Come calcolare i p-value dei coefficienti di regressione di LINESTX

PUBBLICATO LUGLIO 11, 2023 *DI* FRANCESCO BERGAMASCHI E DAVID BIANCONI

La funzione DAX *LINESTX* è stata introdotta recentemente in Power BI Desktop allo scopo di permettere la stima dei coefficienti di una regressione lineare. Da allora sono emersi diversi video e articoli su come usarla. Questo articolo non verterà su questo aspetto, di conseguenza. Per l'uso di *LINESTX* rimandiamo all'ottimo articolo (corredato di video) di sqlbi (link).

In questo articolo ci preoccuperemo, invece, come preannunciato in questo articolo, di calcolare i *p-value* dei coefficienti stimati. È, infatti, molto importante ricordare che le regressioni lineari non possono essere usate per fare stime se non si fanno delle verifiche di ipotesi. Non ci addentreremo nel complesso sistema teorico che sta dietro questa affermazione ma non possiamo non dire che ogni stima potrebbe essere un abbaglio.

La statistica è così: nulla è certo. Ogni affermazione ha un certo livello di confidenza. Senza la stima dei *p-value* (e altre verifiche sui residui che non tratteremo qui ma che sono più semplici in software dedicati come R e che possono essere fatti in Power BI usando la R-*visual*), usare una regressione lineare per prendere decisioni è molto pericoloso.

Useremo un modello molto semplice, come sempre. In figura 1 sono visibili tre tabelle, tra loro scollegate. I dati sono nella tabella importata *SampleData*. Le altre due tabelle (*Criteria* e *Linear Regression Output Enriched*) sono state introdotte per, rispettivamente, creare le soglie di non rifiuto dell'ipotesi nulla tra cui scegliere e mostrare l'output di *LINESTX*, che è una tabella, arricchita dai calcoli dei *p-value*. Il report che verrà creato è mostrato in figura 2, dove tutto è risolto attraverso misure e l'output arricchito dai *p-value* non è osservabile ma dove è visibile l'uso della tabella *Criteria* (riquadro rosso).

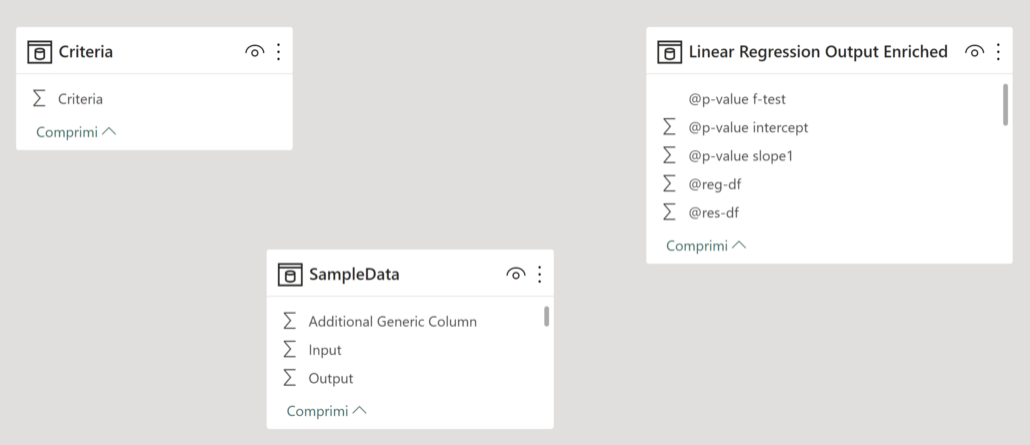

Figura 1

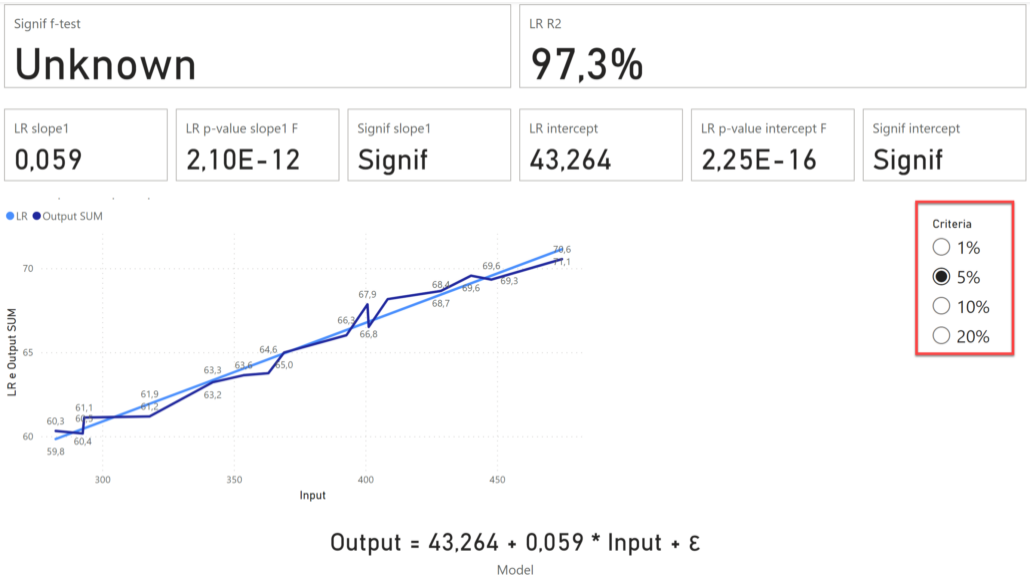

Figura 2

In figura 3 è mostrata la tabella *SampleData*. La tabella contiene un insieme di dati macroeconomici relativi al numero di persone occupate (*Output*) e al prodotto nazionale lordo o PNL – deflazionato ai valori del 1954 – (Input) rilevati negli Stati Uniti negli anni 1947-1962 (*Additional Generic Column*). Questo data set fa parte di uno studio più ampio realizzato da J. W. Longley (1967). Siamo interessati a spiegare l'andamento del numero di occupati (*Output*) in funzione del PNL (*Input*). Più avanti nell'articolo verrà mostrato che la tabella può avere un altro contenuto grazie ad un parametro di Power Query.

Input	Output	Additional Generic Column
282,276	60,323	1947
293,137	61,122	1948
292,578	60,171	1949
317,988	61,187	1950
341,970	63,221	1951
353,720	63,639	1952
369,076	64,989	1953
363,112	63,761	1954
392,756	66,019	1955
400,746	67,857	1956
408,458	68,169	1957
401,215	66,513	1958
428,689	68,655	1959
440,106	69,564	1960
447,859	69,331	1961
474,674	70,551	1962

Figura 3

Sviluppo

L'output della funzione *LINESTX* è un tabella, che riporta una singola riga con diverse colonne. Ognuno dei valori rappresenta un termine che descrive il modello statistico stimato tramite la regressione. Ecco i diversi output forniti nella stima di un modello come il seguente:

y = intercept + slope1 * x1 + slope2 *x2 + … + slopeN * xN + ε

- SlopeN (un output per ogni espressione x)
- Intercept
- StandardErrorSlopeN (un output per ogni espressione x)
- StandardErrorIntercept
- CoefficientOfDetermination (R2)
- StandardError
- FStatistic
- DegreesOfFreedom
- RegressionSumOfSquares
- ResidualSumOfSquares

Nella lista non sono presenti i *p-value* né del test F (intero modello) né delle stime dei singoli coefficienti stimati (*SlopeN*). Tuttavia, è possibile calcolare i p-value dei singoli coefficienti stimati come segue.

Per prima cosa, osserviamo l'output generato da *LINESTX*, nella tabella calcolata *Linear Regression Output Enriched*. Essa è generabile con il seguente codice:

```
Linear Regression Output =
LINESTX ( SampleData, [Output], [Input] )
```

La tabella creata è visibile in figura 4

Slope1	Intercept	StandardErrorSlope1	StandardErrorIntercep	CoefficientOfDetermi	StandardError	FStatistic	DegreesOfFree	RegressionSumOfSc	ResidualSumOfSquares
0,0587255735	43,264225811	0,00260453571229945	0,989312825329342	0,97	0,595114187787254	508,38637248	14	180,05057344892	4,95825255107956

Figura 4

Tra i vari risultati, *Intercept* fornisce la stima dell'intercetta della retta di regressione. Avendo un solo regressore (*Input*) per cercare di spiegare *Output*, la tabella contiene soltanto una colonna di stima di coefficienti di regressione (*Slope1*). Tale valore è una stima di pendenza (quantifica, per ogni incremento unitario dell'input, di quanto aumenta l'output).

Ecco come arricchire la tabella con i *p-value* dell'intercetta e della pendenza:

```
Linear Regression Output Enriched =
VAR SampleTable = SampleData
VAR SampleSize =
   COUNTROWS ( SampleData )
VAR NumOfRegressors = 1
RETURN
   ADDCOLUMNS (
      ADDCOLUMNS (
         ADDCOLUMNS (
            LINESTX ( SampleTable, [Output], [Input] ),
            "@t-test slope1", DIVIDE ( [Slope1], [StandardErrorSlope1] ),
            "@t-test intercept", DIVIDE ( [Intercept], [StandardErrorIntercept] ),
            "@reg-df", NumOfRegressors
         ),
```

```
        "@p-value slope1", T.DIST.2T ( ABS ( [@t-test slope1] ), SampleSize – 2 ),
        "@p-value intercept", T.DIST.2T ( ABS ( [@t-test intercept] ), SampleSize – 2 ),
        "@res-df", [DegreesOfFreedom] – [@reg-df]
    ),
    "@p-value f-test",
        "Unknown" — dovrebbe essere F.DIST.RT([FStatistic],[@reg-df],[@res-
df] ma in DAX questa funzione non esiste —
    )
```

Ecco le colonne addizionali che sono state calcolate (figura 5), con in evidenza i *p-value* di interesse:

@t-test slope1	@p-value slope1	@t-test intercept	@p-value intercept	@reg-df	@res-df	@p-value f-test
22,5474249634048	2,10491700760481E-12	43,7315929840811	2,24957877793893E-16	1	13	Unknown

Figura 5

Le colonne *Linear Regression Output Enriched[@t-test slope1]* e *Linear Regression Output Enriched[@t-test intercept]* rappresentano un passo intermedio per arrivare al calcolo dei *p-value* dell'intercetta e del regressore; le colonne *Linear Regression Output Enriched*[@reg-df] e *Linear Regression Output Enriched*[@res-df] servirebbero al calcolo del *p-value* del test F che, tuttavia, non è possibile al momento e da qui la scelta di attribuire il valore "Unknown" a questa colonna. Il *p-value* del test F è, in effetti, il più importante perché chiarisce la significatività dell'intero modello; la ragione dell'attuale impossibilità del calcolo è la mancanza in DAX della funzione *F.DIST.RT* che è invece disponibile in Excel. Non ci addentriamo ulteriormente in questi complessi concetti in questo articolo.

Adesso che possiamo disporre dei *p-value* dei coefficienti del modello di regressione, passiamo alla sintesi delle misure e della tabella *Criteria*, che permettono di creare il report mostrato in figura 2.

La tabella *Criteria* è stata creata usando il tasto "Immettere i dati" in Power BI Desktop ed ha i tipici valori soglia che si usano in Statistica: 1%, 5%, 10%, 20%.

Ecco le misure:

```
Signif f-test =
"Unknown"
```

```
LR R2 =
VAR Line =
    LINESTX ( ALLSELECTED ( SampleData ), [Output], [Input] )
VAR Result =
    SELECTCOLUMNS ( Line, [CoefficientOfDetermination] )
RETURN
    Result
```

```
LR slope1 =
VAR Line =
    LINESTX ( ALLSELECTED ( SampleData ), [Output], [Input] )
VAR slope =
    SELECTCOLUMNS ( Line, [Slope1] )
RETURN
    slope
```

```
LR p-value slope1 =
VAR SampleSize =
    COUNTROWS ( SampleData )
VAR Line =
    ADDCOLUMNS (
        ADDCOLUMNS (
            LINESTX ( ALLSELECTED ( SampleData ), [Output], [Input] ),
            "@t-test slope1", DIVIDE ( [Slope1], [StandardErrorSlope1] )
        ),
        "@p-value slope1", T.DIST.2T ( ABS ( [@t-test slope1] ), SampleSize – 2 )
    )
VAR Result =
    SELECTCOLUMNS ( Line, [@p-value slope1] )
RETURN
    Result
```

```
LR p-value slope1 F =
FORMAT ( [LR p-value slope1], "Scientific" )
```

```
LR p-value slope1 F =
FORMAT ( [LR p-value slope1], "Scientific" )


Signif slope1 =
IF (
    [LR p-value slope1] < SELECTEDVALUE ( Criteria[Criteria] ),
    "Signif",
    "NOT Signif"
)


LR intercept =
VAR Line =
    LINESTX ( ALLSELECTED ( SampleData ), [Output], [Input] )
VAR Result =
    SELECTCOLUMNS ( Line, [Intercept] )
RETURN
    Result


LR p-value intercept =
VAR SampleSize =
    COUNTROWS ( SampleData )
VAR Line =
    ADDCOLUMNS (
        ADDCOLUMNS (
            LINESTX ( ALLSELECTED ( SampleData ), [Output], [Input] ),
            "@t-test intercept", DIVIDE ( [Intercept], [StandardErrorIntercept] )
        ),
        "@p-value intercept", T.DIST.2T ( ABS ( [@t-test intercept] ), SampleSize - 2 )
    )
VAR Result =
    SELECTCOLUMNS ( Line, [@p-value intercept] )
RETURN
    Result
```

```
LR p-value intercept F =
FORMAT ( [LR p-value intercept], "Scientific" )


Signif intercept =
IF (
    [LR p-value intercept] < SELECTEDVALUE ( Criteria[Criteria] ),
    "Signif",
    "NOT Signif"
)


Output SUM =
SUM ( SampleData[Output] )


LR =
VAR Line =
    LINESTX ( ALLSELECTED ( SampleData ), [Output], [Input] )
VAR slope =
    SELECTCOLUMNS ( Line, [Slope1] )
VAR intercept =
    SELECTCOLUMNS ( Line, [Intercept] )
VAR x =
    SELECTEDVALUE ( SampleData[Input] )
VAR Result = x * slope + intercept
RETURN
    Result



Model =
IF (
    [Signif slope1] = "Signif"
        && [Signif intercept] = "Signif",
    "Output = " & FORMAT ( [LR intercept], "#,0.00#" ) & " + "
        & FORMAT ( [LR slope1], "#,0.00#" ) & " * Input + ε",
```

```
IF (
    [Signif slope1] = "Signif",
    "Output = " & FORMAT ( [LR slope1], "#,0.00#" ) & " * Input + ε",
    IF (
        [Signif intercept] = "Signif",
        "Output = " & FORMAT ( [LR intercept], "#,0.00#" ) & " + ε",
        "No statistical significance"
    )
)
```

In figura 2, tuttavia, i p-value sono talmente piccoli da portare al non rifiuto dell'ipotesi nulla in qualunque condizione di selezione delle soglie (valore selezionato della colonna *Criteria[Criteria]*). Per evidenziare la cosa, mostriamo qui di seguito la figura 2 con una modifica del valore soglia in un valore più restrittivo (dal 5% al 1%), come si vede sia l'intercetta che la pendenza stimate rimangono significative. Valori meno restrittivi, ovviamente, porterebbero a nessuna modifica ancora a maggior ragione.

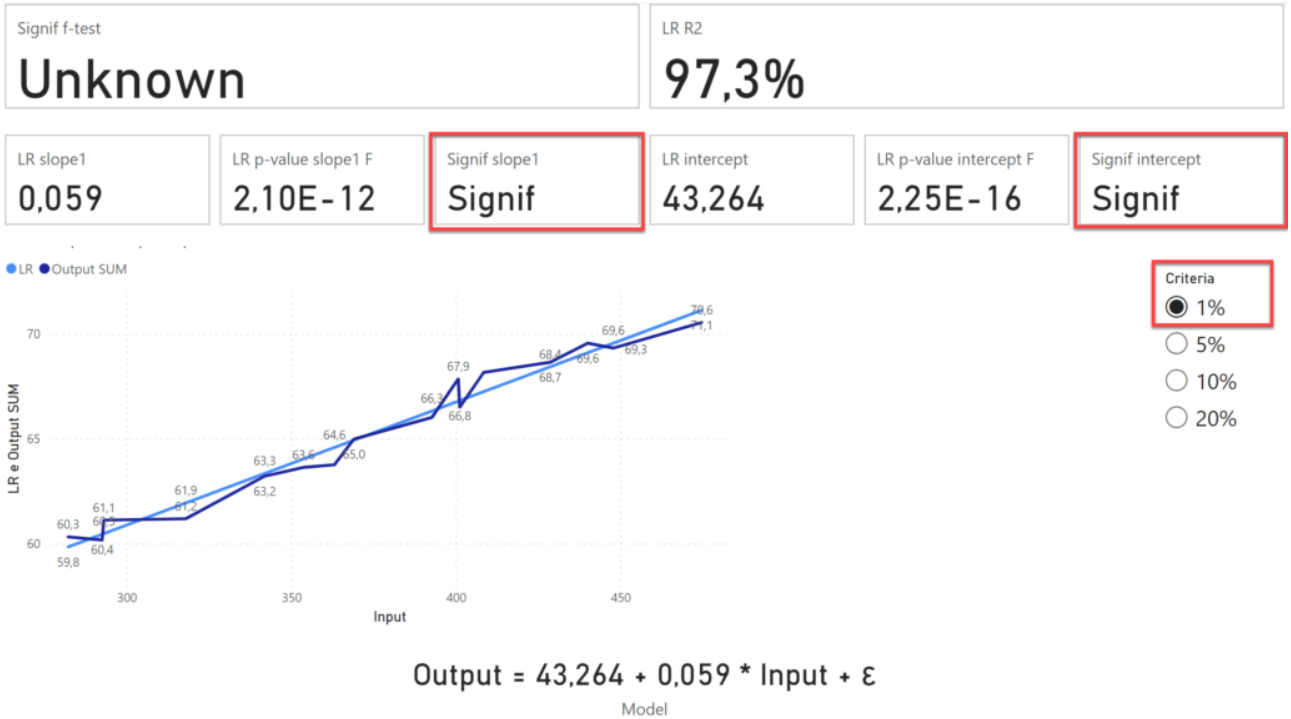

Figura 6

Per permettere di apprezzare la dinamicità del report, come preannunciato ad inizio articolo, è stato creato un parametro in Power Query che permette di modificare la tabella *SampleData* in modo da riempirla con dei dati fake creati da noi e che portano ad una stima di intercetta con p-value più grandi, che permettono di osservare il cambiamento di significatività sulla base della selezione della soglia di non rifiuto. A tale scopo, si proceda come in figura 7.

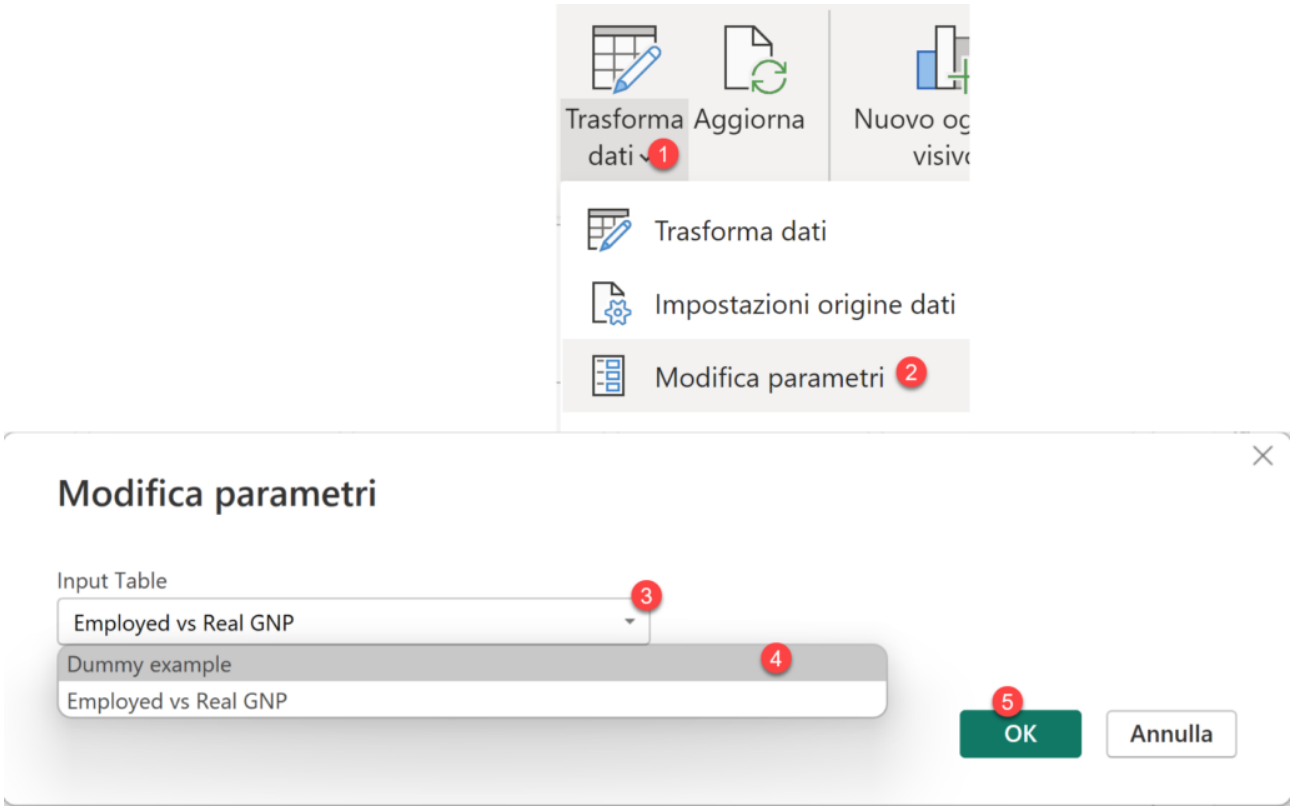

Figura 7

Si clicchi, infine, su "*Applica modifiche*" in Power BI Desktop. Ecco come si presenta adesso la tabella *SampleData*:

Input	Output	Additional Generic Column
1,000	3,000	
2,000	6,000	
3,000	7,000	
4,000	10,000	
5,000	11,000	

Figura 8

Ed ecco come si presenta il report di figura 2:

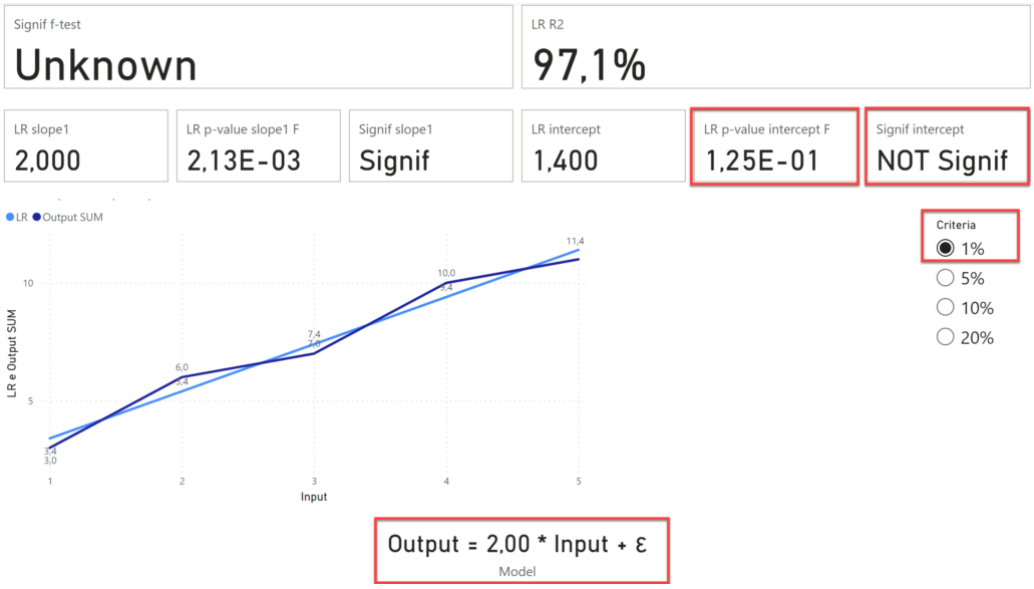

Figura 9

Come si vede in figura 9, il *p-value* dell'intercetta è 0,125 cioè il 12,5%. Questo valore è non significativo (rifiuto ipotesi nulla) quando la soglia di non rifiuto è al 1% come mostrato in figura. Si noti che, in questo caso, l'intercetta non compare nell'equazione del modello. Ecco cosa succede se seleziona una soglia che porta al non rifiuto (20%).

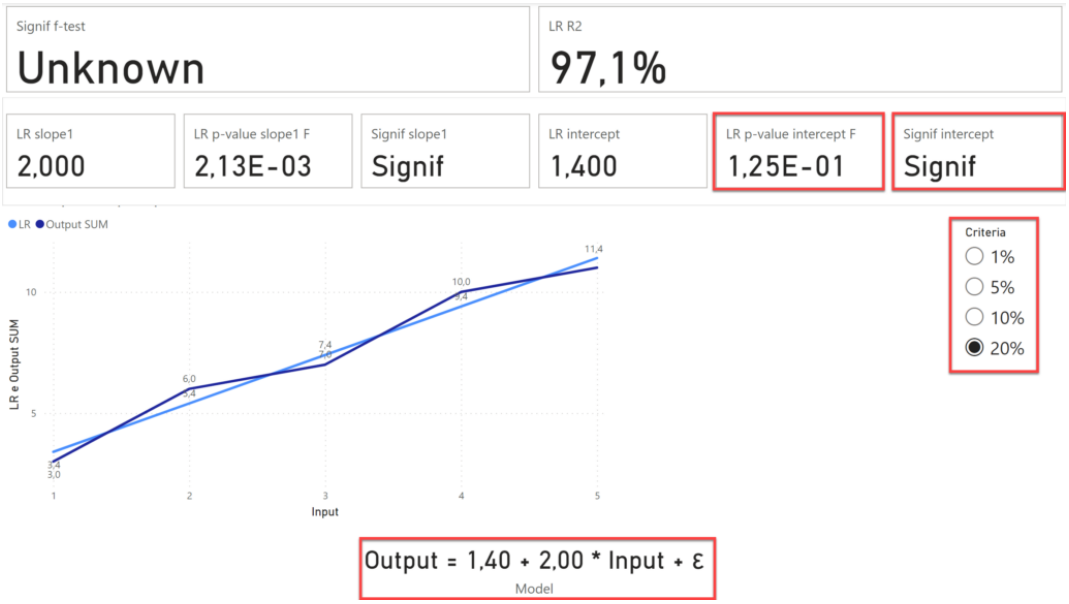

Figura 10

Se ci fosse interesse per l'uso della R-visual in Power BI Desktop fatecelo sapere nei commenti e faremo un articolo al riguardo, come anticipazione, ecco un paio di schermate.

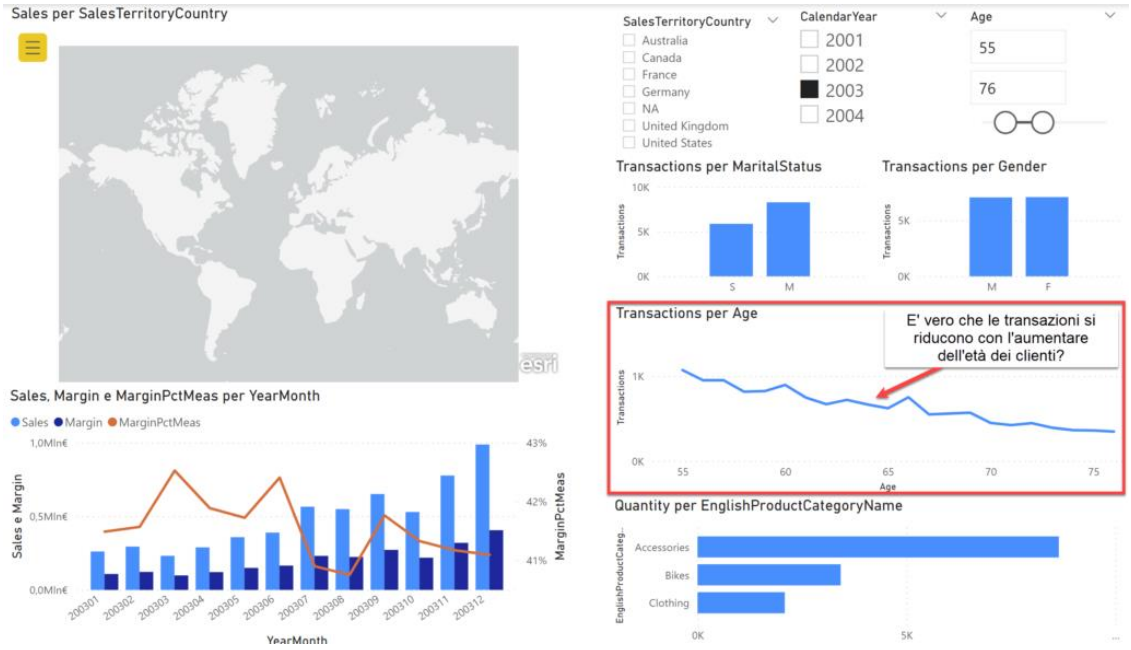

Figura 11

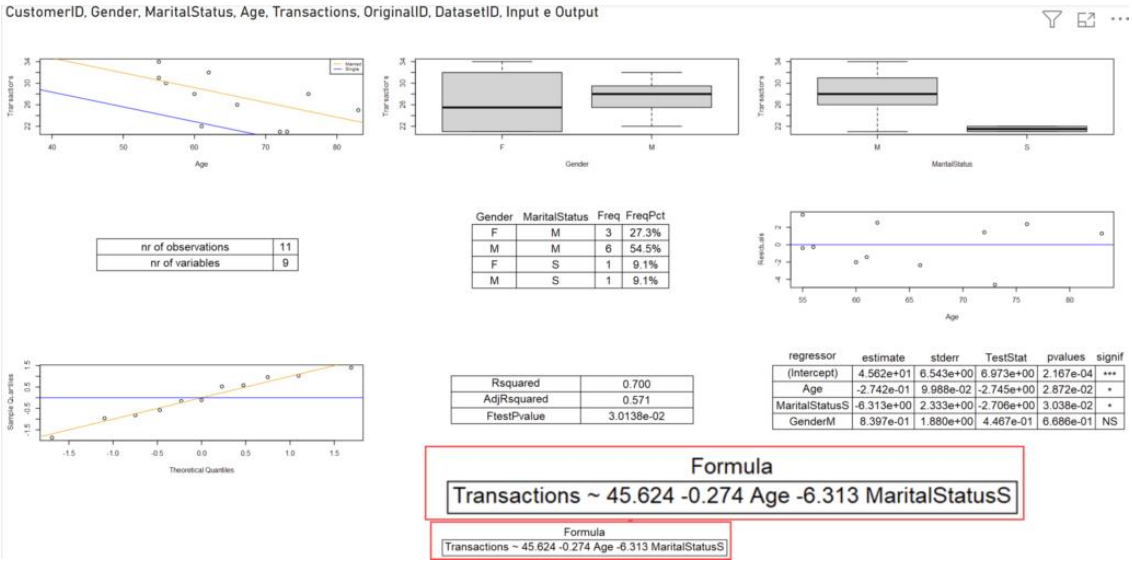

Figura 12

Conclusioni

È possibile calcolare i p-value dell'intercetta e delle pendenze stimate da LINESTX. Purtroppo, ad oggi, non è possibile farlo per il test F. Speriamo Microsoft crei la funzione DAX *F.DIST.RT* che permetterebbe di farlo. Ciò porterebbe alla possibilità di fare tutto in DAX. Tuttavia, restano più comodi i software statistici che possono essere usati dentro Power BI Desktop. Ciò che porta al desideri odi fare tutto in DAX è la bellezza dei report Power BI rispetto a quelli in R e la flessibilità del data model. Si veda questo articolo per apprezzare la possibilità di usare *slicer* in Power BI Desktop e osservare l'intercetta e la pendenza cambiare in real-time nel report.

file .pbixDownload

Autore del Post

Francesco Bergamaschi e David Bianconi

See author's posts