# Trabajo Final de Econometría para Datos Masivos de la Universitat de les Illes Balears

Brian Abad Guadalupe

Universitat de les Illes Balears

Centre d'Estudis de Postgrau

MADM - Màster Universitari en Anàlisi de Dades Massives en Economia i Empresa

Mallorca

2024

**Tabla de contenido**

## Resumen

Este trabajo analiza el conjunto de datos Hitters, que contiene estadísticas de jugadores de béisbol y sus salarios, para identificar las variables más relevantes en la predicción de los ingresos y comparar diferentes métodos de modelado.

## Introducción

Este trabajo académico se centra en el análisis y modelado de datos utilizando el conjunto de datos Hitters, que contiene información relacionada con estadísticas de jugadores de béisbol profesional y sus salarios. El objetivo principal del estudio es identificar las variables más relevantes para predecir los salarios de los jugadores, así como evaluar y comparar diferentes enfoques de modelado para optimizar la precisión de las predicciones. Para ello, se implementan técnicas de análisis exploratorio, selección de variables y ajuste de modelos utilizando métodos avanzados, como la regresión lineal múltiple, la selección de subconjuntos, la regresión regularizada (LASSO y Ridge), y métodos basados en componentes principales (PCA y PLS). Todo ello mediante el uso de herramientas del entorno R y bibliotecas como caret, glmnet, leaps, o corrplot.

Este trabajo no solo tiene como finalidad el análisis técnico, sino también la extracción de conclusiones útiles que demuestren la aplicabilidad práctica de los métodos empleados en el contexto deportivo y económico. Además, se proporciona una comparación entre los enfoques tradicionales y modernos para mejorar la comprensión del impacto de las decisiones metodológicas en los resultados.

## Pregunta 1

El conjunto de datos utilizado es **Hitters**, incluido en el paquete **ISLR2** de **R**, que contiene estadísticas de bateadores de béisbol de la Major League Baseball (MLB) correspondientes a una temporada específica. Este conjunto de datos se utiliza para explorar la relación entre las métricas de rendimiento de los jugadores y sus salarios, facilitando el análisis predictivo y econométrico.

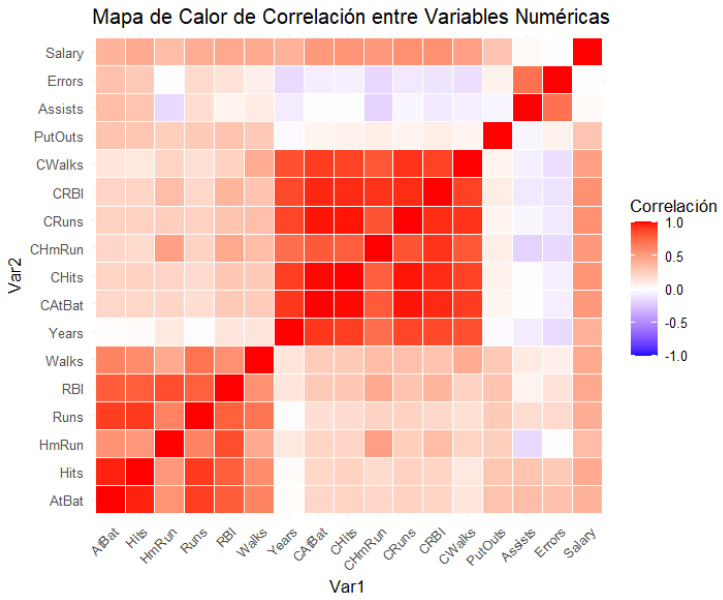
**Descripción de las Variables:**

- **AtBat:** Número de turnos al bate en la temporada actual.
- **Hits:** Número de hits en la temporada actual.
- **HmRun:** Número de home runs en la temporada actual.
- **Runs:** Número de carreras anotadas en la temporada actual.
- **RBI:** Número de carreras impulsadas (Runs Batted In) en la temporada actual.
- **Walks:** Número de bases por bolas en la temporada actual.
- **Years:** Número de años en la liga mayor.
- **CAtBat:** Número total de turnos al bate en la carrera.
- **CHits:** Número total de hits en la carrera.
- **CHmRun:** Número total de home runs en la carrera.
- **CRuns:** Número total de carreras anotadas en la carrera.
- **CRBI:** Número total de carreras impulsadas en la carrera.
- **CWalks:** Número total de bases por bolas en la carrera.
- **League:** Liga en la que juega el bateador (A o N).
- **Division:** División dentro de la liga (E o W).
- **PutOuts:** Número de putouts (jugadas defensivas) en la temporada actual.
- **Assists:** Número de asistencias (jugadas defensivas) en la temporada actual.
- **Errors:** Número de errores defensivos en la temporada actual.
- **Salary:** Salario del bateador en miles de dólares (variable respuesta).
- **NewLeague:** Liga en la que jugó el bateador la temporada anterior (A o N).

**Variables Empleadas:**

Para el análisis, se han considerado todas las variables numéricas disponibles para obtener una visión preliminar del conjunto de datos.

**Figura 1** Mapa de calor de correlación




*Nota.* Fuente: Elaboración propia

El mapa de calor de correlación proporciona una visualización útil para entender las relaciones entre las variables numéricas del conjunto de datos Hitters y puede ayudar a interpretar los resultados de los modelos de mínimos cuadrados ordinarios (OLS) que hemos discutido.

Hay una correlación muy alta entre las variables de carrera acumuladas (CAtBat, CHits, CHmRun, CRuns, CRBI, CWalks) y las variables de temporada actual (AtBat, Hits, HmRun, Runs, RBI, Walks). Esto se evidencia por los colores rojos intensos en el cuadrante superior derecho del mapa, indicando correlaciones cercanas a 1. Estas correlaciones altas sugieren que las métricas de carrera acumuladas son fuertes indicadores de las estadísticas actuales de un jugador, lo cual es lógico ya que las estadísticas de carrera incluyen las de la temporada actual.

**Salary** muestra correlaciones positivas moderadas a fuertes con varias variables, especialmente con CHits, CRuns, CRBI, y CWalks, lo que se refleja en los colores rojizos en la última columna del mapa. Esto indica que la experiencia y el rendimiento a lo largo de la carrera de un jugador son factores importantes en la determinación del salario. Las variables de temporada actual como Hits y Walks también muestran correlaciones positivas con Salary, aunque menos intensas comparadas con las variables de carrera.

## Pregunta 2

En el modelo de **mínimos cuadrados ordinarios (OLS)** con variables seleccionadas, se incluyen las siguientes variables que se considera tienen una relación significativa con el salario: **CHits**, **HmRun**, **Years**, **Walks**, **RBI**, **Hits** y **Runs**. Estas variables fueron elegidas tras un análisis de correlación que sugiere que pueden ser predictoras clave del salario de los jugadores.

Se ha decidido excluir ciertas variables del modelo debido a su impacto limitado o redundante en la predicción del salario:

- **Division:** Excluida porque no parece tener un impacto significativo en el salario según el análisis preliminar.
- **League y NewLeague:** Aunque son factores categóricos interesantes, se excluyen del modelo OLS debido a que la codificación dummy sería necesaria, aumentando la complejidad sin aportar significativamente al modelo según los resultados iniciales.

Las variables seleccionadas reflejan aspectos clave del rendimiento y la experiencia de los jugadores que, en general, se espera que influyan directamente en su salario. La exclusión de **Division** se basa en su aparente falta de significancia estadística, y las ligas se manejan mejor como variables categóricas mediante codificación dummy en otros contextos.

**Interpretación de los Resultados del Modelo OLS con Variables Seleccionadas:**

- **$R^2$ Ajustado:** 0.4214, indicando que alrededor del 42% de la variabilidad en Salary es explicada por el modelo.
- **Coeficientes Significativos:**
    - **CHits:** Positivo y significativo, sugiriendo que más hits a lo largo de la carrera incrementan el salario.
    - **Walks:** Positivo y significativo, indicando que jugadores que reciben más bases por bolas tienden a ganar más.

- **Coeficientes No Significativos:** Variables como HmRun, RBI, Years, Hits, y Runs no son significativas al nivel de 0.05, lo que podría sugerir multicolinealidad o que estas variables no añaden valor predictivo adicional una vez controladas las demás.

Este modelo captura una parte significativa de la variabilidad en el salario, pero hay mucho espacio para mejorar, considerando que más del 50% de la variabilidad no está explicada. La importancia de CHits y Walks sugiere que la experiencia y la disciplina en el plato son valoradas en términos de salario.

## Pregunta 3

Comparación del Modelo OLS con Todas las Variables Explicativas:

- **$R^2$ Ajustado:** 0.4958, mejor que el modelo con variables seleccionadas.
- **Coeficientes Significativos:**
  - **AtBat:** Negativo, sugiriendo que más turnos al bate podrían estar asociados con un menor salario (posiblemente debido a la durabilidad o la posición en el lineup).
  - **Hits, Walks, CRuns, PutOuts:** Positivos y significativos, indicando su contribución positiva al salario.
  - **CWalks:** Negativo, lo cual es contraintuitivo y podría indicar multicolinealidad o interacción con otras variables.

El modelo completo explica casi el 50% de la variabilidad en Salary, superando al 42% del modelo seleccionado, lo que subraya la relevancia de considerar múltiples dimensiones del rendimiento y la carrera de los jugadores. En este modelo, la significancia de más variables sugiere que la interacción entre ellas y la inclusión de diversos aspectos del desempeño son esenciales para la predicción del salario. No obstante, la existencia de variables no significativas puede señalar redundancia o falta de aporte único una vez consideradas otras variables. La diferencia en la significancia de las variables entre ambos modelos podría atribuirse a la multicolinealidad, especialmente entre variables que miden aspectos similares de la carrera, como CHits y CAtBat. Aunque el modelo completo ofrece una visión más amplia y detallada de los factores salariales, el modelo con variables seleccionadas facilita una interpretación más simple y menos afectada por la

multicolinealidad, aunque a costa de una menor precisión explicativa. Juntos, estos enfoques proporcionan valiosas perspectivas sobre los determinantes del salario en el béisbol profesional.

## Pregunta 4

```
> # Usamos la función createDataPartition del paquete caret
> library(caret)
> trainIndex <- createDataPartition(datos_completos$Salary, p = 0.85, list = FALSE)
> datos_entrenamiento <- datos_completos[trainIndex, ]
> datos_prueba <- datos_completos[-trainIndex, ]
> # Paso 2: Ajustar el modelo OLS completo en el conjunto de entrenamiento
> modelo_ols_completo <- lm(Salary ~ . - Division, data = datos_entrenamiento)
> # Paso 3: Evaluar el modelo en el conjunto de prueba
> predicciones <- predict(modelo_ols_completo, newdata = datos_prueba)
> # Paso 4: Calcular el error de prueba (usando RMSE como métrica)
> rmse_prueba <- sqrt(mean((datos_prueba$Salary - predicciones)^2))
> # Mostrar el RMSE del conjunto de prueba
> print(paste("El RMSE en el conjunto de prueba es:", rmse_prueba))
[1] "El RMSE en el conjunto de prueba es: 491.364828305471"
```

Error de Predicción (RMSE): El RMSE (Root Mean Square Error) reportado en el conjunto de prueba es de aproximadamente 491.36. En términos econométricos, este valor representa la desviación estándar de los residuos cuando se usan las predicciones del modelo OLS para estimar los salarios fuera de la muestra de entrenamiento. Un RMSE de este tamaño indica que, en promedio, las predicciones del modelo se desvían en aproximadamente 491.36 miles de dólares de los salarios reales de los jugadores en el conjunto de prueba. Un RMSE tan alto sugiere que el modelo no está capturando adecuadamente la variabilidad en los salarios de los jugadores cuando se aplica a datos no utilizados en su estimación. Esto podría ser debido a varios factores:

1. **Sobreajuste**: es posible que el modelo esté sobre ajustando a los datos de entrenamiento, capturando el ruido específico de este conjunto en lugar de las verdaderas relaciones subyacentes que se generalizarían a otros datos. Esto es especialmente preocupante en modelos con muchas variables explicativas, como en este caso, donde se han incluido todas menos Division.
2. **Multicolinealidad**: dado el análisis de correlación previo que mostró correlaciones altas entre varias variables, la multicolinealidad puede estar inflando la varianza de

los estimadores de los coeficientes, lo que lleva a predicciones menos estables y precisas en el conjunto de prueba.

3. **Heterogeneidad no capturada**: el modelo podría no estar capturando adecuadamente la heterogeneidad en los datos, posiblemente porque las relaciones entre las variables y el salario no son lineales o porque hay interacciones importantes no incluidas en el modelo.

## Pregunta 5

La mejor selección de conjuntos utilizando **validación cruzada de 10 -veces logra un RMSE de 331.0596**. Esto indica que, en promedio, la predicción del salario tiene un error absoluto aproximado de 331.060 dólares. Este método examina todas las posibles combinaciones de variables y selecciona el subconjunto que minimiza el error de predicción. Aunque tiene un buen rendimiento, no es el más preciso entre las alternativas evaluadas, posiblemente debido al sobreajuste inherente al explorar todas las combinaciones posibles de variables.

## Pregunta 6

Este método de **Selección por Pasos Hacia Adelante 10-veces** logra el menor error de prueba entre todos los evaluados, con un **RMSE de 324.364**. La selección por pasos hacia adelante agrega variables al modelo de manera incremental, eligiendo en cada paso la variable que más mejora el ajuste del modelo. Este resultado sugiere que esta estrategia evita incluir variables redundantes o irrelevantes y encuentra una combinación de predictores con mayor precisión.

## Pregunta 7

**El RMSE obtenido con este método utilizando validación cruzada de 5-veces** es más alto (**348.984**), lo que refleja un rendimiento inferior al de la validación con 10 -veces. Esto puede deberse a una menor estabilidad en la estimación del error cuando se utilizan menos veces, lo que introduce mayor variabilidad en los resultados. Además, explorar todas las combinaciones posibles de variables puede generar problemas de sobreajuste, afectando el rendimiento predictivo.

**La selección por pasos hacia adelante con 5-veces** de validación cruzada obtiene un error intermedio (**338.5962**). Aunque tiene mejor rendimiento que la mejor selección de conjuntos en el mismo esquema de validación, su RMSE es mayor que el obtenido con 10 -veces. Esto refuerza la importancia de un esquema de validación más robusto para lograr estimaciones más precisas del error de prueba.

## Pregunta 8

**Tabla 1.** Resultados de Error de Prueba en Modelos Predictivos.

| Método | Error de Prueba (RMSE) | Interpretación Económica (USD) |
|---|---|---|
| Mejor selección de conjuntos (10 veces) | 331.0596 | $331,060 |
| Selección por pasos hacia adelante (10 veces) | 324.3640 | $324,364 |
| Mejor selección de conjuntos (5 veces) | 348.9840 | $348,984 |
| Selección por pasos hacia adelante (5 veces) | 338.5962 | $338,596 |

*Fuente.* Elaboración propia.

El análisis revela que, tanto en validación cruzada de 10 -veces como en la de 5 -veces, la selección por pasos hacia adelante produce un error de prueba más bajo en comparación con la mejor selección de conjuntos. Este resultado indica que, para este conjunto de datos, la estrategia de selección hacia adelante es más eficiente en identificar las variables que contribuyen a un ajuste más preciso del modelo.

Los resultados muestran que, independientemente del método de selección empleado, la validación cruzada de 10 -veces genera un error de prueba menor que la validación cruzada de 5 -veces. Esto es coherente con la teoría estadística, ya que una mayor cantidad de pliegues tiende a proporcionar estimaciones más estables y precisas del error del modelo. No obstante, las diferencias observadas no son extremadamente significativas, aunque sí consistentes.

Las diferencias en el error de prueba entre los dos métodos dentro de la misma configuración de validación cruzada oscilan entre 6 y 10 puntos de RMSE. Aunque esta diferencia no es extremadamente grande, es relevante, dado que el RMSE se encuentra en el rango de los cientos. De manera similar, la validación cruzada de 10 -veces muestra mejoras en el error de prueba frente a la de 5 -veces, con diferencias de aproximadamente 15 a 18 puntos de RMSE para ambos métodos. Esto subraya el valor de emplear más pliegues en la validación cruzada para obtener una evaluación más precisa del rendimiento del modelo, aunque esto conlleva un mayor costo computacional.

Podemos concluir que, la selección por pasos hacia adelante se presenta como el método más adecuado comparado con la mejor selección de conjuntos, basándonos en los menores valores de RMSE obtenidos. Asimismo, la validación cruzada de 10 -veces demuestra ser superior a la de 5 -veces al proporcionar estimaciones más precisas del error de prueba. Por último, aunque las diferencias en RMSE entre los métodos y configuraciones de validación no son drásticas, tienen implicaciones prácticas importantes, especialmente en contextos como la predicción salarial en el ámbito deportivo.

## Pregunta 9

En este apartado, se procede a analizar los coeficientes del modelo en términos de significancia estadística al nivel del 5% y a interpretar su ajuste general, resaltando las implicaciones de los resultados obtenidos.

**Coeficientes**

- **Intercept:** El valor esperado de Salary cuando todas las variables predictoras son cero es 107.34071. No es significativo al nivel de 0.05, pero está cerca (p-valor de 0.099809).
- **AtBat:** Por cada aumento de una unidad en AtBat, Salary disminuye en promedio en 2.30753 unidades. Este efecto es altamente significativo ($p < 0.001$).
- **Hits:** Cada hit adicional está asociado con un aumento en Salary de 7.42174 unidades, también muy significativo.

- **Walks:** Un aumento en Walks está relacionado con un incremento de 5.71641 en Salary, significativo también.
- **CAtBat:** Similar a AtBat, pero para la carrera de un jugador, cada unidad adicional se asocia con una disminución de 0.14941 en Salary, significativo.
- **CRuns:** Cada carrera adicional en la carrera del jugador incrementa Salary en 1.53466 unidades, significativo.
- **CRBI:** Cada RBI adicional está relacionado con un aumento de 0.76838 en Salary, significativo.
- **CWalks:** A diferencia de Walks, un aumento en CWalks (carreras durante la carrera) se asocia con una disminución de 0.80680 en Salary, lo cual es significativo.
- **PutOuts:** Cada putout adicional está asociado con un aumento de 0.30136 en Salary, significativo.
- **Assists:** Aunque el efecto es positivo (0.30222), su significancia es marginal ($p = 0.059650$), justo por encima del umbral típico de 0.05 para la significancia.

**Ajuste del modelo**

- **Residual Standard Error:** De 316.2, lo que indica la desviación estándar de los residuos, sugiriendo que hay bastante variabilidad en la predicción.
- **Multiple R-squared:** 0.5255, indicando que aproximadamente el 52.55% de la variabilidad en Salary puede ser explicada por el modelo.
- **Adjusted R-squared:** 0.5086, que ajusta por el número de predictores en el modelo, es ligeramente menor, reflejando la penalización por la complejidad del modelo.
- **F-statistic:** Con un valor p extremadamente bajo ($< 2.2e-16$), el modelo como un todo es significativamente mejor que un modelo nulo en predecir Salary.

Este modelo sugiere que varios aspectos del rendimiento de un jugador, tanto a corto (temp. actual) como a largo plazo (carrera), tienen un impacto significativo en su salario. Sin embargo, hay que considerar que los residuos muestran una dispersión considerable, lo que indica que hay otros factores no incluidos en el modelo que también influyen en el salario. Además, la presencia de coeficientes negativos para variables como AtBat y CWalks podría sugerir algún fenómeno

específico del deporte o problemas de multicolinealidad que merecerían una investigación más profunda.

## Pregunta 10

El resultado indica que el Error de Prueba (RMSE) para la Regresión Ridge con $\lambda = 0.1$ es **335.7037**. Un RMSE de 335.7037 significa que, en promedio, las predicciones de salario del modelo Ridge con $\lambda = 0.1$ se desvían de los valores reales de salario por aproximadamente esta cantidad en unidades de la variable Salary. La regularización Ridge reduce la magnitud de los coeficientes para prevenir el sobreajuste, pero puede que no elimine variables irrelevantes como lo haría LASSO ($\alpha$=1). Esto podría ser una desventaja si hay muchas variables que no aportan mucho al modelo.

## Pregunta 11

El RMSE de 336.3670 indica que, en promedio, las predicciones de salario del modelo LASSO se desvían de los valores reales de salario por aproximadamente esta cantidad. Comparando este valor con otros modelos que has probado, como el RMSE de 324.364 para la Selección por Pasos Hacia Adelante con Validación Cruzada de 10 Veces o el de 335.7037 para el modelo Ridge, podemos ver que el modelo LASSO tiene un error de prueba ligeramente más alto que Ridge, pero superior al modelo de selección hacia adelante.

```
#Número de coeficientes diferentes de cero
coeficientes_no_cero <- sum(coef(modelo_lasso_final) != 0)
print(paste("Número de coeficientes estimados diferentes de cero:",
coeficientes_no_cero))
[1] "Número de coeficientes estimados diferentes de cero: 18"
```

El hecho de que haya 18 coeficientes diferentes de cero sugiere que el modelo LASSO ha seleccionado 18 variables como significativas para la predicción de Salary. Esto muestra que LASSO ha realizado una selección de variables, reduciendo la complejidad del modelo al eliminar las variables menos importantes (o las que no superan el umbral de regularización impuesto por $\lambda$).

## Pregunta 12

```
[1] "Error de prueba (RMSE) para Regresión Ridge con 5 veces CV: 335.698082501018"
[1] "Error de prueba (RMSE) para Regresión LASSO con 5 veces CV: 335.006560972268"
[1] "Número de coeficientes estimados diferentes de cero para LASSO con 5 veces CV: 18"
```

Con la regresión Ridge, el RMSE con validación cruzada de 5-veces es 335.6981, y con 10-veces es 335.7037, mostrando una diferencia mínima, lo que indica que el número de pliegues no afecta significativamente el rendimiento.

En el modelo LASSO, el RMSE es de 335.0066 con 5-veces CV, mejor que el 336.3670 de 10-veces CV, sugiriendo que LASSO es más preciso con menos pliegues. LASSO seleccionó 18 variables en ambos casos, demostrando consistencia y robustez en la selección de variables.

En comparación, LASSO supera ligeramente a Ridge en RMSE con 5-veces CV, destacando su eficacia en la selección de variables y su estabilidad frente a cambios en la validación cruzada, lo que lo hace preferible para este conjunto de datos.

## Pregunta 13

El modelo de Componentes Principales (PCA) fue evaluado utilizando validación cruzada con configuraciones de 10 y 5-veces para determinar el número óptimo de componentes principales (M) y estimar el error de prueba (RMSE) asociado. En el esquema de validación cruzada de 10-veces, el modelo seleccionó 15 componentes principales como óptimos, indicando que se requieren 15 dimensiones para capturar una parte significativa de la variabilidad en los predictores y lograr un buen ajuste del modelo. El error de prueba correspondiente fue de 333.2915 (RMSE), lo que significa que, en promedio, las predicciones realizadas por el modelo PCA se desvían de los valores reales de salario en esta magnitud. Este resultado refleja que la validación cruzada de 10-veces proporciona una evaluación precisa y robusta del rendimiento del modelo, permitiendo capturar más información de los datos sin comprometer la estabilidad de la estimación del error. En contraste, en la configuración de validación cruzada de 5-veces, el modelo seleccionó un número menor de componentes principales, específicamente 10 componentes, como el valor óptimo para explicar la variabilidad de los datos. Sin embargo, el error de prueba obtenido en este esquema fue de 339.2862 (RMSE), ligeramente mayor al valor registrado con 10-veces. La selección de menos

componentes puede reflejar que, con un menor número de pliegues en la validación cruzada, las estimaciones del rendimiento del modelo son más propensas a variaciones, lo que puede llevar a una evaluación menos precisa del ajuste.

## Pregunta 14

El modelo de Mínimos Cuadrados Parciales (PLS) fue evaluado utilizando validación cruzada con configuraciones de 10 y 5-veces para identificar el número óptimo de componentes principales (M) y estimar el error de prueba (RMSE). En la validación cruzada de 10-veces, se seleccionaron 2 componentes como óptimos, demostrando que PLS puede capturar eficientemente la relación entre variables predictoras y respuesta con una baja complejidad, con un RMSE de 332.0606, lo que indica una desviación promedio de las predicciones del salario real. En contraste, con la validación cruzada de 5-veces, se eligieron 10 componentes, resultando en un RMSE de 339.2862, ligeramente superior. Esto sugiere que, con menos pliegues, el modelo requiere más componentes para capturar la información relevante, lo cual puede llevar a estimaciones menos estables del error de prueba.

## Pregunta 15

**Tabla 2.** Errores de prueba de modelos comparados

| Modelo y Validación Cruzada | Error de Prueba (RMSE) |
|---|---|
| Regresión Ridge (10 veces CV) | 335.7037 |
| Regresión Ridge (5 veces CV) | 335.6981 |
| Regresión LASSO (10 veces CV) | 336.367 |
| Regresión LASSO (5 veces CV) | 335.0066 |
| Componentes Principales (10 veces CV) | 333.2915 |
| Componentes Principales (5 veces CV) | 339.2862 |
| Mínimos Cuadrados Parciales (PLS, 10 veces CV) | 332.0606 |
| Mínimos Cuadrados Parciales (PLS, 5 veces CV) | 339.2862 |

*Fuente.* Elaboración propia.

Los errores de prueba (RMSE) de los diferentes enfoques de modelado varían entre 332.0606 y 339.2862, con una diferencia máxima de 7.2256, lo que sugiere un desempeño similar en predicción fuera de la muestra. El modelo de Mínimos Cuadrados Parciales (PLS) con validación cruzada de 10 particiones es el más preciso, con un RMSE de 332.0606, superando ligeramente al modelo de Componentes Principales (PCR) con 10 particiones, que tuvo un RMSE de 333.2915. La validación cruzada de 10 particiones ofrece evaluaciones más estables y precisas que la de 5 particiones, como se evidencia por los menores RMSE en PLS y PCR. Entre Ridge y LASSO, ambos presentan rendimientos competitivos, aunque con RMSEs ligeramente mayores. En conclusión, PLS con 10 particiones emerge como el mejor ajuste, destacando por su baja RMSE y su eficiencia en capturar la relación predictor-respuesta con solo 2 componentes, lo que lo hace robusto y eficiente para este análisis.

Basado en el RMSE, el mejor modelo entre los dos es la Selección por pasos hacia adelante con 10 veces de validación cruzada. Este resultado sugiere que, para este conjunto de datos y problema específico, la selección de variables paso a paso ha identificado una combinación de variables que predice el salario con mayor precisión que el modelo PLS, el cual, aunque eficiente en la reducción de dimensionalidad, no alcanza la misma precisión predictiva en este caso.

Es importante destacar que, aunque el RMSE es un indicador crucial de la precisión del modelo, la elección final también podría depender de otros factores como la interpretabilidad del modelo, la simplicidad, y la necesidad de reducir la dimensionalidad o de manejar la multicolinealidad. Sin embargo, en cuanto a la precisión de predicción medida por RMSE, la Selección por pasos hacia adelante con 10 veces de validación cruzada es el modelo preferido.

Brian Abad Guadalupe
Universitat de les Illes Balears
Centre d'Estudis de Postgrau
MADM - Màster Universitari en Anàlisi de Dades Massives en Economia i Empresa