Taller RML (Parte 1) - Gr4

Departamento de Estadística - UNALMED

Luis David Hernandez Peréz Daniel Felipe villa Rengifo
Juan Gabriel Carvajal Negrete

1. Realice una descripción de la base de datos. Contextualice el problema y explique cada una de las variables involucradas en el modelo. (https://www.codersarts.com/post/predict-boston-house-prices-using-python-linear-regression).

Descripción de la base de datos

Tomaremos el conjunto de datos de Vivienda que contiene información sobre diferentes casas en Boston. Hay 506 muestras y 13 variables de características en este conjunto de datos. El objetivo es predecir el valor de los precios de la casa utilizando las características dadas.

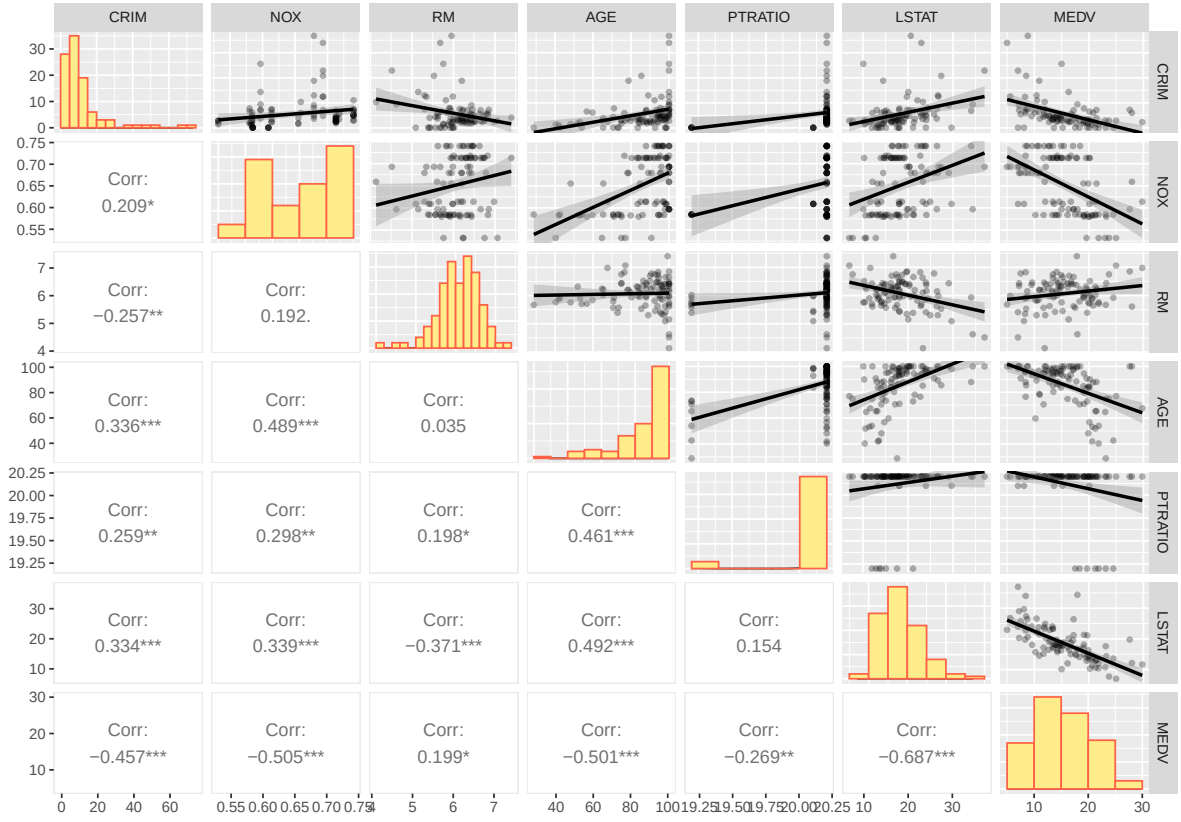
La descripción de todas las características se proporciona a continuación:

- **CRIM**: tasa de criminalidad per cápita por ciudad
- **ZN**: proporción de terrenos residenciales zonificados para lotes de más de 25 000 pies cuadrados
- **INDUS:** proporción de acres de negocios no minoristas por ciudad
- **CHAS**: variable ficticia de Charles River (= 1 si el terreno limita con el río; 0 en caso contrario) NOX: concentración de óxido nítrico (partes por 10 millones)
- **RM**: número promedio de habitaciones por vivienda
- **AGE**: proporción de unidades ocupadas por sus propietarios construidas antes de 1940 DIS: distancias ponderadas a cinco centros de empleo de Boston
- **RAD**: índice de accesibilidad a carreteras radiales
- **TAX**: tasa de impuesto a la propiedad de valor total por cada $10 000
- **PTRATIO**: proporción de alumnos por maestro por ciudad B: $1000(Bk - 0{,}63)^2$, donde Bk es la proporción de [personas de afroamericanas [descendencia] por ciudad

- **LSTAT**: Porcentaje de la población de estatus inferior
- **MEDV**: Valor medio de las viviendas ocupadas por sus propietarios en miles de dólares

Los precios de las viviendas indicados por la variable **MEDV** son nuestra variable objetivo y las restantes son las variables características en función de las cuales predeciremos el valor de una vivienda.

2. Realice un análisis descriptivo de las variables que se van a tener en cuenta en el modelo. Concluya.



Correlaciones: MEDV tiene una fuerte correlación positiva con RM (número promedio de habitaciones por vivienda), lo que indica que las viviendas con más habitaciones tienden a tener un valor más alto. Por otro lado, MEDV tiene una fuerte correlación negativa con LSTAT (porcentaje de la población de estatus inferior), lo que sugiere que las viviendas en áreas con mayor porcentaje de población de bajos ingresos tienden a tener un valor más bajo.

Otras variables: CRIM (tasa de criminalidad): Tiene una correlación negativa moderada con MEDV, lo que indica que las viviendas en áreas con mayor criminalidad tienden a tener un valor más bajo.RM (número de habitaciones): Además de su fuerte correlación con MEDV, RM también muestra una correlación positiva con AGE (proporción de unidades ocupadas por sus

propietarios construidas antes de 1940). Esto podría indicar que las viviendas más antiguas tienden a tener más habitaciones.LSTAT (porcentaje de población de bajos ingresos): Además de su correlación negativa con MEDV, LSTAT también muestra correlaciones negativas con RM y una correlación positiva con CRIM. Esto sugiere que las áreas con mayor porcentaje de población de bajos ingresos tienden a tener viviendas más pequeñas y mayor criminalidad.

```
datos4 %>% ggplot(aes(x=MEDV))+
  geom_histogram(aes(y=..density..), fill="dodgerblue") +
  geom_density(linewidth = 1.2) +
  theme_bw() +
  theme(axis.title.y = element_blank())
```

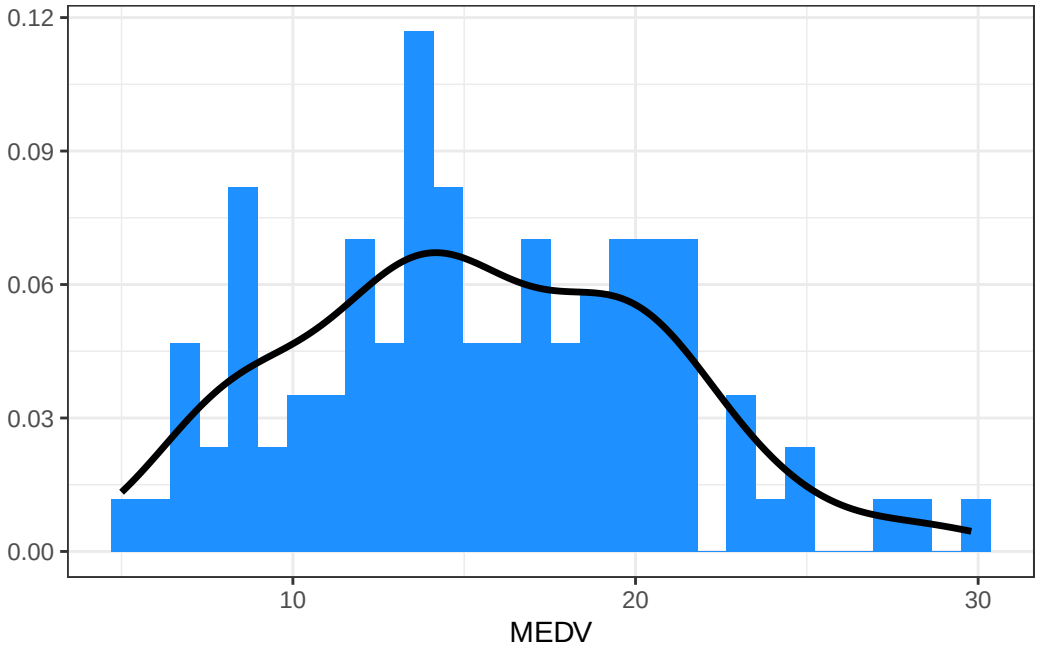


Figura 1: Distribución de la variable respuesta `MEDV`

Vemos que los valores de la variable `MEDV` se distribuyen no normales y en su mayoría agrupados por lo cual podríamos decir que no hay presencia de datos atípicos.

Prueba de normalidad

```
shapiro.test(datos4$MEDV)
```

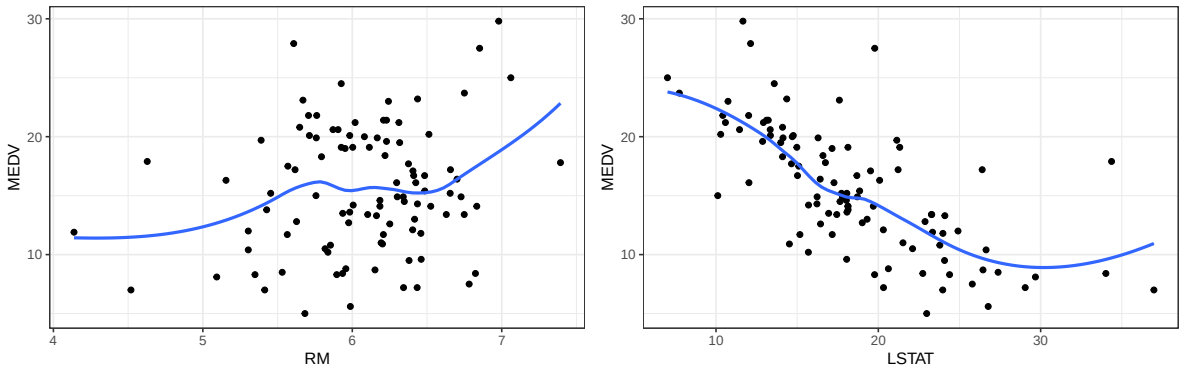
```
    Shapiro-Wilk normality test

data:  datos4$MEDV
W = 0.98536, p-value = 0.3368
```

Distribución: Como se nota en la prueba de shapiro, la variable MEDV se distribuye de manera normal, con algunos valores atípicos en el extremo superior. Esto significa que la mayoría de las viviendas tienen un valor medio cercano a la media, con algunas viviendas que tienen valores significativamente más altos.

Factores influyentes: El análisis sugiere que el valor medio de las viviendas (MEDV) está influenciado por varios factores, incluyendo el número de habitaciones (RM), el porcentaje de población de bajos ingresos (LSTAT), la tasa de criminalidad (CRIM) y posiblemente la edad de la vivienda (AGE). Sabiendo todo esto, podemos ajustar un modelo de regresión múltiple.

Con base en las observaciones anteriores graficaremos la variable`RM` y `LSTAT` usando un diagrama de dispersión, frente a la variable de respuesta `MEDV`.



Podemos notar que los precios de los hogares aumentan a medida que el valor de la variable `RM` aumenta "linealmente". Hay pocos valores atípicos y los datos parecen tener cierto limite en 9.

Los precios tienden a disminuir a medida que aumenta la variable `LSTAT`. aunque no parece seguir exactamente una tendencia lineal "cuadrática o exponencial".

3. Ajuste un modelo de regresión lineal múltiple, muestre la tabla de parámetros ajustados y escriba la ecuación ajustada. Calcule la Anova del modelo ¿Es significativo el modelo? ¿Qué proporción de la variabilidad total de la respuesta es explicada por el modelo? Opine sobre esto último.

```
# Modelo de regresión lineal múltiple con todas las variables
modelo <- lm(MEDV~.,datos4)
```

Coeficientes	Estimación	Error Estándar	Valor t	Pr(>|t|)
(Intercepto)	59.06933	30.51461	1.936	0.05593 .
CRIM	-0.08634	0.03139	-2.750	0.00715 **
NOX	-21.92000	6.51255	-3.366	0.00111 **
RM	0.26285	0.80549	0.326	0.74492
EDAD	-0.01242	0.02954	-0.421	0.67507
PTRATIO	-1.00533	1.59420	-0.631	0.52984
LSTAT	-0.46531	0.08208	-5.669	1.61e-07 *

El modelo de regresión lineal presentado busca predecir el valor medio de las viviendas (MEDV) en Boston, encontrando que la criminalidad (CRIM), la contaminación (NOX) y el nivel socioeconómico (LSTAT) son los predictores más importantes, con coeficientes negativos y significativos. Mientras que el número de habitaciones (RM), la edad de las viviendas (AGE) y la calidad de la educación (PTRATIO) no muestran una influencia significativa en este modelo. El intercepto (59.06933) representa el valor estimado de una vivienda cuando todas las variables predictoras son cero, pero debe interpretarse con cautela. Se podrían realizar mejoras al modelo, como eliminar variables no significativas, verificar la multicolinealidad, explorar otras variables relevantes (distancia a empleos, áreas verdes), considerar interacciones entre variables y evaluar posibles relaciones no lineales.

Ecuación del modelo

$$\hat{y}_0 = 59.06933 - 0.08634x_1 - 21.92000x_2 + 0.26285x_3 - 0.01242x_4 - 1.001242x_5 - 1.00533x_6 - 0.46531x_7$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prueba1	1	0.13	0.13	0.00	0.9628
Prueba2	1	1739.90	1739.90	31.08	0.0001
Prueba3	1	4138.43	4138.43	73.92	0.0000
Prueba4	1	608.37	608.37	10.87	0.0049
Residuals	15	839.72	55.98		

Tabla 1: Resultados de las pruebas estadísticas

Comentarios