UNIVERSIDAD
NACIONAL
DE COLOMBIA

MEJOR MODELO PREDICTIVO SOBRE LOS DATOS DE ADMISIÓN DE LA UNIVERSIDAD NACIONAL DE COLOMBIA 2015

ALUMNOS:

Cardenas Rosas Maria Camila
Estadística
mcardenasro@unal.edu.co

Conde Hernandez Karen Lucianna
Matemáticas
kconde@unal.edu.co

Forero Laiton Angie Daniela
Matemáticas
anforerol@unal.edu.co

TEMA:
Análisis de Regresión
Modelos Con Enfoque Predictivo

FECHA DE ENTREGA:
19 de junio del 2023

1. Criterio de habilidad predictiva:

Para comparar la habilidad predictiva de los modelos estudiados, utilizamos como criterio el error cuadrático medio (**MSE**) por varias razones. Entre estas podemos destacar su sensibilidad a los errores de predicción, pues los penaliza de manera cuadrática. Esto significa que los errores más grandes tienen un impacto significativamente mayor en la métrica en comparación con los errores más pequeños.

Por otro lado, está relacionado con el sesgo y la varianza. Un modelo con alto sesgo puede subestimar o sobreestimar los valores reales, mientras que un modelo con alta varianza puede ser muy sensible a las fluctuaciones en los datos de entrenamiento. Al minimizar el **MSE** , se busca encontrar un modelo que tenga un equilibrio adecuado entre sesgo y varianza. Al elegir el modelo con el **MSE** más bajo, se busca encontrar aquel que minimice la discrepancia entre las predicciones y los valores reales.

2. Validación o de validación cruzada:

Detalle los mecanismos de validación o de validación cruzada que se utilizaron.

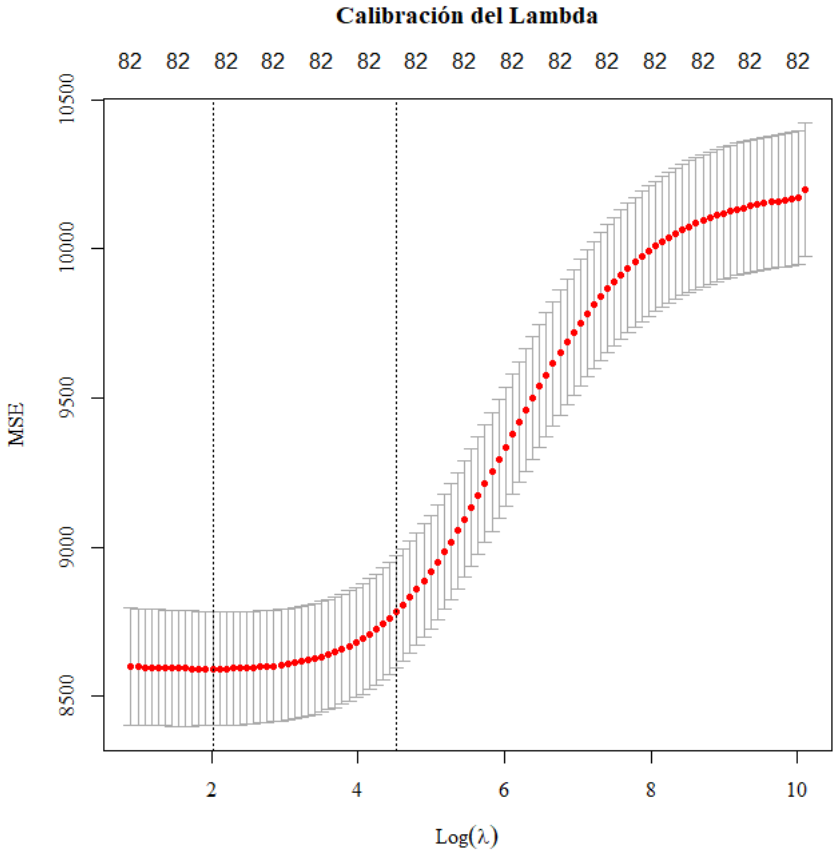
Reporte:

Con el objetivo de predecir el puntaje de admisión a la Universidad Nacional de Colombia de una persona que reúne ciertas características, se utilizaron varios métodos de modelación, algunos de estos son: el método de selección automática de variables, método lineal generalizado con la familia Gamma, método no paramétrico con splines, método no paramétrico con regresión local y finalmente los métodos de regularización.

Este último enfoque busca penalizar los valores elevados de las estimaciones de los coeficientes del modelo de regresión introduciendo un hiperparámetro (λ), este hiperparámetro se calibro en nuestro proyecto con ayuda de la validación cruzada en 10 etapas de los datos que fueron escogidos como entrenamiento, por ejemplo, en los modelos de regresión RIDGE se utilizó el siguiente código (que toma por defecto 10 etapas) con el cual se puede obtener el mejor lambda:

```
cv.out4 <- cv.glmnet(x[train, ], Puntaje[train], alpha = 0)
bestlam4 <- cv.out4$lambda.min
```

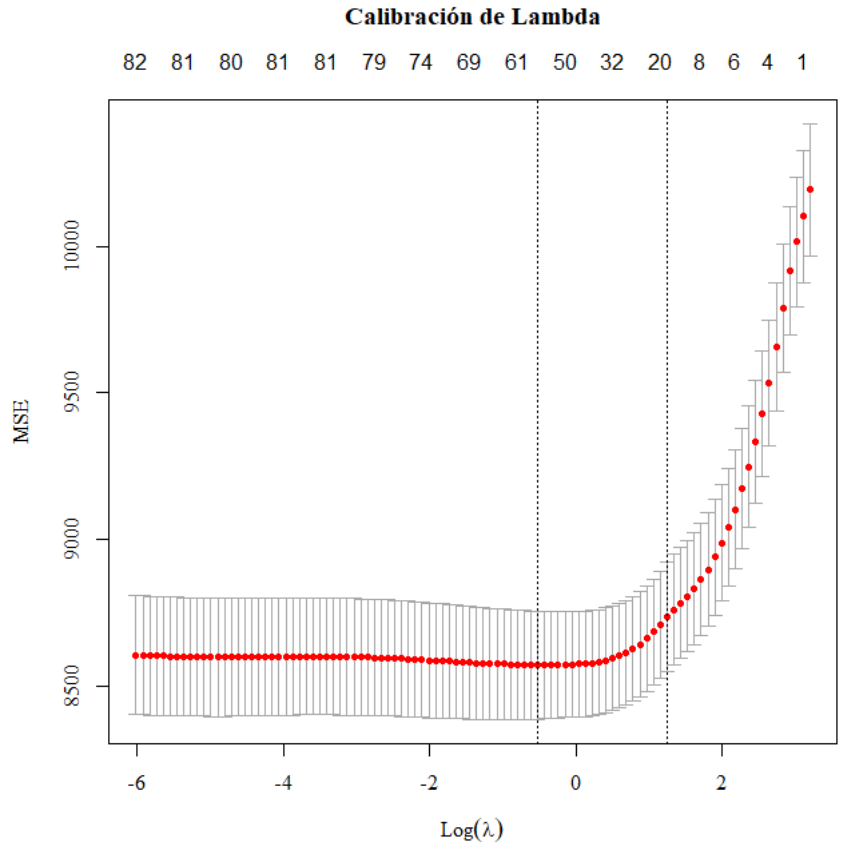
También se realizó la siguiente gráfica de como se comporta el MSE en términos del logaritmo de Lambda.



Ahora, para los modelos en los cuales se utilizó la regresión LASSO se realizo un procedimiento análogo con el siguiente código (solo cambia el valor de $\alpha = 1$ pues es más adecuado para este método) de donde, se obtiene el mejor lambda:

```
cv.out1 <- cv.glmnet(x[train, ], Puntaje[train], alpha = 1)
bestlam1 <- cv.out1$lambda.min
```

De igual manera que en el anterior se realiza una gráfica para ver como se comporta el MSE en términos del logaritmo de Lambda.



Así se obtuvo el hiperparámetro necesario para realizar la regularización y para obtener la mejor habilidad predictiva utilizando este tipo de modelos.

Ahora, en los primeros enfoques no fue necesario utilizar validación pues las estimaciones de nuestros modelos no dependían de algún hiperparámetro que tocara estimar con anterioridad, la escogencia del grado de los splines se realizo en términos de su habilidad predictiva para los datos de testeo.

3. Construcción de los modelos:

Lo primero que hicimos fue modificar la base de datos, definiendo la variable *EDAD_PRESS* como la edad que tenía cada individuo al momento de presentar el examen de admisión. Después de realizar este cálculo, borramos las columna *ASP_FECHANACIMIENTO* y la reemplazamos por *EDAD_PRESS*. Así transformamos esta variable en una cuantitativa.

Como primer paso se realizo un modelo lineal que incluye todas las variables para tener un MSE de referencia.

Como consecuencia de la creación de variables dummies, resultaron 83 variables, por esto no incluimos el método de selección *Best Subset*, pues su peso computacional sería muy alto. Luego, nos concentramos en los métodos de selección *Forward* y *Backward*.

Cuando ejecutamos los métodos *Forward* y *Backward* sin interacciones obtuvimos modelos con 25 y 21 variables respectivamente, ambos tenían valores de MSE similares y menores que el MSE de referencia.

Después de esto intentamos construir un modelo con intercciones que para nosotras tenían sentido. Incluimos el sexo con el estrato, el tipo de admisión, el departamento, la naturaleza del colegio y la naturaleza del colegio con el departamento. Sin embargo, tuvimos problemas ocasionados por la falta de representación en algunas categorías en el subconjunto de testeo. Por ejemplo, no se calcula el coeficiente para *ASP_SEXOM*ASP_RAYA_TIPOMBMP*, pues solo hay una persona que fue admitida por este tipo de admisión y no pertenece al grupo de testeo. Esto se arregló agregando ceros en las columnas correspondientes a estas categorías en la matriz de testeo. Volvimos a estimar el modelo, pero el MSE resultó ser muy grande.

Seguido a esto pasamos a utilizar técnicas de regularización, modelos lineales generalizados y no parametricos, en donde obtuvimos habilidades predictivas mucho mejores que en las anteriores (la explicación de estos modelos se puede ver en el siguiente ítem), en algunos las interacciones mejoraban la habilidad mientras que en otros la empeoraban, adicionalmente a esto no se agregaron transformaciones en las variables explicativas pues la mayoría eran categorías y la única variable cuantitativa tenia muchos valores concentrados en los mismos números.

4. Mejor desempeño:

A continuación se mostrarán 3 de los modelos realizados que tienen el mejor desempeño en habilidad predictiva utilizando como criterio el MSE. La explicación del hallazgo de estos 3 modelos se podrá encontrar inmediatamente después de la tabla.

Modelo	Descripción del modelo	Criterio predicción sobre testeo
1	El mejor modelo que se obtuvo esta dado por uno no paramétrico en el cual se incluyeron las variables *EDAD_PRES* Y *ASP_ANOTERMINACION* como variables cuantitativas con splines de grados 8 y 10 respectivamente, adicionalmente se incluyeron las variables categóricas sin interacciones. Este modelo se puede ver en $\mathcal{R}$ como: `modgambest3<- gam(Puntajetrain~ s(EDAD_PRES,8) +s(ASP_ANOTERMINACION,10)+COL_DEPARTAMENTO +COL_NATURALEZA+ASP_ESTADOCIVIL+ASP_SEXO+ ASP_ESTRATO+ASP_RAYA_TIPO,data = BaseF2)` Este modelo se puede escribir de mejor manera tomando x_i como las variables explicativas escritas en el orden del código de $\mathcal{R}$ $$\begin{cases} Y_k = \mu_k + \epsilon_k \\ \mu_k = \beta_0 + f_1(x_1) + f_2(x_2) + \beta_3 x_3 + \beta_4 x_4 \\ \qquad +\beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \end{cases}$$ En donde f_1 es un polinomio de grado 8 y f_2 es un polinomio de grado 10.	MSE
2	MSE El segundo mejor modelo es un modelo Lasso con todas las variables en el cual se incluyeron algunas interacciones, en este modelo la variable *ASP_ANOTERMINACION* se encuentra de manera cualitativa, mientras que *EDAD_PRES* es una variable que sigue siendo cuantitativa como en el primer modelo, los coeficientes del modelo se pueden ver en $\mathcal{R}$ de la siguiente forma: `lasso.coef3 <- predict(out, type = "coefficients", s = bestlam3)[1:164, ]` Este modelo se puede escribir de mejor manera tomando x_i como las variables explicativas escritas en el orden del código de $\mathcal{R}$	MSE

	$$\begin{cases} \tilde{Y}_k = \tilde{\mu}_k + \epsilon_k \\ \tilde{\mu}_k = \beta_1\tilde{x}_1 + \beta_2\tilde{x}_2 + \beta_3\tilde{x}_3 + \beta_4\tilde{x}_4 + \beta_5\tilde{x}_5 + \beta_6\tilde{x}_6 + \beta_7\tilde{x}_7 \\ \quad +\beta_8\tilde{x}_8 + \beta_9\tilde{x}_5 \cdot \tilde{x}_6 + \beta_{10}\tilde{x}_5 \cdot \tilde{x}_7 + \beta_{11}\tilde{x}_5 \cdot \tilde{x}_2 \\ \quad +\beta_{12}\tilde{x}_5 \cdot \tilde{x}_3 + \beta_{13}\tilde{x}_2 \cdot \tilde{x}_3 \\ \{\epsilon_k\} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$ En donde cada una de las interacciones corresponde a las siguientes: • *ASP_ESTRATO*∗*ASP_SEXO*, • *ASP_SEXO*∗*ASP_RAYA_TIPO*, • *ASP_SEXO*∗*COL_DEPARTAMENTO*, • *ASP_SEXO*∗*COL_NATURALEZA*, • *COL_DEPARTAMENTO*∗*COL_NATURALEZA* respectivamente	
3	El tercer mejor modelo de habilidad predictiva que se realizo fue un modelo lineal generalizado con la familia Gamma incluyendo todas las variables, sin interacciones y en donde la única variable cuantitativa esta dada por *EDAD_PRES*. Este modelo se puede visualizar de la siguiente forma: $$\begin{cases} \{Y_k\} \sim \Gamma(\mu_k, \phi) \\ g(\mu_k) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ \quad +\beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \end{cases}$$ En donde cada x_i es una de las variables que tenemos en la base de datos. Adicionalmente se sigue que la función g esta dada por: $g(\mu_k) = \frac{1}{\mu_k}$.	MSE

Modelo 1

- ¿Cómo fue construido ese modelo?

  Como en el modelo de regresión no paramétrico solamente se pueden incluir splines para las variables cuantitativas, lo primero que se realizo fue cambiar las variables *EDAD_PRES* Y *ASP_ANOTERMINACION* de variables cualitativas a variables cuantitativas que expresan la edad que tiene un individuo en el momento de la presentación del examen y el tiempo que ha pasado entre la presentación del examen y su graduación respectivamente.

  Ahora, se incluyeron todas las variables categóricas pues a lo largo de la modelación se mostró que la habilidad predictiva se veía disminuida cuando se quitaba alguna de las variables. Adicionalmente, como se quiere realizar un modelo predictivo se particionaron los datos en un $70\,\%$ de entrenamiento y un $30\,\%$ de testeo y se entreno un modelo inicial en donde los splines tenían grado 4 y 5 respectivamente, este modelo inicial se puede encontrar en el script de $\mathcal{R}$ en donde se puede ver que el MSE esta dado por 8466.006. A continuación para encontrar el mejor modelo que tuviera estos dos splines y que predijera de mejor manera a los datos se tomaron los datos

de entrenamiento y se entrenaron todos los posibles modelos con un primer spline de grado i para *EDAD_PRES* y un segundo spline de grado j para la variable *ASP_ANOTERMINACION*, se coloca como máximo grado a 10 pues es necesario evitar un sobreajuste, este proceso se realizó con el siguiente código:

```
for(i in 1:10){
  for(j in 1:10){
  modgam2<- gam(Puntajetrain~ s(EDAD_PRES,i)+s(ASP_ANOTERMINACION,j)
  +COL_DEPARTAMENTO+COL_NATURALEZA+ASP_ESTADOCIVIL+ASP_SEXO
  +ASP_ESTRATO+ASP_RAYA_TIPO,data = BaseF2)
  predics3 <- predict(modgam2, newdata = testeo.M)
  Errores3<-cbind(Errores3,
  mean((as.numeric(BaseF$PUN12_TOTAL[testeo])-predics3)^2))
  }
}
```

En donde se nos dice que la mejor combinación (i, j) para los grados de los spline en este modelo es de $(8, 10)$, por tanto se realiza la estimación del modelo mediante gam con estos grados y se procede a mirar las predicciones que este modelo toma para el testeo:

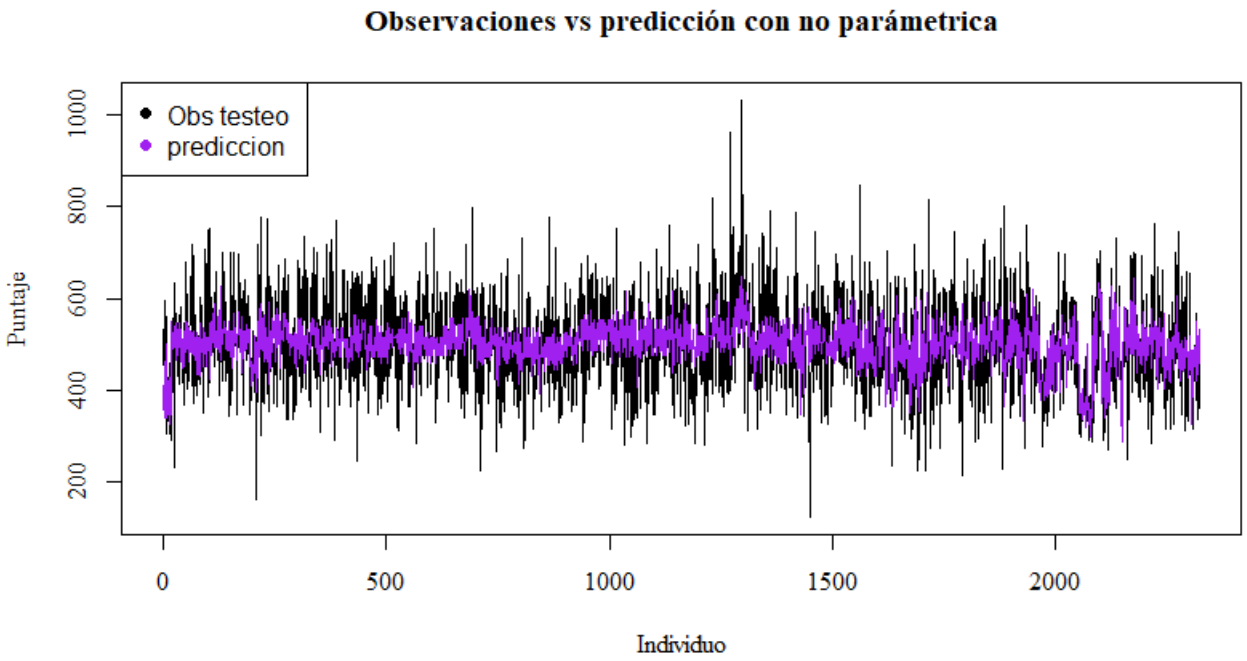
```
predicciones5 <- predict(modgambest3, newdata = testeo.M)
```

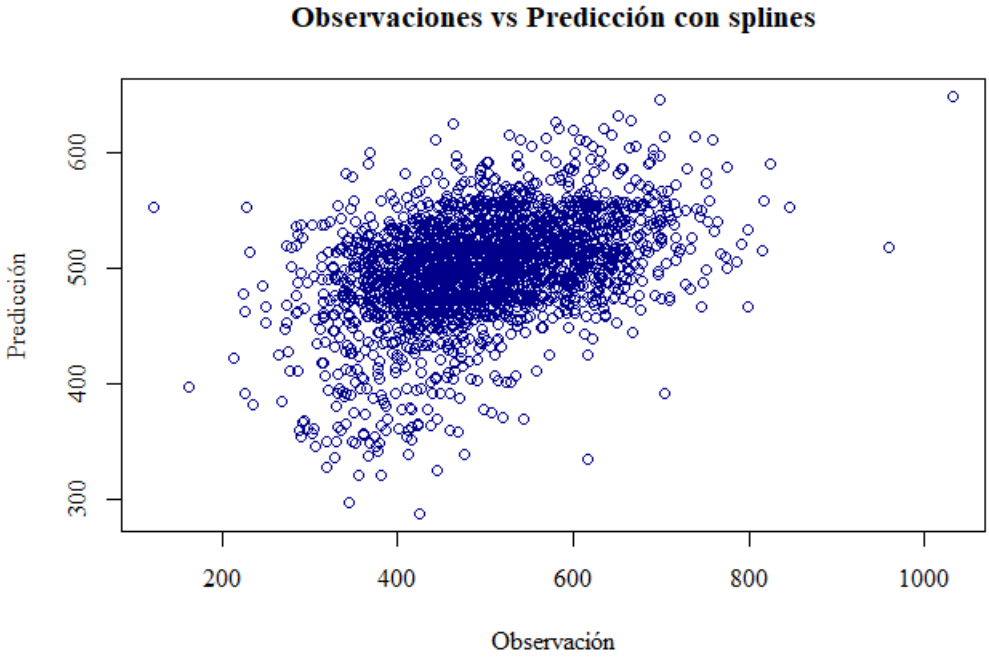
Con estas predicciones ya se puede calcular el MSE del nuevo modelo planteado:

```
val.errors16<- mean((as.numeric(BaseF$PUN12_TOTAL[testeo]) - predicciones5)^2)
val.errors16
```

El cual arroja un valor de 8416.138, este valor es el más pequeño que se obtuvo entre todos los modelos que se entrenaron y se probaron en el testeo, la mayoría de estos casi siempre tuvieron valores de este 8500 y 8600 por tanto se escogió este modelo como el que mejor habilidad predictiva obtuvo.

En las siguientes gráfica se pueden ver las observaciones de los individuos de testeo versus las predicciones que realiza nuestro mejor modelo:





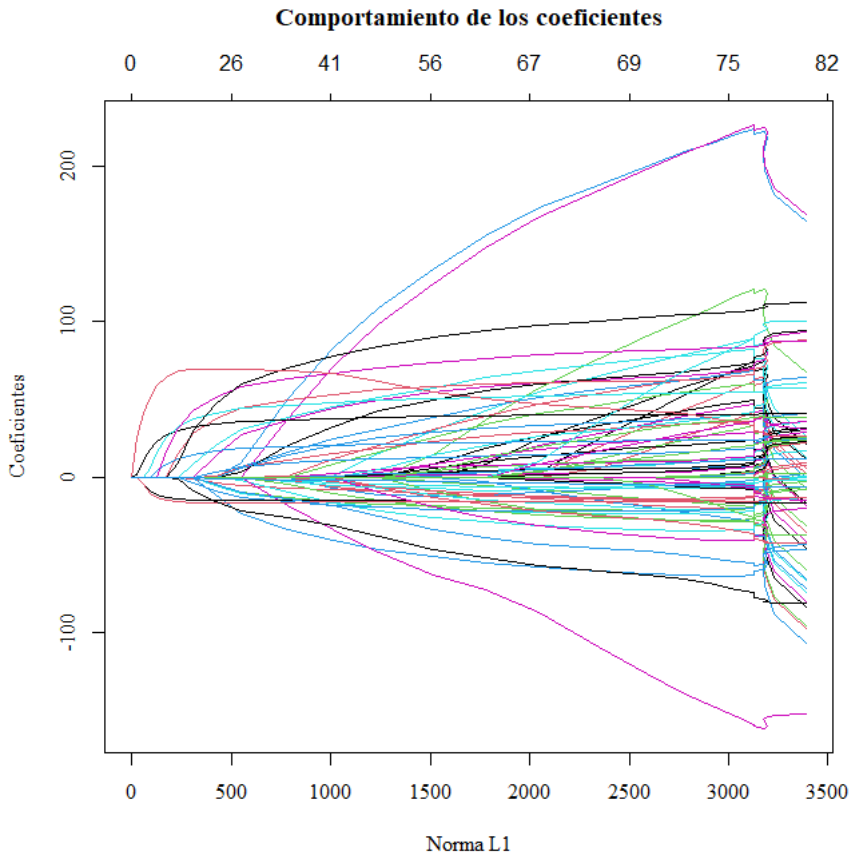
Después de todo este proceso, se pasa a estimar el modelo con todas las observaciones (entrenamiento y testeo) y se evalúa sobre los nuevos datos que pretendemos predecir.

Observación: Note que a pesar de que los grados de los splines son bastante grandes esto no necesariamente indica sobreajuste pues al evaluar estos grados con solo los datos de entrenamiento se tienen los suficientes datos de testeo por fuera de dicha estimación como para evitar que estos polinomios tiendan a dañar la habilidad predictiva de este modelo.

Modelo 2:

- ¿Cómo fue construido ese modelo?

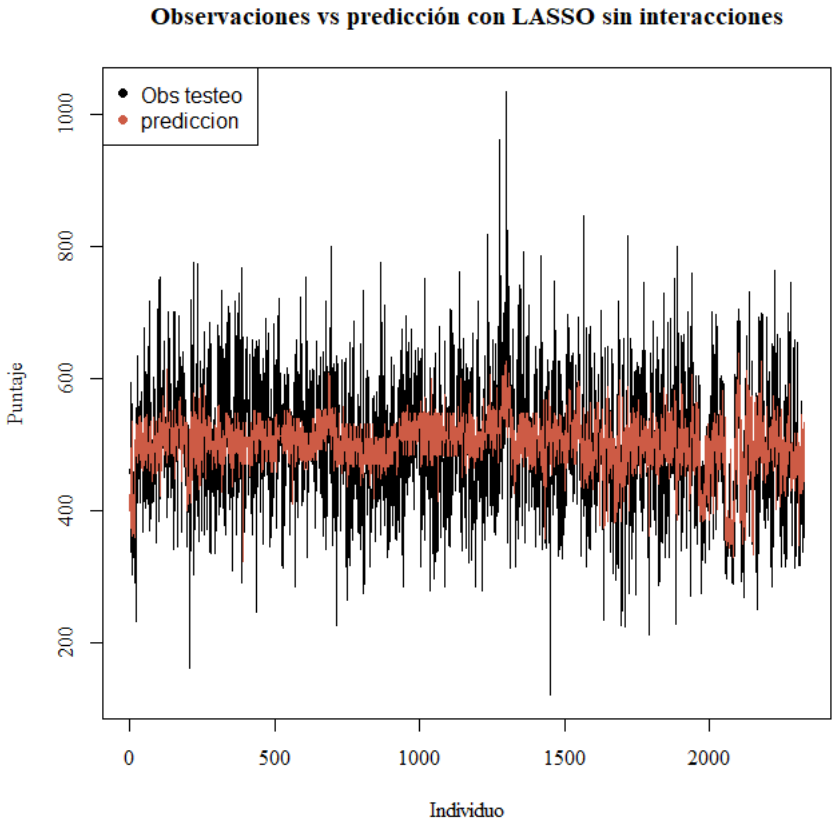
Inicialmente se tomo un modelo Lasso con todas las variables incluidas en donde la variable *EDAD_PRES* se toma como cuantitativa, el resto de variables como cualitativas y no se incluyeron interacciones, para estimarlo primero se calibro λ de la manera mencionada anteriormente, dando como resultado que el hiperparametro del modelo en este caso es 0.593166, adicionalmente se entreno el modelo con los datos de entrenamiento y con distintos lambdas en donde se obtuvo esta imagen que nos deja ver el comportamiento de los coeficientes mediante el crecimiento de la norma.



Se calculan las predicciones de este modelo inicial utilizando el código de $\mathcal{R}$ dado por:

```
lasso.pred <- predict(lasso.mod, s = bestlam,
                      newx = x[testeo, ])
```

Con las cuales se puede calcular el MSE, es decir, la habilidad de predicción de este modelo con respecto a los datos de testeo, lo cual de un resultado de 8538.874, en seguida se puede ver un gráfico de las predicciones del modelo en el testeo vs las verdaderas observaciones (puntajes) de los datos de testeo.



Sin embargo, como este error todavía no es satisfactorio y sabemos que el modelo Lasso hace una eliminación de variables entonces se toma como siguiente modelo un modelo Lasso con todas las posibles interacciones de orden 2.

En este modelo también se realiza la calibración y obtención del mejor lambda el cual en este caso es de $\lambda = 1.811443$, con la obtención de este se realiza el entrenamiento del modelo y posteriormente las predicciones para los datos de testeo:

```
lasso.pred2 <- predict(lasso.mod2, s = bestlam2,
                       newx = x[testeo, ])
```

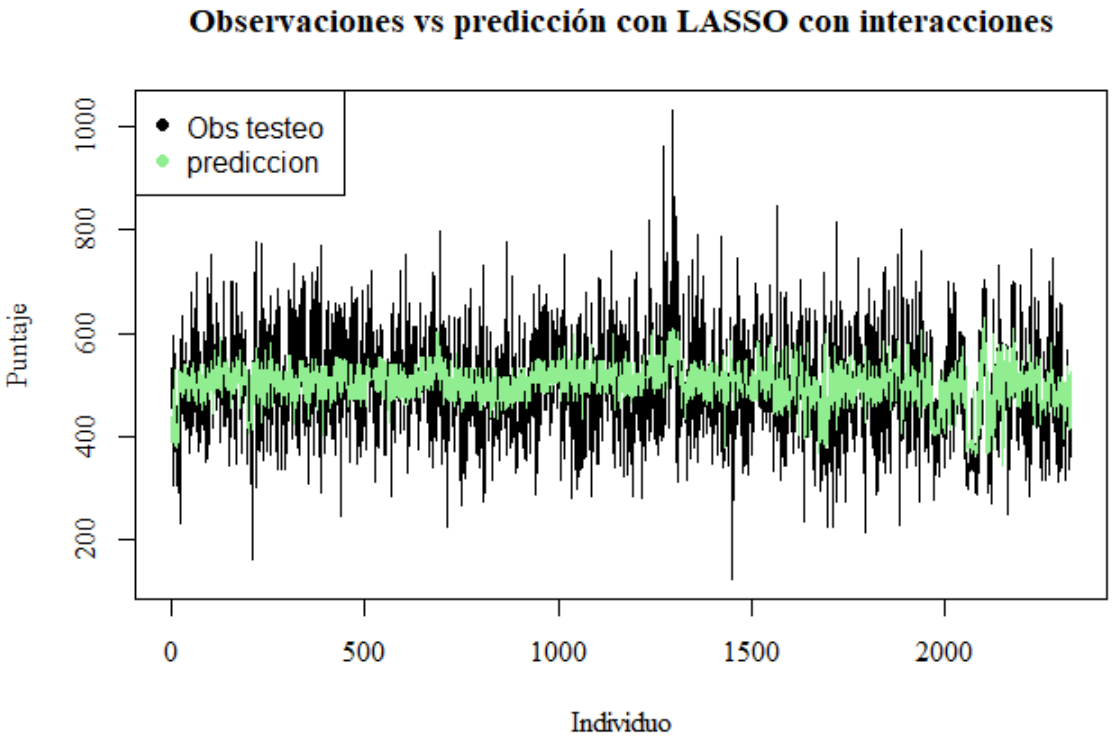
Con estas predicciones se obtiene un MSE de 8558.592, es decir, la habilidad predictiva disminuye, sin embargo, dado el contexto y sabiendo con anterioridad que el sexo puede influir ingresamos en el modelo algunas interacciones entre variables explicativas que podrian explicar de cierta manera la efectividad o no efectividad de la presentación del examen en un individuo de Colombia.

Se realiza un último modelo con las interacciones mencionadas en la anterior tabla, con estas interacciones agregadas se obtiene que el valor de lambda es $\lambda = 1.252842$, seguido a esto entrenamos el modelo y con ello se obtuvieron las nuevas predicciones para los datos de testeo que en $\mathcal{R}$ se hallaron con ayuda del siguiente código.

```
lasso.pred3 <- predict(lasso.mod3, s = bestlam3,
                       newx = x[testeo, ])
```

De donde se obtiene un MSE de $8529{,}745$ el cual mejora la habilidad predictiva de los anteriores dos modelos formados con LASSO, se puede ver el mejoramiento de la habilidad predictiva en

el siguiente gráfico:



Por tanto, este es el segundo mejor modelo con habilidad predictiva y con un método de modelamiento distinto que se obtuvo.

Modelo 3

- ¿Cómo fue construido ese modelo?

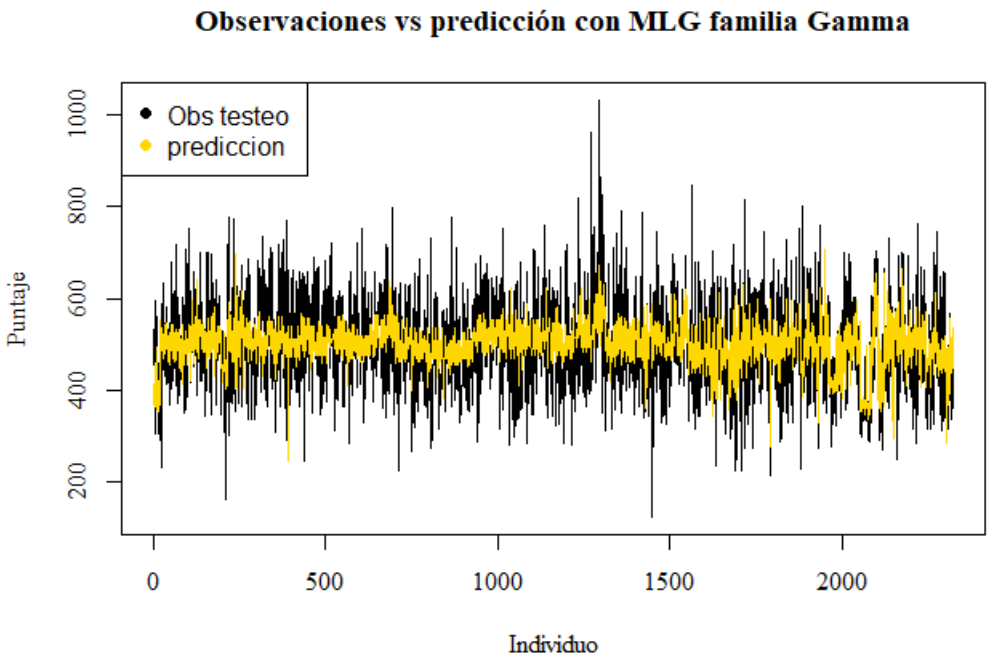
  Inicialmente se escoge un modelo lineal generalizado con familia gamma pues la variable respuesta es continua y por tanto si se coloca la familia normal se esta realizando el mismo modelo lineal múltiple, ahora, para el modelo inicial se colocan todas las variables solas sin ninguna interacción, luego como el MLG no tiene hiperparametros se puede pasar a entrenar el modelo directamente:

```
modelocompletoG<-glm(PUN12_TOTAL~1+ASP_ANOTERMINACION+COL_DEPARTAMENTO
+COL_NATURALEZA+ASP_ESTADOCIVIL+ASP_SEXO+ASP_ESTRATO+ASP_RAYA_TIPO
+EDAD_PRES,family=Gamma("inverse"),subset = train,data = BaseF)
```

  Ahora, las predicciones se calculan de la siguiente forma:

```
glm.probs2 <- predict(modelocompletoG, newdata=test.mat3, type = "response")
```

  De donde se obtiene que el modelo tiene un MSE de 8530.567, la comparación de los datos observados para el testeo y los datos de predicción se puede ver en la siguiente gráfico:



Como se buscaba el menor MSE posible para que la habilidad predictiva mejore entonces se intento realizar varios modelos con interacciones, sin embargo, estos modelos no aumentaban la capacidad predictiva y en cambio la empeoraban demasiado.

5. Variables Explicativas:

¿Qué variables aparecen en el modelo de mejor habilidad predictiva y cómo se interpretan sus respectivas influencias sobre la respuesta?

Reporte:

El summary del mejor modelo que se obtuvo con habilidad predictiva para el testeo, esta dado por:

```
Anova for Parametric Effects
                             Df   Sum Sq Mean Sq  F value    Pr(>F)
s(EDAD_PRES, 8)               1   253416  253416  30.3072 3.807e-08 ***
s(ASP_ANOTERMINACION, 10)     1  1020811 1020811 122.0837 < 2.2e-16 ***
COL_DEPARTAMENTO             32  3725338  116417  13.9228 < 2.2e-16 ***
COL_NATURALEZA                1  2003694 2003694 239.6313 < 2.2e-16 ***
ASP_ESTADOCIVIL               6    84799   14133   1.6902     0.119
ASP_SEXO                      1  3058533 3058533 365.7845 < 2.2e-16 ***
ASP_ESTRATO                   6  3102503  517084  61.8405 < 2.2e-16 ***
ASP_RAYA_TIPO                 8  1242337  155292  18.5721 < 2.2e-16 ***
Residuals                  7674 64166692    8362

Anova for Nonparametric Effects
                          Npar Df Npar F     Pr(F)
(Intercept)
s(EDAD_PRES, 8)                7 19.268 < 2.2e-16 ***
s(ASP_ANOTERMINACION, 10)      9 15.489 < 2.2e-16 ***
COL_DEPARTAMENTO
COL_NATURALEZA
ASP_ESTADOCIVIL
ASP_SEXO
ASP_ESTRATO
ASP_RAYA_TIPO
```

Las variables que tenemos en el modelo de mejor habilidad predictiva son: *EDAD_PRES*, *ASP_ANOTERMINACION*, *COL_DEPARTAMENTO*, *COL_NATURALEZA*, *ASP_ESTADOCIVIL*, *ASP_SEXO*, *ASP_ESTRATO*, *ASP_RAYA_TIPO*. Para ver como influyen sobre la respuesta, tendremos en cuenta la significancia de cada variable. Al ver el resumen del modelo, vemos que todas las variables, a excepción de *ASP_ESTADOCIVIL*, influyen bastante en el modelo que obtuvimos al tener p-valores demasiado cercanos a cero. *ASP_ESTADOCIVIL*, por al contrario, no aporta significativamente al modelo.

Por tanto, se concluye que el grueso de las variables exceptuando por *ASP_ESTADOCIVIL* tiene una incidencia en el puntaje obtenido por un individuo en el examen de la Universidad Nacional de Colombia, lo cual verifica el hecho de que este examen no es del todo equitativo pues las ciertas condiciones de un individuo pueden beneficiar o perjudicar su posible puntaje en el examen.

6. Aplicación:

¿Considerarían "justo" que la universidad decidiera usar ese mejor modelo obtenido por ustedes para calcular los puntajes de los futuros aspirantes, sin necesidad de aplicarles un examen? ¿Por qué sí o por qué no? Justifiquen su respuesta.

No consideramos justo utilizar nuestro modelo para obtener el puntaje de admisión de nuevos aspirantes. En nuestra base de datos vimos que no había muestras representativas de algunas categorías, por ejemplo, de algunos departamentos o años de terminación escolar. Esto puede generar sesgos en nuestras estimaciones. Además, viendo nuestro mejor modelo, las variables cuantitativas están teniendo más influencia que las categóricas, por ejemplo, se asume que, a mayor edad, menos puntaje. Por lo tanto, teniendo en cuenta los datos con los que se construyó y se probó el modelo, podemos afirmas que sea justo utilizarlo para predecir puntajes.