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.

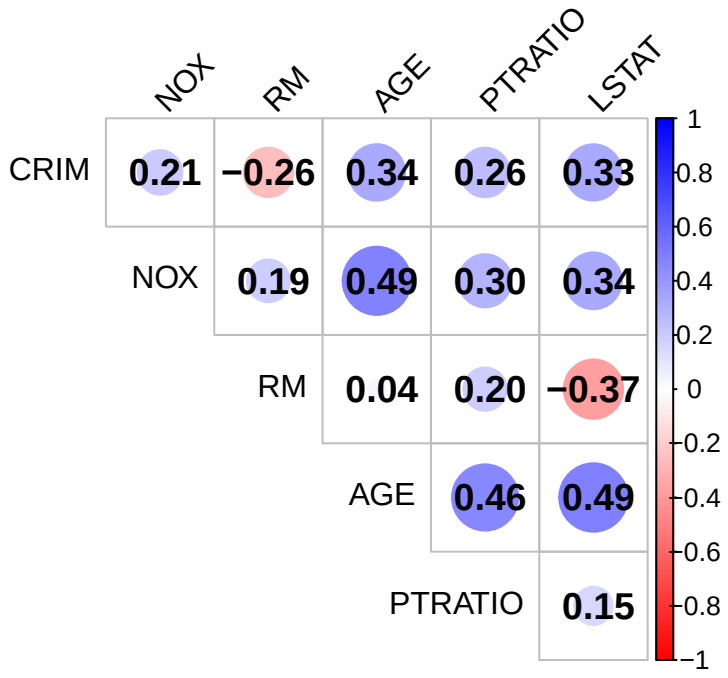

Figura 1: Gráfico de Correlaciones

b) VIF's

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

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

De la Tabla 7 podemos conluir que el modelo final selecciona LSTAT, NOX, y CRIM como las variables predictoras más relevantes.

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		

Fuente	Suma de Cuadrados	DF	Media Cuadrática	F	Significancia
Total	2800.000	99			

De la Tabla 8 podemos concluir que el modelo es significativo y explica una gran parte de la variación total en MEDV.

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

Todas las variables predictoras seleccionadas tienen un impacto significativo y negativo en MEDV.

d) Selección hacia adelante o *forward*

```
ols_step_forward_p(modelo,p_val=0.05,details =TRUE)
```

```
Forward Selection Method
------------------------

Candidate Terms:

1. CRIM
2. NOX
3. RM
4. AGE
5. PTRATIO
6. LSTAT


Step   => 0
Model  => MEDV ~ 1
R2     => 0

Initiating stepwise selection...

                    Selection Metrics Table
-----------------------------------------------------------------
Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-----------------------------------------------------------------
```

```
LSTAT         0.00000        0.472             0.466    559.176
NOX           0.00000        0.255             0.247    593.586
AGE           0.00000        0.251             0.243    594.149
CRIM          0.00000        0.209             0.200    599.623
PTRATIO       0.00684        0.072             0.063    615.506
RM            0.04678        0.040             0.030    618.954
-------------------------------------------------------------

Step      => 1
Selected  => LSTAT
Model     => MEDV ~ LSTAT
R2        => 0.472

                    Selection Metrics Table
-------------------------------------------------------------
Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-------------------------------------------------------------
NOX             5e-05        0.555             0.546    543.986
CRIM          0.00079        0.530             0.520    549.499
AGE           0.01023        0.507             0.497    554.342
PTRATIO       0.02363        0.499             0.489    555.869
RM            0.41624        0.475             0.465    560.491
-------------------------------------------------------------

Step      => 2
Selected  => NOX
Model     => MEDV ~ LSTAT + NOX
R2        => 0.555

                    Selection Metrics Table
-------------------------------------------------------------
Predictor    Pr(>|t|)    R-Squared    Adj. R-Squared      AIC
-------------------------------------------------------------
CRIM          0.00148        0.600             0.587    535.404
PTRATIO       0.17429        0.564             0.550    544.054
AGE           0.24137        0.562             0.548    544.549
RM            0.47555        0.558             0.544    545.453
-------------------------------------------------------------

Step      => 3
Selected  => CRIM
Model     => MEDV ~ LSTAT + NOX + CRIM
R2        => 0.6

                    Selection Metrics Table
-------------------------------------------------------------
```

Predictor	Pr(>|t|)	R-Squared	Adj. R-Squared	AIC
PTRATIO	0.43847	0.602	0.586	536.769
AGE	0.51930	0.602	0.585	536.964
RM	0.93933	0.600	0.583	537.398

No more variables to be added.

Variables Selected:

=> LSTAT
=> NOX
=> CRIM

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	621.008	626.218	335.415	0.00000	0.00000
1	LSTAT	559.176	566.991	274.542	0.47182	0.46643
2	NOX	543.986	554.407	259.946	0.55524	0.54607
3	CRIM	535.404	548.430	252.132	0.59990	0.58740

Final Model Output

Model Summary

R	0.775	RMSE	3.347
R-Squared	0.600	MSE	11.203
Adj. R-Squared	0.587	Coef. Var	22.036
Pred R-Squared	0.562	AIC	535.404
MAE	2.487	SBC	548.430

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of

	Squares	DF	Mean Square	F	Sig.
Regression	1679.722	3	559.907	47.98	0.0000
Residual	1120.278	96	11.670		
Total	2800.000	99			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(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

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*

```
ols_step_backward_p(modelo,p_val=0.05,details = TRUE) # Selección backward
```

```
Backward Elimination Method
---------------------------

Candidate Terms:

1. CRIM
2. NOX
3. RM
4. AGE
5. PTRATIO
6. LSTAT


Step   => 0
Model  => MEDV ~ CRIM + NOX + RM + AGE + PTRATIO + LSTAT
R2     => 0.604

Initiating stepwise selection...

Step     => 1
Removed  => RM
Model    => MEDV ~ CRIM + NOX + AGE + PTRATIO + LSTAT
R2       => 0.60307

Step     => 2
Removed  => AGE
Model    => MEDV ~ CRIM + NOX + PTRATIO + LSTAT
R2       => 0.60243

Step     => 3
Removed  => PTRATIO
Model    => MEDV ~ CRIM + NOX + LSTAT
R2       => 0.5999


No more variables to be removed.

Variables Removed:

=> RM
=> AGE
=> PTRATIO
```

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Full Model	540.494	561.335	257.748	0.60353	0.57795
1	RM	538.608	556.844	255.698	0.60307	0.58196
2	AGE	536.769	552.400	253.694	0.60243	0.58569
3	PTRATIO	535.404	548.430	252.132	0.59990	0.58740

Final Model Output

Model Summary

R	0.775	RMSE	3.347
R-Squared	0.600	MSE	11.203
Adj. R-Squared	0.587	Coef. Var	22.036
Pred R-Squared	0.562	AIC	535.404
MAE	2.487	SBC	548.430

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	1679.722	3	559.907	47.98	0.0000
Residual	1120.278	96	11.670		
Total	2800.000	99			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(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

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

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			

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

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()`.

```
set.seed(21)
cv_model <- cv.glmnet(X, y, alpha = 0) # encontrado el mejor kappa
```

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

```r
# Mejor modelo lambda óptimo + 1sd
# ================================================================================
mejor_modelo_rr <- glmnet(X, y, alpha = 0,lambda = cv_model$lambda.min)
```

```r
# Predicciones modelo de regresion ridge
# ================================================================================
predicciones_rr <- predict(mejor_modelo_rr, newx = X)
```

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

Regresión Lasso

```r
# 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()`.

```r
set.seed(64)
cv_model <- cv.glmnet(X, y, alpha = 1) # encontrado el mejor kappa
```

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

```r
# Mejor modelo kappa óptimo

mejor_model_lasso <- glmnet(X, y, alpha = 1,lambda = cv_model$lambda.min)
```

```r
# Predicciones modelo de regresion ridge

predicciones_lasso <- predict(mejor_model_lasso, newx = X)
```

```
[1] "Error (mse) modelo lasso: 11.3080127422723"
```

Punto 4

Realice el ajuste PCR y comente.

```r
# División de los datos en train y test

entrenamiento <- datos4[401:480, ]
prueba <- datos4[481:500, ]
```