Introducción a la Regresión Lineal Múltiple

Contenido

1. Introducción

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y cuantitativa) se determina a partir de un conjunto de variables independientes llamadas predictores ($X1$, $X2$, $X3$...). Es una extensión de la regresión lineal simple. 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 ella (esto último se debe analizar con cautela para no malinterpretar causa-efecto).

Tanto la variable dependiente como las independientes idealmente deben medirse con una escala de intervalos. Las variables nominales tales como religión, licenciatura o zona de residencia deben recodificarse como variables binarias (ficticios o dummy) u otros tipos de variables de contraste. Si se ha recopilado un gran número de variables independientes y se desea construir un modelo de regresión que incluya sólo variables que estén estadísticamente relacionadas con la variable dependiente, puede utilizar uno de los métodos de selección de variables para seleccionar las variables independientes. Para ver cómo se ajusta el modelo de regresión a los datos, puede examinar los residuos, identificar los puntos de influencia y generar otros tipos de diagnósticos que proporciona este procedimiento

La formulación matemática de estos modelos es la siguiente:

$$Y = m(X_1, X_2, \dots, X_k) + \mu_i$$

Donde μ_i es el error de observación debido a variables no controladas

En el modelo de Regresión Lineal General se "supone" que la función de regresión $Y = m(X_1, X_2, \dots, X_k) + \mu_i$ es lineal. Por tanto, la expresión matemática del modelo de regresión lineal general es:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + e_i$$

Donde:

- β_0: es la ordenada en el origen, el valor de la variable dependiente Y cuando todos los predictores son cero.
- β_i: es el efecto promedio que tiene el incremento en una unidad de la variable predictora Xi sobre la variable dependiente Y, manteniéndose

constantes el resto de variables. Se conocen como coeficientes parciales de regresión.

- *ei*: es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

Un primer objetivo es estimar los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$, y la función de distribución del error a partir de una muestra de n observaciones.

Es importante tener en cuenta que la magnitud de cada coeficiente parcial de regresión depende de las unidades en las que se mida la variable 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 estandarizar (sustraer la media y dividir entre la desviación estándar) las variables predictoras previo ajuste del modelo.

2. Interpretación y estadísticos de la regresión.

El coeficiente β_1 mide el cambio experimentado por Y asociado con una variación de X_{i1} en una unidad. De forma similar β_2 mide la variación experimentada por Y como consecuencia de un incremento de X_{i2} en una unidad. En ambos casos, el supuesto de que las demás variables explicativas permanecen constantes es crucial para esta interpretación de los coeficientes. En base a estos supuestos, los parámetros desconocidos β_1 y β_2 se denominan Coeficientes de regresión parciales, puesto que corresponden a los valores de las derivadas parciales de Y con respecto a X_{i1} y X_{i2}, respectivamente.

3. Condiciones para el uso de la RLM

3.1 Hipótesis relativa a las perturbaciones

1) Toda perturbación aleatoria tiene una media igual a cero, en otras palabras, las perturbaciones son variables aleatorias con esperanza matemática cero.

$$E(u_i) = 0$$

2) Todas las perturbaciones aleatorias tienen la misma varianza es decir que las varianzas son constantes para toda la muestra. Esta característica de la varianza recibe el nombre de "**homocedasticidad**" o sea.

$$E(u_i^2) = \sigma_u^2 = \sigma_i^2; \quad para \quad i = 1,2,3,.....n$$

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, se distribuyen de forma aleatoria manteniendo una misma dispersión y sin ningún patrón específico. Una distribución cónica es un claro identificador de falta de homocedasticidad. También se puede recurrir a contrastes de homocedasticidad como el test de Breusch-Pagan.

3) Las perturbaciones aleatorias son independientes entre sí, lo que equivale a decir que las covarianzas son nulas. (**independencia**)

$$\text{Cov}\,(u_i u_j) = E(u_i u_j) = 0 \quad \text{para } i \neq j$$

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.

4) Las perturbaciones se distribuyen normalmente con $\overline{X} = 0$ y σ^2

Los residuos se tienen que distribuir de forma normal, con media igual a 0. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a test de hipótesis de normalidad.

3.2 HIPOTESIS RELATIVA A LAS VARIABLES

1) Las variables predeterminadas X_1, X_2, $X_{3,..}$ no son aleatorias, son determinadas por el investigador; otra forma de ver es: Relación lineal entre los predictores numéricos y la variable respuesta, cada predictor numérico tiene que estar linealmente relacionado 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.

2) La variable respuesta Y toma valores dados por las respuestas de los observadores, por lo tanto, es una variable dependiente aleatoria, sus parámetros $\overline{Y}$ y σ_y^2

El modelo es: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots + \beta_k X_k + u_i$

Utilizando valores esperados:

$$E(Y_i) = \mathrm{E}(\beta_0) + E(\beta_1 X_1) + E(\beta_2 X_2) + E(\beta_3 X_3) + \ldots\ldots\ldots + E(\beta_k X_k) + E(u_i)$$

El esperado de un producto es igual al producto de los esperados de los factores Y por otro lado sus B son constantes y las X_i son constantes fijas, y como el esperado de una constante es la constante misma, tenemos:

$$E(Y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots + \beta_k X_k + E(u_i)$$

Pero según hipótesis que E(ui)= 0, queda:

$$E(Y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots + \beta_k X_k$$

La varianza es:

$$S_X^2 = \frac{1}{N}\sum(Xi - \overline{X})^2 = E(Xi - \overline{X})^2 = E(X_i - E(X_i))^2$$

Por tanto: $\sigma_Y^2 = E[Y_i - E(Y_i)]^2; \quad reemplazando \quad Y_i; E(Y_i)$

$$\sigma_Y^2 = E[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + u_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_k X_k]^2$$

Reduciendo términos semejantes: $\sigma_Y^2 = E(u^2)$, por hipótesis se conoce $E(u^2) = \sigma_u^2$, entonces:

$$\sigma_Y^2 = E(u^2) = \sigma_u^2$$

Además, las Y_i son independientes entre sí, es decir que sus covarianzas son nulas:

$$E[Y_i - E(Y_i)][Y_{i'} - E(Y_{i'})] = E(u_i, u_{i'}) = 0 \quad (i \neq j)$$

3) Las variables en X_i e Y_i representan magnitudes numéricas que se obtienen sin error de observación, es decir que no contienen ningún error de medida

3.3 HIPOTESIS RELATIVAS A LOS PARAMETROS

1) Los parámetros estructurales B son constantes para todas las unidades de la muestra y para todas las muestras posibles.

Existen otras dos hipótesis relativas a las variables que son llamadas "A" y "B"

Hipótesis "A".

Las variables en X_i son independientes entre sí y no deben tener relación lineal exacta entre sí. Por ejemplo X_1 es independiente de X_2 ….X_k. Cuando existe alguna relación lineal aun cuando no sea exacta entre dos o más X_i entonces decimos que hay "Multicolinealidad" o "Intercorrelación" (vea más abajo)

Hipótesis "B".

El número de observaciones o sea el tamaño de la muestra tiene que ser igual o mayor que el número de parámetros incógnitas a estimar; para poder estimar los valores de los parámetros se deben resolver ecuaciones que deben tener igual o menor número de incógnitas, pero no podrá haber mayor número de incógnitas que de ecuaciones.

3.4 OTRAS CONSIDERACIONES.

1) **Parsimonia**:

Este término hace referencia a que el mejor modelo es aquel capaz de explicar con mayor precisión la variabilidad observada en la variable

respuesta empleando el menor número de predictores, por lo tanto, con menos asunciones.

2) **Valores atípicos, con alto leverage o influyentes:**

   Es importante identificar observaciones que sean atípicas o 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.

3) **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 colineales 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 encontrar 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 a la estimación y contraste de un modelo. Los pasos recomendados a seguir son:

   - Si el coeficiente de determinación R^2 es alto, pero ninguno de los predictores resulta significativo, hay indicios de colinealidad.

- Calcular una matriz 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 multicolinealidad. 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 a 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 límites 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 realmente en la variable respuesta. Esta medida no suele tener mucho

impacto en el modelo en cuanto a su capacidad predictiva 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 colineales en un único predictor, aunque con el riesgo de perder su interpretación.

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.

Dado que las condiciones se verifican a partir de los residuos, primero se suele generar el modelo y después se valida. De hecho, el ajuste de un modelo debe verse como un proceso iterativo en el que se ajusta un modelo inicial, se evalúa mediante sus residuos y se mejora. Así hasta llegar a un modelo óptimo

4. Tamaño de la muestra:

No se trata de una condición de por sí, pero, si no se dispone de suficientes observaciones, predictores que no son realmente influyentes podrían parecerlo. En el libro *Hanbook of biological statistics* recomiendan que el número de observaciones sea 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. De hecho, el ajuste de un modelo debe verse como un proceso iterativo en el que se ajusta el modelo, se evalúan sus residuos y se mejora. Así hasta llegar a un modelo óptimo.

En el siguiente cuadro se resumen las hipótesis del modelo de regresión lineal general

HIPÓTESIS del Modelo de Regresión Lineal General	
En base u_i	**En base a la variable respuesta Y**
$E(u_i) = 0$	$E(Y/X_{i1}, X_{i2}, X_{i3}, \ldots\ldots\ldots, X_{ik}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik}$
Homocedasticidad $Var(u_i) = \sigma^2$	Homocedasticidad $\mathrm{Var}(Y/X_{i1}, X_{i2}, X_{i3}, .., X_{ik}) = \sigma^2$
Independencia, $Cov(u_i, u_j) = 0$ los errores, u_i, son independientes	Independencia las observaciones, Y_i, son independientes
Normalidad $u_i \in N(0, \sigma^2)$	Normalidad $Y/x_{i1}, x_{i2}, \ldots, x_{ik} \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}, \sigma^2)$
$n > k + 1$	$n > k + 1$
Las variables regresoras son linealmente independientes	Las variables regresoras son linealmente independientes

5. Selección de predictores

La evaluación de un modelo de regresión múltiple, así como la elección de qué predictores se deben 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: Selección de predictores y mejor modelo lineal múltiple: subset selection, ridge regression, lasso regression y dimension reduction.

5.1 Selección de los predictores

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

a) Método jerárquico.

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

b) Método de entrada forzada

se introducen todos los predictores simultáneamente.

c) 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 $\beta 0$. 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 Cp , AIC, BIC y R^2-*ajustado*, cada uno de ellos con 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.

Es frecuente encontrar ejemplos en los que la selección de los predictores se basa en el *p-value* asociado a cada uno. Si bien este método es sencillo e intuitivo, presenta múltiples inconvenientes: la inflación del error tipo I debida a las múltiples comparaciones, la eliminación de los predictores menos significativos tiende a incrementar la significancia de los otros predictores … Por esta razón, a excepción de casos muy sencillos con pocos predictores, es preferible no emplear los *p-values* como criterio de selección.

En el caso de variables categóricas, si al menos uno de sus niveles es significativo, se considera que la variable es significativa. Cabe mencionar que, si una variable se excluye del modelo como predictor, significa que no aporta información adicional al modelo, pero sí puede estar relacionada con la variable respuesta.

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

6. Evaluación del modelo en conjunto

Al igual que ocurre en los modelos lineales simples, R^2 (coeficiente de determinación) 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 respecto al total de variabilidad. 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. Es por esto que R^2 no puede utilizarse para comparar modelos con distinto número de predictores.

$R^2 ajustado$ 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. $R^2ajustado$ permite encontrar el mejor modelo, aquel que consigue explicar mejor la variabilidad de Y con el menor número de predictores. Si bien es un método para evaluar la bondad de ajuste muy utilizado, hay otros.

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

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 uno de los 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 mejor de lo esperado por azar.

7. Variables nominales/categóricas como predictores

Cuando se introduce una variable categórica como predictor, un nivel se considera el de referencia (normalmente codificado como 0) y el resto de niveles se comparan con él. En el caso de que el predictor categórico tenga más de dos niveles, se generan lo que se conoce como variables *dummy*, que son 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 regresión βi 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.

La idea de variables *dummy* se entiende mejor con un ejemplo. Supóngase un modelo en el que la variable respuesta *peso* se predice en función de la *altura* y *nacionalidad* del sujeto. La variable nacionalidad es cualitativa con 3 niveles (americana, europea y asiática). A pesar de que el predictor inicial es *nacionalidad*, se crea una variable nueva por cada nivel, cuyo valor puede ser 0 o 1. De tal forma que la ecuación del modelo completo es:

$$peso = \beta_0 + \beta_1 altura + \beta_2 americana + \beta_3 europea + \beta_4 asiatica$$

Si el sujeto es europeo, las variables dummy *americana* y *asiatica* se consideran 0, de forma que el modelo para este caso se queda en:

$$peso = \beta_0 + \beta_1 altura + \beta_3 europea$$

8. Validación cruzada (Cross-Validation)

Una vez seleccionado el mejor modelo, se tiene que comprobar su validez prediciendo nuevas observaciones que no se hayan 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. Para una descripción más detallada de la validación cruzada consultar: Validación de modelos de regresión: Cross-validation, OneLeaveOut, Bootstrap.

9. Identificación de valores atípicos (outliers), de alto leverage o influyentes.

Independientemente de que el modelo se haya podido aceptar, siempre es conveniente identificar si hay algún posible *outlier*, observación con alto *leverage* o 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 suele ser capaz de predecir mejor la mayoría de casos. Sin embargo, es muy importante prestar

atención a estos valores ya que, de no tratarse de 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. En el capítulo Regresión Lineal Simple se describe con detalle cómo realizar el análisis para detectarlos.

10. ESTIMADORES MINIMO CUADRATICOS

(Método de los mínimos cuadrados)

El modelo por analizar es:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + u_i$$

Que se calcula por medio de la ecuación siguiente (muestra):

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \cdots + \hat{\beta}_k X_{ik} + u_i$$

Otra forma de expresar es usando letras minúsculas

$$\hat{Y} = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + \ldots + b_k X_{ik}$$

Donde cada coeficiente de regresión β_i se estima mediante $\hat{\beta}_i$ o b_i de los datos muestrales.

Al ajustar un modelo de regresión lineal múltiple, en particular cuando el número de variables excede de 2, el conocimiento del álgebra matricial facilita considerablemente las manipulaciones matemáticas.

Estimadores mínimos cuadráticos:

Aplicando el principio de los mínimos cuadrados que consiste en obtener los valores de β con la condición de que "la suma de cuadrados de los residuos sea mínima", es decir: $\sum e_i^2 = Mínima$

Partiendo del modelo muestral:

$$\hat{Y} = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + \ldots + b_k X_{ik} + e_i$$

Despejando e_i, elevando al cuadrado ambos miembros y aplicando sumatorias, tenemos:

$$\sum e_i^2 = \sum (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2} - b_3 X_{i3} - \ldots - b_k X_{ik})^2$$

Para minimizar los errores, debemos derivar parcialmente respecto a cada uno de los estimadores e igualar a cero.

$$\frac{\partial \sum e_i^2}{\partial b_o} = -2\sum (Y_i - b_0 - b_1X_{i1} - b_2X_{i2} - b_3X_{i3} - \ldots - b_kX_{ik}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2\sum (Y_i - b_0 - b_1X_{i1} - b_2X_{i2} - b_3X_{i3} - \ldots - b_kX_{ik})\,(X_{1i}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = -2\sum (Y_i - b_0 - b_1X_{i1} - b_2X_{i2} - b_3X_{i3} - \ldots - b_kX_{ik})(X_{2i}) = 0$$

.

.

$$\frac{\partial \sum e_i^2}{\partial b_k} = -2\sum (Y_i - b_0 - b_1X_{i1} - b_2X_{i2} - b_3X_{i3} - \ldots - b_kX_{ik})(X_{ik}) = 0$$

Dividiendo entre -2 todos los residuales, realizando multiplicaciones, ejecutando las sumatorias y pasando al segundo miembro los términos con signo negativo, tendremos:

$$\sum Y_i = nb_o + b_1\sum X_{i1} + b_2\sum X_{i2} + b_3\sum X_{i3} + \ldots + b_k\sum X_{ik}$$

$$\sum X_{i1}Y_i = b_o\sum X_{i1} + b_1\sum X_{i1}^2 + b_2\sum X_{i2}X_{i1} + b_3\sum X_{i3}X_{i1} + \ldots + b_k\sum X_{ik}X_{i1}$$

$$\sum X_{i2}Y_i = b_o\sum X_{i2} + b_1\sum X_{i1}X_{i2} + b_2\sum X_{i2}^2 + b_3\sum X_{i3}X_{i2} + \ldots + b_k\sum X_{ik}X_{i2}$$

.

$$\sum X_{ik}Y_i = b_o\sum X_{ik} + b_1\sum X_{i1}X_{ik} + b_2\sum X_{i2}X_{ik} + b_3\sum X_{i3}X_{ik} + \ldots + b_k\sum X_{ik}^2$$

Conforme al sistema de ecuaciones normales escrito en forma matricial queda:

$$\begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \\ \sum X_{i3}Y_i \\ \vdots \\ \sum X_{ik}Y_i \end{bmatrix} = \begin{bmatrix} N & \sum X_{i1} & \sum X_{i2} & \sum X_{i3} & \ldots\ldots & \sum X_{ik} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i2}X_{i1} & \sum X_{i3}X_{i1} & \ldots\ldots & \sum X_{ik}X_{i1} \\ \sum X_{i2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2 & \sum X_{i3}X_{i2} & \ldots\ldots & \sum X_{ik}X_{i2} \\ \sum X_{i3} & \sum X_{i1}X_{i3} & \sum X_{i2}X_{i3} & \sum X_{i3}^2 & \ldots\ldots & \sum X_{ik}X_{i3} \\ \ldots\ldots & & & & & \\ \sum X_{ik} & \sum X_{i1}X_{ik} & \sum X_{i2}X_{ik} & \sum X_{i3}X_{ik} & \ldots\ldots & \sum X_{ik}^2 \end{bmatrix} x \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_k \end{bmatrix}$$

$$X´Y = X`X . \beta$$

Despejando queda:

$$\beta = \frac{X´Y}{X`X}$$

$$\beta = (X`X)^{-1}X´Y$$

Donde $(X`X)^{-1}$ es la matriz inversa

PROPIEDADES DEL HIPERPLANO DE REGRESION

Como Obtener $\sum Y_i$

Consideremos un vector columna de "unos" al que designamos por t (vector t), obtengamos la transpuesta (t´) que será un vector fila. Multiplicando se obtiene $\sum Y_i$.

Entonces t´ Y es:

$$[1 \quad 1 \quad 1 \ldots \ 1].\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ .. \\ Y_n \end{bmatrix} = Y_1 + Y_2 + y_3 + \ldots\ldots + Y_n = \sum Y_i = t´Y$$

Obtener $\sum Y_i$ en función de los valores de X_i.

Realicemos la siguiente operación: t´ X (matriz)

$$[1 \quad 1 \quad 1 \ldots \ 1].\begin{bmatrix} 1 & X_{11} & X_{12} & \ldots\ldots\ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots\ldots\ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & X_{n2} & \ldots\ldots\ldots & X_{nk} \end{bmatrix} = \begin{bmatrix} N & \sum X_{i1} & \sum X_{i2} & \ldots\ldots & \sum X_{ik} \end{bmatrix}$$

Multiplicando por B.

$$\begin{bmatrix} N & \sum X_{i1} & \sum X_{i2} & \ldots\ldots & \sum X_{ik} \end{bmatrix}\begin{bmatrix} b_o \\ b_1 \\ \ldots\ldots \\ b_k \end{bmatrix}$$

$$= \left[b_o N + b_1 \sum X_{i1} + b_2 \sum X_{i2} + \ \ldots\ldots + b_k \sum X_{ik}\right]$$

$$= t´XB$$

Comparando dentro del sistema de ecuaciones normales:

$t´X\beta = \sum Y_i$ Pero como, $\sum Y_i = t´Y$

$t´Y = t´X\beta$,

Igualmente: $\sum \hat{Y}_i = \sum Y_i$

1ra propiedad:

"La sumatoria del residuo, es igual a cero"

Demostración:

$$e = Y - \hat{Y}; pero \quad \hat{Y} = X\beta; \Rightarrow \quad e = Y\text{-}X\beta$$

Multiplicando por t´: $t´e = t´Y\text{-}t´X\beta$

Reemplazando $t´e \ por \ \sum e$: $\sum e = \sum Y - \sum Y = 0$

De esto podemos decir, que "el valor esperado de los residuos es cero" o "que los residuos tienen una media igual a cero, que se expresa como:

$E(e_i) = 0$

Matricialmente: $\sum e = t´e = 0$

2da propiedad:

"La sumatoria del producto cruzado de cada una de las variables en X_i por los residuos es igual a cero"

Demostración:

$$e = Y - X\beta$$

Multiplicando por la transpuesta: $X´e = X´Y\text{-}X´X\beta$ pero; $X´Y = X´X\beta$

Luego: $X´e = X´Y\text{-}X´Y = 0$

Matricialmente:

$$\begin{bmatrix} 1 & 1 & 1 & \ldots\ldots\ldots 1 \\ X_{11} & X_{21} & X_{31} & \ldots. X_{n1} \\ X_{12} & X_{22} & X_{32} & \ldots. X_{n2} \\ \ldots\ldots\ldots\ldots\ldots\ldots & & & \\ X_{1k} & X_{2k} & X_{3k} & \ldots. X_{nk} \end{bmatrix} x \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_k \end{bmatrix} = \begin{bmatrix} \sum e_i \\ \sum e_i X_{i1} \\ \sum e_i X_{i2} \\ \ldots\ldots \\ \sum e_i X_{ik} \end{bmatrix} = 0$$

3ra Propiedad:

"La sumatoria del producto cruzado de las Y estimadas por los residuos es igual a cero".

Demostración:

$$\hat{Y} = X\beta$$

Tomando la transpuesta: $\hat{Y}' = \beta' X'$

Multiplicando por el vector columna: $\hat{Y}'e = \beta' X' e$; pero: $X' e = 0$

Luego: $\hat{Y}'e = 0$

OBTENIENDO MATRICES MINIMO CUADRATICOS:

Matriz: X´X :

$$\begin{bmatrix} 1 & 1 & 1 & \ldots\ldots & 1 \\ X_{11} & X_{21} & X_{31} & .. & X_{n1} \\ X_{12} & X_{22} & X_{32} & .. & X_{n2} \\ \ldots\ldots & & & & \\ X_{1k} & X_{2k} & X_{3k} & .. & X_{nk} \end{bmatrix} x \begin{bmatrix} 1 & X_{11} & X_{12}\ldots\ldots & X_{1k} \\ 1 & X_{21} & X_{22}\ldots\ldots & X_{2k} \\ 1 & X_{31} & X_{32}\ldots\ldots & X_{3k} \\ \ldots\ldots & & & \\ 1 & X_{n1} & X_{n2}\ldots\ldots & X_{nk} \end{bmatrix} =$$

$$\begin{bmatrix} N & \sum X_{i1} & \sum X_{i2} \ldots\ldots & \sum X_{ik} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} \ldots\ldots & \sum X_{i1}X_{ik} \\ \sum X_{i2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2 \ldots\ldots & \sum X_{i2}X_{ik} \\ \ldots\ldots & & & \\ \sum X_{ik} & \sum X_{i1}X_{ik} & \sum X_{i2}X_{ik} \ldots\ldots & \sum X_{ik}^2 \end{bmatrix}$$

Es una matriz simétrica:

Matriz: X´Y:

$$\begin{bmatrix} 1 & 1 & 1 & \ldots\ldots & 1 \\ X_{11} & X_{21} & X_{31} & .. & X_{n1} \\ X_{12} & X_{22} & X_{32} & .. & X_{n2} \\ \ldots\ldots & & & & \\ X_{1k} & X_{2k} & X_{3k} & .. & X_{nk} \end{bmatrix} x \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \ldots \\ Y_N \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \\ . \\ \sum X_{ik}Y_i \end{bmatrix}$$

Propiedades de los estimadores mínimos cuadráticos (teorema de Markoff)

"Para ser eficiente un estimador debe ser ***lineal, insesgado y optimo****".*

MATRIZ DE VARIANZAS Y COVARIANZAS.

Se sabe que: $VAR(X_i) = E\left[\left(X_i - \overline{X}\right)^2\right]$; esto es:

$$E(\mu\mu^{\iota}) = \begin{bmatrix} E(\mu_1^2) & \mathrm{E}(\mu_1\mu_2) & \ldots\ldots\ldots & \mathrm{E}(\mu_1\mu_n) \\ E(\mu_1\mu_2) & E(\mu_2^2) & \ldots\ldots\ldots & \mathrm{E}(\mu_2\mu_n) \\ . & . & & . \\ E(\mu_1\mu_n) & \mathrm{E}(\mu_2\mu_n) & \ldots\ldots\ldots & E(\mu_n^2) \end{bmatrix} = \begin{bmatrix} \sigma_\mu^2 & 0 & 0 & \ldots\ldots\ldots & 0 \\ 0 & \sigma_\mu^2 & 0 & \ldots\ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & 0 & \ldots\ldots\ldots & \sigma_\mu^2 \end{bmatrix}$$

11. Ejemplos manuales

Ejemplo: En el siguiente ejemplo se hace uso de matrices para hallar un modelo de regresión lineal simple.

Sean las sumas.

$\sum X = 114$ $\sum Y = 312$

$\sum X^2 = 2216$ $\sum XY = 5828$

$\sum Y^2 = 16594$ n= 6

a) Cálculo de la Matriz $X'X$

$$(X'X) = \begin{bmatrix} 6 & 114 \\ 114 & 2216 \end{bmatrix} = \begin{bmatrix} \mathrm{n} & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$$

b) determinante $|X'X|$

$$|X'X| = \begin{vmatrix} 6 & 114 \\ 114 & 2216 \end{vmatrix} = (6)(2216) - (114)^2 = 300$$

Por ser una matriz no singular, tiene inversa

c) Hallando la inversa $(X'X)^{-1}$

Matriz de cofactores:

$$(X'X)^c = \begin{bmatrix} 2216 & -114 \\ -114 & 6 \end{bmatrix}$$

Matriz Adjunta. Adj. $(X´X)$

$$Adj(X´X) = \begin{bmatrix} 2216 & -114 \\ -114 & 6 \end{bmatrix}$$

$$(X´X)^{-1} = \begin{bmatrix} \frac{2216}{300} & \frac{-114}{300} \\ \frac{-114}{300} & \frac{6}{300} \end{bmatrix} = \begin{bmatrix} 0.72 & -0.38 \\ -0.38 & 0.02 \end{bmatrix}$$

d) Cálculo del vector columna X´Y

$$(X´Y) = \begin{bmatrix} 312 \\ 5828 \end{bmatrix} = \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

e) Cálculo de los estimadores β

$\beta = (X`X)^{-1}X´Y$:

$$(X´X)^{-1}X´Y = \begin{bmatrix} 0.72 & -0.38 \\ -0.38 & 0.02 \end{bmatrix} \begin{bmatrix} 312 \\ 5828 \end{bmatrix} = \begin{bmatrix} -1990 \\ 22.96 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\hat{Y} = -1990 + 22.96X$$

Ejemplo paso a paso:

Iniciemos con el siguiente cuadro con una variable dependiente y dos independientes cuantitativas:

Tabla Nro. Datos y cálculos necesarios para la RLM.

Y_i	X_{i1}	X_{i2}	Y_i^2	X_{i1}^2
45	23	16	2025	529
50	21	18	2500	441
40	20	14	1600	400
60	19	12	3600	361
55	17	15	3025	289
62	14	9	3844	196
312	114	84	16594	2216
$\sum Y_i$	$\sum X_{i1}$	$\sum X_{i2}$	$\sum Y_i^2$	$\sum X_{i1}^2$

X_{i2}^2	$X_{i1}Y_i$	$X_{i2}Y_i$	$X_{i1}X_{i2}$
256	1035	720	368
324	1050	900	378
196	800	560	280
144	1140	720	228
225	935	825	255
881	868	558	126

1226 $\sum X_{i2}^2$	5828 $\sum X_{i1}Y_i$	4283 $\sum X_{i2}Y_i$	1635 $\sum X_{i1}X_{i2}$

CALCULO DE LOS ESTIMADORES (Con datos Observados)

a) Modelo matricial: Y = XB +e

$$\begin{bmatrix}Y_1\\Y_2\\Y_3\\Y_4\\Y_5\\Y_6\end{bmatrix}=\begin{bmatrix}1 & X_{11} & X_{12}\\1 & X_{21} & X_{22}\\1 & X_{31} & X_{32}\\1 & X_{41} & X_{42}\\1 & X_{51} & X_{52}\\1 & X_{61} & X_{62}\end{bmatrix}\begin{bmatrix}b_0\\b_1\\b_2\end{bmatrix}+\begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{bmatrix};\quad \begin{bmatrix}45\\50\\40\\60\\55\\62\end{bmatrix}=\begin{bmatrix}1 & 23 & 16\\1 & 21 & 18\\1 & 20 & 14\\1 & 19 & 12\\1 & 17 & 15\\1 & 14 & 9\end{bmatrix}x\begin{bmatrix}b_o\\b_1\\b_2\end{bmatrix}+\begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\end{bmatrix}$$

b) Cálculo de la Matriz X´X

$$\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\\23 & 21 & 20 & 19 & 17 & 14\\16 & 18 & 14 & 12 & 15 & 9\end{bmatrix}\begin{bmatrix}1 & 23 & 16\\1 & 21 & 18\\1 & 20 & 14\\1 & 19 & 12\\1 & 17 & 15\\1 & 14 & 9\end{bmatrix}=$$

$$\begin{bmatrix}6 & 114 & 84\\114 & 2216 & 1635\\84 & 1635 & 1226\end{bmatrix}=\begin{bmatrix}N & \sum X_{i1} & \sum X_{i2}\\\sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2}\\\sum X_{i2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2\end{bmatrix}$$

c) Calculo del determinante $|X´X|$

$$|X´X|=\begin{vmatrix}6 & 114 & 84\\114 & 2216 & 1635\\84 & 1635 & 1226\end{vmatrix}=6\begin{vmatrix}2216 & 1635\\1635 & 1226\end{vmatrix}-114\begin{vmatrix}114 & 1635\\84 & 1226\end{vmatrix}+84\begin{vmatrix}114 & 2216\\84 & 1635\end{vmatrix}$$

$$|X´X|=6(43591)-114(2424)+84(246)=261546-276336+20664=5874$$

Por ser una matriz no singular, tiene inversa.

d) Hallando la inversa $(X´X)^{-1}$

Matriz de cofactores:

$$\begin{bmatrix}+\begin{vmatrix}2216 & 1635\\1635 & 1226\end{vmatrix} & -\begin{vmatrix}114 & 1635\\84 & 1226\end{vmatrix} & +\begin{vmatrix}114 & 2216\\84 & 1635\end{vmatrix}\\-\begin{vmatrix}114 & 84\\1635 & 1226\end{vmatrix} & +\begin{vmatrix}6 & 84\\84 & 1226\end{vmatrix} & -\begin{vmatrix}6 & 114\\84 & 1635\end{vmatrix}\\+\begin{vmatrix}114 & 84\\2216 & 1635\end{vmatrix} & -\begin{vmatrix}6 & 84\\114 & 1635\end{vmatrix} & +\begin{vmatrix}6 & 114\\114 & 2216\end{vmatrix}\end{bmatrix}=\begin{bmatrix}43591 & -2424 & 246\\-2424 & 300 & -234\\246 & -234 & 300\end{bmatrix}$$

Matriz Adjunta. Adj. $(X´X)$

$$Adj(X\,´X) = \begin{bmatrix} 43591 & -2424 & 246 \\ -2424 & 300 & -234 \\ 246 & -234 & 300 \end{bmatrix}$$

$$\text{Inversa: } (X\,´X)^{-1} = \begin{bmatrix} \frac{43591}{5874} & \frac{-2424}{5874} & \frac{246}{5874} \\ \frac{-2424}{5874} & \frac{300}{5874} & \frac{-234}{5874} \\ \frac{246}{5874} & \frac{-234}{5874} & \frac{300}{5874} \end{bmatrix} = \begin{bmatrix} 7.4210 & -0.4127 & 0.0419 \\ -0.4127 & 0.0511 & -0.0398 \\ 0.0419 & -0.0398 & 0.0511 \end{bmatrix}$$

e) Cálculo del vector columna X´Y

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 23 & 21 & 20 & 19 & 17 & 14 \\ 16 & 18 & 14 & 12 & 15 & 9 \end{bmatrix} \begin{bmatrix} 45 \\ 50 \\ 40 \\ 60 \\ 55 \\ 62 \end{bmatrix} = \begin{bmatrix} 312 \\ 5828 \\ 4283 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \end{bmatrix} = X\,´Y$$

f) Cálculo de los estimadores β

$\beta = (X`X)^{-1}X´Y$:

$$\beta = \begin{bmatrix} 7.4210 & -0.4127 & 0.0419 \\ -0.4127 & 0.0511 & -0.0398 \\ 0.0419 & -0.0398 & 0.0511 \end{bmatrix} \begin{bmatrix} 312 \\ 5828 \\ 4283 \end{bmatrix} = \begin{bmatrix} 89.70670 \\ -1.72114 \\ -0.35750 \end{bmatrix} = \begin{bmatrix} b_o \\ b_1 \\ b_2 \end{bmatrix}$$

g) La ecuación de regresión que:

$$\hat{Y} = 89.70670 - 1.722114X_{i1} - 0.35750\,X_{i2}$$

CALCULO DE LOS ESTIMADORES (Utilizando desviaciones)

Transformando formulas:

$$\sum x_{i1}^2 = \sum X_{i1}^2 - \frac{(\sum X_{i1})^2}{N}\,; \qquad \sum x_{i1}^2 = 2216 - \frac{(114)^2}{6} = 50$$

$$\sum x_{i2}^2 = \sum X_{i2}^2 - \frac{(\sum X_{i2})^2}{N}; \qquad \sum x_{i2}^2 = 1226 - \frac{(84)^2}{6} = 50$$

$$\sum y_i^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{N}; \qquad \sum y_i^2 = 16594 - \frac{(312)^2}{6} = 370$$

$$\sum x_{i1}x_{i2} = \sum X_{i1}X_{i2} - \frac{(\sum X_{i1})(\sum X_{i2})}{N}; \qquad \sum x_{i1}x_{i2} = 1635 - \frac{(114)(84)}{6} = 39$$

$$\sum x_{i1}y_i = \sum X_{i1}Y_i - \frac{(\sum X_{i1})(\sum Y_i)}{N}; \qquad \sum x_{i1}y_i = 5828 - \frac{(114)(312)}{6} = -100$$

$$\sum x_{i2}y_i = \sum X_{i2}Y_i - \frac{(\sum X_{i2})(\sum Y_i)}{N}; \qquad \sum x_{i2}y_i = 4283 - \frac{(84)(312)}{6} = -85$$

a) Cálculo de la Matriz $x\acute{}x$

$$(x\acute{}x) = \begin{bmatrix} 50 & 39 \\ 39 & 50 \end{bmatrix} = \begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i1}\,x_{i2} & \sum x_{i2}^{\,2} \end{bmatrix}$$

b) Cálculo del determinante $|x\acute{}x|$

$$|x\acute{}x| = \begin{vmatrix} 50 & 39 \\ 39 & 50 \end{vmatrix} = 2500 - 1521 = 979$$

Por ser una matriz no singular, tiene inversa

c) Hallando la inversa $(X\acute{}X)^{-1}$

Matriz de cofactores:

$$(x\acute{}x)^c = \begin{bmatrix} 50 & -39 \\ -39 & 50 \end{bmatrix}$$

Matriz Adjunta. Adj. $(X\acute{}X)$

$$Adj(x\acute{}x) = \begin{bmatrix} 50 & -39 \\ -39 & 50 \end{bmatrix}$$

Inversa: $(x\acute{}x)^{-1} = \begin{bmatrix} \frac{50}{979} & \frac{-39}{979} \\ \frac{-39}{979} & \frac{50}{979} \end{bmatrix} = \begin{bmatrix} 0.05107 & -0.039837 \\ -0.039837 & 0.05107 \end{bmatrix}$

d) Cálculo del vector columna x´y

$$(x\acute{}y) = \begin{bmatrix} -100 \\ -85 \end{bmatrix} = \begin{bmatrix} \sum x_{i1}y_i \\ \sum x_{i2}y_i \end{bmatrix}$$

e) Cálculo de los estimadores β

$\beta = (x\grave{}x)^{-1}x\acute{}y$:

$$(x\acute{}x)^{-1}x\acute{}y = \begin{bmatrix} 0.05107 & -0.039837 \\ -0.039837 & 0.05107 \end{bmatrix}\begin{bmatrix} -100 \\ -85 \end{bmatrix} = \begin{bmatrix} -1.7211 \\ -0.3575 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

f) cálculo de b_o

$$b_o = \overline{Y} - b_1\overline{X_1} - b_2\overline{X_2} = 52 - (-1.7211)(19) - (-0.3575)(14) = 89.7059$$

La ecuación de regresión que:

$$\hat{Y} = 89.7059 - 1.7211X_{i1} - 0.3576X_{i2}$$

CALCULO DE LA VARIANZA RESIDUAL

(estimador de la varianza poblacional)

a) Con Observaciones:

$$Y\,´Y = \sum Y_i^2 = 16594$$

$$\beta\,´(X\,´Y) = [89.70670 \quad -1.72114 \quad -0.3575]\begin{bmatrix}312\\5828\\4283\end{bmatrix} \rightarrow \beta\,´(X\,´Y) = 16426.514$$

K = 3: tres estimadores (b_o,b_1 y b_2)

$$S_e^2 = \frac{Y\,´Y - \beta\,´(X\,´Y)}{N-K} = \frac{16594 - 16426.514}{6-3} = 55.8286$$

Desviación estándar: $S_e = \sqrt{55.8286} = 7.472$

b) Con desvíos:

$$y\,´y = \sum y_i^2 = 370$$

$$\beta\,´(x\,´y) = [-1.7211 \quad -0.3575]\begin{bmatrix}-100\\-85\end{bmatrix} \rightarrow \beta\,´(x\,´y) = 202.4975$$

K = 3: tres estimadores (b_o,b_1 y b_2)

$$S_e^2 = \frac{y\,´y - \beta\,´(x\,´y)}{N-K} = \frac{370 - 202.4975}{6-3} = 55.834$$

Desviación estándar: $S_e = \sqrt{55.834} = 7.472$

CALCULO DE LAS VARIANZAS DE LOS ESTIMADORES

a) Con Observaciones

$$\sigma_\beta^2 = Var(\beta) = S_\beta^2 = S_e^2(X\,´X)^{-1}$$

$$S_\beta^2 = 55.8286\begin{bmatrix}7.4210 & -0.4127 & 0.0419\\-0.4127 & 0.0511 & -0.0398\\0.0419 & -0.0398 & 0.0511\end{bmatrix} = \begin{bmatrix}414.3040 & & \\ & 2.8513 & \\ & & 2.8513\end{bmatrix} =$$

$$\begin{bmatrix}S_{b_o}^2 & & \\ & S_{b_1}^2 & \\ & & S_{b_2}^2\end{bmatrix}$$ Desviación estándar:

$$S_{b_1} = \sqrt{2.8513} = 1.6887$$

$$S_{b_2} = \sqrt{2.8513} = 1.6887$$

b) Con desviaciones

$$S_{\beta}^2 = S_e^2(x'x)^{-1}$$

$$S_{\beta}^2 = 55.834 = \begin{bmatrix} 0.05107 & -0.039837 \\ -0.039837 & 0.05107 \end{bmatrix} = \begin{bmatrix} 2.8515 & \\ & 2.8515 \end{bmatrix} = \begin{bmatrix} S_{b_1}^2 \\ S_{b_2}^2 \end{bmatrix}$$

Desviación estándar:

$$S_{b_1} = \sqrt{2.8513} = 1.6887$$

$$S_{b_2} = \sqrt{2.8513} = 1.6887$$

PRUEBAS DE HIPOTESIS REFERENTE A LOS PARAMETROS (Inferencia)

Para b_1:

a. Ho: $\beta_1 = 0$ No existe relación lineal entre X e Y, β_1 no debe estar en el modelo

   Ha: $\beta_1 \neq 0$ existe relación entre X e Y, β_1 debe estar en el modelo

b. Nivel de significacia: $\alpha = 0.05$

c. Prueba estadística:

$$t = \frac{b_i - \beta_i}{S_{b_1}} = \frac{-1.7211 - 0}{1.6886} = -1.019$$

d. Decisión:

   /t/ = 1.019 < t_t= 3.18 no se rechaza Ho

   NO hay relación entre X_{1i} y Y_i . El parámetro β_1 no es significativo.

Para b_2:

a. Ho: $\beta_2 = 0$ No existe relación lineal entre X e Y, β_2 no debe estar en el modelo

   Ha: $\beta_2 \neq 0$ existe relación entre X e Y, β_2 debe estar en el modelo

b. Nivel de significancia: $\alpha = 0.05$

c. Prueba estadística:

$$t = \frac{b_2 - \beta_i}{S_{b_2}} = \frac{-0.3575 - 0}{1.6886} = -0.2117$$

d. Decisión:

/t/ = 02117 < t_t= 3.18 No se rechaza Ho

No hay relación entre X_{2i} y Y_i . El parámetro β_2 no es significativo.

INTERVALO DE CONFIANZA PARA LOS PARAMETOS

Para b_1: $\left[b_1 - t_{\alpha/2}(S_{b_1}) < \beta_1 < b_1 + t_{\alpha/2}(S_{b_1})\right] = 1 - \alpha$

Del ejemplo:

$$IC = [-1.7211 - (3.18)(1.6886) < \beta_1 < -1.7211 + (3.18)(1.6886)] = 1 - 0.05$$
$$IC = [-7.0908 < \beta_1 < 3.6486] = 0.95$$

Para b_2: $\left[b_2 - t_{\alpha/2}(S_{b_2}) < \beta_2 < b_2 + t_{\alpha/2}(S_{b_2})\right] = 1 - \alpha$

Del ejemplo:

$$IC = [-0.3575 - (3.18)(1.6886) < \beta_2 < -0.3575 + (3.18)(1.6886)] = 1 - 0.05$$
$$IC = [-5.7272 < \beta_2 < 5.0122] = 0.95$$

CALCULO DE CORRELACION Y DETERMINACION

Correlación.

Al ajustar un modelo de regresión múltiple a una nube de observaciones es importante disponer de alguna medida que permita medir la bondad del ajuste. Esto se consigue con los coeficientes de determinación múltiple.

Coeficiente de determinación.

En general cuando se ajusta un modelo estadístico a una nube de puntos, una medida de la bondad del ajuste es el **coeficiente de determinación.**

Representa el porcentaje de variabilidad de *Y* que explica el modelo de regresión.

Se verifica que 0 ≤ R^2 ≤ 1. Si R^2 = 1 la relación lineal es exacta y si R^2 = 0 no existe relación lineal entre la variable respuesta y las variables regresoras.

Para generalizar el concepto de R^2 como medida de la bondad de ajuste del modelo de regresión lineal múltiple, se generalizará sobre la descomposición de la variación de la variable dependiente.

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Entonces matricialmente es:

a. Como datos observados

$$R^2 = \frac{\beta\,'(X\,'Y) - \frac{(t\,'Y)^2}{N}}{Y\,'Y - \frac{(t\,'Y)^2}{N}} = \frac{16426.514 - \frac{(312)^2}{6}}{16594 - \frac{(312)^2}{6}} = 54.73\%$$

b. Como desviaciones

$$R^2 = \frac{\beta\,'(x\,'y)}{y\,'y} = \frac{202.4975}{370} = 54.73\%$$

ANALISIS DE VARIANZA

Si se quiere ver el modelo en su conjunto, para ver si las variables explicativas consideradas en el modelo ejercen influencia en la variable dependiente.

a. Como observaciones

F.V.	G.L.	S.C.	C.M.	F
Regresión	K-1	SCR	CMR = SCR/(K-1)	CMR/CME
Error (Residual)	N-K	SCE	CME = SCE/(N-K)	-
Total	N-1	SCT	-	-

Donde:

GLR = K-1 = 3-1 = 2 (K = número de parámetros)

GLE = N-K = 6-3 = 3

GLT = N-1 = 6-1 = 5

$$SCR = \beta\,'(X\,'Y) - \frac{(t\,'Y)^2}{N} = 16426.514 - \frac{(312)^2}{6} = 202.514$$

$$SCT = Y\,'Y - \frac{(t\,'Y)^2}{N} = 16594 - \frac{(312)^2}{6} = 370.00$$

$$SCE = SCT - SCR = 370 - 202.514 = 167.486$$

F.V.	G.L.	S.C.	C.M.	F
Regresión	2	202.514	101.257	1.814
Error (Residual)	3	167.486	55.828	-
Total	5	370.00	-	-

b. Como desviaciones:

$$SCR = \beta\,´(x\,´y) = 202.514$$

$$SCT = y\,´y = 370.00$$

$$SCE = SCT - SCR = 370 - 202.514 = 167.486$$

Llegamos al mismo cuadro ANVA

F.V.	G.L.	S.C.	C.M.	F
Regresión	2	202.514	101.257	1.814
Error (Residual)	3	167.486	55.828	-
Total	5	370.00	-	-

PRUEBA DE HIPOTESIS DEL COEFICIENTE DE DETERMINACION POBLACIONAL.

1) Ho: $\rho^2 = 0$. No hay relacion entre las Xs y Y

   Ha: $\rho^2 > 0$. Existe relacion entre las Xs y Y

2) Nivel de significancia

$$\alpha = 0.05$$

3) Prueba estadística

$$F = 1.814$$

4) Decisión:

   $F_t = F_{(2,3)0.05} = 9.55$

   $F = 1.814 < F_t = 9.55$ No se rechaza la Ho. No existe relación entre Y y las Xs.

   El modelo debe ser cambiado, modificado o incrementarle otras variables.

CUADRO ANVA POR ETAPAS

Trata de analizar la contribución de cada variable al modelo.

$\beta_1 = coeficiente\ de\ X_{i1}\ en\ la\ regresion\ lineal\ simple$ de Y respecto a X_{i1}

$\beta_2 = coeficiente\ de\ X_{i2}\ en\ la\ regresion\ lineal\ simple$ de Y respecto a X_{i2}

$$\beta_1 = \frac{\sum x_{i1}y}{\sum x_{i1}^2} = \frac{-100}{50} = -2$$

$$\beta_2 = \frac{\sum x_{i2}y}{\sum x_{i2}^2} = \frac{-85}{50} = -1.7$$

La suma de cuadrados explicada debida solamente a X_{i1} es:

$$\beta_1 \sum x_{i1}y = (-2)(-100) = 200$$

La suma de cuadrados explicada debida solamente a X_{i2} es:

$\beta_2 \sum x_{i2}y = (-1.7)(-85) = 144.5$

Construyamos los siguientes cuadros

F.V.	G.L.	S.C.	C.M.	F
Acción X_{i1}*	1	200.000	200	0.045
Adición X_{i2}**	1	2.514	2.514	
Regresión***	2	202.514	101.257	-
Residual	3	167.468	55.828	-
Total	5	370	74.000	-

* $R(\beta_1/\beta_0)$

** $R(\beta_2/\beta_0, \beta_1)$

*** $R(\beta_1\beta_2/\beta_0)$

F.V.	G.L.	S.C.	C.M.	F
Acción X_{i2}*	1	144.5	144.5	1.039
Adición X_{i1}**	1	58.014	58.014	
Regresión***	2	202.514	101.257	-

Residual	3	167.468	55.828	-
Total	5	370	74.000	-

PREDICCION

Supongamos que el modelo está bien determinado con un r alto, entonces considerando esto realicemos predicciones para efectos de muestra.

a. predicción puntual.

El modelo es: $\hat{Y} = 89.7067 - 1.7211X_{i1} - 0.3575X_{i2} + e_i$

Hallar Yp para cuando $X_{i1} = 26$ y $X_{i2} = 19$

$$\hat{Y}_p = 89.7067 - 1.7211(26) - 0.3575(19) = 38.1656$$

b. predicción por intervalos.

Sabemos que $Se = 7.472$; $t_t = 3.18$; $\overline{X_{i1}} = 19$; $\overline{X_{i2}} = 14$

$x_p = [(26 - 19) \quad (19.14)] = [7 \quad 5]$; $x'_p = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$

Formula:

$$Y_p - t_{\alpha/2} S_e \sqrt{1 + x_p (x'x)^{-1} x'_p} < Y_p < Y_p + t_{\alpha/2} S_e \sqrt{1 + x_p (x'x)^{-1} x'_p}$$

$$x_p (x'x)^{-1} x'_p = [7 \quad 5] \begin{bmatrix} 0.05107 & -0.09837 \\ -0.09837 & 0.05107 \end{bmatrix} \begin{bmatrix} 7 \\ 5 \end{bmatrix} = 0.9908$$

$$38.1656 - 3.18(7.472)\sqrt{1 + 0.9908} < Y_p < 38.1656 - 3.18(7.472)\sqrt{1 + 0.9908}$$

$$4.6389 < Y_p < 71.6923$$

Ejemplo:

En un trabajo de investigación sobre peso vivo en ovinos en relación a la longitud, altura y profundidad se obtuvieron los siguientes datos:

X_{i1}	X_{i2}	X_{i3}	Y
57	56	26	27

68	58	26	35
58	54	25	24
55	48	26	27
60	53	24	28
81	59	22	45
80	58	27	51
81	59	37	63
70	61	25	42
79	60	32	61

Hallar un modelo de regresión lineal completo

Solución

$n = 10$

$\sum X_{i1} = 689\,;$ $\sum X_{i2} = 566\,;$ $\sum X_{i3} = 270$ $\sum Y_i = 403\,;$

$\sum X_{i1}X_{i2} = 39296\,;$ $\sum X_{i1}X_{i3} = 18787;$ $\sum X_{i2}X_{i3} = 15326\,;$ $\sum X_{i1}^2 = 48525;$

$\sum X_{i2}^2 = 32176\ ;$ $\sum X_{i3}^2 = 7460\,;$ $\sum Y_i^2 = 18123;$

$\sum X_{i1}Y_i = 29063;$ $\sum X_{i2}Y_i = 23170;$ $\sum X_{i3}Y_i = 11286;$

$\overline{X}_{i1} = 68.9;$ $\overline{X}_{i2} = 56.6;$ $\overline{X}_{i1} = 27;$ $\overline{Y}_i = 40.3$

X'X es una matriz 4x4; para reducir los cálculos, usaremos desviaciones:

Transformaciones:

$$\sum x_{i1}x_{i2} = \sum X_{i1}X_{i2} - \frac{\sum X_{i1}\sum X_{i2}}{n} = 39296 - \frac{(689)(566)}{10} = 298.6$$

$$\sum x_{i1}x_{i3} = \sum X_{i1}X_{32} - \frac{\sum X_{i1}\sum X_{i3}}{n} = 18787 - \frac{(689)(270)}{10} = 184$$

$$\sum x_{i2}x_{i3} = \sum X_{i2}X_{i3} - \frac{\sum X_{i2}\sum X_{i3}}{n} = 15326 - \frac{(566)(270)}{10} = 44$$

$$\sum x_{i1}^2 = \sum X_{i1}^2 - \frac{(\sum X_{i1})^2}{n} = 48525 - \frac{(689)^2}{10} = 1052.9$$

$$\sum x_{i2}^2 = \sum X_{i2}^2 - \frac{(\sum X_{i2})^2}{n} = 32176 - \frac{(566)^2}{10} = 140.4$$

$$\sum x_{i3}^2 = \sum X_{i3}^2 - \frac{(\sum X_{i3})^2}{n} = 7460 - \frac{(270)^2}{10} = 170$$

$$\sum y_i^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 18123 - \frac{(403)^2}{10} = 1882.1$$

$$\sum x_{i1} y_i = \sum X_{i1} Y_i - \frac{\sum X_{i1} \sum Y_i}{n} = 29063 - \frac{(689)(403)}{10} = 1296.3$$

$$\sum x_{i2} y_i = \sum X_{i2} Y_i - \frac{\sum X_{i2} \sum Y_i}{n} = 23170 - \frac{(566)(403)}{10} = 360.2$$

$$\sum x_{i3} y_i = \sum X_{i3} Y_i - \frac{\sum X_{i3} \sum Y_i}{n} = 11286 - \frac{(270)(403)}{10} = 405$$

a) Calculo de la Matriz X´X

$$\begin{bmatrix} 1052.9 & 298.6 & 184 \\ 298.6 & 140.4 & 44 \\ 184 & 44 & 170 \end{bmatrix} = \begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \sum x_{i1}x_{i3} \\ \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \sum x_{i2}x_{i3} \\ \sum x_{i1}x_{i3} & \sum x_{i2}x_{i3} & \sum x_{i3}^2 \end{bmatrix}$$

b) Calculo del determinante $|X´X|$

$$|X´X| = \begin{bmatrix} 1052.9 & 298.6 & 184 \\ 298.6 & 140.4 & 44 \\ 184 & 44 & 170 \end{bmatrix}$$

$$= 1052.9 \begin{vmatrix} 140.4 & 44 \\ 44 & 170 \end{vmatrix} - 298.6 \begin{vmatrix} 298.6 & 44 \\ 184 & 170 \end{vmatrix} + 184 \begin{vmatrix} 298.6 & 140.4 \\ 184 & 44 \end{vmatrix}$$

$$|X´X| = 1052.9(21932) - 298.6(-42666) + 184(-12695.2) = 8016218.4$$

Por ser una matriz no singular, tiene inversa

c) Hallando la inversa $(X´X)^{-1}$

Matriz de cofactores:

$$\begin{bmatrix} + \begin{vmatrix} 2216 & 1635 \\ 1635 & 1226 \end{vmatrix} - \begin{vmatrix} 114 & 1635 \\ 84 & 1226 \end{vmatrix} + \begin{vmatrix} 114 & 2216 \\ 84 & 1635 \end{vmatrix} \\ - \begin{vmatrix} 114 & 84 \\ 1635 & 1226 \end{vmatrix} + \begin{vmatrix} 6 & 84 \\ 84 & 1226 \end{vmatrix} - \begin{vmatrix} 6 & 114 \\ 84 & 1635 \end{vmatrix} \\ + \begin{vmatrix} 114 & 84 \\ 2216 & 1635 \end{vmatrix} - \begin{vmatrix} 6 & 84 \\ 114 & 1635 \end{vmatrix} + \begin{vmatrix} 6 & 114 \\ 114 & 2216 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} 43591 & -2424 & 246 \\ -2424 & 300 & -234 \\ 246 & -234 & 300 \end{bmatrix}$$

Matriz Adjunta. Adj. (X´X)

$$Adj(X´X) = \begin{bmatrix} 43591 & -2424 & 246 \\ -2424 & 300 & -234 \\ 246 & -234 & 300 \end{bmatrix}$$

$$\text{Inversa: } (X´X)^{-1} = \begin{bmatrix} \frac{43591}{5874} & \frac{-2424}{5874} & \frac{246}{5874} \\ \frac{-2424}{5874} & \frac{300}{5874} & \frac{-234}{5874} \\ \frac{246}{5874} & \frac{-234}{5874} & \frac{300}{5874} \end{bmatrix} = \begin{bmatrix} 7.4210 & -0.4127 & 0.0419 \\ -0.4127 & 0.0511 & -0.0398 \\ 0.0419 & -0.0398 & 0.0511 \end{bmatrix}$$

d) Cálculo del vector columna X´Y

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 23 & 21 & 20 & 19 & 17 & 14 \\ 16 & 18 & 14 & 12 & 15 & 9 \end{bmatrix} \begin{bmatrix} 45 \\ 50 \\ 40 \\ 60 \\ 55 \\ 62 \end{bmatrix} = \begin{bmatrix} 312 \\ 5828 \\ 4283 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \end{bmatrix} = X´Y$$

e) Cálculo de los estimadores β

$$\beta = (\text{X`X})^{-1}\text{X´Y}: \beta = \begin{bmatrix} 7.4210 & -0.4127 & 0.0419 \\ -0.4127 & 0.0511 & -0.0398 \\ 0.0419 & -0.0398 & 0.0511 \end{bmatrix} \begin{bmatrix} 312 \\ 5828 \\ 4283 \end{bmatrix} = \begin{bmatrix} 89.70670 \\ -1.72114 \\ -0.35750 \end{bmatrix} = \begin{bmatrix} b_o \\ b_1 \\ b_2 \end{bmatrix}$$

f) La ecuación de regresión que: $\boldsymbol{Y = 89.70670 - 1.72114X_{i1} - 0.35750X_{i2}}$

CALCULO DE LOS ESTIMADORES (Utilizando desviaciones)

Transformando formulas:

$$\sum x_{i1}^2 = \sum X_{i1}^2 - \frac{(\sum X_{i1})^2}{N}$$

$$\sum x_{i1}^2 = 2216 - \frac{(114)^2}{6} = 50$$

$$\sum x_{i2}^2 = \sum X_{i2}^2 - \frac{(\sum X_{i2})^2}{N} \qquad \sum x_{i2}^2 = 1226 - \frac{(84)^2}{6} = 50$$

$$\sum y_i^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{N}$$

$$\sum y_i^2 = 16594 - \frac{(312)^2}{6} = 370$$

$$\sum x_{i1}x_{i2} = \sum X_{i1}X_{i2} - \frac{(\sum X_{i1})(\sum X_{i2})}{N} \quad \sum x_{i1}x_{i2} = 1635 - \frac{(114)(84)}{6} = 39$$

$$\sum x_{i1}y_i = \sum X_{i1}Y_i - \frac{(\sum X_{i1})(\sum Y_i)}{N} \qquad \sum x_{i1}y_i = 5828 - \frac{(114)(312)}{6} = -100$$

$$\sum x_{i2}y_i = \sum X_{i2}Y_i - \frac{(\sum X_{i2})(\sum Y_i)}{N} \qquad \sum x_{i2}y_i = 4283 - \frac{(84)(312)}{6} = -85$$

a) Cálculo de la Matriz $x´x$

$$(x´x) = \begin{bmatrix} 50 & 39 \\ 39 & 50 \end{bmatrix} = \begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i1}\,x_{i2} & \sum x_{i2}^2 \end{bmatrix}$$

b) Cálculo del determinante $|x´x|$

$$|x´x| = \begin{vmatrix} 50 & 39 \\ 39 & 50 \end{vmatrix} = 2500 - 1521 = 979$$

Por ser una matriz no singular, tiene inversa

c) Hallando la inversa $(X´X)^{-1}$

Matriz de cofactores:

$$(x´x)^c = \begin{bmatrix} 50 & -39 \\ -39 & 50 \end{bmatrix}$$

Matriz Adjunta. Adj. (X´X)

$$Adj(x´x) = \begin{bmatrix} 50 & -39 \\ -39 & 50 \end{bmatrix}$$

Inversa: $(x´x)^{-1} = \begin{bmatrix} \frac{50}{979} & \frac{-39}{979} \\ \frac{-39}{979} & \frac{50}{979} \end{bmatrix} = \begin{bmatrix} 0.05107 & -0.039837 \\ -0.039837 & 0.05107 \end{bmatrix}$

d) Cálculo del vector columna x´y

$$(x´y) = \begin{bmatrix} -100 \\ -85 \end{bmatrix} = \begin{bmatrix} \sum x_{i1}y_i \\ \sum x_{i2}y_i \end{bmatrix}$$

e) Cálculo de los estimadores β

$\beta = (x`x)^{-1}x´y$:

$$(x´x)^{-1}x´y = \begin{bmatrix} 0.05107 & -0.039837 \\ -0.039837 & 0.05107 \end{bmatrix}\begin{bmatrix} -100 \\ -85 \end{bmatrix} = \begin{bmatrix} -1.7211 \\ -0.3575 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

f) cálculo de b_o

$$b_o = \overline{Y} - b_1\overline{X_1} - b_2\overline{X_2} = 52 - (-1.7211)(19) - (-0.3575)(14) = 89.7059$$

g) La ecuación de regresión que: $\boldsymbol{Y = 89.7059 - 1.7211X_{i1} - 0.3575X_{i2}}$

CALCULO DE LA VARIANZA RESIDUAL (estimador de la varianza poblacional)

a) Con Observaciones:

$$Y\,'Y = \sum Y_i^2 = 16594$$

$$\beta\,'(X\,'Y) = [89.70670 \quad -1.72114 \quad -0.3575]\begin{bmatrix}312\\5828\\4283\end{bmatrix} \rightarrow \beta\,'(X\,'Y) = 16426.514$$

K = 3: tres estimadores (b_o,b_1 y b_2)

$$S_e^2 = \frac{Y\,'Y - \beta\,'(X\,'Y)}{N-K} = \frac{16594 - 16426.514}{6-3} = 55.8286$$

Desviación estándar: $S_e = \sqrt{55.8286} = 7.472$

b) Con desvíos:

$$y\,'y = \sum y_i^2 = 370$$

$$\beta\,'(x\,'y) = [-1.7211 \quad -0.3575]\begin{bmatrix}-100\\-85\end{bmatrix} \rightarrow \beta\,'(x\,'y) = 202.4975$$

K = 3: tres estimadores (b_o,b_1 y b_2)

$$S_e^2 = \frac{y\,'y - \beta\,'(x\,'y)}{N-K} = \frac{370 - 202.4975}{6-3} = 55.834$$

Desviación estándar: $S_e = \sqrt{55.834} = 7.472$

CALCULO DE LAS VARIANZAS DE LOS ESTIMADORES

a) Con Observaciones

$$\sigma_\beta^2 = Var(\beta) = S_\beta^2 = S_e^2(X\,'X)^{-1}$$

$$S_\beta^2 = 55.8286 \begin{bmatrix} 7.4210 & -0.4127 & 0.0419 \\ -0.4127 & 0.0511 & -0.0398 \\ 0.0419 & -0.0398 & 0.0511 \end{bmatrix} = \begin{bmatrix} 414.3040 & & \\ & 2.8513 & \\ & & 2.8513 \end{bmatrix} =$$

$\begin{bmatrix} S_{b_o}^2 \\ S_{b_1}^2 \\ S_{b_2}^2 \end{bmatrix}$ Desviación estándar:

$$S_{b_1} = \sqrt{2.8513} = 1.6887$$

$$S_{b_2} = \sqrt{2.8513} = 1.6887$$

b) Con desviaciones

$$S_\beta^2 = S_e^2 (x'x)^{-1}$$

$$S_\beta^2 = 55.834 = \begin{bmatrix} 0.05107 & -0.039837 \\ -0.039837 & 0.05107 \end{bmatrix} = \begin{bmatrix} 2.8515 & \\ & 2.8515 \end{bmatrix} = \begin{bmatrix} S_{b_1}^2 \\ S_{b_2}^2 \end{bmatrix}$$

Desviación estándar:

$$S_{b_1} = \sqrt{2.8513} = 1.6887$$

$$S_{b_2} = \sqrt{2.8513} = 1.6887$$

PRUEBAS DE HIPOTESIS REFERENTE A LOS PARAMETROS (Inferencia)

Para b_1:

a. Ho: $\beta_1 = 0$ No existe relación lineal entre X e Y, β_1 no debe estar en el modelo.

Ha: $\beta_1 \neq 0$ existe relación entre X e Y, β_1 debe estar en el modelo.

b. Nivel de significacia:

$$\alpha = 0.05$$

c. Prueba estadística:

$$t = \frac{b_i - \beta_i}{S_{b_1}} = \frac{-1.7211 - 0}{1.6886} = -1.019$$

d. Decisión:

/t/ = 1.019 < t_t= 3.18 no se rechaza Ho

NO hay relación entre X_{1i} y Y_i . El parámetro β_1 no es significativo.

Para b_2:

a. Ho: $\beta_2 = 0$ No existe relación lineal entre X e Y, β_2 no debe estar en el modelo

   Ha: $\beta_2 \neq 0$ existe relación entre X e Y, β_2 debe estar en el modelo

b. Nivel de significacia:

$$\alpha = 0.05$$

c. Prueba estadística:

$$t = \frac{b_2 - \beta_i}{S_{b_2}} = \frac{-0.3575 - 0}{1.6886} = -0.2117$$

d. Decisión:

   /t/ = 02117 < t_t= 3.18 No se rechaza Ho

   No hay relación entre X_{2i} y Y_i . El parámetro β_2 no es significativo.

INTERVALO DE CONFIANZA PARA LOS PARAMETOS

Para b_1: $\left[b_1 - t_{\alpha/2}(S_{b_1}) < \beta_1 < b_1 + t_{\alpha/2}(S_{b_1})\right] = 1 - \alpha$

Del ejemplo:

$$IC = [-1.7211 - (3.18)(1.6886) < \beta_1 < -1.7211 + (3.18)(1.6886)] = 1 - 0.05$$
$$IC = [-7.0908 < \beta_1 < 3.6486] = 0.95$$

Para b_2: $\left[b_2 - t_{\alpha/2}(S_{b_2}) < \beta_2 < b_2 + t_{\alpha/2}(S_{b_2})\right] = 1 - \alpha$

Del ejemplo:

$$IC = [-0.3575 - (3.18)(1.6886) < \beta_2 < -0.3575 + (3.18)(1.6886)] = 1 - 0.05$$
$$IC = [-5.7272 < \beta_2 < 5.0122] = 0.95$$

CALCULO DE CORRELACION Y DETERMINACION

Correlación.

Al ajustar un modelo de regresión múltiple a una nube de observaciones es importante disponer de alguna medida que permita medir la bondad del ajuste. Esto se consigue con los coeficientes de determinación múltiple.

Coeficiente de determinación.

En general cuando se ajusta un modelo estadístico a una nube de puntos, una medida de la bondad del ajuste es el **coeficiente de determinación.**

Representa el porcentaje de variabilidad de *Y* que explica el modelo de regresión.

Se verifica que $0 \leq R^2 \leq 1$. Si $R^2 = 1$ la relación lineal es exacta y si $R^2 = 0$ no existe relación lineal entre la variable respuesta y las variables regresoras.

Para generalizar el concepto de R^2 como medida de la bondad de ajuste del modelo de regresión lineal múltiple, se generalizará sobre la descomposición de la variación de la variable dependiente.

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Entonces matricialmente es:

a. Como datos observados

$$R^2 = \frac{\beta\,´(X\,´Y) - \frac{(t\,´Y)^2}{N}}{Y\,´Y - \frac{(t\,´Y)^2}{N}} = \frac{16426.514 - \frac{(312)^2}{6}}{16594 - \frac{(312)^2}{6}} = 54.73\%$$

b. Como desviaciones

$$R^2 = \frac{\beta\,´(x\,´y)}{y\,´y} = \frac{202.4975}{370} = 54.73\%$$

CONSTRUCCION DEL CUADRO DE ANALISIS DE VARIANZA

Si se quiere ver el modelo en su conjunto, para ver si las variables explicativas consideradas en el modelo ejercen influencia en la variable dependiente.

A continuación, se expone como descomponer la variabilidad de la variable de interés *Y* cuando se ajusta un modelo de regresión múltiple.

Razonando como en el modelo de Regresión Lineal Simple, en cada observación muestral se puede hacer la siguiente descomposición.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \qquad i = 1, 2, \ldots n.$$

El contraste que se desea resolver es el siguiente:

$$C_M \equiv \left\{ \begin{array}{ll} H_0 \equiv & \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0 \\ H_1 \equiv & \text{algún } \alpha_i \neq 0 \quad \text{para algún } i \end{array} \right\} \qquad \begin{array}{c} \textbf{contraste} \\ \textbf{conjunto de la F} \end{array}$$

Si H_0 es cierto ninguna de las variables regresoras influye en la variable respuesta (el modelo no influye).

a. Como observaciones

F.V.	G.L.	S.C.	C.M.	F
Regresión	K-1	SCR	CMR = SCR/(K-1)	CMR/CME
Error (Residual)	N-K	SCE	CME = SCE/(N-K)	-
Total	N-1	SCT	-	-

Donde:

GLR = K-1 = 3-1 = 2 (K = número de parámetros)

GLE = N-K = 6-3 = 3

GLT = N-1 = 6-1 = 5

$$SCR = \beta'(X'Y) - \frac{(t'Y)^2}{N} = 16426.514 - \frac{(312)^2}{6} = 202.514$$

$$SCT = Y'Y - \frac{(t'Y)^2}{N} = 16594 - \frac{(312)^2}{6} = 370.00$$

$$SCE = SCT - SCR = 370 - 202.514 = 167.486$$

F.V.	G.L.	S.C.	C.M.	F
Regresión	2	202.514	101.257	1.814
Error (Residual)	3	167.486	55.828	-
Total	5	370.00	-	-

b. Como desviaciones:

$$SCR = \beta\,'(x\,'y) = 202.514$$
$$SCT = y\,'y = 370.00$$
$$SCE = SCT - SCR = 370 - 202.514 = 167.486$$

Llegamos al mismo cuadro ANVA

F.V.	G.L.	S.C.	C.M.	F
Regresión	2	202.514	101.257	1.814
Error (Residual)	3	167.486	55.828	-
Total	5	370.00	-	-

PRUEBA DE HIPOTESIS DEL COEFICIENTE DE DETERMINACION POBLACIONAL.

1) Ho: $\rho^2 = 0$. No hay relacion entre las Xs y Y

   Ha: $\rho^2 > 0$. Existe relacion entre las Xs y Y

2) Nivel de significancia $\alpha = 0.05$

3) Prueba estadística

   $F = 1.814$

4) Decisión: $F_t = F_{(2,3)0.05} = 9.55$

   $F = 1.814 < F_t = 9.55$ No se rechaza la Ho. No existe relación entre Y y las Xs.

   El modelo debe ser cambiado, modificado o incrementarle otras variables.

CUADRO ANVA POR ETAPAS

Trata de analizar la contribución de cada variable al modelo.

$\beta_1 = coeficiente\ de\ X_{i1}\ en\ la\ regresion\ lineal\ simple$ de Y respecto a X_{i1}

$\beta_2 = coeficiente\ de\ X_{i2}\ en\ la\ regresion\ lineal\ simple$ de Y respecto a X_{i2}

$$\beta_1 = \frac{\sum x_{i1} y}{\sum x_{i1}^2} = \frac{-100}{50} = -2$$

$$\beta_2 = \frac{\sum x_{i2} y}{\sum x_{i2}^2} = \frac{-85}{50} = -1.7$$

La suma de cuadrados explicada debida solamente a X_{i1} es:

$$\beta_1 \sum x_{i1} y = (-2)(-100) = 200$$

La suma de cuadrados explicada debida solamente a X_{i2} es:

$$\beta_2 \sum x_{i2} y = (-1.7)(-85) = 144.5$$

Construyamos los siguientes cuadros

F.V.	G.L.	S.C.	C.M.	F
Acción X_{i1}*	1	200.000	200	0.045
Adición X_{i2}**	1	2.514	2.514	
Regresión***	2	202.514	101.257	-
Residual	3	167.468	55.828	-
Total	5	370	74.000	-

* $R(\beta_1/\beta_0)$

** $R(\beta_2/\beta_0, \beta_1)$

*** $R(\beta_1\beta_2/\beta_0)$

F.V.	G.L.	S.C.	C.M.	F
Acción X_{i2}*	1	144.5	144.5	1.039
Adición X_{i1}**	1	58.014	58.014	
Regresión***	2	202.514	101.257	-
Residual	3	167.468	55.828	-
Total	5	370	74.000	-

PREDICCION

Supongamos que el modelo está bien determinado con un r alto, entonces considerando esto realicemos predicciones para efectos de muestra.

a. predicción puntual.

El modelo es: $Y = 89.7067 - 1.7211X_{i1} - 0.3575X_{i2} + e_i$

Hallar Yp para cuando $\mathbf{X_{i1} = 26}$ **y** $\mathbf{X_{i2} = 19}$

$$Y_p = 89.7067 - 1.7211(26) - 0.3575(19) = 38.1656$$

b. predicción por intervalos.

Sabemos que $Se = 7.472$; $t_t = 3.18$; $\overline{X_{i1}} = 19$; $\overline{X_{i2}} = 14$

$$x_p = [(26 - 19) \quad (19.14)] = [7 \quad 5] \quad ; \quad x'_p = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$$

Formula:

$$Y_p - t_{\alpha/2} S_e \sqrt{1 + x_p (x'x)^{-1} x'_p} < Y_p < Y_p + t_{\alpha/2} S_e \sqrt{1 + x_p (x'x)^{-1} x'_p}$$

$$x_p (x'x)^{-1} x'_p = [7 \quad 5] \begin{bmatrix} 0.05107 & -0.09837 \\ -0.09837 & 0.05107 \end{bmatrix} \begin{bmatrix} 7 \\ 5 \end{bmatrix} = 0.9908$$

$$38.1656 - 3.18(7.472)\sqrt{1 + 0.9908} < Y_p < 38.1656 - 3.18(7.472)\sqrt{1 + 0.9908}$$

$$4.6389 < Y_p < 71.6923$$

12. Ejemplo 1 en R. 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 como: habitantes, analfabetismo, ingresos, asesinatos, universitarios, heladas, área y densidad poblacional.

Importando y mostrando datos.

```
# Regresión lineal múltiple
# Y: esperanza de vida media--
# Xs: habitantes, analfabetismo, ingresos, esperanza de vida, asesinatos,
# universitarios, heladas, área y densidad poblacional.
require(dplyr)
datos <- read.csv("state.csv", head=T, sep=";")

head(datos)
```

```
##   life_exp population income illiteracy murder hs_grad frost   area   density
## 1    69.05       3615   3624        2.1   15.1    41.3    20  50708  71.29053
## 2    69.31        365   6315        1.5   11.3    66.7   152 566432   1.20000
## 3    70.55       2212   4530        1.8    7.8    58.1    15 113417  19.50325
## 4    70.66       2110   3378        1.9   10.1    39.9    65  51945  40.61989
## 5    71.71      21198   5114        1.1   10.3    62.6    20 156361 135.57089
## 6    72.06       2541   4884        0.7    6.8    63.9   166 103766  24.48779
```

La data set empleada es state. Para facilitar su interpretación se renombra las variables.

```
# renombrar variables

datos <- rename(habitantes = population, analfabetismo = illiteracy,
ingresos = income, esp_vida = "life_exp", asesinatos = murder, universita
rios = "hs_grad", heladas = frost, area = area,. data = datos)

datos <- mutate(.data = datos, densidad_pobl = habitantes * 1000/area)

head(datos)
```

```
##   esp_vida habitantes ingresos analfabetismo asesinatos universitarios heladas
## 1    69.05       3615     3624           2.1       15.1           41.3      20
## 2    69.31        365     6315           1.5       11.3           66.7     152
## 3    70.55       2212     4530           1.8        7.8           58.1      15
## 4    70.66       2110     3378           1.9       10.1           39.9      65
## 5    71.71      21198     5114           1.1       10.3           62.6      20
## 6    72.06       2541     4884           0.7        6.8           63.9     166
##     area   density densidad_pobl
## 1  50708  71.29053    71.2905261
## 2 566432   1.20000     0.6443845
## 3 113417  19.50325    19.5032491
## 4  51945  40.61989    40.6198864
## 5 156361 135.57089   135.5708904
## 6 103766  24.48779    24.4877898
```

```
str(datos)
```

```
## 'data.frame':    50 obs. of  10 variables:
##  $ esp_vida      : num  69 69.3 70.5 70.7 71.7 ...
```

```
##  $ habitantes    : int  3615 365 2212 2110 21198 2541 3100 579 8277 4931 ...
##  $ ingresos      : int  3624 6315 4530 3378 5114 4884 5348 4809 4815 4091 ...
##  $ analfabetismo : num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ asesinatos    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ universitarios: num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ heladas       : int  20 152 15 65 20 166 139 103 11 60 ...
##  $ area          : int  50708 566432 113417 51945 156361 103766 4862 1982 5409
0 58073 ...
##  $ density       : num  71.3 1.2 19.5 40.6 135.6 ...
##  $ densidad_pobl : num  71.291 0.644 19.503 40.62 135.571 ...
```

```
# datos <- scale(datos[,2:10])
# datos
```

Todas las variables son numéricas, donde la variable esp_vida es la dependiente y las demás variables las independientes

```
summary(datos)
```

```
##     esp_vida       habitantes       ingresos    analfabetismo
##  Min.   :67.96   Min.   :  365   Min.   :3098   Min.   :0.500
##  1st Qu.:70.12   1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625
##  Median :70.67   Median : 2838   Median :4519   Median :0.950
##  Mean   :70.88   Mean   : 4246   Mean   :4436   Mean   :1.170
##  3rd Qu.:71.89   3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575
##  Max.   :73.60   Max.   :21198   Max.   :6315   Max.   :2.800
##    asesinatos     universitarios     heladas            area
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
##  Median : 6.850   Median :53.25   Median :114.50   Median : 54277
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
##     density       densidad_pobl
##  Min.   :  1.20   Min.   :  0.6444
##  1st Qu.: 25.34   1st Qu.: 25.3352
##  Median : 73.02   Median : 73.0154
```

```
##  Mean   :149.24   Mean   :149.2245
##  3rd Qu.:144.28   3rd Qu.:144.2828
##  Max.   :975.00   Max.   :975.0033
```

La esperanza de vida media es 70.88 años, la muestra de estudios tuvo como mínimo 67.97 años y máxima 73.60 años.....

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, qué 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.

Correlación de Pearson bivariada

```
round(cor(x = datos, method = "pearson"),3) # correlaciones con redondeo
```

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

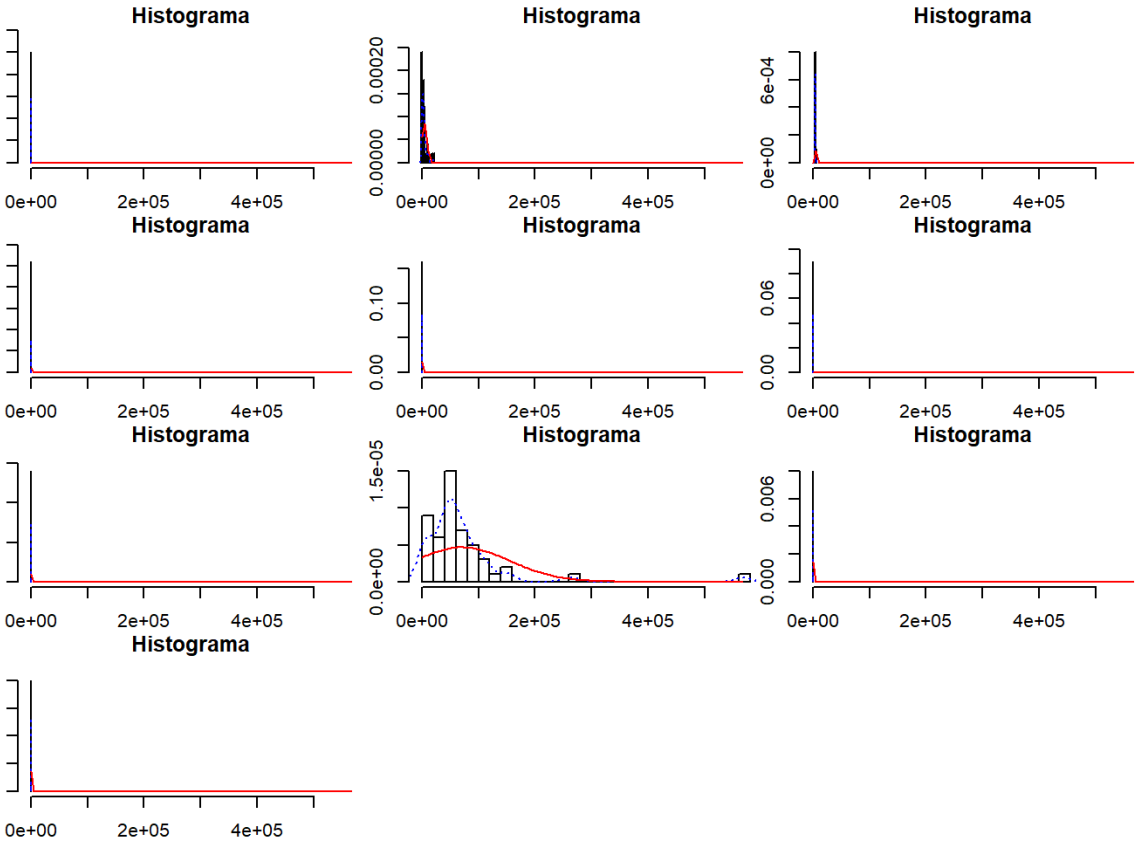
```
## area                0.334    0.059  1.000  -0.341         -0.341
## density            -0.088    0.002 -0.341   1.000          1.000
## densidad_pobl      -0.088    0.002 -0.341   1.000          1.000
```

Las variables que tienen relación con la esperanza de vida son: analfabetismo, asesinatos y universitarios, existen otras variables relacionadas en forma más débil y otras que no muestran relación.
Así mismo podemos ver correlaciones entre variables independientes.

Histograma de frecuencias de las variables

```
require(psych)
```

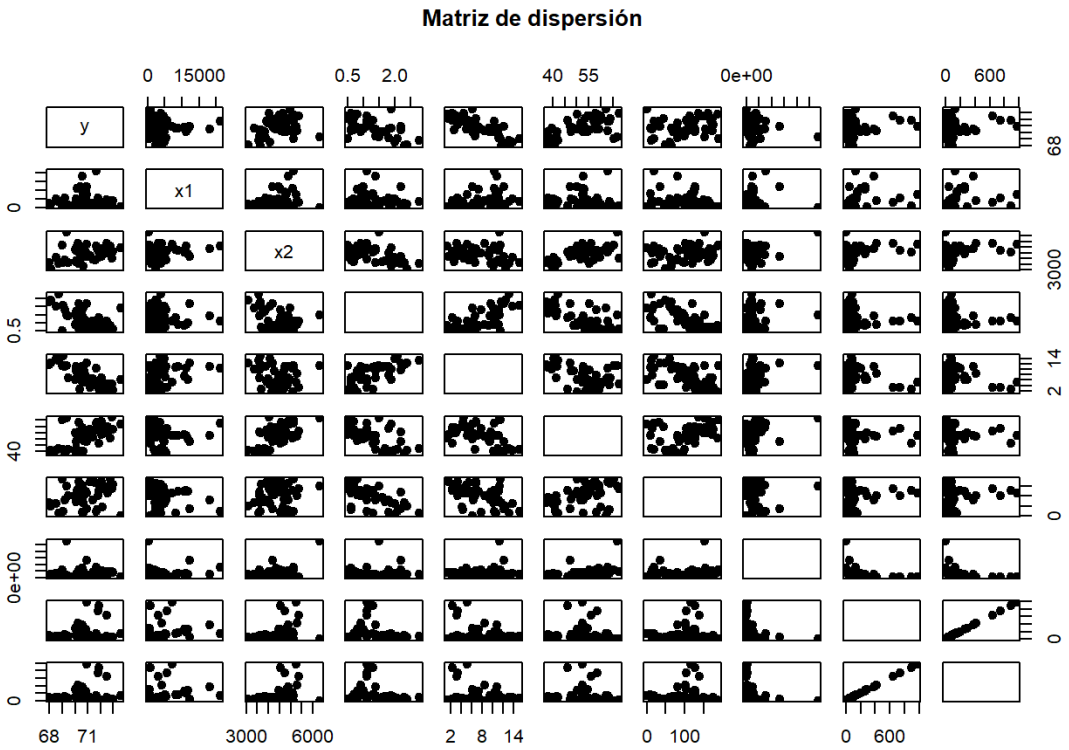
```
multi.hist(x = datos, dcol = c("blue", "red"), dlty = c("dotted", "solid"),
           main = "Histograma")
```



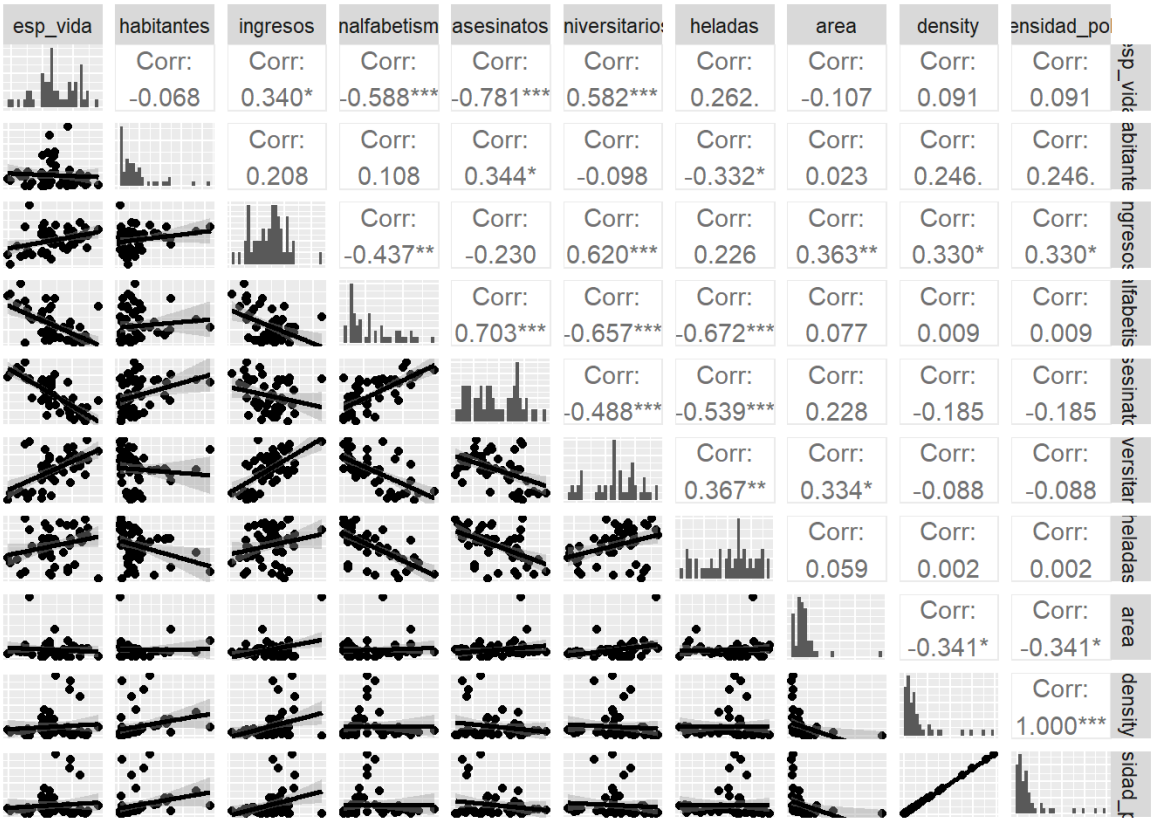
No es muy usual debido a las escalas de las variables.

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.

```
# matriz de dispersion

# x11()

pairs(datos, labels=c("y","x1","x2"), main='Matriz de dispersión', cex.ma
in=0.8, cex = 1.5, pch = 20, bg="light blue", cex.labels = 1, font.labels
=1)
```




```
require(GGally)

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

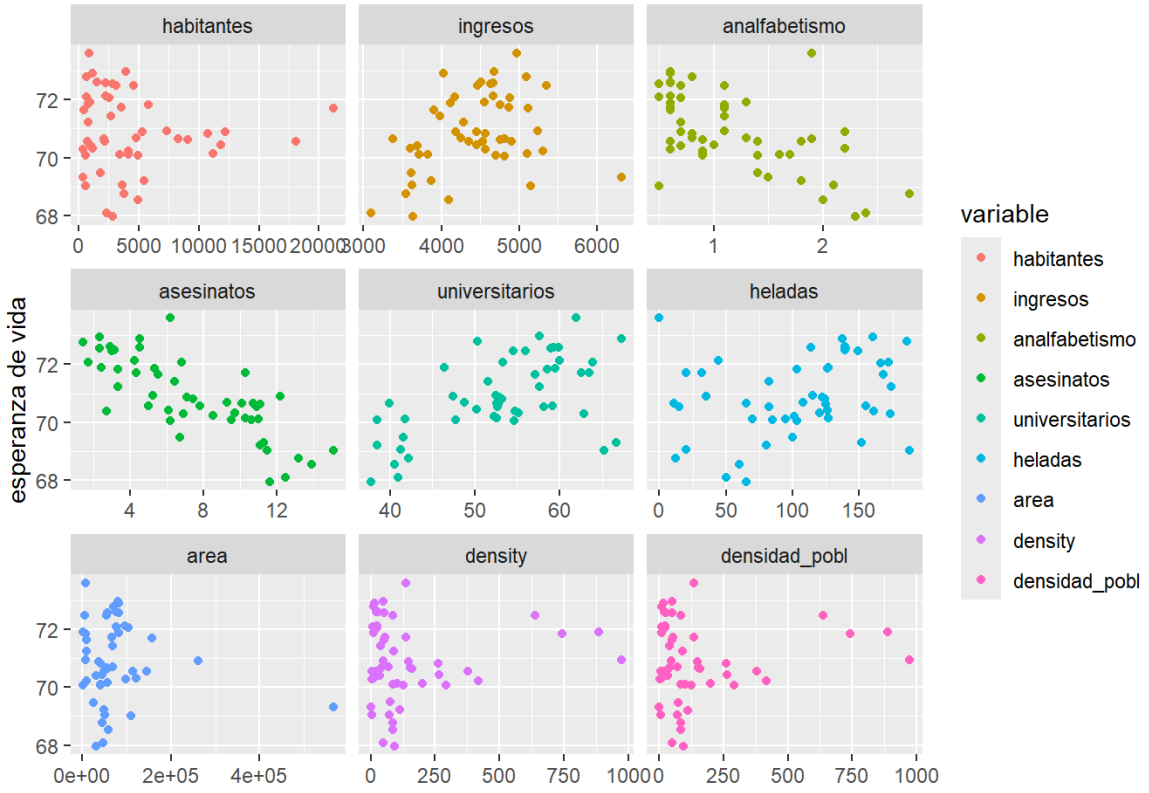


Del análisis preliminar se pueden extraer las siguientes conclusiones:

- 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).

Gráficos parciales.

```
# Gráficos parciales
# Gráfico de dispersión de la esperanza de vida respecto de cada predictora
library(reshape2)
datacomp = melt(datos, id.vars = 'esp_vida')
ggplot(datacomp) +
  geom_jitter(aes(value, esp_vida, colour = variable)) +
  facet_wrap(~variable, scales = "free_x") +
  labs(x = "", y = "esperanza de vida")
```



La figura muestra la relación entre la variable dependiente con las variables independientes, la variable área, densidad y densidad poblacional muestran puntos alejados, serían las variables me nos relacionadas con la esperanza de vida.

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 *mixto* iniciando el modelo con todas las variables como predictores y realizando la selección de los mejores predictores con la medición *Akaike(AIC)*. No entra al análisis densidad por haber sido mutada a densidad poblacional.

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

```
## Call:
## lm(formula = esp_vida ~ habitantes + analfabetismo + asesinatos +
##     universitarios + heladas + area + densidad_pobl, data = datos)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.59949 -0.41411 -0.04057  0.61060  1.41540
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.073e+01  1.608e+00  44.000  < 2e-16 ***
## habitantes      6.659e-05  2.986e-05   2.230   0.0311 *
## analfabetismo   1.817e-01  3.767e-01   0.482   0.6321
## asesinatos     -3.201e-01  4.830e-02  -6.628 4.97e-08 ***
## universitarios  5.064e-02  2.152e-02   2.353   0.0234 *
## heladas        -4.959e-03  3.151e-03  -1.574   0.1230
## area           -7.468e-07  1.670e-06  -0.447   0.6571
## densidad_pobl  -7.045e-04  5.687e-04  -1.239   0.2223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7316 on 42 degrees of freedom
## Multiple R-squared:  0.7454, Adjusted R-squared:  0.703
## F-statistic: 17.57 on 7 and 42 DF,  p-value: 1.233e-10
```

El modelo con todas las variables introducidas como predictores tiene un R^2 alta (0.7454), es capaz de explicar el 75% de la variabilidad observada en la esperanza de vida. El *p-value* del modelo en su conjunto (ANVA) es significativo (1.233e-10<α) por lo que se puede aceptar que el modelo no es por azar, al menos uno de los coeficientes parciales es distinto de 0. En nuestra salida, muchas variables no son significativas, lo que es un indicativo de que podrían no contribuir al modelo, eliminando estas variables, podría mejorar el modelo.

Hasta aquí demos una lectura de los coeficientes.

La interpretación de los coeficientes nos indica que la esperanza de vida aumenta en 0.00006659 unidades por el incremento de una unidad de habitantes, aumenta en 0.1817 unidades por cada unidad de la tasa de analfabetismo, disminuye en 0.03201 unidades por cada unidad de asesinato, sucesivamente.

```
AIC(modelo)
```

```
## [1] 119.9248
```

```
BIC(modelo)
```

```
## [1] 137.133
```

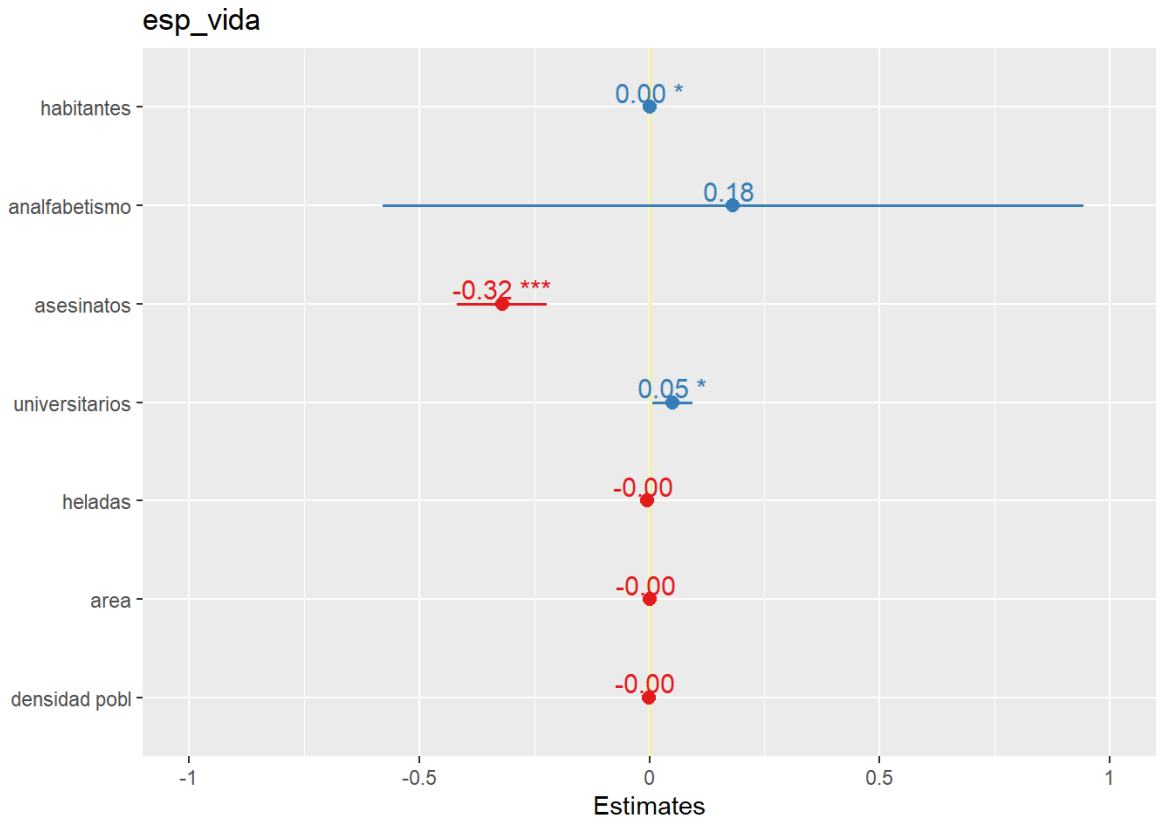
```
# C?lculo del error cuadr?tico medio RMSE:
sqrt(mean(modelo$residuals^2))
```

```
## [1] 0.6705272
```

A continuación, se presenta el ajuste obtenido para cada variable de forma marginal.

Representamos gráficamente la estimación e intervalo de confianza de los coeficientes del modelo para apreciar los efectos descritos:

```
plot_model(modelo,
           show.values = TRUE,
           vline.color = "yellow")
```



Comparamos los resultados con los del modelo estandarizado. La tabla proporciona las estimaciones e intervalo de confianza al 95% de los parámetros en la escala original (Estimates y CI), las estimaciones y CI de los coeficientes del modelo estandarizado (std.Beta y standarized CI), y el p-valor asociado a cada coeficiente. Observamos coeficientes significativos como habitantes, asesinatos y universitarios.

```
# Inferencia sobre los parámetros del modelo
tab_model(modelo,
          show.std = TRUE,
          show.r2 = FALSE)
```

	esp_vida				
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*P*
(Intercept)	70.73	-0.00	67.49 – 73.98	-0.16 – 0.16	**<0.001**
Habitantes	0.00	0.22	0.00 – 0.00	0.02 – 0.42	**0.031**
Analfabetismo	0.18	0.08	-0.58 – 0.94	-0.26 – 0.43	0.632
Asesinatos	-0.32	-0.88	-0.42 – -0.22	-1.15 – -0.61	**<0.001**
Universitarios	0.05	0.30	0.01 – 0.09	0.04 – 0.57	**0.023**

Heladas	-0.00	-0.19	-0.01 – 0.00	-0.44 – 0.05	0.123
Área	-0.00	-0.05	-0.00 – 0.00	-0.26 – 0.17	0.657
densidad pobl	-0.00	-0.12	-0.00 – 0.00	-0.30 – 0.07	0.222
Observations	50				

Se aprecia como la variable más relevante para predecir la esperanza de vida es asesinatos, seguido de Universitarios y finalmente Habitantes.

Resumen de bondad de ajuste:

```
library (broom)
glance(modelo)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.745         0.703 0.732      17.6 1.23e-10     7  -51.0  120.  137.
## # ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# Tabla ANOVA
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: esp_vida
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## habitantes      1  0.4089  0.4089  0.7640  0.387056
## analfabetismo   1 30.1716 30.1716 56.3695 2.752e-09 ***
## asesinatos      1 27.2885 27.2885 50.9831 9.166e-09 ***
## universitarios  1  5.0359  5.0359  9.4086  0.003771 **
## heladas         1  2.0900  2.0900  3.9047  0.054745 .
## area            1  0.0026  0.0026  0.0048  0.945101
## densidad_pobl   1  0.8212  0.8212  1.5343  0.222350
## Residuals      42 22.4803  0.5352
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Según el análisis de varianza del modelo, las variables mas importantes para predecir la esperanza de vida serian: analfabetismo, asesinatos y universitarios.

Modelo final.

```
# modelo final.
modelof <- lm(esp_vida ~ analfabetismo + asesinatos +
                universitarios, data = datos)
summary(modelof)
```

```
## Call:
## lm(formula = esp_vida ~ analfabetismo + asesinatos + universitarios,
##     data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65922 -0.46400  0.08517  0.59643  1.77657
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69.73545    1.22208  57.063  < 2e-16 ***
## analfabetismo   0.25398    0.30508   0.833  0.40942
## asesinatos     -0.25813    0.04350  -5.934 3.63e-07 ***
## universitarios  0.05179    0.01876   2.761  0.00825 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7985 on 46 degrees of freedom
## Multiple R-squared:  0.6679, Adjusted R-squared:  0.6462
## F-statistic: 30.83 on 3 and 46 DF,  p-value: 4.444e-11
```

```
AIC(modelof)
```

```
## [1] 125.2206
```

3. Selección de los mejores predictores

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

```
# usemos dos métodos
#Seleccion de los mejores predictores steepwise mixto
step(object = modelo, direction = "both", trace = 1)
```

```
# Start:  AIC=-23.97
## esp_vida ~ habitantes + analfabetismo + asesinatos + universitarios +
##     heladas + area + densidad_pobl
##
##                  Df Sum of Sq    RSS      AIC
## - area            1    0.1070 22.587 -25.7317
## - analfabetismo   1    0.1245 22.605 -25.6929
## - densidad_pobl   1    0.8212 23.302 -24.1752
## <none>                        22.480 -23.9691
## - heladas         1    1.3258 23.806 -23.1040
## - habitantes      1    2.6621 25.142 -20.3733
## - universitarios  1    2.9633 25.444 -19.7778
## - asesinatos      1   23.5109 45.991   9.8214
##
## Step:  AIC=-25.73
## esp_vida ~ habitantes + analfabetismo + asesinatos + universitarios +
##     heladas + densidad_pobl
##
##                  Df Sum of Sq    RSS      AIC
## - analfabetismo   1    0.0607 22.648 -27.5976
## - densidad_pobl   1    0.7168 23.304 -26.1696
## <none>                        22.587 -25.7317
## + area            1    0.1070 22.480 -23.9691
## - heladas         1    1.8206 24.408 -23.8557
## - habitantes      1    2.5631 25.150 -22.3574
## - universitarios  1    3.5390 26.126 -20.4539
## - asesinatos      1   25.2413 47.829   9.7801
##
```

```
## 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
## + 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
## + 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
heladas
##      7.103e+01       5.014e-05      -3.001e-01       4.658e-02      -5
.943e-03
```

El mejor modelo resultante del proceso de selección ha sido:

$$esp_{vida} = 71.03 + 0.00005014(habitantes) - 0.3001(asesinatos) + 0.04658(universitarios) - 0.00594(heladas)$$

Con un AIC de -28.16

Utilizando Steep Wise Backward

```
#Seleccion de los mejores predictores steep wise backward
step(object = modelo, direction = "backward", trace = 1)
```

```
## Start:  AIC=-23.97
## esp_vida ~ habitantes + analfabetismo + asesinatos + universitarios +
##     heladas + area + densidad_pobl
##
##                  Df Sum of Sq    RSS      AIC
## - area            1    0.1070 22.587 -25.7317
## - analfabetismo   1    0.1245 22.605 -25.6929
## - densidad_pobl   1    0.8212 23.302 -24.1752
## <none>                        22.480 -23.9691
## - heladas         1    1.3258 23.806 -23.1040
## - habitantes      1    2.6621 25.142 -20.3733
## - universitarios  1    2.9633 25.444 -19.7778
## - asesinatos      1   23.5109 45.991   9.8214
##
## Step:  AIC=-25.73
## esp_vida ~ habitantes + analfabetismo + asesinatos + universitarios +
##     heladas + densidad_pobl
##
##                  Df Sum of Sq    RSS      AIC
## - analfabetismo   1    0.0607 22.648 -27.5976
## - densidad_pobl   1    0.7168 23.304 -26.1696
## <none>                        22.587 -25.7317
## - heladas         1    1.8206 24.408 -23.8557
## - habitantes      1    2.5631 25.150 -22.3574
## - universitarios  1    3.5390 26.126 -20.4539
```

```
## - asesinatos      1   25.2413 47.829   9.7801
##
## 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
## - 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
## - 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
```

resumen:

1) AIC= 23.97 : esp_vida ~ habitantes + analfabetismo + asesinatos + universitarios + heladas + area + densidad_pobl
2) AIC= 25.73: esp_vida ~ habitantes + analfabetismo + asesinatos + universitarios + heladas + densidad_pobl
3) AIC=-27.6: esp_vida ~ habitantes + asesinatos + universitarios + heladas + densidad_pobl
4) AIC=-28.16: esp_vida ~ habitantes + asesinatos + universitarios + heladas

Finalmente:

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

```
## Coefficients:
##    (Intercept)     habitantes      asesinatos  universitarios        heladas
##     7.103e+01      5.014e-05     -3.001e-01      4.658e-02     -5.943e-03
```

$$esp_{vida} = 71.03 + 0.00005014(habitantes) - 0.3001(asesinatos) + 0.04658(universitarios) - 0.00594(heladas)$$

Similar al otro método.

Con esto se corre el modelo

```
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
```

$$esp_{vida} = 71.03 + 0.00005014(habitantes) - 0.3001(asesinatos) + 0.04658(universitarios) - 0.00594(heladas)$$

En el modelo final con las variables seleccionadas como predictores, se tiene un R^2 alta (0.736), es capaz de explicar el 73,6% de la variabilidad observada en la esperanza de vida. El p-value del modelo es significativo (1.696e-12<α) por lo que se puede aceptar que el modelo no es por azar, al menos uno de los coeficientes parciales es distinto de 0. Existe una variable en el límite de no significancia al 5% (habitantes).

Es recomendable mostrar el intervalo de confianza para cada uno de los coeficientes parciales de regresión:

```
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 regresión 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. Para este ejemplo, por cada unidad que aumenta el predictor *universitario*, la esperanza de vida aumenta en promedio 0.04658 unidades, manteniéndose constantes el resto de predictores.

```
AIC(modelo1)
```

```
## [1] 115.7326
```

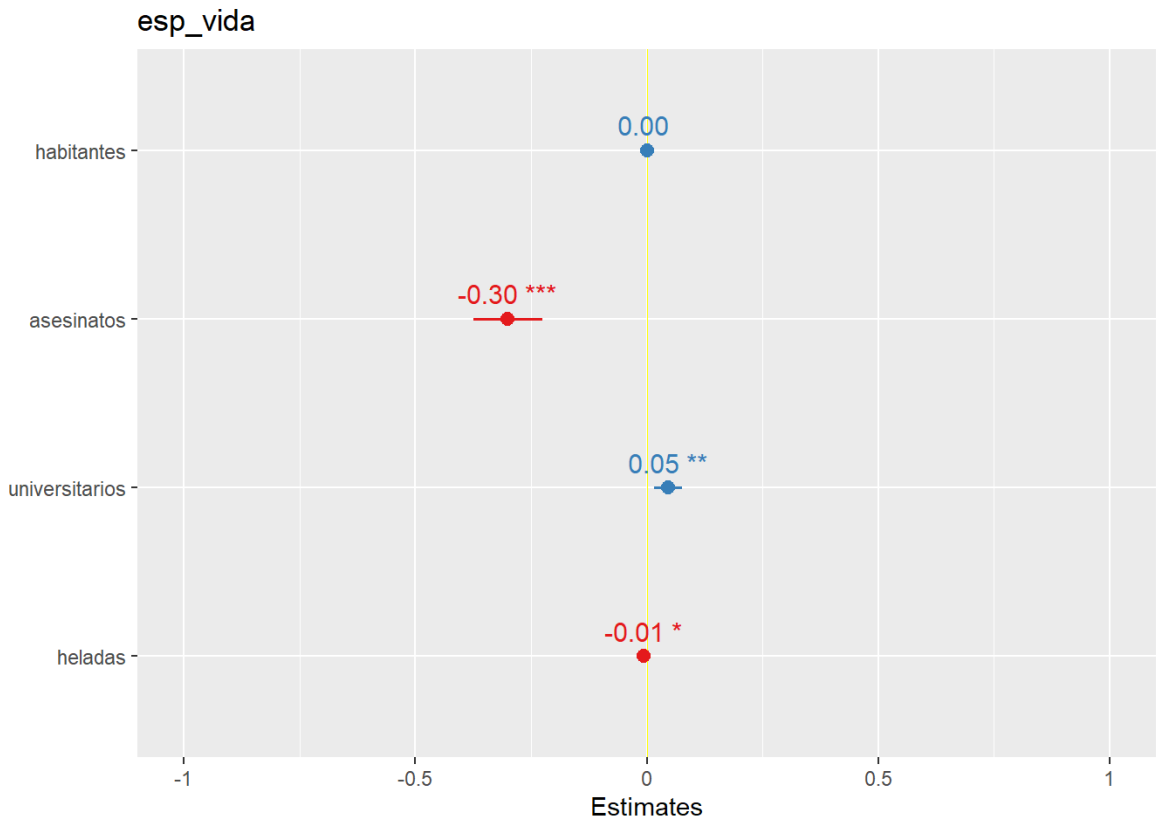
```
BIC(modelo1)
```

```
## [1] 127.2048
```

```
# Cálculo del error cuadrático medio RMSE:
sqrt(mean(modelo1$residuals^2))
```

```
## [1] 0.6827597
```

```
# Representando gráficamente la estimación e intervalo de confianza de lo
s coeficientes del modelo para apreciar los efectos descritos:

# Gráfico del ajuste

library (sjPlot)

plot_model(modelo1,

           show.values = TRUE,

           vline.color = "yellow")
```




```
# Inferencia sobre los parámetros del modelo

tab_model(modelo1,

          show.std = TRUE,

          show.r2 = FALSE)
```

	esp_vida				
Predictors	*Estimates*	*std. Beta*	*CI*	*standardized CI*	*p*
(Intercept)	71.03	-0.00	69.11 – 72.95	-0.15 – 0.15	**<0.001**
habitantes	0.00	0.17	-0.00 – 0.00	-0.00 – 0.34	0.052
asesinatos	-0.30	-0.83	-0.37 – -0.23	-1.03 – -0.62	**<0.001**

universitarios	0.05	0.28	0.02 – 0.08	0.10 – 0.46	**0.003**
heladas	-0.01	-0.23	-0.01 – -0.00	-0.42 – -0.04	**0.018**
Observations	50				

```
library (broom)

glance(modelo1)
```

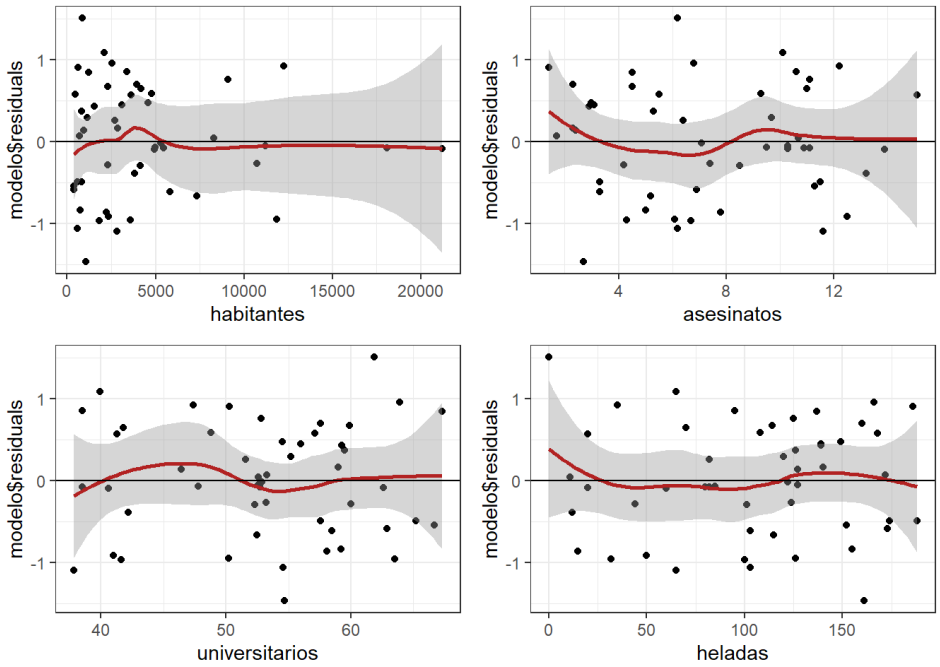
```
## # A tibble: 1 × 12

##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC
BIC

##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl>
<dbl>

## 1     0.736         0.713 0.720      31.4 1.70e-12     4  -51.9  116.
127.

## # ⓘ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

4. Validación de condiciones para la regresión múltiple lineal

Relación lineal entre los predictores numéricos y la variable respuesta:

Esta condición se puede validar 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.

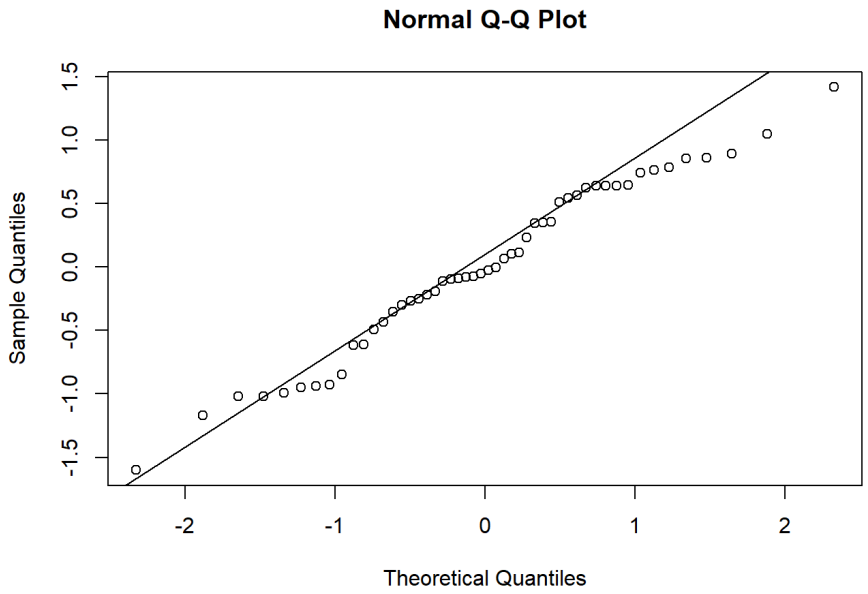
```
require(ggplot2)
require(gridExtra)
plot1 <- ggplot(data = datos, aes(habitantes, modelo$residuals)) + geom_point() +
  geom_smooth(color = "firebrick") + geom_hline(yintercept = 0) + theme_bw()
plot2 <- ggplot(data = datos, aes(asesinatos, modelo$residuals)) + geom_point() +
  geom_smooth(color = "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)
```



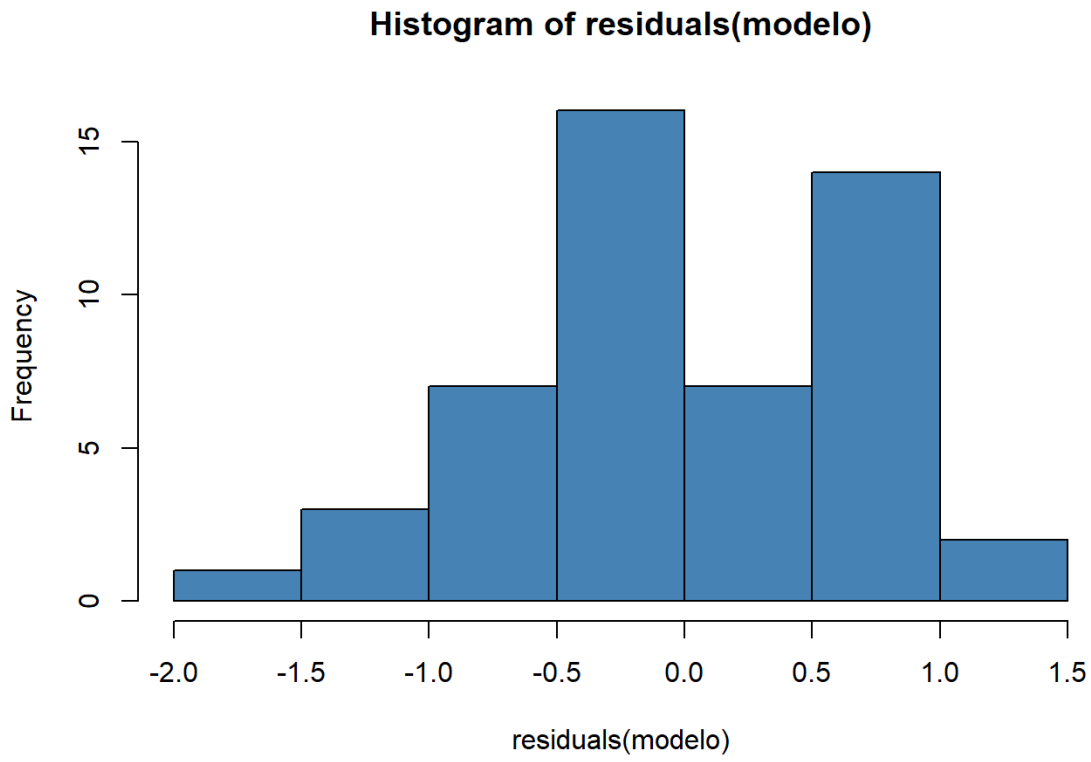
Se cumple la linealidad para todos los predictores, el ajuste es casi lineal (residuales frente a variables originales)

Distribución normal de los residuos:

```
# grafico de normalidad
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```



```
hist (residuals (modelo), col = "steelblue")
```




```
shapiro.test(modelo$residuals) #test de normalidad
```

```
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.97493, p-value = 0.3627
```

Tanto el análisis gráfico como es test de hipótesis confirma la normalidad $p(0.3627) > \alpha(0.05)$ con un valor puntual de w=0.97493.

```
ks.test(modelo$residuals,"pnorm")
```

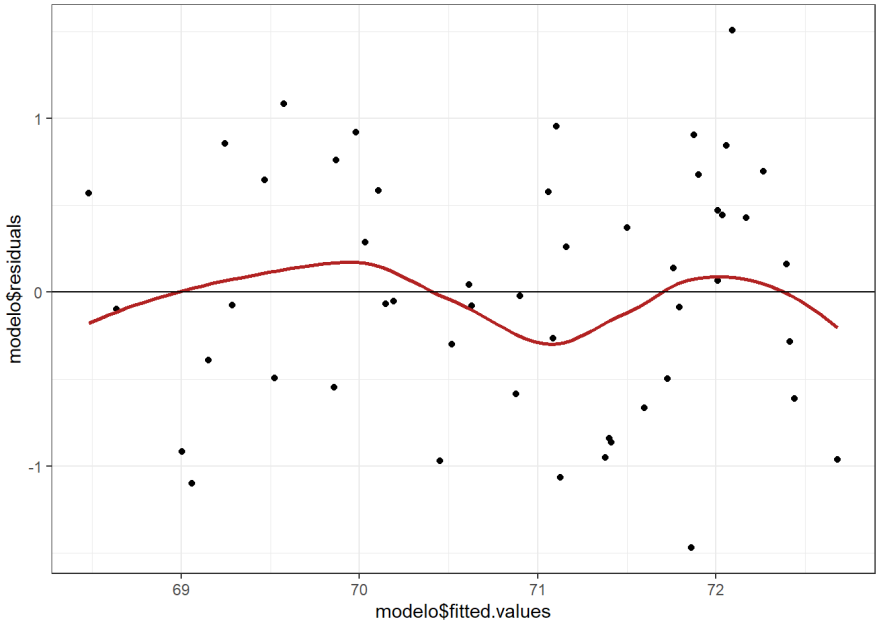
```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
```

```
## data:  modelo$residuals
## D = 0.14651, p-value = 0.2116
## alternative hypothesis: two-sided
```

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.

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



test de pagan

```
library(lmtest)
bptest(modelo)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo
```

```
## BP =10.605, df = 7, p-value = 0.1568
```

la figura y la prueba estadística evidencian que no hay falta de homocedasticidad $p(0.1568) > \alpha(0.05)$.

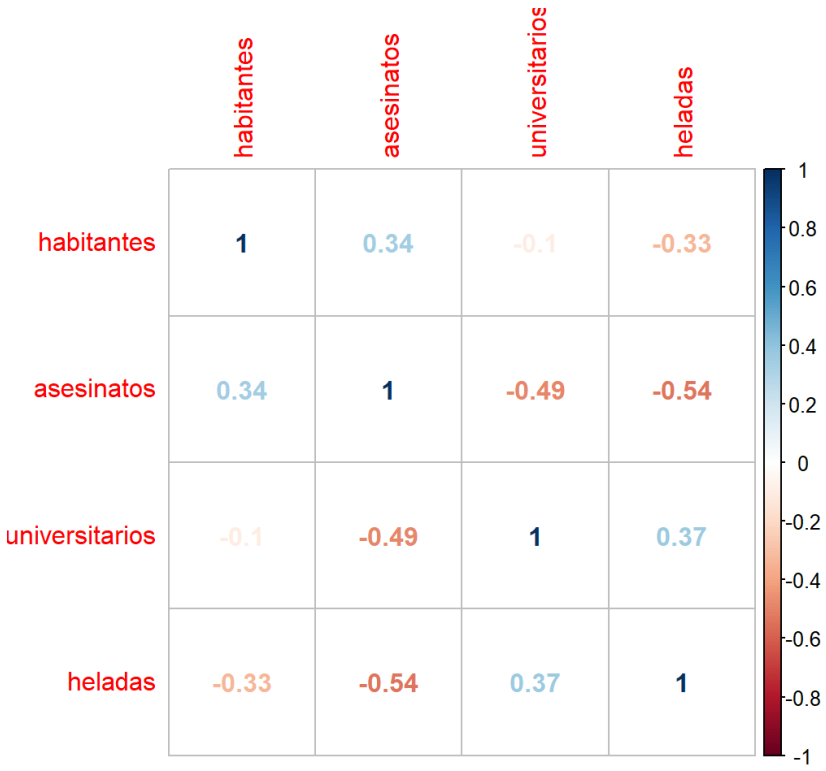
```
#Usabdo la prueba de varianza constante (Non-constant Variance Score Test
),
# H0:la varianza es constante en el ámbito de la predicción de Y
# prueba de homocedasticidad
# prueba varianza constante de errores
ncvTest(modelo)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.1232113, Df = 1, p = 0.72558
```

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):

```
# Analisis de Inflacion de Varianza (VIF):
require(car)
vif(modelo)
```

```
##     habitantes  analfabetismo     asesinatos universitarios        hel
adas
##       1.626706       4.827006       2.910884       2.765855       2.45
5787
##           area  densidad_pobl
##       1.859609       1.446292
```

```
# vif < 5 ausencia de multicolinealidad
# vif 5 a 10 multicolinealidad leve a moderada
# vif > 10 grave
```

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

Autocorrelación

```r
# Autocorrelacion:
require(car)
dwt(modelo, alternative = "two.sided") # durwin watson autocorr
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.03768782      2.040917    0.86
##  Alternative hypothesis: rho != 0
```

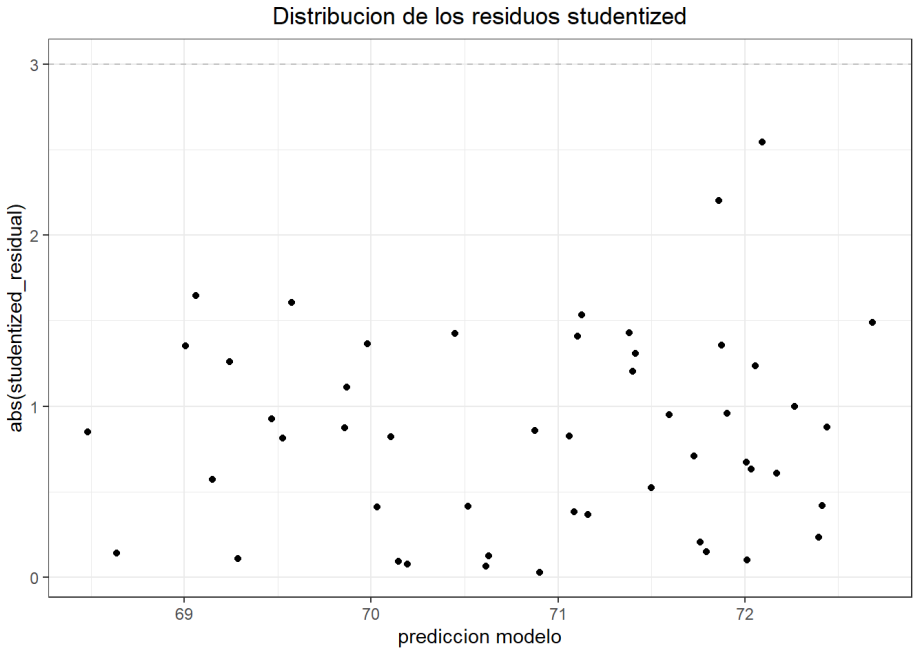
No hay evidencia de autocorrelación $p(0.86) > \alpha(0.05)$.

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.

4. Identificación de posibles valores atípicos o influyentes

```r
# Identificacion de posibles valores atipicos 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 = "Distribucion de los residuos studentized",
                                x = "prediccion modelo") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```




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

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

No se identifica ninguna observación atípica.

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

```
## Potentially influential observations of
##   lm(formula = esp_vida ~ habitantes + analfabetismo + asesinatos +
universitarios + heladas + area + densidad_pobl, data = datos) :
##
##    dfb.1_ dfb.hbtn dfb.anlf dfb.assn dfb.unvr dfb.hlds dfb.area dfb.dn
s_
## 2   0.01   0.41     0.09    -0.13     0.11    -0.02    -1.44_*  -0.52
## 5   0.04  -0.20    -0.01     0.04    -0.06     0.04    -0.01     0.05
## 11 -0.83  -0.22     0.89    -0.35     1.23_*  -0.44    -0.79    -0.23
## 28  0.26   0.18     0.09    -0.46    -0.24    -0.19     0.16    -0.06
## 30  0.00   0.03     0.02    -0.02     0.01     0.01    -0.03    -0.22
## 31 -0.09   0.02     0.11    -0.01     0.08     0.08    -0.04    -0.03
## 32  0.04  -0.08    -0.03     0.00    -0.03    -0.03     0.03     0.00
##    dffit   cov.r   cook.d hat
```

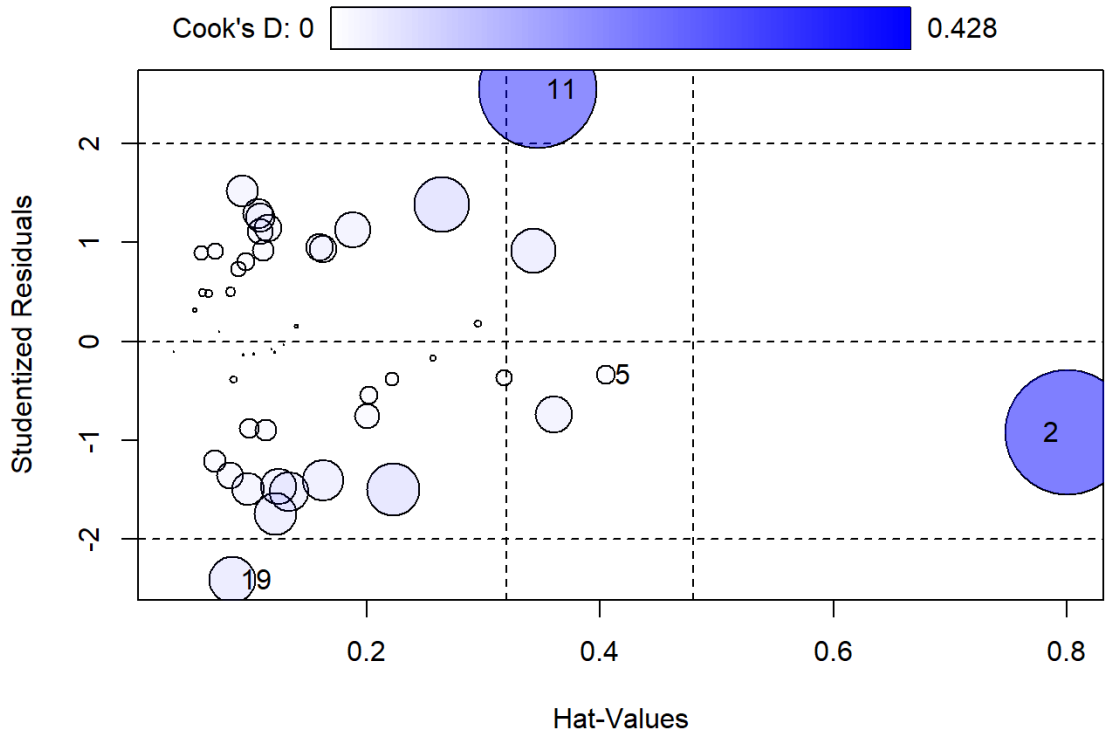
```
## 2  -1.85_*  5.17_*  0.43   0.80_*
## 5  -0.28    1.99_*  0.01   0.41
## 11  1.85_*  0.57    0.38   0.35
## 28 -0.55    1.71_*  0.04   0.36
## 30 -0.25    1.73_*  0.01   0.32
## 31  0.12    1.71_*  0.00   0.30
## 32 -0.10    1.62_*  0.00   0.26
```

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 $2.5((p+1)/n)$, 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
## 2  -0.9216865 0.8008140 0.42845807
## 5  -0.3407549 0.4050804 0.01009519
## 11  2.5443916 0.3465017 0.37960582
## 19 -2.4124148 0.0844454 0.06019022
```

Los análisis muestran varias observaciones influyentes (posición 5, 11, 19, 28 y 47) 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 es:

$$esp_{vida} = 71.03 + 0.00005014(habitantes) - 0.3001(asesinatos) + 0.04658(universitarios) - 0.00594(heladas)$$

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. Cinco observaciones (posición 5, 11, 19, 28 y 47) podrían estar influyendo de forma notable en el modelo.

```
# predicciones:

nuevas_observaciones <- data.frame(habitantes=3615,analfabetismo=2.1,ases
inatos=15.1,universitarios=41.3,heladas=20,area=50708,densidad_pobl=0.246
)

predict(object = modelo, newdata = nuevas_observaciones)
```

```
##     1
## 68.47425
```

```
# Punto atipico (outlier) y punto influyente

# Punto atipico (outlier): es una observacion que es numericamente distan
te del resto de los datos.

# Punto influyente: punto que tiene impacto en las estimativas del modelo
.


# Prueba de Bonferroni para detectar outliers

library(car)

outlierTest(modelo, cutoff=Inf, n.max=4)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 11  2.544392           0.014809      0.74047
## 19 -2.412415           0.020406           NA
## 40 -1.742964           0.088836           NA
## 4   1.523409           0.135330           NA
```

13. Ejemplo 2. 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
```

```
str(datos)
```

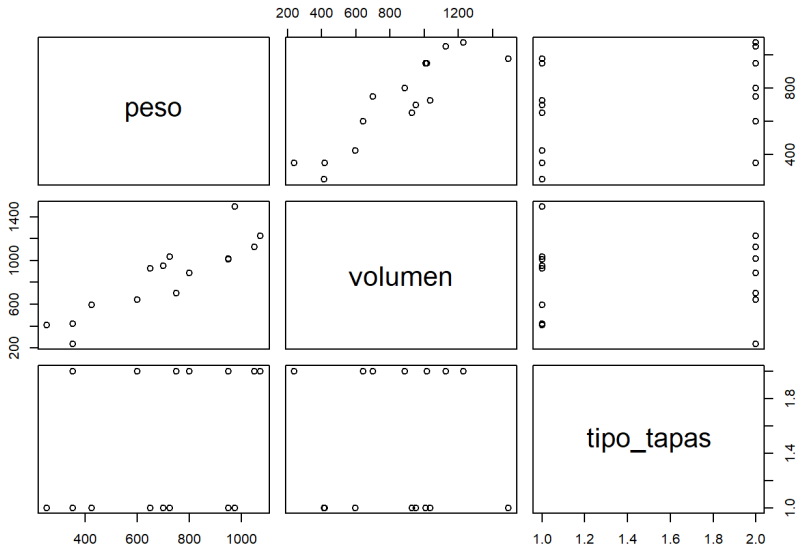
```
## 'data.frame'   :    15 obs. of  3 variables:
##  $ peso         : num  800 950 1050 350 750 ...
##  $ volumen   : num  885 1016 1125 239 701 ...
##  $ tipo_tapas : Factor w/ 2 levels "blandas","duras": 2 2 2 2 2 2 2 1 1 1 ...
```

La data tiene 3 variables y 15 observaciones de las cuales dos son numéricas y tipo_tapas, cualitativa.

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)
```

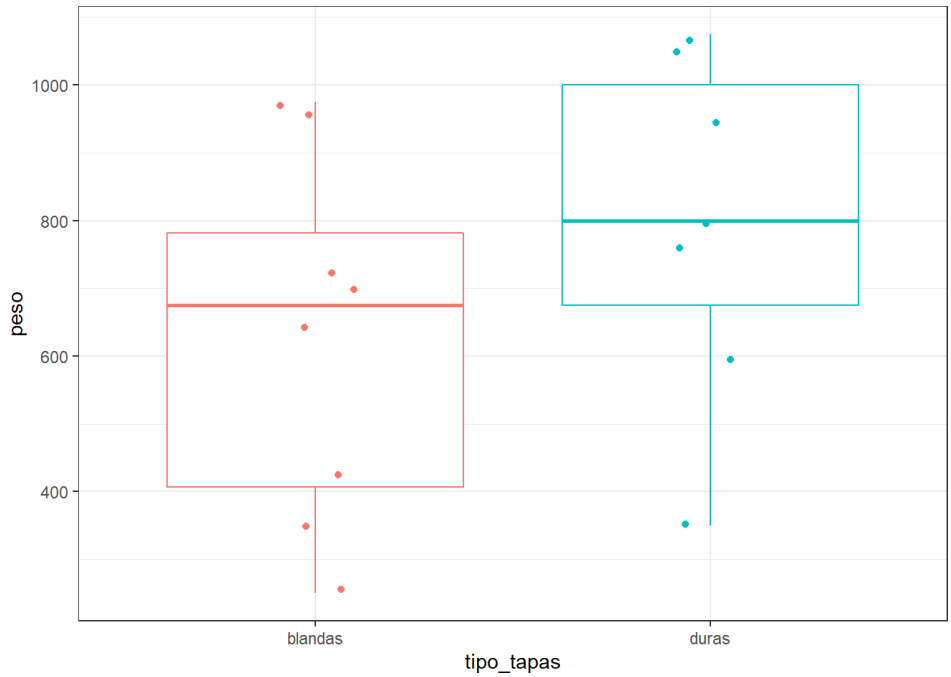



```
cor.test(datos$peso, datos$volumen, method = "pearson")
```

```
##  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
```

El análisis gráfico y de correlación (0.8959) 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.

```
library(ggplot2)
ggplot(data = datos, mapping = aes(x = tipo_tapas, y = peso, color =tipo_tapas)) +
  geom_boxplot() + geom_jitter(width = 0.1) +
  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
```

El modelo tiene un R^2 de 0.9275 (alto) y es significativo en su conjunto $p(0.000001455) < \alpha(0.05)$. Las variables individualmente son significativas excepto el intercepto.

```
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 dicho 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.

$$Peso\ libro = 13.91557 + 0.71795(volumen) + 184.04727(tipotapasduras)$$

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.

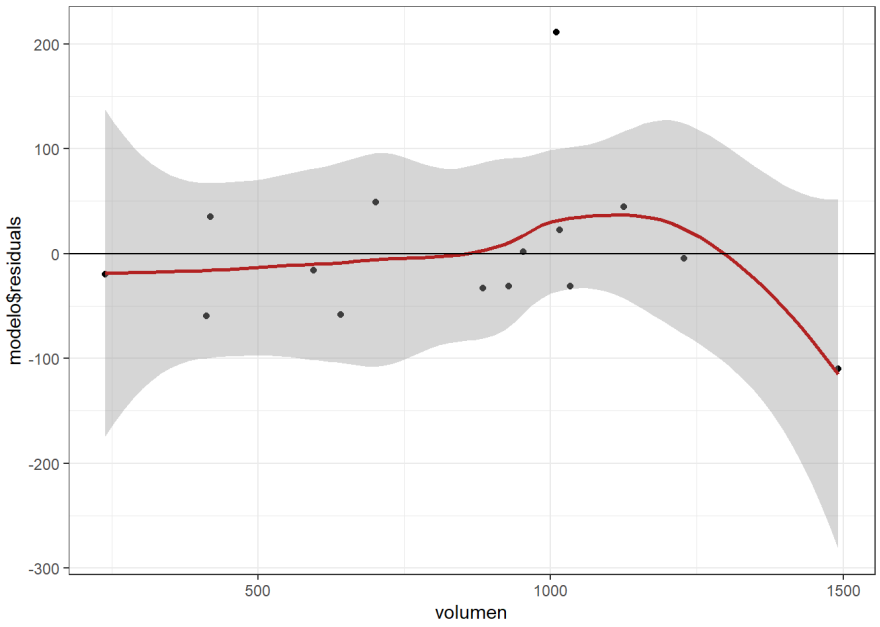
3. Elección de los predictores

En este caso, al solo haber dos predictores y ser significativos 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:

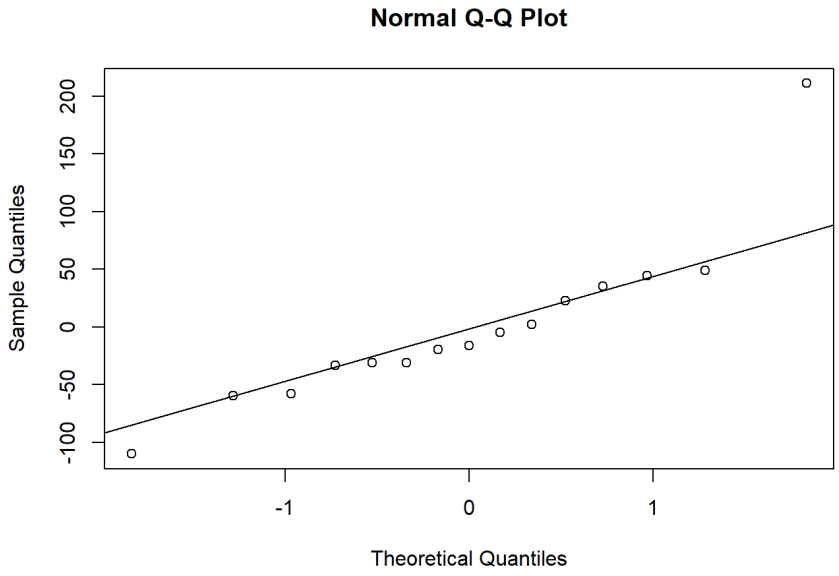
```
require(ggplot2)
ggplot(data = datos, aes(x = volumen, y = modelo$residuals)) +
  geom_point() +  geom_smooth(color = "firebrick") + geom_hline(yintercept = 0)
+ theme_bw()
```



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

2. Distribución normal de los residuos:

```
# Distribucion 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.

```
# ecluyendo valor atipico
which.max(modelo$residuals)
```

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

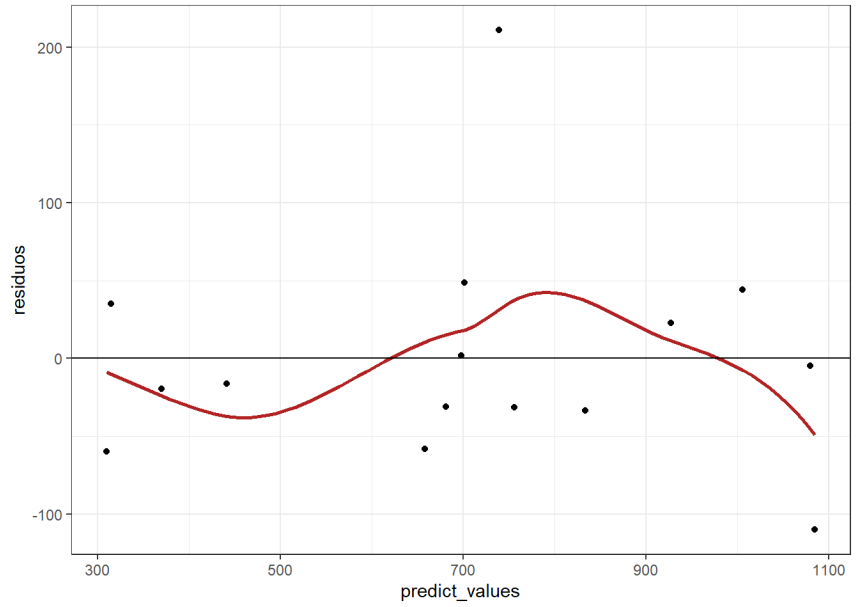
```
shapiro.test(modelo$residuals[-13])
```

```
##  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 ($p(0.7263) > \alpha(0.05)$) 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:

```
# 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", se = FALSE) + geom_hline(yintercept = 0) +
  theme_bw()
```




```
library(lmtest)
bptest(modelo)
```

```
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 2.0962, df = 2, p-value = 0.3506
```

No hay evidencias que indiquen falta de homocedasticidad $p(0.3506) > \alpha(0.05)$

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   0.816
##  Alternative hypothesis: rho != 0
```

No hay evidencia de autocorrelación $p(0.816) > \alpha(0.05)$

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. Identificación de posibles valores atípicos o influyentes

```
# Identificacion de posibles valores atipicos o influyentes
require(car)
outlierTest(modelo)
```

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

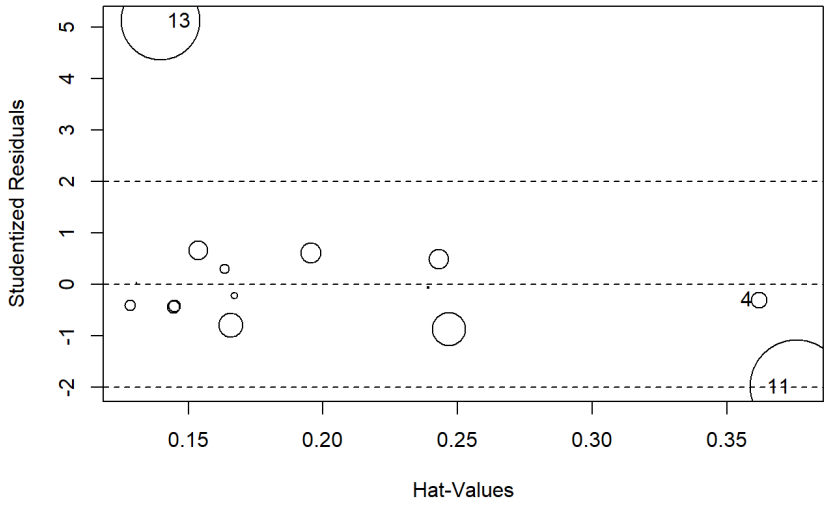
Tal como se apreció en el estudio de normalidad de los residuos, la observación 13 tiene un residuo estandarizado >3 (más de 3 veces la desviación estándar de los residuos) por lo que se considera un dato atípico. El siguiente paso es determinar si es influyente.

```
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
## 4     -0.3008849  0.3616894  0.01850173
## 11    -1.9897106  0.3757842  0.63729788
## 13     5.1268330  0.1397761  0.45819722
```

El análisis muestran varias observaciones influyentes aunque ninguna excede los límites de preocupación para los valores de *Leverages hat* (>2.5(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 es:

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

es capaz de explicar el 92.75% de la variabilidad observada en el peso de los libros (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.

14. 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 Xj 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 Xj es constante, independientemente del valor de Xj. Existen dos aproximaciones clásicas que permiten relajar estas condiciones cuando se trabaja con modelos lineales.