
Pontificia Universidad Católica de Valparaíso

Instituto de Estadística - Magíster en Estadística

EST741 - Taller de consultoría - Ciclo 02 2019

# Caso de aplicación 2.

# Identificación de factores de riesgo en pacientes de diabetes post pabellón.

*Matías F. Rebolledo G.*

*Martes, 8 de Octubre de 2019*

# Índice general



# Índice de cuadros



# Índice de gráficos

# 1. Definición del caso.

El presente informe busca indentificar los factores de riesgo asociados a pacientes diagnosticados de diabetes en un centro médico determinado. Se ha definido como **caso** a aquel paciente que después de 30 días de realizarle una biopsia ha presentado algún tipo de síntoma que haya requerido hospitalización. Mientras que se definió como **control** a aquel paciente que no presentó ninguna de complicación en un periodo posterior a 30 días que haya requerido hospitalización.

La metodología consiste en identificar los factores de riesgo por medio de la construcción de un modelo de clasificación. Gracias a este modelo podremos someter un paciente ficticio con características predefinidas y poder medir cuales son los factores de riesgo asociados al vector de características de ese paciente o de un conjunto de prueba que desee incorporarse en cuestión.

La muestra ha sido obtenida mediante la creación de un cuestionario que fue respondido por cada paciente y cuyas preguntas están asociadas a procedimientos que fueron aplicados según la sintomatología de cada paciente. El tamaño de la muestra es de 284 observaciones y un número de 21 variables. El número de pacientes que corresponden a casos es de 12, mientras que el que corresponde a control es de 272.

Años del estudio (toma de observaciones) - 2012, 2013

# 2. Análisis y propuesta de solución.

En esta sección plantearemos posible soluciones en virtud del análisis que realizaremos abordando tanto principios estadísticos como aprovechando la misma estructura de la muestra del estudio. Por un lado, realizaremos todo análisis posible en temas descriptivos y por otro, evaluaremos varias opciones para construir modelos y las formas de abordar la validez de los mismos.

Una de las características está dado por el pequeño tamaño muestral. A eso se suma que la categoría de interés que corresponde a un paciente tipo caso, equivale a solamente un 4.23 %, mientras que aquellos que son control corresponden a un 95.77 %.

## 2.1. Análisis exploratorio.

### 2.1.1. Combinaciones lineales con la variable dependiente.

Como paso inicial a nuestro análisis descriptivo evidenciamos que existen variables que son combinación lineal casi perfecta con la variable de respuesta del estudio. Inicialmente esto está planteado en el mismo enunciado, al definir como **caso** a aquellos pacientes que sufrieron fiebre e infección urinaria en los primeros 30 días después del procedimiento de la biopsia. Este es el primer indicio lo que vamos a sustentar realizando algunas tabulaciones y visualizaciones para tener una evidencia más sólida.

Cuadro 1: Matriz de correlaciones aplanada y ordenada decreciente

| cor_x | cor_y | cor | Pval |
|---|---|---|---|
| HOSPITALIZACION | FIEBRE | 0,9590 | 0 |
| PATRON.DE.RESISTENCIA | TIPO.DE.CULTIVO | 0,9421 | 0 |
| FIEBRE | NUMERO.DE.DIAS.POST.BIOPSIA | 0,9019 | 0 |
| PATRON.DE.RESISTENCIA | AGENTE.AISLADO | 0,8958 | 0 |
| AGENTE.AISLADO | ITU | 0,8714 | 0 |
| PATRON.DE.RESISTENCIA | ITU | 0,8415 | 0 |
| HOSPITALIZACION | DIAS.HOSPITALIZACION.MQ | 0,8075 | 0 |
| HOSPITALIZACION | NUMERO.DE.DIAS.POST.BIOPSIA | 0,7871 | 0 |
| DIAS.HOSPITALIZACION.MQ | AGENTE.AISLADO | 0,7781 | 0 |
| AGENTE.AISLADO | TIPO.DE.CULTIVO | 0,7781 | 0 |
| DIAS.HOSPITALIZACION.MQ | FIEBRE | 0,7744 | 0 |
| TIPO.DE.CULTIVO | ITU | 0,7425 | 0 |

| cor_x | cor_y | cor | Pval |
|---|---|---|---|
| DIAS.HOSPITALIZACION.MQ | NUMERO.DE.DIAS.POST.BIOPSIA | 0,7132 | 0 |
| DIAS.HOSPITALIZACION.MQ | PATRON.DE.RESISTENCIA | 0,6231 | 0 |
| DIAS.HOSPITALIZACION.MQ | ITU | 0,6092 | 0 |
| HOSPITALIZACION | AGENTE.AISLADO | 0,5373 | 0 |
| HOSPITALIZACION | PATRON.DE.RESISTENCIA | 0,5188 | 0 |
| AGENTE.AISLADO | FIEBRE | 0,5153 | 0 |
| DIAS.HOSPITALIZACION.UPC | TIPO.DE.CULTIVO | 0,5139 | 0 |
| HOSPITALIZACION | TIPO.DE.CULTIVO | 0,5136 | 0 |

El cuadro 1 muestra las primeras 20 correlaciones correspondientes a la matriz de correlacion en formato aplanado y ordenadas en forma decreciente por el coeficiente de correlación de Pearson. Los valores de probabilidad para el coeficiente de Pearson resultaron ser demasiado pequeños y muy significativos, por lo que resultaron ser cero. El tamaño de la matriz de correlaciones es de $P(P-1)/2$ y equivale a 210. Las restantes 190 filas poseen un coeficiente de Pearson inferior a 0,5 y el coeficiente más negativo es de -0.21, lo que no indica una fuerte correlación negativa en la muestra. Por su parte, la correlación positiva más elevada es de 0.96 y corresponde a que la variable FIEBRE está es casi una combinación lineal de la variable dependiente HOSPITALIZACION. Veremos la frecuencia conjunta en la siguiente tabulación cruzada.

```
with( dfmbio , table( HOSPITALIZACION , FIEBRE ) )
```

```
##                FIEBRE
## HOSPITALIZACION  NO  SI
##              NO 271   1
##              SI   0  12
```

Lo representado en la tabulación anterior implica que hubo solo un paciente que sí tuvo fiebre pero no fue identificado como caso. A pesar de esto, esta variable tiene una implicancia negativa en la construcción de un modelo por lo que optaremos por excluirla de nuestro análisis. Volveremos a inspeccionar la matriz aplanada de correlaciones para observar otras variables que estén correlacionadas con la variable HOSPITALIZACION y evaluar su pertinencia.

```
head( dfmbio_num_cor [ dfmbio_num_cor$cor_x == "HOSPITALIZACION" , ] ,  6 )
```

```
##              cor_x                       cor_y       cor Pval
## 1  HOSPITALIZACION                      FIEBRE 0.9590012    0
## 7  HOSPITALIZACION     DIAS.HOSPITALIZACION.MQ 0.8074868    0
```

```
## 8  HOSPITALIZACION NUMERO.DE.DIAS.POST.BIOPSIA 0.7870693    0
## 16 HOSPITALIZACION              AGENTE.AISLADO 0.5372837    0
## 17 HOSPITALIZACION       PATRON.DE.RESISTENCIA 0.5188452    0
## 20 HOSPITALIZACION             TIPO.DE.CULTIVO 0.5135695    0
```

Se observa que la segunda variable con la mayor correlación positiva es DIAS.HOSP.MQ. Observaremos la tabulación cruzada de frecuencia con la variable dependiente.

```
with( dfmbio , table( HOSPITALIZACION , DIAS.HOSP.MQ ) )
```

```
##                DIAS.HOSP.MQ
## HOSPITALIZACION   0   2   3   4   5  12
##              NO 272   0   0   0   0   0
##              SI   0   5   2   3   1   1
```

Esta tabulación tiene dos problema visibles. Por un lado, la variable DIAS.HOSP.MQ tiene alta correlación con HOSPITALIZACION, mientras que por otro, si se categoriza esta variable para todos número de días mayor que cero y se vuelve a hacer esta tabla de contingencia tendríamos colinealidad perfecta con la variable dependiente. Por lo tanto, esta variable tampoco muestra utilidad y optamos por excluirla de nuestros análisis. Para ejemplificar lo antes señalado volvemos a mostrar la tabulación cruzada esta vez con una variable recodificada nueva DIAS.HOSP.MQ.DUM versus HOSPITALIZACION.

```
dfmbio_num_dum <- dfmbio_num [ , c(21,19,13) ]
#str(dfmbio_num_dum)
dfmbio_num_dum$DIAS.HOSP.MQ.DUM <- 0
dfmbio_num_dum$DIAS.HOSP.MQ.DUM [ dfmbio_num_dum$DIAS.HOSP.MQ > 0 ] <- 1
with( dfmbio_num_dum , table( HOSPITALIZACION , DIAS.HOSP.MQ.DUM ) )
```

```
##                DIAS.HOSP.MQ.DUM
## HOSPITALIZACION   0   1
##               0 272   0
##               1   0  12
```

De donde observamos una colinealidad perfecta entre ambas variables. Repetiremos la misma inspección con la variable NUM.DIAS.POST.BIOP para evaluar si a priori sería conveniente continuar con esta variable dentro de nuestra base de datos.

```
with( dfmbio_num , table( HOSPITALIZACION , NUM.DIAS.POST.BIOP ) )
```

```
##                NUM.DIAS.POST.BIOP
## HOSPITALIZACION   0   1   2   3   4   5
##               0 271   0   0   0   0   1
##               1   0   3   4   4   1   0
```

Observamos un problema similar al de las últimas dos variables analizadas. En primera instancia ambas muestran una correlación de Pearson de 0.79 y en caso de recodificar esta variable obtendría colinealidad casi perfecta como veremos a continuación.

```
dfmbio_num_dum$NUM.DIAS.POST.BIOP.DUM <- 0
dfmbio_num_dum$NUM.DIAS.POST.BIOP.DUM [ dfmbio_num_dum$NUM.DIAS.POST.BIOP > 0 ] <- 1
with( dfmbio_num_dum , table( HOSPITALIZACION , NUM.DIAS.POST.BIOP.DUM ) )
```

```
##                NUM.DIAS.POST.BIOP.DUM
## HOSPITALIZACION   0   1
##               0 271   1
##               1   0  12
```

En síntesis, optaremos por dejar fuera de análisis a las últimas tres variables FIEBRE, DIAS.HOSP.MQ y NUM.DIAS.POST.BIOP por tener una colinealidad muy elevada con la variable dependiente lo que empobrece los principios del modelo que se quiere construir.

### 2.1.2. Combinaciones lineales entre variables independientes.

De forma similar nos compete realizar la misma verificación esta vez entre variables independientes de la muestra con la finalidad de identificar posibles combinaciones lineales entre ellas. A continuación mostramos la tabulación cruzada para las variables con correlación de Pearson positiva más elevada de la matriz de correlación que fue previamente aplanada.

```
head( dfmbio_num_cor [ dfmbio_num_cor$cor_x != "HOSPITALIZACION" , ] , 14 )
```

```
##                       cor_x                      cor_y       cor Pval
## 2     PATRON.DE.RESISTENCIA           TIPO.DE.CULTIVO 0.9421373    0
## 3                    FIEBRE NUMERO.DE.DIAS.POST.BIOPSIA 0.9019470    0
## 4     PATRON.DE.RESISTENCIA             AGENTE.AISLADO 0.8958262    0
## 5            AGENTE.AISLADO                        ITU 0.8714283    0
## 6     PATRON.DE.RESISTENCIA                        ITU 0.8415226    0
## 9   DIAS.HOSPITALIZACION.MQ             AGENTE.AISLADO 0.7781219    0
```

```
## 10           AGENTE.AISLADO             TIPO.DE.CULTIVO 0.7780602    0
## 11  DIAS.HOSPITALIZACION.MQ                      FIEBRE 0.7743808    0
## 12          TIPO.DE.CULTIVO                        ITU 0.7425254    0
## 13  DIAS.HOSPITALIZACION.MQ NUMERO.DE.DIAS.POST.BIOPSIA 0.7131912    0
## 14  DIAS.HOSPITALIZACION.MQ       PATRON.DE.RESISTENCIA 0.6231013    0
## 15  DIAS.HOSPITALIZACION.MQ                        ITU 0.6092084    0
## 18           AGENTE.AISLADO                      FIEBRE 0.5152557    0
## 19 DIAS.HOSPITALIZACION.UPC             TIPO.DE.CULTIVO 0.5138981    0
```

Se trata de un fragmento del cuadro 1 que presenta los pares de variables con la correlación de Pearson más elevada. En este caso corresponde a las primeras 16 combinaciones de variables que no incluyen a la variable dependiente para analizar la colinealidad solamente entre covariables. La correlación positiva más alta es entre PATR.RESIST y TIPO.CULTIVO que es de 0.94. Veremos a continuación una tabla de contigencia de la frecuencia entre ambas variables.

```
with( dfmbio , table( abbreviate(PATR.RESIST) ,
abbreviate(TIPO.CULTIVO) )[c(3,1,4,5,2),c(3,1,4,2)] )
```

```
##
##                NO HEMO UROC HEYU
##   NO          280    0    0    0
##   ARCRGRSMR     0    1    0    0
##   RAACYG        0    0    1    0
##   RAASCCCYCC    0    0    0    1
##   MULS          0    0    1    0
```

Se puede observar que la distribución conjunta es muy similar a aquellas en que hubo colinealidad con la variable dependiente. La mayor parde de las observaciones se concentran en aquellas que no presentan patrón de resistencia, puesto que la mayoría de pacientes no presentó a su vez infecciones urinarias, ni fiebre, o bien, tuvieron resultado negativo en la biopsia, por lo que no se puede medir en ellos si después de aplicar un tratamiento con antibióticos, se puede decir que haya habido una resistencia de parte de los parásitos al tratamiento prescrito según la profilaxis de cada paciente. A continuación analizaremos la frecuencia conjunta entre las covariables PATR.RESIST y AGENTE.AISLADO, cuyo coeficiente de correlación de Pearson es de 0.9.

```
with( dfmbio , table( abbreviate(PATR.RESIST) ,
abbreviate(AGENTE.AISLADO) )[c(3,1,4,5,2),c(2,1,3)] )
```

```
##
```

```
##                NO E.CO PSEA
##   NO          280    0    0
##   ARCRGRSMR     0    1    0
##   RAACYG        0    1    0
##   RAASCCCYCC    0    1    0
##   MULS          0    0    1
```

### 2.1.3. Identificación de variables con varianza cero o casi cero.

Esta sección nos ayudará a sintetizar las últimos dos análisis, puesto que podemos clasificar todas las variables independientes según si su varianza es casi cero o no. Esto ha sido señalado en la siguiente fuente[1] y por los autores (Wing et al. 2019) y (Zorn 2005), quienes señalan que, por un lado, se puede optar por excluir toda variable que estropee las propiedades del modelo o bien que al ser colineales impidan ejecutar la inversión de la matriz de información, pero por otro, tomando en cuenta que a veces hay variables que por razones de ámbito deben permanecer en el modelo aunque no tengan las mejores propiedades o no aporten mucha información, vale la pena abordar en forma definitiva este último enfoque para poder concluir con decisiones sobre las variables que resultan ser combinaciones lineales entre ellas o con la variable objetivo del estudio. Con estas últimas tres secciones y con apoyo un modelo logístico inicial aplicaremos unas pruebas de significancia parcial entre posibles variables a excluir y mediremos las desvianzas de cada modelo posible para ver cual es mejor al momento de incluir o no una variable con poca capacidad explicativa.

En el siguiente código observamos las variables que presentan una varianza casi cero, lo que está propuesto mediante la función `caret::nearZeroVar` del autor (Wing et al. 2019).

```
dfmbio_fac_nzv <- caret::nearZeroVar( dfmbio_fac , saveMetrics=T )
dfmbio_fac_nzv [ dfmbio_fac_nzv$nzv > 0 , ]
```

```
##                    freqRatio percentUnique zeroVar  nzv
## HOSP.ULT.MES       141.00000     0.7042254   FALSE TRUE
## CUP                141.00000     0.7042254   FALSE TRUE
## EPOC                45.66667     1.4084507   FALSE TRUE
## NUM.DIAS.POST.BIOP  67.75000     2.1126761   FALSE TRUE
## FIEBRE              20.84615     0.7042254   FALSE TRUE
## ITU                 93.66667     0.7042254   FALSE TRUE
## TIPO.CULTIVO       140.00000     1.4084507   FALSE TRUE
## AGENTE.AISLADO      93.33333     1.0563380   FALSE TRUE
```

[1] Near-zero variance predictors. Should we remove them?

```
## PATR.RESIST        280.00000     1.7605634   FALSE TRUE
## DIAS.HOSP.MQ        54.40000     2.1126761   FALSE TRUE
## DIAS.HOSP.UPC      283.00000     0.7042254   FALSE TRUE
## HOSPITALIZACION     22.66667     0.7042254   FALSE TRUE
```

Se puede observar en este listado las variables que contienen varianza casi cero, lo que se corresponde con varias de las tabulaciones que se ha revisado hasta ahora, tales como, FIEBRE, DIAS.HOSP.MQ, NUM.DIAS.POST.BIOP, que son combinación lineal casi perfecta con la variable dependiente y otras que tienen la misma propiedad entre variables independientes.

En la etapa de modelación realizaremos un tanteo incluyendo y excluyendo alguna de las variables que tienen varianza casi cero para decidir finalmente si tienen, por un lado, un efecto negativo, o bien, son útiles en la identificación de los factores de riesgo.

## 2.2. Métodos de clasificación pertinentes.

Dentro de la gran variedad de métodos que existen queremos abordar los problemas relevantes para nuestra muestra de datos. Hemos visto necesitaremos abordar el fenómeno de la inflación de varianza para evitar que la estimación de errores estándar de los estimadores sea elevada y por ende le quite estabilidad a los estimadores. Además queremos abordar una estrategia para el efecto que conlleva el falta de balanceo entre las categorías de pacientes. Por último, dedicaremos una sección para abordar una solución para validar cada uno de los modelos que estimemos considerando que tenemos una muestra pequeña y no tenemos una muestra de prueba. En este sentido usaremos alguna estrategia adicional en la que podamos prescindir de una muestra de prueba, tal como, validación cruzada. Veremos como hacer sinergia entre el uso de validación cruzada, sub muestreo y sobre muestreo para lo que usaremos en toda la construcción de modelos el paquete **caret**.

### 2.2.1. Modelo logístico.

En esta sección analizaremos un modelo de clasficación logística inicial en el que visualizaremos, por un lado, el efecto de las combinaciones lineales previamente identificadas y, por otro, el efecto dado por el desbalance entre los casos y los controles en la muestra de pacientes. En este caso observaremos los factores de inflación de varianza, las desvianzas y usaremos los paquetes **caret** y **glm**. Probaremos con una variante inicial en la que construiremos un ponderador que represente las proporciones desbalanceadas entre casos y controles e incorporaremos esta ponderación en alguno de los modelos logísticos para comparar sus desvianzas con los que no consideren este ponderador.

En relación con la inflación de varianza tenemos que en las primeras estimaciones que realizamos, R nos alerta del

siguiente mensaje: `Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred`. Esto quiere decir que a pesar de que el modelo logístico puede invertir la matriz de diseño, todavía existen regresores que separan perfectamente la variable dependiente y que los estimadores de los parámetros siguen inflados. Según la opinión de analistas,[2] una de las opciones más recomendadas es, por un lado, un modelo logístico penalizado y, por el otro, un modelo logístico bayesiano. Los autores son (Zorn 2005) y (Gelman et al. 2008) para cada modelo propuesto, respectivamente.

#### 2.2.2. Clasificación binaria penalizada.

En esta sección buscamos incorporar el efecto de la inflación de varianza dada por las variables que poseen varianza cercana a cero, puesto que hemos optado por decidir sobre la inclusión final de las que presenten menos deterioro, lo que haremos usando una clasificación binaria penalizada usando los paquetes **caret** y **glmnet**. Este último paquete también es sensible a la inversión de la matriz de confusión por lo que toda exclusión previamente hecha cuenta para esta sección de manera de evaluar un último conjunto de variable que puedan prestar alguna capacidad de indentificación de los factores de riesgo más importantes para el estudio.

#### 2.2.3. Bosque aleatorio.

Según el autor (Wing et al. 2019) el problema de covariables con varianza cero o casi cero, puede bloquear la ejecución de varios modelos o que sus estimadores sean inestables, excepto aquellos basados en árboles de clasificación.

### 2.3. Métodos para validar el modelo de clasificación.

#### 2.3.1. Usando muestra de entrenamiento y de prueba.

#### 2.3.2. Usando validación cruzada.

#### 2.3.3. Usando técnicas para muestras no balanceadas.

Según el autor (Wing et al. 2019) y también señalado en esta[3] fuente, un modelo con muestra imbalanceada tenderá siempre a predecir con mayor sesgo la clase más frecuente y, por lo tanto, tenderá a estimar un alto valor para la sensitividad y uno muy bajo para la especificidad.

[2] How to deal with perfect separation in logistic regression?
[3] Illustrative Example 5: Optimizing probability thresholds for class imbalances

# 3. Experimentación.

# 4. Conclusiones.

## Glosario.

- PSA: Antígeno prostático específico.
- PROFILAXIS: Es el tratamiento preventivo de la enfermedad.
- ITU: Infección en el tracto urinario o sepsis.
- NEG: Negativo (BIOPSIA, resultado negativo).
- PROSTATITIS: Inflamación de la próstata.
- ASMA: Enfermedad respiratoria que ocurre como respuesta del sistema inmune a una reacción anómala.
- EPOC: Enfermedad respiratoria obstructiva crónica causada por agentes tóxicos o irritantes como el tabaco.
- CUP: Cateter urinario por paciente, implemento requerido como vía de evauación de orina en forma externa para pacientes con dificultad en orinar.

## Anexos.

## Referencias.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2 (4): 1360–83. doi:10.1214/08-AOAS191.

Wing, Max Kuhn Contributions from Jed, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2019. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13 (2): 157–70. doi:10.1093/pan/mpi009.