Taller 3: Seleccion de variables - Regresión de LASSO.

Andrés Felipe Palomino - David Stiven Rojas
Códigos: 1922297 - 1924615
Universidad del Valle

30 de abril de 2023

Universidad del Valle

Ejercicio 1: Los datos Hitters de la librería ISLR contienen información del salario y medidas de rendimiento de 322 jugadores de baseball profesionales en 1986 a 1988. El objetivo principal es determinar proponer un modelo predictivo para el salario de los jugadores.
Variables: **Salary(y)**, **AtBat**(x_1), **CAtBat**(x_2), **Hits**(x_3), **CHits**(x_4), **Runs**(x_5), **CRuns**(x_6), **HmRun**(x_7), **CmRun**(x_8), **RBI**(x_9), **CRBI**(x_{10}), **Errors**(x_{11}), **Assists**(x_{12}), **Walks**(x_{13}), **CWalks**(x_{14}), **Years**(x_{15}).
Nota: La base de datos contaba con datos faltantes, los cuales fueron omitidos, dado que esto es un ejercicio pedagógico y no tenemos forma de consultar el porqué ocurrió esta situación, aun así se pueden hacer métodos de imputación como KNN, por regresión, múltiple, etc. Los cuales no son el objetivo de este informe.

1) Ajuste un modelo para el salario en función de las demás variables (exprese claramente el modelo). Evalúe los supuestos. Si no se cumple alguno, haga transformaciones para corregirlo.
Modelo:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{5i}\beta_5 + x_{6i}\beta_6 + x_{7i}\beta_7 + x_{8i}\beta_8 + x_{9i}\beta_{10} + x_{10i}\beta_{11}$$

$$+x_{11i}\beta_{12} + x_{12i}\beta_{13} + x_{13i}\beta_{14} + x_{14i}\beta_{15} + x_{15i}\beta_{16} + \varepsilon_i$$

Se procederá a ajustar, evaluar los supuestos del modelo y en caso de ser necesario, realizar las respectivas transformaciones para su corrección.

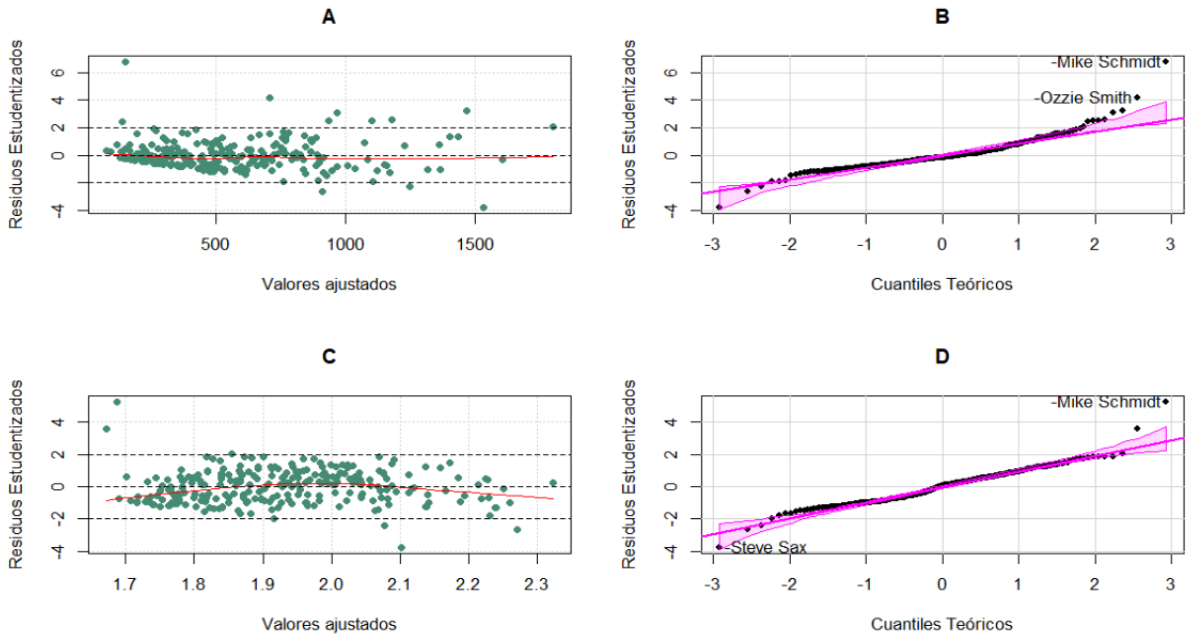

Figura 1: Validación de supuestos general para el modelo ordinario (A,B) y con transformación de Box-Cox (C,D)

Después de ajustar el modelo en general y realizar su validación de supuestos, antes de proceder a la interpretación de los coeficientes β_i encontramos el incumplimiento de estos mismos. En la Figura 1 en las sub figuras A y B se evidencia que el supuesto de homoscedasticidad y normalidad no se cumplen, afirmaciones corroboradas en la Tabla 1 dónde al realizar las pruebas de hipótesis formales rechazamos las hipótesis anteriormente mencionadas, cabe aclarar que utilizaremos una significancia de 0.05. Para solucionar esta problemática realizamos la transformación de Box-Cox dónde obtuvimos un $\lambda = 0{,}11$, valor que es muy cercano a 0 por lo cual realizamos la transformación monótona del logaritmo como se sugiere al aplicar este método y obtener resultados similares, en los supuestos para este segundo modelo en las sub figuras C y D se evidencia la corrección del problema de heteroscedasticidad y que aunque no se logró corregir por completo la normalidad, los valores se ajusten mejor a los anchos de confianza en D. Asumimos que la muestra es aleatoria y que no existe ningún tipo de correlación entre los individuos, es decir independencia.

	Shapiro Wilk(p-value)	Breusch Pagan(p-value)
Modelo	0.000000000001431	0.02494
Modelo Box-Cox	0.000003703	0.09959

Tabla 1: Resultados pruebas de hipótesis para la validación de supuestos

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6660	0.0344	48.48	0.0000
AtBat	-0.0007	0.0003	-2.59	0.0102
CAtBat	0.0000	0.0001	0.19	0.8518
Hits	0.0030	0.0010	3.05	0.0026
CHits	-0.0001	0.0003	-0.25	0.7993
Runs	-0.0007	0.0012	-0.54	0.5909
CRuns	0.0003	0.0003	1.14	0.2573
HmRun	0.0022	0.0026	0.87	0.3870
CHmRun	-0.0002	0.0007	-0.25	0.8030
RBI	-0.0002	0.0011	-0.16	0.8749
CRBI	0.0001	0.0003	0.41	0.6847
Errors	-0.0016	0.0018	-0.89	0.3757
Assists	0.0001	0.0001	1.04	0.2972
Walks	0.0026	0.0008	3.43	0.0007
CWalks	-0.0003	0.0001	-2.10	0.0365
Years	0.0096	0.0051	1.87	0.0624

Tabla 2: Modelo ajustado por el método de Box-Cox

Adicional a esto contamos con un valor de R^2 de 0.5331 es decir que aproximadamente 53.1 % de la variabilidad del salario de los jugadores está siendo explicada por este conjunto de covariables, ademas se cuenta con un R^2_{adj} de 0.5048. El valor del estadístico F es de 18.71 con un valor p asociado de aproximadamente 0 lo que indica que por lo menos una de estas estimaciones de los β_i es diferente de 0, particularmente vemos que solo 4 covariables en las pruebas individuales son significantes, esto nos permite considerar la presencia de multicolinealidad dado el gran conjunto de covariables y que por la naturaleza de las mismas presentan posibles correlaciones(esto lo evaluaremos al final del informe). También en la Tabla 2 observamos tanto relaciones positivas como negativas entre el salario y el conjunto de covariables si asumimos que se presentan aumentos con las demas covariables constantes. No hacemos enfásis en la interpretación individual de cada βi debido a que deberíamos hablar en términos del logaritmo del salario debido a la transformación realizada y esto podría causar confusión.

2) Realice un proceso de selección de variables. Proponga al menos dos modelos e interprete los resultados.

En el proceso de selección de variables se procede a ajustar todos los posibles modelos (32767), del cual se observa el R^2_{adj}, el AIC y el BIC. Además se realizan los algoritmos de selección (forward selection, backward selection, stepwise selection) para identificar si hay diferencias en modelos.

	n	predictors	adjr	aic	sbc
11831	7	AtBat Hits CRuns HmRun Walks CWalks Years	0.5160960	-316.73	-284.58
11896	7	AtBat Hits CRuns CRBI Walks CWalks Years	0.5157923	-316.57	-284.42
11886	7	AtBat Hits CRuns RBI Walks CWalks Years	0.5157425	-316.54	-284.39
11866	7	AtBat Hits CRuns CHmRun Walks CWalks Years	0.5155160	-316.42	-284.27
6027	6	AtBat Hits CRuns Walks CWalks Years	0.5154508	-317.35	-288.77

Tabla 3: Valores de los criterios de selección segun la selección de variables.

Ajustando todos los posibles modelos, y realizando la selección de los 5 mejores ajustes segun el R^2_{adj} y el AIC, se encontraron la misma selección de variables predictoras consignadas en la Tabla 3. Al realizar los algoritmos de selección se evidenció que los 3 algoritmos coinciden en la misma selección de variables consolidados en la Tabla 4, las cuales corresponden al modelo 6027 de la Tabla 3

Forward selection	Backward selection	Stepwise selection
CRuns	Full Model	CRuns
Hits	RBI	Hits
Years	CAtBat	Years
Walks	CHits	Walks
AtBat	CHmRun	AtBat
CWalks	Runs	CWalks
	CRBI	
	Errors	
	Assists	
	HmRun	

Tabla 4: Resumen de las variables seleccionadas según el algoritmo de selección

Con base en los resultados anteriores, y debido a que no se cuenta con una opinión del experto del tema, se seleccionaran las variables solamente con el análisis estadístico realizado, las cuales corresponden al modelo que presenta el menor AIC (6 variables) y el modelo que presenta el mejor R^2_{adj} (7 variables) los cuales son los siguientes:

$$ModeloAIC : log(y_i) = \beta_0 + x_{6i}\beta_1 + x_{3i}\beta_2 + x_{15i}\beta_3 + x_{13i}\beta_4 + x_{1i}\beta_5 + x_{14i}\beta_6 + \epsilon_i$$

$$ModeloR2 : log(y_i) = \beta_0 + x_{6i}\beta_1 + x_{3i}\beta_2 + x_{15i}\beta_3 + x_{13i}\beta_4 + x_{1i}\beta_5 + x_{14i}\beta_6 + x_{7i}\beta_7 + \epsilon_i$$

	ModeloAIC			ModeloR2		
	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
(Intercept)	1.6501	0.0320	0.0000	1.6530	0.0320	0.0000
CRuns	0.0003	0.0001	0.0008	0.0003	0.0001	0.0008
Hits	0.0027	0.0007	0.0001	0.0027	0.0007	0.0001
Years	0.0111	0.0039	0.0050	0.0112	0.0039	0.0045
Walks	0.0023	0.0006	0.0003	0.0023	0.0006	0.0004
AtBat	-0.0006	0.0002	0.0066	-0.0006	0.0002	0.0042
CWalks	-0.0002	0.0001	0.0307	-0.0002	0.0001	0.0262
HmRun				0.0013	0.0011	0.2479

Tabla 5: Estimación de los modelos seleccionados

En la Tabla 4 se presentan las estimaciones de los coeficientes, además se cuenta con un R^2_{adj} de 0.5155 para el modeloAIC y de 0.5161 para el modeloR2. Un estadistico F de 47.45 para el modeloAIC y junto con un estadistico F de 40.92 para el modeloR2, ambos con un valor p asociado de aproximadamente 0 lo que indica que por lo menos una de estas estimaciones de los β_i es diferente de 0. Dada la transformación en variable "Y", se interpreta las variables en términos de relaciones, para el modeloAIC se tiene relaciones positivas en el salario de los jugadores y el conjunto de las variables, si asumimos que se presentan aumentos con las demás variables constantes, excepto en CWalks y AtBat que presentan relaciones negativas. Para el modeloR2 se mantienen las mismas relaciones entre las el salario y las covariables que se presentan en el otro modelo, solo que a este se le incluye la variable HmRun que presenta una relación positiva. En términos de pruebas individuales (t) se observa que para el modelo AIC todos los coeficientes son significantes, mientras que para el otro modelo propuesto la variable HmRun no es significante, es decir que el coeficiente asociado a la variable HmRun es 0 si ya tengo incluidas las demás covariables. Como conclusión, el mejor modelo a plantear sería el modeloAIC si solo tomamos en cuenta el análisis estadístico en cuestión de coeficientes significativos. Reiteramos al final realizaremos la respectiva descripción general de todos los hallazgos encontrados.

3) Realice un proceso de selección utilizando la regresión de LASSO.

Primeramente, se realizará la estimación de lambda por medio de CV.

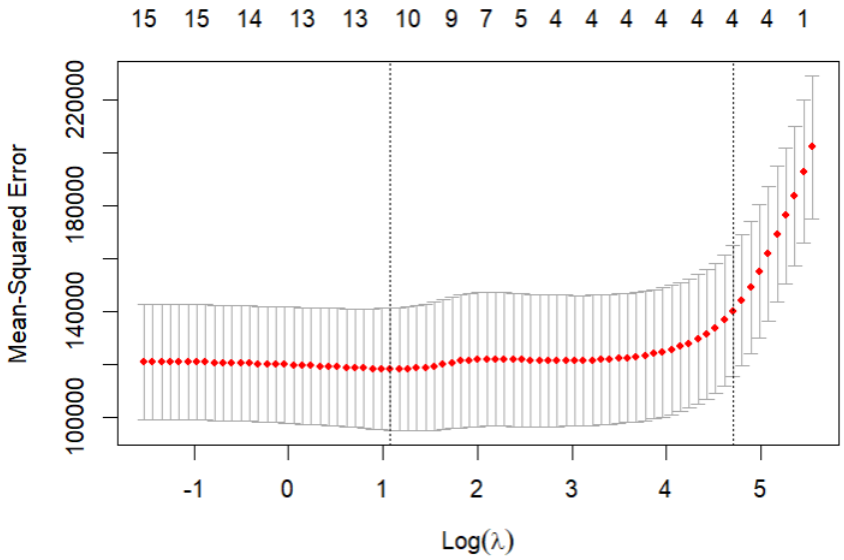

Figura 2: Validación cruzada para diferentes valores de λ

Las covariables seleccionadas al usar el estimador de LASSO con el λ óptimo(0.03462219) (regla de una desviación estandar) son: Hits(x_3), CHits(x_4), CRuns(x_6), CRBI(x_{10}), Walks(x_{13}).

Con base en la regresión de LASSO se estima el siguiente modelo:

$$log(y_i) = \beta_0 + x_{3i}\beta_1 + x_{4i}\beta_2 + x_{6i}\beta_3 + x_{10i}\beta_4 + x_{13i}\beta_5$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6650	0.0222	74.95	0.0000
Hits	0.0011	0.0002	4.87	0.0000
CRuns	0.0002	0.0001	2.56	0.0111
CRBI	0.0001	0.0001	1.33	0.1834
Walks	0.0009	0.0005	1.93	0.0548

Tabla 6: Resumen del modelo seleccionado por LASSO

En la Tabla 6 se presentan las estimaciones de los coeficientes, además se cuenta con un R^2 de 0.4913 para el ajustado por regresión de LASSO. Un estadístico F de 62.3 con un valor p asociado de aproximadamente 0 lo que indica que por lo menos una de estas estimaciones de los β_i es diferente de 0. En general podemos observar que todas las variables presentan la relación lineal positiva con el salario. Hits, CRuns, Walks presentan un aporte significativo dentro del modelo.

4) Compare los modelos propuestos con el modelo completo. ¿Cuáles características de los jugadores tienen mayor influencia sobre el salario?

En general podemos observar que las variables que están presentes independientes del método utilizado son CRuns, Hits y Walks, por lo cual tenemos un indicio sólido para considerar un aporte significativo general de ellas dentro del modelo. A su vez, los modelos ajustados presentan valores en sus pruebas de hipótesis individuales para β_i que nos permiten ver cuáles tienen aportes significativos, presentan R^2 parecidos, lo que nos describe que del modelo completo hay muchas variables con aportes irrelevantes que solo nos aportan inflación de varianza y ruido que no son compensados por su significancia estadística ni por explicación de variabilidad. Aun así, debemos ser conscientes de que todas estas metodologías utilizando algoritmos de selección o procesos de selección a través de procesos de optimización no deben ser nuestro único recurso inmediato para la generación de un modelo, sino que siempre debemos considerar la opinión de expertos en el tema, la complejidad de sus mediciones y en general ser conscientes de los procesos de validez externa e interna del modelo, también estos criterios no tiene en consideración la multicolinealidad dentro de las variables así que sería óptimo considerar métodos no solo de selección sino también de reducción de varianza si es detectada esta problemática, como lo puede ser la regresión por PLS o Elastic Net, donde claramente independiente del modelo a utilizar debemos considerar siempre toda la validación de supuestos necesarias para su validez. A continuación ilustraremos las tablas de los VIF's asociados a los modelos, dónde evidenciamos problemas de multicolinealidad altos, también ilustramos en general un resumen de las variables seleccionadas para cada método.

Modelo	CRuns	Hits	Years	Walks	AtBat	CWalks	HmRun	CRBI
AIC	14.15	14.51	5.47	2.97	15.92	12.99	1.51	
R2	14.15	14.51	5.47	2.96	15.92	12.95		
LASSO	9.62	1.53	1.62					9.47

Tabla 7: VIF de los modelos ajustados.