Modulo 5: Técnicas Avanzadas de Predicción

Modelo lineal Gaussiano. Elementos Básicos

Leandro Gutierrez

30/09/2024

Descripción de la tarea

Dentro del paquete de R "car" se encuentra una base de datos de salarios de profesorado de universidad con las siguientes variables:

- rank - 3 niveles de cargos de profesor.
- discipline - tipo de enseñanza que imparte.
- yrs.since.phd - años desde el doctorado.
- yrs.service - años de servicio.
- sex – género.
- salary - salario en dólares.

1. Propón la regresión para explicar el salario a través de los años de servicio y los años desde el doctorado. Justifica si era lo esperado o no y si difiere justificar la razón de dicho diferimiento. Obtén la suma de residuos al cuadrado, el coeficiente de determinación y el coeficiente de determinación corregido del modelo.
2. Incluye el género en el modelo. Valora la nueva suma de residuos al cuadrado.
3. Justifica, a través del coeficiente de determinación corregido, si el género es una variable a tener en cuenta para mejorar el modelo de predicción del salario.
4. Indica cómo incrementa el salario ante una variación en los años de servicio.
5. Indica cómo afecta a las betas del modelo si dividimos el salario por mil para expresarlo en miles.
6. Con el modelo anterior, teniendo en cuenta años de servicio y años desde el doctorado, realiza el mismo modelo, pero con el logaritmo neperiano del salario. Indica si se mantienen los signos de las betas obtenidas.
7. Indica cómo incrementa el salario ante una variación, en los años de servicio en este nuevo modelo.
8. Utilizando un modelo de regresión lineal (lm), realiza una modelización correcta del salario (utilizando las variables que desees de la base de datos) y presenta los resultados argumentando, desde tu conocimiento, las razones por las que eliges dicho modelo.

Solución

Carga de los datos

```
# cargamos el dataset salaries
data(Salaries)

# creamos un dataframe tibble auxiliar para trabajar
df <- as_tibble(Salaries)

# previsualizamos datos
pander(head(df))
```

rank	discipline	yrs.since.phd	yrs.service	sex	salary
Prof	B	19	18	Male	139.750
Prof	B	20	16	Male	173.200
AsstProf	B	4	3	Male	79.750
Prof	B	45	39	Male	115.000
Prof	B	40	41	Male	141.500
AssocProf	B	6	6	Male	97.000

```
pander(summary(df))
```

rank	discipline	yrs.since.phd	yrs.service	sex	salary
AsstProf : 67	A:181	Min. : 1.00	Min. : 0.00	Female: 39	Min. : 57800
AssocProf: 64	B:216	1st Qu.:12.00	1st Qu.: 7.00	Male :358	1st Qu.: 91000
Prof :266		Median :21.00	Median :16.00		Median :107300
		Mean :22.31	Mean :17.61		Mean :113706
		3rd Qu.:32.00	3rd Qu.:27.00		3rd Qu.:134185
		Max. :56.00	Max. :60.00		Max. :231545

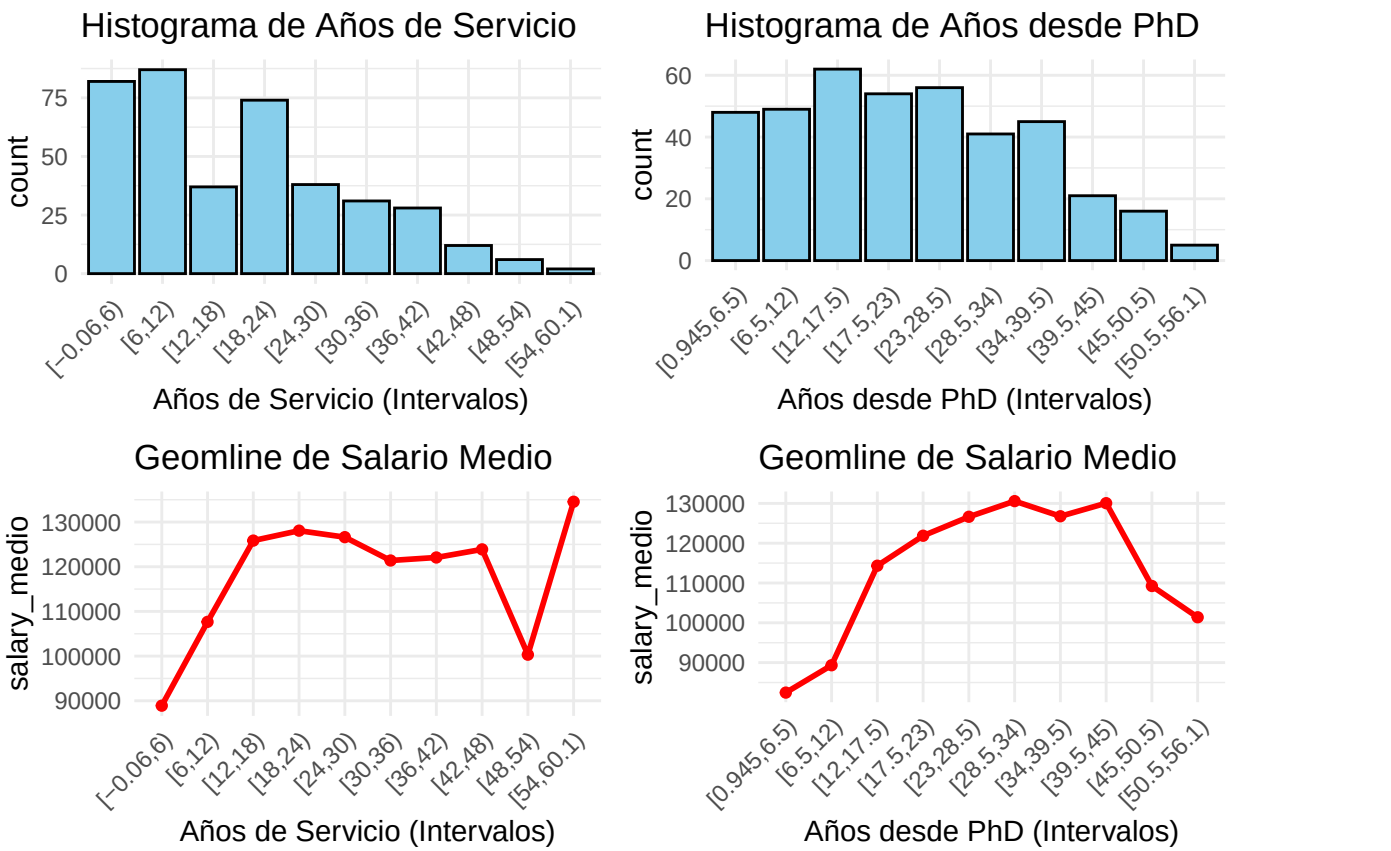
Podemos observar que contamos con un set de datos de **397 observaciones**, con **6 variables**, donde 3 de ellas son `factores` y 3 son numéricas de tipo `integer`. No se observan valores nulos.

Visualización de los datos

Visualizamos histograma y evolución de `salary` en función de las variables `yrs.service` y `yrs.since.phd`

```
# calculamos los bines para cada variable a analizar
df$yrs.service.bin <- cut(df$yrs.service, breaks = 10, right = FALSE)
df$yrs.since.phd.bin <- cut(df$yrs.since.phd, breaks = 10, right = FALSE)

# calculamos salario medio para cada bin
salary_by_service <- df %>%
  group_by(yrs.service.bin) %>%
  summarise(salary_medio = mean(salary), count = n())

# visualizamos los datos
# salary_by_service

# calculamos salario medio para cada bin
salary_by_phd <- df %>%
  group_by(yrs.since.phd.bin) %>%
  summarise(salary_medio = mean(salary), count = n())

# visualizamos los datos
# salary_by_phd

# plot histograma de años de servicio
plot1 <- ggplot(df, aes(x = yrs.service.bin)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Años de Servicio", x = "Años de Servicio (Intervalos)") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

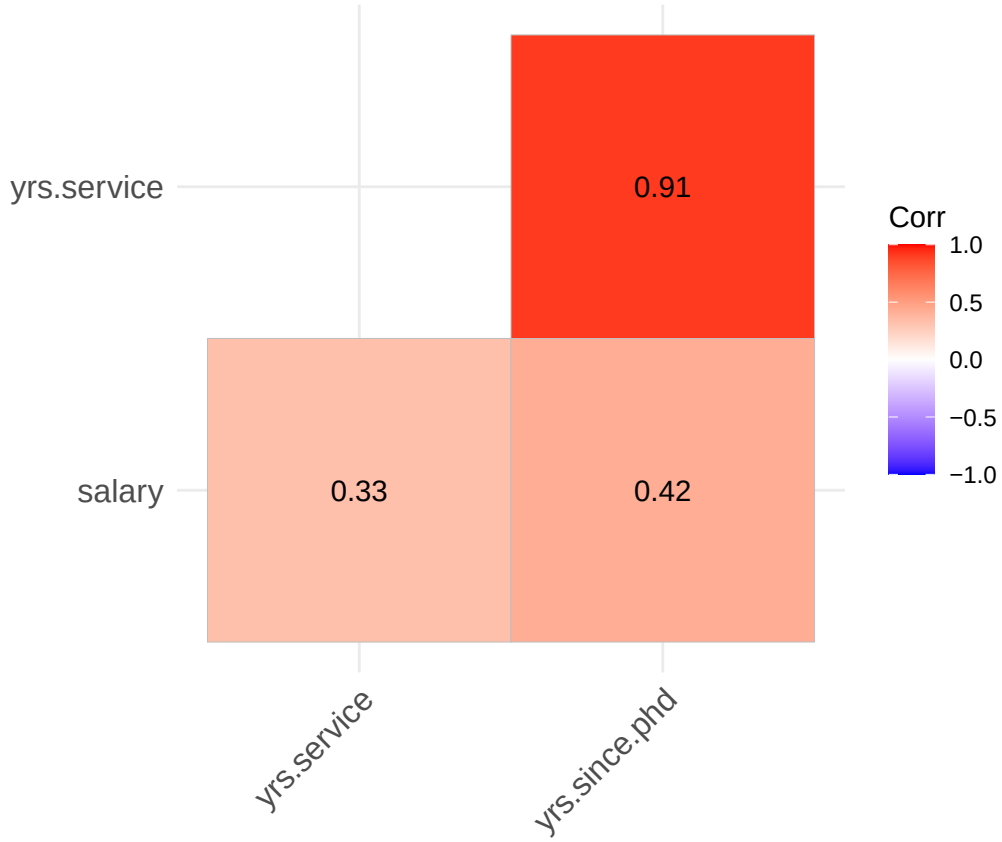
```r
# plot geomline de salario medio de cada bin
plot2 <- ggplot(df, aes(x = yrs.service.bin)) +
  geom_line(data = salary_by_service, aes(y = salary_medio), group = 1, color = "red", size = 1) +
  geom_point(data = salary_by_service, aes(y = salary_medio), group = 1, color = "red") +
  labs(title = "Geomline de Salario Medio", x = "Años de Servicio (Intervalos)") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot histograma de años dede phd + salario medio de cada bin
plot3 <- ggplot(df, aes(x = yrs.since.phd.bin)) +
  geom_bar(aes(y = ..count..), fill = "skyblue", color = "black") +
  labs(title = "Histograma de Años desde PhD", x = "Años desde PhD (Intervalos)") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# plot histograma de años dede phd + salario medio de cada bin
plot4 <- ggplot(df, aes(x = yrs.since.phd.bin)) +
  geom_line(data = salary_by_phd, aes(y = salary_medio), group = 1, color = "red", size = 1) +
  geom_point(data = salary_by_phd, aes(y = salary_medio), group = 1, color = "red") +
  labs(title = "Geomline de Salario Medio", x = "Años desde PhD (Intervalos)") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# armamos grilla con los plots
plot_grid(plot1, plot3, plot2, plot4, ncol = 2)
```



```
# df %>% filter(yrs.service.bin == '[48,54)')
# df %>% filter(yrs.service.bin == '[54,60.1)')
```

Ahora veremos la matriz de correlación para las variables que queremos estudiar

```
# seleccionamos solo columnas que nos interesan
tabla <- df %>% select(salary, yrs.service, yrs.since.phd)

# graficamos matriz de correlaciones
cr <- cor(tabla, use="complete.obs")
ggcorrplot(cr, hc.order = TRUE, type = "lower", lab = TRUE)
```



Podemos notar una gran correlación entre las variables `yrs.service` y `yrs.since.phd`, esto es indicador que deberíamos considerar dispensar de alguna de ellas. De todas formas y a fin de continuar con el trabajo práctico las mantendremos por el momento.

Apartado 1

En primer lugar definiremos el modelo propuesto para predecir el salario en función de los años de servicio y los años desde el doctorado:

$$\hat{salary} = \beta_0 + \beta_1 * yrs.service + \beta_2 * yrs.since.phd + \epsilon$$

```
# definimos la fórmula de nuestro modelo
formula <- as.formula('salary ~ yrs.service + yrs.since.phd')
formula
```

```
## salary ~ yrs.service + yrs.since.phd
```

```r
# creamos el modelo y visualiamos
modelo1 <- lm(formula=formula, data=df)
pander(summary(modelo1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	89.912	2.844	31,62	3,811e-110
yrs.service	-629,1	254,5	-2,472	0,01385
yrs.since.phd	1.563	256,8	6,086	2,754e-09

Table 4: Fitting linear model: formula

Observations	Residual Std. Error	R^2	Adjusted R^2
397	27.357	0,1883	0,1842

Ahora encontramos la suma de los cuadrados de los residuos de nuestro modelo 1

```r
# encontramos ssr
sum(residuals(modelo1)^2)
```

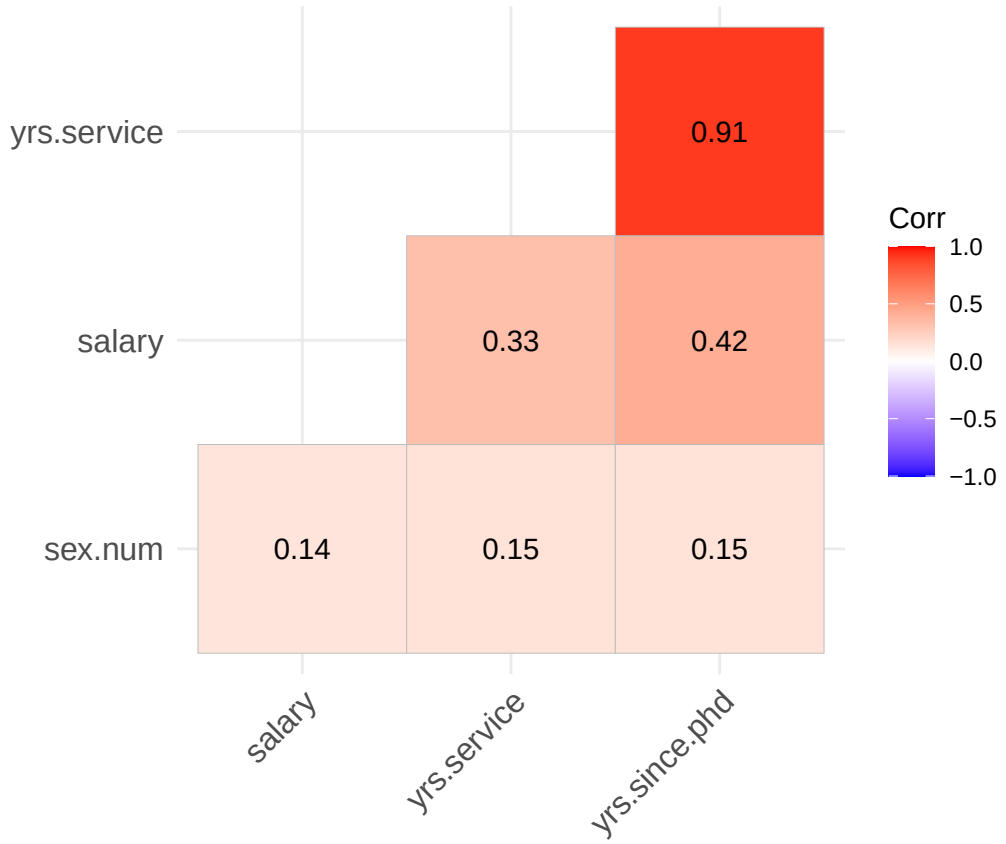
```
## [1] 294874679053
```

Para nuestro primer modelo obtuvimos un coeficiente de determinación (R^2) de **0,1883**, lo que indica que nuestras variables `yrs.service` y `yrs.since.phd` explican el **18,83%** de la variabilidad de nuestra variable dependiente `salary`. Además encontramos un coeficiente de determinación ajutado de **0,1842**. Y obtuvimos un `SSR` de **2,94e+11**.

Apartado 2

En primer lugar crearemos en nuestro dataset una variable auxiliar numérica `sex.num`, de tal manera que esta indique 1 para `Hombres` y 0 para `Mujeres`. Además visualizaremos la matriz de correlación con la nueva variable agregada.

```r
# hacemos una copia de nuestro dataset original con una nueva col
df <- df %>% mutate(sex.num = if_else(df$sex == "Male", 1, 0))

# seleccionamos solo columnas que nos interesan
tabla <- df %>% select(salary, yrs.service, yrs.since.phd, sex.num)

# graficamos matriz de correlaciones
cr1 <- cor(tabla, use="complete.obs")
ggcorrplot(cr1, hc.order = TRUE, type = "lower", lab = TRUE)
```



Definimos nuestro nuevo modelo:

$$\hat{salary} = \beta_0 + \beta_1 * yrs.service + \beta_2 * yrs.since.phd + \beta_3 * sex.num + \epsilon$$

```
formula2 <- as.formula('salary ~ yrs.service + yrs.since.phd + sex.num')
formula2
```

```
## salary ~ yrs.service + yrs.since.phd + sex.num
```

```
modelo2<-lm(formula=formula2, data=df)
pander(summary(modelo2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.876	4.801	17,26	4,194e-50
yrs.service	-649,8	254	-2,558	0,01089
yrs.since.phd	1.553	256,1	6,062	3,15e-09
sex.num	8.457	4.656	1,816	0,07008

Table 6: Fitting linear model: formula2

Observations	Residual Std. Error	R^2	Adjusted R^2
397	27.278	0,1951	0,189

Encontramos la suma de los cuadrados de los residuos de nuestro modelo 2

```r
# encontramos ssr
sum(residuals(modelo2)^2)
```

```
## [1] 2.9242e+11
```

El segundo modelo propuesto, donde incorporamos en género del individuo, nos entrega un coeficiente de determinación (R^2) de **0,1951**, el cual indica que las variables independientes del modelo explican al rededor de un **19,51%** de la variabilidad del salario, valor levemente mayor al primer modelo. También el coef. de determinación ajustado del nuevo modelo es levemente mayor al del modelo 1, con un valor de **0,189**. Para el segundo modelo el valor de `SSR` es levemente menor que para el modelo 1, con un valor de **2,92e+11**. A pesar de la mejor, el **p-value** (0,07008) de la nueva variable predicora `sex.num` es considerablemente bajo ($p < 0{,}05$), lo que nos indica que puede no ser significativo para nuestro análisis.

Apartado 3

En el **modelo 2** donde se incluyó la variable predictora `sex.num`, la cual es una representación numérica del sexo del individuo escrutado, obtuvimos un valor para el coeficiente de determinación un tanto mayor que para el **modelo 1**, esto nos dice que el segundo modelo explica levemente de mejor manera la variabilidad de la variable dependiente.

Apartado 4

Según los summary de nuestros dos modelos anteriores, la variable a predecir `salary` tiene una realción de proporcionalidad inversa contra la variable predictora `yrs.service`, para el modelo 1 el beta correspondiente a `yrs.service` toma un valor de **-629,1**, mientras que para el modelo 2 obtenemos un beta de **-649,8**. Ambos valores indican que ante cada año transcurrido en servicio el valor estimado del salario decae **$629,1** y **$649,8** respectivamente.

Ahora bien, sabemos por lo desarrollado en la sección de preparación de datos y visualización, que la relación entre la variable independiente `años de servicio (yrs.sevice)` y la variable dependiente `salario (salary)` no parece guardar una relación lineal a lo largo de todo el recorrido de la variable predictora `yrs.sevice`. Mas bien, parecen distinguirse 3 segmentos bien diferenciados: el primero de ellos, entre los 0 y 18 años de servicio, en el cual se percibe una relación lineal de pendiente positiva, donde a medida que aumenta la variable años de servicio aumenta en correspondencia la variable predicha salario. Un segundo segmento de pendiente aproximadamente nula, entre los 18 y 50 años de servicio, donde el salario parece mantenerse constante a pesar del aumento de años de servico. Y por último un tercer segmento entre los 50 y 60 años de servicio, donde la variable salario parece aproximarse a una función parabólica (x^2), donde se ve caer considerablemente el salario al rededor de los 55 años de servico, esto posiblemente debido a la falta de muestras en dicho segmento, podemos ver que solo contamos 8 lecturas para este último segmento (2 bines: [48,54) + [54,60.1)).

Apartado 5

Utilizaremos el modelo 2 como base para desarrollar este apartado

```r
# creamos un dataframe auxiliar
df <- df %>% mutate(salary.esc = salary/1000)

# definimos la formula con salario escalado
formula3 <- as.formula('salary.esc ~ yrs.service + yrs.since.phd + sex.num')

# definimos el modelo
modelo3 <- lm(formula=formula3, data=df)
pander(summary(modelo3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82,88	4,801	17,26	4,194e-50
yrs.service	-0,6498	0,254	-2,558	0,01089
yrs.since.phd	1,553	0,2561	6,062	3,15e-09
sex.num	8,457	4,656	1,816	0,07008

Table 8: Fitting linear model: formula3

Observations	Residual Std. Error	R^2	Adjusted R^2
397	27,28	0,1951	0,189

Vemos como reescalar nuestra variable dependiente `salary` afectó proporcionalmente a nuestras variables independientes, modificando en la misma escala los betas y las desviaciones standard asociados a ellas, lo mismo se percibe con el valor del β_0 el cual también sufre el reescalado.

Apartado 6

Utilizamos el modelo 1 como base para el desarrollo del siguiente apartado

```
# creamos un dataframe auxiliar
df <- df %>% mutate(salary.log = log(salary))

# definimos la formula con salario escalado
formula4 <- as.formula('salary.log ~ yrs.service + yrs.since.phd')

# definimos el modelo
modelo4<-lm(formula=formula4, data=df)
pander(summary(modelo4))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11,4	0,02435	468,2	0
yrs.service	-0,005316	0,002179	-2,439	0,01515
yrs.since.phd	0,01347	0,002199	6,125	2,196e-09

Table 10: Fitting linear model: formula4

Observations	Residual Std. Error	R^2	Adjusted R^2
397	0,2343	0,1932	0,1892

Podemos apreciar como al realizar una transformación de nuestra variable dependiente a través del logaritmo natural de su valor afecta a los betas de las variables independientes. En los tres casos (beta0, beta1 y beta2) los valores absolutos de los coeficientes se vieron reducidos, mientras que sus signos se mantuvieron idénticos.

Apartado 7

Cabe mencionar que es complejo comparar ambos modelos de manera directa, dado que la transforación recae sobre la variable dependiente `salary`, implicando esto que ahora nuestros betas nos informan sobre cuanto influyen cada una de las variables independientes en la **variación porcentual** de la variable a predecir y no

sobre el valor absoluto de la misma. Entonces por ejemplo, ahora podemos decir que por cada año de servicio vemos disminuir el salario en un 0,5%. De igual manera, ante la variación de un año desde el doctorado el salario incrementa en un 1% si las demás variables se mantuviesen constantes.

Apartado 8

Utilizaremos el paquete `earth` para determinar los parámetros adecuados de nuestro modelo de regresión y los segmentos de corte que mejor permitan explicar nuestra variable `salary`. Utilizaremos un valor thresh de **0.01** haciendo mas estricto nuestra selección de términos, haciendo que solo permanezcan en nuestro modelo las variables que al incluirlas al menos mejoren la explicabilidad en un 1%.

```
formula.all <- as.formula('salary ~ yrs.service + yrs.since.phd + discipline + sex.num')

modelo <- earth(formula = formula.all, data=df, thresh=0.01)

summary(modelo)
```

```
## Call: earth(formula=formula.all, data=df, thresh=0.01)
##
##                     coefficients
## (Intercept)           125913.287
## disciplineB            15148.414
## h(25-yrs.since.phd)    -2575.776
## h(yrs.since.phd-25)     -605.036
##
## Selected 4 of 4 terms, and 2 of 4 predictors
## Termination condition: RSq changed by less than 0.01 at 4 terms
## Importance: yrs.since.phd, disciplineB, yrs.service-unused, sex.num-unused
## Number of terms at each degree of interaction: 1 3 (additive model)
## GCV 598848227    RSS 229432784007    GRSq 0.3488958    RSq 0.3684768
```

Podemos ver como la función `earth` nos entregó el conjunto de 3 variables mas significativas para nuestro modelo lineal `yrs.since.phd.below.25`, `yrs.since.phd.above.25`, y `discipline`. Además, tal cual lo habíamos visto en la sección de análisis preliminar el alto valor de correlación entre las variables `yrs.since.phd` y `yrs.service` hacen que esta última carezca de significancia para nuestro modelo.

Ahora que earth nos ha facilitado una propuesta, llevamos los puntos de corte y variables sugeridas a nuestro modelo

```
# creamos las nuevas variables de acuerdo a lo que nos entregó `earth`
df <- df %>% mutate(yrs.since.phd.below.25 = if_else(df$yrs.since.phd < 25, yrs.since.phd, 0))

df <- df %>% mutate(yrs.since.phd.above.25 = if_else(df$yrs.since.phd >= 25, yrs.since.phd, 0))

# df %>% select(yrs.since.phd, yrs.since.phd.above.25, yrs.since.phd.below.25)

# definimos nuestra formula final
formula.final <- as.formula('salary ~ yrs.since.phd.below.25 + yrs.since.phd.above.25 + discipline')

# definimos el modelo final
modelo.final <- lm(formula=formula.final, data=df)

# visualizamos resumen del modelo
pander(summary(modelo.final))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	72.218	4.014	17,99	3,173e-53
yrs.since.phd.below.25	1.895	248,8	7,617	1,951e-13
yrs.since.phd.above.25	1.281	114,5	11,18	2,127e-25
disciplineB	15.199	2.702	5,625	3,519e-08

Table 12: Fitting linear model: formula.final

Observations	Residual Std. Error	R^2	Adjusted R^2
397	26.115	0,2623	0,2566

Como pudimos observar en el análisis visual preliminar la variable `yrs.sice.phd` no mantenía una relación completamente lineal con `salary` a lo largo de su recorrido. En su lugar pudimos apreciar al menos dos segmentos diferenciados si consideramos la pendiente de una posible recta lineal que exprese la relación entre las variables. Resultado al que también arriba el algortimo MARS implementado por la función `earth`, dividiendo la variable `yrs.since.phd` en dos, aquellos cuyo valor está por debajo de los 25 años y aquellos cuyo valor es igual o superior a dicha cantidad de años.

Habiendo tomado las sugerencias realizadas por `earth`, dividiendo la variable original `yrs.since.phd` en dos variables (`yrs.since.phd.below.25` y `yrs.since.phd.above.25`) y agregandolas a nuestro modelo junto a la variable `discipline`, cuyos valores posibles son `A` o `B`, obtuvimos nuestro modelo final con un coeficiente de determinación ajustado R_a^2 de `0,2566`, lo que nos dice que nuestro modelo es capaz de explicar al menos el **25,66%** de la variabilidad total de la variable dependiente `salary`. Mejorando en un **39,29%** ((26,23-18,83)/18,83) respecto al primer modelo considerado (modelo1) y un **34,44%** ((26.23-19.51)/19.51) respecto al segundo modelo (modelo2).