

Mayo 2018

Cristina Gil Martínez

REGRESIÓN LINEAL MÚLTIPLE

Apuntes personales sobre regresión lineal múltiple

CONTENIDO

INTRODUCCIÓN.....	1
ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN.....	1
Precisión de los coeficientes de regresión.....	2
TEST DE HIPÓTESIS.....	3
SELECCIÓN DE VARIABLES.....	4
BONDAD DE AJUSTE DEL MODELO	5
R^2 ajustado	5
Error estándar residual (RSE)	6
PREDICTORES CUALITATIVOS.....	6
Predictores con dos niveles	6
Predictores con más de dos niveles.....	6
EXTENSIONES DEL MODELO LINEAL	7
Interacción de predictores.....	7
Regresión polinomial	8
PROBLEMAS POTENCIALES: VIOLACIÓN DE CONDICIONES	8
No linealidad en los datos.....	8
Correlación de los errores.....	9
Varianza no constante de los errores (heterocedasticidad).....	9
<i>Outliers</i> (valores atípicos)	10
Puntos con alta influencia (<i>high leverage</i>).....	10
Colinealidad	11
Tamaño de la muestra	12
comparación entre regresión lineal y <i>K-nearest neighbors</i>	12
EJEMPLO EN R.....	14
1. Exploración de los datos.....	14
2. Analizar relación entre variables	16
3. Generar el modelo.....	18
4. Selección de predictores	19
5. Validación de condiciones	21
6. Reajuste/mejora del modelo.....	25
BIBLIOGRAFÍA.....	27

INTRODUCCIÓN

La regresión lineal múltiple representa una extensión de la regresión lineal simple en la que podemos incluir más de un predictor a la vez.

En el caso de contar con más de una variable predictora, podríamos pensar en que una opción sería ajustar un modelo de regresión a cada uno por separado. Sin embargo, este enfoque puede no llegar a resultar del todo satisfactorio, ya que cada ecuación de regresión estaría ignorando las demás a la hora de estimar los coeficientes de regresión. Además, si se diera que los predictores estuvieran correlacionados entre sí ello podría llevar a estimaciones erróneas haciendo el ajuste por separado. Por tanto, una ventaja de la regresión lineal múltiple es que evalúa el efecto de cada predictor en presencia del resto, evitando el fenómeno de **confusión** que puede aparecer cuando la asociación observada entre un predictor y la variable respuesta se explica por otra variable (factor de confusión) de manera total o parcial.

Por tanto, en una ecuación de regresión lineal múltiple, se asociará cada predictor (X_1, X_2, \dots, X_p) un coeficiente β que cuantificará la asociación entre el predictor en cuestión y la variable respuesta Y :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

donde

β_j = efecto medio que tiene sobre Y el incremento en una unidad de X_j , manteniendo fijos el resto de predictores.

β_0 = ordenada en el origen, valor esperado de Y cuando todos los predictores son cero.

ϵ = residuo o error del modelo, diferencia entre lo observado y lo estimado.

ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN

Al igual que en el caso de regresión lineal simple, los verdaderos coeficientes $\beta_0, \beta_1, \dots, \beta_p$ son desconocidos, por lo que han de ser también estimados. Una vez estimados, podemos llevar a cabo las predicciones con la fórmula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

Al igual también que en la regresión lineal simple, se sigue el método de mínimos cuadrados para estimar estos coeficientes

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2$$

Precisión de los coeficientes de regresión

La imprecisión en los coeficientes de regresión estimados se relaciona con el error reducible (*bias* del modelo), a diferencia del error irreducible o aleatorio ϵ en el modelo, que se relaciona con el grado de incertidumbre asociado a cuánto difiere cada punto individual de la verdadera recta de regresión, o lo que es lo mismo, la diferencia entre lo observado y estimado por el modelo.

Error estándar (SE)

El error estándar se utiliza para estimar cómo de precisos son nuestros estimadores de los coeficientes β_0 ... β_j , y como, de media, difieren del valor de los verdaderos valores de los parámetros β_0 ... β_j .

[Ver explicación más detallada en el capítulo Regresión Lineal Simple]

Intervalo de confianza y de predicción

Mediante el cálculo del **intervalo de confianza** podemos obtener una estimación de qué exactitud tienen nuestros estimadores de los coeficientes, o el intervalo para el valor **medio** de Y dado un valor de X.

Por otra parte, el **intervalo de predicción** es una estimación del intervalo en el cual se encontrarán futuras observaciones, con una determinada probabilidad, dado lo que ya ha sido observado

$$\hat{y}(x^*) \pm t_{1-\alpha/2, n-2} SE_{\hat{y}(x^*)}$$

El intervalo de predicción es siempre más amplio que el de confianza, ya que incorporan a la vez el error de la estimación y el error irreducible ϵ .

TEST DE HIPÓTESIS

En un escenario de regresión múltiple con p predictores hacemos uso del test de hipótesis para determinar si todos los coeficientes de regresión son iguales o diferentes a 0.

Las hipótesis nula y alternativa corresponden a

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_a : \text{al menos un } \beta_j \text{ es diferente a } 0$$

Este test de hipótesis requiere de un análisis de varianza mediante el cálculo del estadístico F (F test), a partir del cual se obtiene el p -value a partir de una distribución F .

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

donde

$$TSS = \sum (y_i - \bar{y})^2$$

$(TSS - RSS)/p$ = **intervarianza** o medida de variabilidad entre grupos

$RSS/(n - p - 1)$ = **intravarianza** o estimación de la varianza residual no explicada

A diferencia del estadístico T , el estadístico F se ajusta al número de predictores, y reduce en gran medida que por azar se acepte una asociación falsa entre los predictores y la variable respuesta, probabilidad que aumenta al tener en cuenta cada estadístico t y su correspondiente p -value de cada coeficiente, puesto que por azar alrededor del 5% de los p -values asociados con cada variable se encontrarán por debajo de 0,05, es decir, aumenta el error de tipo I.

Cuanto más alejado esté el estadístico F de 1, mayor será la evidencia de que al menos uno de los predictores incluidos en el modelo es útil, rechazando la hipótesis nula. Cuando el tamaño muestral n es grande, un estadístico F algo mayor de 1 podría todavía dar muestra de evidencia en contra de H_0 . Por lo contrario, se necesitará un estadístico F mayor para rechazar H_0 si n es pequeño. Sin embargo, si el test F no resulta significativo no se puede aceptar el modelo como válido.

En el supuesto caso en el que estemos interesados un subgrupo q de coeficientes del modelo, utilizaríamos la hipótesis nula

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \dots = \beta_p = 0$$

En este caso ajustamos un segundo modelo que use todas las variables excepto el subgrupo q seleccionado. Suponiendo que la suma de residuos al cuadrado para ese modelo sea RSS_0 , el estadístico F se calcularía como

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n - p - 1)}$$

IMPORTANTE: Cuando el número de predictores es mayor que el número de observaciones ($p > n$), no es apropiado ajustar el modelo por mínimos cuadrados, por lo que el estadístico F tampoco lo es.

SELECCIÓN DE VARIABLES

Si el estadístico F del modelo resulta significativo, se considera que el modelo es útil y se puede proceder a identificar los predictores que sí contribuyen al modelo, ya que es muy frecuente que no todos contribuyan.

Son varios los estadísticos que se pueden utilizar para juzgar la calidad de modelos alternativos con distintos subgrupos de los predictores y elegir el mejor:

- C_p de Mallow
- Criterio de información de Akaike (AIC)
- Criterio de información bayesiano (BIC)
- R^2 ajustado

Desafortunadamente, existe un total 2^p modelos que contengan subgrupos de p predictores, por lo que incluso para un número moderado de p , evaluar cada combinación posible resulta inviable. Por esta razón, necesitamos un enfoque más eficiente para elegir un grupo de modelos más pequeños a tener en cuenta:

- Método **Forward**: se parte del modelo nulo (sin predictores, solo con β_0). A continuación se ajustan p modelos de regresión simple y se añade al modelo nulo la variable con la que se obtiene el menor RSS. Se van añadiendo variables una a una de esta forma hasta que alguna condición se satisfaga.
- Método **Backward**: se parte del modelo completo con todos los predictores, y se elimina el que tenga el mayor p -value (estadísticamente menos significativa). El nuevo modelo ($p - 1$) se reajusta y se continúan eliminando variables no significativas o hasta que cierta condición se satisfaga (por ejemplo, un valor límite de p -value establecido).
- **Selección mixta**: combinación entre el método *forward* y *backward*. Se parte del modelo nulo, y a medida que se añaden predictores, si en algún momento alguno de ellos deja de ser significativo, se elimina del modelo.

IMPORTANTE: el método de selección *backward* no puede usarse si $p > n$, mientras que el método de selección *forward* siempre se puede usar.

No hay garantías de que con todas las estrategias acabemos obteniendo el mismo modelo final. En este caso podemos guiarnos de qué modelo tiene mayor R^2_{ajustado} para seleccionar el mejor.

BONDAD DE AJUSTE DEL MODELO

Dos de las medidas más usadas para determinar como de bien se ajusta el modelo son el R^2 y RSE, calculadas e interpretadas de la misma manera que para la regresión lineal simple.

R^2 ajustado

En regresión lineal simple, el **coeficiente de determinación R^2** se corresponde al cuadrado de la correlación entre la variable respuesta y el predictor. En regresión lineal múltiple, se corresponde al cuadrado de la correlación entre la variable respuesta y el modelo lineal ajustado.

$$\text{Cor}(Y, \hat{Y})^2$$

El R^2 normal puede llegar a ser una estimación sesgada de la cantidad de variabilidad explicada por el modelo. Por cada predictor introducido en el modelo, R^2_{ajustado} introduce una penalización a R^2 , que depende de los grados de libertad, número de predictores y tamaño de la muestra. Por ello, R^2_{ajustado} ofrece una mejor estimación. Un valor de R^2_{ajustado} próximo a 1 también indica que el modelo es capaz de explicar una gran proporción de la varianza en la variable respuesta.

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n - d - 1)}{\text{TSS}/(n - 1)}$$

Un aspecto a tener en cuenta es que R^2 aumenta cuando más predictores son incluidos en el modelo, aun cuando éstos solo estuvieran levemente asociados con la variable respuesta (RSS siempre disminuye conforme más predictores se incluyan en el modelo). El hecho por ejemplo de que un predictor provoque solo un muy pequeño aumento en R^2 puede ser indicativo de que podría ser excluida del modelo, ya que no sería útil en explicar la variabilidad observada en Y . Por ello, R^2 no puede utilizarse para comparar modelos con distinto número de predictores.

Consejo: usar R^2_{ajustado} en lugar de los p -values para seleccionar el mejor modelo entre varios posibles.

Error estándar residual (RSE)

El error estándar residual es una estimación de la desviación estándar de la respuesta en relación a la recta de regresión poblacional. En regresión lineal múltiple, aquellos modelos que incluyen más variables pueden tener mayor RSE si la disminución en del RSS es pequeña en relación con el incremento de p .

$$\text{RSE} = \sqrt{\frac{1}{n - p - 1} \text{RSS}}$$

Las unidades en que se mide RSE son las mismas que las de la variable Y.

[Ver explicación en el capítulo [Regresión Lineal Simple](#)]

PREDICTORES CUALITATIVOS

No necesariamente todos los predictores en un modelo de regresión lineal han de ser **cuantitativos**, también se pueden incluir **cualitativos**.

Predictores con dos niveles

Si un factor o variable cualitativa tiene solo dos niveles o valores posibles, podemos incluirla en el modelo como una variable **dummy** (toma dos valores numéricos). La decisión de cómo codificar los niveles del factor es arbitraria (ej. 0 y 1, siendo el nivel codificado como 0 el nivel de referencia) y no tiene efecto en el ajuste de la regresión, pero sí determina la interpretación de los coeficientes.

El valor del coeficiente de correlación β_j correspondiente a un nivel de una variable **dummy** (codificado como 1) indica el promedio con el que influye dicho nivel sobre la variable respuesta en comparación con el nivel de referencia no codificado como variable **dummy** (β_0).

Predictores con más de dos niveles

En el caso de un predictor cualitativo con más de dos niveles, una sola variable **dummy** no puede representar todos los niveles posibles. En esta situación, podemos crear **variables dummy adicionales**. De nuevo, el nivel seleccionado como referencia es arbitrario.

EXTENSIONES DEL MODELO LINEAL

El modelo estándar de regresión lineal aporta resultados fácilmente interpretables, pero establece una serie de suposiciones restrictivas que a menudo no se cumplen en la práctica. Dos de las más importantes establecen que la relación entre los predictores y la variable Y es aditiva y lineal.

- **Aditividad:** el efecto de los cambios en el predictor X_j en la respuesta Y es independiente de los valores de otros predictores.
- **Linealidad:** el cambio en la respuesta Y debido a un cambio en una unidad de X_j es constante, independientemente del valor de X_j .

Interacción de predictores

De acuerdo con la ecuación estándar de regresión lineal con dos variables predictoras

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

si incrementamos X_1 en una unidad, entonces Y aumentará en β_1 unidades, de media, e independientemente de la presencia de X_2 y su valor. En este ejemplo, una manera de extender el modelo para permitir la efectos de interacción y relajar la condición de aditividad es incluir un tercer predictor o **término de interacción**, siendo en este caso el producto de X_1 y X_2 , lo cual resultaría en el modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$

En este caso, el efecto de X_1 sobre Y ya no es constante, depende del valor de X_2 .

Es importante seguir el principio de jerarquía que establece que si una interacción entre factores resulta significativa y se quiere introducir en el modelo, hay que introducir los factores participantes también por separado aun si no son significativos. Excluirlos podría alterar el significado de la interacción.

El concepto de interacción no es aplicable solo a variables cuantitativas, también a combinaciones entre variables cuantitativas y cualitativas.

Regresión polinomial

En algunos casos, la verdadera relación entre la variable respuesta y los predictores puede no ser lineal, por lo que podemos aplicar por ejemplo una regresión polinomial (existen métodos más complejos).

Una forma simple de incorporar asociaciones no lineales en un modelo lineal es incluir versiones transformadas de los predictores, elevándolos a distintas potencias, evitando un exceso de grados para evitar el sobreajuste o *overfitting*.

PROBLEMAS POTENCIALES: VIOLACIÓN DE CONDICIONES

Los siguientes casos han de ser evaluados a la hora de considerar válido un modelo lineal:

1. No linealidad entre predictores y variable respuesta
2. Correlación de los errores
3. Variabilidad no constante en los residuos (heterocedasticidad)
4. Outliers (valores atípicos)
5. Puntos con alta influencia (*high leverage*)
6. Colinealidad
7. Tamaño de la muestra

No linealidad en los datos

Una condición principal de un modelo lineal es que cada variable esté linealmente relacionada con la variable respuesta.

¿Cómo detectarla?

Graficar los residuos en función de los valores de Y predichos por el modelo puede ser un método útil para detectar falta de linealidad en los datos.

Posible solución:

Si se detectan patrones que indican una falta clara de linealidad, un enfoque simple sería el de usar transformaciones no lineales de los predictores, como $\log X$, \sqrt{X} , X^2 . Existen métodos más avanzados para estos casos.

Correlación de los errores

Una de las condiciones que asume el modelo de regresión lineal es que los errores $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ no estén correlacionados o que sean independientes. La correlación de los errores (errores adyacentes tienden a tener valores similares), que no solo, pero ocurre frecuentemente en el contexto de series temporales de datos, afecta directamente a la estimación del error estándar (SE) de los coeficientes de regresión (se subestiman, por lo que los intervalos de confianza y predicción tienden a ser más estrechos de lo que deberían). Si hay correlación entre los errores, los *p-values* de los predictores dejan de ser fiables (son más bajos de lo que deberían, por lo que podríamos considerar erróneamente significativos parámetros que en realidad no lo son).

¿Cómo detectarla?

Graficar los residuos en función del tiempo o registro de las observaciones. Ante la falta de correlación, no se debería diferenciar ningún patrón claro en el gráfico.

Varianza no constante de los errores (heterocedasticidad)

Otra de las suposiciones de la regresión lineal es la varianza constante de los errores. De esto dependen el error estándar, los intervalos de confianza y tests de hipótesis.

A menudo se da el caso de la varianza de los errores no es constante, por ejemplo un aumento de la varianza con valores ajustados más altos en \hat{Y} .

¿Cómo detectarla?

Graficar los residuos frente a los valores ajustados por el modelo, e identificar si existe un patrón cónico u otro patrón. Idealmente deberían distribuirse de forma aleatoria en torno a 0. También podemos recurrir al **test de Breusch-Pagan** como contraste de homocedasticidad. La hipótesis nula de este test es que los residuos poseen una varianza constante.

✓ Posible solución:

Transformar la respuesta Y con una función cóncava como por ejemplo $\log Y$ o \sqrt{Y} . Este tipo de transformaciones ayudan a encoger en mayor medida los valores mayores en la respuesta Y , reduciendo la heterocedasticidad.

(Alternativa: *weighted least squares*)

Outliers (valores atípicos)

Un **outlier** o valor atípico es un valor de y_i que se encuentra alejado del valor predicho por el modelo. Pueden aparecer por varias razones, como por ejemplo debido a un error a la hora de la obtención del dato, pero también pueden ser casos interesantes o deberse por ejemplo a predictores ausentes. Aunque los *outliers* puedan no tener gran efecto en la recta de regresión, pueden llegar a causar otros problemas.

¿Cómo detectarlos?

Graficar los residuos puede ayudarnos a identificar *outliers*. Podemos también evaluar cómo cambia el modelo y los valores de RSE y R^2 cuando se excluyen del modelo.

En la práctica, nos puede resultar difícil decidir cómo de grande un residuo ha de ser para considerar ese dato como un valor atípico. Para ello es útil graficar los **residuos estudentizados** (*studentized residuals*), que se calculan dividiendo cada residuo entre su error estándar. Aquellas observaciones con residuos estudentizados > 3 serán considerados como posibles *outliers*.

Puntos con alta influencia (*high leverage*)

Si los *outliers* son puntos con un valor inusual de y_i , los **puntos influyentes** tienen valores inusuales de x_i . Sin embargo, un *outlier* puede ser a la vez un punto influyente, siendo esta una combinación peligrosa para la validez del modelo. Es decir, el valor del predictor es grande/pequeño en relación con el resto de observaciones. Los puntos influyentes suelen tener mayor efecto sobre la recta de regresión, por ello es importante identificarlos.

¿Cómo detectarlos?

Un **gráfico de los residuos estudentizados vs el grado de influencia** o *leverage* para cada observación es muy útil para identificar observaciones que son *outliers*, puntos influyentes o ambas cosas a la vez. Para cuantificar el grado de influencia de una observación para el caso de una regresión lineal simple, se calcula el estadístico de influencia

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x_{i'} - \bar{x})^2}$$

h_i aumenta con el aumento de la distancia entre la observación x_i y la media de x . Su valor siempre se encuentra entre $1/n$ y 1. Un valor alto de este estadístico es indicativo de que la observación para el que se ha calculado es influyente.

Colinealidad

La **colinealidad** se refiere a la situación en la que dos o más predictores están estrechamente relacionados uno con otro. Se dice que dos predictores correlacionados son colineales. La presencia de colinealidad puede suponer un problema, pues puede dificultar la separación del efecto individual de variables colineales sobre la variable respuesta. Además, la colinealidad provoca el aumento del error estándar y reduce la precisión de las estimaciones de los coeficientes de regresión, reduciendo la probabilidad de considerarlos significativos cuando realmente lo son (disminuye la potencia del test de hipótesis). Esto se explica porque el estadístico t para cada predictor se calcula dividiendo β_j entre el error estándar, y si el error estándar aumenta, disminuye el estadístico t .

¿Cómo detectarla?

Una forma simple de detectar colinealidad es inspeccionar una **matriz de correlación de los predictores**. Sin embargo, con esta matriz no podríamos detectar colinealidad entre más de dos predictores (multicolinealidad). En este último caso lo que ayudaría sería calcular el **factor de inflación de la varianza (VIF)**. Su valor menor posible es 1. Para cada variable puede calcularse como

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_{X_j|X_{-j}}^2}$$

donde $R_{X_j}^2$ es la R^2 de la regresión de X_j sobre el resto de predictores.

VIF = 1 (ausencia de colinealidad)

$1 < VIF < 5$ (cierta colinealidad)

$5 < VIF < 10$ (alta colinealidad)

Otras opciones son generar un **modelo de regresión lineal simple** entre cada predictor frente a los demás y detectar si en algún caso el R^2 es alto.

✓ Posible solución:

Se podría eliminar una de las variables problemáticas ya que la presencia de colinealidad implica que la información que esta variable proporciona sobre la respuesta Y es redundante en presencia de otra variable/s. Otra opción sería combinar las variables colineales en un solo predictor (aunque se corre el riesgo de perder su interpretación).

Tamaño de la muestra

Tener en cuenta que si no disponemos de suficientes observaciones, predictores que no son realmente influyentes lo podrían parecer. Se recomienda que el número de observaciones sea en torno a 10 – 20 veces el número de predictores. Cuantas más observaciones, más predictores se pueden incorporar.

COMPARACIÓN ENTRE REGRESIÓN LINEAL Y K-NEAREST NEIGHBORS

La regresión lineal es un ejemplo de **método paramétrico** que asume una función específica lineal para $f(X)$. Las ventajas de los métodos paramétricos es que son fáciles de ajustar ya que solo es necesario estimar un cierto número de coeficientes. En el caso de la regresión lineal, estos coeficientes son fáciles de interpretar, y los test de significancia estadística son fáciles de llevar a cabo. Sin embargo, también cuentan con desventajas, como el hecho de que asumen ciertas condiciones sobre la forma de $f(X)$. Si dicha $f(X)$ se aleja de la realidad y el objetivo es la predicción, el método paramétrico no dará buenos resultados ni podremos obtener conclusiones fiables de él. Por el contrario, los **métodos no paramétricos** no asumen de manera explícita una forma particular de $f(X)$, por lo que suponen una alternativa más flexible para llevar a cabo la regresión. Aun así, el método paramétrico superará al no paramétrico si la función seleccionada es próxima a la verdadera forma de f , además de que los métodos no paramétricos sufren las consecuencias de la varianza.

Uno de los métodos no paramétricos más simples es la **regresión KNN** o K vecinos más cercanos (*K-nearest neighbors*). El método de regresión KNN está estrechamente relacionado con el clasificador KNN. Dado un valor para K y un valor del predictor x_0 , el método de regresión KNN primero identifica las K observaciones de entrenamiento que se encuentran más cerca a x_0 (representado por \mathcal{N}_0). Se estima $f(x_0)$ usando la media de todos los valores en \mathcal{N}_0 .

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

De manera general, el valor a escoger de K dependerá del equilibrio bias-varianza. Un valor de K pequeño proporciona el ajuste más flexible, con lo que tendremos un bajo bias pero una alta varianza. Esto se debe a que la predicción en una determinada región dependerá únicamente de una observación. Por lo contrario, valores altos de K proporcionan un ajuste menos variable (la predicción dependerá de la media de varios valores). Podemos utilizar el MSE del test set para identificar el valor óptimo de K.

Los resultados de KNN son ligeramente peores que la regresión lineal cuando la relación es lineal, pero es mucho mejor ante la falta de linealidad. Tener en cuenta que un problema común que sufre KNN es que a medida que aumentan los predictores, KNN suele dar peores resultados que la regresión lineal aun cuando hay falta de linealidad. Una razón es que a mayores dimensiones se requiere de más observaciones. Este problema se conoce como *curse of dimensionality*. Por ejemplo, las K observaciones más próximas a una nueva observación x_0 pueden llegar a estar muy lejos en un determinado espacio p -dimensional cuando p es alto, lo cual lleva a una mala predicción de $f(x_0)$ y a un mal ajuste KNN.

Además, si el test MSE es ligeramente peor en la regresión KNN en comparación con la regresión lineal, podemos preferir escoger esta última ya que nos aporta una mayor facilidad de interpretar el modelo.

De manera general, los métodos paramétricos tenderán a dar mejores resultados cuando hay un número bajo de observaciones por predictor.

EJEMPLO EN R

En el siguiente ejemplo empleamos el set de datos *Auto* que contiene información sobre vehículos para ajustar un modelo de regresión lineal múltiple para predecir *mpg* o consumo de combustible (millas/galón) en función del resto de variables.

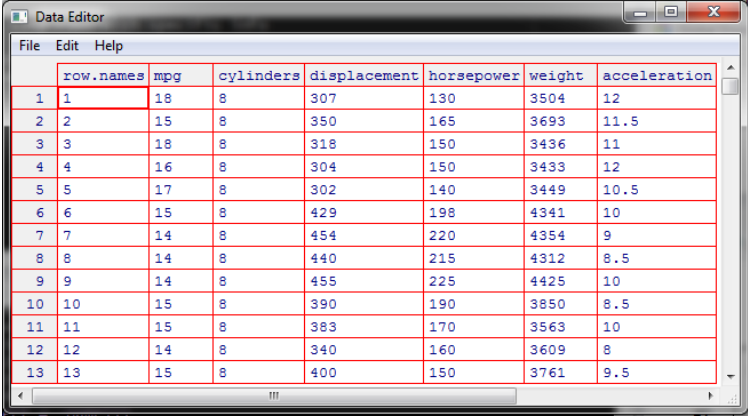
1. Análisis exploratorio de los datos

- **fix()** -> Muestra la matriz de datos en una nueva ventana
- **str()** -> Muestra de manera compacta la estructura interna de un objeto
- **summary()** -> Resumen de los datos
- **pairs()** -> Matriz con gráficos de dispersión de las variables cuantitativas

Como primer paso realizamos una exploración de los datos:

```
library(MASS)
library(ISLR)

fix(Auto)
```



	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration
1	1	18	8	307	130	3504	12
2	2	15	8	350	165	3693	11.5
3	3	18	8	318	150	3436	11
4	4	16	8	304	150	3433	12
5	5	17	8	302	140	3449	10.5
6	6	15	8	429	198	4341	10
7	7	14	8	454	220	4354	9
8	8	14	8	440	215	4312	8.5
9	9	14	8	455	225	4425	10
10	10	15	8	390	190	3850	8.5
11	11	15	8	383	170	3563	10
12	12	14	8	340	160	3609	8
13	13	15	8	400	150	3761	9.5

```
str(Auto)
```

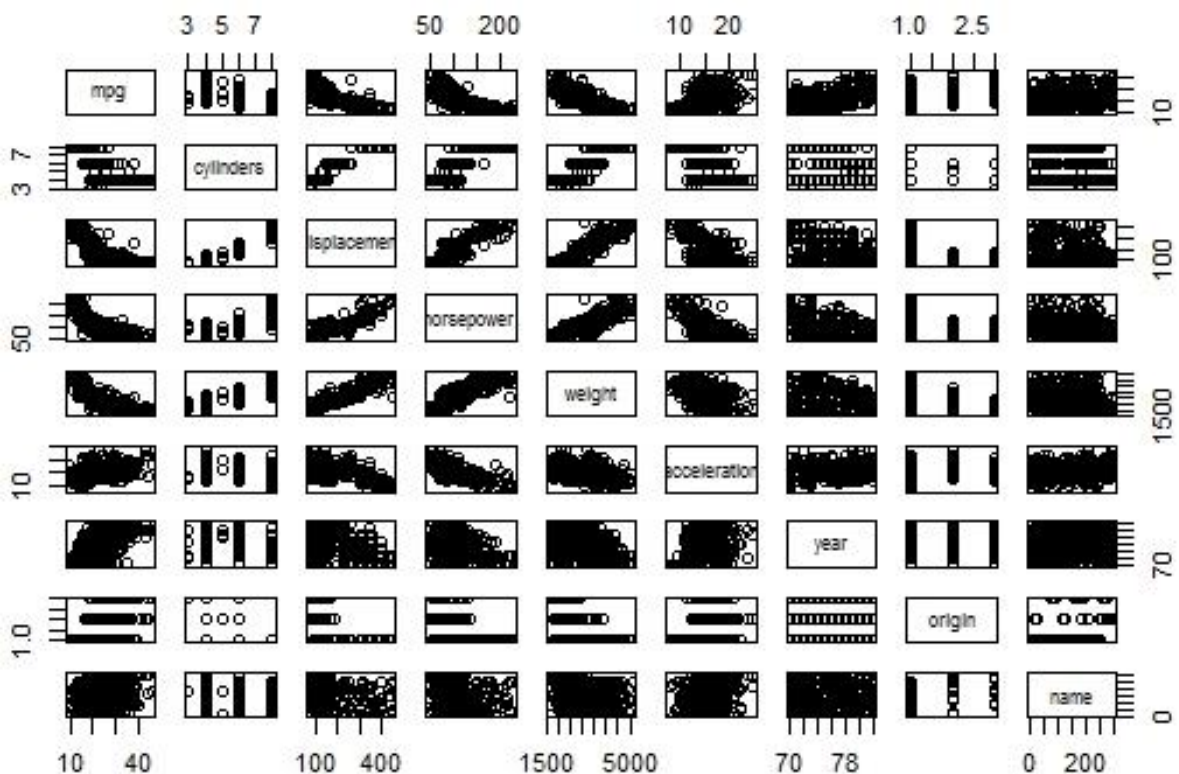
```
## 'data.frame':  392 obs. of  9 variables:
## $ mpg      : num  18 15 18 16 17 15 14 14 15 ...
## $ cylinders : num   8  8  8  8  8  8  8  8  8 ...
## $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
## $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
## $ weight      : num  3504 3693 3436 3433 3449 ...
## $ acceleration: num   12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
## $ year        : num   70  70  70  70  70  70  70  70  70  70 ...
## $ origin      : num    1  1  1  1  1  1  1  1  1  1 ...
```

```
## $ name : Factor w/ 304 levels "amc ambassador brougham",...: 49 36 231 1
4 161 141 54 223 241 2 ...
```

```
summary(Auto)
```

```
##      mpg      cylinders  displacement  horsepower      weight
##  Min.   : 9.00    Min.   :3.000    Min.   : 68.0    Min.   : 46.0    Min.   :1613
## 1st Qu.:17.00    1st Qu.:4.000    1st Qu.:105.0    1st Qu.: 75.0    1st Qu.:2225
##  Median :22.75    Median :4.000    Median :151.0    Median : 93.5    Median :2804
##  Mean   :23.45    Mean   :5.472    Mean   :194.4    Mean   :104.5    Mean   :2978
## 3rd Qu.:29.00    3rd Qu.:8.000    3rd Qu.:275.8    3rd Qu.:126.0    3rd Qu.:3615
##  Max.   :46.60    Max.   :8.000    Max.   :455.0    Max.   :230.0    Max.   :5140
##
##  acceleration      year      origin
##  Min.   : 8.00    Min.   :70.00    Min.   :1.000
## 1st Qu.:13.78    1st Qu.:73.00    1st Qu.:1.000
##  Median :15.50    Median :76.00    Median :1.000
##  Mean   :15.54    Mean   :75.98    Mean   :1.577
## 3rd Qu.:17.02    3rd Qu.:79.00    3rd Qu.:2.000
##  Max.   :24.80    Max.   :82.00    Max.   :3.000
##
##      amc matador : 5
##      ford pinto  : 5
##      toyota corolla : 5
##      amc gremlin  : 4
##      amc hornet   : 4
##      chevrolet chevette: 4
##      (Other)      :365
```

```
# Matriz con gráficos de dispersión de todas las variables del set de datos
pairs(Auto)
```



2. Analizar relación entre variables

- `cor()` -> Calcula los coeficientes de correlación entre variables (solo acepta vectores numéricos)
- `corrplot()` -> Representación gráfica de los coeficientes de correlación
- `ggpairs()` -> Combina en un único gráfico diagramas de dispersión, distribución de las variables y los valores de correlación.

A continuación, estudiamos la relación entre las variables para identificar cuáles pueden ser los mejores predictores o si hay alguna con una relación de tipo no lineal o detectar indicios de colinealidad. Excluimos la variable cualitativa *name*.

Matriz de correlación entre predictors

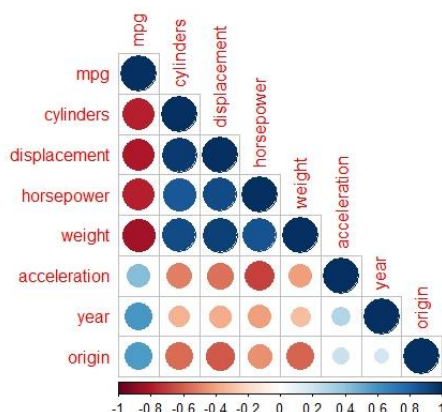
```
round(cor(subset(Auto, select = -name), method = "pearson"), digits = 3)
```

```
##           mpg cylinders displacement horsepower weight acceleration
## mpg          1.000    -0.778      -0.805      -0.778 -0.832      0.423
## cylinders    -0.778     1.000       0.951       0.843  0.898     -0.505
## displacement -0.805     0.951       1.000       0.897  0.933     -0.544
## horsepower   -0.778     0.843       0.897       1.000  0.865     -0.689
## weight        -0.832     0.898       0.933       0.865  1.000     -0.417
## acceleration  0.423    -0.505      -0.544      -0.689 -0.417      1.000
## year          0.581    -0.346      -0.370      -0.416 -0.309      0.290
## origin        0.565    -0.569      -0.615      -0.455 -0.585      0.213
##
##           year origin
## mpg          0.581  0.565
## cylinders    -0.346 -0.569
## displacement -0.370 -0.615
## horsepower   -0.416 -0.455
## weight        -0.309 -0.585
## acceleration  0.290  0.213
## year          1.000  0.182
## origin        0.182  1.000
```

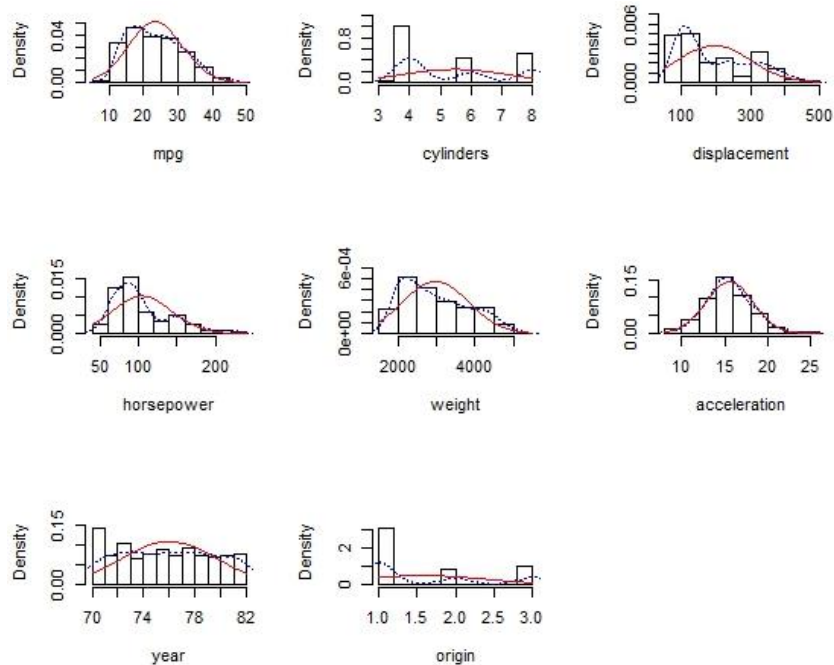
Valores de correlación r próximos a 1 o -1 indican una alta correlación entre variables. También podemos representarlos gráficamente con la siguiente función:

```
require(corrplot)
```

```
corrplot(round(cor(subset(Auto, select = -name)), digits = 3), type = "lower")
```

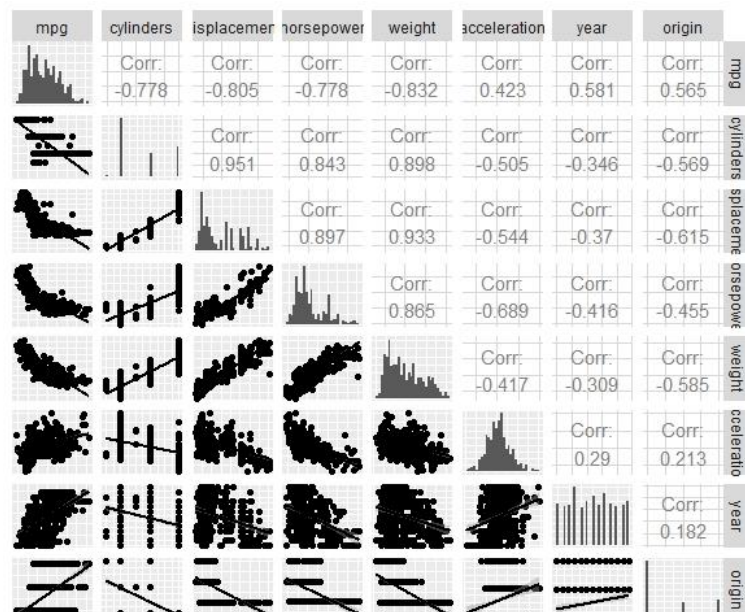


```
# Distribución de densidad de las variables cuantitativas del modelo
library(dplyr)
require(psych)
multi.hist(x = select(Auto, - name), dcol = c("blue", "red"),
           dltty = c("dotted", "solid"), main = "" )
```



El siguiente paquete permite combinar los diagramas de dispersión, la distribución y los valores de correlación:

```
Library(dyplr)
require(GGally)
ggpairs(select(Auto, - name), lower = list(continuous = "smooth"),
        diag = list(continuous = "bar"), axisLabels = "none")
```



De lo analizado hasta ahora podemos concluir que:

- I. Las variables que mayor relación (no siendo del todo lineal) tienen con *mpg* son: *displacement* ($r = -0,8$), *weight* ($r = -0,83$), *horsepower* ($r = -0,77$) y *cylinders* ($r = -0,77$), siendo la relación en todas, negativa.
- II. Se observa una alta correlación (colinealidad) entre pares de variables como *displacement* y *cylinders* ($r = 0,95$) y *displacement* y *weight* ($r = 0,93$). Con ello, posiblemente no sería útil introducir ambos pares en el modelo.
- III. La distribución de las variables parece acercarse bastante a una distribución normal, dado el número de observaciones con las que disponemos.

3. Generar el modelo

- `lm(y ~ x1 + x2 + x3 ..., data)` -> Modelo de regresión lineal múltiple
- `lm(y ~ x1 + x2 + x3...+0, data)` -> Modelo sin ordenada en el origen
- `lm(y ~ 1, data)` -> Modelo nulo (sin predictores). La predicción es la media de y .
- `summary(modelo)` -> Resumen del ajuste del modelo
- `contrasts(variable)` -> Variables dummy que R ha generado para cada nivel de una variable cualitativa

Vamos a generar el modelo con todos los predictores a excepción de la variable *name* que proporciona el nombre del modelo del coche, y que en este caso es prescindible ya que no aporta información importante al modelo. R generaría variables *dummy* automáticamente para las variables cualitativas. Con la función `contrasts()` podríamos conocer qué valor R ha asociado a cada nivel del factor.

```
modelo.lineal <- lm(mpg ~ . - name, data = Auto)
summary(modelo.lineal)

##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.218435   4.644294  -3.707  0.00024 ***
## cylinders    -0.493376   0.323282  -1.526  0.12780
## displacement  0.019896   0.007515   2.647  0.00844 **
## horsepower   -0.016951   0.013787  -1.230  0.21963
## weight       -0.006474   0.000652  -9.929 < 2e-16 ***
## acceleration  0.080576   0.098845   0.815  0.41548
## year         0.750773   0.050973  14.729 < 2e-16 ***
```

```
## origin          1.426141    0.278136    5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

De lo analizado hasta ahora podemos concluir que:

- I. El modelo con todas las variables introducidas como predictores es capaz de explicar el 82,15% de la varianza observada en el consumo de combustible ($R^2_{\text{ajustado}} = 0,818$).
- II. El p -value del modelo es significativo ($2,2e-16$), por lo que podemos decir que el modelo es útil y que existe una relación entre los predictores y la variable respuesta (al menos uno de los coeficientes es distinto a 0).
- III. Los predictores que parecen tener una relación estadísticamente significativa con la variable respuesta son: *displacement*, *weight*, *year* y *origin*, a diferencia de *cylinders*, *horsepower*, y *acceleration*.
- IV. Ejemplo de interpretación de coeficiente: por cada año que pasa, se recorre más distancia por volumen de combustible ($\beta_{\text{year}} = 0,75$) manteniéndose el resto de predictores constante, es decir, aumenta la eficiencia.

4. Selección de predictores

- **step(modelo)** -> Selección de predictores por stepwise y AIC
- **update()** -> Reajustar y actualizar el modelo
- **confint(modelo)** -> Intervalos de confianza para los coeficientes del modelo

En este ejemplo vamos a utilizar el método del *stepwise* mixto y el AIC con la función `step()` para determinar la calidad del modelo:

```
step(modelo.lineal, direction = "both", trace = 1)

## Start:  AIC=950.5
## mpg ~ (cylinders + displacement + horsepower + weight + acceleration +
##        year + origin + name) - name
##
##              Df Sum of Sq    RSS    AIC
## - acceleration  1      7.36 4259.6  949.18
## - horsepower    1     16.74 4269.0  950.04
## <none>              4252.2  950.50
## - cylinders     1     25.79 4278.0  950.87
```

```
## - displacement 1      77.61 4329.8  955.59
## - origin       1      291.13 4543.3  974.46
## - weight       1     1091.63 5343.8 1038.08
## - year         1     2402.25 6654.5 1124.06
##
## Step: AIC=949.18
## mpg ~ cylinders + displacement + horsepower + weight + year +
##      origin
##
##              Df Sum of Sq    RSS    AIC
## <none>                4259.6  949.18
## - cylinders          1      27.27 4286.8  949.68
## + acceleration       1       7.36 4252.2  950.50
## - horsepower         1      53.80 4313.4  952.10
## - displacement       1      73.57 4333.1  953.89
## - origin             1     292.02 4551.6  973.17
## - weight             1    1310.43 5570.0 1052.32
## - year              1    2396.17 6655.7 1122.13
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year + origin, data = Auto)
##
## Coefficients:
## (Intercept)      cylinders  displacement  horsepower      weight
##   -15.563492   -0.506685      0.019269   -0.023895   -0.006218
##      year      origin
##    0.747516    1.428242
```

acceleration (la variable con mayor *p-value*) ha sido la única variable excluida en el proceso de selección.

Reajustamos el modelo excluyendo dicha variable:

```
modelo.lineal2 <- update(modelo.lineal, formula = ~ . -acceleration)
summary(modelo.lineal2)

##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year + origin, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.7604 -2.1791 -0.1535  1.8524 13.1209
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.556e+01  4.175e+00  -3.728 0.000222 ***
## cylinders    -5.067e-01  3.227e-01  -1.570 0.117236
## displacement  1.927e-02  7.472e-03   2.579 0.010287 *
## horsepower   -2.389e-02  1.084e-02  -2.205 0.028031 *
## weight       -6.218e-03  5.714e-04 -10.883 < 2e-16 ***
## year         7.475e-01  5.079e-02  14.717 < 2e-16 ***
## origin       1.428e+00  2.780e-01   5.138 4.43e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.326 on 385 degrees of freedom
## Multiple R-squared:  0.8212, Adjusted R-squared:  0.8184
## F-statistic: 294.6 on 6 and 385 DF,  p-value: < 2.2e-16
```

Tras excluir la variable *acceleration*, R^2 apenas se ha modificado, lo que indica que la variable excluida no ayudaba en gran medida a explicar la variabilidad de la variable Y.

Los intervalos de confianza para cada uno de los coeficientes serían:

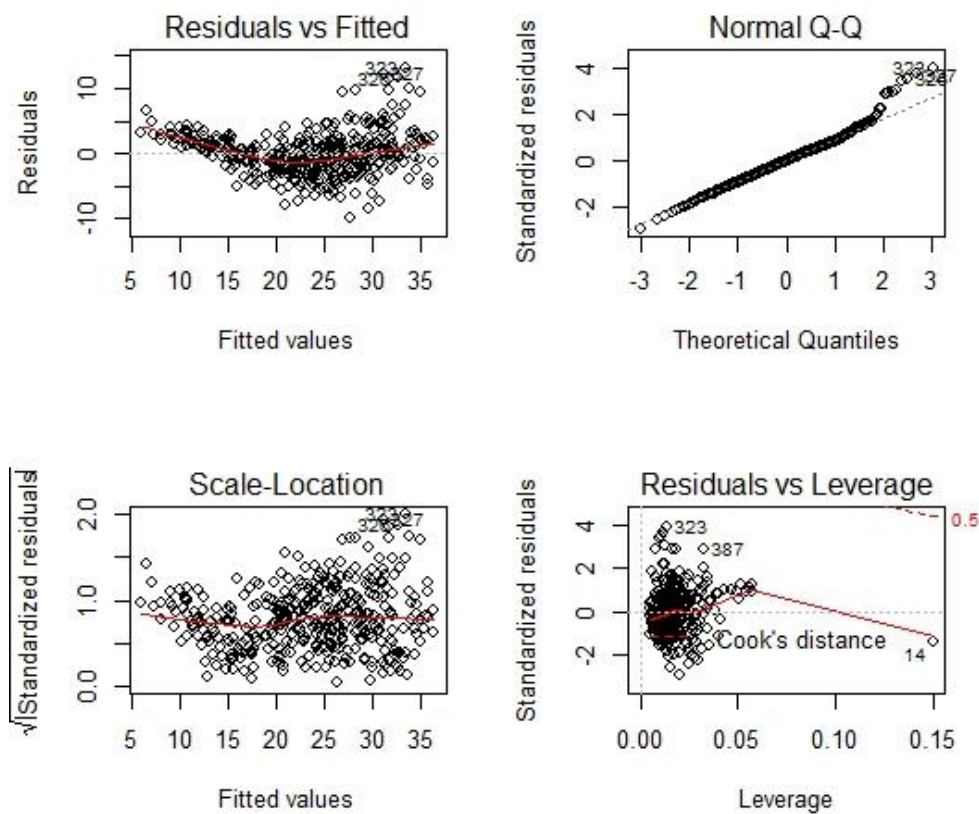
```
confint(modelo.lineal2)

##                2.5 %        97.5 %
## (Intercept) -23.772628686 -7.354355925
## cylinders   -1.141217264  0.127846990
## displacement 0.004577392  0.033961179
## horsepower   -0.045199801 -0.002590258
## weight       -0.007341708 -0.005094914
## year         0.647647254  0.847384650
## origin       0.881647846  1.974835924
```

5. Validación de condiciones

- `plot(modelo)` -> Análisis de los residuos (distribución, variabilidad...)
- `shapiro.test(modelo$residuals)` -> Test de hipótesis de Shapiro Wilk para el análisis de normalidad
- `plot(predict(modelo), rstudent(modelo))` -> Residuos estudentizados para detección de outliers o puntos influyentes
- `bptest(modelo)` -> Test de contraste de homocedasticidad Breusch-Pagan
- `influence.measures(modelo)` -> Detección de observaciones influyentes
- `influencePlot(modelo)` -> Visualización de observaciones influyentes
- `outlierTest(modelo)`` -> Test de detección de *outliers*
- `vif(modelo)` -> Calcula VIFs (factor de inflación de la varianza)

```
par(mfrow=c(2,2))
plot(modelo.lineal2)
```

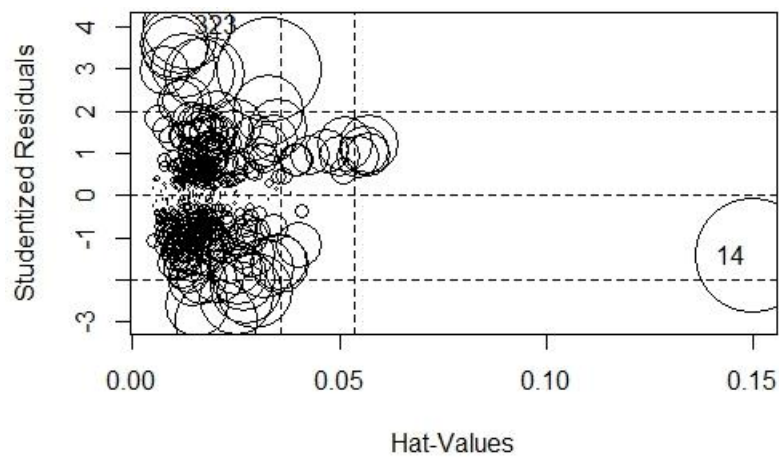



Detección y visualización de observaciones influyentes

`require(car)`

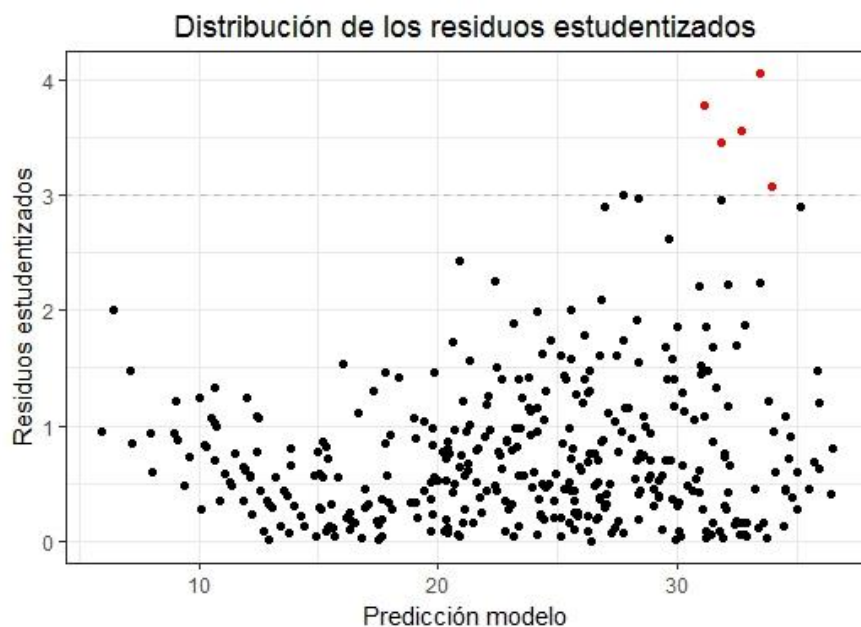
`influencePlot(modelo.lineal2)`

```
##      StudRes      Hat      CookD
## 14  -1.416771 0.15008894 0.05050589
## 323  4.049553 0.01317737 0.03007971
```



```
library(ggplot2)

# Gráfico residuos estudentizados frente a valores ajustados por el modelo
ggplot(data = Auto, aes(x = predict(modelo.lineal2),
                        y = abs(rstudent(modelo.lineal2))))+
  geom_hline(yintercept = 3, color = "grey", linetype = "dashed")+
  # se identifican en rojo las observaciones con residuos estandarizados absolutos > 3
  geom_point(aes(color = ifelse(abs(rstudent(modelo.lineal2)) > 3, "red", "black")))+
  +
  scale_color_identity()+
  labs(title = "Distribución de los residuos estudentizados",
       x = "Predicción modelo",
       y = "Residuos estudentizados")+
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```



```
# Detección de los residuos estudentizados > 3 considerados como outliers
which(rstudent(modelo.lineal2) > 3)

## 245 323 326 327 394
## 243 321 324 325 389

outlierTest(modelo.lineal2)

##      rstudent unadjusted p-value Bonferonni p
## 323 4.049553      6.2098e-05      0.024343
```

```
# Test de hipótesis para el análisis de normalidad de Los residuos
shapiro.test(modelo.lineal2$residuals)
```

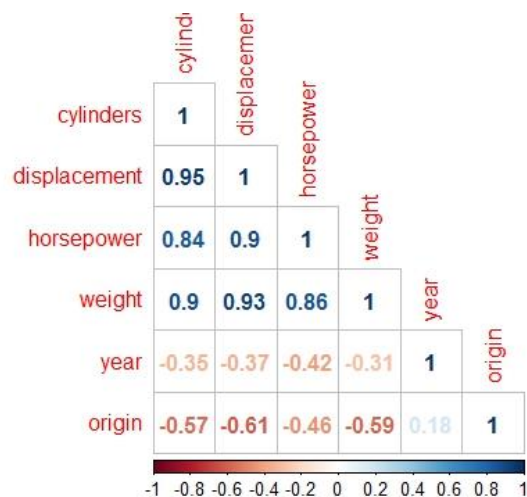
```
##
## Shapiro-Wilk normality test
##
## data:  modelo.lineal2$residuals
## W = 0.97461, p-value = 2.327e-06
```

```
# Test de contraste de homocedasticidad Breusch-Pagan
library(lmtest)
bptest(modelo.lineal2)
```

```
##
## studentized Breusch-Pagan test
##
## data:  modelo.lineal2
## BP = 23.063, df = 6, p-value = 0.0007755
```

Como hemos visto en el segundo paso del análisis, hay evidencias de alta colinealidad entre algunas variables. Podríamos utilizar la función `vif()` para calcular el factor de inflación de la varianza y detectar variables con mayor colinealidad.

```
corrplot(cor(select(Auto, cylinders, displacement, horsepower, weight, year,
                    origin)), method = "number", type = "lower")
```



```
# Factores de inflación de la varianza
vif(modelo.lineal2)
```

```
## cylinders displacement horsepower weight year origin
## 10.710150 21.608513 6.147752 8.324047 1.237304 1.772234
```

Hasta el momento podemos concluir que:

- I. El ajuste lineal parece no ser del todo preciso, ya que se observa un patrón curvo en los residuos frente a los valores ajustados por el modelo, además de que no acaban de distribuirse de forma homogénea en torno a 0. El test de Breusch-Pagan también proporciona evidencias de falta de homocedasticidad ($p\text{-value} = 0,0007$).
- II. El Q-Q plot refleja que hay indicios de falta de normalidad en los residuos (aquellos de mayor valor), corroborado también por el test de hipótesis de Shapiro Wilk ($p\text{-value} = 2,32\text{e-}06$).
- III. La observación 14 parece tener un nivel alto de influencia, aunque no se considere como residuo de alta magnitud. La observación 323 también se considera influyente. Un análisis más exhaustivo consistiría en excluir las observaciones y ver el impacto sobre el modelo.
- IV. Los predictores *cylinders* y *displacement* muestran una alta inflación de la varianza.
- V. Cuatro de las seis variables que incluye el modelo están muy correlacionadas.

6. Reajuste/mejora del modelo

- `poly(x = , degree =)` -> Añadir términos polinómicos al modelo
- `*` , `:` -> Interacción entre variables
- `log(x)` , \sqrt{x} , x^2 -> Transformaciones de variables
- `anova(modelo1, modelo2)` -> Comparar modelos (anidados)

Ya que algunas de las condiciones para el ajuste lineal no acaban de satisfacerse, y observando la matriz de correlación podemos ver como la distribución de las variables *horsepower*, *displacement* y *weight* tiene un patrón no lineal parecido frente a *mpg*, podríamos aproximar el ajuste utilizando un polinomio de grado 2. En el siguiente intento podemos incluir términos polinómicos a estas variables y estudiar si el modelo mejora. Es importante no excederse en el grado de polinomio para evitar el “*overfitting*”, ya que cuanto mayor es el polinomio, más flexible es el modelo.

```
modelo.lineal.poli <- update(modelo.lineal2, formula = ~ . + poly(displacement, 2)
                             + poly(horsepower, 2) + poly(weight, 2))
summary(modelo.lineal.poli)

##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     year + origin + poly(displacement, 2) + poly(horsepower,
##     2) + poly(weight, 2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.0799 -1.5267 -0.0789  1.4437 11.8994
```

```
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.957e+01  3.671e+00  -5.332 1.67e-07 ***
## cylinders      3.633e-01  3.271e-01   1.110 0.267519
## displacement  -2.480e-03  7.493e-03  -0.331 0.740795
## horsepower    -4.358e-02  1.043e-02  -4.177 3.66e-05 ***
## weight        -4.710e-03  5.739e-04  -8.206 3.54e-15 ***
## year          7.790e-01  4.487e-02  17.362 < 2e-16 ***
## origin        5.704e-01  2.674e-01   2.133 0.033561 *
## poly(displacement, 2)1      NA         NA         NA      NA
## poly(displacement, 2)2  1.223e+01  6.397e+00   1.912 0.056593 .
## poly(horsepower, 2)1      NA         NA         NA      NA
## poly(horsepower, 2)2  1.466e+01  4.326e+00   3.388 0.000777 ***
## poly(weight, 2)1          NA         NA         NA      NA
## poly(weight, 2)2  1.599e+01  4.670e+00   3.424 0.000683 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 382 degrees of freedom
## Multiple R-squared:  0.8643, Adjusted R-squared:  0.8611
## F-statistic: 270.3 on 9 and 382 DF,  p-value: < 2.2e-16
```

```
# Test de hipótesis para evaluar si un modelo se ajusta mejor que el otro
anova(modelo.lineal2, modelo.lineal.poli)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cylinders + displacement + horsepower + weight + year +
##      origin
## Model 2: mpg ~ cylinders + displacement + horsepower + weight + year +
##      origin + poly(displacement, 2) + poly(horsepower, 2) + poly(weight,
##      2)
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     385 4259.6
## 2     382 3232.8   3    1026.7 40.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONCLUSIÓN:

Incluyendo términos polinómicos (siendo en *displacement* menos significativo) hemos conseguido mejorar el modelo y que explique casi un 5% más de la variabilidad ($R^2_{\text{ajustado}} = 0,8611$ y $p\text{-value}$ de ANOVA = $2.2e-16$). Las observaciones 323 y 14 podrían estar influyendo en el modelo.

Modelo lineal múltiple:

$$\text{mpg} = -19,57 + 0,363\text{cylinders} - 2,4 \times 10^{-3}\text{displacement} - 4,358 \times 10^{-2}\text{horsepower} + 14,66\text{horsepower}^2 - 4,71 \times 10^{-3}\text{weight} + 15,99\text{weight}^2 + 0,779\text{year} + 0,57\text{origin}$$

BIBLIOGRAFÍA

OpenIntro Statistics: Third Edition, David M Diez, Christopher D Barr, Mine Çetinkaya-Rundel

An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)