Caso Estudio solucionado

Regresión con R

CNE/ISCIII

1 Preambulo

1.1 Objetivo

A partir de los controles del estudio de casos y controles EURAMIC se pretende evaluar el efecto de distintos factores sobre el riesgo de hipertensión en hombres adultos. La base incluye las siguientes variables:

- **statcc**: 1 "caso de infarto", 0 "control"
- **edad**: edad del paciente en años
- **peso**: peso en kg
- **altura**: altura en cm
- **tipfum**: 1 "no fumador", 2 "ex fumador", 3 "fumador actual"
- **alcohol**: consumo de alcohol en g/día (0 si no es bebedor actual)
- **hta**: 1 "hipertenso", 0 "normotenso"
- **diabetes**: 1 "diabético", 0 "no diabético"
- **hdlcol**: colesterol HDL en mmol/l
- **totcol**: colesterol Total en mmol/l

1.2 Importación de los datos

```
euramic = read.csv("data/euramic.csv")
euramic$tipfum = factor(euramic$tipfum, levels = 1:3, labels = c("nunca", "ex", "activo"))
str(euramic)
```

```
## 'data.frame':    1339 obs. of  10 variables:
##  $ statcc  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ edad    : int  67 66 51 52 69 43 61 47 66 60 ...
##  $ peso    : int  62 70 68 70 72 78 81 90 84 103 ...
##  $ altura  : int  164 178 158 172 173 172 175 173 165 180 ...
##  $ tipfum  : Factor w/ 3 levels "nunca","ex","activo": 1 3 1 2 3 3 3 3 3 2 ...
##  $ alcohol : int  0 12 21 80 0 32 4 96 216 73 ...
##  $ hta     : int  0 0 0 0 0 1 0 0 0 1 ...
##  $ diabetes: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hdlcol  : num  0.89 NA 1.58 0.79 1.29 ...
##  $ totcol  : num  6.29 NA 6.96 5.01 4.79 ...
```

2 Correlación y regresión simple

Utilizando la muestra de los controles (representativos de la población general,

1. Representar en un diagrama de dispersión el colesterol HDL y el índice de masa corporal (imc = peso (kg)/altura (m)^2).
2. Evaluar la asociación lineal entre estas dos variables (interpretar)
3. Estimar mediante un modelo de regresión lineal el efecto del IMC sobre el nivel de HDL e interpretarlo
4. Chequear las asunciones del modelo
5. Calcular el intervalo de confianza de este efecto (interpretar)
6. Representar la recta de regresión

```
euramic$imc = euramic$peso/(euramic$altura/100)^2
controles = subset(euramic, statcc == 0)
plot(hdlcol ~ imc, data = controles)
cor.test(controles$hdlcol, controles$imc)
```

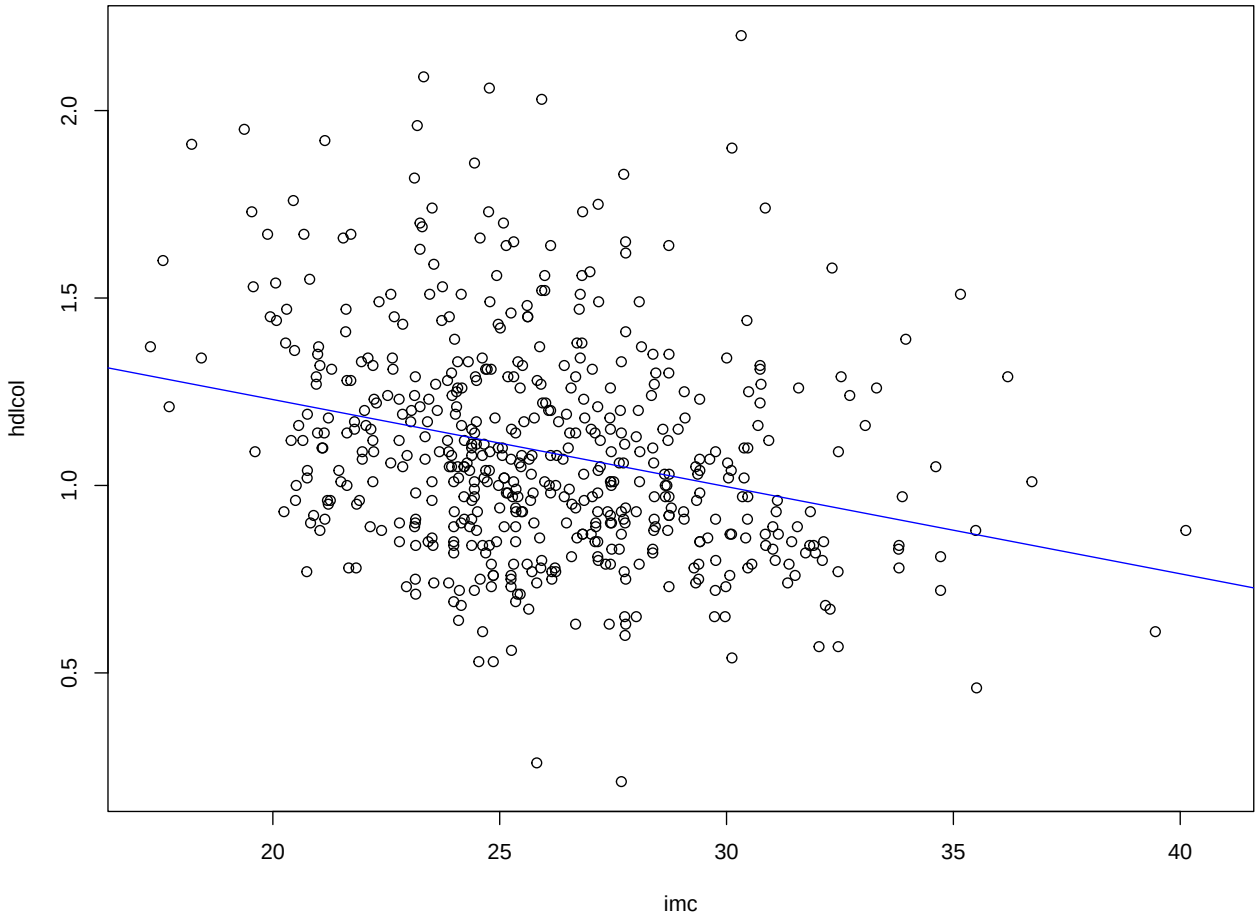
```
##
##  Pearson's product-moment correlation
##
## data:  controles$hdlcol and controles$imc
## t = -6.6278, df = 531, p-value = 8.385e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3530587 -0.1960900
## sample estimates:
##        cor
## -0.2764168
```

```
simple = lm(hdlcol ~ imc, data = controles)
summary(simple)
```

```
##
## Call:
## lm(formula = hdlcol ~ imc, data = controles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8407 -0.1844 -0.0463  0.1505  1.2107
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.693956   0.092042  18.404  < 2e-16 ***
## imc         -0.023238   0.003506  -6.628 8.38e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2833 on 531 degrees of freedom
##   (167 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.07641,    Adjusted R-squared:  0.07467
## F-statistic: 43.93 on 1 and 531 DF,  p-value: 8.385e-11
```

```
abline(reg = simple, col = "blue")
```



3 Regresión múltiple

1. Evaluar la no-linealidad del efecto del IMC sobre el nivel de HDL
2. Estimar el efecto del IMC ajustando por el hábito tabáquico (interpretar)
3. Contrastar si el efecto del IMC está modificado por el consumo de tabaco.
4. Representar en un grafico, la relación entre HDL e IMC según el hábito tabáquico

```
require(splines)
for (k in 1:4) {
    fit = lm(hdlcol ~ ns(imc, k), data = controles)
    cat("AIC para spline con", k, "grados de libertad:", AIC(fit), "\n")
}
```

```
## AIC para spline con 1 grados de libertad: 172.1984
## AIC para spline con 2 grados de libertad: 169.3571
## AIC para spline con 3 grados de libertad: 168.1503
## AIC para spline con 4 grados de libertad: 169.778
```

```r
summary(multiple <- lm(hdlcol ~ imc + tipfum, data = controles))
```

```
##
## Call:
## lm(formula = hdlcol ~ imc + tipfum, data = controles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86822 -0.17503 -0.04668  0.14984  1.17846
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.722466   0.093152  18.491  < 2e-16 ***
## imc          -0.023672   0.003492  -6.778 3.27e-11 ***
## tipfumex      0.016882   0.030930   0.546   0.5854
## tipfumactivo -0.065806   0.030983  -2.124   0.0341 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2816 on 528 degrees of freedom
##   (168 observations deleted due to missingness)
## Multiple R-squared:  0.09198,    Adjusted R-squared:  0.08683
## F-statistic: 17.83 on 3 and 528 DF,  p-value: 4.9e-11
```

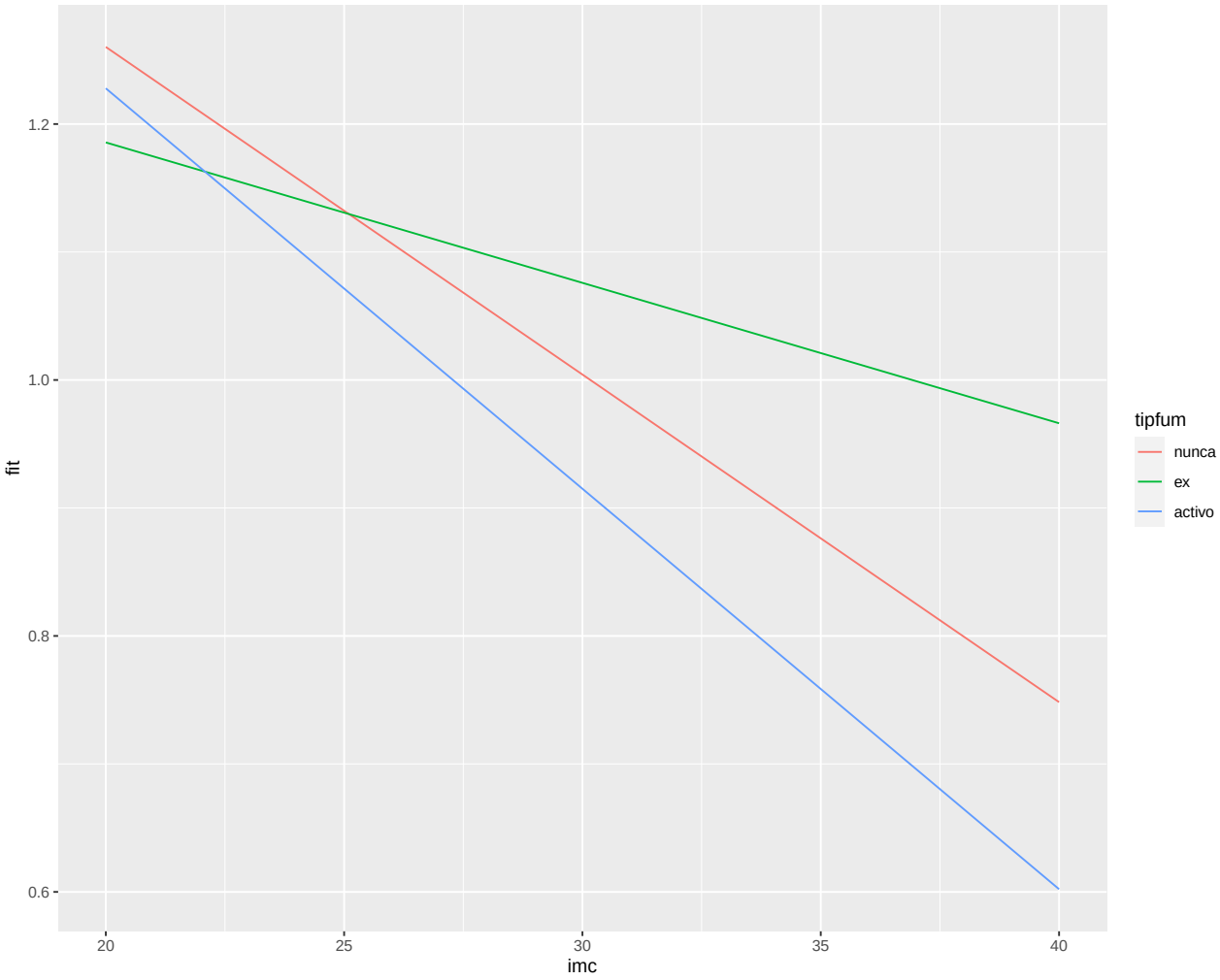
```r
summary(multiple <- lm(hdlcol ~ imc * tipfum, data = controles))
```

```
##
## Call:
## lm(formula = hdlcol ~ imc * tipfum, data = controles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86180 -0.18397 -0.04755  0.14896  1.12765
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.772183   0.186892   9.482  < 2e-16 ***
## imc              -0.025597   0.007182  -3.564 0.000399 ***
## tipfumex         -0.367112   0.249876  -1.469 0.142383
## tipfumactivo      0.081457   0.230374   0.354 0.723791
## imc:tipfumex      0.014625   0.009524   1.536 0.125241
## imc:tipfumactivo -0.005691   0.008837  -0.644 0.519911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2804 on 526 degrees of freedom
##   (168 observations deleted due to missingness)
## Multiple R-squared:  0.1029, Adjusted R-squared:  0.09434
```

```
## F-statistic: 12.06 on 5 and 526 DF,  p-value: 4.468e-11
```

```
nuevos = expand.grid(imc = 20:40, tipfum = c("nunca", "ex", "activo"))
nuevos$fit = predict(multiple, nuevos)

ggplot2::qplot(imc, fit, data = nuevos, color = tipfum, geom = "line")
```



4 Regresión logistica

1. Evaluar la asociación (OR) entre ser fumador actual y el riesgo de hipertensión
2. Estimar el efecto crudo del colesterol HDL y su efecto ajustado por IMC, edad y hábito tabaquico
3. Dar la tabla de resultados de esta regresión logística en un formato apto para publicación
4. Representar el riesgo de hipertension en función del IMC para hombres de 30, 50 y 70 años

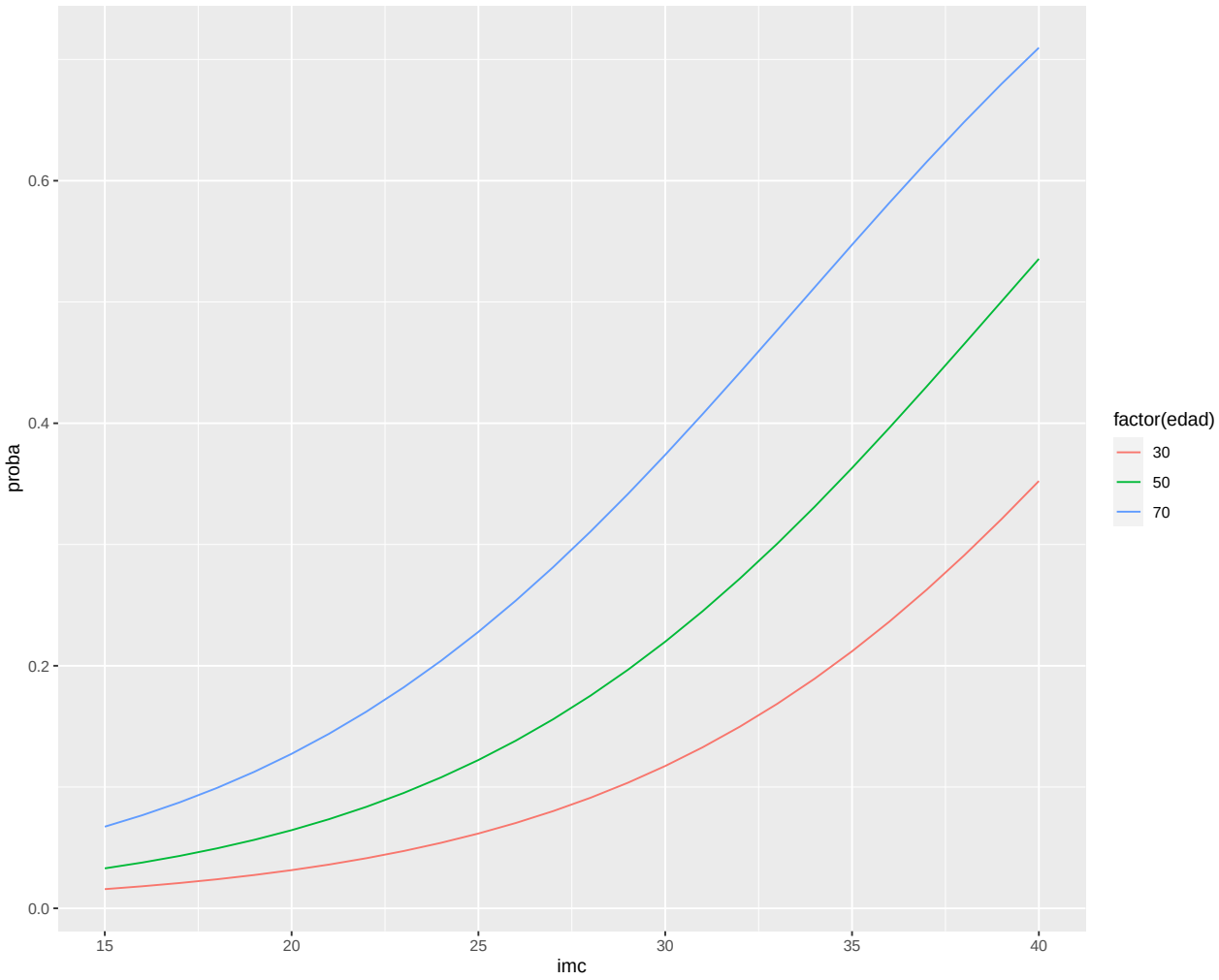
```
or = glm(hta ~ hdlcol + imc + edad + factor(tipfum), data = controles, family = "binomial")
summary(or)
```

```
##
## Call:
```

```
## glm(formula = hta ~ hdlcol + imc + edad + factor(tipfum), family = "binomial",
##     data = controles)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3063  -0.6645  -0.5132  -0.3357   2.4510
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.61287    1.40076  -5.435 5.49e-08 ***
## hdlcol               -0.34430    0.42890  -0.803  0.42212
## imc                   0.15780    0.03440   4.588 4.48e-06 ***
## edad                  0.04141    0.01388   2.984  0.00285 **
## factor(tipfum)ex      0.08918    0.29732   0.300  0.76423
## factor(tipfum)activo -0.14341    0.31319  -0.458  0.64703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 496.19  on 531  degrees of freedom
## Residual deviance: 459.73  on 526  degrees of freedom
##   (168 observations deleted due to missingness)
## AIC: 471.73
##
## Number of Fisher Scoring iterations: 5
```

```
or2 = glm(hta ~ imc + edad, data = controles, family = "binomial")
nuevos = expand.grid(edad = c(30, 50, 70), imc = 15:40)
nuevos$proba = predict(or2, nuevos, type = "response")

ggplot2::qplot(imc, proba, data = nuevos, color = factor(edad), geom = "line")
```


0.6
0.4
0.2
0.0
proba
15
20
25
30
35
40
imc
factor(edad)
30
50
70