Cómputo estadístico (Tarea 3)

J. Antonio García Ramirez

Septiembre 26, 2018

Ejercicio 1

Considerando el conjunto de datos ***College*** *de la libreria ISLR, vamos a predecir el número de solicitudes recibidas(Apps) usando las otras variables del conjunto de datos.*

a) *Divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba y ajusta un modelo lineal usando mínimos cuadrados en el conjunto de entrenamiento. Reporta el error de prueba obtenido.*

Dividimos el conjunto de datos aleatoriamente con un conjunto de prueba correspondiente al 30% de la muestra original.

Para poder comparar resultados entre la estimación de mínimos cuadrados y los métodos de Ridge, Lasso, PCR y PLS se escala el conjunto de datos para que tengan media cero y varianza igual a uno, excepto la variable **Private** ya que es categórica.

Después de fijar la semilla y efectuar la división del conjunto de datos se obtuvo un error cuadrático medio de 0.06297405 en el conjunto de prueba utilizando un modelo lineal con todas las variables y los datos escalados.

b) *Ajusta un modelo de regresión Ridge sobre el conjunto de entrenamiento, con λ elegido por validación cruzada. Reporta el error de prueba obtenido.*

Utilizando validación cruzada con 10 folds, regresión Ridge y un $\lambda = 0.001$ se obtiene un error de prueba de 0.05405734 sobre los datos escalados

```
## [1] 0.001
```

```
## [1] 0.05405734
```

c) *Ajusta un modelo de lasso en el conjunto de entrenamiento, con λ elegido por validación cruzada. Reporta el error de prueba obtenido, junto con el número de estimaciones de coeficientes no nulos.*

Utilizando validación cruzada, regresión Lasso y un $\lambda = 0.001$ se obtiene un error de prueba de 0.08525753 sobre los datos escalados.

d) *Ajusta un modelo de* ***PCR*** *en el conjunto de entrenamiento, con* $\boldsymbol{M}$ *elegido por la validación cruzada. Reporta el error de prueba obtenido, junto con el valor de* $\boldsymbol{M}$ *seleccionado mediante validación cruzada.*

Utilizando validación cruzada con 10 folds, regresión PCR y 7 componentes principales (en vista de que en el conjunto de entrenamiento la octava componente es de poca fignificancia) se obtiene un error de prueba de 0.09222398 sobre los datos escalados.

e) *Ajusta un modelo* ***PLS*** *en el conjunto de entrenamiento, con* $\boldsymbol{M}$ *elegido por la validación cruzada. Informe el error de prueba obtenido, junto con el valor de* $\boldsymbol{M}$ *seleccionado mediante validación cruzada*

Utilizando validación cruzada, PLS como método de regresión y 7 componentes principales (en vista de que en el conjunto de entrenamiento la octava componente de ppls es de poca significancia) se obtiene un error de prueba de 0.06016777 sobre los datos escalados.

```
## [1] 0.06016777
```

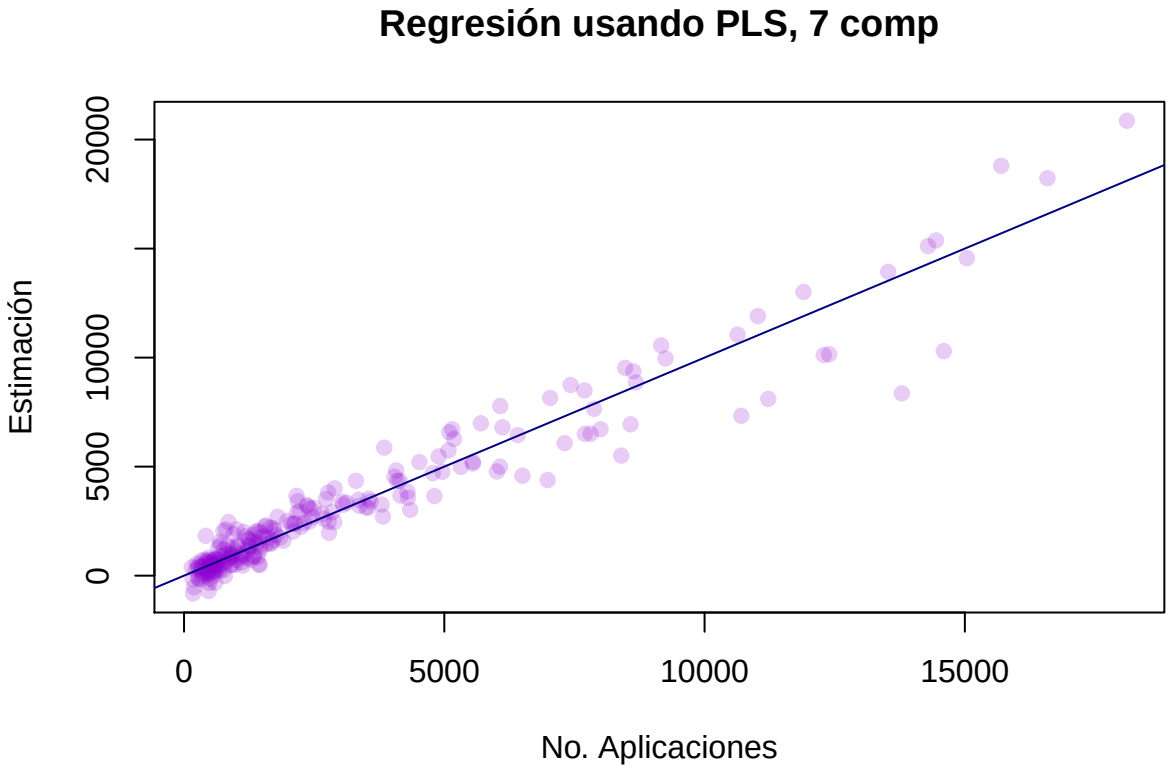
f) *Comenta los resultados obtenidos. ¿Con qué precisión podemos predecir el número de solicitudes de estudios universitarios recibidas? ¿Hay mucha diferencia entre los errores de prueba resultantes de estos cinco enfoques?*

Considerando los resultados obtenidos y en aras de retener un modelo altamente preciso y que también posea una interpretación sencilla. Reportamos los resultados de la regresión PLS.

Como podemos apreciar en la gráfica siguiente el ajuste es apropiado, los valores de la gráfica se presentan en sus unidades originales. Si bien el MSE del modelo es del orden de 0.06 las predicciones poseen un error relativo mayor al 200% en vista de que la predicción del modelo se ve afectada por casos aislados.

Con el modelo PLS como base la estimación de Ridge posee mejores resultados pero los consideramos marginales.

```
## [1] 0.2452912
```



Ejercicio 2

Se ha visto que a medida que aumenta el número de características de un modelo, el error de entrenamiento disminuirá necesariamente, pero el error de prueba no. Explorar esto con datos simulados

a) *Genera un conjunto de datos con* $p = 20$ *características,* $n = 1000$ *observaciones y un vector de respuesta cuantitativo generado de acuerdo con el modelo:*

$$Y = X\beta + \epsilon$$

Donde β *tiene algunos elementos que son exactamente iguales a cero.*

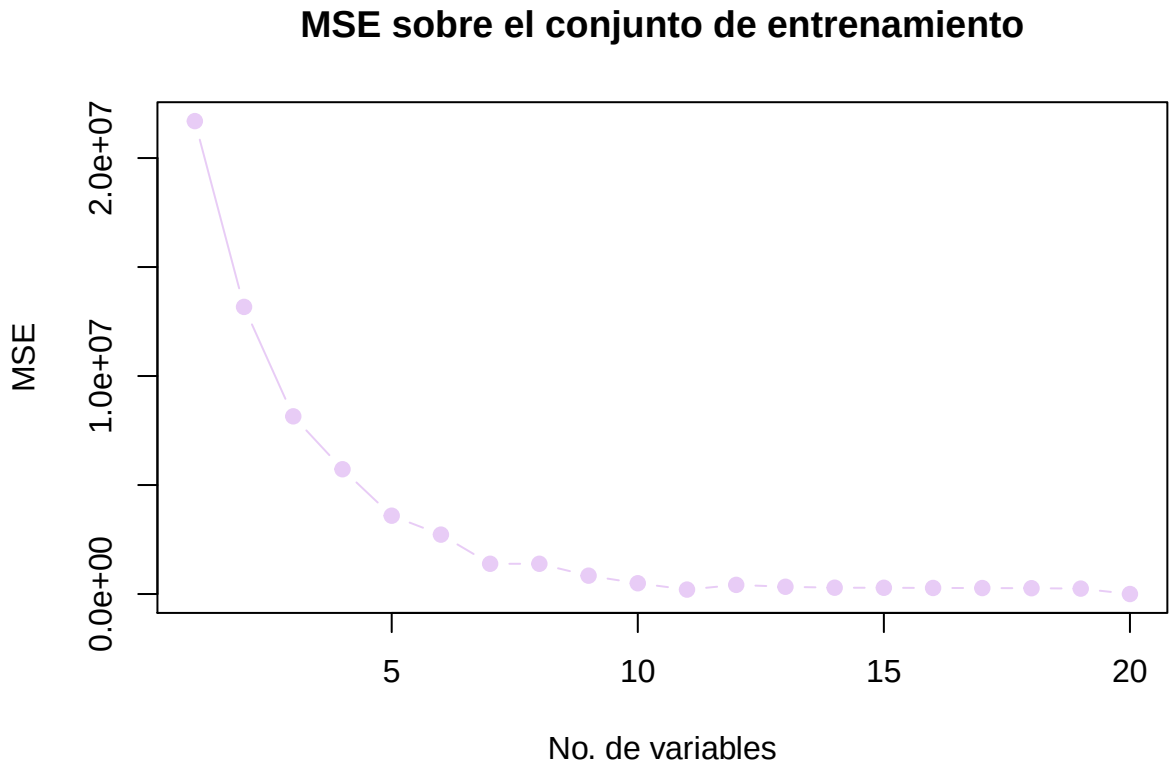
Generamos las características, las 20 variables son normales con media 1,2,. . . ,20 y desviación estándar $1, 2^2, ..., 20^2$. Y los primeros coeficientes $\beta_1, \beta_2, ...\beta_5$ son cero.

b) *Divide tu conjunto de datos en un conjunto de entrenamiento que contenga* 100 *observaciones y un conjunto de pruebas que contenga* 900 *observaciones.*

Realizamos la partición mencionada, del conjunto $\{1, ..., 1000\}$ escogemos aleatoriamente 100 observaciones para el conjunto de prueba y 900 para el conjunto de prueba.

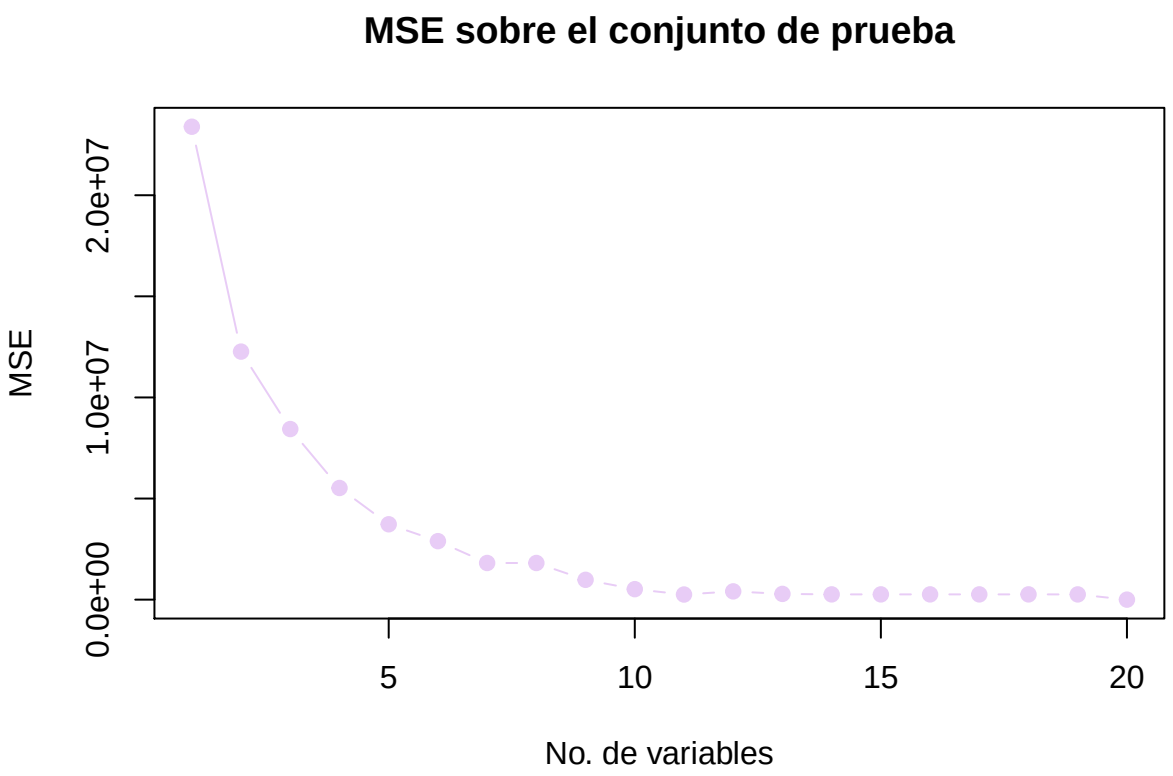
c) *Realiza la selección del mejor subconjunto sobre el conjunto de entrenamiento y gráfica el error de entrenamiento MSE asociado con el mejor modelo en cada tamaño.*

Utilizando el método exhaustivo y el criterio BIC, se procedió a seleccionar el mejor subconjunto de variables para un modelo lineal con p =1 hasta p=20 con el conjunto de prueba. Los MSE para cada modelo sobre el conjunto de prueba se encuentran en la siguiente gráfica, como es de esperar el MSE sobre el conjunto de prueba disminuye conforme aumenta el número de regresores:



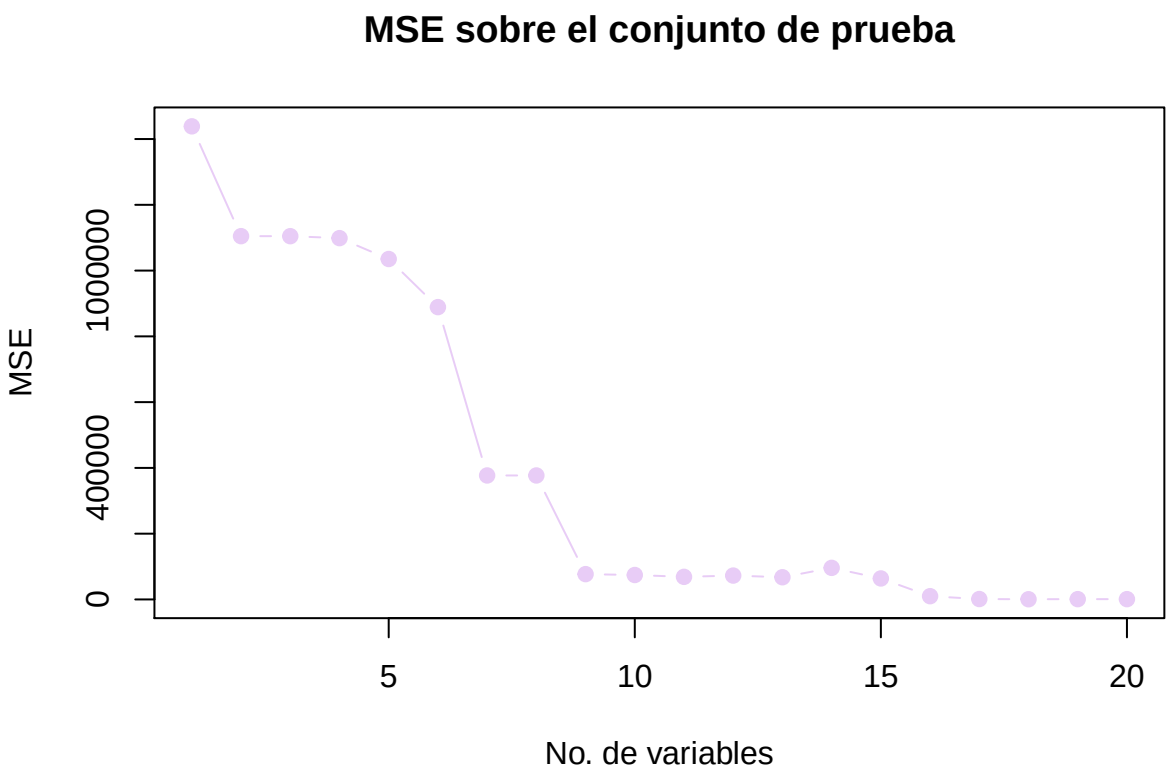
d) *Grafica el error de prueba **MSE** asociado con el mejor modelo de cada tamaño.*

Los MSE sobre el conjunto de prueba para cada número de regresores se presentan en la siguiente gráfica, es de notar que al igual que con el conjunto de entrenamiento este tiende a disminuir conforme aumenta el número de regresores (aunque en diferente escala).



e) *¿Para qué tamaño de modelo el error de prueba MSE toma su valor mínimo? Comenta tus resultados. Si toma su valor mínimo en un modelo que sólo contiene una interceptación o un modelo que contenga todas las características, entonces juega con la forma en la que estás generando los datos en (a) hasta que aparezca un escenario en el que el error de prueba **MSE** se minimiza para un tamaño de modelo intermedio.*

En vista de los resultados se volvió a generar el conjunto de datos de prueba, las variables se construyeron de nuevo como normales con media cero y desviación estándar de $1, 2^2, 3^2, ..., 20^2$ con ceros los últimos 6 coeficientes $\beta_{15}, \beta_{16}, ., \beta_{20}$, el menor MSE se da con el modelo con 18 características que no engloba a las variables $x \sim Norm(0, 16^2)$ y $x \sim Norm(0, 20^2)$.

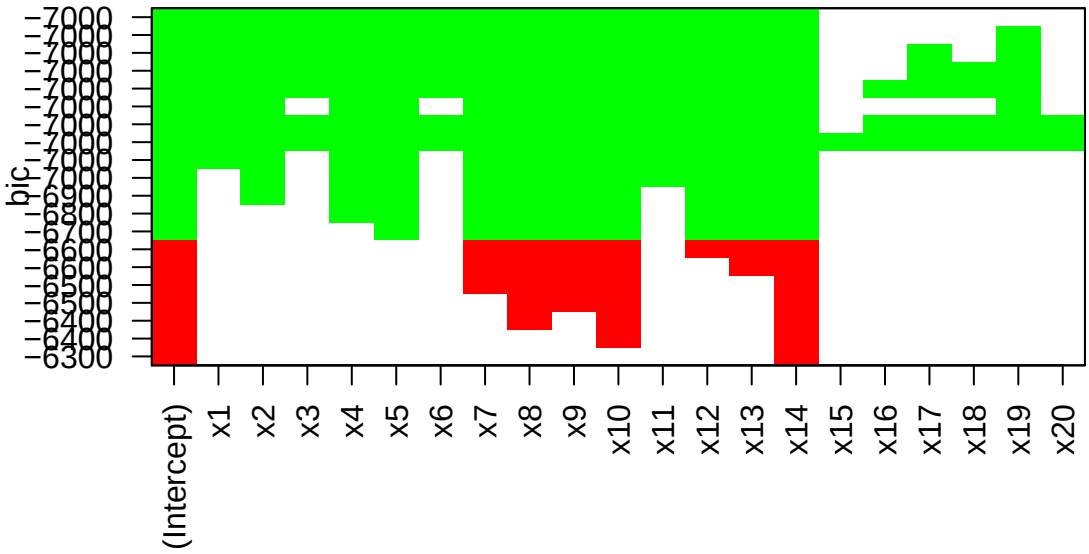



```
## [1] 18
```

f) *¿Cómo se compara el modelo con el que se minimiza el error de prueba con el modelo verdadero utilizado para generar los datos? Comenta sobre los valores de los coeficientes*

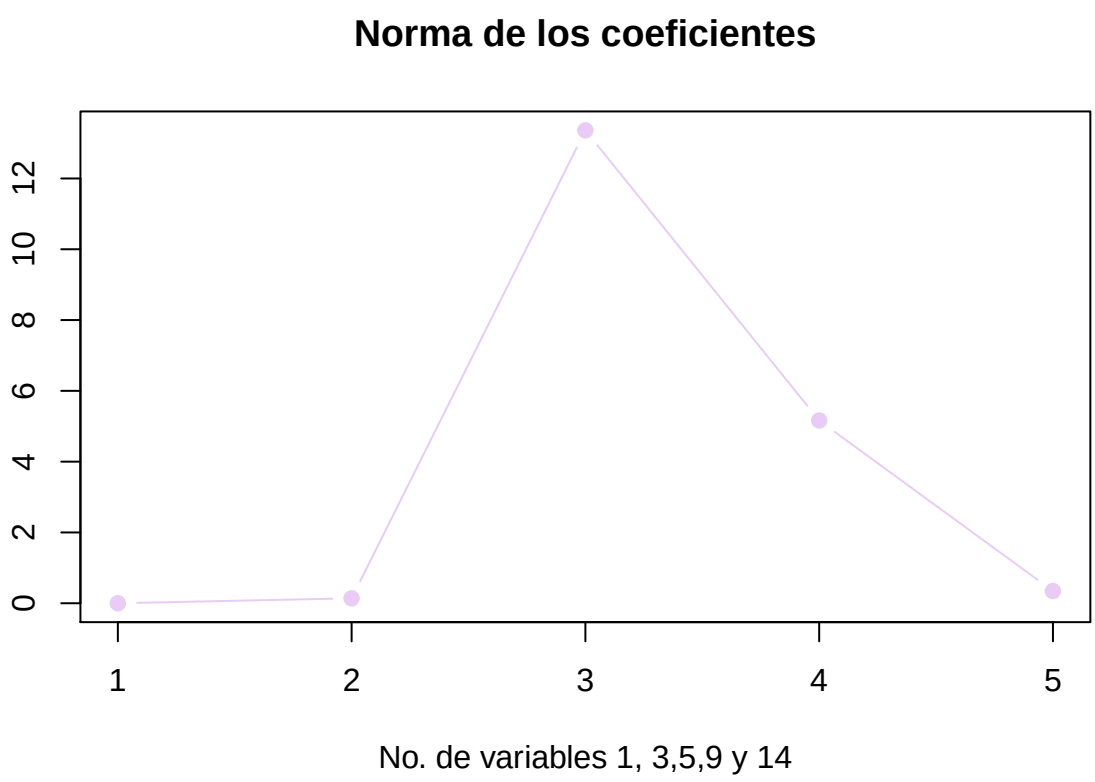
Los modelos son muy diferentes. El modelo que genera los datos es una combinación lineal de las primeras 15 variables normales con media cero y desviación estándar de $1, 2^2, ...$ hasta 15^2, $x_i \sim Norm(0, i^2)$ y los cinco últimos coeficientes son cero. Si bien con el método exhaustivo el modelo que tiene menor MSE en el conjunto de prueba se logra incluyendo 18 variables, todas las variables excepto la que tiene desviación de 16^2 y la de 20^2. En el gráfico siguiente se logra apreciar que las últimas variables a considerar son las de desviaciones $17^2, 18^2$ y 19^2 que son precisamente las que tienen un coeficiente de cero en el modelo.

variables que en el modelo de datos tienen un coeficiente de cero las últimas seis variables.



g) *Crea un gráfico que muestre* $\sqrt{\sum_{j=1}^{p}(\beta - \hat{\beta}_j^r)^2}$ *para un rango de valores de* r*, donde* β_j^r *es el j-esimo coeficiente estimado para el mejor modelo que contiene* r *coeficientes. Comenta lo que observas. ¿Cómo se compara esto con el gráfico del error de prueba de d)?*

Se eligio $r = 1, 3, 5, 9, 14$ en el gráfico anterior podemos notar que conforme el número de regresores aumenta la norma de la diferencia entre los coeficientes estimados y los reales tiende a disminuir lo cual es ad hoc con el resultado del inciso d pues al aumentar el número de regresores el error de prueba tiende a disminuir.


Norma de los coeficientes
0
2
4
6
8
10
12
1
2
3
4
5
No. de variables 1, 3,5,9 y 14