Universidades

TP Final - Regresión Avanzada

Joaquin Bermejo, Franco Scarafia y Gerard Seward

Introducción

Se nos presenta una base de datos sobre universidades públicas y privadas con las siguientes variables

Variable	Descripción
`privada`	indica si la universidad es privada o no.
`aplicaciones`	cantidad de aplicaciones recibidas por la universidad durante el último año (cada estudiante que aspira a ingresar debe presentar una aplicación formal, a partir de la cual es admitido/a o rechazado/a), medida en miles de personas.
`ingresantes`	cantidad de aplicaciones aceptadas, medida en miles de personas.
`estudiantes`	cantidad total de estudiantes en carreras de grado, medida en miles de personas.
`top10`	porcentaje de ingresantes que fueron parte del 10% de estudiantes con mejores calificaciones en sus respectivas escuelas secundarias.
`cuota`	costo de la cuota de la universidad, medida en miles de dólares.
`prof_dr`	porcentaje de profesores de la universidad que poseen título de doctorado.
`razon`	tasa de estudiantes por profesor.
`tasa_grad`	porcentaje de estudiantes que se gradúan.

La variable de interés es `tasa_grad` que indica el porcentaje de estudiantes que se gradúan.

Regresión Lineal

División en entrenamiento y prueba

```
set.seed(1234)
filas_train <- sample(x = 1:nrow(df), size = nrow(df)*0.7) #asignacion aleatoria

df_train <- slice(df, filas_train)
df_test <- slice(df, -filas_train)
```

Ajustes de modelos

El **primer modelo** propuesto surge de aplicar un método de selección *stepwise* considerando solamente las variables originales, sin interacciones.

```
Call:
lm(formula = tasa_grad ~ cuota + top10, data = df_train)

Coefficients:
(Intercept)        cuota        top10
    39.1618       1.8600       0.2487
```

El **segundo modelo** también surge de aplicar el método *stepwise* pero considerando como modelo maximal aquel con todas las interacciones de segundo orden.

```
Call:
lm(formula = tasa_grad ~ cuota + top10 + cuota:top10, data = df_train)

Coefficients:
(Intercept)        cuota        top10  cuota:top10
   33.99731      2.33018      0.43030     -0.01453
```

El **tercer modelo** surge de aplicar la técnica de mejores subconjuntos. Visto que el modelo anterior incluye tres términos (dos efectos principales y una interacción entre ellos) se elige el mejor modelo con 3 variables explicativas.

```
Subset selection object
Call: regsubsets.formula(x = tasa_grad ~ ., data = df_train)
8 Variables  (and intercept)
             Forced in Forced out
privadaTRUE      FALSE      FALSE
aplicaciones     FALSE      FALSE
ingresantes      FALSE      FALSE
estudiantes      FALSE      FALSE
top10            FALSE      FALSE
cuota            FALSE      FALSE
prof_dr          FALSE      FALSE
razon            FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         privadaTRUE aplicaciones ingresantes estudiantes top10 cuota prof_dr
1  ( 1 ) " "         " "          " "         " "         " "   "*"   " "
2  ( 1 ) " "         " "          " "         " "         "*"   "*"   " "
3  ( 1 ) "*"         " "          " "         " "         "*"   "*"   " "
4  ( 1 ) " "         "*"          " "         "*"         "*"   "*"   " "
5  ( 1 ) "*"         "*"          " "         "*"         "*"   "*"   " "
6  ( 1 ) "*"         "*"          " "         "*"         "*"   "*"   "*"
7  ( 1 ) "*"         "*"          "*"         "*"         "*"   "*"   "*"
8  ( 1 ) "*"         "*"          "*"         "*"         "*"   "*"   "*"
         razon
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) " "
4  ( 1 ) " "
5  ( 1 ) " "
6  ( 1 ) " "
7  ( 1 ) " "
8  ( 1 ) "*"
```

Comparación de modelos

```
       CME    PRESS       Cp      AIC      BIC
1 177.1570 96972.46 92.92821 4359.097 4376.286
2 175.8549 96567.48 90.29824 4357.091 4378.577
3 176.6482 97082.87 93.11895 4359.535 4381.021
```

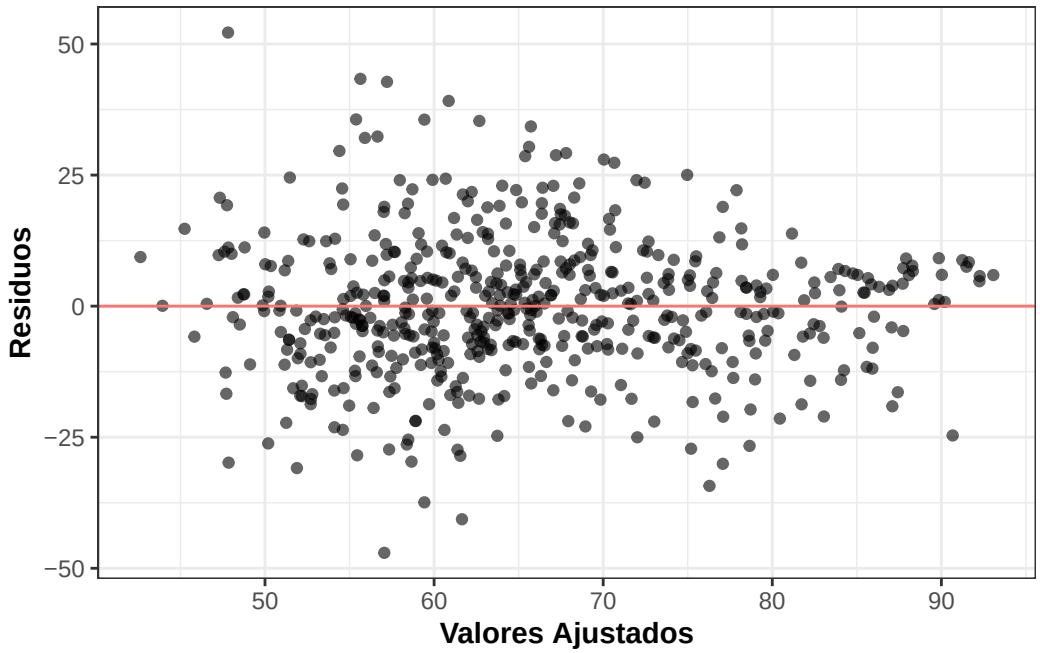
Puede verse que para todas las métricas salvo BIC, el mejor modelo (en términos de desempeño) es el segundo: aquel que considera dos explicativas y su interacción. Por lo tanto, el modelo

seleccionado queda de la forma:

$tasa\_grad = \beta_0 + \beta_1\ cuota + \beta_2\ top10 + \beta_3\ cuota * top10 + \epsilon$

Análisis de residuos

Residuos versus valores ajustados



Se puede ver que la variancia de los residuos no es constante para todos los valores ajustados. En particular, se evidencia una mayor variabilidad para tasas de graduación predichas en el rango 55% a 65%.

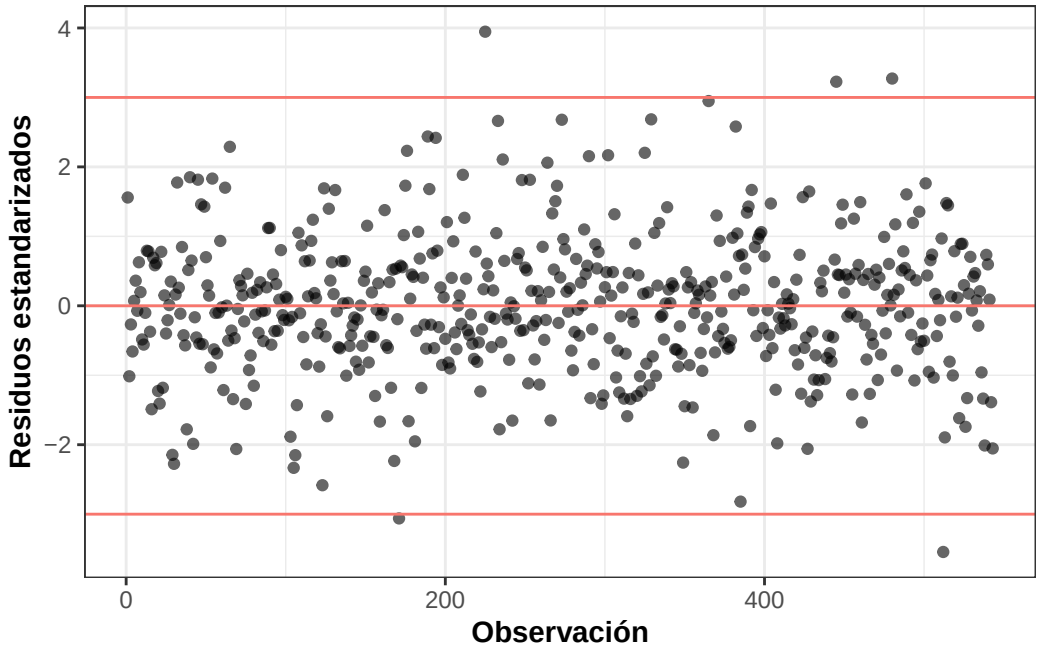
La hipótesis anterior puede evaluarse mediante el test de Breusch-Pagan.

```
    studentized Breusch-Pagan test

data:  sel_mod
BP = 11.965, df = 3, p-value = 0.007503
```

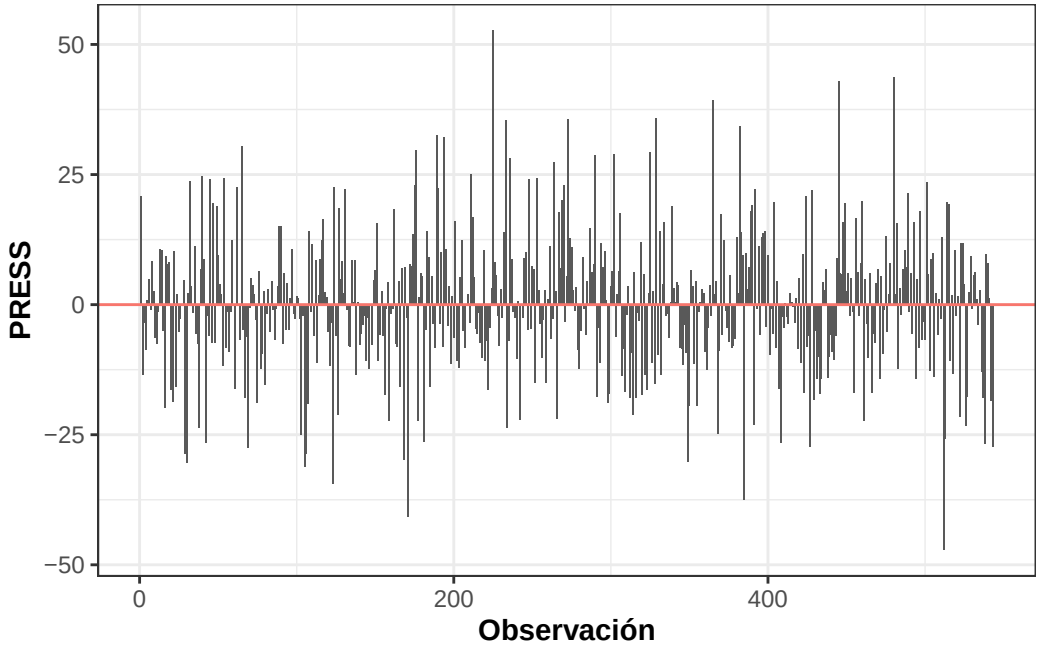
Como el p-value resulta inferior al nivel de significación 5%, se rechaza la hipótesis nula, indicando que posiblemente no se esté cumpliendo el supuesto de homocedasticidad de los residuos.

Residuos estandarizados



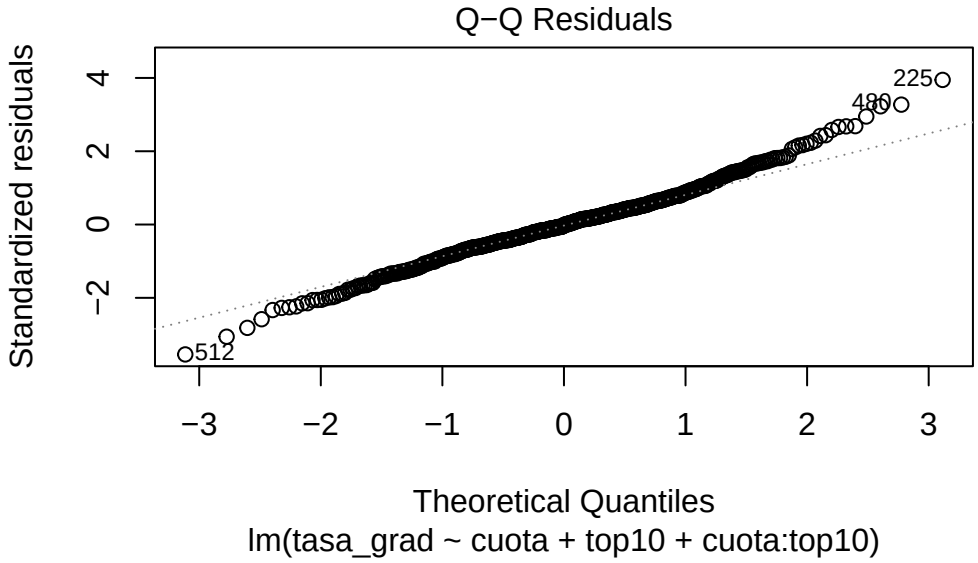
Se encuentran 5 valores con residuos estandarizados mayores a 3 unidades, en valor absoluto. Esto corresponde a un 0.92% de la totalidad de las observaciones de entrenamiento.

Residuos PRESS



Se observa un mayor valor absoluto de los residuos PRESS para las observaciones que tenían errores estandarizados mayores a 3 unidades en el gráfico anterior.

Análisis de normalidad




```
    Anderson-Darling normality test

data:  sel_mod$residuals
A = 1.6877, p-value = 0.0002482
```

Dado que el p-value es inferior al nivel de significación del 5%, se rechaza la hipótesis nula de distribución Normal para los errores.

Análisis de colinealidad

```
      cuota       top10 cuota:top10
   4.178070    9.384239   16.856884
```

Los términos `top10` y `cuota:top10` presentan un valor de VIF mayor a 5 unidades. Esto indicaría una colinealidad entre estos términos, lo cual resulta lógico dado que el segundo término refiere a la interacción entre el primer término y la variable explicativa restante. De hecho, puede verse que los valores de VIF para el modelo sin interacción se ven reducidos.

```
  cuota   top10
1.46823 1.46823
```

Interpretación de los predictores

```
Call:
lm(formula = tasa_grad ~ cuota + top10 + cuota:top10, data = df_train)

Residuals:
    Min      1Q  Median      3Q     Max
-47.054  -7.871  -0.285   7.113  52.186

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.997311   3.043330  11.171  < 2e-16 ***
cuota        2.330176   0.292218   7.974 9.24e-15 ***
top10        0.430303   0.098965   4.348 1.64e-05 ***
cuota:top10 -0.014531   0.007274  -1.998   0.0462 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.3 on 539 degrees of freedom
Multiple R-squared:  0.3907,    Adjusted R-squared:  0.3873
F-statistic: 115.2 on 3 and 539 DF,  p-value: < 2.2e-16
```

Los tres términos reusltan significativos al 5%. Por lo tanto, debido a la presencia de interacción, las interpretaciones de los coeficientes del modelo son las siguientes:

- Aumentar mil dólares la cuota se asocia con un incremento promedio en la tasa de graduación igual a `2,33 - 0,015 * top10` por la interacción en unidades porcentuales.
- Aumentar en una unidad porcentual el porcentaje de ingresantes que fueron parte del top 10% de estudiantes en sus escuelas secundarias se asocia con un incremento promedio en la tasa de graduación igual a `0,43 - 0,015 * cuota` por la interacción en unidades porcentuales.

Regularización y Predicción

Ajuste con técnica Ridge

```
[1] "Mejor valor de lambda: 0.9"
```

Ajuste con técnica Lasso

```
[1] "Mejor valor de lambda: 0"
```

Para la técnica Lasso, el valor óptimo del parámetro de regularización es $\lambda = 0$, lo cual implica estimaciones equivalentes a Mínimos Cuadrados Ordinarios. En otras palabras, bajo la técnica Lasso se concluye que no sería necesario aplicar regularización.

Comparación de modelos

Ajuste

```
2 x 4 sparse Matrix of class "dgCMatrix"
      (Intercept)    cuota     top10  cuota:top10
MCO      33.99731 2.330176 0.4303031 -0.014530641
Ridge    40.35142 1.750391 0.2400150  0.000611321
```

Los coeficientes asociados a los efectos principales se ven reducidos al aplicar regularización por Ridge.

Capacidad predictiva

```
# A tibble: 1 x 2
  rmse_MCO rmse_ridge
     <dbl>      <dbl>
1     14.1       14.1
```

Los valores de RMSE son muy similares para ambos métodos de estimación, aunque es menor para Mínimos Cuadrados Ordinarios, indicando que la regularización no mejora la capacidad predictiva del modelo.

Regresión Logística

Definición de variable respuesta (dicotómica)

```
df <- df %>% mutate(tasa_grad_binaria = if_else(tasa_grad < 75, F, T))
```

División en entrenamiento y prueba

```
set.seed(1492)
particion_logreg <- createDataPartition(df$tasa_grad_binaria, p = 0.7, list = F)
logreg_train <- df[particion_logreg,]
logreg_test <- df[-particion_logreg,]
```

Ajuste e interpretación del modelo

```
Call:
glm(formula = tasa_grad_binaria ~ privada + aplicaciones + ingresantes +
    estudiantes + top10 + cuota + prof_dr + razon, family = binomial(link = "logit"),
    data = logreg_train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1625097  0.9516771  -4.374 1.22e-05 ***
privadaTRUE   0.1911998  0.4721906   0.405   0.6855
aplicaciones  0.1508995  0.0751209   2.009   0.0446 *
ingresantes   0.8752909  0.6650412   1.316   0.1881
estudiantes  -0.3377068  0.1429975  -2.362   0.0182 *
top10         0.0206241  0.0086945   2.372   0.0177 *
cuota         0.2008847  0.0457842   4.388 1.15e-05 ***
prof_dr       0.0005592  0.0093520   0.060   0.9523
razon         0.0266573  0.0362013   0.736   0.4615
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 673.29  on 544  degrees of freedom
Residual deviance: 526.96  on 536  degrees of freedom
```

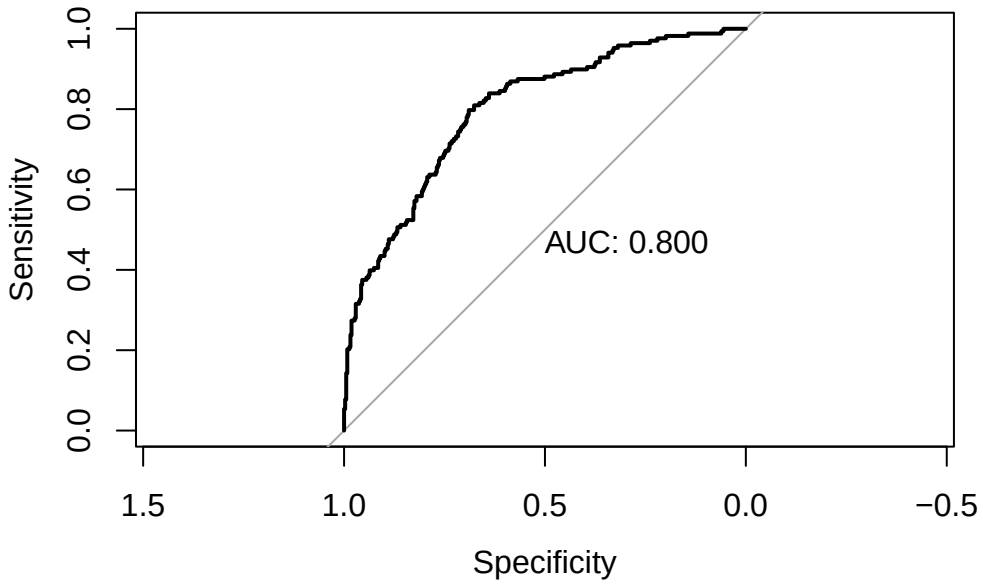
```
AIC: 544.96

Number of Fisher Scoring iterations: 5

aplicaciones  estudiantes        top10        cuota
   1.1628798    0.7134045    1.0208383    1.2224839
```

- Ante un aumento de mil aplicaciones recibidas, la chance de que una universidad tenga una buena tasa de graduación aumenta en un 16%.
- Ante un aumento de mil estudiantes en carreras de grado, la chance de que una universidad tenga una buena tasa de graduación disminuye en un 29%.
- Ante un aumento en una unidad porcentual del porcentaje de ingresantes que fueron parte del top 10% de estudiantes en sus escuelas secundarias, la chance de que una universidad tenga una buena tasa se graduación aumenta en un 2%.
- Ante un aumento de mil dólares en la cuota, la chance de que una universidad tenga una buena tasa de graduación aumenta en un 22%.

Curva ROC y punto de corte óptimo



Se obtiene un valor de AUC (área bajo la curva) igual a 0,8, lo cual habla de un buen clasificador.

Bajo el método de Youden se obtiene un punto de corte óptimo igual a 0.257. Este valor es lejano al punto de corte por defecto: 0,5.

Métricas de capacidad predictiva

```
Confusion Matrix and Statistics

            Reference
Prediction   Mala tasa Buena tasa
  Mala tasa        152         44
  Buena tasa         9         27

               Accuracy : 0.7716
                 95% CI : (0.7121, 0.8239)
    No Information Rate : 0.694
    P-Value [Acc > NIR] : 0.005386

                  Kappa : 0.3762

 Mcnemar's Test P-Value : 3.008e-06

            Sensitivity : 0.3803
            Specificity : 0.9441
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.7755
              Precision : 0.7500
                 Recall : 0.3803
                     F1 : 0.5047
             Prevalence : 0.3060
         Detection Rate : 0.1164
   Detection Prevalence : 0.1552
      Balanced Accuracy : 0.6622

       'Positive' Class : Buena tasa
```

- **Precisión:** El modelo clasifica correctamente al 77% de las universidades del conjunto de prueba según si tienen o no una buena tasa de graduación.
- **Sensibilidad:** Entre las universidades con buena tasa de graduación, sólo un 38% de ellas fueron clasificadas correctamente.
- **Especificidad:** Entre las universidades con mala tasa de graduación, un 94% fueron clasificadas correctamente.
- **VPP:** Cuando el modelo predice que una universidad tiene una buena tasa de graduación, acierta un 75% de las veces.
- **VPN:** Cuando el modelo predice que una universidad tiene una mala tasa de graduación, acierta un 78% de las veces.

- **F1:** La media armónica entre la sensibilidad y el VPP resulta igual a 50%.
- **Kappa:** La capacidad predictiva del modelo propuesto es aceptable.