Introducción a la Regresión Lineal Múltiple

Joaquín Amat Rodrigo *j.amatrodrigo@gmail.com*

Julio, 2016

Índice

Introducción

La información aquí presente se ha obtenido principalmente de OpenIntro Statistics, Tutorials by William B. King Coastal Carolina Universit y de Introduction to Statistical Learning. En este último libro se puede encontrar información mucho más detallada sobre la regresión lineal múltiple.

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (*Y*) se determina a partir de un conjunto de variables independientes llamadas predictores (*X1, X2, X3*...). Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre la variable dependiente (esto último se debe que analizar con cautela para no malinterpretar causa-efecto).

Los modelos lineales múltiples siguen la siguiente ecuación:

$$\hat{Y}_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}) + e_i$$

- β_0: la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
- β_i: el efecto medio que tiene el incremento en una unidad de la variable predictora X_i sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.
- e_i: el residuo, la diferencia entre el valor observado y el estimado por el modelo.

Es importante tener en cuenta que la magnitud de cada coeficiente parcial de correlación depende de las unidades en las que se mida la variables predictora a la que corresponde, por lo que su magnitud no está asociada con la importancia de cada predictor. Para poder determinar qué impacto tienen en el modelo cada una de las variables se emplean los *coeficientes parciales estandarizados* que se obtienen al transformar los coeficientes parciales de correlación en *Z-factors* dividiéndolos entre su error estándar.

Condiciones para la regresión lineal múltiple

Los modelos de correlación lineal múltiple requieren de las mismas condiciones que los modelos lineales simples más otras adicionales.

No colinealidad o multicolinealidad:

En los modelos lineales múltiples los predictores deben ser independientes, no debe de haber colinealidad entre ellos. La colinealidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores. Como consecuencia de la colinealidad no se puede identificar de forma precisa el efecto individual que tiene cada una de las variables conlineales sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto que resulta prácticamente imposible establecer su significancia estadística. Además, pequeños cambios en los datos provocan grandes cambios en las estimaciones de los coeficientes. Si bien la colinealidad propiamente dicha existe solo si el coeficiente de correlación simple o múltiple entre algunas de las variables independientes es 1, esto raramente ocurre en la realidad. Sin embargo, es frecuente encintar la llamada *casi-colinealidad* o *multicolinealidad no perfecta*.

No existe un método estadístico concreto para determinar la existencia de colinealidad o multicolinealidad entre los predictores de un modelo de regresión, sin embargo, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida afecta gravemente a la estimación y contraste de un modelo. Los pasos recomendados a seguir son:

- Calcular una matriz de coeficientes de correlación en la que se estudia la relación lineal entre cada par de predictores. Es importante tener en cuenta que, a pesar de no obtenerse ningún coeficiente de correlación alto, no está asegurado que no exista muticolinealidad. Se puede dar el caso de tener una relación lineal casi perfecta entre tres o más variables y que las correlaciones simples entre pares de estas mismas variables no sean mayores que 0.5.

- Generar un modelo de regresión lineal simple entre cada uno de los predictores frente al resto. Si en alguno de los modelos el *coeficiente de determinación* R^2 es alto, estaría señalando a una posible colinealidad.

- Tolerancia (TOL) y Factor de Inflación de la Varianza (VIF): Se trata de dos parámetros que vienen cuantificar lo mismo (uno es el inverso del otro). El VIF de cada predictor se calcula según la siguiente fórmula.

$$VIF_{\hat{\beta}_j} = \frac{1}{1 - R^2}$$

$$Tolerancia_{\hat{\beta}_j} = \frac{1}{VIF_{\hat{\beta}_j}}$$

Donde R^2 se obtiene de la regresión del predictor X_j sobre los otros predictores. Esta es la opción más recomendada, los límites de referencia que se suelen emplear son:

- VIF = 1: Ausencia total de colinealidad
- 1 < VIF < 5: La regresión puede verse afectada por cierta colinealidad.
- 5 < VIF < 10: Causa de preocupación
- El termino tolerancia es 1/VIF por lo que los limites recomendables están entre 1 y 0.1.

En caso de encontrar colinealidad entre predictores, hay dos posibles soluciones. La primera es excluir uno de los predictores problemáticos intentando conservar el que, a juicio del investigador, está influyendo de forma directa a la variable respuesta. Esta medida no suele tener mucho impacto en el modelo ya que al existir colinealidad, la información que aporta uno de los predictores es redundante en presencia del otro. La segunda opción consiste en combinar las variables coloniales en un único predictor.

Cuando se intenta establecer relaciones causa-efecto, la colinealidad puede llevar a conclusiones muy erróneas haciendo creer que una variable es la causa cuando en realidad es otra la que está influenciando sobre ese predictor.

Parsimonia:

Este término hace referencia a que el mejor modelo es aquel que es capaz de explicar con mayor precisión la variabilidad observada en la variable dependiente con el menor número de predictores, por lo tanto con menos asunciones.

Relación lineal entre los predictores numéricos y la variable respuesta:

Cada variable explicatoria numérica tiene que estar linealmente relacionada con la variable respuesta *Y* mientras los demás predictores se mantienen constantes, de lo contrario

no se puede introducir en el modelo. La forma más recomendable de comprobarlo es representando los residuos del modelo frente a cada uno de los predictores. Si la relación es lineal, los residuos se distribuyen de forma aleatoria entorno a cero. Estos análisis son solo aproximados, ya que no hay forma de saber si realmente la relación es lineal cuando el resto de predictores se mantienen constantes.

Distribución normal de los residuos:

Los residuos se deben distribuir de forma normal con media cero. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a test de hipótesis de normalidad.

Variabilidad constante de los residuos (homocedasticidad):

La varianza de los residuos debe de ser constante en todo el rango de observaciones. Para comprobarlo se representan los residuos, si la varianza es constante, los residuos se deben de distribuir de forma aleatoria manteniendo una misma dispersión y sin ningún patrón específico. Una distribución cónica es un claro identificador de falta de homocedasticidad. *En algunos libros se menciona el test de Goldfeld-Quandt y el test de Breusch-Pagan como contraste de homocedasticidad en correlación y regresión*

No autocorrelación (Independencia):

Los valores de cada observación son independientes de los otros, esto es especialmente importante de comprobar cuando se trabaja con mediciones temporales. Se recomienda representar los residuos ordenados acorde al tiempo de registro de las observaciones, si existe un cierto patrón hay indicios de autocorrelación. También se puede emplear el test de hipótesis de Durbin-Watson.

Valores atípicos y altamente influyentes:

Es importante identificar observaciones que sean atípicas y que puedan estar influenciando al modelo. La forma más fácil de detectarlas es a través de los residuos, tal como se explica en el capítulo de regresión lineal simple.

Tamaño de la muestra:

No se trata de una condición de por sí pero, si no se dispone de suficientes observaciones, variables que no son predictores influyentes podrían parecerlo. En el libro *Hanbook of biological statistics* recomiendan que el número de observaciones debe ser como mínimo entre 10 y 20 veces el número de predictores del modelo.

La gran mayoría de condiciones se verifican utilizando los residuos, por lo tanto, se suele generar primero el modelo y posteriormente validar las condiciones.

Elección de predictores para generar el mejor modelo

La evaluación de un modelo de regresión múltiple así como la elección de qué predictores se debe de incluir en el modelo es uno de los pasos más importantes en la modelización estadística. En los siguientes apartados se introduce este tema de forma muy simplificada. Para un desarrollo más detallado ver capítulo dedicado a la elección de predictores.

Evaluación del modelo en conjunto

Al igual que ocurre en los modelos lineales simples, R^2 es un cuantificador de la bondad de ajuste del modelo. Se define como el porcentaje de varianza de la variable *Y* que se explica mediante el modelo lineal respecto al total de variabilidad observada en *Y*. Por lo tanto, permite cuantificar como de bueno es el modelo para predecir el valor de las observaciones.

En los modelos lineales múltiples, cuantos más predictores se incluyan en el modelo mayor es el valor de R^2, ya que, por poco que sea, cada predictor va a explicar una parte de la variabilidad observada en *Y*.

R^2-ajustado introduce una penalización al valor de R^2 por cada predictor que se introduce en el modelo. El valor de la penalización depende del número de predictores utilizados y del tamaño de la muestra, es decir, del número de grados de libertad. Cuanto mayor es el tamaño de la muestra, más predictores se pueden incorporar en el modelo. R^2-ajustado permite encontrar el mejor modelo, aquel que consigue explicar mejor la variabilidad de la variable dependiente con el menor número de predictores. Si bien es un método para evaluar la bondad de ajuste muy utilizado, hay otros.

$$R^2_{ajustado} = 1 - \frac{SSE}{SST} x \frac{n-1}{n-k-1} = R^2 - (1-R^2)\frac{n-1}{n-k-1} = 1 - \frac{SSE/df_e}{SST/df_t}$$

- Siendo: SSE la variabilidad explicada por el modelo (*sum of squares explained*), SST la variabilidad total de *Y* (*sum of squares total*), *n* el tamaño de la muestra y *k* el número de predictores introducidos en el modelo.

Para conocer la variabilidad que explica cada una de las variables (predictores) incorporadas en el modelo se recurre a un ANOVA, ya que es el método que se encarga de analizar la varianza.

Tal y como ocurre en los modelos lineales simples o en los estudios de correlación, por muy alta que sea la bondad de ajuste, si el test F no resulta significativo no se puede aceptar el modelo como válido puesto que no es capaz de explicar la varianza observada.

Elección de los predictores

A la hora de seleccionar los predictores que deben formar parte del modelo se pueden seguir varios métodos:

Método jerárquico: basándose en el criterio del analista, se introducen unos predictores determinados en un orden determinado.

Método de entrada forzada: se introducen todos los predictores simultáneamente.

Método paso a paso (*stepwise*): emplea criterios matemáticos para decidir qué predictores contribuyen significativamente al modelo y en qué orden se introducen. Dentro de este método se diferencias tres estrategias:

- Dirección *forward*: El modelo inicial no contiene ningún predictor, solo el parámetro β_0. A partir de este se generan todos los posibles modelos introduciendo una sola variable de entre las disponibles. Aquella variable que mejore en mayor medida el modelo se selecciona. A continuación se intenta incrementar el modelo probando a introducir una a una las variables restantes. Si introduciendo alguna de ellas mejora, también se selecciona. En el caso de que varias lo hagan, se selecciona la que incremente en mayor medida la capacidad del modelo. Este proceso se repite hasta llegar al punto en el que ninguna de las variables que quedan por incorporar mejore el modelo.

- Dirección *backward*: El modelo se inicia con todas las variables disponibles incluidas como predictores. Se prueba a eliminar una a una cada variable, si se mejora el modelo, queda excluida. Este método permite evaluar cada variable en presencia de las otras.

- Doble o mixto: Se trata de una combinación de la selección *forward* y *backward*. Se inicia igual que el *forward* pero tras cada nueva incorporación se realiza un test de extracción de predictores no útiles como en el *backward*. Presenta la ventaja de que si a medida que se añaden predictores, alguno de los ya presentes deja de contribuir al modelo, se elimina.

El método paso a paso requiere de algún criterio matemático para determinar si el modelo mejora o empeora con cada incorporación o extracción. Existen varios parámetros empelados, de entre los que destacan el *R^2-ajustado*, el *p-value* de cada predictor y el *Akaike(AIC)*, cada uno de ellos tiene ventajas e inconvenientes. El método *Akaike(AIC)* tiende a ser más restrictivo e introducir menos predictores que el *R^2-ajustado*. Para un mismo set de datos, no todos los métodos tienen porque concluir en un mismo modelo.

Cuando el criterio de elección se basa en el *p-value* de los predictores, el proceso tiene que seguir el método *backward*. Generar el modelo incluyendo todas las variables. De entre las variables cuyo *p-value* no sea significativo se excluye aquella que tenga mayor *p-value*. Se recalcula el modelo y se repite el proceso hasta que solo queden predictores con significancia estadística. En el caso de variables categóricas, si al menos uno de sus niveles es significativo, se considera que la variable es significativa.

En R la función `step()` permite encontrar el mejor modelo basado en AIC utilizando cualquiera de las 3 variantes del método paso a paso.

Variables nominales/categóricas como predictores

Cuando se introduce una variable categórica como predictor, un nivel se considera el de referencia (normalmente codificado como 0) contra el que se van a comparar el resto de niveles. En el caso de que el predictor categórico tenga más de dos niveles se generan lo que se conoce como variables *dummy*, que consisten en variables creadas para cada uno de los niveles del predictor categórico y que pueden tomar el valor de 0 o 1. Cada vez que se emplee el modelo para predecir un valor, solo una variable *dummy* por predictor adquiere el valor 1 (la que coincida con el valor que adquiere el predictor en ese caso) mientras que el resto se consideran 0. El valor del coeficiente parcial de correlación β_i de cada variable *dummy* indica el porcentaje promedio en el que influye dicho nivel sobre la variable dependiente Y en comparación con el nivel de referencia de dicho predictor.

Validación cruzada

Una vez seleccionado el mejor modelo que se puede crear empleando los datos disponibles se tiene que comprobar su validez prediciendo nuevas observaciones que no se han empleado para entrenarlo, de este modo se verifica si el modelo se puede generalizar. La validación cruzada consiste en estudiar la precisión de un modelo a través de diferentes muestras. Una estrategia comúnmente empleada es dividir aleatoriamente los datos en dos grupos (70%-30%), ajustar el modelo con el primer grupo y evaluar la precisión de las predicciones con el segundo. *Ver capítulo dedicado a este tema.*

Identificación de valores atípicos (*outliers*) o influyentes

Independientemente de que el modelo se haya podido aceptar, siempre es conveniente identificar si hay algún posible *outlier* u observación altamente influyente, puesto que podría estar condicionando en gran medida el modelo. La eliminación de este tipo de observaciones debe de analizarse con detalle y dependiendo de la finalidad del modelo. Si el fin es predictivo, un modelo sin *outliers* ni observaciones altamente influyentes puede ser más útil para predecir con mayor precisión la mayoría de casos. Sin embargo, es muy importante prestar atención a estos valores ya que de no ser errores de medida pueden ser los casos más interesantes. El

modo adecuado a proceder cuando se sospecha de algún posible valor atípico o influyente es calcular el modelo de regresión incluyendo y excluyendo dicho valor.

Si se quiere hacer un estudio analítico de los posibles valores atípicos se recurre a los *studentized residuals* que se obtienen al dividir los residuos de cada observación entre una estimación de su error estándar. Si se supera un valor absoluto de 3 debe de considerarse esa observación como posible valor atípico. En R se pueden calcular con la función `rstudent()`.

Además de que un valor sea atípico, es necesario estudiar como de influyente es en el conjunto del modelo. Si un valor es atípico pero no influyente, no es crítico que se considere su eliminación. Dos de las medidas más empleadas para cuantificar la influencia son:

- Leverages (hat): Se consideran observaciones influyentes aquellas cuyos valores *hat* superen *2.5x((p+1)/n)*, siendo *p* el número de predictores y *n* el número de observaciones.
- Distancia Cook (cook.d): Se consideran influyentes valores superiores a 1.

En R se dispone de la funcione `outlierTest()` del paquete *car* y de las funciones `influence.measures()`, `influence.plot()` y `hatvalues()` para identificar las observaciones más influyentes en el modelo.

Ejemplo1. Predictores numéricos

Un estudio quiere generar un modelo que permita predecir la esperanza de vida media de los habitantes de una ciudad en función de diferentes variables. Se dispone de información sobre: habitantes, analfabetismo, ingresos, esperanza de vida, asesinatos, universitarios, heladas, área y densidad poblacional.

```
# El data set empleado es el state.x77 Para facilitar su interpretación se
# renombra y se modifica
require(dplyr)
datos <- as.data.frame(state.x77)
datos <- rename(habitantes = Population, analfabetismo = Illiteracy, ingresos =
Income, esp_vida = `Life Exp`, asesinatos = Murder, universitarios = `HS Grad`,
heladas = Frost, area = Area, .data = datos)
datos <- mutate(.data = datos, densidad_pobl = habitantes * 1000/area)
```

1.Analizar la relación entre variables

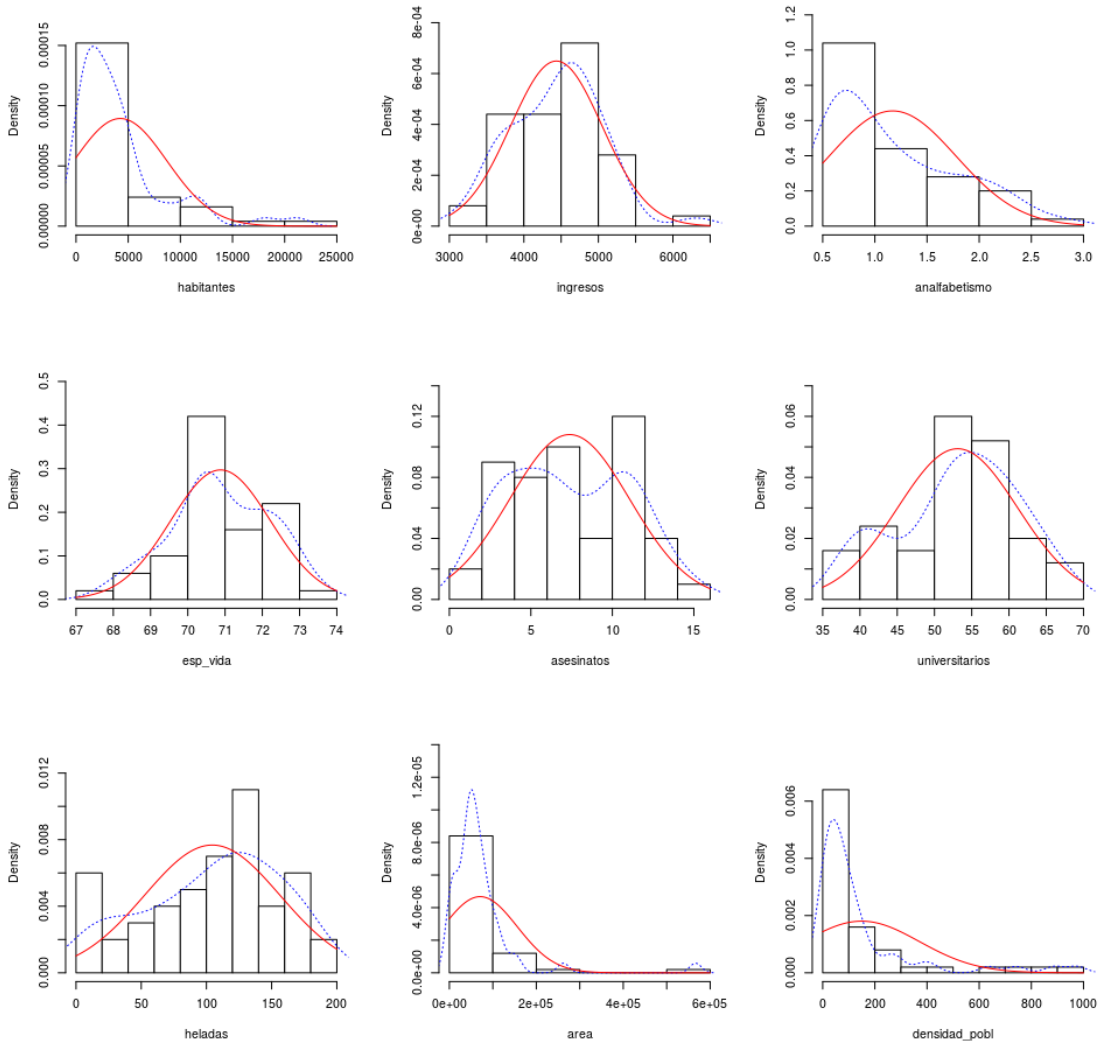
El primer paso a la hora de establecer un modelo lineal múltiple es estudiar la relación que existe entre variables. Esta información es crítica a la hora de identificar cuáles pueden ser los mejores predictores para el modelo, que variables presentan relaciones de tipo no lineal (por lo que no pueden ser incluidas) y para identificar colinealidad entre predictores. A modo complementario, es recomendable representar la distribución de cada variable mediante histogramas.

Las dos formas principales de hacerlo son mediante representaciones gráficas (gráficos de dispersión) y el cálculo del coeficiente de correlación de cada par de variables.

```
round(cor(x = datos, method = "pearson"), 3)
```

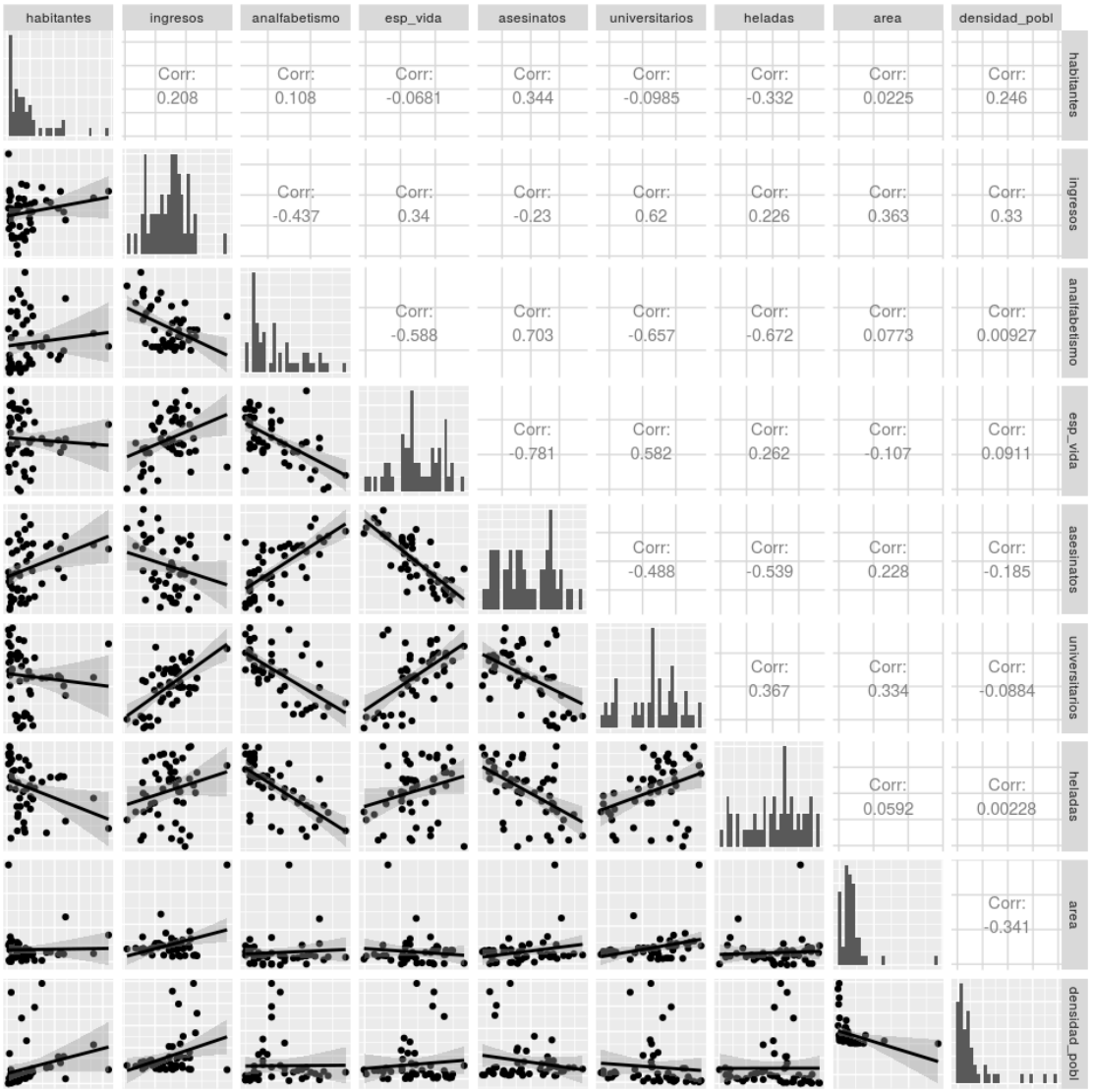
```
##                habitantes ingresos analfabetismo esp_vida asesinatos
## habitantes          1.000    0.208         0.108   -0.068      0.344
## ingresos            0.208    1.000        -0.437    0.340     -0.230
## analfabetismo       0.108   -0.437         1.000   -0.588      0.703
## esp_vida           -0.068    0.340        -0.588    1.000     -0.781
## asesinatos          0.344   -0.230         0.703   -0.781      1.000
## universitarios     -0.098    0.620        -0.657    0.582     -0.488
## heladas            -0.332    0.226        -0.672    0.262     -0.539
## area                0.023    0.363         0.077   -0.107      0.228
## densidad_pobl       0.246    0.330         0.009    0.091     -0.185
##                universitarios heladas   area densidad_pobl
## habitantes             -0.098  -0.332  0.023         0.246
## ingresos                0.620   0.226  0.363         0.330
## analfabetismo          -0.657  -0.672  0.077         0.009
## esp_vida                0.582   0.262 -0.107         0.091
## asesinatos             -0.488  -0.539  0.228        -0.185
## universitarios          1.000   0.367  0.334        -0.088
## heladas                 0.367   1.000  0.059         0.002
## area                    0.334   0.059  1.000        -0.341
## densidad_pobl          -0.088   0.002 -0.341         1.000
```

```
require(psych)
multi.hist(x = datos, dcol = c("blue", "red"), dlty = c("dotted", "solid"),
    main = "")
```



Otros paquetes permiten representar a la vez los diagramas de dispersión, los valores de correlación para cada par de variables y la distribución de cada una de las variables.

```
require(GGally)
ggpairs(datos, lower = list(continuous = "smooth"), diag = list(continuous =
"bar"), axisLabels = "none")
```



Del análisis preliminar se pueden extraer las siguientes conclusiones:

- Las variables que tienen una mayor relación lineal con la esperanza de vida son: asesinatos (r= -0.78), analfabetismo (r= -0.59) y universitarios (r= 0.58).
- Universitarios y analfabetismo están medianamente correlacionados (r = -0.66) por lo que posiblemente no sea útil introducir ambos predictores en el modelo.
- Las variables habitantes, área y densidad poblacional muestran una distribución exponencial, una transformación logarítmica posiblemente haría más normal su distribución.

2.Generar el modelo

Como se ha explicado en la introducción, hay diferentes formas de llegar al modelo final más adecuado. En este caso se va a emplear el método *backward* iniciando el modelo con todas las variables como predictores y realizando la selección de los mejores predictores con la medición *Akaike(AIC)*.

```
modelo <- lm(esp_vida ~ habitantes + ingresos + analfabetismo + asesinatos +
    universitarios + heladas + area + densidad_pobl, data = datos)
summary(modelo)
```

```
##
## Call:
## lm(formula = esp_vida ~ habitantes + ingresos + analfabetismo +
##     asesinatos + universitarios + heladas + area + densidad_pobl,
##     data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47514 -0.45887 -0.06352  0.59362  1.21823
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.995e+01  1.843e+00  37.956  < 2e-16 ***
## habitantes      6.480e-05  3.001e-05   2.159   0.0367 *
## ingresos        2.701e-04  3.087e-04   0.875   0.3867
## analfabetismo   3.029e-01  4.024e-01   0.753   0.4559
## asesinatos     -3.286e-01  4.941e-02  -6.652 5.12e-08 ***
## universitarios  4.291e-02  2.332e-02   1.840   0.0730 .
## heladas        -4.580e-03  3.189e-03  -1.436   0.1585
## area           -1.558e-06  1.914e-06  -0.814   0.4205
## densidad_pobl  -1.105e-03  7.312e-04  -1.511   0.1385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7337 on 41 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7013
## F-statistic: 15.38 on 8 and 41 DF,  p-value: 3.787e-10
```

El modelo con todas las variables introducidas como predictores tiene una R^2 alta (0.7501), es capaz de explicar el 75,01% de la variabilidad observada en la esperanza de vida. El p-value del modelo es significativo (3.787e-10) por lo que se puede aceptar que el modelo no es

por azar, al menos uno de los coeficientes parciales es distinto de 0. Muchos de ellos no son significativos, lo que es un indicativo de que podrían no contribuir al modelo.

3.Selección de los mejores predictores

En este caso se van a emplear la estrategia de *step wise mixto*. El valor matemático empleado para determinar la calidad del modelo va a ser *Akaike(AIC)*.

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

```
## Start:  AIC=-22.89
## esp_vida ~ habitantes + ingresos + analfabetismo + asesinatos +
##     universitarios + heladas + area + densidad_pobl
##
##                  Df Sum of Sq    RSS     AIC
## - analfabetismo   1    0.3050 22.373 -24.208
## - area            1    0.3564 22.425 -24.093
## - ingresos        1    0.4120 22.480 -23.969
## <none>                        22.068 -22.894
## - heladas         1    1.1102 23.178 -22.440
## - densidad_pobl   1    1.2288 23.297 -22.185
## - universitarios  1    1.8225 23.891 -20.926
## - habitantes      1    2.5095 24.578 -19.509
## - asesinatos      1   23.8173 45.886  11.707
##
## Step:  AIC=-24.21
## esp_vida ~ habitantes + ingresos + asesinatos + universitarios +
##     heladas + area + densidad_pobl
##
##                  Df Sum of Sq    RSS     AIC
## - area            1    0.1427 22.516 -25.890
## - ingresos        1    0.2316 22.605 -25.693
## <none>                        22.373 -24.208
## - densidad_pobl   1    0.9286 23.302 -24.174
## - universitarios  1    1.5218 23.895 -22.918
## + analfabetismo   1    0.3050 22.068 -22.894
## - habitantes      1    2.2047 24.578 -21.509
## - heladas         1    3.1324 25.506 -19.656
## - asesinatos      1   26.7071 49.080  13.072
##
## Step:  AIC=-25.89
## esp_vida ~ habitantes + ingresos + asesinatos + universitarios +
##     heladas + densidad_pobl
##
```

```
##                   Df Sum of Sq    RSS     AIC
## - ingresos         1     0.132 22.648 -27.598
## - densidad_pobl    1     0.786 23.302 -26.174
## <none>                         22.516 -25.890
## - universitarios   1     1.424 23.940 -24.824
## + area             1     0.143 22.373 -24.208
## + analfabetismo    1     0.091 22.425 -24.093
## - habitantes       1     2.332 24.848 -22.962
## - heladas          1     3.304 25.820 -21.043
## - asesinatos       1    32.779 55.295  17.033
##
## Step:  AIC=-27.6
## esp_vida ~ habitantes + asesinatos + universitarios + heladas +
##     densidad_pobl
##
##                   Df Sum of Sq    RSS     AIC
## - densidad_pobl    1     0.660 23.308 -28.161
## <none>                         22.648 -27.598
## + ingresos         1     0.132 22.516 -25.890
## + analfabetismo    1     0.061 22.587 -25.732
## + area             1     0.043 22.605 -25.693
## - habitantes       1     2.659 25.307 -24.046
## - heladas          1     3.179 25.827 -23.030
## - universitarios   1     3.966 26.614 -21.529
## - asesinatos       1    33.626 56.274  15.910
##
## Step:  AIC=-28.16
## esp_vida ~ habitantes + asesinatos + universitarios + heladas
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         23.308 -28.161
## + densidad_pobl    1     0.660 22.648 -27.598
## + ingresos         1     0.006 23.302 -26.174
## + analfabetismo    1     0.004 23.304 -26.170
## + area             1     0.001 23.307 -26.163
## - habitantes       1     2.064 25.372 -25.920
## - heladas          1     3.122 26.430 -23.877
## - universitarios   1     5.112 28.420 -20.246
## - asesinatos       1    34.816 58.124  15.528

##
## Call:
## lm(formula = esp_vida ~ habitantes + asesinatos + universitarios +
##     heladas, data = datos)
##
## Coefficients:
##    (Intercept)      habitantes      asesinatos  universitarios
##      7.103e+01       5.014e-05      -3.001e-01       4.658e-02
##        heladas
##     -5.943e-03
```

El mejor modelo resultante del proceso de selección ha sido:

```
modelo <- (lm(formula = esp_vida ~ habitantes + asesinatos + universitarios +
    heladas, data = datos))
summary(modelo)
```

```
##
## Call:
## lm(formula = esp_vida ~ habitantes + asesinatos + universitarios +
##     heladas, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.103e+01  9.529e-01  74.542  < 2e-16 ***
## habitantes      5.014e-05  2.512e-05   1.996  0.05201 .
## asesinatos     -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## universitarios  4.658e-02  1.483e-02   3.142  0.00297 **
## heladas        -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

Es recomendable mostrar el intervalo de confianza para cada uno de los coeficientes parciales de correlación.

```
confint(lm(formula = esp_vida ~ habitantes + asesinatos + universitarios + heladas,
    data = datos))
```

```
##                        2.5 %        97.5 %
## (Intercept)     6.910798e+01 72.9462729104
## habitantes     -4.543308e-07  0.0001007343
## asesinatos     -3.738840e-01 -0.2264135705
## universitarios  1.671901e-02  0.0764454870
## heladas        -1.081918e-02 -0.0010673977
```

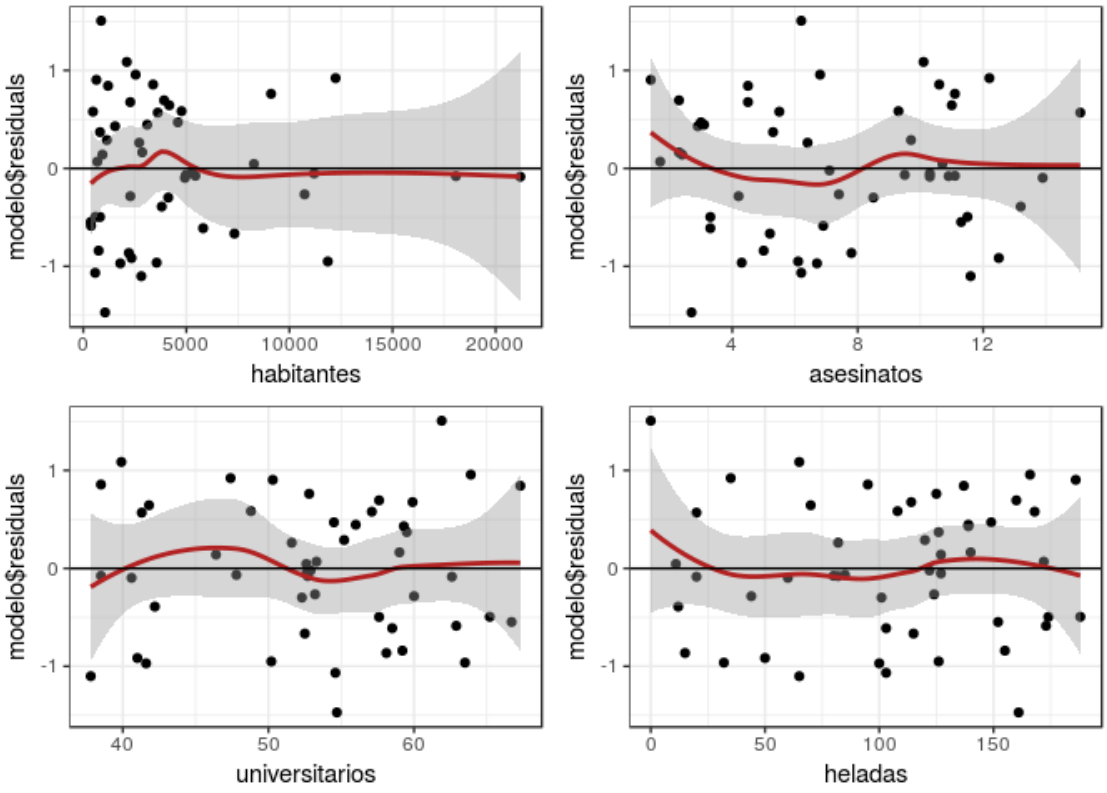
Cada una de las pendientes de un modelo de regresión lineal múltiple (coeficientes parciales de correlación de los predictores) se define del siguiente modo: Si el resto de variables se mantienen constantes, por cada unidad que aumenta el predictor en cuestión, la variable (*Y*) varía en promedio tantas unidades como indica la pendiente.

4.Validación de condiciones para la regresión múltiple lineal

Relación lineal entre los predictores numéricos y la variable respuesta:

Esta condición se puede validar o bien mediante diagramas de dispersión entre la variable dependiente y cada uno de los predictores (como se ha hecho en el análisis preliminar) o con diagramas de dispersión entre cada uno de los predictores y los residuos del modelo. Si la relación es lineal, los residuos deben de distribuirse aleatoriamente en torno a 0 con una variabilidad constante a lo largo del eje X. Esta última opción suele ser más indicada ya que permite identificar posibles datos atípicos.

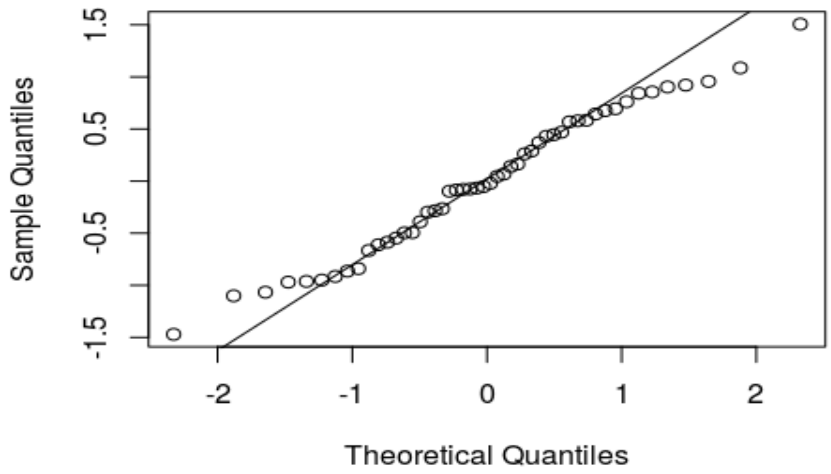
```
require(ggplot2)
require(gridExtra)
plot1 <- ggplot(data = datos, aes(habitantes, modelo$residuals)) + geom_point() +
    geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()


plot2 <- ggplot(data = datos, aes(asesinatos, modelo$residuals)) + geom_point() +
    geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()


plot3 <- ggplot(data = datos, aes(universitarios, modelo$residuals)) + geom_point()
+ geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()


plot4 <- ggplot(data = datos, aes(heladas, modelo$residuals)) + geom_point() +
    geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()


grid.arrange(plot1, plot2, plot3, plot4)
```



Se cumple la linealidad para todos los predictores

Distribución normal de los residuos:

```
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```



```
shapiro.test(modelo$residuals)
```

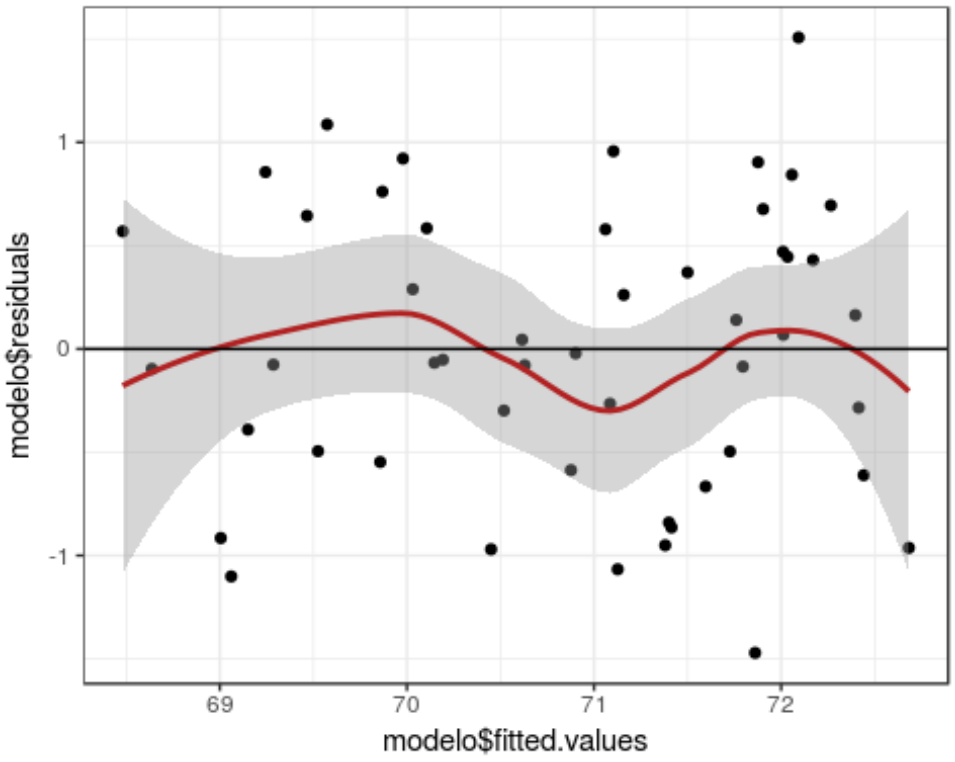
```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.97935, p-value = 0.525
```

Tanto el análisis gráfico como es test de hipótesis confirman la normalidad.

Variabilidad constante de los residuos (homocedasticidad):

Al representar los residuos frente a los valores ajustados por el modelo, los primeros se tienen que distribuir de forma aleatoria en torno a cero, manteniendo aproximadamente la misma variabilidad a lo largo del eje X. Si se observa algún patrón específico, por ejemplo forma cónica o mayor dispersión en los extremos, significa que la variabilidad es dependiente del valor ajustado y por lo tanto no hay homocedasticidad.

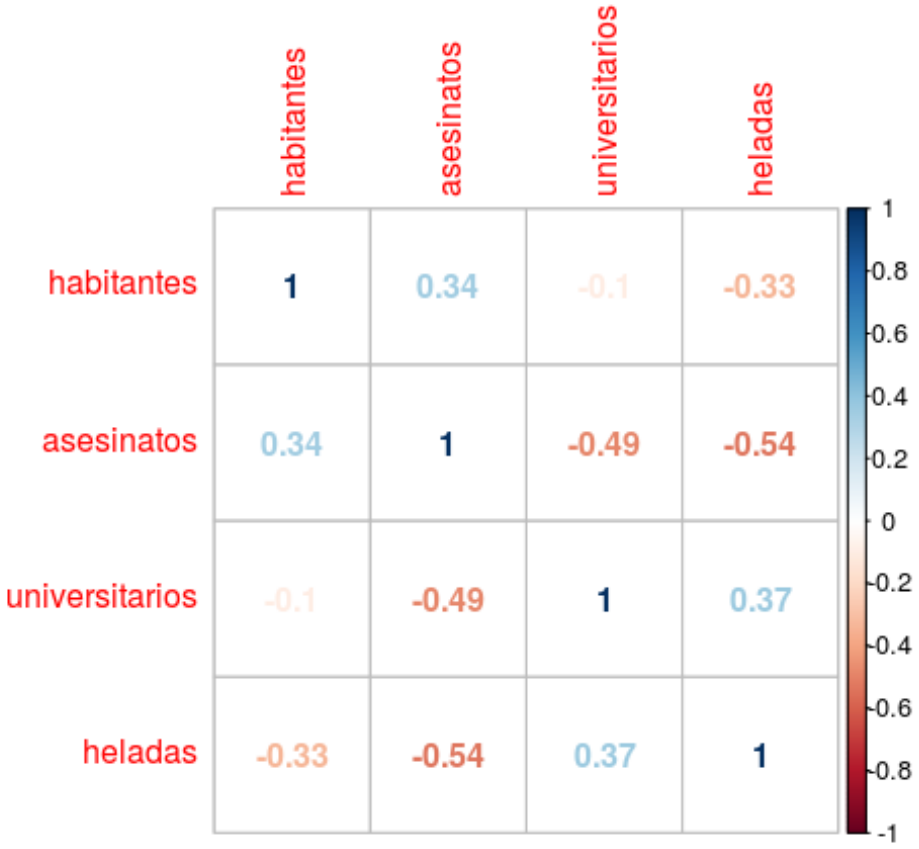
```
ggplot(data = datos, aes(modelo$fitted.values, modelo$residuals)) + geom_point() +
    geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()
```



No multicolinealidad:

Matriz de correlación entre predictores.

```
require(corrplot)
corrplot(cor(select(datos, habitantes, asesinatos, universitarios, heladas)),
    method = "number")
```



Análisis de Inflación de Varianza (VIF):

```
require(car)
vif(modelo)
```

```
##     habitantes     asesinatos universitarios        heladas
##       1.189835       1.727844       1.356791       1.498077
```

No hay predictores que muestren una correlación lineal muy alta ni inflación de varianza.

Autocorrelación:

```
require(car)
dwt(modelo, alternative = "two.sided")
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02867262      1.913997   0.778
##  Alternative hypothesis: rho != 0
```

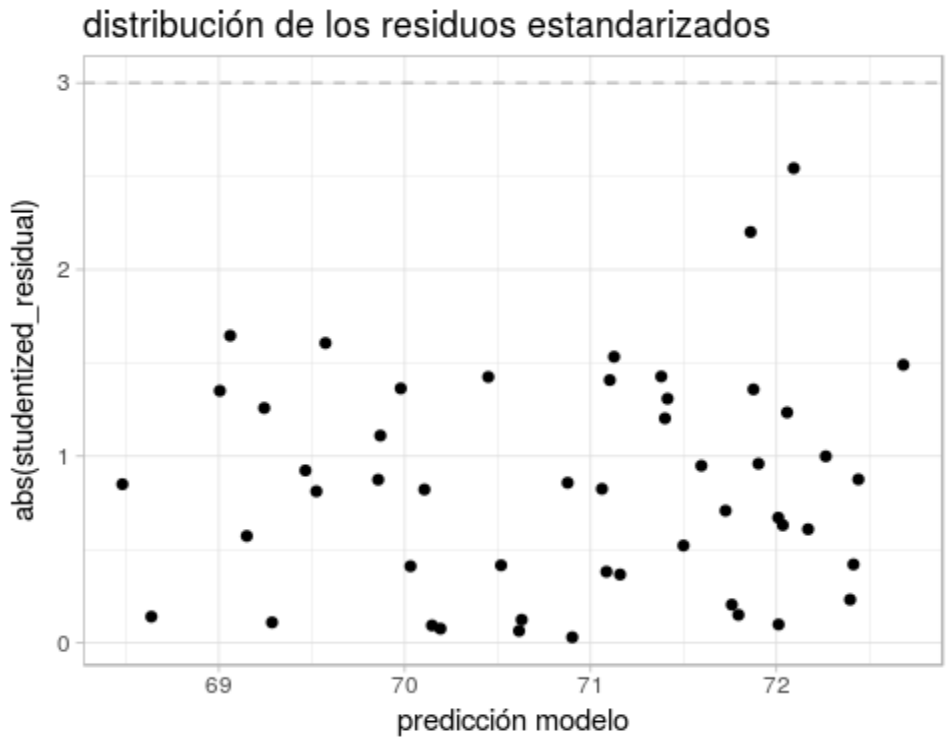
No hay evidencia de autocorrelación

Tamaño de la muestra:

No existe una condición establecida para el número mínimo de observaciones, pero para prevenir que una variable resulte muy influyente cuando realmente no lo es, se recomienda que la cantidad de observaciones sea entre 10 y 20 veces el número de predictores. En este caso debería haber como mínimo 40 observaciones y se dispone de 50 por lo que es apropiado.

5.Identficación de posibles valores atípicos o influyentes

```
library(dplyr)
datos$studentized_residual <- rstudent(modelo)
ggplot(data = datos, aes(x = predict(modelo), y = abs(studentized_residual))) +
    geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +
    #se identifican en rojo observaciones con residuos estandarizados absolutos > 3
    geom_point(aes(color = ifelse(abs(studentized_residual) >3, 'red', 'black'))) +
    scale_color_identity() +
    labs(title = "distribución de los residuos estandarizados", x = "predicción
modelo") +
    theme_light()
```




```
which(abs(datos$studentized_residual) > 3)
```

```
## integer(0)
```

No se identifica ninguna observación extrema.

```
summary(influence.measures(modelo))
```

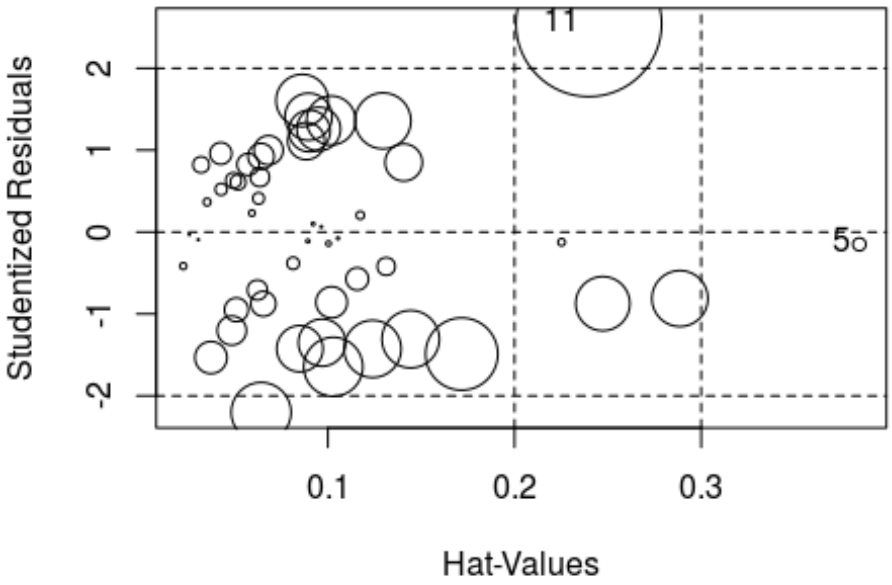
```
## Potentially influential observations of
##   lm(formula = esp_vida ~ habitantes + asesinatos + universitarios +
heladas, data = datos) :
##
##    dfb.1_ dfb.hbtn dfb.assn dfb.unvr dfb.hlds dffit   cov.r   cook.d
## 2   0.41   0.18    -0.40    -0.35    -0.16    -0.50    1.36_*  0.05
## 5   0.04  -0.09     0.00    -0.04     0.03    -0.12    1.81_*  0.00
## 11 -0.03  -0.57    -0.28     0.66    -1.24_*   1.43_*  0.74    0.36
## 28  0.40   0.14    -0.42    -0.29    -0.28    -0.52    1.46_*  0.05
## 32  0.01  -0.06     0.00     0.00    -0.01    -0.07    1.44_*  0.00
##    hat
## 2   0.25
## 5   0.38_*
## 11  0.24
## 28  0.29
## 32  0.23
```

En la tabla generada se recogen las observaciones que son significativamente influyentes en al menos uno de los predictores (una columna para cada predictor). Las tres últimas columnas son 3 medidas distintas para cuantificar la influencia. A modo de guía se pueden considerar excesivamente influyentes aquellas observaciones para las que:

- Leverages (hat): Se consideran observaciones influyentes aquellas cuyos valores *hat* superen *2.5x((p+1)/n)*, siendo *p* el número de predictores y *n* el número de observaciones.
- Distancia Cook (cook.d): Se consideran influyentes valores superiores a 1.

La visualización gráfica de las influencias se obtiene del siguiente modo:

```
influencePlot(modelo)
```




```
##       StudRes       Hat       CookD
## 5  -0.1500614 0.3847592 0.002879053
## 11  2.5430162 0.2397924 0.363778638
```

Los análisis muestran varias observaciones influyentes (posición 5 y 11) que exceden los límites de preocupación para los valores de *Leverages* o *Distancia Cook*. Estudios más exhaustivos consistirían en rehacer el modelo sin las observaciones y ver el impacto.

6.Conclusión

El modelo lineal múltiple

$$Esperanza\ de\ vida = 5.014e^{-05} habitantes - 3.001e^{-01} asesinatos + 4.658e^{-02} universitarios - 5.943e^{-03} heladas$$

es capaz de explicar el 73.6% de la variabilidad observada en la esperanza de vida (R^2: 0.736, R^2-Adjusted: 0.7126). El test F muestra que es significativo (*p-value*: 1.696e-12). Se satisfacen todas las condiciones para este tipo de regresión múltiple. Dos observaciones (posición 5 y 11) podrían estar influyendo de forma notable en el modelo.

Ejemplo2. Predictores numéricos y categóricos.

Se dispone de un dataset que contiene información de 30 libros. Se conoce del peso total de cada libro, el volumen que tiene y el tipo de tapas (duras o blandas). Se quiere generar un modelo lineal múltiple que permita predecir el peso de un libro en función de su volumen y del tipo de tapas.

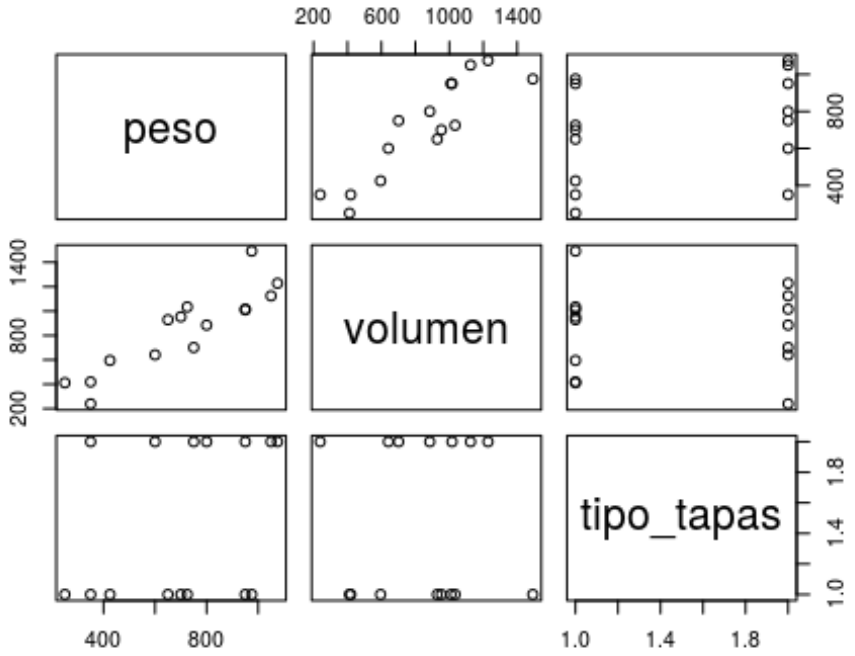
```
datos <- data.frame(peso = c(800, 950, 1050, 350, 750, 600, 1075, 250, 700,
    650, 975, 350, 950, 425, 725), volumen = c(885, 1016, 1125, 239, 701, 641,
    1228, 412, 953, 929, 1492, 419, 1010, 595, 1034), tipo_tapas = c("duras",
    "duras", "duras", "duras", "duras", "duras", "duras", "blandas", "blandas",
    "blandas", "blandas", "blandas", "blandas", "blandas", "blandas"))
head(datos, 4)
```

```
##   peso volumen tipo_tapas
## 1  800     885      duras
## 2  950    1016      duras
## 3 1050    1125      duras
## 4  350     239      duras
```

1.Analizar la correlación entre cada par de variables cuantitativas y diferencias del valor promedio entre las categóricas

Se enfrentan cada par de variables cuantitativas mediante un diagrama de dispersión múltiple (*pairwise scatterplot*) para intuir si existe relación lineal o monotónica con la variable respuesta. Si no la hay, no es adecuado emplear un modelo de regresión lineal. Además, se estudia la relación entre variables para detectar posible colinealidad. Para las variables de tipo categórico se genera un *boxplot* con sus niveles para intuir su influencia en la variable dependiente.

```
pairs(datos)
```

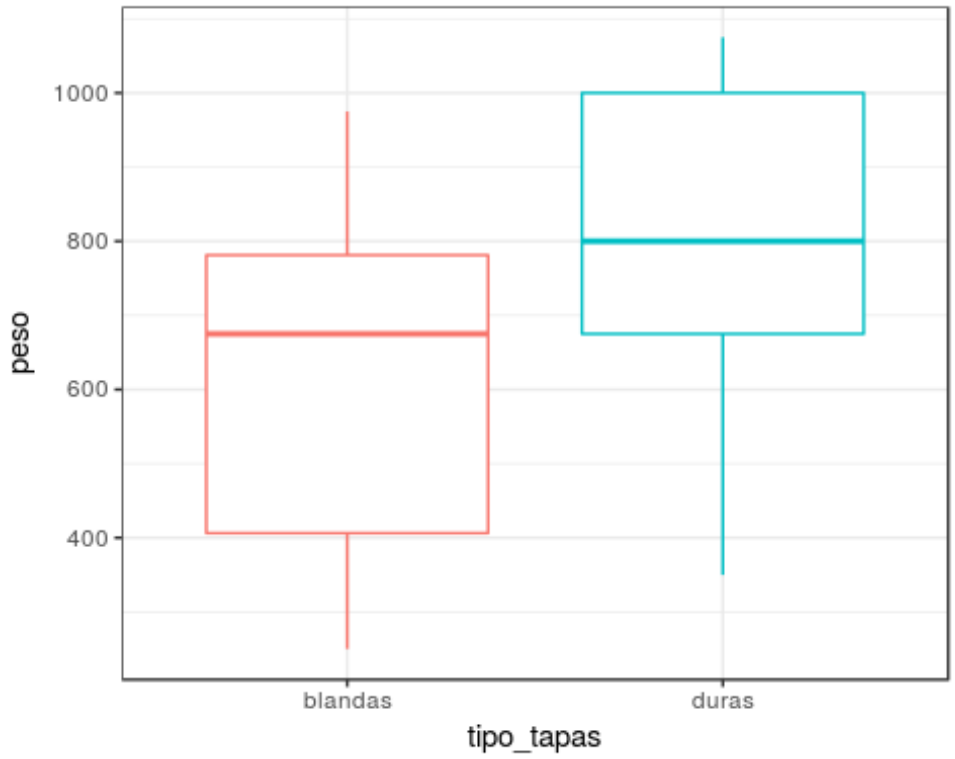



```
cor.test(datos$peso, datos$volumen, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$peso and datos$volumen
## t = 7.271, df = 13, p-value = 6.262e-06
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.7090393 0.9651979
## sample estimates:
##       cor
## 0.8958988
```

```
ggplot(data = datos, mapping = aes(x = tipo_tapas, y = peso, color = tipo_tapas)) +
    geom_boxplot() + theme_bw() + theme(legend.position = "none")
```



El análisis gráfico y de correlación muestran una relación lineal significativa entre la variable *peso* y *volumen*. La variable *tipo_tapas* parece influir de forma significativa en el peso. Ambas variables pueden ser buenos predictores en un modelo lineal múltiple para la variable dependiente peso.

2.Generar el modelo lineal múltiple

```
modelo <- lm(peso ~ volumen + tipo_tapas, data = datos)
summary(modelo)
```

```
##
## Call:
## lm(formula = peso ~ volumen + tipo_tapas, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -110.10  -32.32  -16.10   28.93  210.95
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      13.91557   59.45408   0.234 0.818887
## volumen           0.71795    0.06153  11.669  6.6e-08 ***
## tipo_tapasduras 184.04727   40.49420   4.545 0.000672 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.2 on 12 degrees of freedom
## Multiple R-squared:  0.9275, Adjusted R-squared:  0.9154
## F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

```
confint(modelo)
```

```
##                        2.5 %      97.5 %
## (Intercept)     -115.6237330 143.4548774
## volumen            0.5839023   0.8520052
## tipo_tapasduras   95.8179902 272.2765525
```

Cada una de las pendientes de un modelo de regresión lineal múltiple se define del siguiente modo: Si el resto de variables se mantienen constantes, por cada unidad que aumenta el predictor en cuestión, la variable (*Y*) varía en promedio tantas unidades como indica la pendiente. En el caso del predictor *volume*, si el resto de variables no varían, por cada unidad de *volumen* que aumenta el libro el peso se incrementa en promedio 0.71795 unidades.

Cuando un predictor es cualitativo, uno de sus niveles se considera de referencia (el que no aparece en la tabla de resultados) y se le asigna el valor de 0. El valor de la pendiente de cada nivel de un predictor cualitativo se define como el promedio de unidades que cada nivel está por encima o debajo del nivel de referencia. Para el predictor *tipo_tapas*, el nivel de referencia es *tapas blandas* por lo que si el libro tiene este tipo de tapas se le da a la variable el

valor 0 y si es de *tapas duras* el valor 1. Acorde al modelo generado, los libros de tapa dura son en promedio 184.04727 unidades de peso superiores a los de tapa blanda.

$$Peso\ libro = 13.91557 + 0.71795\ volumen + 184.04727\ tipotapas$$

El modelo es capaz de explicar el 92.75% de la variabilidad observada en el peso de los libros *(R-squared: 0.9275)*. El valor de R^2-ajustado es muy alto y cercano al R^2 *(Adjusted R-squared: 0.9154)* lo que indica que el modelo contiene predictores útiles. El test F muestra un *p-value* de 1.455e-07 por lo que el modelo en conjunto es significativo. Esto se corrobora con el *p-value* de cada predictor, en ambos casos significativo.

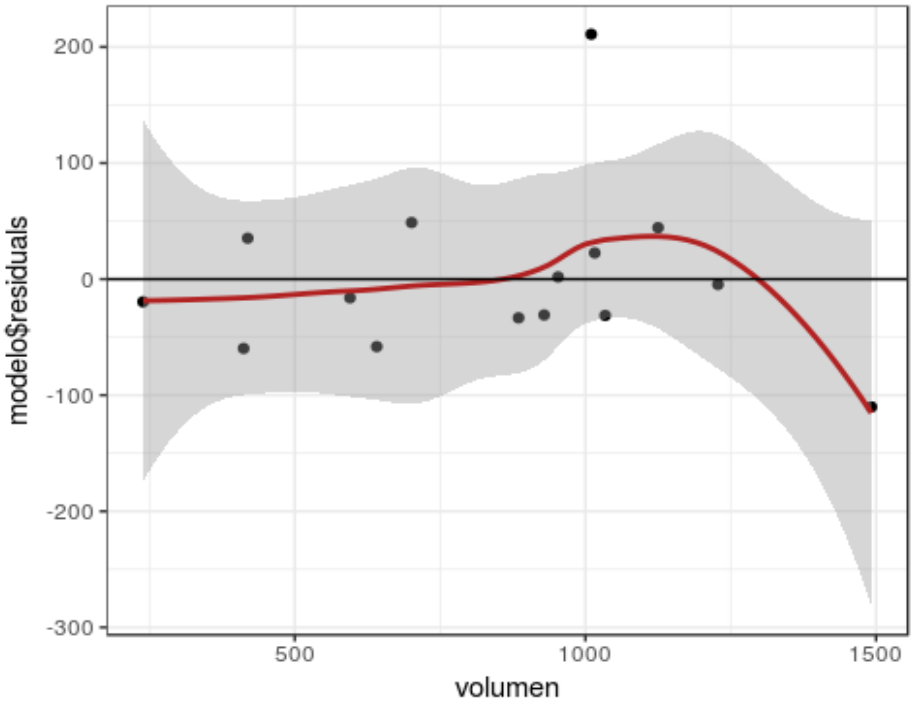
3.Elección de los predictores

En este caso, al solo haber dos predictores, a partir del *summary* del modelo se identifica que ambas variables incluidas son importantes.

4.Condiciones para la regresión múltiple lineal

1.Relación lineal entre los predictores numéricos y la variable dependiente:

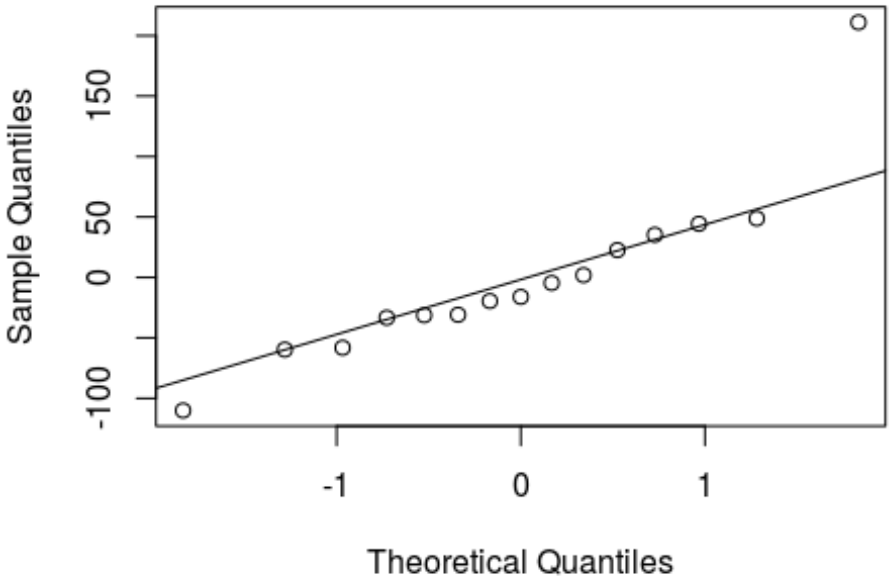
```
require(ggplot2)
ggplot(data = datos, aes(x = volumen, y = modelo$residuals)) + geom_point() +
    geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()
```



Se satisface la condición de linealidad. Se aprecia un posible dato atípico.

2.Distribución normal de los residuos:

```
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```



```
shapiro.test(modelo$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.85497, p-value = 0.02043
```

La condición de normalidad no se satisface, posiblemente debido a un dato atípico. Se repite el análisis excluyendo la observación a la que pertenece el residuo atípico.

```
which.max(modelo$residuals)
```

```
## 13
## 13
```

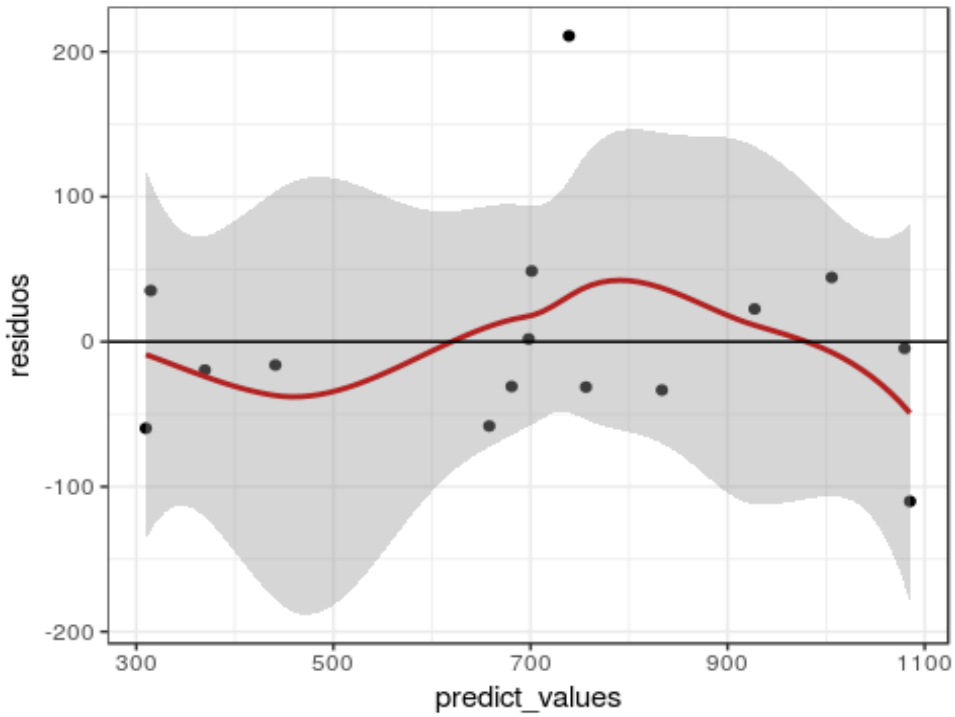
```
shapiro.test(modelo$residuals[-13])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals[-13]
## W = 0.9602, p-value = 0.7263
```

Se confirma que los residuos sí se distribuyen de forma normal a excepción de un dato extremo. Es necesario estudiar en detalle la influencia de esta observación para determinar si el modelo es más preciso sin ella.

3.Variabilidad constante de los residuos:

```
ggplot(data = data.frame(predict_values = predict(modelo), residuos =
residuals(modelo)), aes(x = predict_values, y = residuos)) + geom_point() +
geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()
```



4.No multicolinealidad:

Dado que solo hay un predictor cuantitativo no se puede dar colinealidad.

5.Autocorrelación:

```
require(car)
dwt(modelo,alternative = "two.sided")
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    0.0004221711      1.970663    0.76
##  Alternative hypothesis: rho != 0
```

No hay evidencia de autocorrelación

6.Tamaño de la muestra:

No existe una condición establecida para el número mínimo de observaciones pero, para prevenir que una variable resulte muy influyente cuando realmente no lo es, se recomienda que la cantidad de observaciones sea entre 10 y 20 veces el número de predictores. En este caso debería haber como mínimo 20 observaciones y se dispone de 15 por lo que se debería considerar incrementar la muestra.

5.Identficación de posibles valores atípicos o influyentes

```
require(car)
outlierTest(modelo)
```

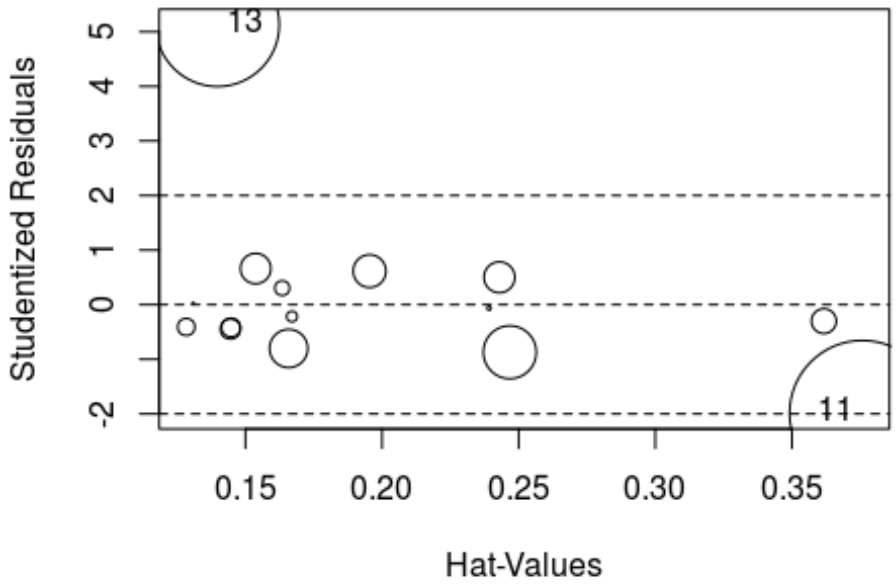
```
##    rstudent unadjusted p-value Bonferonni p
## 13 5.126833         0.00032993     0.004949
```

Tal como se apreció en el estudio de normalidad de los residuos, la observación 13 tiene un residuo estandarizado >3 (más de 3 veces la desviación estándar de los residuos) por lo que se considera un dato atípico. El siguiente paso es determinar si es influyente.

```
summary(influence.measures(modelo))
```

```
## Potentially influential observations of
##   lm(formula = peso ~ volumen + tipo_tapas, data = datos) :
##
##    dfb.1_ dfb.vlmn dfb.tp_t dffit   cov.r   cook.d hat
## 4  -0.16   0.18    -0.10    -0.23    1.98_*  0.02   0.36
## 11  0.70  -1.26_*   0.57    -1.54_*  0.83    0.64   0.38
## 13  0.31   0.67    -1.31_*   2.07_*  0.04_*  0.46   0.14
```

```
influencePlot(modelo)
```




```
##      StudRes       Hat     CookD
## 11 -1.989711 0.3757842 0.6372979
## 13  5.126833 0.1397761 0.4581972
```

El análisis muestran varias observaciones influyentes aunque ninguna excede los límites de preocupación para los valores de *Leverages hat* (>2.5x(2+1)/15 = 0.5) o *Distancia Cook* (>1). Estudios más exhaustivos consistirían en rehacer el modelo sin las observaciones y ver el impacto.

6.Conclusión

El modelo lineal múltiple

$$Peso\ libro = 13.91557 + 0.71795\ volumen + 184.04727\ tipotapas$$

es capaz de explicar el 92.75% de la variabilidad observada en la esperanza de vida (R-squared: 0.9275, Adjusted R-squared: 0.9154). El test F muestra que es significativo (1.455e-07). Se satisfacen todas las condiciones para este tipo de regresión.

Extensión del modelo lineal

Los modelos de regresión lineal presentan dos grandes ventajas, que son capaces de describir con suficiente precisión muchos escenarios que se dan en el mundo real y que los resultados son fácilmente interpretables. Sin embargo, para que sean totalmente válidos se tienen que satisfacer una serie de condiciones muy restrictivas que en la práctica no siempre se cumplen. Dos de las condiciones más importantes son que la relación entre los predictores y la variable respuesta debe ser *aditiva* y *lineal*. La aditividad implica que el efecto que tienen los cambios en el predictor X_j sobre la variable respuesta Y es independiente de los valores que tomen los otros predictores del modelo. La condición de linealidad implica que la variación en la variable respuesta Y debida al cambio de una unidad en el predictor X_j es constante, independientemente del valor de X_j. Existen dos aproximaciones clásicas que permiten relajar estas condiciones cuando se trabaja con modelos lineales.

Interacción de predictores

Supóngase que el departamento de ventas de una empresa quiere estudiar la influencia que tiene la publicidad a través de distintos canales sobre el número de ventas de un producto. Se dispone de un conjunto de datos que contiene el ingreso (en millones) conseguido por ventas en 200 regiones, así como la cantidad de presupuesto, también en millones, destinado a anuncios por ratio, TV y periódicos en cada una de ellas.

```
tv <- c(230.1, 44.5, 17.2, 151.5, 180.8, 8.7, 57.5, 120.2, 8.6, 199.8, 66.1,
    214.7, 23.8, 97.5, 204.1, 195.4, 67.8, 281.4, 69.2, 147.3, 218.4, 237.4,
    13.2, 228.3, 62.3, 262.9, 142.9, 240.1, 248.8, 70.6, 292.9, 112.9, 97.2,
    265.6, 95.7, 290.7, 266.9, 74.7, 43.1, 228, 202.5, 177, 293.6, 206.9, 25.1,
    175.1, 89.7, 239.9, 227.2, 66.9, 199.8, 100.4, 216.4, 182.6, 262.7, 198.9,
    7.3, 136.2, 210.8, 210.7, 53.5, 261.3, 239.3, 102.7, 131.1, 69, 31.5, 139.3,
    237.4, 216.8, 199.1, 109.8, 26.8, 129.4, 213.4, 16.9, 27.5, 120.5, 5.4,
    116, 76.4, 239.8, 75.3, 68.4, 213.5, 193.2, 76.3, 110.7, 88.3, 109.8, 134.3,
    28.6, 217.7, 250.9, 107.4, 163.3, 197.6, 184.9, 289.7, 135.2, 222.4, 296.4,
    280.2, 187.9, 238.2, 137.9, 25, 90.4, 13.1, 255.4, 225.8, 241.7, 175.7,
    209.6, 78.2, 75.1, 139.2, 76.4, 125.7, 19.4, 141.3, 18.8, 224, 123.1, 229.5,
    87.2, 7.8, 80.2, 220.3, 59.6, 0.7, 265.2, 8.4, 219.8, 36.9, 48.3, 25.6,
    273.7, 43, 184.9, 73.4, 193.7, 220.5, 104.6, 96.2, 140.3, 240.1, 243.2,
    38, 44.7, 280.7, 121, 197.6, 171.3, 187.8, 4.1, 93.9, 149.8, 11.7, 131.7,
    172.5, 85.7, 188.4, 163.5, 117.2, 234.5, 17.9, 206.8, 215.4, 284.3, 50,
    164.5, 19.6, 168.4, 222.4, 276.9, 248.4, 170.2, 276.7, 165.6, 156.6, 218.5,
    56.2, 287.6, 253.8, 205, 139.5, 191.1, 286, 18.7, 39.5, 75.5, 17.2, 166.8,
    149.7, 38.2, 94.2, 177, 283.6, 232.1)
```

```
radio <- c(37.8, 39.3, 45.9, 41.3, 10.8, 48.9, 32.8, 19.6, 2.1, 2.6, 5.8, 24,
    35.1, 7.6, 32.9, 47.7, 36.6, 39.6, 20.5, 23.9, 27.7, 5.1, 15.9, 16.9, 12.6,
    3.5, 29.3, 16.7, 27.1, 16, 28.3, 17.4, 1.5, 20, 1.4, 4.1, 43.8, 49.4, 26.7,
    37.7, 22.3, 33.4, 27.7, 8.4, 25.7, 22.5, 9.9, 41.5, 15.8, 11.7, 3.1, 9.6,
    41.7, 46.2, 28.8, 49.4, 28.1, 19.2, 49.6, 29.5, 2, 42.7, 15.5, 29.6, 42.8,
    9.3, 24.6, 14.5, 27.5, 43.9, 30.6, 14.3, 33, 5.7, 24.6, 43.7, 1.6, 28.5,
    29.9, 7.7, 26.7, 4.1, 20.3, 44.5, 43, 18.4, 27.5, 40.6, 25.5, 47.8, 4.9,
    1.5, 33.5, 36.5, 14, 31.6, 3.5, 21, 42.3, 41.7, 4.3, 36.3, 10.1, 17.2, 34.3,
    46.4, 11, 0.3, 0.4, 26.9, 8.2, 38, 15.4, 20.6, 46.8, 35, 14.3, 0.8, 36.9,
    16, 26.8, 21.7, 2.4, 34.6, 32.3, 11.8, 38.9, 0, 49, 12, 39.6, 2.9, 27.2,
    33.5, 38.6, 47, 39, 28.9, 25.9, 43.9, 17, 35.4, 33.2, 5.7, 14.8, 1.9, 7.3,
    49, 40.3, 25.8, 13.9, 8.4, 23.3, 39.7, 21.1, 11.6, 43.5, 1.3, 36.9, 18.4,
    18.1, 35.8, 18.1, 36.8, 14.7, 3.4, 37.6, 5.2, 23.6, 10.6, 11.6, 20.9, 20.1,
    7.1, 3.4, 48.9, 30.2, 7.8, 2.3, 10, 2.6, 5.4, 5.7, 43, 21.3, 45.1, 2.1,
    28.7, 13.9, 12.1, 41.1, 10.8, 4.1, 42, 35.6, 3.7, 4.9, 9.3, 42, 8.6)
periodico <- c(69.2, 45.1, 69.3, 58.5, 58.4, 75, 23.5, 11.6, 1, 21.2, 24.2,
    4, 65.9, 7.2, 46, 52.9, 114, 55.8, 18.3, 19.1, 53.4, 23.5, 49.6, 26.2, 18.3,
    19.5, 12.6, 22.9, 22.9, 40.8, 43.2, 38.6, 30, 0.3, 7.4, 8.5, 5, 45.7, 35.1,
    32, 31.6, 38.7, 1.8, 26.4, 43.3, 31.5, 35.7, 18.5, 49.9, 36.8, 34.6, 3.6,
    39.6, 58.7, 15.9, 60, 41.4, 16.6, 37.7, 9.3, 21.4, 54.7, 27.3, 8.4, 28.9,
    0.9, 2.2, 10.2, 11, 27.2, 38.7, 31.7, 19.3, 31.3, 13.1, 89.4, 20.7, 14.2,
    9.4, 23.1, 22.3, 36.9, 32.5, 35.6, 33.8, 65.7, 16, 63.2, 73.4, 51.4, 9.3,
    33, 59, 72.3, 10.9, 52.9, 5.9, 22, 51.2, 45.9, 49.8, 100.9, 21.4, 17.9,
    5.3, 59, 29.7, 23.2, 25.6, 5.5, 56.5, 23.2, 2.4, 10.7, 34.5, 52.7, 25.6,
    14.8, 79.2, 22.3, 46.2, 50.4, 15.6, 12.4, 74.2, 25.9, 50.6, 9.2, 3.2, 43.1,
    8.7, 43, 2.1, 45.1, 65.6, 8.5, 9.3, 59.7, 20.5, 1.7, 12.9, 75.6, 37.9, 34.4,
    38.9, 9, 8.7, 44.3, 11.9, 20.6, 37, 48.7, 14.2, 37.7, 9.5, 5.7, 50.5, 24.3,
    45.2, 34.6, 30.7, 49.3, 25.6, 7.4, 5.4, 84.8, 21.6, 19.4, 57.6, 6.4, 18.4,
    47.4, 17, 12.8, 13.1, 41.8, 20.3, 35.2, 23.7, 17.6, 8.3, 27.4, 29.7, 71.8,
    30, 19.6, 26.6, 18.2, 3.7, 23.4, 5.8, 6, 31.6, 3.6, 6, 13.8, 8.1, 6.4, 66.2,
    8.7)
ventas <- c(22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, 13.2, 4.8, 10.6, 8.6, 17.4,
    9.2, 9.7, 19, 22.4, 12.5, 24.4, 11.3, 14.6, 18, 12.5, 5.6, 15.5, 9.7, 12,
    15, 15.9, 18.9, 10.5, 21.4, 11.9, 9.6, 17.4, 9.5, 12.8, 25.4, 14.7, 10.1,
    21.5, 16.6, 17.1, 20.7, 12.9, 8.5, 14.9, 10.6, 23.2, 14.8, 9.7, 11.4, 10.7,
    22.6, 21.2, 20.2, 23.7, 5.5, 13.2, 23.8, 18.4, 8.1, 24.2, 15.7, 14, 18,
    9.3, 9.5, 13.4, 18.9, 22.3, 18.3, 12.4, 8.8, 11, 17, 8.7, 6.9, 14.2, 5.3,
    11, 11.8, 12.3, 11.3, 13.6, 21.7, 15.2, 12, 16, 12.9, 16.7, 11.2, 7.3, 19.4,
    22.2, 11.5, 16.9, 11.7, 15.5, 25.4, 17.2, 11.7, 23.8, 14.8, 14.7, 20.7,
    19.2, 7.2, 8.7, 5.3, 19.8, 13.4, 21.8, 14.1, 15.9, 14.6, 12.6, 12.2, 9.4,
    15.9, 6.6, 15.5, 7, 11.6, 15.2, 19.7, 10.6, 6.6, 8.8, 24.7, 9.7, 1.6, 12.7,
    5.7, 19.6, 10.8, 11.6, 9.5, 20.8, 9.6, 20.7, 10.9, 19.2, 20.1, 10.4, 11.4,
    10.3, 13.2, 25.4, 10.9, 10.1, 16.1, 11.6, 16.6, 19, 15.6, 3.2, 15.3, 10.1,
    7.3, 12.9, 14.4, 13.3, 14.9, 18, 11.9, 11.9, 8, 12.2, 17.1, 15, 8.4, 14.5,
    7.6, 11.7, 11.5, 27, 20.2, 11.7, 11.8, 12.6, 10.5, 12.2, 8.7, 26.2, 17.6,
    22.6, 10.3, 17.3, 15.9, 6.7, 10.8, 9.9, 5.9, 19.6, 17.3, 7.6, 9.7, 12.8,
    25.5, 13.4)

datos <- data.frame(tv, radio, periodico, ventas)
```

El modelo lineal múltiple que se obtiene empleando las variables *tv*, *radio* y *periodico* como predictores de ventas es el siguiente:

```
modelo <- lm(formula = ventas ~ tv + radio + periodico, data = datos)
summary(modelo)
```

```
##
## Call:
## lm(formula = ventas ~ tv + radio + periodico, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## tv           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## periodico   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

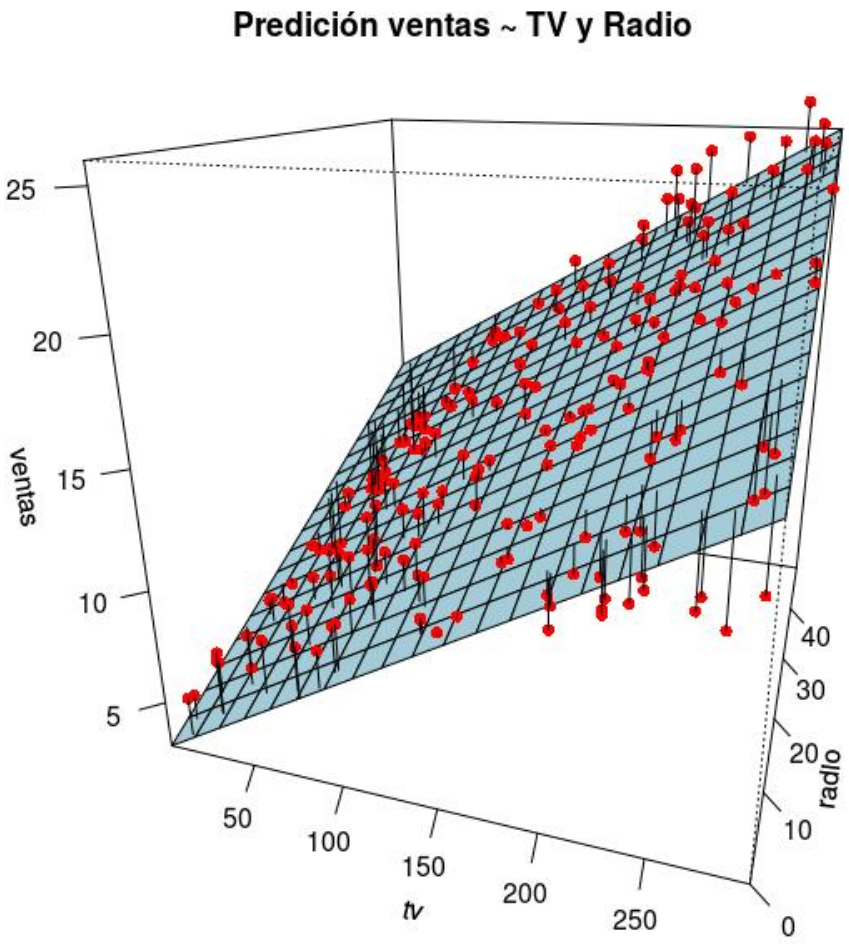
Acorde al *p-value* obtenido para el coeficiente parcial de correlación de *periodico*, esta variable no contribuye de forma significativa al modelo. Como resultado de este análisis se concluye que las variables *tv* y *radio* están asociadas con la cantidad de ventas.

```
modelo <- update(modelo, .~. -periodico)
summary(modelo)
```

```
##
## Call:
## lm(formula = ventas ~ tv + radio, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## tv           0.04575    0.00139  32.909   <2e-16 ***
## radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

```r
# Al ser un modelo con dos predictores continuos se puede representar en 3D
rango_tv <- range(datos$tv)
nuevos_valores_tv <- seq(from = rango_tv[1], to = rango_tv[2], length.out = 20)
rango_radio <- range(datos$radio)
nuevos_valores_radio <- seq(from = rango_radio[1], to = rango_radio[2], length.out
= 20)

predicciones <- outer(X = nuevos_valores_tv, Y = nuevos_valores_radio, FUN =
function(tv, radio) {
    predict(object = modelo, newdata = data.frame(tv, radio))
})

superficie <- persp(x = nuevos_valores_tv, y = nuevos_valores_radio, z =
predicciones, theta = 18, phi = 20, col = "lightblue", shade = 0.1, xlab = "tv",
ylab = "radio", zlab = "ventas", ticktype = "detailed", main = "Predición ventas ~
TV y Radio")

observaciones <- trans3d(datos$tv, datos$radio, datos$ventas, superficie)
error <- trans3d(datos$tv, datos$radio, fitted(modelo), superficie)
points(observaciones, col = "red", pch = 16)
segments(observaciones$x, observaciones$y, error$x, error$y)
```



El modelo lineal a partir del cual se han obtenido las conclusiones asume que el efecto sobre las ventas debido a un incremento en el presupuesto de uno de los medios de comunicación es independiente del presupuesto gastado en los otros. Por ejemplo, el modelo lineal considera que el efecto promedio sobre las ventas debido a aumentar en una unidad el presupuesto de anuncios en TV es siempre de 0.04575, independientemente de la cantidad invertida en anuncios por radio. Sin embargo, la representación gráfica muestra que el modelo tiende a sobrevalorar las ventas cuando el presupuesto es muy alto en uno de los medios pero muy bajo en el otro. Por contra, los valores de ventas predichos por el modelo están por debajo de las ventas reales cuando el presupuesto está repartido de forma equitativa entre ambos medios. Este comportamiento sugiere que existe interacción entre los predictores, por lo que el efecto de cada uno de ellos sobre la variable respuesta depende en cierta medida del valor que tome el otro predictor.

Tal y como se ha definido previamente, un modelo lineal con dos predictores sigue la ecuación:

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

Acorde a esta definición, el incremento de una unidad en el predictor X_1 produce un incremento promedio de la variable Y de β_1. Modificaciones en el predictor X_2 no alteran este hecho, y lo mismo ocurre con X_2 respecto a X_1. Para que el modelo pueda contemplar la interacción se introduce un tercer predictor, llamado *interaction term*, que se construye con el producto de los predictores X_1 y X_2.

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

La reorganización de los términos resulta en:

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

El efecto de X_1 sobre Y ya no es constante, sino que depende del valor que tome X_2.

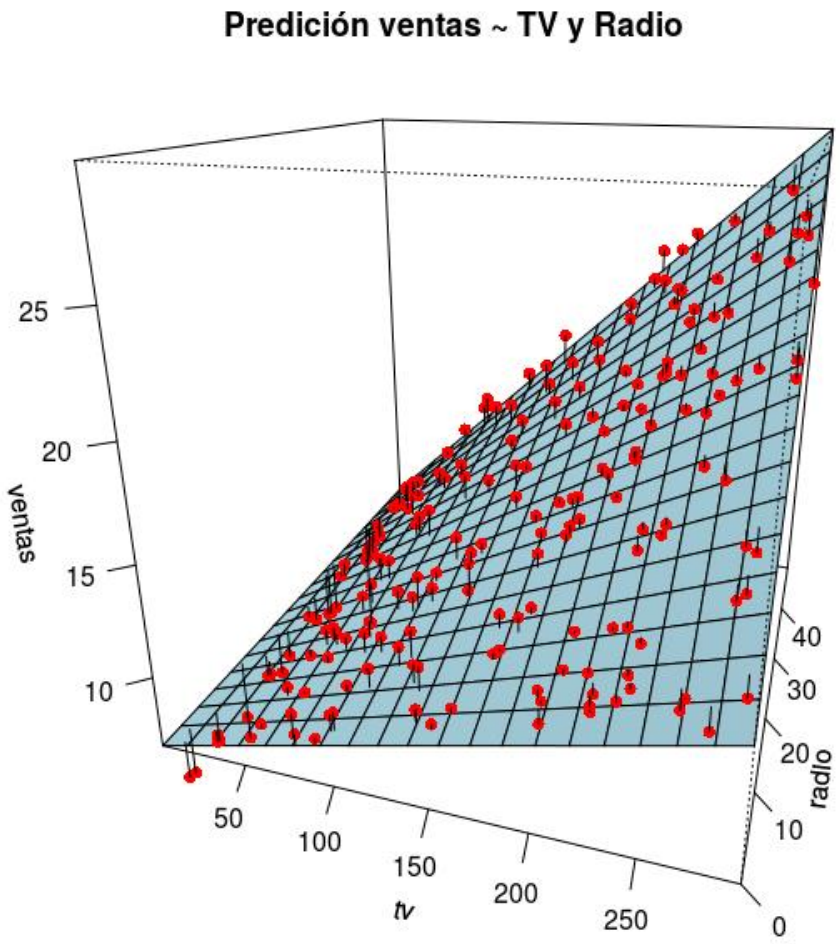
En R se puede introducir interacción entre predictores de dos formas, indicando los predictores individuales y entre cuales se quiere evaluar la interacción, o bien de forma directa.

```
modelo_interaccion <- lm(formula = ventas ~ tv + radio + tv:radio, data = datos)
summary(modelo_interaccion)
```

```
##
## Call:
## lm(formula = ventas ~ tv + radio + tv:radio, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3366 -0.4028  0.1831  0.5948  1.5246
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
## tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
## radio       2.886e-02  8.905e-03   3.241   0.0014 **
## tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9435 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```r
#lm(formula = ventas ~ tv * radio, data = datos) es equivalente.

# Al ser un modelo_interaccion con dos predictores continuos se puede
# representar en 3D
rango_tv <- range(datos$tv)
nuevos_valores_tv <- seq(from = rango_tv[1], to = rango_tv[2], length.out = 20)
rango_radio <- range(datos$radio)
nuevos_valores_radio <- seq(from = rango_radio[1], to = rango_radio[2], length.out
= 20)

# La funcion outer() permite aplicar una funcion a cada combinacion de los
# parametros x, y pasados como argumento, es una alternativa a utilizar
# expand.grid()

predicciones <- outer(X = nuevos_valores_tv, Y = nuevos_valores_radio, FUN =
function(tv, radio) {
    predict(object = modelo_interaccion, newdata = data.frame(tv, radio))
})

superficie <- persp(x = nuevos_valores_tv, y = nuevos_valores_radio, z =
predicciones, theta = 18, phi = 20, col = "lightblue", shade = 0.1, xlab = "tv",
ylab = "radio", zlab = "ventas", ticktype = "detailed", main = "Predición ventas ~
TV y Radio")
# Se pueden representar las observaciones a partir de las cuales se ha
# creado la superficie así como segmentos que midan la distancia respecto al
# modelo_interaccion generado.
observaciones <- trans3d(datos$tv, datos$radio, datos$ventas, superficie)
error <- trans3d(datos$tv, datos$radio, fitted(modelo_interaccion), superficie)
points(observaciones, col = "red", pch = 16)
segments(observaciones$x, observaciones$y, error$x, error$y)
```



Los resultados muestran una evidencia clara de que la interacción *tv* x *radio* es significativa y de que el modelo que incorpora la interacción (*Adjusted R-squared* = 0.9673) es superior al modelo que solo contemplaba el efecto de los predictores por separado (*Adjusted R-squared* = 0.8956).

Se puede emplear un ANOVA para realizar un test de hipótesis y obtener un *p-value* que evalúe la hipótesis nula de que ambos modelos se ajustan a los datos igual de bien.

```r
anova(modelo, modelo_interaccion)
```

```
## Analysis of Variance Table
##
## Model 1: ventas ~ tv + radio
## Model 2: ventas ~ tv + radio + tv:radio
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 556.91
## 2    196 174.48  1    382.43 429.59 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

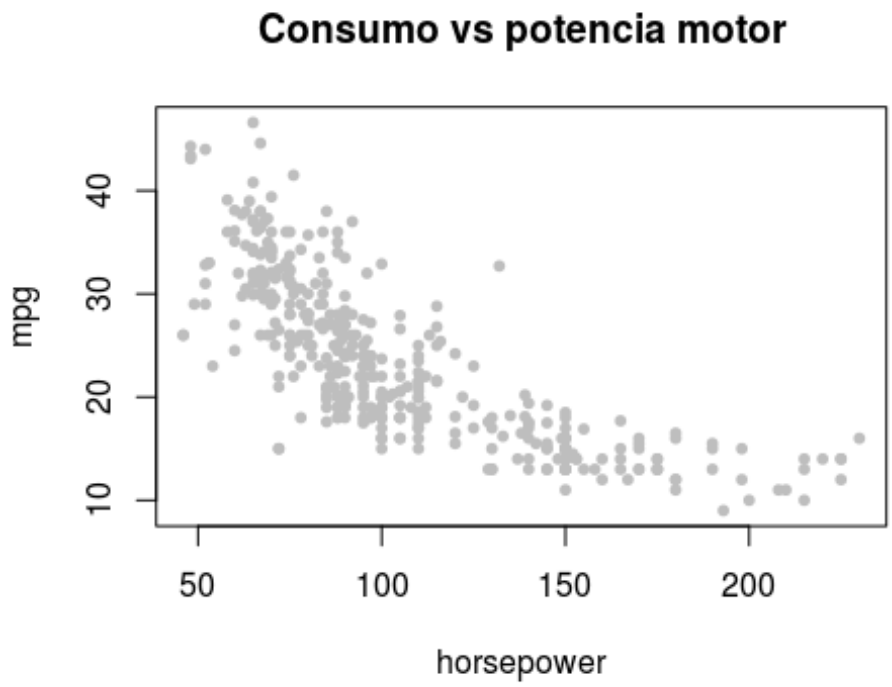
En los modelos de regresión lineal múltiple que incorporan interacciones entre predictores hay que tener en cuenta el *hierarchical principle*, según el cual, si se incorpora al modelo una interacción entre predictores, se deben incluir siempre los predictores individuales que participan en la interacción, independientemente de que su *p-value* sega significativo o no.

La interacción entre predictores no está limitada a predictores cuantitativos, también puede crearse interacción entre predictor cuantitativo y cualitativo.

Regresión polinomial

*Una marca de coches quiere generar un modelo de regresión que permita predecir el consumo de combustible (*mpg*) en función de la potencia del motor (*horsepower).

```r
library(ISLR)
attach(Auto)

plot(x = horsepower, y = mpg, main = "Consumo vs potencia motor", pch = 20,
    col = "grey")
```



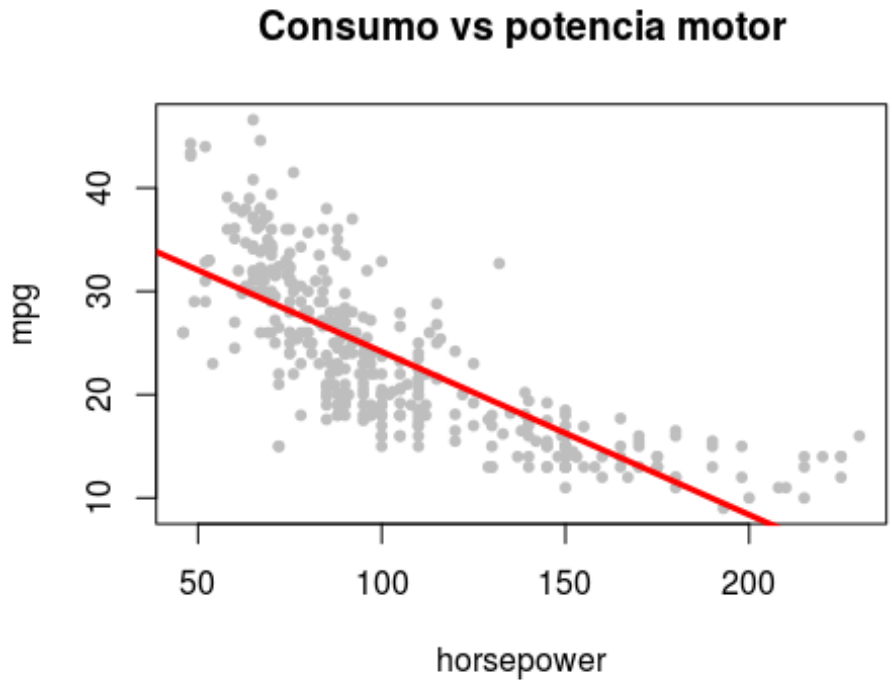
La representación gráfica de los datos muestra una fuerte asociación entre el consumo y la potencia del motor. La distribución de las observaciones apunta a que la relación entre ambas variables tiene cierta curvatura, por lo que un modelo lineal no puede captarla por completo.

```
attach(Auto)
modelo_lineal <- lm(formula = mpg ~ horsepower, data = Auto)
summary(modelo_lineal)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
plot(x = horsepower, y = mpg, main = "Consumo vs potencia motor", pch = 20,
    col = "grey")
abline(modelo_lineal, lwd = 3, col = "red")
```



Una forma de incorporar asociaciones no lineales a un modelo lineal es mediante transformaciones de los predictores incluidos en el modelo, por ejemplo, elevándolos a distintas potencias. En este caso, el tipo de curvatura es de tipo cuadrática, por lo que un polinomio de segundo grado podría mejorar el modelo.

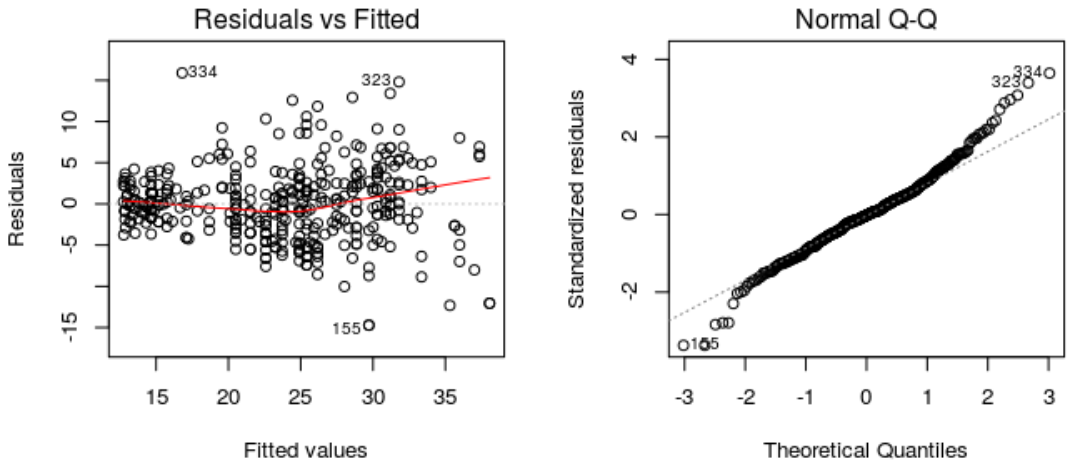
En R se pueden generar modelos de regresión polinómica de diferentes formas:

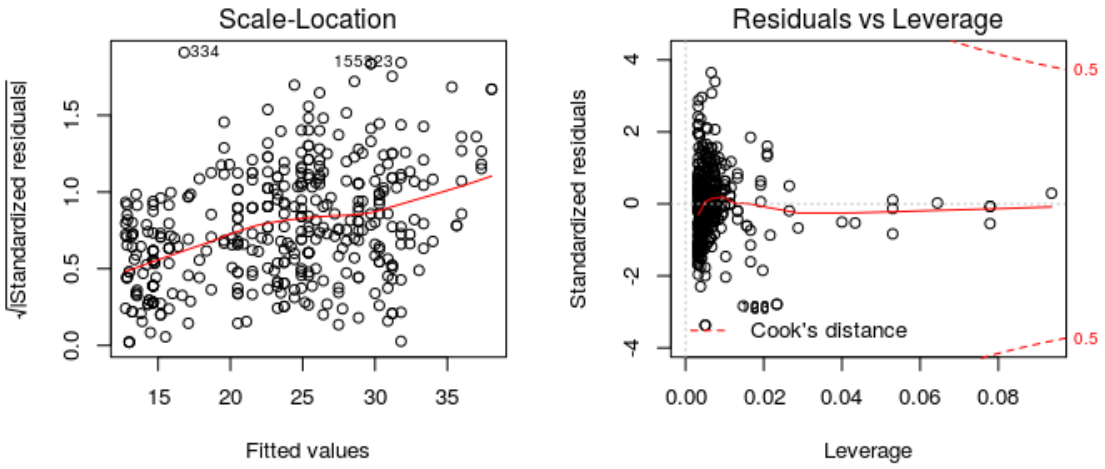
- Identificando cada elemento del polinomio: `modelo_pol2 <- lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)` El uso de `I()` es necesario ya que el símbolo ^ tiene otra función dentro de las formula de R.
- Con la función `poly()`: `lm(formula = mpg ~ poly(horsepower, 2),data = Auto)`

```r
modelo_cuadratico <- lm(formula = mpg ~ poly(horsepower, 2), data = Auto)
summary(modelo_cuadratico)
```

```
##
## Call:
## lm(formula = mpg ~ poly(horsepower, 2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.4459     0.2209  106.13   <2e-16 ***
## poly(horsepower, 2)1 -120.1377     4.3739  -27.47   <2e-16 ***
## poly(horsepower, 2)2   44.0895     4.3739   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686
## F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(modelo_cuadratico)
```





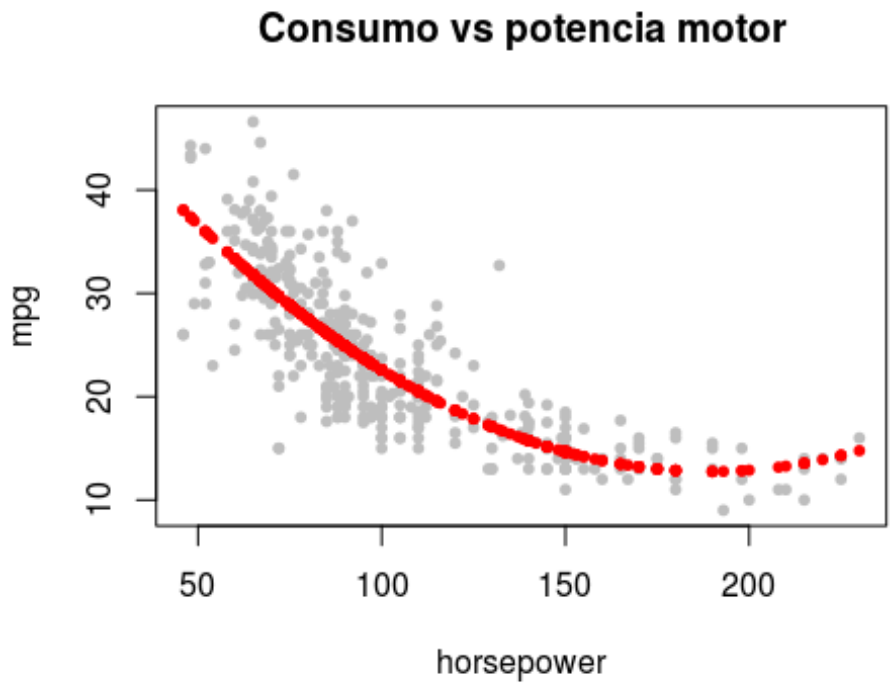
El valor R^2 del modelo cuadrático (0.6876) es mayor que el obtenido con el modelo lineal simple (0.6059) y el *p-value* del término cuadrático es altamente significativo. Se puede concluir que el modelo cuadrático recoge mejor la verdadera relación entre el consumo de los vehículos y la potencia de su motor.

Es posible emplear un ANOVA para realizar un test de hipótesis y obtener un *p-value* que evalúe la hipótesis nula de que ambos modelos se ajustan a los datos igual de bien.

```
anova(modelo_lineal, modelo_cuadratico)
```

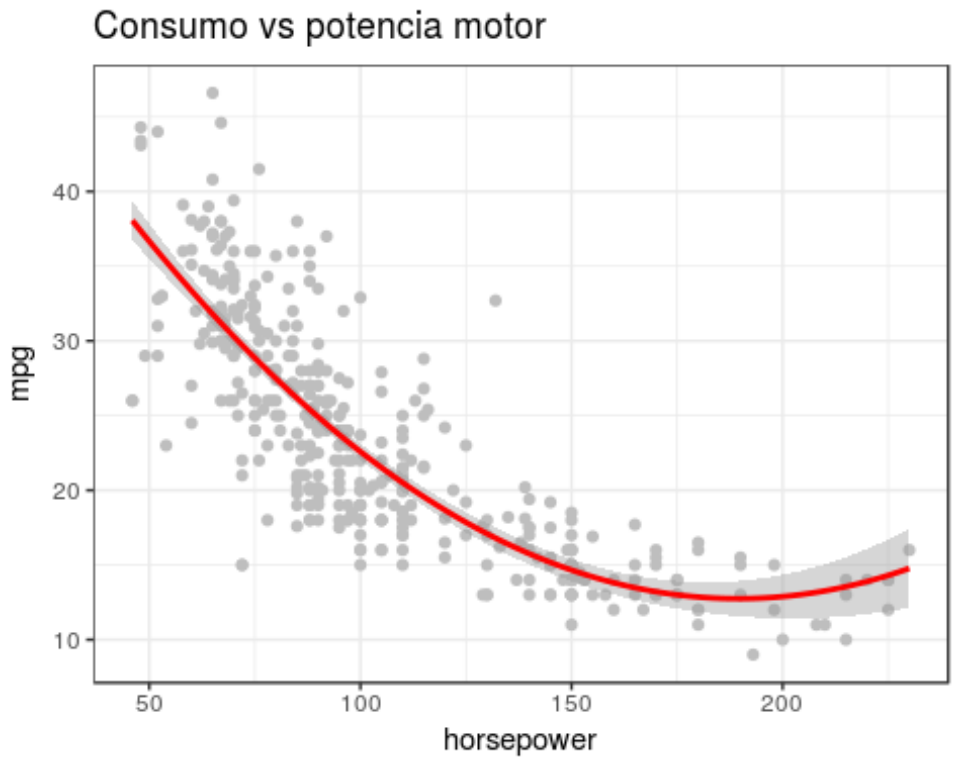
```
## Analysis of Variance Table
##
## Model 1: mpg ~ horsepower
## Model 2: mpg ~ poly(horsepower, 2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    390 9385.9
## 2    389 7442.0  1    1943.9 101.61 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(x = horsepower, y = mpg, main = "Consumo vs potencia motor", pch = 20,
    col = "grey")
points(x = horsepower, fitted(modelo_cuadratico), col = "red", pch = 20)
```



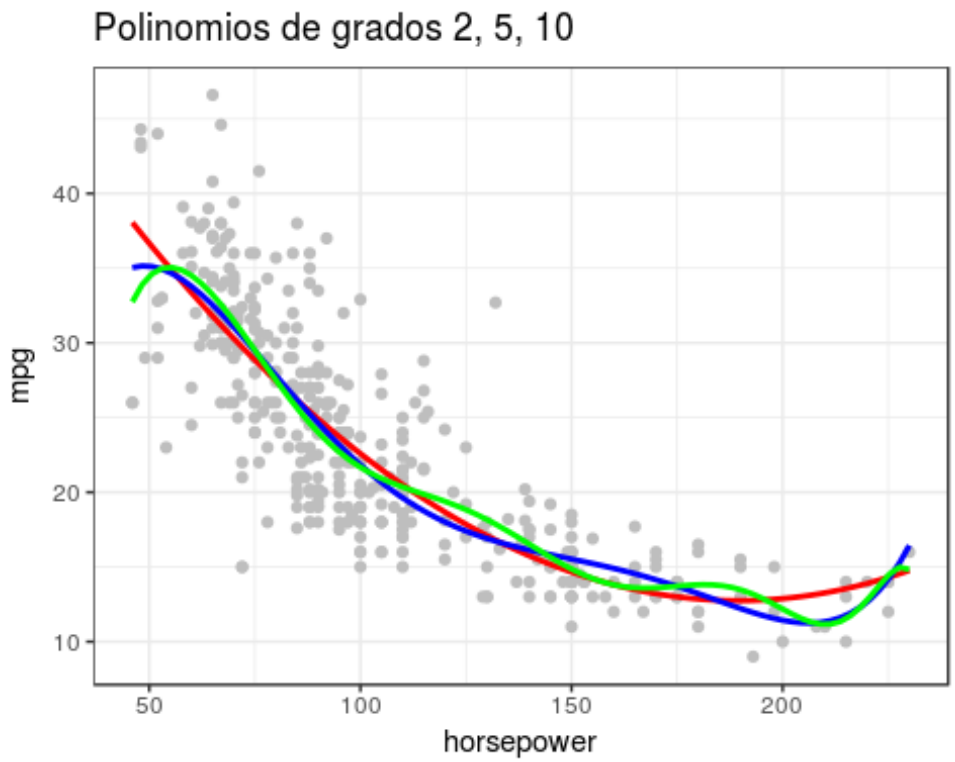
Misma representación con GGPLOT2

```
library(ggplot2)
ggplot(Auto, aes(x = horsepower, y = mpg)) +
 geom_point(colour = "grey") +
 stat_smooth(method = "lm",formula = y ~ poly(x, 2), colour = "red") + labs(title =
"Consumo vs potencia motor") +
 theme_bw()
```



La elección del grado del polinomio influye directamente en la flexibilidad del modelo. Cuanto mayor es el grado del polinomio más se ajusta el modelo a las observaciones, un polinomio de grado *nºobservaciones-1* pasa por todos los puntos. Por lo tanto, es importante no excederse en el grado del polinomio para no causar problemas de *overfitting*. En el capítulo *Validación de modelos de regresión: Cross-validation, OneLeaveOut, Bootstrap* se explica cómo emplear la validación-cruzada para identificar el grado adecuado.

```
library(ggplot2)
ggplot(Auto, aes(x = horsepower, y = mpg)) +
 geom_point(colour = "grey") +
 stat_smooth(method = "lm", formula = y ~ poly(x, 2), colour = "red", se = FALSE) +
 stat_smooth(method = "lm",formula = y ~ poly(x, 5), colour = "blue", se = FALSE) +
 stat_smooth(method = "lm", formula = y ~ poly(x, 10), colour = "green", se =
FALSE) +
 labs(title = "Polinomios de grados 2, 5, 10") +
 theme_bw()
```


Polinomios de grados 2, 5, 10
mpg
40
30
20
10
50
100
150
200
horsepower