798e-15
```

Como podemos observar, el p-valor obtenido en todos los tests resulta menor que 0.05, por lo que rechazaremos la hipótesis nula, que indica que las variables son independientes, para concluir en que todas influyen en la presencia de infecciones.

Para calcular el *Odds Ratio* (OR), que puede ser definido como una medida de asociación enre variables binarias (sus valores son sí y no, 1 y 0), se deben seguir los pasos siguientes. En primer lugar, realizaremos el estudio con varias funciones[1] , de las que extraeremos el cómputo final. Una OR=1 quiere decir que no hay asociación entre las variables. Una OR<1 quiere decir que el valor con el que se está comparando es un factor de protección frente a la infección. Por último, una OR>1 indica que la exposición es un factor de riesgo, tanto mayor cuanto mayor sea el valor de la OR.

```
oddsratioWald.proc <- function(n00, n01, n10, n11, alpha = 0.05){
  #
  #  Compute the odds ratio between two binary variables, x and y,
  #  as defined by the four numbers nij:
  #
  #    n00 = number of cases where x = 0 and y = 0
  #    n01 = number of cases where x = 0 and y = 1
  #    n10 = number of cases where x = 1 and y = 0
  #    n11 = number of cases where x = 1 and y = 1
  #
  OR <- (n00 * n11)/(n01 * n10)
  #
  #  Compute the Wald confidence intervals:
  #
  siglog <- sqrt((1/n00) + (1/n01) + (1/n10) + (1/n11))
  zalph <- qnorm(1 - alpha/2)
  logOR <- log(OR)
  loglo <- logOR - zalph * siglog
  loghi <- logOR + zalph * siglog
  #
  ORlo <- exp(loglo)
  ORhi <- exp(loghi)
  #
  oframe <- data.frame(LowerCI = ORlo, OR = OR, UpperCI = ORhi, alpha = alpha)
  oframe
}
```

[1] https://www.r-bloggers.com/computing-odds-ratios-in-r/

```
AutomaticOR.proc <- function(x,y,alpha=0.05){
  #
  xtab <- table(x,y)
  n00 <- xtab[1,1]
  n01 <- xtab[1,2]
  n10 <- xtab[2,1]
  n11 <- xtab[2,2]
  #
  rawOR <- (n00*n11)/(n01*n10)
  if (rawOR < 1){
    n01 <- xtab[1,1]
    n00 <- xtab[1,2]
    n11 <- xtab[2,1]
    n10 <- xtab[2,2]
    iLevel <- 2
  }
  else{
    iLevel <- 1
  }
  outList <- vector("list",2)
  outList[[1]] <- paste("Odds ratio between the level [",dimnames(xtab)[[1]][1],"] of the first variable
  outList[[2]] <- oddsratioWald.proc(n00,n01,n10,n11,alpha)
  outList
}
```

De esta manera, al llamar a la última función, obtendremos el valor Odds para el primer estudio, la relación entre una infección postquirúrgica y tener diabetes.

```
AutomaticOR.proc(data$INFEC, data$DIABETES)
```

```
## [[1]]
## [1] "Odds ratio between the level [ no ] of the first variable and the level [ no ] of the second va
##
## [[2]]
##    LowerCI       OR  UpperCI alpha
## 1 1.371639 1.984106 2.870051  0.05
```

De esta forma, obtenemos que el *Odds ratio* es 1.984. Así, podemos proceder con el estudio del resto de variables, a excepción de la edad y el nivel de hematocritos, cuyo resultado no será relevante dado que no se trata de variables binarias.

```
AutomaticOR.proc(data$INFEC, data$DESNUTR)
```

```
## [[1]]
## [1] "Odds ratio between the level [ no ] of the first variable and the level [ no ] of the second va
##
## [[2]]
##    LowerCI       OR  UpperCI alpha
## 1 2.304606 3.488083 5.279306  0.05
```

```
AutomaticOR.proc(data$INFEC, data$OBES)
```

```
## [[1]]
## [1] "Odds ratio between the level [ no ] of the first variable and the level [ no ] of the second va
##
## [[2]]
```

```
##   LowerCI       OR  UpperCI alpha
## 1 1.38965 1.809495 2.356186  0.05
```

```
##   Diabetes Desnutricion Obesidad
## 1     1.98         3.48      1.8
```

Podemos interpretar los resultados indicando que las tres variables binarias influyen en la aparición de infecciones postquirúrgicas, siendo la presencia de desnutrición la más determinante, multiplicando por 3.48 la probabilidad de que aparezca la infección en caso de mostrar desnutrición.

b) Calculo de OR para variables contínuas

Edad y hematocrito son variables continuas: ¿podríamos seguir el procedimiento anterior para el cálculo de la OR?

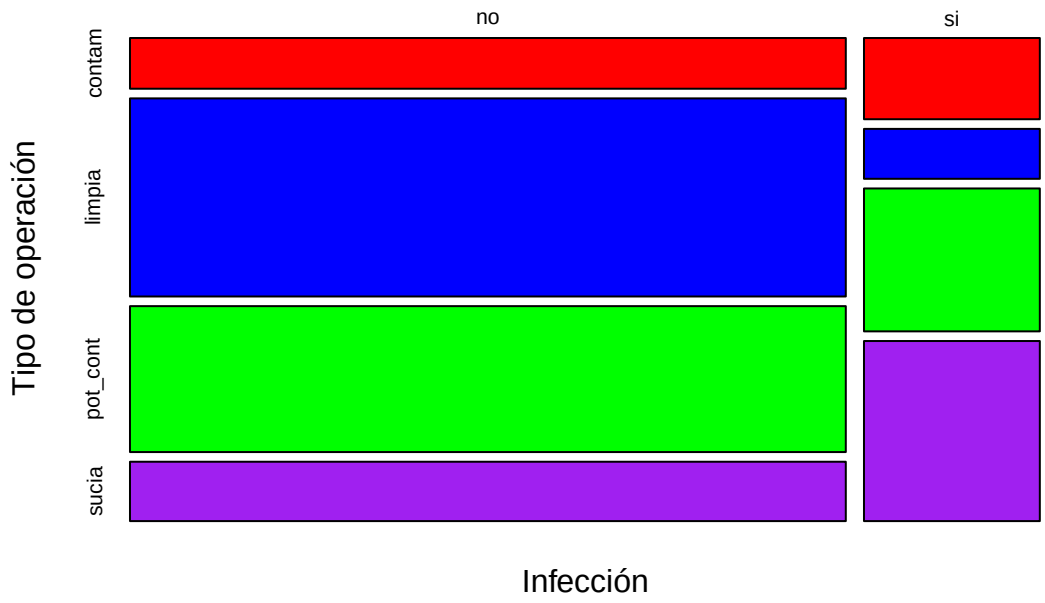
Dado que la edad y el nivel de hematocritos no son variables binarias, no podrían seguirse los pasos anteriores, ya que el cálculo del OR únicamente es válido para tablas de contingencia 2x2. Sin embargo, haciendo uso de este método, podrá estudiarse la relación entre la presencia de infección y una cierta edad, o una categorización de edades.

c) Relación entre infección y tipo de operación

Si queremos ver la relación entre INFEC (Infección) y TIP_OPER (tipo de operación), ¿podríamos seguir el procedimiento anterior, para el cáculo de la OR? En el caso que la respuesta fuese negativa, ¿cuál sería una solución?

Al igual que en el caso anterior, no podremos establecer una medida de relación entre la presencia de infección y el tipo de operación que se ha llevado a cabo dada la naturaleza del método, al no ser el tipo de operación una variable binaria. Dado que existen 4 tipos de operaciones, la tabla de contingencias se mostraría de la siguiente forma.

```
##
##      contam limpia pot_cont sucia
##   no    211    824      607   247
##   si     83     51      146   184
```



Se puede observar que el tipo de operación que más determina que exista una infección postquirúrgica es la *sucia*, seguida de la *pot cont*, *contam* y *limpia* en último lugar.

Existen otras formas de obtener el *Odds Ratio*, sin embargo no son tan directas como la anterior. Por ejemplo, podemos hacer un modelo de regresión logística y de sus coeficientes obtener el *Odds ratio* como vemos a continuación.

```
model <- glm(INFEC ~TIP_OPER, data=data, family="binomial")
odds <- exp(coef(model))
odds
```

```
##      (Intercept)   TIP_OPERlimpia TIP_OPERpot_cont    TIP_OPERsucia
##        0.3933649        0.1573430        0.6114607        1.8937613
```

Efectivamente, podemos observar que la variable que presenta una mayor relación con la presencia de infección es el tipo de operación sucia, que aumenta 1.89 unidades las posibilidades de generar una infección tras la operación.

2.2. Model de regresión logística

La regresión logística forma parte de los llamados modelos lineales generalizados, que tratan de estimar la probabilidad de que ocurra un evento binario en base a ciertas variables predictoras. Para realizar el modelo de regresión logística en R, utilizaremos la función glm (general lineal models), que se diferencia de la función lm en que hay que especificar el tipo de familia. En nuestro caso, aplicaremos el modelo a la variable Infección, que es dicotómica (solo puede adoptar dos valores), por lo que estableceremos la familia "binomial".

a) Infección y diabetes

Estimar el modelo de regresión logística donde la variable dependiente es "INFEC" y la explicativa es tener diabetes o no. ¿Podemos considerar que el hecho de tener diabetes es un factor de riesgo de infección? Justifica tu respuesta. Tiene relación con lo obtenido en el apartado anterior? Se recodificará la variable DIABETES en 0=NO y 1=SI.

```
model <- glm(INFEC ~ DIABETES, data=data, family="binomial")
summary(model)
```

```
##
## Call:
## glm(formula = INFEC ~ DIABETES, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8731  -0.6482  -0.6482  -0.6482   1.8239
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.45322    0.05426 -26.780  < 2e-16 ***
## DIABETESsi   0.68517    0.18835   3.638 0.000275 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2336.5  on 2352  degrees of freedom
## Residual deviance: 2324.3  on 2351  degrees of freedom
## AIC: 2328.3
##
## Number of Fisher Scoring iterations: 4
```

La interpretación del modelo es similar al anterior. Se puede observar que la variable DIABETESsi, tiene un p-valor inferior a 0.05, por lo que, como anteriormente, concluiremos que la presencia de diabetes influye en la posible presencia de una infección postquirúrgica. En cambio, la interpretación de los coeficientes varía ligeramente. Existen funciones para facilitar la interpretación de los coeficientes, exponenciandolos, y obteniendo así los *Odds ratio*.

```
exp(coef(model))
```

```
## (Intercept)  DIABETESsi
##    0.233817    1.984106
```

Esto indica que la presencia de diabetes aumenta en 1.98 unidades la probabilidad de obtener una infección. Como podemos observar, es el mismo resultado que obtuvimos en el punto 2.1.a.

b) Infección, diabetes, edad y nivel de hematocritos

Posteriormente se añadirá al modelo las variable explicativa desnutrición. ¿Se observa una mejora del modelo? Explicar.

En este caso, la variable dependiente es dicotómica, mientras que las variables explicativas son cuantitativa discreta en el caso de la edad, cualitativa en el caso de la diabetes (sí = 1 y no = 0) y continua en el caso del nivel de hematocritos.

```r
model2 <- glm(INFEC ~ DIABETES + EDAD + HEMAT, data=data, family="binomial")
summary(model2)
```

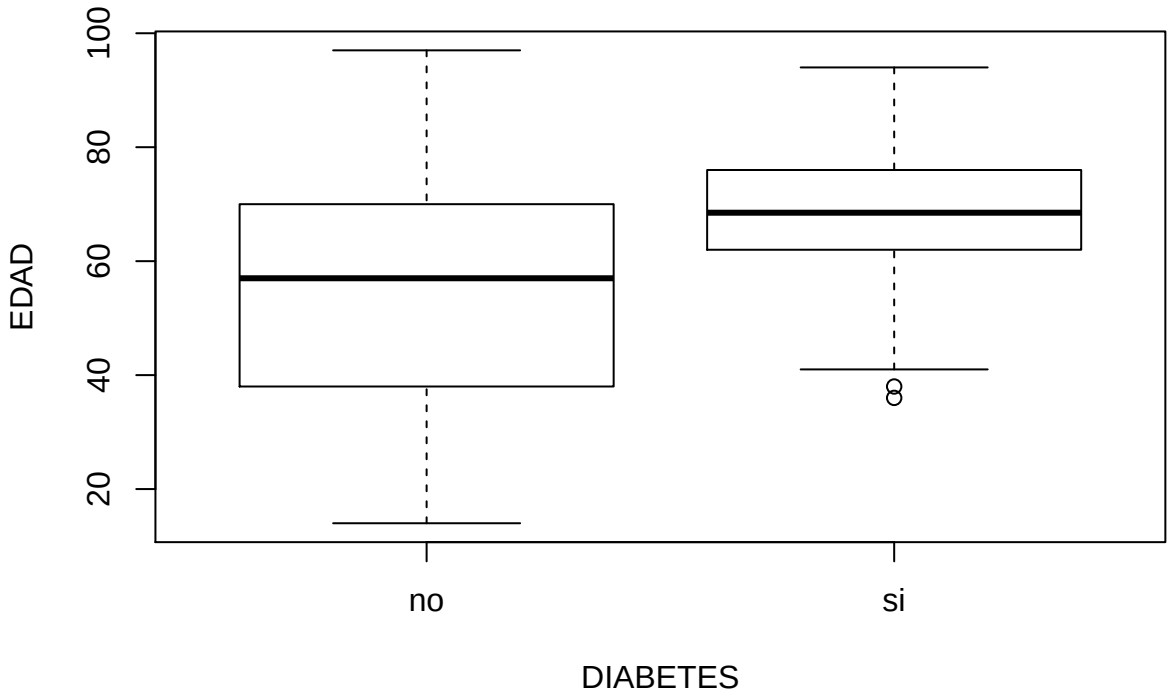
```
##
## Call:
## glm(formula = INFEC ~ DIABETES + EDAD + HEMAT, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1805  -0.7020  -0.5810  -0.4327   2.5107
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.814496   0.399414  -2.039   0.0414 *
## DIABETESsi   0.387511   0.199228   1.945   0.0518 .
## EDAD         0.018358   0.002986   6.148 7.83e-10 ***
## HEMAT       -0.378479   0.076409  -4.953 7.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2231.4  on 2247  degrees of freedom
## Residual deviance: 2144.3  on 2244  degrees of freedom
##   (105 observations deleted due to missingness)
## AIC: 2152.3
##
## Number of Fisher Scoring iterations: 4
```

```r
odds <- exp(coefficients(model2))
odds
```

```
## (Intercept)  DIABETESsi        EDAD       HEMAT
##   0.4428623   1.4733095   1.0185275   0.6849025
```

En este caso, la variable diabetes deja de ser relevante en la presencia de infecciones, lo que evidencia este modelo es que la edad y el nivel de hematocritos es más relevante, es decir, tiene una dependencia más fuerte a la edad y el nivel de hematocritos que a la diabetes. Como podemos ver a continuación, esto se debe a que la diabetes sí que está ligada a la edad, siendo más frecuente en edades más altas. Se puede observar tanto en el gráfico, como en el resultado del *Odds range*. Éste es un ejemplo de la paradoja de Simpson.

```r
plot(data$EDAD~data$DIABETES, ylab="EDAD", xlab="DIABETES")
```




```
model3 <- glm(DIABETES ~ EDAD, data=data, family="binomial")
odds <- exp(coefficients(model3))
odds
```

```
## (Intercept)        EDAD
## 0.003404904 1.048614817
```

Concluiremos con que, teniendo en cuenta la edad y el nivel de hematocritos, además de la presencia de la enfermedad diabética, diremos que la diabetes deja de ser tan relevante en la presencia de infecciones, mientras que la edad y el nivel de hematocritos obtienen un p-valor menor que 0.05. La presencia de diabetes se mantiene moderadamente influyente, ya que aumenta 1.47 unidades la probabilidad de infectarse tras la operación, sin embargo la edad multiplica la probabilidad por algo más de una unidad por cada año del paciente. Cuanto menor es el nivel de hematocritos, se aumenta la probabilidad de subrir una infección en 1.46 puntos, que es la inversa de 0.68.

2.3. Mejora del modelo

a) Categorización variables continuas

Entrenamos el mismo modelo anterior, pero categorizando ambas variables continuas: Edad: (edad>=65 y edad<65) y Hematocrito: (hb <37 y hb>=37). Explicar los resultados. ¿De qué forma influye la edad y los niveles de hematocritos en este modelo? Explicar como se interpretan los resultados del modelo.

En este caso, convertiremos la edad y el nivel de hematocritos en variables binarias, dividiendolas en dos agrupaciones, estableceremos el valor 1 para los casos en los que la edad es mayor o igual a 65, y 0 en el caso contrario, al igual que estableceremos el valor 1 para los casos en los que el nivel de hematocritos es superior o igual a 37 y 0 si es menor.

```r
edad_categorica <- c()
edad_categorica[which(data$EDAD >= 65)] <- 1
edad_categorica[which(data$EDAD < 65)] <- 0
hematocritos <- c()
hematocritos[which(data$HCTO >= 37)] <- 1
hematocritos[which(data$HCTO < 37)] <- 0

new_data <- data.frame(data$INFEC, data$DIABETES, factor(edad_categorica), factor(hematocritos))
names(new_data) <- c("INFEC", "DIABETES", "EDAD_65", "HCTO_37")
summary(new_data)
```

```
##  INFEC     DIABETES  EDAD_65     HCTO_37
##  no:1889   no:2211   0   :1455   0   : 619
##  si: 464   si: 142   1   : 896   1   :1727
##                      NA's:   2   NA's:   7
```

Con este nuevo data frame llamado new_data, aplicamos el modelo de regresión logística.

```r
model4 <- glm(INFEC ~ DIABETES + EDAD_65 + HCTO_37, data=new_data, family="binomial")
summary(model4)
```

```
##
## Call:
## glm(formula = INFEC ~ DIABETES + EDAD_65 + HCTO_37, family = "binomial",
##     data = new_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1404  -0.6786  -0.5177  -0.5177   2.0377
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1459     0.1084 -10.569  < 2e-16 ***
## DIABETESsi    0.4674     0.1946   2.402   0.0163 *
## EDAD_651      0.5908     0.1086   5.441 5.31e-08 ***
## HCTO_371     -0.7962     0.1115  -7.138 9.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2332.5  on 2343  degrees of freedom
## Residual deviance: 2225.4  on 2340  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 2233.4
##
## Number of Fisher Scoring iterations: 4
```

```r
odds <- exp(coefficients(model4))
odds
```

```
## (Intercept)  DIABETESsi    EDAD_651    HCTO_371
##   0.3179474   1.5957814   1.8054029   0.4510478
```

Así, podemos comprobar que la diabetes vuelve a ser más relevante que anteriormente, siendo su p-valor menor que 0.05, al categorizarse el resto de variables, y no haber tanta casuística. En este caso la probabilidad

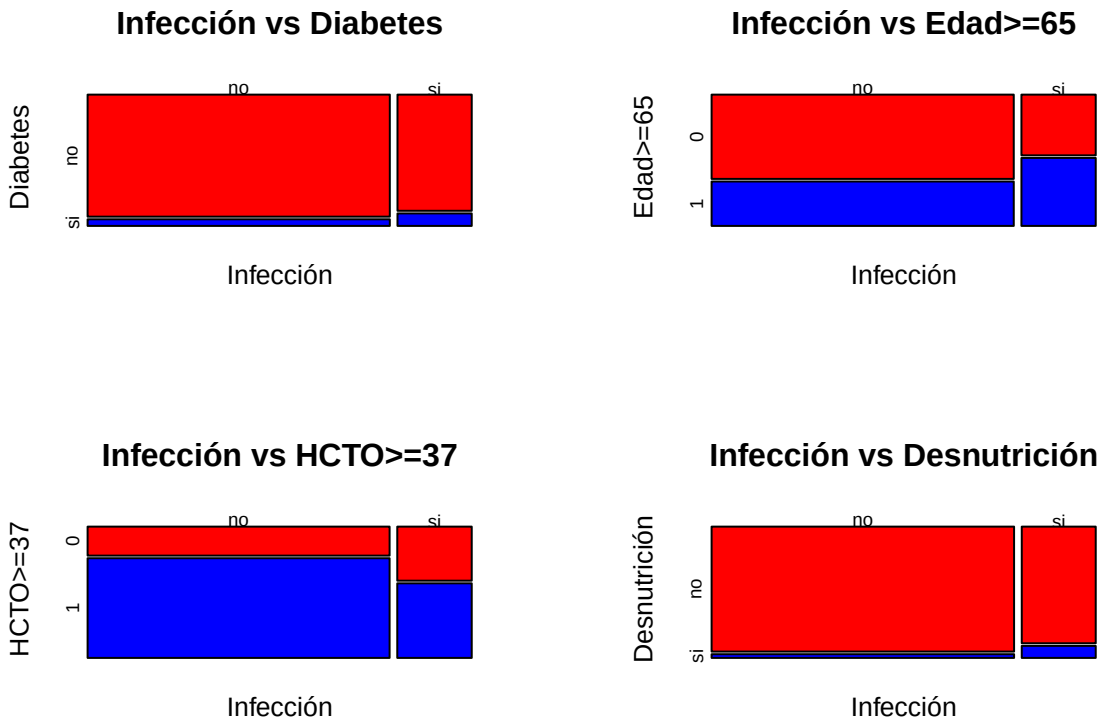
de tener una infección aumenta en casi 1.6 unidades si el paciente tiene diabetes, así como aumenta en 1.8 si el paciente tiene 65 años o más. Sin embargo, cuanto menor sea el nivel de hematocritos, la probabilidad de que surga una infección postoperatoria aumenta en 2.21 unidades, que es la inversa de 0.45.

b) Infección, diabetes, edad, nivel de hematocritos y desnutrición

Posteriormente se añadirá al modelo las variable explicativa desnutrición. ¿Se observa una mejora del modelo? Explicar.

Añadiremos al modelo anterior ya categorizado la variable destrutrición. En primer lugar, observaremos las gráficas que contienen la relación de cada variable con la existencia de una futura infección para poder anticipar qué variables van a ser más significativas.

```
new_data <- data.frame(data$INFEC, data$DIABETES, factor(edad_categorica), factor(hematocritos), data$DE
names(new_data) <- c("INFEC", "DIABETES", "EDAD_65", "HCTO_37", "DESNUTR")
```



Como podemos ver, se dan más infecciones en personas diabéticas que las que no lo son, o en personas de 65 años o mayores. También apreciamos que la presencia de desnutrición resalta los casos de infección, al igual que cuando el nivel de hematocritos es menor a 37.

```
model5 <- glm(INFEC ~ DIABETES + EDAD_65 + HCTO_37 + DESNUTR, data=new_data, family="binomial")
summary(model5)
```

```
##
## Call:
## glm(formula = INFEC ~ DIABETES + EDAD_65 + HCTO_37 + DESNUTR,
##     family = "binomial", data = new_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4266  -0.6694  -0.5163  -0.5163   2.0402
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2408     0.1123 -11.052  < 2e-16 ***
## DIABETESsi    0.4463     0.1952   2.286 0.022262 *
## EDAD_651      0.5661     0.1095   5.171 2.33e-07 ***
## HCTO_371     -0.7071     0.1145  -6.175 6.61e-10 ***
## DESNUTRsi     0.7975     0.2204   3.619 0.000296 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2321.0  on 2336  degrees of freedom
## Residual deviance: 2204.2  on 2332  degrees of freedom
##   (16 observations deleted due to missingness)
## AIC: 2214.2
##
## Number of Fisher Scoring iterations: 4
```

```
odds <- exp(coefficients(model5))
odds
```

```
## (Intercept)  DIABETESsi    EDAD_651    HCTO_371   DESNUTRsi
##   0.2891641   1.5624612   1.7613372   0.4930876   2.2200541
```

Podemos observar que en este caso todas las variables observadas son determinantes en la aparición de futuras infecciones tras una operación. Los valores presentes en el modelo anterior se mantienen similares, mientras que la presencia de desnutrición destaca siendo el atributo más asociado a la aparición de una infección, aumentanto en 2.22 unidades la probabilidad de que ésta exista.

2.4. Predicción

Según el modelo del apartado anterior, ¿cuál será la probabilidad de infección postquirúrgica de un paciente de 50 años, con diabetes, concentración de hematocritos de 34, y que no presente desnutrición?

La probabilidad de infectarse para un paciente con las siguientes características es 0.311.

```
paciente <- data.frame(EDAD_65 = factor(0), DIABETES = "si", HCTO_37 =factor(0), DESNUTR = "no")
paciente
```

```
##   EDAD_65 DIABETES HCTO_37 DESNUTR
## 1       0       si       0      no
```

```
predict(model5, paciente, type = "response")
```

```
##         1
## 0.3112035
```

2.5. Conclusiones

En este apartado deberéis exponer, de las variables explicativas estudiadas, cuáles pueden considerarse factores de riesgo en la infección postquirúrgica. Razonar en base a los resultados obtenidos.

Finalmente, hemos estudiado varios modelos para establecer una relación entre variables explicativas y la presencia de una infección postquirúrgica. De las variables explicativas observadas (EDAD, HCTO, DESNUTR, DIABETES y TIPO_OPER), descaca la dependencia de la desnutrición frente al resto, ya que es la más determinante en la futura infección. Según se puede ver en el último estudio, aumenta 2.22 unidades la probabilidad de infección. En la siguiente tabla se puede observar las variables obtenidas en el último modelo ordenadas de menor a mayor en su nivel de influencia en la presencia de una infección.

	DESNUTRICION_SI	HCTO_MENOR_37	EDAD_MAYOR_65	DIABETES_SI
1	2.22	2	1.76	1.56

Como se ha comentado anteriormente, la presencia de desnutrición es el atributo del que más depende la presencia de infeccón, seguido del nivel de hematocritos, que resultara más relevante si es menor. A continuación, encontramos la edad, que estará más vinculada a la infección postquirúrgica a medida que aumenta la misma, así como la presencia de diabetes también es un factor determinante.