Trabajo Práctico I

Aprendizaje Estadístico - (84:04)

A. Verri Kozlowski y F. Patitucci

Contents

El dataset `diabetes2.csv` es original del National Institute of Diabetes and Digestive and Kidney Diseases. Se creó con el objetivo de predecir si un paciente tiene o no diabetes, basado en diferentes medidas contenidas en el dataset. En particular, todos los pacientes son de sexo femenino mayores de 21 a~nos de herencia Pima. Para este primer estudio, lo que buscaremos es encontrar relaciones entre otras de las variables contenidas en el dataset.

Ejercicio 1

Cargar los datos de diabetes2.csv. La variable Outcome indica si la persona es diabética (1) o no (0). Transformarla en un factor. Finalmente revisar que todas las variables contenidas en el dataframe estén correctamente definidas.

```
# data <- read.csv("data/diabetes2.csv", header = TRUE)
DT <-fread(file=file.path("data","diabetes2.csv"))
DT[,V1:=NULL]
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
integer	integer	integer	integer	integer	numeric	numeric	integer	integer
6	148	72	35	0	33.6	0.6	50	1
1	85	66	29	0	26.6	0.4	31	0
1	89	66	23	94	28.1	0.2	21	0
0	137	40	35	168	43.1	2.3	33	1
3	78	50	32	88	31.0	0.2	26	1
2	197	70	45	543	30.5	0.2	53	1
1	189	60	23	846	30.1	0.4	59	1
5	166	72	19	175	25.8	0.6	51	1
0	118	84	47	230	45.8	0.6	31	1
1	103	30	38	83	43.3	0.2	33	0

n: 539

```
DT[,Outcome:=as.factor(Outcome)]
```

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
integer	integer	integer	integer	integer	numeric	numeric	integer	factor
6	148	72	35	0	33.6	0.6	50	1
1	85	66	29	0	26.6	0.4	31	0
1	89	66	23	94	28.1	0.2	21	0
0	137	40	35	168	43.1	2.3	33	1
3	78	50	32	88	31.0	0.2	26	1
2	197	70	45	543	30.5	0.2	53	1
1	189	60	23	846	30.1	0.4	59	1
5	166	72	19	175	25.8	0.6	51	1
0	118	84	47	230	45.8	0.6	31	1
1	103	30	38	83	43.3	0.2	33	0

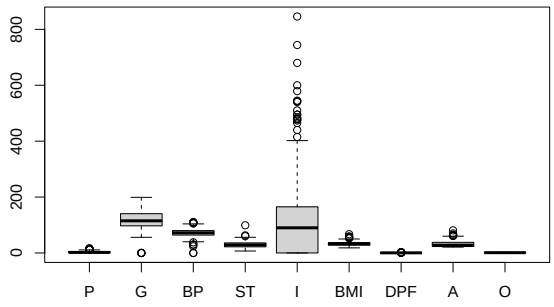
n: 539

```
setnames(DT,
         old=c("Pregnancies","Glucose" ,"BloodPressure", "SkinThickness" ,"Insulin" ,"DiabetesPedigreeFunction", "A
         new=c("P","G" ,"BP", "ST" ,"I" ,"DPF", "A" ,"O"))
```

Mediante un análisis de los datos del set, se observan muchas variables definidas como enteros, `int` que en algunos casos representan variables discretas cardinales tales como la cantidad de embarazos(`Pregnancies,`), ordinales tales como la edad `(Age)` o nominales binarias tales como`(Outcomes)` y en otros casos, representan variable reales, tales como `Age`. Para facilitar la representación de algunas figuras, se elimina la columna de índice, se renombran las variables del dataset original. El tipo de dato `int` del dataset no afecta los algoritmos del modelo de regresión y no requieren tratamiento

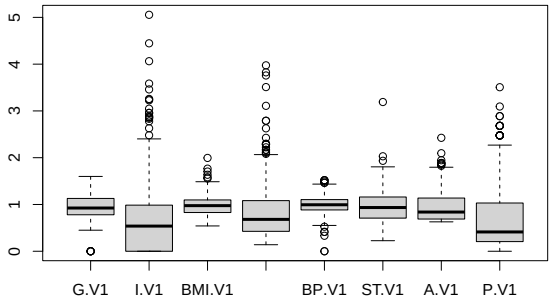
La figura siguiente muestra un diagrama de cajas (boxPlot) para las variables continuas

```
boxplot(DT)
```



La inspección del dataset indica que existen 146 observaciones con `I=0`y el primer valor diferente a 0 es 14. A partir del boxPlot puede apreciarse que la variable `I (Insulin)` tiene algunos valores que en apariencia parecen ser outliers pero eso es debido a las escalas diferentes. Los datos estandarizados (no centrados) muestran el siguiente diagrama de cajas

```
DTS <- DT[,.(G=scale(G,center = FALSE),I=scale(I,center = FALSE),BMI=scale(BMI,center = FALSE),DPF=scale(DPF,center
boxplot(DTS)
```



Ejercicio 2

Se desea ajustar un modelo de regresión múltiple para predecir la variable BMI en función del resto de las variables en el data set. Escribir el modelo propuesto, indicando los supuestos del mismo.

Sea una variable aleatoria Y definida como una función de un vector de variables aleatorias X según el siguiente modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \cdot X_2 + ... + \beta_p X_p + \varepsilon$$

,

donde $\varphi(X)$ es la mediana condicional de Y y ε es una variable aleatoria con valor medio nulo $\mathbb{E}[\varepsilon] = 0$ y varianza constante $\mathbb{VAR}[\varepsilon] = \sigma^2$

La mediana condicional $ Y (X)$ es una función expresada en términos de los parámetros β_i y para la cual nos interesa conocer un estimador $\hat{y} = \hat{\varphi}(x)$ expresado en función de un conjunto de observacions x definido como

$$Y \approx \varphi(X) = \beta_0 + \beta_1 X_1 + \beta_2 \cdot X_2 + ... + \beta_p X_p$$

El objeto de este análisis será poder encontrar un estimador de la mediana condicional a partir de los estimadores $\hat{\beta}_i$ y un vector de observaciones x según

$$\hat{\varphi}(x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \cdot x_2 + ... + \hat{\beta}_p x_p$$

Para el problema de Diabetes, se busca estimar la mediana condicional del parámetro $Y = BMI$ en función de las variables restanes y el modelo lineal queda expresado según:

$\hat{\varphi}(x) = \hat{BMI} = \beta_0 + \beta_1\ P + \beta_2 G + \beta_3\ BP + \beta_4\ ST + \beta_5\ I + \beta_6\ DPF + \beta_7\ A + \beta_8\ O + \varepsilon$

en donde $\hat{y} = \hat{BMI}$ es un estimador de mediana condicional de la variable aleatoria $Y = BMI$ y - `P (Pregnancies)`,`G (Glucose)`,`BP (BloodPressure)` `ST (SkinThickness)` `I (Insulin)` `DPF (DiabetesPedigreeFunction)`, `A (Age)`, y O es una variable ordinal binaria (lógica) $\mathbb{I}\,(Outcome = 1)$

Los supuestos del modelo son

- Los errores ε_i tienen media cero, $\mathbb{E}[\varepsilon_i] = 0$
- Los errores ε_i tienen idnéntica varianza $\mathbb{VAR}[\varepsilon_i] = \sigma^2$
- Los errores ε_i tienen distribución Normal.
- Los errores ε_i son independientes entre sí y no están correlacionados con las covariables x_i

```
Modelo <- lm( BMI ~ P + G + BP + ST + I + DPF + A + O, data = DT)
Coeff <- Modelo$coefficients
```

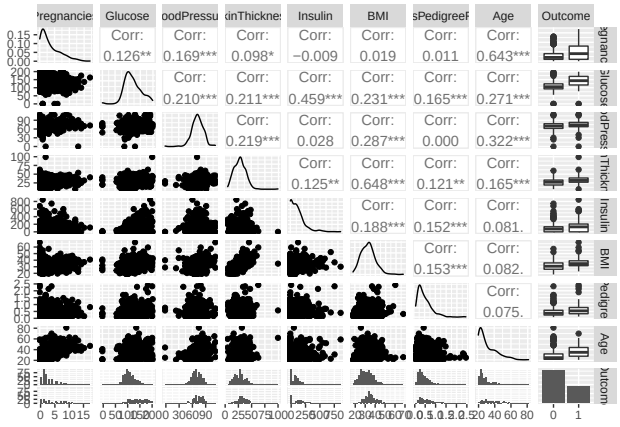
```
t(Coeff) %>% as.data.table() %>% flextable()
```

(Intercept)	P	G	BP	ST	I	DPF	A	O1
16.16141	-0.08327593	-0.002452021	0.09214472	0.3785759	0.00481634	0.8976746	-0.06367733	2.184222

Ejercicio 3

Realizar un scatterplot de las variables con la función ggpairs

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Ejercicio 4

A partir de la tabla de correlaciones estimadas entre las variables, si tuviera que elegir una sola variable para proponer un modelo de regresión simple, ¿cuál eligiríaa y por qué?

Para que se obtenga un buen modelo de predicción mediante una única variable lo que se debe buscar es que el coeficiente de correlación entre las variables en valor absoluto sea lo más cercano a 1 posible. Ya que mientras mas cerca del 1 nos encontremos implicara que existe una relación "mas lineal" entre estas dos. En

caso de ser positiva nos habla de que ambas crecen en conjunto, en caso de ser negativa nos indica que si una decrece la otra crece. Cuando el coeficiente de correlación se acerca en modulo a cero implica que las variables no tienen una relación lineal. Hasta llegar al caso en que la correlación es nula que nos indica que las variables están descorrelacionadas.

- Si se quiere predecir la variable Pregnancies elegiría Age
- Si se quiere predecir la variable Glucose elegiría Insulin
- Si se quiere predecir la variable Blood Pressure elegiría Age
- Si se quiere predecir la variable Skin Thickness elegiría BMI
- Si se quiere predecir la variable Insulin elegiría Glucose
- Si se quiere predecir la variable BMI elegiría Skin Thickness
- Si se quiere predecir la variable Diabetes Pedigree Function elegiría Glucose (en caso de facilitar el análisis podrían ser utilizadas BMI o Insulin)
- Si se quiere predecir la variable Age elegiría Pregnancies

Se debe tener en cuenta que para este análisis se tuvo en cuenta el coeficiente de correlación de las variables, pero para un análisis más exhaustivo se debe tener presente la complejidad de medición de cada variable, además de otros factores particulares de esta área.

Ejercicio 5

Realizar un ajuste de regresión lineal múltiple. A partir de la tabla de coeficientes estimados, ¿Qué variables resultan significativas? ¿A qué nivel? ¿Cuál es el valor de la estimación para σ^2? Especificar las hipótesis nulas y alternativas para alguno de los test t reportados en la tabla, el estadístico del test y la regla de decisión. ¿Cómo se calcula el p-valor para este test?

Antes de plantear el test de significancia, es importante notar que para esto se agrega el supuesto a nuestro modelo de que los errores tienen una distribución conjuntamente normal. Por lo tanto para un X fijo

$Y \sim N_n\left(X\beta, \sigma^2 I\right)$

A partir de esto se plantea los siguientes test de hipótesis ($i \in [1, 7]$, no se testea para β_0 ya que el mismo actúa como un pasabajos, mejorando nuestro modelo)

Test

$H_0 : \beta_i = 0 \, vs \, H_1 : \beta_i \neq 0$

La regla de decision a utilizar es

$$\phi_i(X) = \begin{cases} 1 & si \quad |T_i| > K_{\alpha_i} \\ \\ 0 & caso\ contrario \end{cases}$$

Donde T_i es nuentro estadistico, el cual tiene distribucion T de Student con $n-p$ grado de libertad, en este caso el grado de libertad es 530.

$T_i = \frac{\hat{\beta}_i}{S \cdot d_{ii}}$

Donde d_{ii} es el elemento ii de la matriz $D = (X^t X)^{-1}$

De esta manera calculando el estadístico T para cada uno de los betas bajo la suposición de que la hipótesis nula es correcta, se debe calcular mediante la distribución T de Student, la probabilidad de obtener una realización igual o más extrema como la obtenida con la realización observada, esta probabilidad es el p valor.

Las variables que se consideran significativas son todas aquellas las cuales tienen un Nivel de Significancia $NS \leq 0.05$, para este análisis basta con entrar al summary de la regresión realizada en R y ver uno por uno el p_{valor} asociado a cada variable. Por lo tanto:

- BloodPressure ($p \approx 0$)
- SkinThickness ($p \approx 0$)
- Insulin ($p = 0.0157$)
- Age ($p = 0.0234$)
- Outcome1 ($p \approx 0$)

La estimación de σ^2 es el cuadrado de Residual standard error brindado por summary, en este caso es de 24.8

```
summary(Modelo)
```

```
##
## Call:
## lm(formula = BMI ~ P + G + BP + ST + I + DPF + A + O, data = DT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4933  -3.3207  -0.6014   3.1331  22.2634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.161413   1.440169  11.222  < 2e-16 ***
## P           -0.083276   0.085507  -0.974   0.3305
## G           -0.002452   0.008304  -0.295   0.7679
## BP           0.092145   0.017893   5.150 3.68e-07 ***
## ST           0.378576   0.021711  17.437  < 2e-16 ***
## I            0.004816   0.001987   2.424   0.0157 *
## DPF          0.897675   0.650422   1.380   0.1681
## A           -0.063677   0.028000  -2.274   0.0234 *
## O1           2.184222   0.546829   3.994 7.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.98 on 530 degrees of freedom
## Multiple R-squared:  0.4825, Adjusted R-squared:  0.4747
## F-statistic: 61.78 on 8 and 530 DF,  p-value: < 2.2e-16
```

Ejercicio 6

Evaluar la bondad del ajuste realizado, a través del coeficiente de determinación. Indicar cuánto vale y qué significa.

La bondad del ajuste puede ser medida por el estadístico R^2. El cual se define como