Trabajo Práctico I

Aprendizaje Estadístico - (84:04)

A. Verri Kozlowski y F. Patitucci

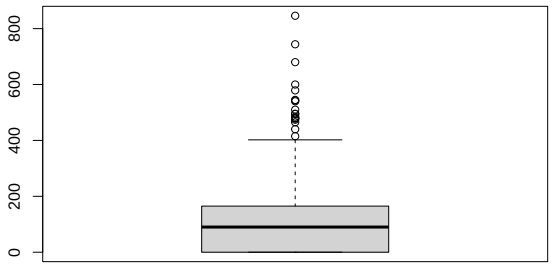
Contents

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)
data$Outcome<-factor(data$Outcome)
data$Glucose<-as.numeric(data$Glucose)
data$BloodPressure<-as.numeric(data$BloodPressure)
data$SkinThickness<-as.numeric(data$SkinThickness)
data$Insulin<-as.numeric(data$Insulin)
boxplot(data[6])
```



Se ve que las siguientes variables debieron ser definidas como num en vez de int, ya que son continuas y no discretas.

- Glucose

- BloodPressure
- SkinThickness
- Insulin

Además se tiene presente que la variable X no tiene relevancia en nuestro estudio, ya que la misma indica el número de paciente.

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.

Se plantea el modelo

$Y = \beta_0 + \beta_1 \cdot x_{Pregnancies} + \beta_2 \cdot x_{Glucose} + \beta_3 \cdot x_{BloodPressure} + \beta_4 \cdot x_{SkinThickness} + \beta_5 \cdot x_{Insulin} + \beta_6 \cdot x_{DiabetesPedigreeFunction} + \beta_7 \cdot x_{Age} + \beta_8 \cdot \mathbb{I}(Outcome = 1) + \varepsilon$

Los supuestos del modelo son

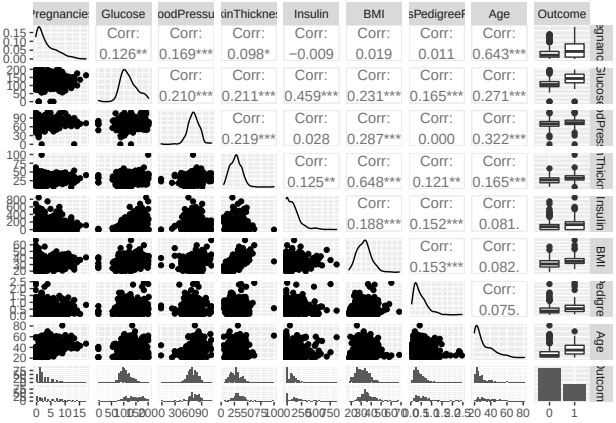
- Los errores ε_i tienen media cero. Esto es $\mathbb{E}[\varepsilon_i] = 0$
- Los errores ε_i tienen todos la misma varianza. Esto es $\mathbb{VAR}[\varepsilon_i] = \sigma^2$
- Los errores ε_i tienen distribuci´on Normal. Los errores ε_i son independientes entre si y no están correlacionados con las covariables x_i

```
Modelo <- lm( BMI ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + DiabetesPedigreeFunction + Ag
```

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

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

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 $p_{valor} \leq 0.05$, 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 (Nivel de significancia ≈ 0)
- SkinThickness (Nivel de significancia ≈ 0)
- Insulin (Nivel de significancia $= 0.0157$)
- Age (Nivel de significancia $= 0.0234$)
- Outcome1 (Nivel de significancia ≈ 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 ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
##     Insulin + DiabetesPedigreeFunction + Age + Outcome, data = data)
##
## 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 ***
## Pregnancies              -0.083276   0.085507  -0.974   0.3305
## Glucose                  -0.002452   0.008304  -0.295   0.7679
## BloodPressure             0.092145   0.017893   5.150 3.68e-07 ***
## SkinThickness             0.378576   0.021711  17.437  < 2e-16 ***
## Insulin                   0.004816   0.001987   2.424   0.0157 *
## DiabetesPedigreeFunction  0.897675   0.650422   1.380   0.1681
## Age                      -0.063677   0.028000  -2.274   0.0234 *
## Outcome1                  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

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

$R^2 = \frac{||\hat{Y}-\bar{Y}||^2}{||Y-\bar{Y}||^2}$

$R^2 = \frac{||Y-\bar{Y}||^2-||Y-\hat{Y}||^2}{||Y-\bar{Y}||^2}$

Donde