Introducción a la Analítica - Tarea 1

Universidad Nacional de Colombia - Escuela de Estadística

Luisa María Acosta—Laura Camila Agudelo—Sebastián Agudelo
Andrea Amaya—Estefanía Echeverry

Agosto 2020

1.(10 pts. Teórico) Considere un estimador $\widehat{f}$ y un input X con los que se obtiene la predicción $\widehat{Y} = \widehat{f}(X)$. Asuma que $\widehat{f}$ y X son fijos. Demuestre que:

$$E(Y - \widehat{Y})^2 = E[f(X) + \varepsilon - \widehat{f}(X)]^2 = [f(X) - \widehat{f}(X)]^2 + \text{var}(\varepsilon)$$

Solución

Tenemos que

$$E[(f(X) + \varepsilon - \hat{f}(X))^2]$$

reagrupando los términos y expandiendo el polinomio nos queda

$$E[(f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\varepsilon + \varepsilon^2]$$

como $\widehat{f}$ y X son fijos

$$(f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))E[\varepsilon] + E[\varepsilon]^2$$

sabemos que el error es una variable aleatoria con media cero, $E[\varepsilon] = 0$, entonces

$$2(f(X) - \hat{f}(X))E[\varepsilon] = 0$$

nos queda que

$$(f(X) - \hat{f}(X))^2 + E[\varepsilon^2]$$

donde $E[\varepsilon^2]$=$Var(\varepsilon)$, así

$$[f(X) - \hat{f}(X)]^2 + Var(\varepsilon)]$$

2 (25 pts. Teórico) Es posible demostrar que la tasa de error Average $(I(y_0 \neq \widehat{y}_0))$ se minimiza en promedio por medio de un clasificador muy simple que clasifica cada observacion a la clase mas probable o factible, dados los valores de sus predictores. En otras palabras, se asigna simplemente una observacion del conjunto de prueba con predictor x_0 a la clase j para la cual

$$\text{P}(Y = j, X = x_0)$$

sea la mas grande. Esta es una probabilidad condicional. Este clasificador tan simple se conoce como CLASIFICADOR DE BAYES. Demuestre que el clasificador de Bayes produce la menor tasa de error de prueba (Test Error Rate) posible y se conoce como

Tasa de Error de Bayes.

Solución

Estamos interesados en demostrar que el clasificador de Bayes:

$$\hat{y}_0 = m\acute{a}x \quad Pr(Y = j|X = x_0),$$

el cual consiste en clasificar cada observación x_0 a la clase j más probable, produce la menor tasa de error de prueba $\frac{1}{n}\sum(I(y_0 \neq \hat{y}_0))$, es decir el valor mínimo de $E[I(y_0 \neq \hat{y}_0))]$. Así, podemos reescribir la tasa de error de la siguiente manera:

$$\begin{aligned}
E[I(y_0 \neq \hat{y}_0)] =& \sum_{j=0}^{1} jP(I(y_0 \neq (\hat{y}_0 = j))) \\
=& 0P(I(y_0 \neq \hat{y}_0 = 0)) + 1P(I(y_0 \neq (\hat{y}_0 = 1)) \\
=& P(I(y_0 \neq \hat{y}_0 = 1)) = P(y_0 \neq \hat{y}_0) \\
=& P(y_0 \neq \hat{y}_0|x_0) = \sum_{i \neq j} P(y_0 = i|x_0) \\
=& 1 - P(y_0 = i|x_0) \\
=& 1 - \underset{i}{m\acute{a}x} \quad P(y_0 = i|x_0)
\end{aligned}$$

La probabilidad de error es 1 menos la probabilidad de decisión correcta. Así, esto último se conoce como tasa de error de Bayes.

3. (20 pts. Teórico y practico)
Ejercicio 7, texto guía pagina 53. Hágalo también en R.
La siguiente tabla proporciona un conjunto de datos de entrenamiento que contiene seis observaciones, tres predictores y una variable de respuesta cualitativa.
Supongamos que deseamos utilizar este conjunto de datos para hacer una predicción para Y cuando X1 = X2 = X3 = 0 utilizando K-vecinos más cercanos K.
(a) Calcular la distancia euclidiana entre cada observación y el punto de prueba, X1 = X2 = X3 = 0.
(b) ¿Cuál es nuestra predicción con K = 1? ¿por qué?
(c) ¿Cuál es nuestra predicción con K = 3? ¿por qué?

(d) Si el límite de decisión de Bayes en este problema es altamente no lineal, entonces ¿esperamos que el mejor valor para K sea grande o pequeño? ¿por qué?

Solución

Obs.	**X1**	**X2**	**X3**	**Y**
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	- 1	0	1	Green
6	1	1	1	Red

(a) Teniendo en cuenta que la ecuación para la distancia euclidiana es:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ki} - x_{kj})^2} \tag{1}$$

Obtenemos con la nueva observación (X1= X2= X3= 0):

Obs.	**X1**	**X2**	**X3**	**Y**	**d(Obs., New Obs.)**
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	3.16
4	0	1	2	Green	2.24
5	- 1	0	1	Green	1.41
6	1	1	1	Red	1.73

(b) En el caso de K= 1. Teniendo en cuenta la distancia euclidiana, se observa que el vecino mas cercano es la Obs. 5. Por lo cual, la predicción haciendo uso de KNN será **Green**.
(c) En el caso de K=3. Se observa que los vecinos mas cercanos son las Obs. 2,5 y 6, en las cuales una es **Green**, y dos son **Red**. Por lo cual, con las probabilidades 1/3 para **Green**, y 2/3 para **Red**, la predicción haciendo uso de KNN será **Red**.

(d) Para un límite de decisión no lineal esperaríamos una **K pequeña** porque de este modo se observaría mayor "flexibilidad", con una frontera curveada al observar pocos puntos al clasificar. Esto, pues con una **K** grande se observaría un efecto menos "flexible", y produciría una frontera mas lineal porque tendría más puntos en cuenta al momento de clasificar.

(En R.)

```
    ## (a)
library(MASS)
library(class)
library(naivebayes)
library(car)

xm <-read.table(file.choose(), header=T)
xm <- data.frame(xm)
str(xm)
normalize <- function(x) {
  norm <- ((x - min(x))/(max(x) - min(x)))
  return (norm)
}
train <- normalize(xm[,1:3])
test <- data.frame("X1"=0,
                   "X2"=0,
                   "X3"=0)
y_train <- xm$Y
xm$dist<- sqrt((xm[,1]-0)^2 +(xm[,2]-0)^2+(xm[,3]-0)^2)
View(xm)
```

	X1	X2	X3	Y	dist
1	0	3	0	Red	3.000000
2	2	0	0	Red	2.000000
3	0	1	3	Red	3.162278
4	0	1	2	Green	2.236068
5	-1	0	1	Green	1.414214
6	1	1	1	Red	1.732051

Figura 1: Salida de R para la *Distancia euclidiana* calculada.

(b) fit.knn$_train < -knn(train = train, test = train, cl = y_train, k = 1, prob = TRUE)$
$fit.knn_Test1 < -knn(train = train, test = test, cl = y_train, k = 1, prob = TRUE)$
$fit.knn_Test1$

```
> fit.knn_Test1
[1] Green
attr(,"prob")
[1] 1
Levels: Red Green
```

Figura 2: Salida de R para la *K=1*.

(c) fit.knn$_Test3 < -knn(train = train, test = test, cl = y_train, k = 3, prob = TRUE)$
$fit.knn_Test3$

```
> fit.knn_Test3
[1] Red
attr(,"prob")
[1] 0.6666667
Levels: Red Green
```

Figura 3: Salida de R para la *K=3*.

4. (45 pts. Practico)

Ejercicio 8, texto guia pagina 54. Los datos se cargan con la libreria de R llamada ISLR y la instruccion college=ISLR::College

Este ejercicio se relaciona con el conjunto de datos de College, que se puede encontrar en el archivo College.csv. Contiene una serie de variables para 777 diferentes universidades y colegios en los Estados Unidos.

(a) Utilice la función read.csv() para leer los datos en R. Asegúrese de que tiene el directorio establecido en la ubicación correcta para los datos.

Cargamos la función **read.csv()** para cargar los datos y los llamamos **college**, para esto usamos el siguiente comando:

```
college <- read.csv("College.csv", stringsAsFactors = TRUE)
```

Y con la función **head()** le damos un primer vistazo a la base de datos, el cual se muestra a continuación:

```
head(college)
```

	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30
3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66
4	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97
5	Alaska Pacific University	Yes	193	146	55	16	44	249	869						
6	Albertson College	Yes	587	479	158	38	62	678	41						

```
5     7560        4120   800      1500  76         72
6    13500        3335   500       675  67         73
  S.F.Ratio perc.alumni Expend Grad.Rate
1      18.1           12   7041         60
2      12.2           16  10527         56
3      12.9           30   8735         54
4       7.7           37  19016         59
5      11.9            2  10922         15
6       9.4           11   9727         55
```

(b) Observe los datos utilizando la función fix(). Usted debe notar que la primera columna es sólo el nombre de cada universidad. Realmente no queremos que R trate esto como datos. Sin embargo, puede ser útil tener estos nombres para más adelante.

```
rownames(college) -> college[,1]
college <- college[,-1]
fix(college)
```

La función **fix()** nos muestra lo siguiente:

	X	Private	Apps	Accept	Enroll
1	Abilene Christian University	Yes	1660	1232	721
2	Adelphi University	Yes	2186	1924	512
3	Adrian College	Yes	1428	1097	336
4	Agnes Scott College	Yes	417	349	137
5	Alaska Pacific University	Yes	193	146	55
6	Albertson College	Yes	587	479	158
7	Albertus Magnus College	Yes	353	340	103
8	Albion College	Yes	1899	1720	489
9	Albright College	Yes	1038	839	227
10	Alderson-Broaddus College	Yes	582	498	172
11	Alfred University	Yes	1732	1425	472
12	Allegheny College	Yes	2652	1900	484
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290
14	Alma College	Yes	1267	1080	385
15	Alverno College	Yes	494	313	157
16	American International College	Yes	1420	1093	220
17	Amherst College	Yes	4302	992	418

Figura 4: Salida de R de la función *fix()*.

c)

i. Utilice la función summary() para producir un resumen numérico de las variables del conjunto de datos.

Usando la función **summary()** se obtiene lo siguiente:

```
   summary(college)
                             X       Private
 Abilene Christian University:  1   No :212
 Adelphi University          :  1   Yes:565
 Adrian College              :  1
 Agnes Scott College         :  1
 Alaska Pacific University   :  1
 Albertson College           :  1
 (Other)                     :771
      Apps           Accept          Enroll
 Min.   :   81   Min.   :   72   Min.   :  35
 1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
 Median : 1558   Median : 1110   Median : 434
 Mean   : 3002   Mean   : 2019   Mean   : 780
 3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
 Max.   :48094   Max.   :26330   Max.   :6392

   Top10perc       Top25perc      F.Undergrad
 Min.   : 1.00   Min.   :  9.0   Min.   :  139
 1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
 Median :23.00   Median : 54.0   Median : 1707
 Mean   :27.56   Mean   : 55.8   Mean   : 3700
 3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
 Max.   :96.00   Max.   :100.0   Max.   :31643

  P.Undergrad         Outstate       Room.Board
 Min.   :    1.0   Min.   : 2340   Min.   :1780
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
 Median :  353.0   Median : 9990   Median :4200
 Mean   :  855.3   Mean   :10441   Mean   :4358
```

```
3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
Max.   :21836.0   Max.   :21700   Max.   :8124

    Books           Personal         PhD
Min.   :  96.0   Min.   : 250   Min.   :  8.00
1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
Median : 500.0   Median :1200   Median : 75.00
Mean   : 549.4   Mean   :1341   Mean   : 72.66
3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
Max.   :2340.0   Max.   :6800   Max.   :103.00

   Terminal       S.F.Ratio      perc.alumni
Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
Median : 82.0   Median :13.60   Median :21.00
Mean   : 79.7   Mean   :14.09   Mean   :22.74
3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
Max.   :100.0   Max.   :39.80   Max.   :64.00

    Expend        Grad.Rate
Min.   : 3186   Min.   : 10.00
1st Qu.: 6751   1st Qu.: 53.00
Median : 8377   Median : 65.00
Mean   : 9660   Mean   : 65.46
3rd Qu.:10830   3rd Qu.: 78.00
Max.   :56233   Max.   :118.00
```

ii. Utilice la función pairs() para producir una matriz de diagrama de dispersión de las primeras diez columnas o variables de los datos. Recuerde que puede hacer referencia a las primeras diez columnas de una matriz A utilizando A[,1:10].

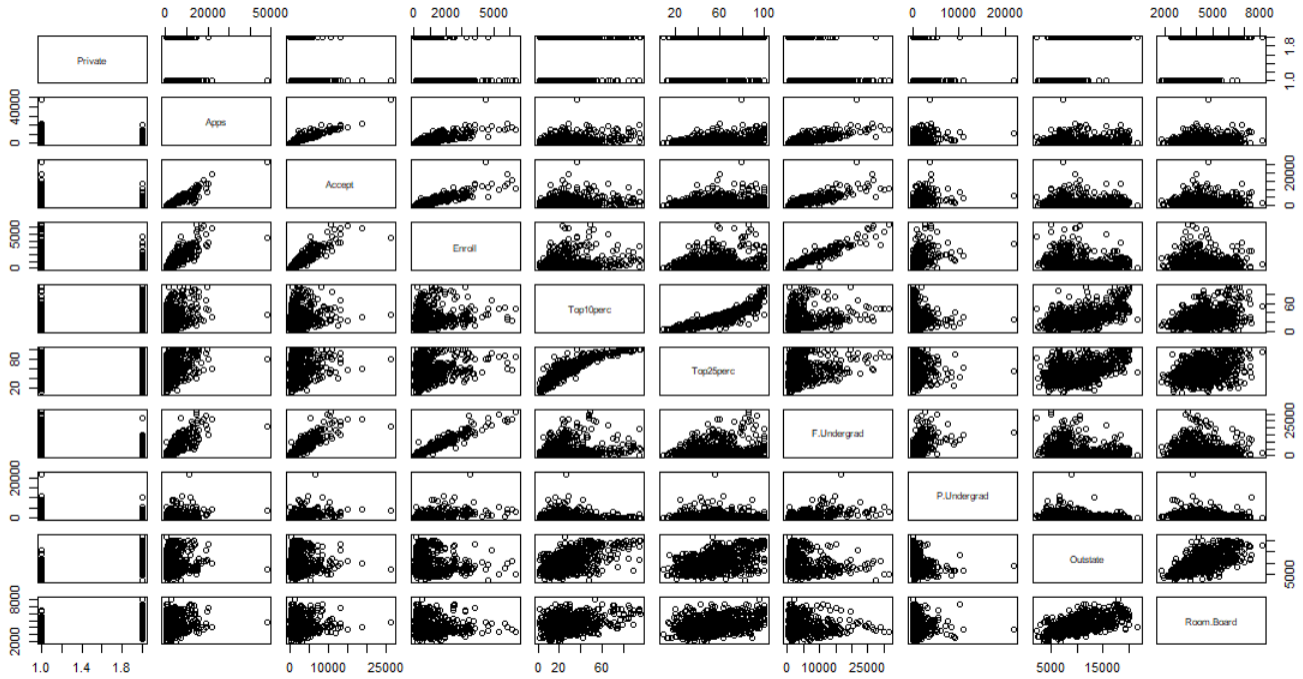


Figura 5: Salida de R de la función *pairs()*.

De la figura 5 se puede observar inicialmente que hay algunas relaciones entre algunas variables (que se podrían explicar mediante un modelo lineal sencillo) como lo son Enroll y F.Undergrad y Top10perc y Top25perc. Además se ven nubes de puntos sin ningún patrón en específico como en las variables Top25perc y Room.Board y Top10perc y Room.Board. Note que las variables Top10perc y Top25perc "se están comportando"de una manera similar ya que al ser el porcentaje de nuevos estudiantes superior de la clase de la escuela secundaria podrían estar compartiendo algunos datos. También se ve una gran dispersión en los puntos de la mayoría de variables y parece ser que la variable Private es bastante (equilibrada) con respecto a todas las otras variables.

iii. Utilice la función plot() para producir diagramas de caja en paralelo de Outstate versus Private.

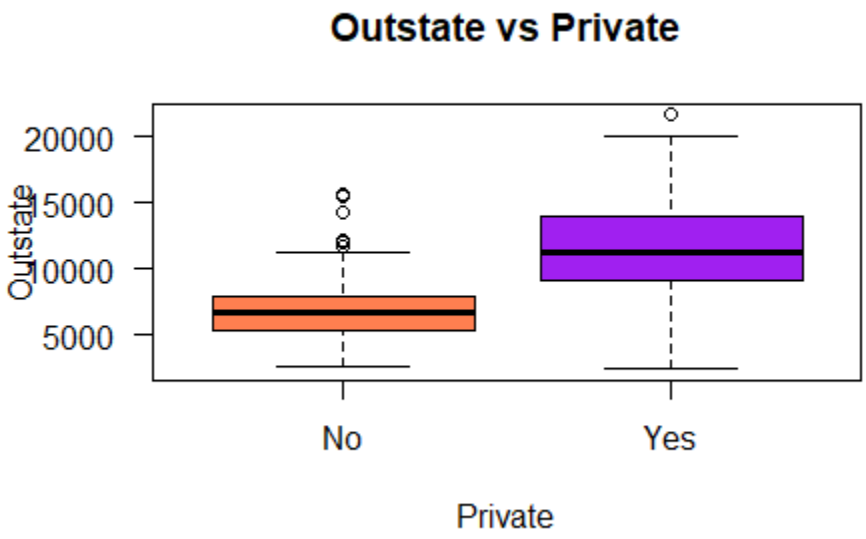


Figura 6: Boxplot: Outstate vs Private.

De la figura 6 se puede observar que en cuanto a cantidad son más la universidades privadas que reciben matriculas fuera del estado que las públicas. También se observa que en cuando a dispersión ambos tipos de universidad parecen ser constantes y que las universidades privadas tienen una dispersión un poco más alta que las universidades públicas. En ambos casos se presentan valores atípicos siendo en el caso público muchos más, lo cual podría ser considerado como valores a estudiar ya sea porque sean errores o algo particular pasa con estas universidades. Se ve también que las matriculas fuera del estado son diferentes en universidades públicas y privadas. Finalmente parece ser que en ambos casos (público/privado) los datos son simétricos por lo que podrían seguir una distribución normal.

iv. Crear una nueva variable cualitativa, llamada Elite, agrupando la variable Top10perc. Vamos a dividir las universidades en dos grupos en función de si la proporción de estudiantes que provienen del 10 % superior de sus clases de bachillerato supera el 50 %.

Para la creación de la nueva variable Elite se uso e siguiente código:

```
Elite <- rep("No",nrow(college))
```

```
Elite[college$Top10perc>50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college,Elite)
```

Utilice la función summary() para ver cuántas unidades de élite hay. Ahora utilice la función plot() para producir diagramas de caja en paralelo de Outstate versus Elite.

```
summary(college)
Elite
No :699
Yes: 78
```

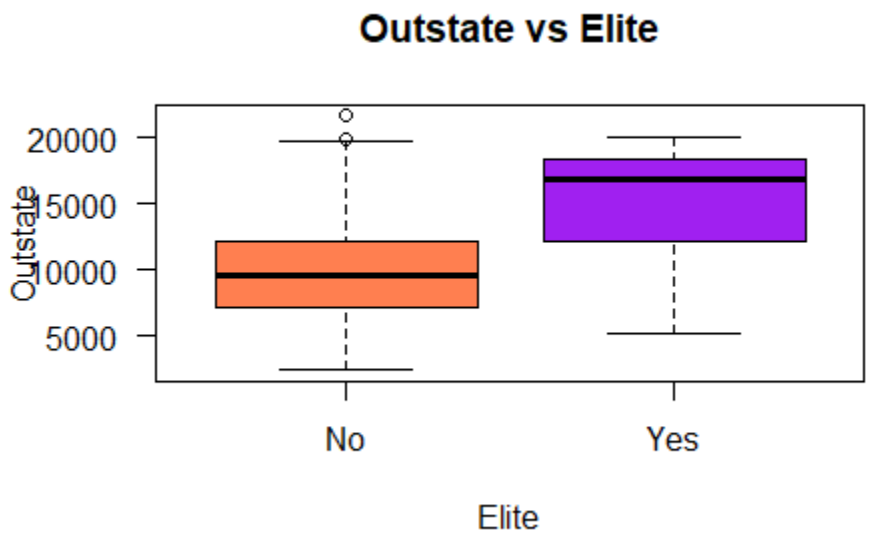


Figura 7: Boxplot: Outstate vs Elite.

En la figura 7 se puede observar que hay más universidades elite con matriculas fuera del estado que universidades públicas. Se ve que las universidades que si son de elite tienen una mayor dispersión que las universidades que no los son (aunque la diferencia no es muy grande gráficamente). También se observan dos valores atípicos en

las universidades que no son elite, los cuales son más grandes que todos los valores incluyendo los valores de las universidades que si son privadas, por lo cual son puntos de estudio y cuidado. Además, se puede apreciar que las matriculas fuera del estado en universidades de elite y las que no, son diferentes ya que una caja esta más arriba que la otra y no se traslapan y sus medianas estás a diferentes alturas. Finalmente, se ve que las universidades que no son de elite con matriculas fuera del estados siguen una distribución simétrica mientras que las que son de elite con matriculas fuera del estado no y sigue una distribución de una cola a la izquierda.

v. Utilice la función hist() para producir algunos histogramas con diferentes números de bins para algunos de los variables cuantitativos. Puede encontrar útil el comando par(mfrow-c(2,2)): dividirá la ventana de impresión en cuatro regiones para que se puedan realizar cuatro trazados simultáneamente. Modificar los argumentos de esta función dividirá la pantalla de otras maneras.

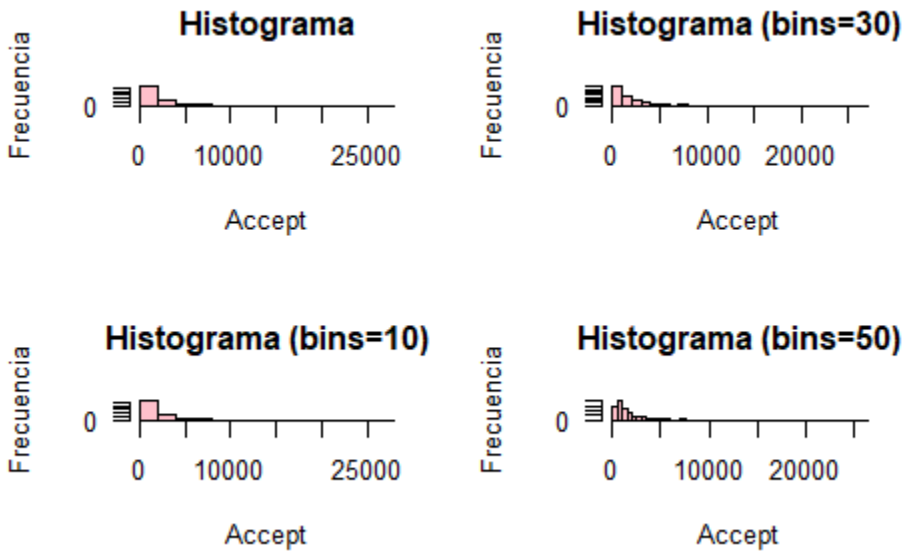


Figura 8: Histograma de la variable Accept.

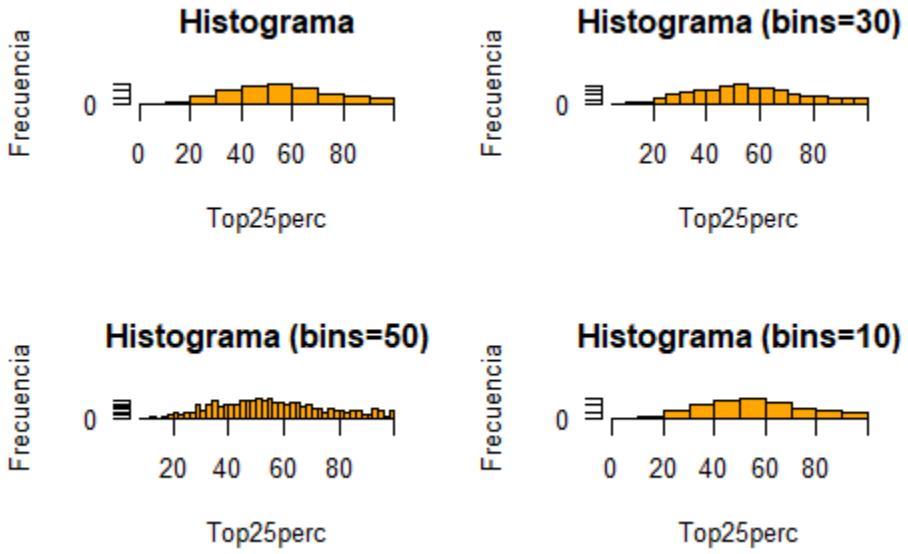


Figura 9: Histograma de la variable Top25perc.

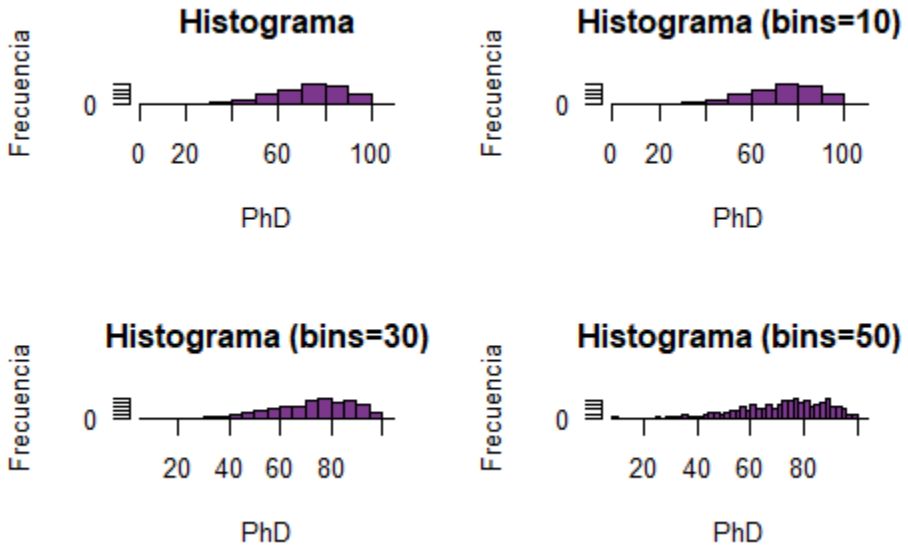


Figura 10: Histograma de la variable PhD.

En las figuras 8, 9 y 10 se pueden apreciar los histogramas de las variables Accept,

Top25perc y PhD respectivamente. Se observa que la variable Accept parece seguir una distribución con cola a la derecha ya que la mayoría de sus datos se encuentran en el intervalo de cero a 2500 (aproximadamente), también parece ser que hay una alta dispersión ya que los datos van mas allá de 25000 y por esto mismo puede ser que estos datos sean puntos atípicos. En cuanto a la variable Top25perc, se aprecia que los datos parecen seguir una distribución simétrica con una aglomeración de los datos alrededor del 50 (aproximadamente). Parece ser que su dispersión no es tan alta como en el caso de la variable Accept y no parece que hayan puntos atípicos en esta variable. Finalmente para la variable PhD, se ve que los datos parecen seguir una distribución con cola a la derecha (asimétrica), con un centro o pico alrededor de 80 (aproximadamente). Al igual que la variable Top25perc no parece que haya una dispersión muy alta, además parece ser que hay ciertos valores alejados del centro por lo cual estos sería candidatos a ser puntos atípicos (los cuales podrían ser responsables de la dispersión en esta variable).

vi. Continúe explorando los datos y proporcione un breve resumen de lo que descubra.

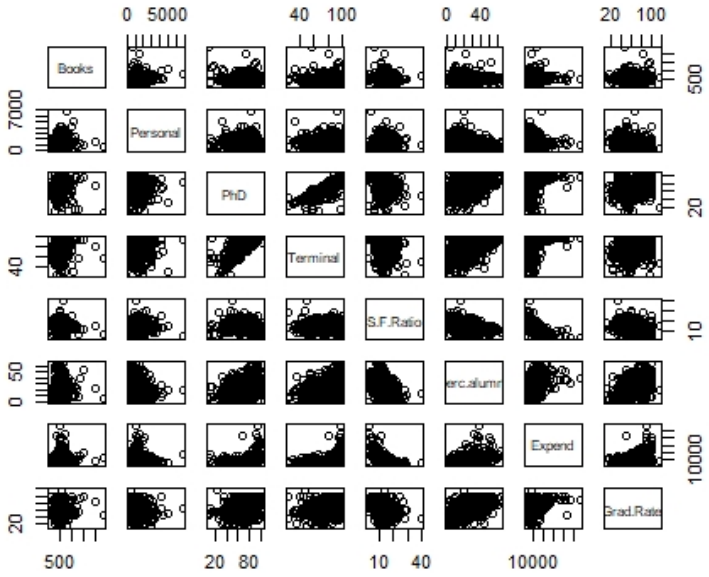


Figura 11: Salida de R de la Función pairs()

En la figura 11 podemos observar las demás variables que no se habían graficado con la función pairs() a simple vista no se logra ver alguna relación lineal, la única que se podia creer que tienen una relación positiva serian las varibales Terminal con PhD,pues sus datos van e ascenso pero no esta del todo calro si hay alguna otra relación.

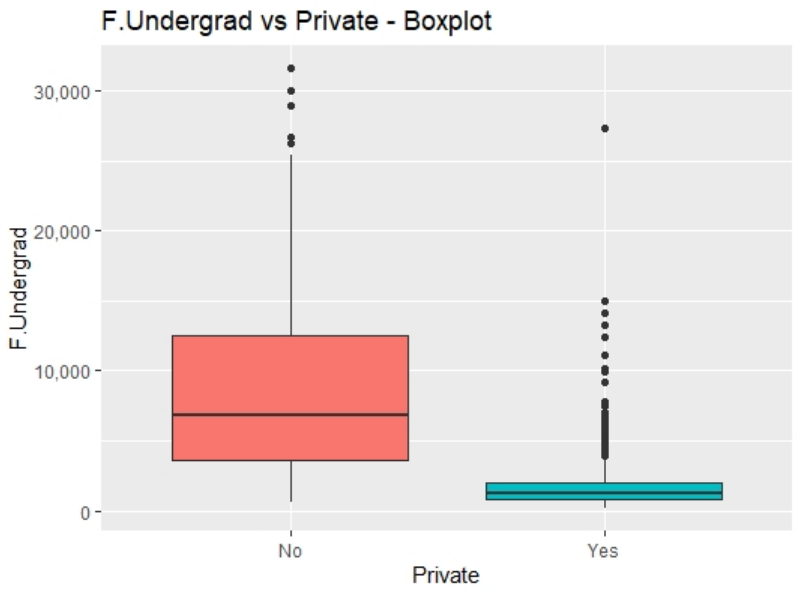
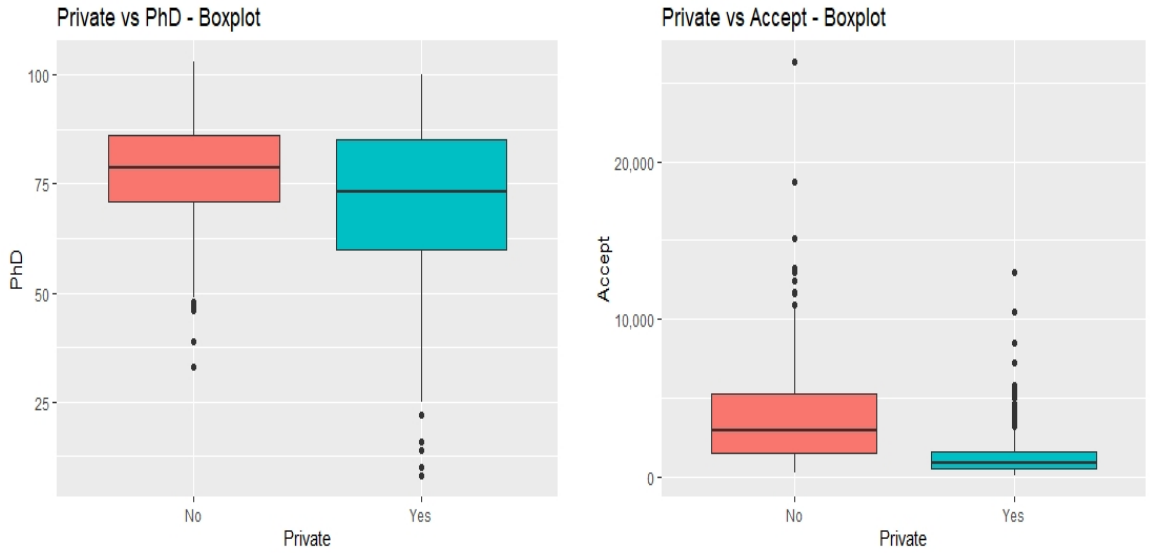


Figura 12: BoxPlot: F.Undergrad vs Private

En esta figura 12 se quiso comparar el número de estudiantes(F.Undergrad) diferenciado por si es de Universidad Privada o no, se nota claramente que las Universidades públicas tienden a tener un número significativamente mayor de estudiantes universitarios, ya que su media es mayor a comparación de la media de las Universidades privadas, sin embargo tiene muchos valores atípicos, en especifico uno que sobresale bastante, también las Universidad Publicas tienen unos valores atípicos.

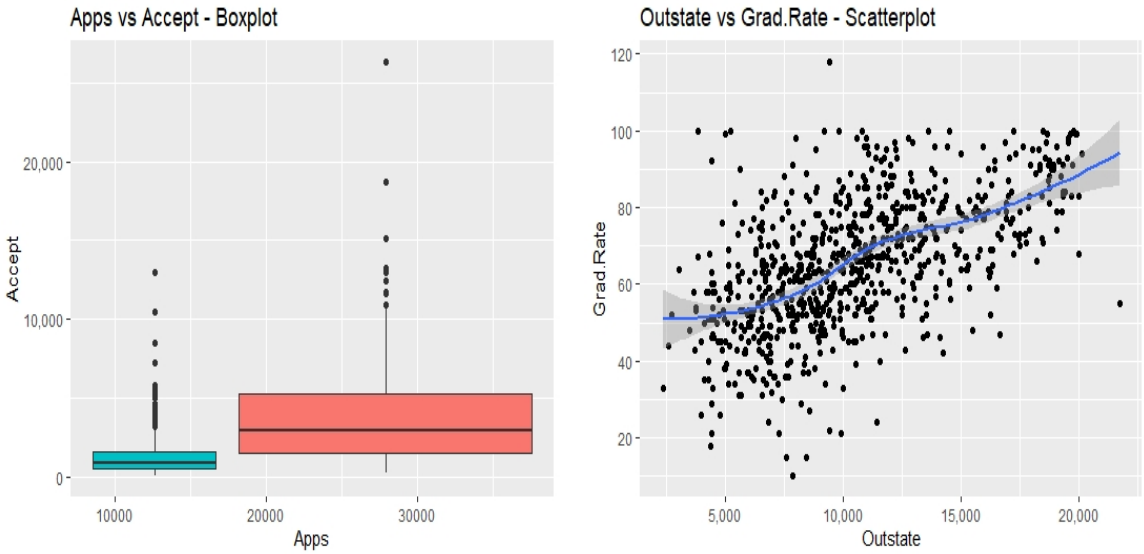


(a) Private vs PhD

(b) Private vs Accept

Figura 13: BoxPlot de Variables con Private.

En la figura 13 se realizaron dos BoxPlot para comparar el número de profesores que tienen doctorado y el número de solicitudes aceptadas (estudiantes aceptados) dependiendo de la Universidad (Private), en la primera imagen no se logra ver una gran diferencia entre las Universidades Publicas y Priavadas pues sus medias no se diferencian de gran manera, aunque en ambas tienen valores atípicos debajo de la caja, es decir, hay algunas Universidades que no cuentan con muchos profesores con doctorado (aunque también debe afectar lo grande que sea la Universidad, puede que haya Universidades mas pequeñas), la cola inferior del Box Plot de las Universidades Privadas es mucho mas larga y tomas valores moenores a comparación de las Universidades públicas aunque tiene algo de sentido pues hay mas estudiantes en Universidades públicas que privadas.
Por otro lado en la segunda imagen las Universidades Públicas hay mayor aceptación de solicitudes para ingresar a la Universidad que las Universidades privadas, pero esto puede ser también por lo antes mencionado, cabe destacar que hay varios atípicos en las dos cajas pero esto ya puede ser que unas Universidades sean más grandes que otras.



(a) BoxPlot: Apps vs Accept.

(b) ScatterPlot de Outstate vs Grad.Rate.

Figura 14: Dos imágenes en la misma figura.

En la figura 14 vemos dos imágenes, en la primera un Box Plot el cual se parece bastante a los anteriores pero este compara las solicitudes que se hicieron versus el número de aceptados, dependiendo si es Universidad Pública o Privada
En la segunda imagen vemos un Gráfico de dispersión de Outstate con Grad.Rate pues se vio que estas dos variables tienen una relación positiva, se puede decir que su relación puede llegar a ser linealpositiva ,sin embargo, también especular que la verdadera relación podría ser logarítmica o cuadrática en un rango mayor de Outstate.

Con base en lo anterior visto se quiso realizar un modelo lineal para explicar el Grad.Rate (Tasa de Graduación) con las demás variables planteadas, para poder ver que variables son las que ayudan a aumentar esa tasa, y si ser de una Universidad priavada o pública puede afectar esto.

Como primero se realizo un modelo lineal simple con todas las variables que tiene los datos dando un $R^2 = 44.88\,\%$, sin embargo viendo los valores-p de los estimadores se vio que se podían eliminar variables, por esto se tomo la decisión de realizar

regresión por segmentos, donde se eligieran las variables que afectaban a Grad.Rate; en la siguiente imagen se muestra las variables escogidas.

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Outstate	addition	0.326	0.326	175.3210	6322.0286	14.1077
2	Top25perc	addition	0.378	0.376	104.9720	6262.3667	13.5676
3	perc.alumni	addition	0.404	0.402	70.0990	6230.9510	13.2876
4	P.Undergrad	addition	0.415	0.412	56.2350	6218.1207	13.1699
5	Apps	addition	0.432	0.428	34.4750	6197.3655	12.9868
6	Room.Board	addition	0.440	0.436	25.0240	6188.1555	12.9019
7	Expend	addition	0.448	0.443	15.8680	6179.0772	12.8185
8	Personal	addition	0.453	0.447	11.6390	6174.8225	12.7754
9	Private	addition	0.456	0.450	8.6940	6171.8200	12.7426

Figura 15: Selección de Variables por Regresión por Segementos

Se puede notar que la primera variable es Outsate como se vio en el diagrama de dispersión pues tienen una relación positiva, otras variables son Top25perc, perp.alumni, P.Undergrad, Apps entre otras, se realizo nuevamente el modelo lineal con solo las variables escogidas dando el siguiente resumen.

```
Call:
lm(formula = Grad.Rate ~ Outstate + Top25perc + perc.alumni +
    P.Undergrad + Apps + Room.Board + Expend + Personal + Private,
    data = college)

Residuals:
    Min      1Q  Median      3Q     Max
-52.345  -7.551  -0.426   7.040  51.789

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.9174907  2.5635570  12.841  < 2e-16 ***
Outstate     0.0010226  0.0002203   4.641 4.07e-06 ***
Top25perc    0.1763996  0.0304582   5.792 1.02e-08 ***
perc.alumni  0.2876422  0.0483114   5.954 3.98e-09 ***
P.Undergrad -0.0016678  0.0003611  -4.619 4.52e-06 ***
Apps         0.0009022  0.0001609   5.606 2.89e-08 ***
Room.Board   0.0018262  0.0005732   3.186  0.00150 **
Expend      -0.0003888  0.0001288  -3.019  0.00262 **
Personal    -0.0018394  0.0007491  -2.455  0.01429 *
PrivateYes   3.3935160  1.5246563   2.226  0.02632 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.74 on 767 degrees of freedom
Multiple R-squared:  0.4561,    Adjusted R-squared:  0.4497
F-statistic: 71.46 on 9 and 767 DF,  p-value: < 2.2e-16
```

Figura 16: Summary de Modelo Lienal para Grad.Rate

EL summary se puede notar que tiene un $R^2 = 44.97\,\%$, dando un R^2 parecido al anterior y que sus variables son todas significativas, puede que se pueda mejorar este modelo intentando poner algunas de sus variables cuadráticas, logrando explicar mejor el modelo.

```
Call:
lm(formula = Grad.Rate ~ Outstate + I(Outstate^2) + Top25perc +
    I(Top25perc^2) + perc.alumni + I(perc.alumni^2) + P.Undergrad +
    I(P.Undergrad^2) + Apps + Room.Board + Expend + Personal +
    Private, data = college)

Residuals:
    Min      1Q  Median      3Q     Max
-52.121  -7.243  -0.447   7.184  52.761

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.611e+01  4.732e+00   5.517 4.73e-08 ***
Outstate          1.669e-03  7.101e-04   2.351 0.019002 *
I(Outstate^2)    -2.923e-08  2.957e-08  -0.989 0.323165
Top25perc         2.558e-01  1.282e-01   1.995 0.046369 *
I(Top25perc^2)   -6.941e-04  1.123e-03  -0.618 0.536886
perc.alumni       5.059e-01  1.460e-01   3.466 0.000559 ***
I(perc.alumni^2) -4.252e-03  2.590e-03  -1.641 0.101117
P.Undergrad      -2.276e-03  6.651e-04  -3.421 0.000657 ***
I(P.Undergrad^2)  4.674e-08  4.456e-08   1.049 0.294488
Apps              9.024e-04  1.659e-04   5.438 7.25e-08 ***
Room.Board        1.833e-03  5.826e-04   3.147 0.001716 **
Expend           -2.951e-04  1.392e-04  -2.120 0.034363 *
Personal         -1.874e-03  7.510e-04  -2.495 0.012808 *
PrivateYes        2.150e+00  1.634e+00   1.316 0.188510
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.72 on 763 degrees of freedom
Multiple R-squared:  0.4612,    Adjusted R-squared:  0.452
F-statistic: 50.24 on 13 and 763 DF,  p-value: < 2.2e-16
```

Figura 17: Summary del Modelo Lineal Cuadrático

Por último se realizo un modelo cuadrático el cual da un $R^2 = 45.52\,\%$, es decir, el modelo no mejoro mucho con las variables al cuadrado, puede que otro modelo sea mejor.