|  |
| --- |
|  |
| Modelo de regresión para ajuste de estimación del esfuerzo del análisis funcional |
|  |
| CDISM |
|  |

**CONTROL DE VERSIONES**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Versión** | **Apartado** | **Descripción** | **Fecha** | **Elaborado** | | **Revisado** |
| 01.00 | Todo | Primera versión | 06/04/2015 | Pedro Dulce S. |  | |
| 02.00 | Todo | Segunda versión | 09/04/2019 |  | Pedro Dulce S. | |
| 02.01 | Todo | Ajuste del modelo solo para aplicarlo en proyectos Pros@ | 12/04/2019 |  | Pedro Dulce S. | |
|  |  |  |  |  |  | |
|  |  |  |  |  |  | |

**ÍNDICE**

[1. introducción 2](#_Toc5727085)

[1.1. Motivaciones 2](#_Toc5727086)

[2. ELECCIÓN DEL MODELO 3](#_Toc5727087)

[2.1. Correlación entre tiempo calculado por HEPLLEM y tiempo de análisis 3](#_Toc5727088)

[2.2. Variables regresoras 4](#_Toc5727089)

[2.3. Desglose de variables regresoras 4](#_Toc5727090)

[2.4. Interacciones entre variables regresoras 5](#_Toc5727091)

[2.5. Modelo de regresión inicial 5](#_Toc5727092)

[2.6. Consideraciones de implementación 5](#_Toc5727093)

[2.7. Análisis detallado del modelo de regresión 6](#_Toc5727094)

[2.8. Significación global de los parámetros 7](#_Toc5727095)

[2.9. Significación individual de los parámetros 7](#_Toc5727096)

[2.10. Gráfico de la distribución normal de residuos 7](#_Toc5727097)

[2.11. Dispersión de la variable Y respecto a las regresoras 8](#_Toc5727098)

[2.12. Ajustes del modelo de regresión actual 9](#_Toc5727099)

[3. ANEXO. TEORÍA SOBRE BONDAD DE UN MODELO DE REGRESIÓN 9](#_Toc5727100)

[3.1. Aplicación del método MCO 9](#_Toc5727101)

[3.2. Bondad de un modelo de regresión 9](#_Toc5727102)

[3.3. Significación individual de variables explicativas 10](#_Toc5727103)

[3.4. Uso del valor p de probabilidad de aceptación de hipótesis 10](#_Toc5727104)

[3.5. Significación global bajo análisis con distribución F de Snedecor 11](#_Toc5727105)

# 

# introducción

## Motivaciones

Dentro de la estructura organizativa de GISS, que diferencia Asistencia Técnica de Desarrollo Gestionado, los responsables de proyectos en los Centros de Desarrollo se encuentran ante la necesidad de estimar con precisión los análisis funcionales, enfrentándose al factor de incertidumbre que supone calibrar el esfuerzo real de futuros análisis.

Con objeto de mitigar esa incertidumbre GISS liberó la herramienta de *Estimación de Proyectos Llave en Mano*, en adelante HEPLLEM, como un intento de cuantificar ese esfuerzo.

Las métricas que maneja dicha herramienta son la tecnología (Pros@, Java, Host), % de usabilidad, actores y casos de uso implicados, entidades de lectura y escritura por Caso de Uso,…

Ya se detectó en su momento que, para estimar evolutivos, la herramienta arrojaba un número excesivo de horas de esfuerzo estimado de análisis. Ese aumento desproporcionado es menor en estimaciones de nuevos desarrollos.

Para aminorar ese efecto a la hora de estimar las tareas de análisis, en el CDISM en 2015 se nos platea la posibilidad de que apliquemos porcentajes de reducción (coeficientes de ajuste) a esas estimaciones de HEPLLEM.

La información utilizada en este estudio parte de análisis funcionales efectuados en proyectos del CDISM, en el año 2015. En próximas versiones, enriqueceremos estos datos con los extraídos de la herramienta de imputación, ARTEMIS, para aquellas tareas de análisis de las que se han efectuado estimaciones con la herramienta HEPLLEM, dentro del ámbito del proyecto FOM2, hojas localizadas en la carpeta *O:\externos\PROSA\FOM2\2018-2019\Hojas de Estimación*.

El primer paso, es analizar si se puede ajustar con un sencillo coeficiente de reducción (ajuste lineal). Para ello, siempre ayuda plasmar en un gráfico cada valor de X (tiempo calculado por esa herramienta, en horas) con su correspondiente valor de tiempo realmente dedicado en horas, variable Y. Uniendo los puntos de la curva obtenida, ya se vislumbra si corresponde a una función lineal, exponencial, u oscilante.

Si no fuera posible obtener un coeficiente de reducción, diseñaremos un modelo de regresión lineal múltiple que explique, con más variables, cómo ajustar las horas de estimación de una forma más realista.

Ese modelo preliminar lo iremos testeando, partiendo en un comienzo de la interacción entre todas las variables elegidas, e iremos eliminando aquellas variables que no resulten significativas para explicar el tiempo de análisis que en realidad se dedica.

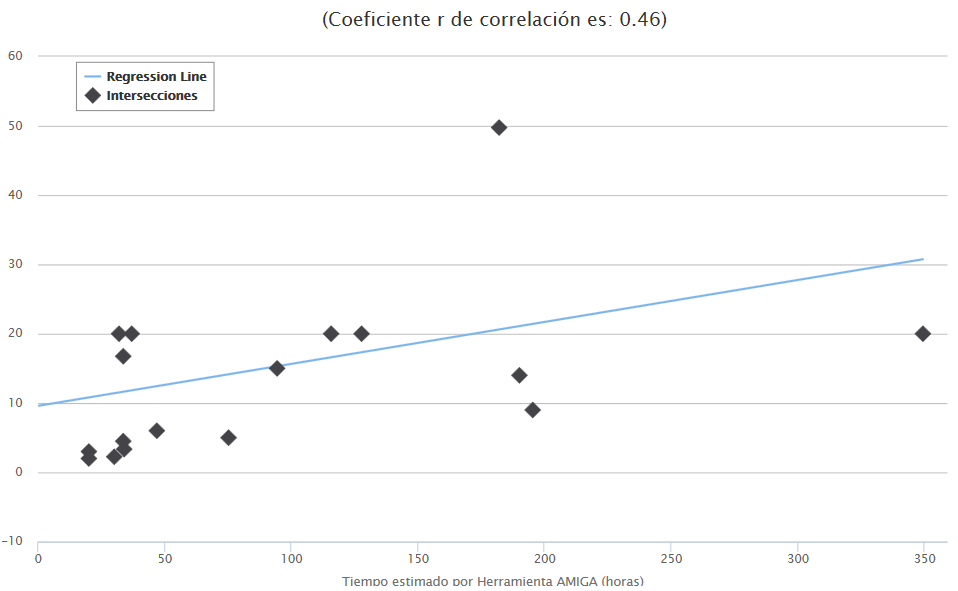
Finalmente, implementaremos en java ese modelo, y lo pondremos en marcha con datos que iremos extrayendo de proyectos Pros@ del CDISM. A más datos, los coeficientes de cada variable se irán ajustando para mejorar la fiabilidad de las estimaciones que arroje este modelo.

# ELECCIÓN DEL MODELO

## Correlación entre tiempo calculado por HEPLLEM y tiempo de análisis

Tomando datos de diferentes análisis de varios proyectos, dibujamos, en el eje Y la variable dependiente que deseamos hallar, *tiempo de análisis*, y en el eje X, el *tiempo calculado por HEPLLEM*, ambas variables continuas.

En base a la forma de la curva, intuimos el tipo de función que la define.



Unimos los puntos (x,y) para pintar la curva.

* Si la curva es lineal, no será necesario realizar transformación alguna, y habrá que calcular el coeficiente de correlación, b, y el término independiente: *y = b\* x + a*

Normalmente, esto viene determinado por un coeficiente de correlación cercano a 1, en el rango [0..1].

* Si la curva mostrase aspecto de función exponencial, *y=aebX*, tendríamos que realizar las siguientes transformaciones al modelo:

*Y =* β*1\* X +* β*0*

*ln y = ln b\*x + a*

Es decir, habría que resolver el conjunto de ecuaciones donde para cada una, habrá que aplicar el logaritmo a cada valor de *x* e *y*.

* Si la curva es oscilante, como sucede para este estudio, entonces existen otras variables que influyen para determinar la variabilidad de Y, no es suficiente con *la variable tiempo calculado por HEPLLEM*, nuestra variable X del estudio.

## Variables regresoras

Estudiamos un posible modelo de regresión lineal múltiple a partir de estas cuatro variables explicativas, para estimar la variable dependiente ***tiempo análisis***:

* ***tiempo calculado HEPLLEM***
* ***tipo de tarea***
* ***tipo componente***

Inicialmente incorporaremos todas las variables y sus interacciones; posteriormente haremos un estudio que nos permita decidir cuáles conviene mantener en el modelo en el caso de que tengan significancia para explicar la variable **tiempo análisis**, que llamamos Y.

En forma matricial, el modelo siempre sigue la expresión **Y = β\*X + µ**

## Desglose de variables regresoras

* Consideramos la variable **TIME** como el tiempo calculado por la herramienta corporativa HEPLLEM. Esta variable es cuantitativa y continua, y su unidad de medida es *hora*.
* La variable cualitativa **TASK** representa el tipo de tarea; es cualitativa, y tiene 3 categorías. Las categorías identificadas en el estudio para el tipo de tarea son:
* Evolutivo Medio
* Evolutivo Largo
* Nuevo Desarrollo

Precisamos dos variables indicadoras para su definición formal dentro del modelo, ya que haremos combinaciones binarias.

Asignamos valores diferentes de cada variable indicadora para cada categoría, manteniendo la independencia lineal de los vectores resultantes.

Y éstas, las combinaciones de las tres variables indicadoras por cada categoría:

|  |  |
| --- | --- |
| TASK \_I TASK \_II  Nuevo Desarrollo 1 0  Evolutivo Medio 0 1  Evolutivo Largo 0 0 | Valores de constantes  1  2  3 |

No consideramos ni los correctivos ni los pequeños evolutivos, al estar tasados en DG, y por tanto, considerarse sus análisis de tiempo poco significativo.

OJO: observar que si hubiéramos escogido la dupla (1,1) para la categoría ‘Evolutivo Largo’, sería combinación lineal de (1,0), (0,1), con [1,1] por tanto, no es válida dicha elección.

* Consideramos la variable cualitativa **COMPONENTE**. Las categorías identificadas en el estudio para el tipo de componente son:
* Pros@ BATCH
* Procedimiento WF
* Tarea WF
* Servicio Pros@

Precisamos tres variables indicadoras para su definición formal dentro del modelo.

Asignamos valores diferentes de cada variable indicadora para cada categoría, manteniendo la independencia lineal de los vectores resultantes.

Estas son las combinaciones resultantes:

|  |
| --- |
| COMP \_I COMP \_II COMP\_III  Pros@ BATCH 1 0 0  Procedimiento WF 0 1 0  Tarea WF 0 0 1  Servicio Pros@ 0 0 0 |

|  |
| --- |
| Valores de constantes  1  2  3  4 |

Nota: usar este algoritmo de reparto a la hora de distribuir las categóricas dummy en el modelo y en sus interacciones:

* asignar valor 1 a las variables dummy en posición = (valor\_constante -1 )
* Si esa posición no existe, o no coincide con (valor\_constante -1), otorgar valor 0.

## Interacciones entre variables regresoras

Al haber incorporado variables dummy o indicadoras, tenemos que evaluar las posibles interacciones, y posteriormente, analizaremos si tienen incidencia en la variabilidad que explica el tiempo real de análisis, Y.

**Las interacciones a tener en cuenta han de contener la variable continua, TIME.**

## Modelo de regresión inicial

**Y = β0 + β1\*TIME + β2\*TASK\_I + β3\*TASK\_II + β4\*COMP\_I + β5\*COMP \_II + β6\*COMP\_III +**

**β7\*(TIME \* TASK\_I) + …+ β23\*(TIME \* TASK\_II\* COMP\_III) + µ**

Así, β5 explica la variabilidad del tiempo de análisis respecto a análisis funcionales sobre definición e implementación de un procedimiento WF en Pros@, mientras que β7 explica la variabilidad concreta del tiempo de análisis respecto a un nuevo desarrollo.

## Consideraciones de implementación

Se aplicará el método **MCO**, ***Mínimos Cuadrados Ordinarios*** (ver anexo), haciendo uso de la librería Java *Open Source* de Apache, Commons org., **Commons-Math3** ([*http://commons.apache.org/proper/commons-math/index.html*](http://commons.apache.org/proper/commons-math/index.html)).

Para la resolución del modelo planteado, vamos a tomar como datos aquellos trabajos de análisis finalizados del proyecto GESE (CDINSS), en el año 2015.

En la clase **MultipleRegressionModelTester** del package de PCM, se define en la variable estática **nombres\_Regresoras** las variables que definen nuestro modelo.

En el método **makeLinearRegressionAnalysisTeams** se pasan los *arrays* de los datos siguientes, que son ingredientes de nuestro modelo planteado:

* tiempos de análisis (dato real, variable *y*)
* tiempo calculado por HEPLLEM
* tipo análisis
* tecnología

Desde este método se invoca a setAnalysisXvarsWithInteractions, donde se da valor tanto a las variables continuas y dicotómicas, como a las indicadores de las variables categorizadas, en el momento de resolver los coeficientes del modelo (las betas).

Para hallar el valor esperado/estimado de Y de un futuro análisis, usaremos el método getXVarsNormalized2Model, al que pasaremos los valores de tiempos de análisis, tiempo calculado por HEPLLEM, tipo análisis y tecnología.

<parar e implementar este modelo en la clase Java, y luego seguir con el documento>

## Analizamos de forma detallada el modelo de regresión usando una muestra de ejemplo

Para la muestra dada:

*<extraer los datos, necesitas Artemis y GEDEON…> pintar solo las variables independientes, no las combinaciones, por sencillez*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Grados de libertad de la regresión: 7 | Grados de libertad residuos: 10 | Grados de libertad total: 17 |  |
| **Error Squared Sum (SCE)** |  |  |  | observaciones: 18 | | |  |
| 2.044,4190 |  |  |  | Valor crítico T (alfa=0.025, 12) | | Valor crítico F (alfa=0.05, 7, 10) |  |
| µ0: 2,2064 |  |  |  | 2,1788 | | 3,1355 |  |
| µ1: 15,9274 |  |  |  |  |  |  | intervalos confianza |
| µ2: 7,9759 |  |  | **Coeficientes (Beta)** | **Standard Error de Beta** | **Estadístico t-Student** | **Valores p-Beta** | **IC Betas** |
| µ3: -6,3368 |  |  | Beta0: 102,7085 | Standard Err Beta0: 47,3476 | t de Beta0: 2,1692 | p de Beta0: 0,0509 | IC Beta0: [ -0,453, 205,87 ] |
| µ4: -5,0909 |  | X1 | Beta1: -0,0226 | Standard Err Beta1: 0,455 | t de Beta1: -0,0496 | p de Beta1: 0,9613 | IC Beta1: [ -1,0139, 0,9688 ] |
| µ5: 0,0 |  | X2 | Beta2: -67,5257 | Standard Err Beta2: 16,7791 | t de Beta2: -4,0244 | p de Beta2: 0,0017 | IC Beta2: [ -104,0842, -30,9672 ] |
| µ6: 7,6104 |  | X3 | Beta3: -68,5546 | Standard Err Beta3: 94,5163 | t de Beta3: -0,7253 | p de Beta3: 0,4822 | IC Beta3: [ -274,488, 137,3787 ] |
| µ7: -35,8786 |  | X4 | Beta4: -62,9163 | Standard Err Beta4: 41,4441 | t de Beta4: -1,5181 | p de Beta4: 0,1549 | IC Beta4: [ -153,2153, 27,3828 ] |
| µ8: 1,1688 |  | X5 | Beta5: 0,5346 | Standard Err Beta5: 0,0832 | t de Beta5: 6,4224 | p de Beta5: 0,0 | IC Beta5: [ 0,3532, 0,7159 ] |
| µ9: 3,0972 |  | X6 | Beta6: 0,5885 | Standard Err Beta6: 1,0244 | t de Beta6: 0,5745 | p de Beta6: 0,5763 | IC Beta6: [ -1,6436, 2,8206 ] |
| µ10: 1,0487 |  | X7 | Beta7: 0,3856 | Standard Err Beta7: 0,4331 | t de Beta7: 0,8902 | p de Beta7: 0,3909 | IC Beta7: [ -0,5582, 1,3293 ] |
| µ11: -2,7473 |  |  |  |  |  |  |  |
| µ12: 0,0 |  |  |  |  |  |  |  |
| µ13: 6,044 |  |  |  |  |  |  |  |
| µ14: -3,2967 |  |  |  |  |  |  |  |
| µ15: -2,9325 |  |  |  |  |  |  | Desviación típica del error |
| µ16: 14,8923 |  |  | **R2/R corregido** | **F value** | **Squared Sum Regression (SCR)** | **Total Squared Sum (SCT)** | **St Dev. Err Regression** |
| µ17: -3,6883 |  |  | 0,9966/0,9943 | 424,7232 | 607.818,5255 | 609.862,9444 | 14,2983    **Analizar si este dato es correcto, aplicando la fórmula me arroja 10,3, ¿qué se me está escapando?** |

## Significación global de los parámetros

Dado que el valor muestral, para nuestro modelo, F (7,10) = **3.1355** es menor que el de la distribución Ƒ (tabla f-snedecor) en ese punto, Ƒ (7, 10)0.05 = **3.64**, no existe evidencia en la muestra para no rechazar H0: β1=β2= β3= β4=…=β8=β9=0, por tanto, aún podemos mejorar este modelo. *Ver apartado 3.5 del anexo.*

Observamos que el valor de **R2** es bastante cercano a 1, en concreto, **0,9966**; esto se interpreta como que el **99.66%** de la variabilidad de la variable **Y** respecto a su promedio es explicada por el modelo de regresión planteado.

## Significación individual de los parámetros

En cuanto a la significación individual de cada variable e interacciones, analizamos si el valor 0 (*para la hipótesis nula*) está contenido en el intervalo de confianza (IC) del parámetro de esa variable, o si su valor *p* es inferior/superior a α (0.05), para rechazar o no la hipótesis nula. *Ver apartados 3.3, 3.4 del Anexo.*

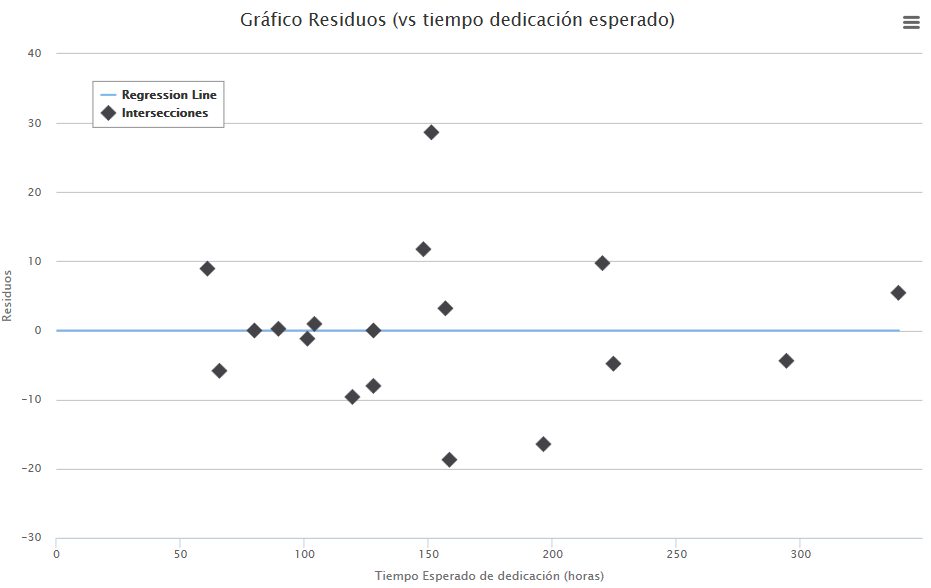
Por ejemplo, el valor *p* para β2 (*estimador del parámetro que explica la variabilidad de TECH sobre Y*), está por debajo de 0.05, en concreto, 0.0017; rechazamos la hipótesis nula ya que contamos con suficiente certeza (más del 95%) de que la variable *TECH* incide en la variabilidad del tiempo de análisis real (variable *y*).

Para el estimador β7 bajo la muestra analizada, su intervalo de confianza contiene al valor 0, por tanto, no podemos rechazar la hipótesis nula H0: β7=0.

Es importante hacer notar que cuando se decide mantener una interacción en el modelo es necesario siempre mantener cada una de las variables (por separado) que intervienen en dicha interacción, por tanto, si decidimos mantener la interacción *TIME\*TECH\*TASK\_I*, hemos de mantener en el modelo las interacciones de orden inferior, *TIME\*TECH, TECH\*TASK\_I, TIME\*TASK\_I,* así como las variables por separado, *TIME*, *TECH y TASK\_I,* aunque no resulten a priori significativas bajo este test. Dicho de otra forma, **Si se introducen en un modelo de regresión términos de interacción y resultan estadísticamente significativos, no se podrán eliminar del modelo los términos de interacción de orden inferiores ni los términos independientes de las variables que participan en la interacción para simplificarlo, deben mantenerse, aunque no resulten estadísticamente significativos.**

## Gráfico de la distribución normal de residuos

Dibujamos el gráfico de la distribución que siguen los residuos por aplicación del modelo de regresión elegido, para reconocer si es adecuado o no nuestro modelo. Comprobamos que ***µ* ~N (0, σ²)**, donde **σ²** es la varianza residual del apartado 7:

****

Observamos que existen atípicos, por debajo de -10, y por encima de 10, que podrían indicar la presencia de una componente cuadrática no incluida.

## Dispersión de la variable Y respecto a las regresoras

Vamos a dibujar el gráfico de la distribución que siguen tanto cada una de las interacciones, como la variable continua (*tiempo estimado de la herramienta*), con respecto a la variable dependiente tiempo de análisis.

Observamos un coeficiente de correlación alto (próximo a 1) para los evolutivos bajo Pros@.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Observando el resto de gráficos, no podemos extraer conclusiones contundentes por falta de datos.

## Ajustes del modelo de regresión actual

En el futuro, dispondremos de más datos que nos permitirán ajustar, si procede, el modelo actual.

## Modelo de regresión para estimación de carga de trabajo a futuro.

Sería interesante reflejar de algún modo en las presentaciones al comité, no sólo gráficas de los datos con la variable X siendo los meses, y la variable y, el volumen de cada tipo de actuación, sino además, un estudio relativo a la previsión de determinada carga de trabajo.

Plantear una tabla dataset que tenga columnas genéricas, para dar cabida a las gráficas de dispersión y timeline. Dicha tabla debería tener el campo *fecha\_entrada\_al\_CD*.

# ANEXO. TEORÍA SOBRE BONDAD DE UN MODELO DE REGRESIÓN

## Aplicación del método MCO

A partir del método de mínimos cuadrados (MCO, OLS en inglés, *ordinary least square*), se iguala a CERO la suma de todas las ecuaciones lineales de la derivada del cuadrado del error de cada observación *i* de la muestra:

*∑ I n (yi – β0– β1\* X1 – β2\*X2 - ……. - βK\*XK) ) = 0*

El estimador MCO para el vector β es insesgado y de mínima varianza, y sigue la expresión: ***βˆ = (X′X )-1 X′Y***

El promedio del error ha de ser 0:  ***E (µ) = 0***

Donde ***µ* ~N (0, σ²I)**, que se deduce:

* por la propiedad de *homoscedasticidad*: Var (ui) = Var (un) = σ²; en el gráfico de distribución de residuos
* Esta propiedad determina una dispersión uniforme/simétrica de los errores de estimación.
* Por la propiedad de no autocorrelación, el resto de elementos de la diagonal es 0 en la matriz de varianzas covarianzas.

## Bondad de un modelo de regresión

El **coeficiente de determinación** (**R2**) en la regresión lineal es una medida de la bondad de ajuste de la recta estimada a los datos reales.

Suma de cuadrados debida al error: SCE = Σ (*yi* – ŷi )2

Suma de cuadrados total: SCT = Σ (*yi* – ***ȳ****)*2

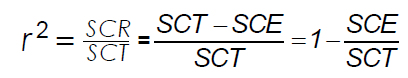
Suma de cuadrados debida a la regresión: SCR = Σ (ŷi - ***ȳ****)*2

Relación entre SCT, SCR y SCE: SCT = SCR + SCE

*yi* *es el valor que toma la variable y en la i-ésima observación*

***ȳ*** *es la media de los datos observados de la variable y*

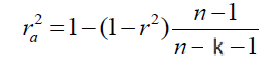
ŷi *es el valor estimado para la observación i-ésima aplicando el modelo de regresión*

**Coeficiente de determinación: **

Expresado en porcentaje, se puede interpretar como el porcentaje de la variabilidad total de “Y” que se puede explicar aplicando la ecuación de regresión.

Cuando la varianza residual (SCE) es muy próxima a cero, r2 se acerca a 1; el modelo explica aprox. el 100% de la variabilidad del valor de la variable *y*; en este caso la interpretación sería que la suma de los residuos es en proporción muy pequeña en relación a la suma de cuadrados total (SCT) o variabilidad total de *y*.

A veces es interesante calcular el **coeficiente de determinación ajustado**, a partir de la fórmula:



## Significación individual de variables explicativas

Realizaremos un estudio basado en el contraste de hipótesis para determinar si el valor de Y es atribuible a cada variable regresora elegida, a partir del estadístico *T-Student*, que sigue una distribución cercana a una Normal centrada en 0, más aproximada a la Normal cuanto mayor es el tamaño de la población del estudio.



*En este caso, y para este tamaño de muestra, hacemos uso de la clase* ***TDistribution*** *del paquete* ***org.apache.commons.math3.stat.distribution****.*

|  |  |
| --- | --- |
| **Figura 1**  *k: grados de libertad + 1*  *α representa el nivel de significancia, y suele tomarse el valor 0.05 (representa el 95% de certeza).* | Una forma de contraste se realiza obteniendo el intervalo de confianza (IC) de un *βi*, definido a partir del rango: ***[βi +/- (StandardErr\_βi\* t(n-k) (α/2))]***  Si el valor 0 dado a *βi* en la H0 está dentro del rango de no rechazo, aceptamos H0.  Otra forma de contraste se realiza calculando el valor *p* de probabilidad que se obtiene al hallar la ordenada en esa distribución para el valor ***ti*** de cada *βi*, donde:    Si el valor de |ti| es mayor que la ordenada de ***t(α/2),*** o *p* es inferior al nivel de significancia elegido, rechazamos H0, es decir, esa variable resulta significativa para explicar la variabilidad de y. |

## Uso del valor p de probabilidad de aceptación de hipótesis



Decisión: Se rechaza H0 en favor de H1 si |ti |> ***t(α/2)*** o si su p-valor < α.

En otras palabras, admitiremos con una certeza igual o superior al 95% el rechazo de la hipótesis nula (H0), es decir, mantendremos la variable X*i*.

Interpretación de p: un valor p=0.002 denota que rechazaríamos la hipótesis H0 siendo verdadera 2 veces de cada 1000 (error de tipo I).

El valor *p* lo hemos calculado a partir de la coordenada de la distribución T, ya que estamos calculando la probabilidad en las dos colas, por tanto, el valor obtenido con esa función es p/2.

***double*** *pValue = CommonUtils.redondearA4Decimales(t\_Beta\_i > 0 ? (1 - T\_distribution\_n\_k.cumulativeProbability(t\_Beta\_i)) \* 2 : T\_distribution\_n\_k.cumulativeProbability(t\_Beta\_i) \* 2);*

**Es importante reseñar que no podremos eliminar del modelo ninguna de las variables Xi, Xj, cuando se decida mantener la interacción entre ambas, Xi\*Xj.**

## Significación global bajo análisis con distribución F de Snedecor

Este estadístico global de contraste usa dos estimaciones de la varianza, una basada en CME, otra basada en CMR.



  
CME es un estimador insesgado de la varianza, mientras que CMR lo es sólo si H0 es cierta. Si H0 es falsa, CMR tiende a sobreestimar la varianza de la regresión.

El estadístico de contraste, bajo H0 es una F, y cumple esta ecuación:

Despejando en dicha ecuación, siendo *k* el número de variables independientes o grados de libertad, y *n* el de observaciones, tenemos:

**= =** H0 ˜ Ƒ (k, n – k – 1)

Recordemos que y 

Los valores en la distribución Ƒ los obtenemos mediante la clase FDistribution del paquete ***org.apache.commons.math3.stat.distribution***.

|  |  |
| --- | --- |
| Si el valor muestral de F para α=0.05 es mayor que su valor crítico en la función de distribución Ƒ (k, n-k-1) para 0,05, rechazamos H0; concluiremos que existe al menos uno de los coeficientes βi, donde i> 0, que contribuye a explicar Y. | El objetivo es comparar nuestro modelo contra el modelo restringido, en el que suponemos que el único parámetro que mantenemos es β0, el intercepto, y el resto los ignoramos por no ser significantes para la variabilidad de Y, es decir, la hipótesis es H0: β1= β2= … βk = 0 |