Institución Universitaria
Reacreditada en Alta Calidad

Métodos de discriminación estadística

Prof.
David E Rodriguez Guevara
PhD.(c) Economia y Finanzas - UBJ

Somos Innovación Tecnológica con Sentido Humano

Alcaldía de Medellín

Discriminación de información: Lo básico

- Las bases de datos comprenden información cuantitativa (datos continuos) y cualitativa (datos discretos)
- R entiende a la información de tipo cuantitativo de forma muy simple.
- Toda base de datos si tiene implícito un nombre que identifique la variable y muestra información estrictamente numérica R interpreta a la serie como una serie cuantitativa.
- Los datos discretos de dividen en variable y categorías de la variable.
- Cuando se usa series cualitativas R no interpreta dicha información, es capaz de leerla, pero no identifica sus características

Discriminación de información: Lo básico

Variable cuantitativa

```
> summary(data[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   33.00   39.00   40.94   48.00   95.00
> class(data[,2])
[1] "integer"
```

Variable cualitativa

```
> summary(data[,3])
   Length     Class      Mode
    45211 character character
> class(data[,3])
[1] "character"
>
```

Discriminación de información: Lo básico

- Para R la forma básica de identificación de la serie cualitativa se realiza por medio de: *as.factor*
- Esta función convierte la información cualitativa a valores funcionales, pasándola a criterio de niveles contables.
- La información en formato “factor” ´permite:
    - Ordenar la información de forma alfabética
    - Cuantificar la información de las características de una variable
    - Permite subdividir la información a conveniencia para modelar

Discriminación de información: Lo básico

```
> job1 <- as.factor(data[,3])
> class(job1)
[1] "factor"
> levels(job1)
 [1] "admin"         "blue-collar"   "entrepreneur"
 [4] "housemaid"     "management"    "retired"
 [7] "self-employed" "services"      "student"
[10] "technician"    "unemployed"    "unknown"
> summary(job1)
       admin   blue-collar  entrepreneur     housemaid
        5171          9732          1487          1240
  management       retired self-employed      services
        9458          2264          1579          4154
     student    technician    unemployed       unknown
         938          7597          1303           288
> |
```

Discriminación de información: Lo básico

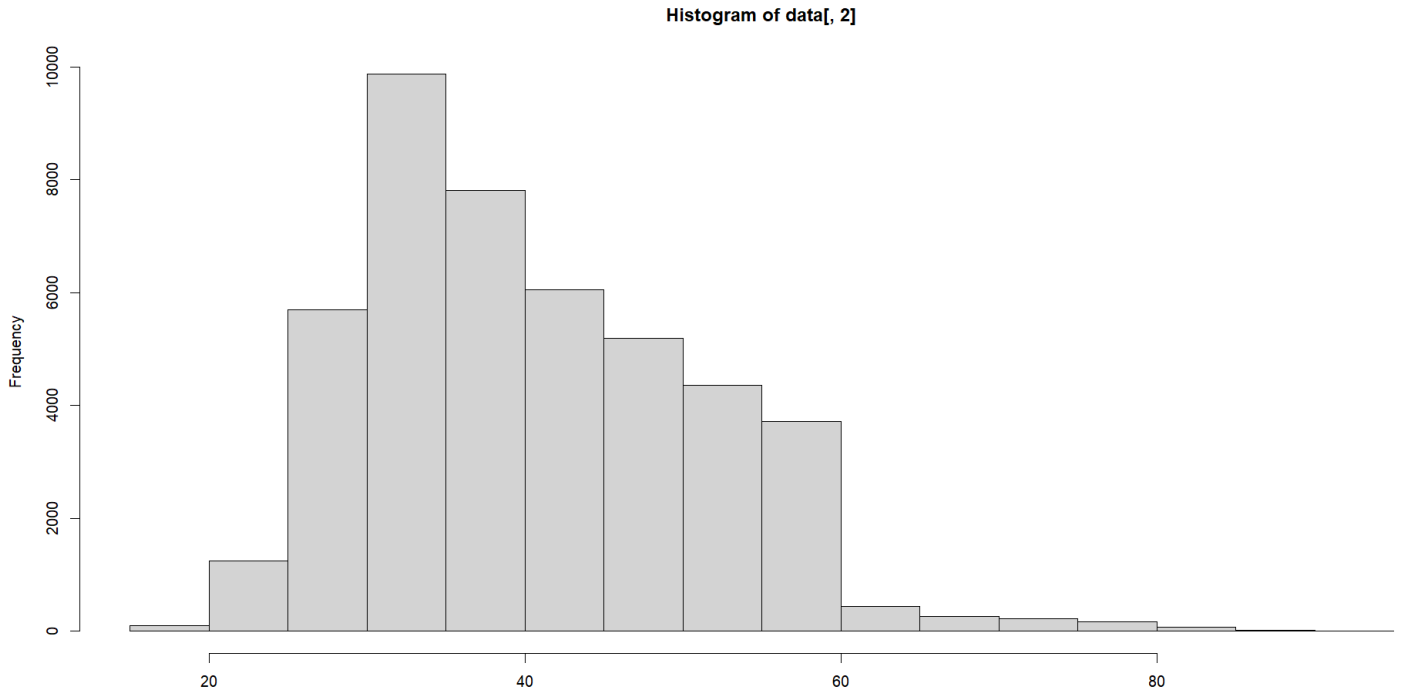
Para transformar las varias variables cualitativas se puede utilizar una función SAPPLY enfocada en el detalle que las variables inicialmente son as.character

```
#convertir varias variables en una base independiente
fac_cols <- sapply(data, is.character)
data[fac_cols] <- lapply(data[fac_cols], as.factor)
```

Como transformar cuantitativas a cualitativas

En muchas ocasiones los datos cuantitativos no son representativos en el termino del uso de su distribución o no tienen aplicabilidad en modelos econométricos no son representativos. Una forma de explorar una visión puede ser convirtiendo la serie cuantitativa a cualitativa y ver los efectos en términos de modelación

```
> summary(data[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   33.00   39.00   40.94   48.00   95.00
```



Como transformar cuantitativas a cualitativas

Para realizar la transformación de una variable cuantitativa se puede utilizar as.factor y una serie secuencial de ifelse, según se identifique los rangos de un histograma (o a discreción del estudio).

```
> hist(data[,2])
> age1 <- as.factor(ifelse(data[,2]<=25,'18-24',
+                          ifelse(data[,2]<=32,'25-32',
+                           ifelse(data[,2]<=39,'33-39',
+                            ifelse(data[,2]<=46,'40-46',
+                              ifelse(data[,2]<=53,'47-53',
+                                ifelse(data[,2]<=60,'54-60',
+                                  '60+')))))))
>
> table(age1)
age1
18-24 25-32 33-39 40-46 47-53 54-60   60+
 1336  9775 12251  8576  6756  5329  1188
>
```

Análisis de datos: Análisis estadísticos

En un análisis discriminatorio es necesario comparar con la serie que se quiere contrastar el estudio, una variable endógena vs una variable exógena, para ello se puede utilizar la función revalue cuando existen variables cualitativas pero se muestran en valores numéricos. (Piense en una estructura de estratos sociales).

```
> riesgo <- as.factor(data[,1])
> riesgo <- revalue(riesgo, c("0"="no_riesgo", "1"="si_riesgo"))
> summary(riesgo)
no_riesgo si_riesgo
    39922      5289
>
```

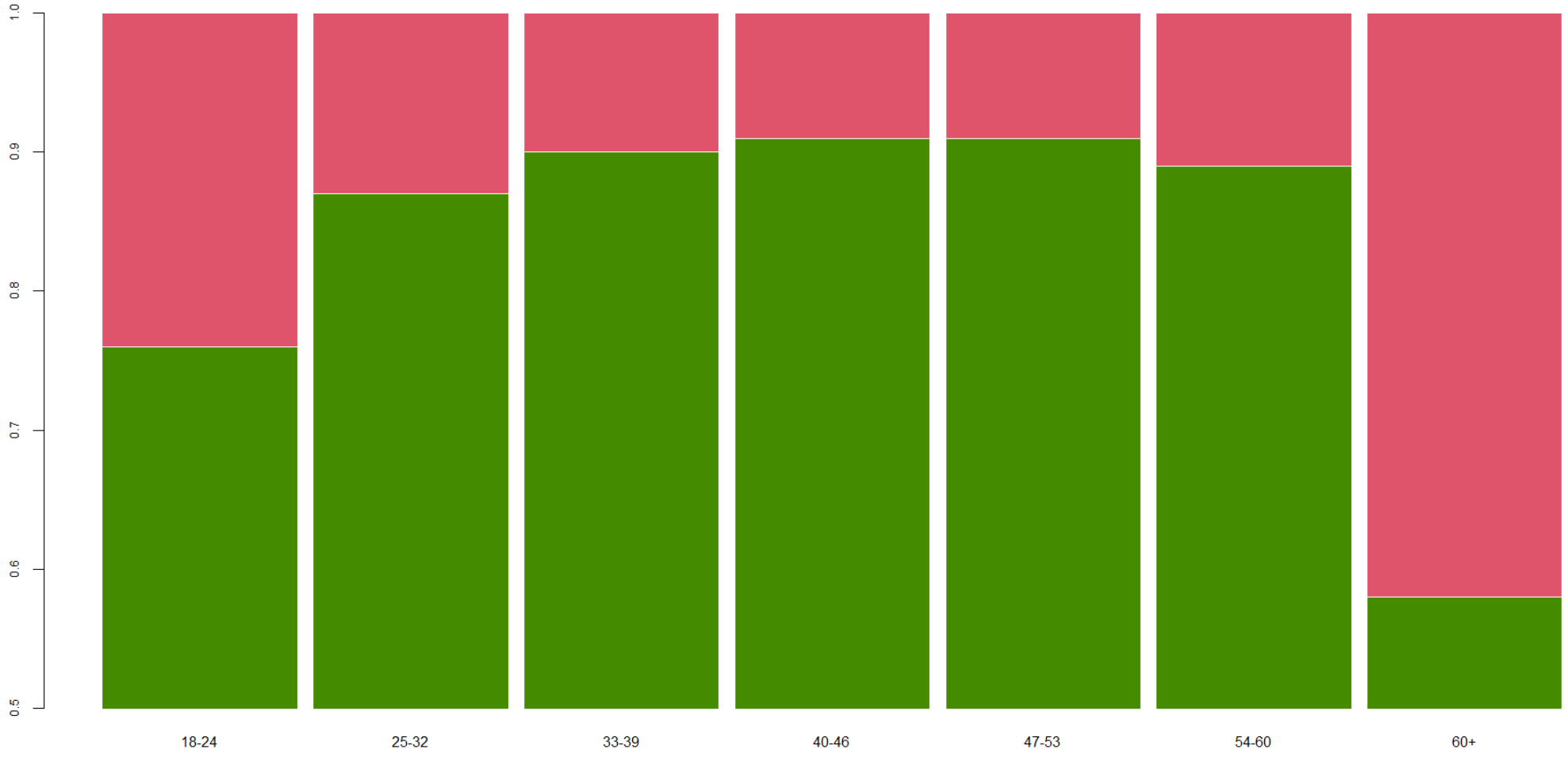
Análisis de datos: Análisis estadísticos

Análisis de datos: Análisis estadísticos

Como se puede observar se muestra una información discriminada en una tabla de contingencia, que permite analizar rápidamente la realidad de la variable exógena sobre la endógena.

```
> agetab
            age1
riesgo       18-24 25-32 33-39 40-46 47-53 54-60   60+
  no_riesgo   1016  8523 10982  7809  6141  4765   686
  si_riesgo    320  1252  1269   767   615   564   502
> agepor
            age1
riesgo       18-24 25-32 33-39 40-46 47-53 54-60  60+
  no_riesgo   0.76  0.87  0.90  0.91  0.91  0.89 0.58
  si_riesgo   0.24  0.13  0.10  0.09  0.09  0.11 0.42
>
```

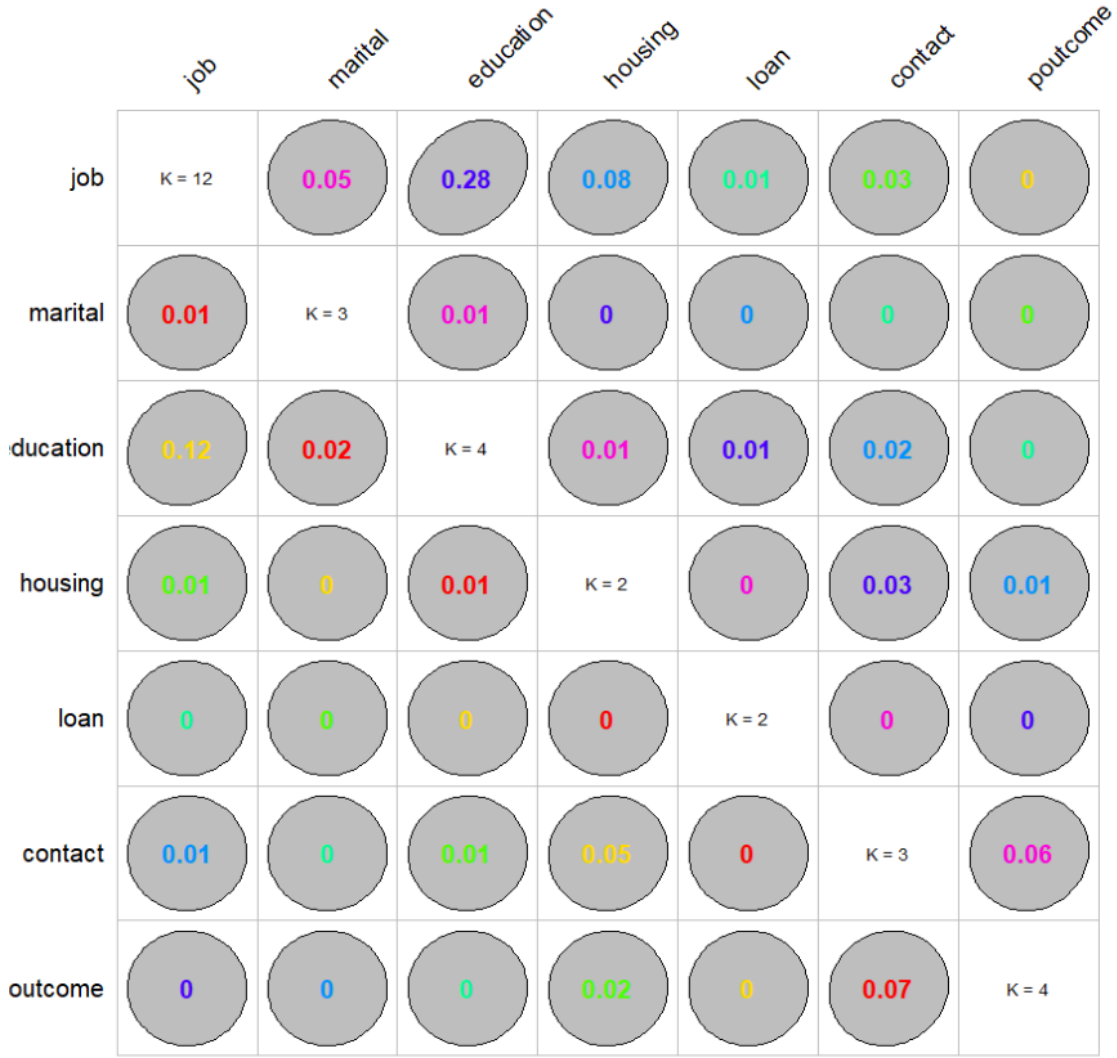
Análisis de datos: Análisis estadísticos



Análisis de datos: Análisis estadísticos

Análisis de datos: Análisis estadísticos

Para analizar la correlación de variables cualitativas, se puede utilizar el Lambda de Goodman-Kruskal (Rodríguez et al, 2022) que permite analizar de forma similar las variables cualitativas entre si para verificar si hay alto nivel de correlación entre variables y así poder descartar las que generen esta alta correlación



Modelos para discriminar: Logit / Probit

Modelo Logit

$$P_i(Y = 1 \mid X_{ik}) = \frac{1}{1 + e^{-z_i}}; \mathrm{z_i} = \alpha + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik}$$

$$P_i(Y = 1 \mid X_{ik}) = \frac{e^{z_i}}{1 + e^{z_i}}$$

Modelo Probit

$$F(x) = \frac{1}{\sqrt{2\pi\sigma_I^2}} \int_{-\infty}^{I_i} e^{\frac{Z^2}{2\sigma^2}} dZ; \mathrm{I_i} = \alpha + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik}$$

Modelos de discriminación

Modelo Logit

```
Call:
glm(formula = formula, family = binomial(logit), data = train)

Coefficients:
                    Estimate   Std. Error z value             Pr(>|z|)
(Intercept)       -1.926128057  0.132283158 -14.561 < 0.0000000000000002 ***
age                0.003837672  0.002035664   1.885             0.059400 .
jobblue-collar    -0.340446193  0.068734928  -4.953    0.000000730665255 ***
jobentrepreneur   -0.502724921  0.120757010  -4.163    0.000031393970316 ***
jobhousemaid      -0.398486859  0.124161643  -3.209             0.001330 **
jobmanagement     -0.179802027  0.068286231  -2.633             0.008462 **
jobretired         0.546634040  0.086819052   6.296    0.000000000304942 ***
jobself-employed  -0.221119486  0.102775000  -2.151             0.031437 *
jobservices       -0.225990670  0.078542314  -2.877             0.004011 **
jobstudent         0.596237875  0.100765277   5.917    0.000000003276743 ***
jobtechnician     -0.204809755  0.064416101  -3.179             0.001475 **
jobunemployed      0.028795233  0.101593095   0.283             0.776842
jobunknown        -0.511154686  0.214611541  -2.382             0.017230 *
maritalmarried    -0.158212174  0.054249897  -2.916             0.003541 **
maritalsingle      0.215710022  0.061742778   3.494             0.000476 ***
educationsecondary 0.208931532  0.059500679   3.511             0.000446 ***
educationtertiary  0.501021462  0.068745258   7.288    0.000000000000314 ***
educationunknown   0.250370827  0.095766900   2.614             0.008939 **
balance            0.000025654  0.000004418   5.807    0.000000006361848 ***
housingyes        -0.735903339  0.035973222 -20.457 < 0.0000000000000002 ***
loanyes           -0.567088146  0.056586294 -10.022 < 0.0000000000000002 ***
---
```

Modelos de discriminación

Modelo Probit

```
Call:
glm(formula = formula, family = binomial(probit), data = train)

Coefficients:
                    Estimate   Std. Error z value             Pr(>|z|)
(Intercept)       -1.123185729  0.069772514 -16.098 < 0.0000000000000002 ***
age                0.001681582  0.001079693   1.557             0.119361
jobblue-collar    -0.170423266  0.035268718  -4.832      0.00000135075483 ***
jobentrepreneur   -0.258264358  0.061098937  -4.227      0.00002368422694 ***
jobhousemaid      -0.208951165  0.063881278  -3.271             0.001072 **
jobmanagement     -0.096715267  0.036566915  -2.645             0.008172 **
jobretired         0.311343483  0.047853198   6.506      0.00000000007706 ***
jobself-employed  -0.119900517  0.054612926  -2.195             0.028131 *
jobservices       -0.120205882  0.040563099  -2.963             0.003042 **
jobstudent         0.347055622  0.058188135   5.964      0.00000000245578 ***
jobtechnician     -0.107624765  0.033967562  -3.168             0.001533 **
jobunemployed      0.013185781  0.055233755   0.239             0.811317
jobunknown        -0.265094949  0.112213246  -2.362             0.018156 *
maritalmarried    -0.082203191  0.028681328  -2.866             0.004156 **
maritalsingle      0.112893655  0.032804915   3.441             0.000579 ***
educationsecondary 0.107430784  0.030495581   3.523             0.000427 ***
educationtertiary  0.262405171  0.036113181   7.266      0.00000000000037 ***
educationunknown   0.130968924  0.050801648   2.578             0.009936 **
balance            0.000015402  0.000002529   6.090      0.00000000112819 ***
housingyes        -0.380934058  0.018721634 -20.347 < 0.0000000000000002 ***
loanyes           -0.274452851  0.027724389  -9.899 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelos para discriminar: Logit / Probit

Identificar la probabilidad de un evento con Modelo Logit y Probit

```
#Calculo de un evento de probabilidad
info = c(v1 = 1, v2 = 35, v3 = 0, v4 = 0, v5 = 1, v6 = 0, v7 = 0, v8 = 0,
         v9 = 0, v10 = 0, v11 = 0, v12 = 0, v13 = 0, v14 = 1, v15 = 0,
         v16 = 0, v17 = 0, v18 = 1, v19 = 5000, v20 = 1, v21 = 0)
minfo = as.matrix(data.frame(info))

#extracción de coeficientes de modelos
coeflogl = as.matrix(summary(model1)$coefficients[,1])
coeflogp = as.matrix(summary(model2)$coefficients[,1])
```

Modelos para discriminar: Logit / Probit

Identificar la probabilidad de un evento con Modelo Logit y Probit

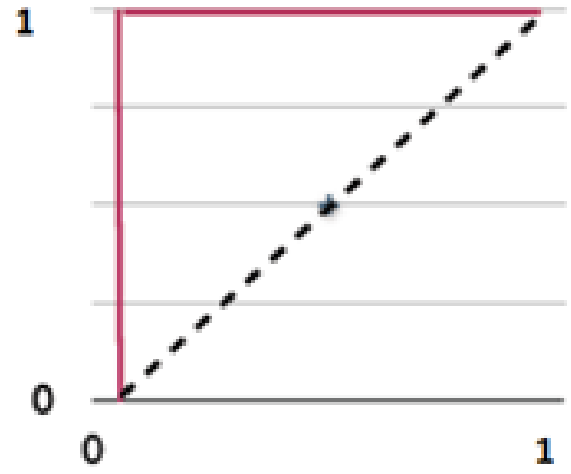
```
> #identificar la probabilidad logit
> zi = crossprod(coeflogl, minfo)
> pi = 1/(1+exp(-zi))
>
> #identificar la probabilidad probit
> ii = crossprod(coeflogp, minfo)
> pip = pnorm(ii, mean = 0, sd = 1, lower.tail = T) #CDF
> pi
             info
[1,] 0.06263353
> pip
             info
[1,] 0.06320183
> |
```

Bondad de Ajuste de modelos discriminatorios

Curva AUC-ROC

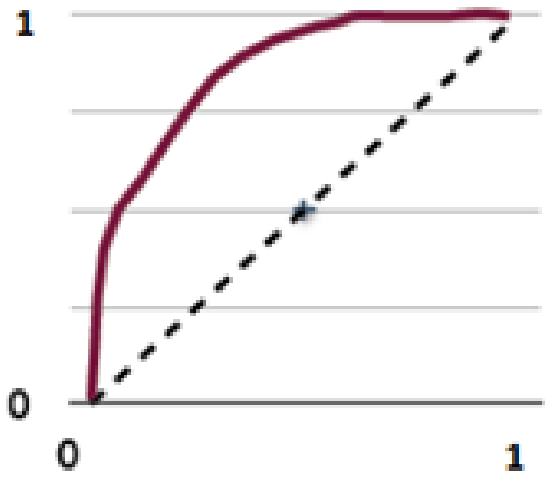
AUC=1

+ valor diagnóstico perfecto



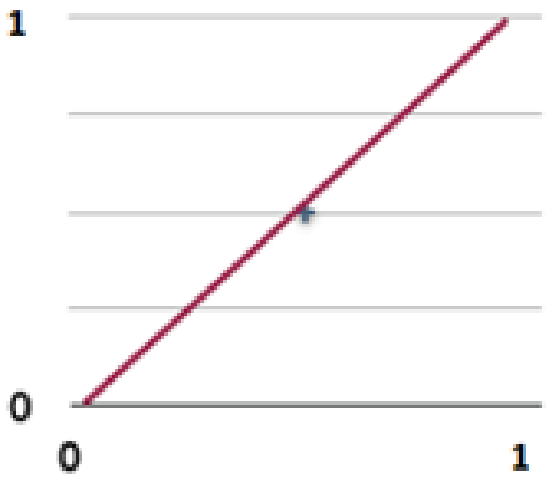
AUC=0,8

+ valor diagnóstico



AUC=0,5

+ sin valor diagnóstico



Bondad de Ajuste de modelos discriminatorios

Curva AUC-ROC

$$
\text{If} \begin{cases} ROC = 0.5 & \text{This suggests no discrimination, so we might as well flip a coin.} \\ 0.5 < ROC < 0.7 & \text{We consider this poor discrimination, not much better than a coin toss.} \\ 0.7 \leq ROC < 0.8 & \text{We consider this acceptable discrimination.} \\ 0.8 \leq ROC < 0.9 & \text{We consider this excellent discrimination.} \\ ROC \geq 0.9 & \text{We consider this outstanding discrimination.} \end{cases}
$$

Bondad de Ajuste de modelos discriminatorios

Tabla de confusión (Especificidad y Sensibilidad)

		Predicted	
		0	1
Actual	0	a	b
	1	c	d

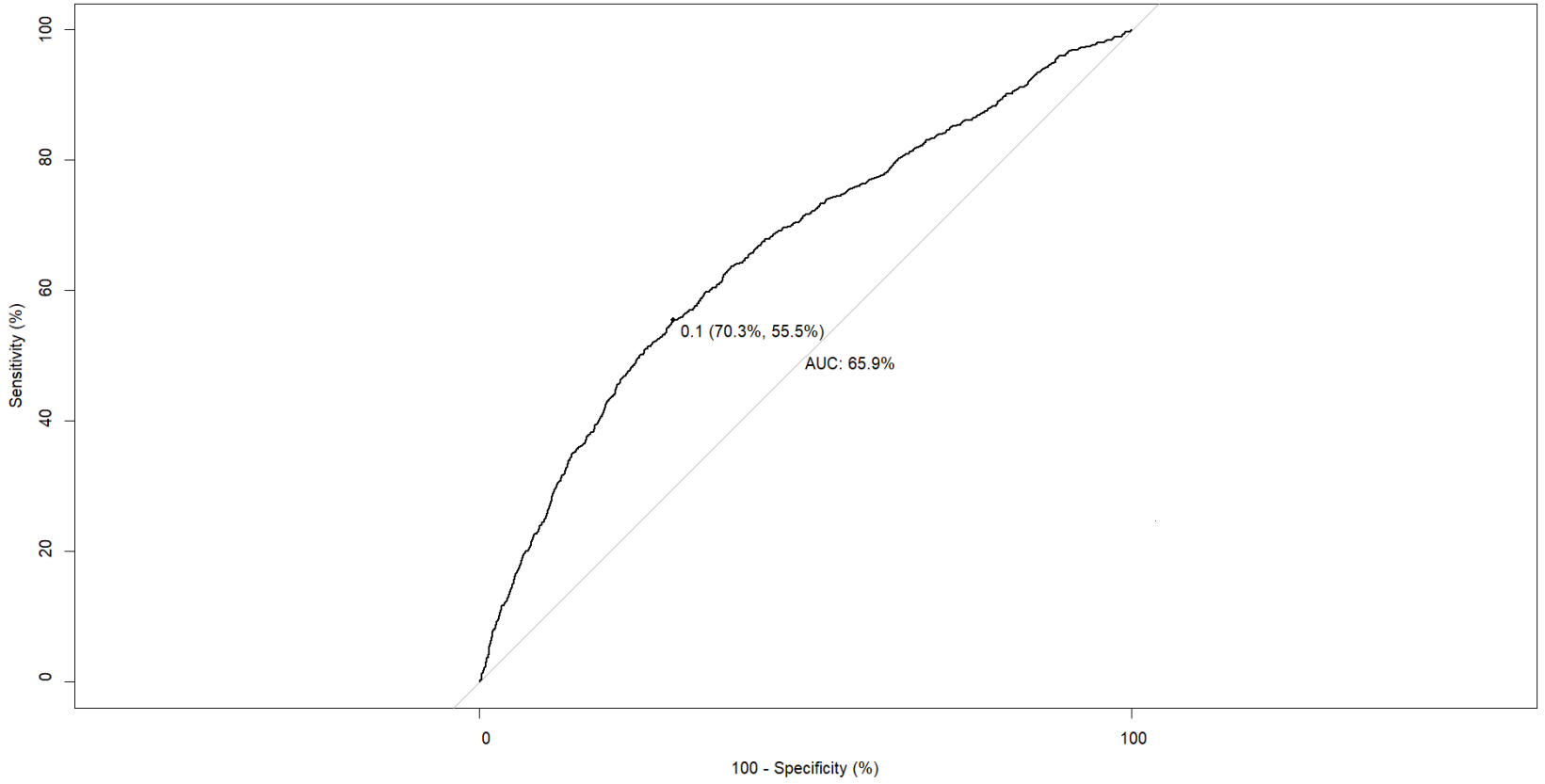
$$Sensibilidad = \frac{d}{c+d}$$

$$Especificidad = \frac{a}{a+b}$$

Bondad de Ajuste de modelos discriminatorios

Curva AUC-ROC
Logit (test data)

```
> probvl = cut(predl, bre
estimado
> yvl = cut(test[,1], bre
real
> discrim = table(yvl,pro
> addmargins(discrim)
      probvl
yvl       0    1  Sum
  0    5609 2369 7978
  1     474  590 1064
  Sum 6083 2959 9042
>
```



Bondad de Ajuste de modelos discriminatorios

Curva AUC-ROC
Logit (test data)

```
> probvl = cut(predl, bre
estimado
> yvl = cut(test[,1], bre
real
> discrim = table(yvl,pro
> addmargins(discrim)
      probvl
yvl       0    1  Sum
  0    5609 2369 7978
  1     474  590 1064
  Sum 6083 2959 9042
>
```

```
> sens = discrim[4]/(discrim[2]+discrim[4])
> espc = discrim[1]/(discrim[1]+discrim[3])
> sens
[1] 0.5545113
> espc
[1] 0.7030584
>
```

Referencias

Breu, F., Guggenbichler, S., & Wollmann, J. (2008). Modelos de Regresion Cualitativa. *Vasa*. http://medcontent.metapress.com/index/A65RM03P4874243N.pdf

Greene, W. H. (1992). A Statistical Model for Credit Scoring. In *NYU Working Paper No. EC-92-29*. http://papers.ssrn.com/abstract=1293124

Hosmer, D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression Third Edition* (John Wiley & Sons, Ed.; 3rd ed., Vol. 1). John Wiley & Sons. www.wiley.com

Mileris, R. (2010). Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian classifier. *Economics and Management*, *15*(1), 1078–1084. http://www.ktu.lt/lt/mokslas/zurnalai/ekovad/15/1822-6515-2010-1078.pdf

Moreno, S. (2013). *El Modelo Logit Mixto para la construcción de un Scoring de Crédito* [Estadística Financiera]. Universidad Nacional de Colombia.

Rodríguez, D., Rendón, J., Trespalacios, A., & Jiménez, E. (2022). Modelación de riesgo de crédito de personas naturales. Un caso aplicado a una caja de compensación familiar colombiana. *REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA*, *33*(1), 29–48. www.upo.es/revistas/index.php/RevMetCuant/article/view/514

Rodríguez-Guevara, D. E., & Gonzalez-Uribe, G. J. (2017). *Principios de Econometría* (Fondo Editorial ITM, Ed.; Fondo Edit). Fondo Editorial ITM. https://books.google.com.co/books?id=BbE-DwAAQBAJ

Institución Universitaria
Reacreditada en Alta Calidad

¡Gracias!

Somos Innovación Tecnológica con Sentido Humano

Alcaldía de Medellín