Universidad Nacional De Colombia

Sede Medellín

Facultad de Ciencias

Departamento de Estadística

Taller RML (Parte 2)

Daniel Felipe Villa Rengifo

Luis David Hernández Pérez

Juan Gabriel Carvajal Negrete

Modelos de Regresión

Enero, 2025

Primeramente recordaremos el contexto de la base de datos y las covariables a ser incluidas para el análisis de nuestro grupo.

El conjunto de datos `boston.csv` contiene información recopilada por el Servicio del Censo de EE. UU. con respecto a la vivienda en el área de Boston,Massachusetts.

Las variables a incluir en el análisis son:

- `CRIM`: tasa de criminalidad per cápita por ciudad.
- `NOX`: concentración de óxidos nítricos (partes por 10 millones).
- `RM`: Número medio de habitaciones por vivienda.
- `AGE`: Proporción de unidades ocupadas por el propietario construidas antes de 1940.
- `PTRATIO`: Proporción de alumnos por profesor por ciudad.
- `LSTAT`: Porcentaje de población con nivel socio-económico bajo.
- `MEDV`: Valor medio de las viviendas ocupadas por el propietario en $1000.

Punto 1

Realice diagnósticos de multicolinealidad mediante.

a) Matriz de correlación de las variables predictoras.

Tabla 1: Matriz de Correlaciones

variables	corr
CRIM-NOX	0.2089774
CRIM-RM	-0.2573178
CRIM-AGE	0.3356346
CRIM-PTRATIO	0.2594995
CRIM-LSTAT	0.3336717
NOX-RM	0.1923926
NOX-AGE	0.4889553
NOX-PTRATIO	0.2975224
NOX-LSTAT	0.3394929
RM-AGE	0.0350345
RM-PTRATIO	0.1976719
RM-LSTAT	-0.3713510
AGE-PTRATIO	0.4611442
AGE-LSTAT	0.4921953
PTRATIO-LSTAT	0.1537090

De acuerdo a Tabla 1 y Figura 1 concluimos que no hay evidencia clara de multicolinealidad severa.

b) VIF's

`Coeficientes estimados, sus I.C, Vifs y Coeficientes estimados estandarizados`

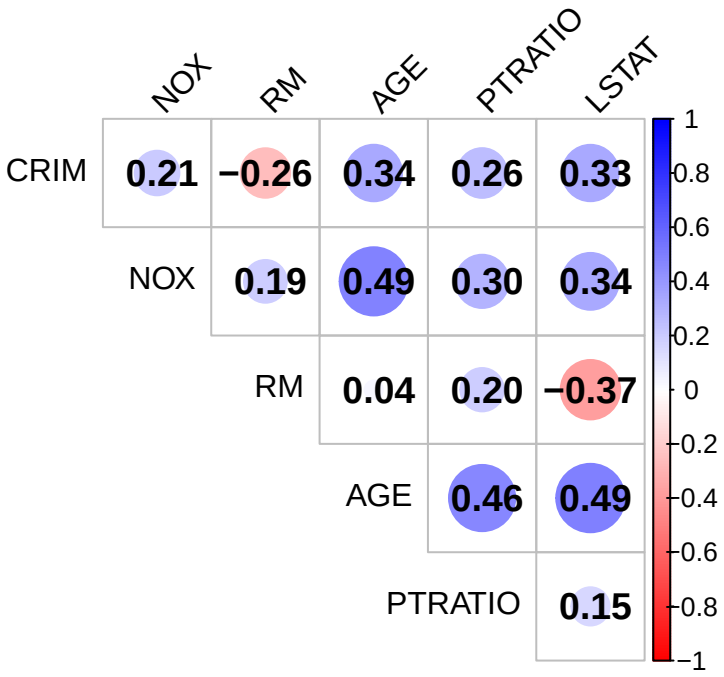

Figura 1: Gráfico de Correlaciones

Tabla 2: Factores de inflación de varianza (VIF) para las variables predictoras del modelo

	Vif
(Intercept)	0.000000
CRIM	1.298144
NOX	1.478590
RM	1.513394
AGE	1.885380
PTRATIO	1.381249
LSTAT	1.768228

Según la Tabla 2 la multicolinealidad no parece ser un problema significativo en este conjunto de datos según los valores de VIF.

c) Indice de condición

Para calcular en indice de condición tendremos en cuenta centrar los datos (Montgomeryet.al.,2021) ya que intercepto solo tendrá una interpretación útil si los valores de las variables independientes igual a cero tienen sentido en el contexto del conjunto de datos. En este caso particular, dado que algunas variables como `RM` número promedio de habitaciones), no tienen sentido físico o práctico como cero, ya que una casa no puede tener cero habitaciones, por tanto el intercepto no tendría una interpretación práctica clara.

Tabla 3: Índice de condición para las variables predictoras del modelo

	Condition.Index
CRIM	1.000000
NOX	1.282998
RM	1.721514
AGE	1.978785
PTRATIO	2.389880
LSTAT	2.641337

Según la Tabla 3 todos los valores del índice de condición están muy por debajo de 10, lo que indica que no hay problemas significativos de multicolinealidad entre las variables predictoras.

d) Proporciones de varianza.

Para el criterio de proporciones de varianza tambien se hizo lo expuesto en el criterio anterior (Indice de condición).

Tabla 4: Proporciones de varianza para las variables predictoras del modelo

	Eigenvalue	CRIM	NOX	RM	AGE	PTRATIO	LSTAT
CRIM	2.3821	0.0498	0.0542	0.0012	0.0650	0.0461	0.0492
NOX	1.4472	0.0417	0.0371	0.2591	0.0052	0.0631	0.0526
RM	0.8038	0.3134	0.1880	0.0010	0.0072	0.2625	0.1027
AGE	0.6084	0.4940	0.2430	0.0393	0.0513	0.2685	0.0657
PTRATIO	0.4171	0.0583	0.4152	0.2582	0.4433	0.3404	0.0064
LSTAT	0.3414	0.0428	0.0624	0.4413	0.4280	0.0195	0.7234

Según la Tabla 4 no hay proporciones π_{ij} altas (> 0.5) para dos o más coeficientes de regresión asociados con un mismo valor propio pequeño, por tanto no hay evidencia de multicolinealidad entre las variables correspondientes a tales coeficientes.

Punto 2

Construya modelos de regresión utilizando los métodos de selección (muestre de cada método solo la tabla de resumen de este y la tabla ANOVA y la de parámetros estimados del modelo finalmente resultante):

a) Selección según el $R^2_{\mathbf{adj}}$

Tabla 5: Selección de Modelos según el Estadístico R^2_{adj}

predictors	adjr
CRIM NOX LSTAT	0.587
CRIM NOX PTRATIO LSTAT	0.586
CRIM NOX AGE LSTAT	0.585
CRIM NOX RM LSTAT	0.583
CRIM NOX AGE PTRATIO LSTAT	0.582
CRIM NOX RM PTRATIO LSTAT	0.582

Según la Tabla 5 podemos concluir que:

- El modelo con las variables predictoras `CRIM`, `NOX`y `LSTAT` es el mejor segund el R^2_{adj} ya que maximiza este estadístico con un número mínimo de predictores.
- La inclusión de más variables predictoras como `PTRATIO` o `AGE` no parece justificar un mejor desempeño,ya que el R^2_{adj} no mejora sustancialmente.

b) Selección según el estadístico C_p.

Selección según el estadístico C_p.

Tabla 6: Selección de Modelos según el Estadístico C_p

predictors	cp
CRIM NOX LSTAT	1.850
CRIM NOX PTRATIO LSTAT	3.256
CRIM NOX AGE LSTAT	3.439
CRIM NOX RM LSTAT	3.845
CRIM NOX AGE PTRATIO LSTAT	5.106
CRIM NOX RM PTRATIO LSTAT	5.177

Según la Tabla 6 el mejor modelo, según el criterio C_p es con las variables predictoras `CRIM`, `NOX`, `LSTAT`.

c) Stepwise

Tabla 7: Resumen Stepwise

Paso	Variable Agregada	AIC	R^2	R^2 Ajustado
0	Base Model	621.008	0.00000	0.00000
1	LSTAT (+)	559.176	0.47182	0.46643
2	NOX (+)	543.986	0.55524	0.54607
3	CRIM (+)	535.404	0.59990	0.58740

Tabla 8: ANOVA Stepwise

Fuente	Suma de Cuadrados	DF	Media Cuadrática	F	Significancia
Regresión	1679.722	3	559.907	47.98	0.0000
Residual	1120.278	96	11.670		
Total	2800.000	99			

Tabla 9: Parámetros estimados del modelo final Stepwise

Variable	Beta	Std. Error	Std. Beta	t	Significancia	Inferior	Superior
(Intercept)	40.703	3.495		11.646	0.000	33.765	47.640
LSTAT	-0.488	0.068	-0.516	-7.201	0.000	-0.622	-0.353
NOX	-23.187	5.663	-0.283	-4.094	0.000	-34.428	-11.946
CRIM	-0.095	0.029	-0.225	-3.274	0.001	-0.153	-0.037

d) Selección hacia adelante o *forward*

Tabla 10: Resumen Forward

Paso	Variable Agregada	AIC	R^2	R^2 Ajustado
0	Base Model	621.008	0.00000	0.00000
1	LSTAT	559.176	0.47182	0.46643
2	NOX	543.986	0.55524	0.54607
3	CRIM	535.404	0.59990	0.58740

Tabla 11: ANOVA Forward

Fuente	Suma de Cuadrados	DF	Media Cuadrática	F	p-valor
Regresión	1679.722	3	559.907	47.98	0.0000
Residual	1120.278	96	11.670		
Total	2800.000	99			

Tabla 12: Parámetros estimados del modelo final Forward

Variable	Beta	Std. Error	Std. Beta	t	p-valor	Inferior	Superior
(Intercept)	40.703	3.495		11.646	0.000	33.765	47.640
LSTAT	-0.488	0.068	-0.516	-7.201	0.000	-0.622	-0.353
NOX	-23.187	5.663	-0.283	-4.094	0.000	-34.428	-11.946
CRIM	-0.095	0.029	-0.225	-3.274	0.001	-0.153	-0.037

e) Selección hacia atrás o *backward*

Tabla 13: Resumen Backward

Paso	Variable Eliminada	AIC	R^2	R^2 Ajustado
0	Full Model	540.494	0.60353	0.57795
1	RM	538.608	0.60307	0.58196
2	AGE	536.769	0.60243	0.58569
3	PTRATIO	535.404	0.59990	0.58740

Tabla 14: ANOVA Backward

Fuente	Suma de Cuadrados	DF	Media Cuadrática	F	p-valor
Regresión	1679.722	3	559.907	47.98	0.0000
Residual	1120.278	96	11.670		
Total	2800.000	99			

Tabla 15: Parámetros estimados del modelo final Backward

Variable	Beta	Std. Error	Std. Beta	t	p-valor	Inferior	Superior
(Intercept)	40.703	3.495		11.646	0.000	33.765	47.640
CRIM	-0.095	0.029	-0.225	-3.274	0.001	-0.153	-0.037
NOX	-23.187	5.663	-0.283	-4.094	0.000	-34.428	-11.946
LSTAT	-0.488	0.068	-0.516	-7.201	0.000	-0.622	-0.353

De los tres métodos de selección podemos concluir que:

- Los tres métodos seleccionaron las mismas tres variables (LSTAT, NOX y CRIM), lo que sugiere que estas son las más relevantes para predecir MEDV.
- Los tres métodos seleccionaron el mismo modelo final, lo que confirma la estabilidad y robustez del proceso de selección.
- El modelo final es significativo y bien ajustado, con un R^2 ajustado de 0.587 y un AIC mínimo de 535.404.
- Las variables clave (LSTAT, NOX y CRIM) tienen un impacto negativo en MEDV, indicando que el nivel socio-económico, la contaminación y la criminalidad son determinantes en los precios de la vivienda. Dado el desempeño similar de los métodos, cualquiera de ellos es válido, pero **Stepwise** combinado puede considerarse el más flexible al evaluar tanto adiciones como eliminaciones de variables en el proceso.

Punto 3

Realice el ajuste utilizando los métodos RR y LASSO. Compare los resultados y comente.

Regresión Ridge

```
# Creando matriz de diseño

X <- as.matrix(model.matrix(modelo))[,-1]
y <- datos4$MEDV

# Ajuste del modelo de regresion ridge

model_ridge <- glmnet(X, y, alpha = 0)
```

Para identificar el valor de **kappa** que da lugar al mejor modelo,recurriremos a validación cruzada con la función `cv.glmnet()`.

```
[1] "Mejor valor de kappa encontrado: 1.21820567454074"
```

Ahora ajustamos el modelo de regresión ridge nuevamente con el valor de kappa optimo.

Tabla 16: Coeficientes regresión ridge con el mejor kappa

Variable	Coeficientes
(Intercept)	55.64931433
CRIM	-0.07570871
NOX	-19.09952447
RM	0.55650949
AGE	-0.03113410
PTRATIO	-1.03489490
LSTAT	-0.36365557

```
[1] "Error (mse) mejor modelo ridge: 11.403038443457"
```

Regresión Lasso

```
# Ajuste del modelo de regresion lasso
model_lasso <- glmnet(X, y, alpha = 1)
```

Para identificar el valor de **kappa** que da lugar al mejor modelo,recurriremos a validación cruzada con la función `cv.glmnet()`.

```
[1] "Mejor valor de kappa encontrado: 0.294821398528895"
```

Ahora ajustamos el modelo de regresión lasso nuevamente con el valor de kappa optimo .

Tabla 17: Coeficientes regresión lasso con el mejor kappa

Variable	Coeficientes
(Intercept)	29.47864426
CRIM	-0.02935753
NOX	-10.47212401
RM	.
AGE	.
PTRATIO	.
LSTAT	-0.36882565

```
[1] "Error (mse) mejor modelo lasso: 11.3077507039502"
```

De la Tabla 16 y Tabla 17 que representa los coeficientes de los modelos con el mejor valor optimo de kappa podemos notar que:

- Ridge mantiene todas las variables en el modelo, pero penaliza sus coeficientes, reduciendo su magnitud.
- Lasso seleccionó solo tres variables (CRIM, NOX, LSTAT) como las más relevantes, eliminando las demás.

Después de ajustar los modelos de Ridge y LASSO utilizando los valores óptimos de Kappa y realizar predicciones, calculamos los errores medios cuadrados (MSE) para cada modelo.

Tabla 18: MSE mejores modelos Ridge y Lasso

Modelo	MSE (Error Medio Cuadrático)
Ridge	11.4030
LASSO	11.3078

De la Tabla 18 podemos decir que ambos modelos tienen errores similares, el modelo LASSO es ligeramente superior en términos de precisión predictiva debido a su menor MSE, la diferencia en el MSE entre ambos métodos es pequeña (~0.1), lo que sugiere que ambos modelos son adecuados para el problema (predicción).

Punto 4

Realice el ajuste PCR y comente.

Primera mente aplicamos analisis de componetes principales a las variables predictoras.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.543 1.2030 0.8965 0.7800 0.64581 0.58433
Proportion of Variance 0.397 0.2412 0.1340 0.1014 0.06951 0.05691
Cumulative Proportion  0.397 0.6382 0.7722 0.8736 0.94309 1.00000
```

De acuedo a la salidad anterior basado en la proporción acumulada de varianza, 3 componentes principales son suficientes para capturar la mayor parte de la información relevante (77.2%).

Ahora ajustemos sel modelo PCR realizando validación cruzada para elegir el número óptimo de componentes principales.

```
set.seed(6475)
pcr_model <- pcr(MEDV ~ ., data = datos4, scale = TRUE, validation = "CV")
summary(pcr_model)
```

```
Data:   X dimension: 100 6
    Y dimension: 100 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           5.345    3.747    3.647    3.564    3.584    3.621    3.639
adjCV        5.345    3.740    3.641    3.555    3.575    3.619    3.623

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X       39.70    63.82    77.22    87.36    94.31   100.00
MEDV    50.86    54.59    57.04    57.47    57.71    60.35
```

De la salida anterior podemos concluir que:

- El error predicho disminuye significativamente hasta los 3 componentes principales y luego se estabiliza.
- 3 componentes principales parece ser el número óptimo, ya que tiene el menor RMSEP ajustado (adjCV) de 3.632.

- Los 3 primeros componentes principales explican el 77.22% de la varianza en las variables predictoras X y el 57.04% de la varianza en MEDV.
- Agregar más componentes (4, 5, o 6) no mejora significativamente la varianza explicada en MEDV.

Después de seleccionar el número óptimo de componentes, realizaremos predicciones y calcularemos el error medio cuadrático (MSE).

```
# Predicciones utilizando los componentes seleccionados
predicciones <- predict(pcr_model, ncomp = 3, newdata = datos4)

# Calcular el MSE
mse_pcr <- mean((predicciones - datos4$MEDV)^2)
print(paste("MSE del modelo PCR:", mse_pcr))
```

```
[1] "MSE del modelo PCR: 12.0300663468632"
```

El MSE de PCR (12.03) es ligeramente mayor que el de Ridge (11.403) y LASSO (11.308). Esto indica que Ridge y LASSO son marginalmente más precisos en este caso específico