MODELIZACIÓN ESTADÍSTICA 2021-2022
Máster en modelización e investigación matemática, estadística y computación

Evaluación continua 2

Aitor Lorenzo Ramírez Cabrera

1. Teoría: Define los siguientes conceptos estadísticos y explica su significado

- **Puntuación tipificada:** Se utiliza para comparar las posiciones relativas de varios elementos con respecto al conjunto de observaciones. Una puntuación tipificada se puede calcular como:

$$z_i = \frac{x_i - \bar{X}}{s} \tag{1}$$

donde $\bar{X}$ es la media de las observaciones y s la desviación estándar. Como podemos observar, z_i no es más que el número de desviaciones estándar que x_i se desvía de la media. La puntuación tipificada es útil para comparar datos procedente de diferentes muestras.

- **Coeficiente de correlación de Pearson:** Es una prueba que mide la dependencia lineal entre dos variables cuantitativas continuas.
Este puede tomar valores en un rango $[-1, 1]$, siendo 0 la prueba de que no hay asociación entre las variables. Un valor mayor que cero indica una asociación positiva, esto es, a medida que aumenta una variable también lo hace la otra. Por otra parte, un valor menor que cero indica una asociación negativa, esto es, a medida que aumenta una variable la otra disminuye.
Para una población, dado un par de variables aleatorias (X, Y) se define como:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{2}$$

Mientras que, para una muestra dada por n pares de datos $\{(x_i, y_i)\}_{i=1}^n$ se define como:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{3}$$

- **Intervalo de confianza:** Es un par de números entre los cuales se encontrará la estimación puntual buscada. El intervalo de confianza nos permite calcular dos valores alrededor de una media muestral que acoten un rango dentro del cual se va a localizar el parámetro poblacional.

- **Región crítica de un test:** La región crítica a un nivel de significación α representa el subconjunto del espacio muestral tal que la probabilidad de que la muestra aleatoria simple pertenezca a esta. Cuando se cumple la hipótesis nula, H_0, esta es igual a α, es decir:

$$Pr((\xi_1, \xi_2, ..., \xi_n) \in C | H_0) = \alpha \tag{4}$$

La regla de decisión en un test quedará definida de acuerdo a una región crítica. Si la muestra obtenida se ubica dentro de la región crítica, rechazamos la hipótesis nula H_0. En caso contrario, no rechazamos la hipótesis nula.

- **p-valor del contraste:** Es la probabilidad de obtener un valor del estadístico al menos tan extremo como el que se ha observado si la hipótesis nula es cierta. Se puede decir que representa la probabilidad de observar la muestra cuando la hipótesis nula es cierta.
Si el p-valor es muy pequeño ($p < 0{,}05$) la muestra es poco compatible con que H_0 sea cierta y se rechaza H_0.
Si el p-valor no es pequeño ($p \geq 0{,}05$), la muestra es compatible con que H_0 sea cierta y no se rechaza.

- **Regresión logística:** Es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa.

2. Simulación

2.1. Analiza mediante simulación el efecto que producen los modelos de regresión diversos factores.

Para empezar, generamos las variables independientes de la simulación tal y como se muestra a continuación:

```
n <- 40 #Tamaño de muestra al menos de 3 observaciones

set.seed(1); x1 <- rnorm(n) #Genero la variable aleatoria x1
#Para generar la segunda variable con correlación de poisson >0.1 creo la función
corr.data<- function(x1, rho){
  set.seed(7);xr <- rnorm(length(x1)) #Variable random con distribución normal

  xcorr<- rho*x1 + sqrt(1-rho^2)*xr #Genero la variable correlada

  return(xcorr)
}
x2 <- corr.data(x1, 0.2) #Genero x2 con correlación 0.2 con respecto a x1
set.seed(2); x3 <- rnorm(n) #Genero la variable aleatoria x3
```

Listing 1: Generación de variables independientes

A continuación, para calcular la variable dependiente defino unos valores cualesquiera para los β, unos residuales aleatorios de media nula y desviación típica σ y aplico la fórmula de regresión lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + res \tag{5}$$

para el caso de estudio, los valores de β elegidos son:

$$\beta_0 = 10 \quad ; \quad \beta_1 = 5 \quad ; \quad \beta_2 = 23 \quad ; \quad \beta_3 = 15$$

Ahora, teniendo la variable dependiente y las independientes, pasamos a crear el modelo de regresión lineal con la función *lm()* de R de manera que obtenemos:

```
> summary(reg1)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3047 -3.2820 -0.5669  2.6613  9.3872

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.9420     0.7825  12.706 7.20e-15 ***
x1            5.2190     0.8972   5.817 1.22e-06 ***
x2           22.8265     0.7461  30.596  < 2e-16 ***
x3           14.3133     0.6934  20.641  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.755 on 36 degrees of freedom
Multiple R-squared:  0.978, Adjusted R-squared:  0.9762
F-statistic: 533.7 on 3 and 36 DF,  p-value: < 2.2e-16
```

Listing 2: Resultado de la regresión lineal

En Listing 2 podemos observar que, como era de esperar, los valores de β son similares a los que habíamos definido anteriormente. Además, en todos los casos el p-valor es mucho menor que 0,05 por lo tanto rechazamos la hipótesis nula, es decir, rechazamos que alguno de estos β sea nulo.

Procedemos a analizar el efecto que producen en los modelos de regresión los siguientes factores:

2.1.1. Un punto de influencia

Para comprobar esto, sabemos que un punto de influencia es un valor atípico que afecta a la pendiente de la línea de regresión. Por lo tanto, podemos alterar alguno de los valores del cualquier variable independiente de manera que sea mucho mayor que el resto.

Empezaremos comprobando que pasa si el punto se altera en la variable que tiene el menor β. En el caso de estudio se tomó la quinta componente de x_1 y se le sumó quince unidades para obtener un valor mucho mayor que el resto. Los resultados de la regresión obtenida en este caso fueron:

```
> summary(reg2)

Call:
lm(formula = y ~ x1.1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9560 -4.9148  0.0496  3.1637 14.5616

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6070     1.0415   9.224 5.14e-11 ***
x1.1          0.8618     0.3888   2.216   0.0331 *
x2           24.0877     0.9464  25.452  < 2e-16 ***
x3           14.8765     0.8957  16.609  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.213 on 36 degrees of freedom
Multiple R-squared:  0.9625,  Adjusted R-squared:  0.9593
F-statistic: 307.7 on 3 and 36 DF,  p-value: < 2.2e-16
```

Listing 3: Regresión lineal con un punto de influencia

Comparando los resultados obtenidos en Listing 3 con los de Listing 2 observamos que apenas hay diferencia en los valores de β_0, β_2 y β_3 (asociados a las variables que no han sido modificadas). Pero sí que hay gran diferencia entre las β_1 ya que vemos que esta pasa de ser 5,22 a ser 0,86. Además, se observa diferencia en el p-valor de esta última, que es mucho mayor, aunque seguimos rechazando la hipótesis de que β_1 pueda ser nula.

A modo de curiosidad podemos comprobar que pasaría si el punto se altera en la variable que tiene el β mayor. En el caso de estudio se tomó la quinta componente de β_2 y se le sumó veinte unidades. Los resultados de la regresión obtenida en este caso fueron:

```
> summary(reg3)

Call:
lm(formula = y ~ x1 + x2.1 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-45.536 -15.191  -4.421  12.978  62.121

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.069      4.014   3.754 0.000613 ***
x1            11.629      4.507   2.580 0.014107 *
x2.1           1.062      1.256   0.846 0.403396
```

```
x3            13.571      3.569   3.802 0.000534 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.47 on 36 degrees of freedom
Multiple R-squared:  0.4178,  Adjusted R-squared:  0.3693
F-statistic: 8.612 on 3 and 36 DF,  p-value: 0.0001931
```

Listing 4: Regresión lineal con un punto de influencia

En este caso se observa que el efecto es mucho mayor ya que, para empezar, existe más error en el cálculo de los coeficientes de regresión β. Con respecto a los p-valores, se observa que estos son mucho mayores que en Listing 2, llegando incluso a no rechazar la hipótesis nula para β_2, es decir, esta variable no sería significativa en el estudio. Se observa, además, que la desviación estándar de los predictores es mucho mayor que la obtenida en el modelo inicial. Por último, llama la atención el R^2, que es mucho menor que el de la regresión original.

2.1.2. La multicolinealidad

La multicolinealidad es la alta correlación entre dos variables explicativas de una regresión. Para simular una regresión en la que existe multicolinealidad podemos definir que una de las variables sea igual a otra más unos residuos random de manera que obtengamos una correlación alta. En el caso de estudio se definió:

$$x_{2.2} = 2x_1 + N(0, 1)$$

Los resultados obtenidos para la regresión en este caso fueron:

```
> summary(reg4)

Call:
lm(formula = y ~ x1 + x2.2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-41.107 -16.572  -5.198  12.772  64.116

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.0409     4.0206   3.990 0.000311 ***
x1           13.7511     8.9001   1.545 0.131081
x2.2         -0.9659     4.5815  -0.211 0.834215
x3           13.5471     3.6409   3.721 0.000675 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.7 on 36 degrees of freedom
Multiple R-squared:  0.407, Adjusted R-squared:  0.3576
F-statistic: 8.235 on 3 and 36 DF,  p-value: 0.000266
```

Listing 5: Regresión lineal con multicolinealidad

En este caso se observa que, tanto el coeficiente independiente de correlación como los asociados a las dos variables correladas, varían bastante con respecta a las definidas inicialmente. Con respecto a los p-valores, estos son mucho mayores que los obtenidos en Listing 2, llegando incluso a no rechazar la hipótesis de que β_1 o β_2 puedan ser nulas. Al igual que en el caso anterior, se observa un valor muy bajo para R^2 que, junto a los altos errores estándar y a los p-valores, nos podrían indicar la existencia de multicolinealidad.

2.1.3. Una especificación inadecuada del modelo

Una especificación inadecuada del modelo puede venir dada por la omisión de regresores relevantes en el modelo. Teniendo en cuenta lo obtenido en Listing 2 podemos plantear que pasaría si quitamos una de las variables con un p-valor menor, por ejemplo x_3:

```
> summary(reg5)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-43.287  -8.914  -1.812  11.128  33.322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.261      2.756   4.086 0.000226 ***
x1             8.047      3.133   2.568 0.014388 *
x2            22.185      2.634   8.422 3.97e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.8 on 37 degrees of freedom
Multiple R-squared:  0.7178,  Adjusted R-squared:  0.7025
F-statistic: 47.05 on 2 and 37 DF,  p-value: 6.852e-11
```

Listing 6: Especificación inadecuada del modelo

En este caso, la estimación de los coeficientes de regresión es similar a la propuesta inicialmente pero se observa que el error estándar es mayor que el obtenido en Listing 2. Comparando los p-valores, en todos los casos se rechaza la hipótesis nula, es decir, todas las variables son significativas, pero estos son mayores que los obtenidos en el modelo inicial.

A continuación comprobamos que ocurre si la especificación inadecuada se produce por la falta de la variable que definimos con un mayor coeficiente de regresión, es decir, si quitamos del análisis x_2.

```
> summary(reg7)

Call:
lm(formula = y ~ x1 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-40.85 -16.45  -5.28  12.88  64.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.869      3.886   4.084 0.000228 ***
x1            12.133      4.450   2.726 0.009728 **
x3            13.430      3.551   3.782 0.000551 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.37 on 37 degrees of freedom
Multiple R-squared:  0.4062,  Adjusted R-squared:  0.3741
F-statistic: 12.66 on 2 and 37 DF,  p-value: 6.483e-05
```

Listing 7: Especificación inadecuada del modelo

En este caso se observa que el R^2 es mucho mejor que en los casos anteriores, demostrando así la especificación inadecuada del modelo.

2.2. Simula un conjunto de datos a partir de un modelo de regresión logística

Para empezar, conocemos que la probabilidad para una regresión logística tiene la forma:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \beta_{2s} x_{2s,i} + \beta_{2n} x_{2n,i})}} \tag{6}$$

de manera que, para acotar el programa, nos interesa tomar un rango de valores pequeño que asegure que p_i abarque todo el intervalo $[0, 1]$. Si definimos:

$$K_i = \beta_0 + \beta_1 x_{1,i} + \beta_{2s} x_{2,i} + \beta_{2n} x_{2,i} \tag{7}$$

Obtenemos que, para que la probabilidad abarque todo ese intervalo, los K_i se deben encontrar en el intervalo $[-5, 5]$.

A continuación, podemos definir una variable cualitativa que tome, de manera aleatoria, el valor s o el valor n. Y, para definir la variable cuantitativa, podemos despejar de la expresión de K_i de manera que obtenemos:

$$x_{1,i} = \frac{K_i - \beta_{2s} x_{2,i} - \beta_{2n} x_{2,i} - \beta_0}{\beta_1} \tag{8}$$

En el caso de estudio, los valores de β elegidos son:

$$\beta_0 = 5 \quad ; \quad \beta_{1s} = 3 \quad ; \quad \beta_{1n} = 2 \quad ; \quad \beta_2 = 3$$

A continuación, se generan unos nuevos valores K' a partir de estas variables y se le suma una componente aleatoria con una distribución $N(0, 0.8)$. Finalmente, para obtener la variable respuesta, se calculan nuevas probabilidades aleatorias a partir de este nuevo K' y se simula la asignación de $Y = 1$ según una distribución binomial dependiente de x. En R esta simulación se consigue haciendo uso de la función *rbinom()*.

Teniendo el conjunto de datos simulados, podemos hacer una regresión logística en R de manera que obtenemos:

```
> modelo <- glm(y~x1+x2, family = binomial)
> summary(modelo)

Call:
glm(formula = y ~ x1 + x2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3836  -0.8201  -0.3268   0.5984   2.0879

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.8485     2.0162   2.405   0.0162 *
x1s           0.4319     1.0133   0.426   0.6699
x2            2.0268     0.7885   2.570   0.0102 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 40.381  on 29  degrees of freedom
Residual deviance: 28.835  on 27  degrees of freedom
AIC: 34.835

Number of Fisher Scoring iterations: 5
```

Listing 8: Regresión logística

En este caso observamos que los β obtenidos son similares a los definidos inicialmente, salvo para el caso con $x_1 = s$ en el que es mucho menor que el indicado. Además, para este caso, se obtiene un p-valor de 0,67, es decir, este coeficiente podría ser nulo 67 de cada 100 veces.

3. Anexo: Código R

El código usado para resolver los ejercicios se muestra a continuación:

```r
##################
#Primer ejercicio
##################

n <- 40 #Tamaño de muestra al menos de 3 observaciones

set.seed(1); x1 <- rnorm(n) #Genero la variable aleatoria x1
#Para generar la segunda variable con correlación de poisson >0.1 creo la función
corr.data<- function(x1, rho){
  set.seed(7);xr <- rnorm(length(x1)) #Genero una variable random con distribución
    normal

  xcorr<- rho*x1 + sqrt(1-rho^2)*xr #Genero la variable correlada

  return(xcorr)
}
x2 <- corr.data(x1, 0.2) #Genero la variable aleatoria x2 con correlación 0.1 con
    respecto a x1
set.seed(2); x3 <- rnorm(n) #Genero la variable aleatoria x3

cor(x1,x2)

#Defino los coeficientes de la regresión
beta0 <- 10; beta1 <- 5; beta2 <- 23; beta3 <- 15

#Defino los residuales
sigma <- 4
residuales = rnorm(n, 0, sigma)

#Calculo la variable respuesta
y <- beta0 + beta1*x1 + beta2*x2 + beta3*x3 + residuales

#Creo el modelo de regresión
reg1 <- lm(y~x1+x2+x3)
summary(reg1) #Vemos que en todos los casos las betas son parecidas a las que hemos
    definidos y los p-valores las consideran significativas


#Si se añade un punto de influencia
#Al punto con el beta menor
x1.1 <- x1
x1.1[5] <- x1[5]+20

plot(x1.1, y)
#Hacemos la regresión para este caso
reg2 <- lm(y.1~x1.1+x2+x3)
summary(reg2)

#Al punto con el beta mayor
x2.1 <- x2
x2.1[5] <- x2[5]+20

plot(x2.1, y)
#Hacemos la regresión para este caso
reg3 <- lm(y~x1+x2.1+x3)
summary(reg3)
```

```r
#Multicolinealidad: relación de dependencia lineal fuerte entre más de dos variables explicativas de una regresión múltiple
#Vamos a suponer multicolienalidad entre x1 y x2
set.seed(16); x2.2 <- 2*x1 + rnorm(n, mean = 0, sd = 1) #Defino x2 como x1 más una variable aleatoria
cor(x1, x2.2) #Veo que la correlación es alta, por lo tanto, hay multicolinealidad
#Hacemos la regresión para este caso
reg4 <- lm(y~x1+x2.2+x3)
summary(reg4)


#Especificación inadecuada del modelo
#Una posible causa de la especificación inadecuada del modelo puede se devida a la omisión de regresores relevante
reg5 <- lm(y~x1+x2)  #Quitamos la tercera que tenía un p-valor muy pequeño
summary(reg5)  #Vemos que los p aumentan y el valor de beta varía

reg6 <- lm(y~x2+x3)
summary(reg6)

#Pruebo a quitar valores de x2
reg7 <- lm(y~x1+x3)
summary(reg7)









##################
#Segundo ejercicio
##################

n <- 30 #Número de muestras

#Por la forma que tiene la función logit sabemos que la probabilidad abarca el rango [0,1] cuando logit se encuentra en el intervalo [-5,5]
set.seed(10);k <- rnorm(n, mean = 0, sd = 3)

#Defino los coeficientes de la regresión
beta0 <- 5; beta1 <- 3; beta2 <- 2; beta3 <- 3

#Defino la variable cualitativa con dos modalidades
set.seed(15); x1 <- sample(c("s", "n"), n, replace=TRUE)

#Para definir la variable cuantítativa hago uso del logit: logit = beta0 + beta*x1s + beta*x1n + beta3*x2
x2<- (k - beta0 - beta1*as.numeric(x1=='s') - beta2*as.numeric(x1=='n'))/beta3

#Los residuales se comportan como una distribución normal con media cero y desviación típica sigma
sigma <- 0.8
residuales <- round(rnorm(n, 0, sigma), 1) #Los redondeo a un decimal

#Calculo el logit
k_aleatorio <- beta0 + beta1*as.numeric(x1=='s') + beta2*as.numeric(x1=='n') + beta3*x2 + residuales

#Calulo la probabilidad
p <- 1/(1+exp(-k_aleatorio))

#Simulo las variables respuesta según una distribución binomial dependiente de x
```

```r
y <- rbinom(n, 1, prob = p )

#Estimo el modelo a partir de los parámetros simulados
modelo <- glm(y~x1+x2, family = binomial(logit))
summary(modelo)
exp(coef(modelo))
```

Este se puede consultar también en el siguiente *link*: https://github.com/AitorLRC/Evcont2_ME/blob/2aed07162522b074ef2356dd28a0fdc7a0061ad2/continua2.R