Taller 2: Multicolinealidad, regresión ridge y por componentes principales.

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

14 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 los factores que afectan el salario mediante las siguientes covariables: **Salary(y)**, **Hits**(x_1), **CHits**(x_2), **Runs**(x_3), **Cruns**(x_4), **HmRun**(x_5), **CmRun**(x_6), **RBI**(x_7), **CRBI**(x_8).
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:

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

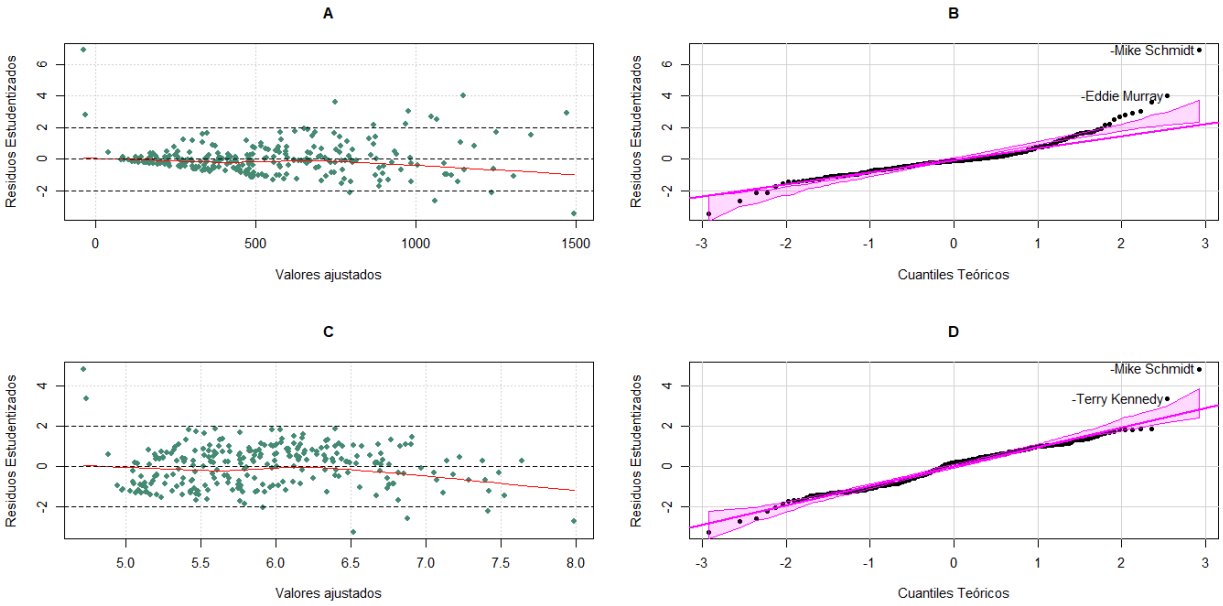

Figura 1: Validación de supuestos general para el modelo ordinario y con transformación de Box-Cox

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 al realizar las pruebas de hipótesis formales, cabe aclarar que utilizaremos una significancia de 0.05. Para solucionar esta problemática realizamos la transformación de Box-Cox dónde obtuvimos un $\lambda = 0{,}12$, 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.

Modelo - Prueba de hipotésis	Shapiro Wilk(p-value)	Breusch Pagan(p-value)
Modelo	0.00000000000005608	0.0459
Modelo Box-Cox	0.00002591	0.4471

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

β_i	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.7095	0.1062	44.35	0.0000
Hits	0.0036	0.0032	1.13	0.2594
CHits	0.0005	0.0008	0.65	0.5175
Runs	0.0044	0.0053	0.83	0.4064
CRuns	0.0002	0.0012	0.14	0.8926
HmRun	0.0016	0.0119	0.13	0.8936
CHmRun	-0.0003	0.0031	-0.10	0.9177
RBI	0.0007	0.0051	0.13	0.8934
CRBI	0.0004	0.0013	0.28	0.7820

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

Adicional a esto contamos con un valor de R^2 de 0.6458 es decir que aproximadamente 65 % de la variabilidad del salario de los jugadores está siendo explicada por este conjunto de covariables. El valor del estadístico F es de 30.34 con un valor p asociado de aproximadamente 0 por lo cual por lo menos una de estas estimaciones de los β_i es diferente de 0, pero particularmente vemos que ninguna prueba individual nos permite evidenciar relaciones significativas, esto nos permite afirmar la presencia de multicolinealidad. También en la Tabla 2 observamos en general relaciones positivas entre el salario y el conjunto de covariables si asumimos que se presentan aumentos con las demas covariables constantes, es decir esperamos que este aumente cuando estan tengan aumentos en sus unidades de medida a excepción de CHmRun , que presenta una relación negativa. 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) Evalúe si hay problemas de multicolinealidad usando los indicadores vistos en clase.

	Hits	CHits	Runs	CRuns	HmRun	CHmRun	RBI	CRBI
VIF	12.89	172.59	11.61	94.45	6.82	40.94	10.76	116.24

Tabla 3: VIF de las covariables del modelo ajustado por el método de Box-Cox

Nota: Si se realizan transformaciones en potencia a la variable de respuesta, estas no interfieren en las covariables por lo cual los indicadores de multicolinealidad no deberían cambiar.

Trabajaremos con el modelo ajustado por el método de Box-Cox puesto que fue el que al realizarlo tuvimos mejores resultados en cuestión de supuestos. En la Tabla 3 podemos evidenciar que existen problemas de multicolinealidad como lo sospechabamos anteriormente en el modelo con valores bastante elevados todos (por encima de 10) a excepción de la variable HmRun que cuenta con un coeficiente cercano a 7. A su vez el indice de condición $n_j = 1726{,}867$, lo que reafirma la presencia de multicolinealidad por ser mayor a 100.

3) Utilice la regresión de ridge para corregir los problemas de multicolinealidad. Considere varios valores de k y seleccione un valor óptimo.

Antes de realizar este método junto con el de la regresión a partir de componentes principales es pertinente aclarar que ambos serán efectuados en la transformación

obtenida a partir del método de Box-Cox puesto que esta fue la que nos presentó el mejor desempeño en cuestión de supuestos del modelo, por lo cual al realizarlos debemos contemplar esta situación y no hacer caso omiso a ella y pensar solo en la regresión a partir del salario sin transformaciones. Se procede a seleccionar un K óptimo por medio de validación cruzada y validación cruzada generalizada. Se compararán modelos y se escogerá el mas óptimo.

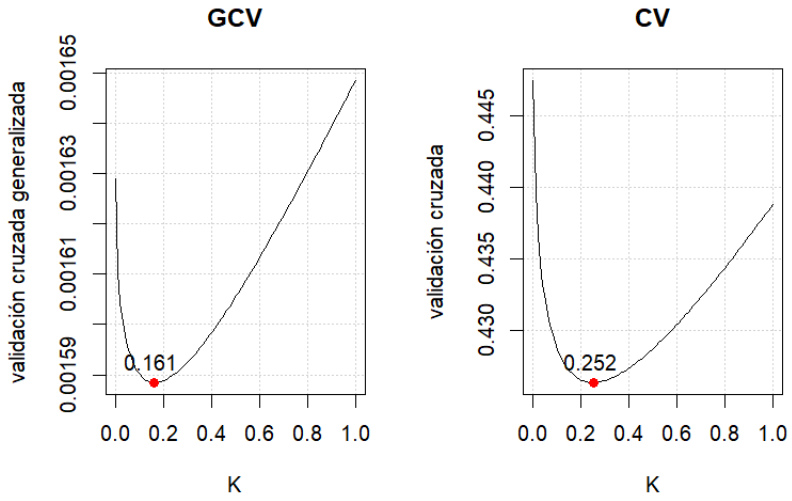

Figura 2: Validación cruzada generalizada (GCV) y Validación cruzada (CV)

Después de evaluar distintos valores de K, se encontró en la Figura 2 que los valores mínimos para la GCV Y CV son 0.161 y 0.252 respectivamente. el cual se escogerá el modelo mas óptimo con base en aquel que posea mejor R^2 y mejor criterio de información AIC.

Ridge	Hits	CHits	Runs	CRuns	HmRun	CHmRun	RBI	CRBI
VIF (K=0.161)	0.890	0.486	1.007	0.544	0.990	0.998	1.038	0.412
VIF (K=0.252)	0.558	0.315	0.628	0.320	0.674	0.625	0.617	0.226

Tabla 3: VIF de las covariables del modelo estimado por RIDGE

En la Tabla 3 se evidencia que efectivamente se corrigieron los problemas de multicolinealidad, ya que estos son bastante inferiores a los ilustrados en la Tabla 3. A su vez, el modelo con menor criterio de información AIC fue el generado por el K de GCV con un valor de -229.28.

β_i	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.7810	401.7560	-10.0465	0.0000
Hits	0.0032	0.6065	3.7961	0.0002
CHits	0.0003	0.4481	6.6007	0.0000
Runs	0.0039	0.6453	2.4822	0.0137
CRuns	0.0005	0.4744	5.2440	0.0000
HmRun	-0.0005	0.6396	-0.1182	0.9060
CHmRun	0.0002	0.6423	0.4325	0.6657
RBI	0.0019	0.6550	1.2150	0.2255
CRBI	0.0004	0.4128	4.4867	0.0000

Tabla 4: Resumen modelo estimado por Ridge

En la Tabla 4 observamos que se mantuvieron las relaciones entre las covariables y el salario entre el modelo por MCO y por Ridge, excepto en las covariables HmRun y CHmRun, las cuales se invirtieron en signo las estimaciones de las covariables, es decir presentaron una relación inversa a la que tenían en el modelo inicial. Adicional a esto contamos con un R^2 de 0.4435.

4) Utilice la regresión por componentes principales para corregir el problema de multicolinealidad.

Para realizar la regresión por componentes principales (PCR) vamos a realizar un screeplot para entender el porcentaje de variabilidad explicada acumulada por los componentes principales.

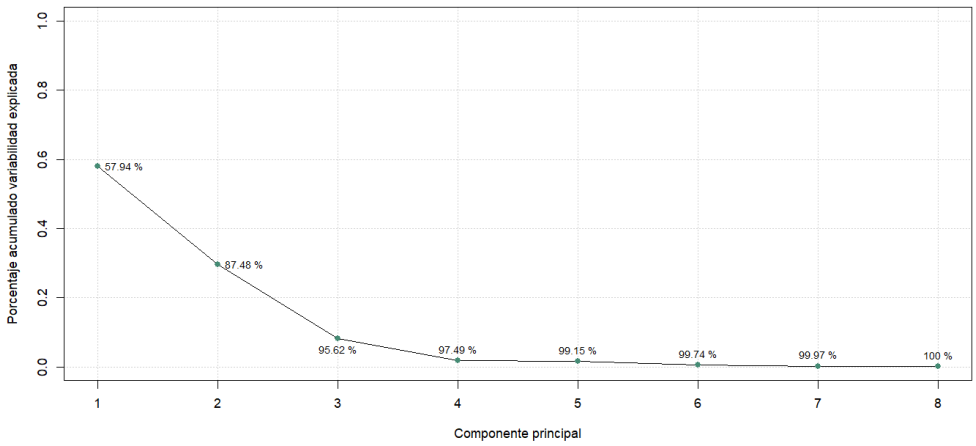

Figura 3: Screeplot

En la Figura 3 observamos que el porcentaje de inercia explicada por los 5 primeros componentes principales es de 99.15 % dónde decidimos evaluar el punto de corte para descartar los valores propios asignados a componentes principales más pequeños, puesto que la variabilidad restante explicada por los tres últimos era muy baja. A su vez, en la Tabla 5 podemos observar que estos 3 últimos valores eran pequeños.

	1	2	3	4	5	6	7	8
Valores propios	4.64	2.36	0.65	0.15	0.13	0.05	0.02	0.00

Tabla 5: Valores propios

Procedemos a realizar la regresión con 5 componentes principales.

Tabla 6: Resumen del modelo por componentes principales

En la Tabla 6 observamos las estimaciones utilizando el método de PCR, donde podemos evidenciar en primer lugar que la relación entre las covariables y el salario, es positiva, a excepción de HmRun y CHmRun que presenta una relación negativa. Las variables HmRun, CHmRun, RBI y Runs presentan valores p asociados al valor

	Estimate	Std.Error	t value	Pr(> |t|)
Hits	0.17	0.05	3.31	0.00
CHits	0.25	0.05	4.62	0.00
Runs	0.12	0.08	1.49	0.14
CRuns	0.22	0.05	4.67	0.00
HmRun	-0.01	0.07	-0.19	0.85
CHmRun	-0.04	0.09	-0.47	0.64
RBI	0.07	0.08	0.93	0.35
CRBI	0.12	0.03	3.96	0.00

del estadístico t por encima de la significancia estipulada, lo que nos describe que con respecto a pruebas individuales, que cada una de ellas no tiene un aporte significativo dentro del modelo si ya hemos incluido las demás. Por último contamos con un R^2 de 0.488.

4) Compare los modelos por mínimos cuadrados ordinarios, regresión de ridge, y regresión por componentes principales. ¿Cuáles son las covariables que tienen un efecto significativo sobre el salario del jugador?

Observamos que en general los cálculos de los R^2 de los modelos presentan diferencias, puesto que en el modelo utilizando el método de Box-Cox contábamos con uno de aproximadamente de 0.65 y a través del uso de los dos métodos para corregir el problema de multicolinealidad encontramos una disminución significativa con valores de 0.4435 y 0.4888 respectivamente. Aun así, esto genera un proceso de compensación al poder ver qué variables dentro del modelo, cuando ya hemos incluido las otras, presentan un aporte significativo para entender el salario del jugador, bajo la transformación, claramente, la corrección de esta problemática nos brinda una mejor contextualización debido a que, las diferentes variables Bajo esta premisa, las variables con la regresión ridge que no presentan un aporte significativo dentro del modelo son HmRun, CHmRun y RBI, covariables que repiten la misma condición en el método de componentes principales, pero añadiendo Runs. También es pertinente aclarar que independiente del método utilizado para la corrección del problema de multicolinealidad, la evaluación de supuestos