Federico Luo 882912

Nicoleta Mihalachi 875130

Nadia Ouadi 887572

**LOAN DEFAULTS**

*Dati anonimizzati di Loan Default per prevedere se un prestito sarà insolvente o meno*

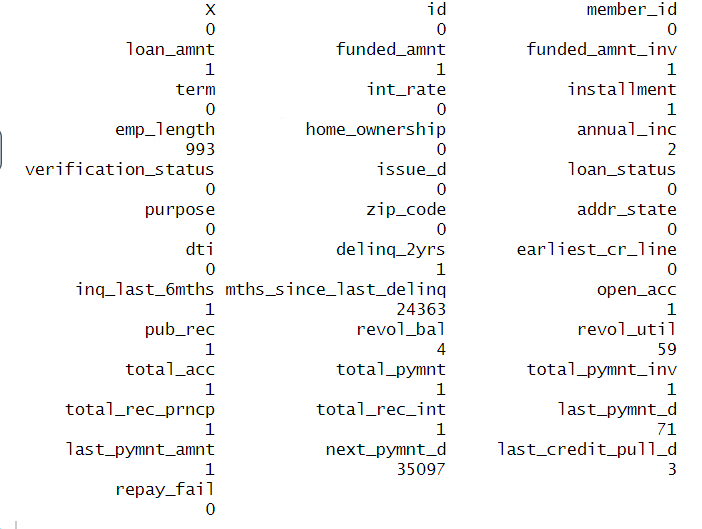
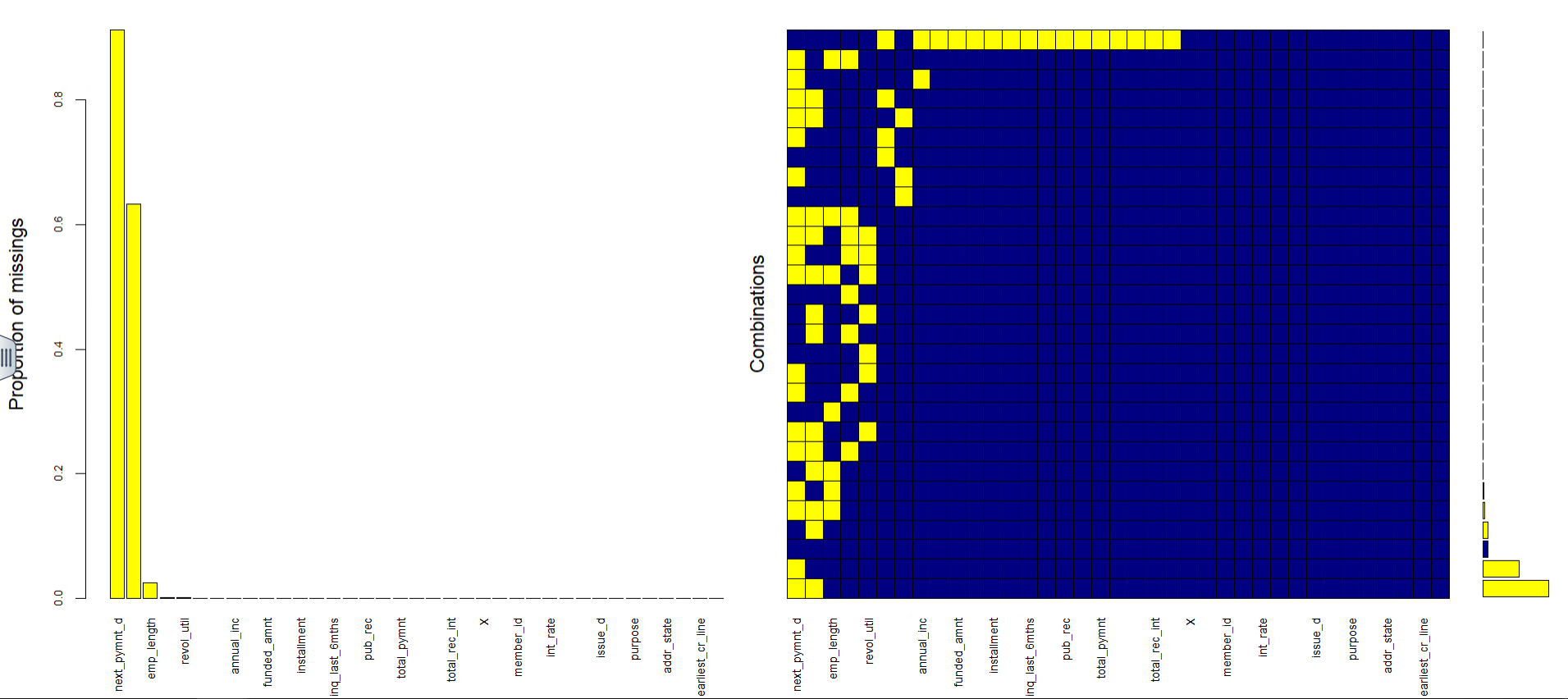
**SCOPO:**

* Analizzare i fattori che influenzano l'ammontare effettivamente finanziato attraverso un modello di regressione lineare
* Analizzare i fattori che influenzano la capacità di un debitore nel riuscire a rimborsare o meno il prestito (tenendo conto degli interessi)

**MATERIALE E DESCRIZIONE:**

* Il dataset utilizzato, reperito da Kaggle, contiene molte osservazioni : [LOAN DEFAULTS](https://www.kaggle.com/datasets/joebeachcapital/loan-default)
* Descrizione variabili : **`id`**: Un identificatore unico per ciascun prestito, **`member\_id`**: Un identificatore unico per ciascun membro,`**loan\_amnt**`: L'ammontare del prestito richiesto,`**funded\_amnt**`:L'ammontare effettivamente finanziato,`**funded\_amnt\_inv**`: L'ammontare effettivamente investito nel prestito,`**term**`: La durata del prestito in mesi (ad esempio, 36 o 60 mesi),`**int\_rate**`: Il tasso di interesse del prestito,`**installment**`: L'importo delle rate mensili del prestito,`**emp\_length**`: L'anzianità lavorativa dell'individuo (ad esempio, "5 years"),`**home\_ownership**`: Il tipo di proprietà dell'individuo (ad esempio, "own", "mortgage", "rent", ecc.),`**annual\_inc**`: Il reddito annuale dell'individuo,`**verification\_status**`: Lo stato di verifica del reddito (ad esempio, "verified" o "not verified"),`**issue\_d**`: La data di emissione del prestito,`**loan\_status**`: Lo stato attuale del prestito (ad esempio, "Fully Paid", "Charged Off", ecc.),`**purpose**`: Lo scopo del prestito (ad esempio, "debt\_consolidation", "credit\_card", ecc.),`**zip\_code**`: Il codice postale dell'individuo, `**addr\_state**`: Lo stato in cui risiede l'individuo, `**dti**`: Il rapporto debito/reddito dell'individuo,`**delinq\_2yrs**`: Il numero di pagamenti in ritardo negli ultimi 2 anni,`**earliest\_cr\_line**`: La data di apertura del primo conto di credito,**`inq\_last\_6mths`**: Il numero di richieste di informazioni del credito negli ultimi 6 mesi,`**mths\_since\_last\_delinq`**: Il numero di mesi dall'ultimo pagamento in ritardo, **`open\_acc**`: Il numero di conti aperti, **`pub\_rec**`: Il numero di record pubblici, `**revol\_bal**`: Il saldo totale delle carte di credito revolving, `**revol\_util**`: L'utilizzo del credito sulle carte revolving, `**total\_acc**`: Il numero totale di conti,`**total\_pymnt**`: L'importo totale pagato,`**total\_pymnt\_inv**`: L'importo totale pagato dagli investitori,`**total\_rec\_prncp**`: L'importo totale del capitale ricevuto,`**total\_rec\_int**`: L'importo totale degli interessi ricevuti, `**last\_pymnt\_d**`: La data dell'ultimo pagamento,`**last\_pymnt\_amnt**`: L'importo dell'ultimo pagamento, `**next\_pymnt\_d**`: La data del prossimo pagamento,`**last\_credit\_pull\_d**`: La data dell'ultima richiesta di informazioni del credito, `**repay\_fail**`: Una variabile che rappresenta il fallimento del rimborso (una variabile binaria che indica se un prestito è stato rimborsato o meno).

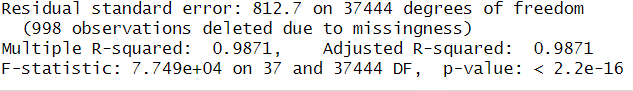
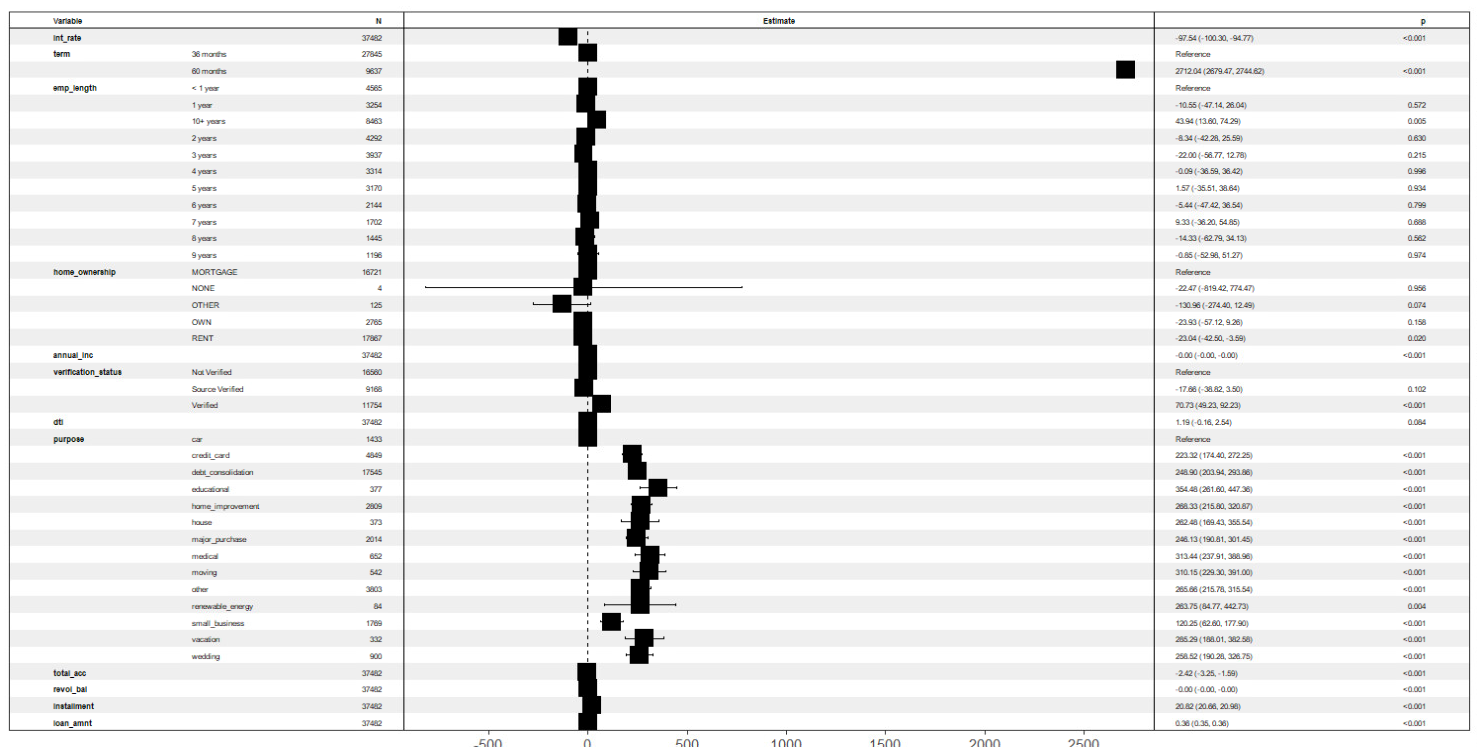
## PRELIMINARE MISSING:

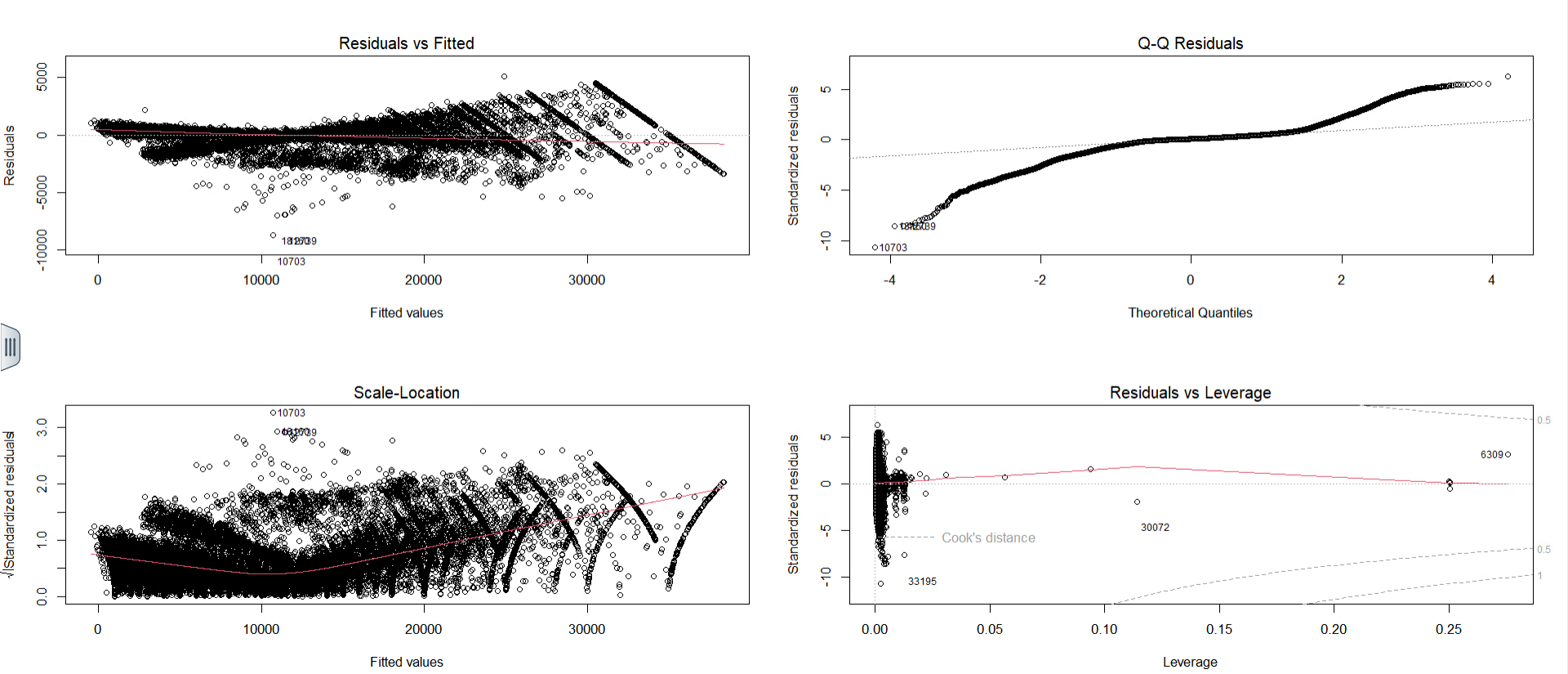


Essendoci una riga che presenta un’unità di missing value, è stato deciso di “omittarlo” perché sembrerebbe essere un errore sistematico di importazione dati ed avendo un dataset molto ampio non influenza in modo significativo l’analisi

## MODELLO INIZIALE:

Attraverso la “conoscenza del dominio” si definiscono le variabili iniziali da studiare, ignorando le variabili che presentano più del 20% di missing values(‘next\_pymnt\_d’,’mths\_since\_last\_delinq’).

****



**Residual vs Fitted :**

* Il modello appare approssimativamente lineare, siccome la LOESS si trova quasi in corrispondenza della linea nera orizzontale.

**QQ-Plot:**

* Gli errori non sono distribuiti in modo normale, possiamo dire che assume una forma leptocurtica.
* Le cause possibili possono essere : Outliers, eccesso di varianza residua (code più pesanti).

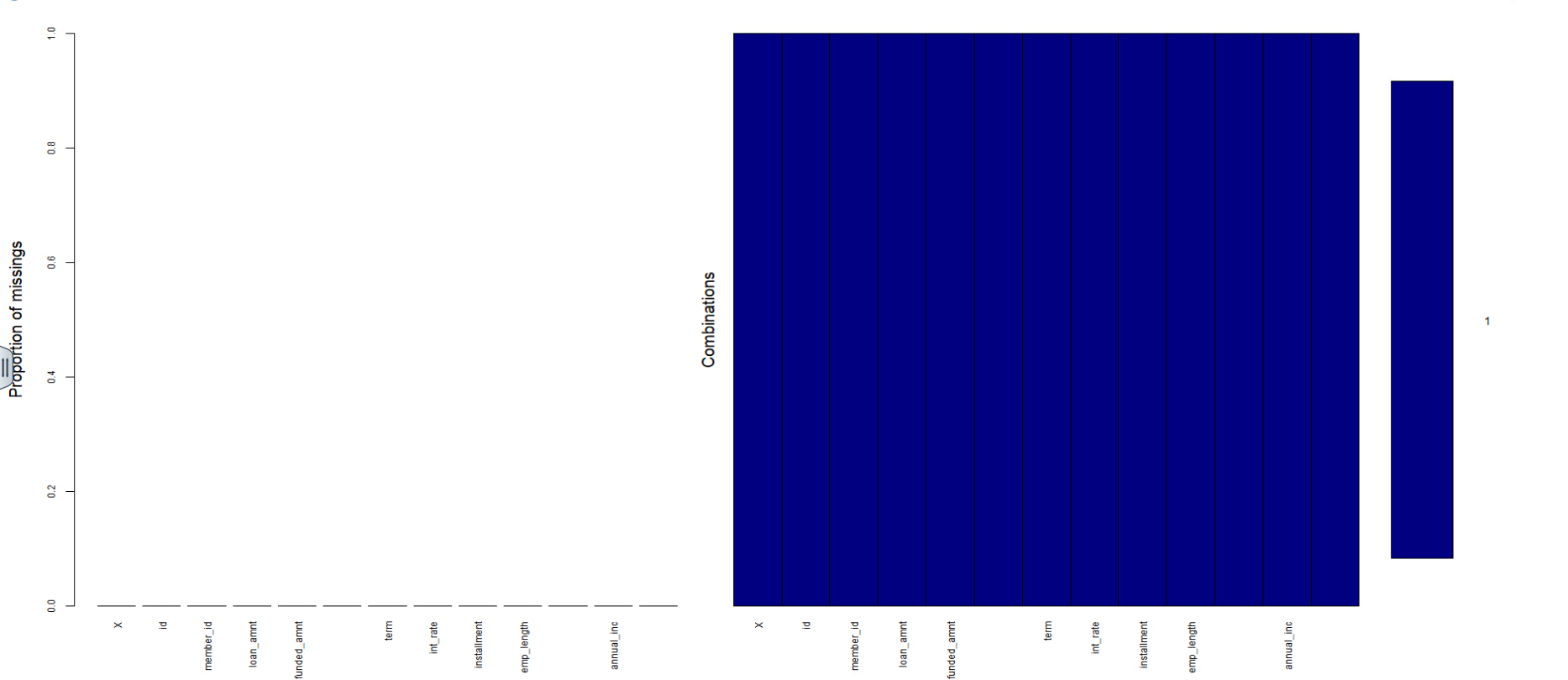
**Scale-Location:**

* Non è presente una retta, si nota una tendenza nei residui e di conseguenza si presenta l’eteroschedasticità
* Problemi possibili : OLS lineari e non distorti, ma non efficienti, quindi standard error sbagliati e statistiche test sbagliate, correlazione tra i residui e i valori interpolati

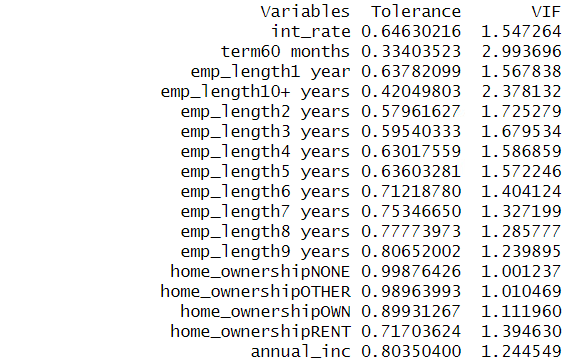
**Residuals vs Leverage:**

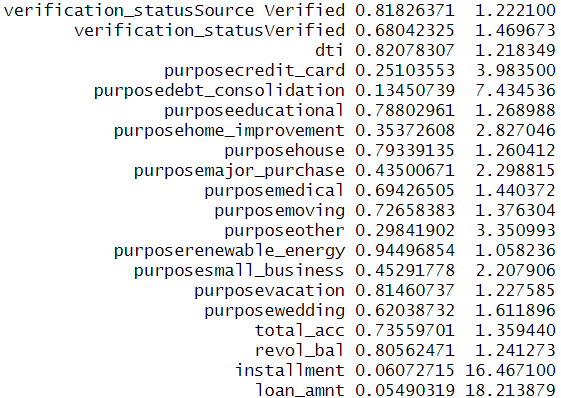
* Il grafico sembra non presentare punti influenti, è possibile notare alcuni punti di alto leverage e alcuni outlier
* Può provocare violazione delle Assunzioni del Modello: se le assunzioni del modello, come l'omoschedasticità o la normalità dei residui, sono violate, ciò potrebbe portare a punti di alto leverage e outlier.

## DIAGNOSTICA MISSING VALUES:

Si utilizzano le colonne del modello iniziale e si applica FCS con metodo “pmm” per ottenere un dataset con dati imputati completi

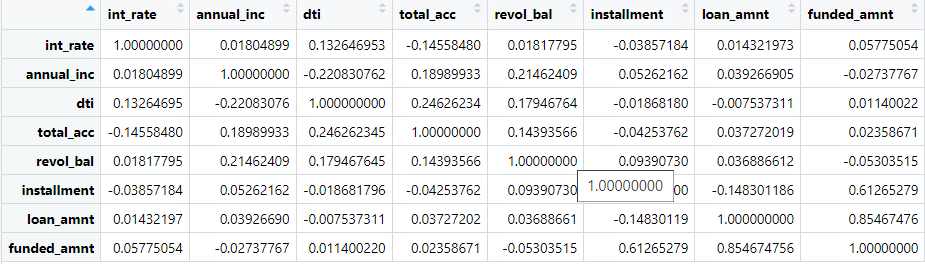
## **COLLINEARITA’**



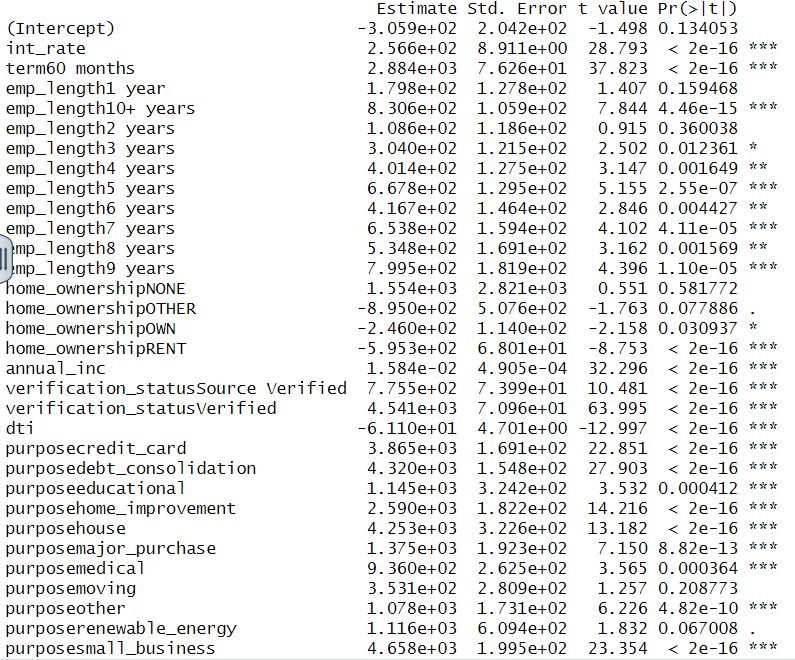


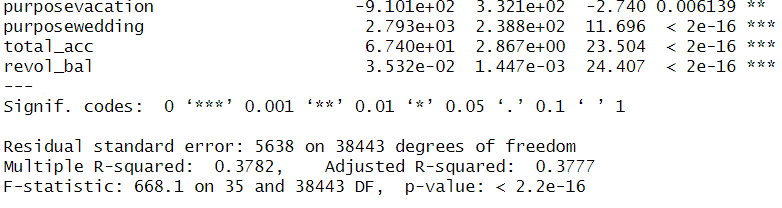
Nel modello si nota la presenza di due variabili collineari, si rimuove la variabile **‘installment’**. Osservando la variabile **‘loan\_amnt’** si nota la sua corrispondenza alla variabile dipendente in termini di valori e di significato. Quindi si rimuovono le variabili sopracitate dal modello.

E’ possibile confermare la collinearità anche attraverso la matrice di correlazione parziale:

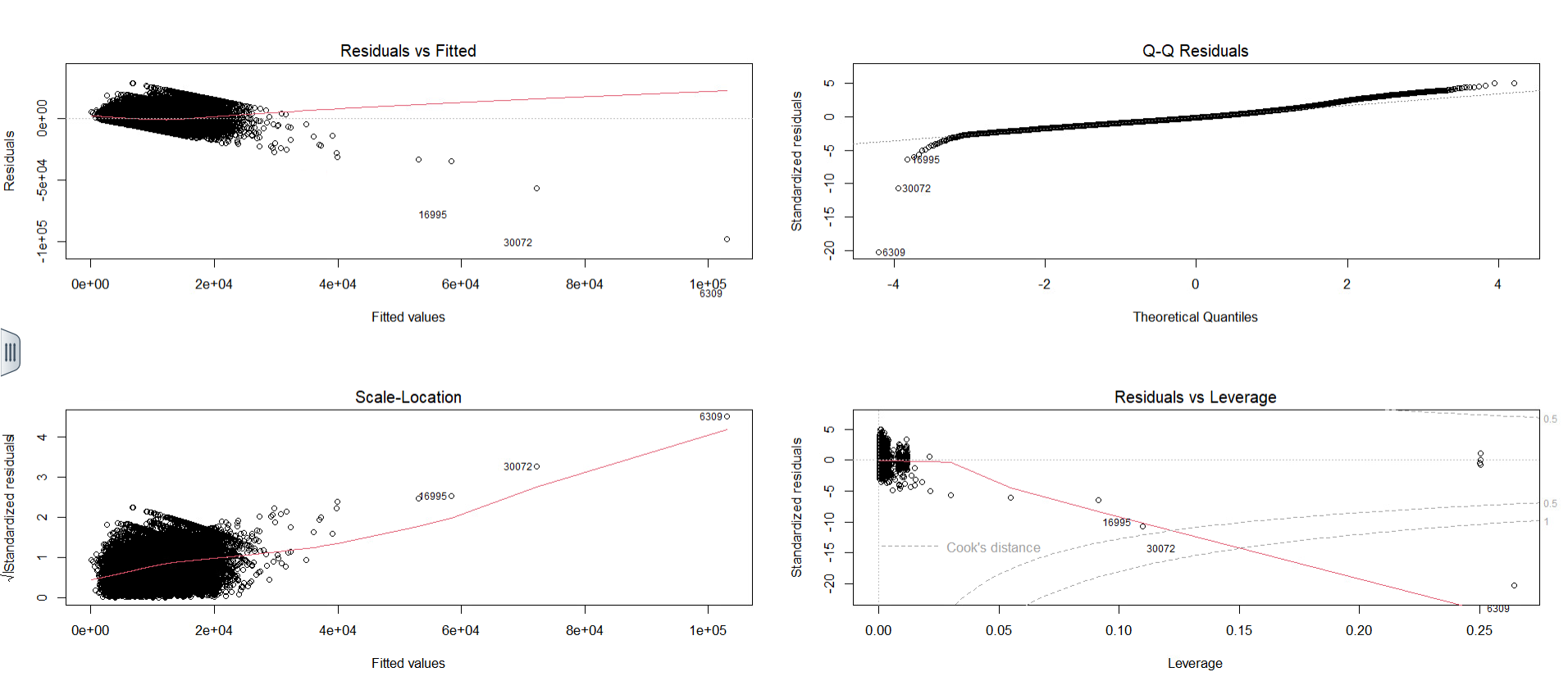


Di conseguenza il modello diventa:



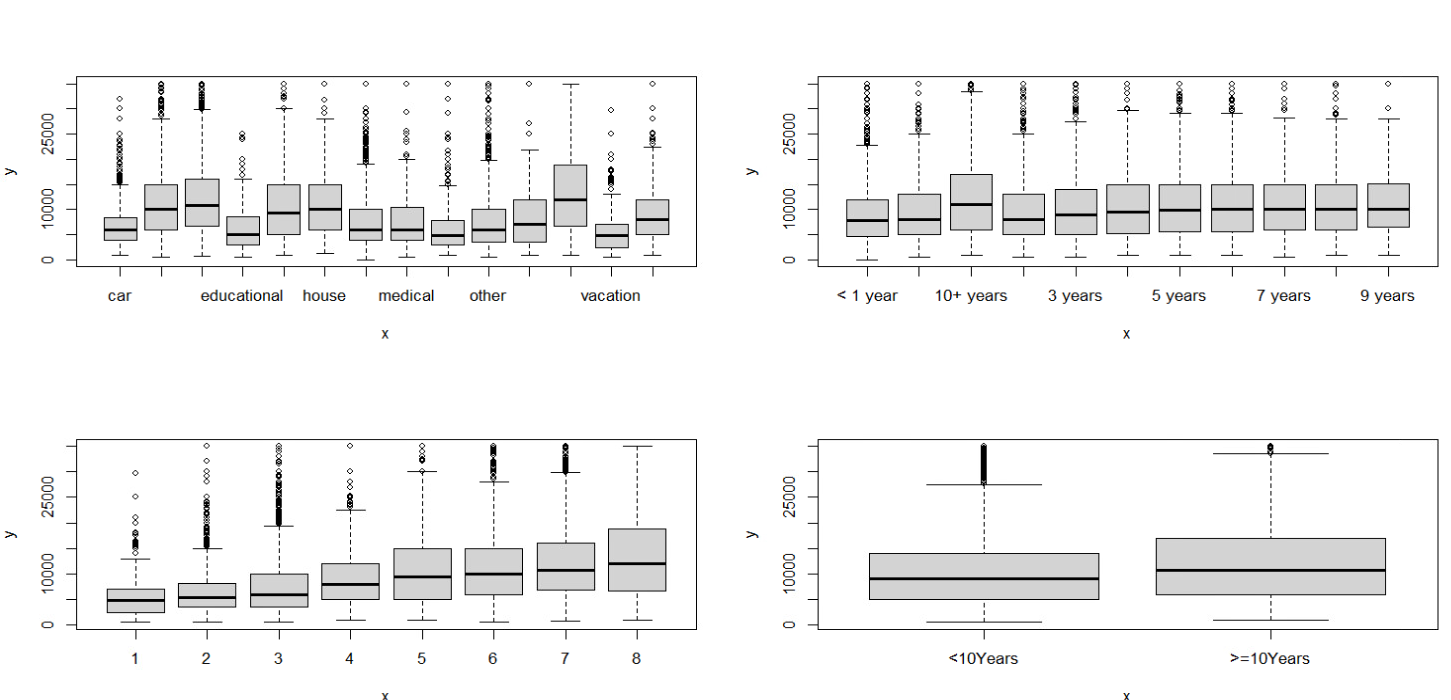


E i grafici diventano :



E’ possibile osservare che il qq-plot non presenta la stessa forma del precedente, infatti sono presenti i punti influenti nel grafico Residuals vs Leverage. E’ presente una tendenza che provoca eteroschedasticità. Dopo l’eliminazione delle variabili collineari,la relazione fra le variabili target e indipendenti è cambiata,si nota infatti un abbassamento di R^2 e delle statistiche F.

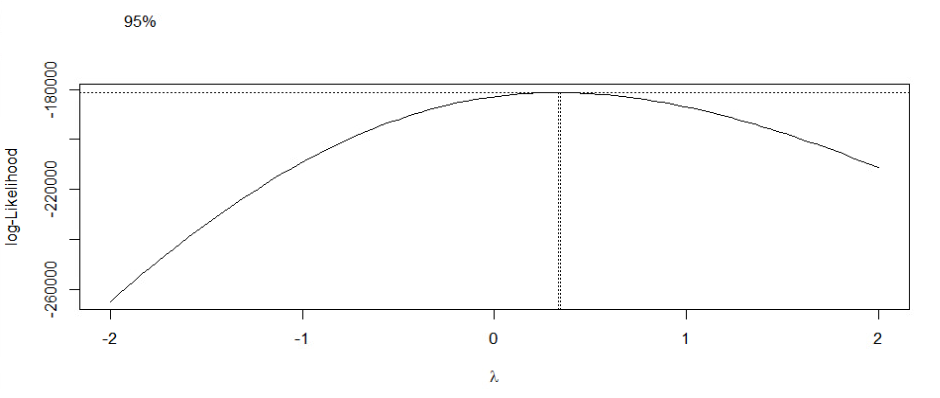
## OPTIMAL GROUPING



Si nota la presenza di un numero elevato di livelli della variabile **Purpose** e la presenza di alcune variabili non significative. Quindi si prova a eseguire l’optimal grouping ottenendo un buon raggruppamento.

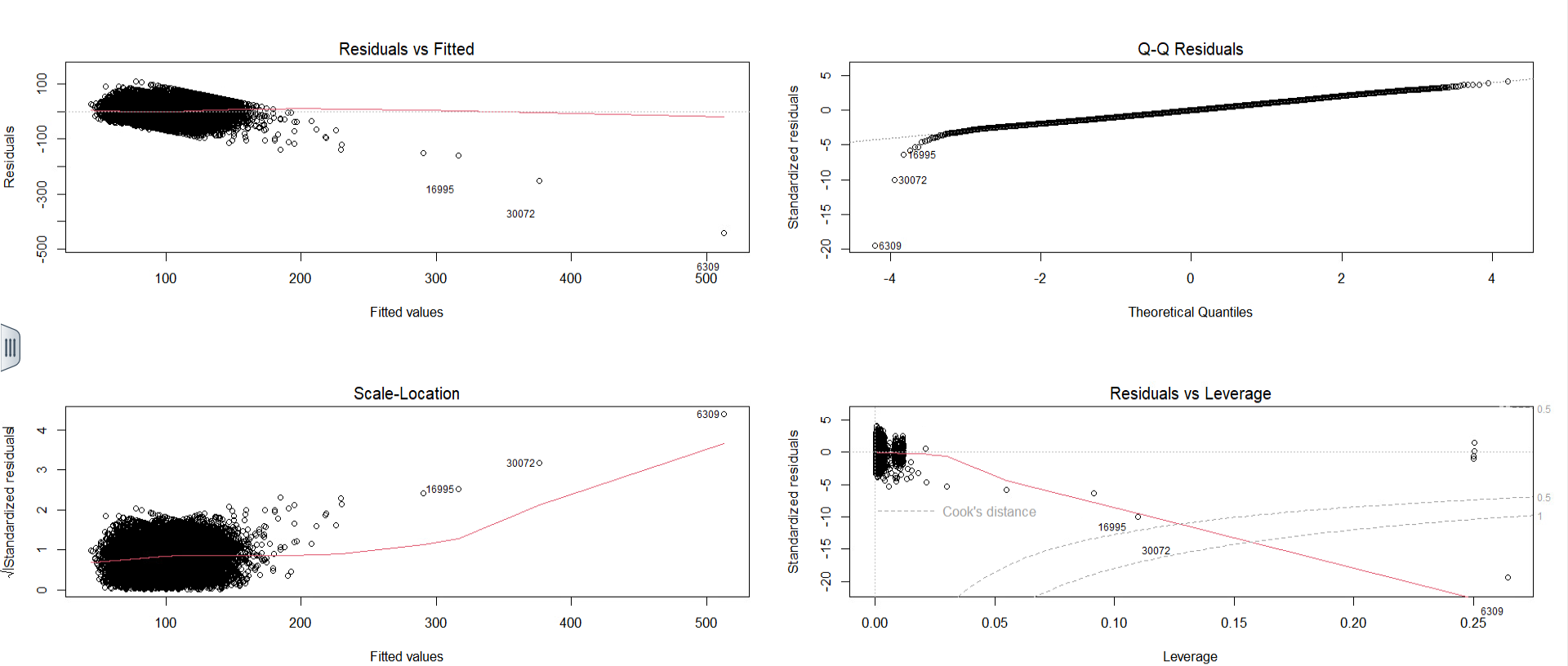
Per quanto riguarda la variabile **emp\_length**, l’optimal grouping non ha un buon effetto, quindi semplicemente si esegue un raggruppamento a due livelli manualmente 10+ Years come “10+Years” e tutte le altre come “<10 Years”.

## BOX COX



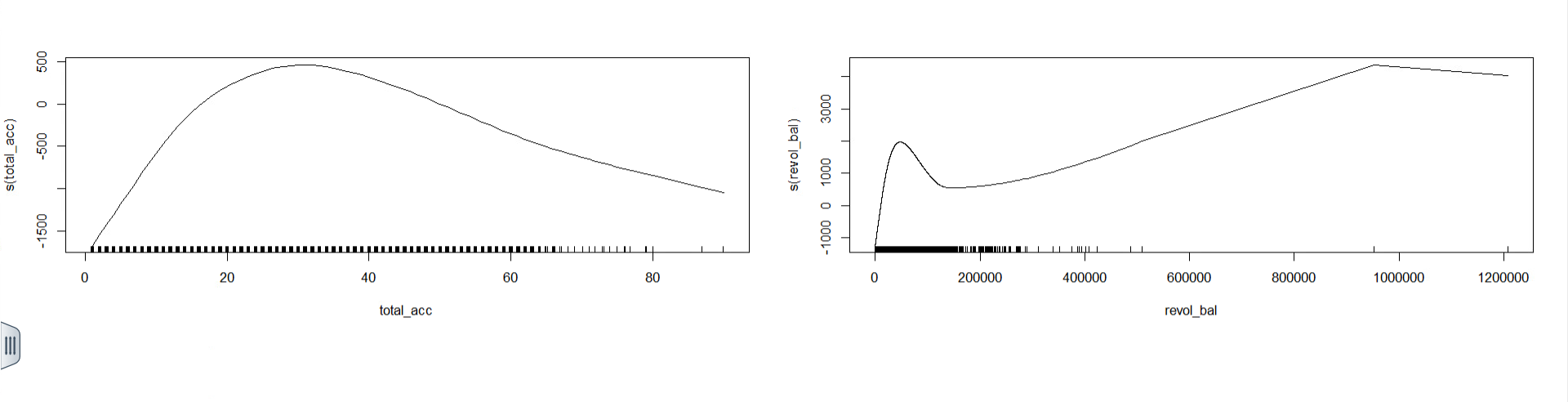
lambda = 0.3434343 quindi si approssima a 0.5

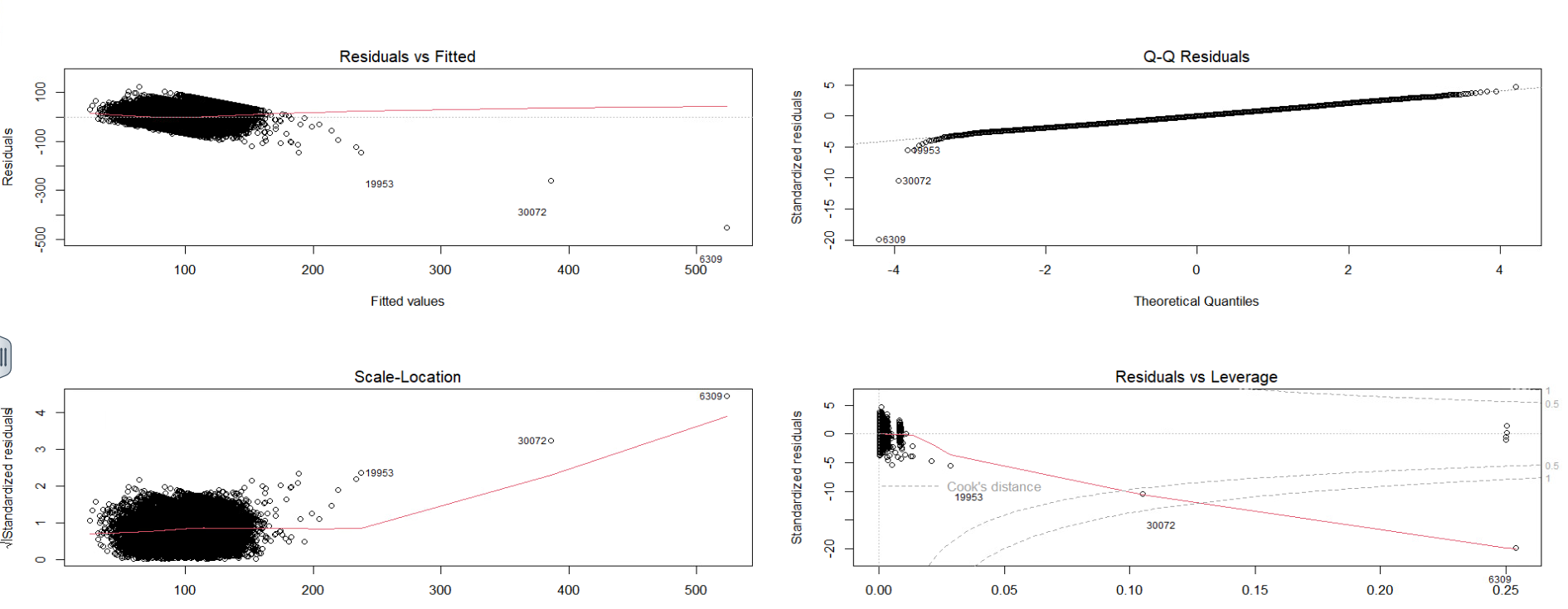
### TRASFORMAZIONE BOX COX (GRAFICI)



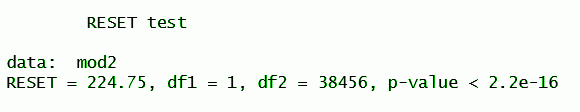
Si vede un miglioramento nella linearità del modello.

### GAM+BOX COX



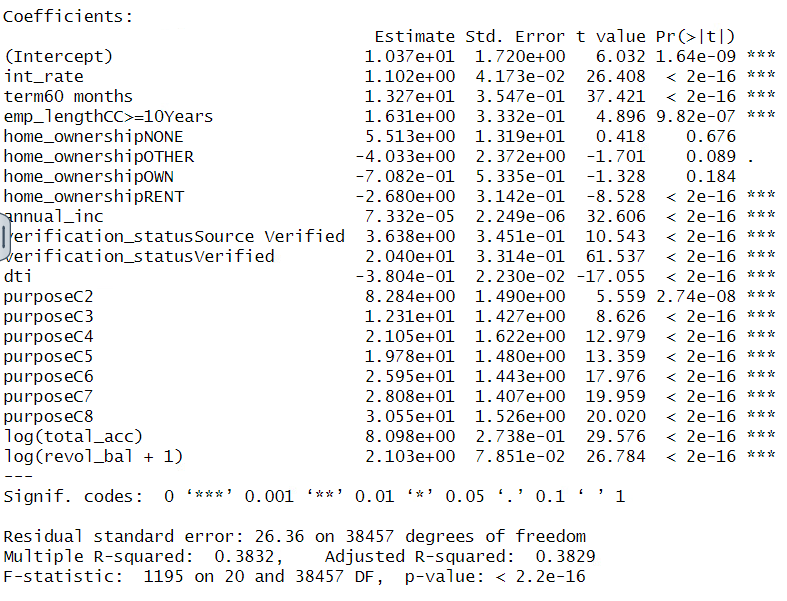


Non si nota nessun miglioramento sostanziale della linearità dopo aver eseguito BOX-COX e GAM insieme.

****

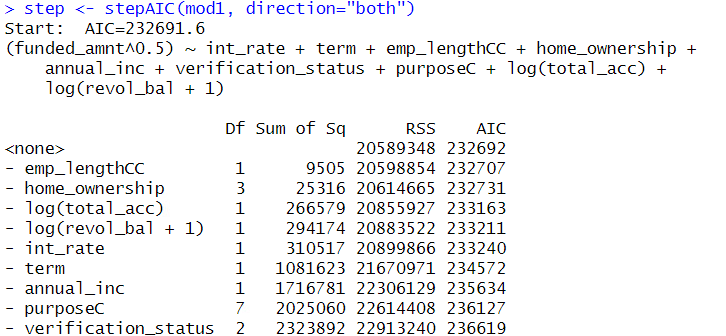
Infatti con il Ramsey Test si rifiuta l’ipotesi nulla , quindi è possibile aggiungere altri parametri non lineari per migliorare il modello.

POST BOX COX E GAM MODELLO:



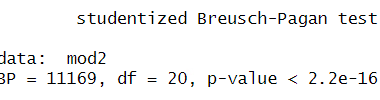
Si nota che i coefficienti cambiano, e gli standard error diminuiscono .

## AIC

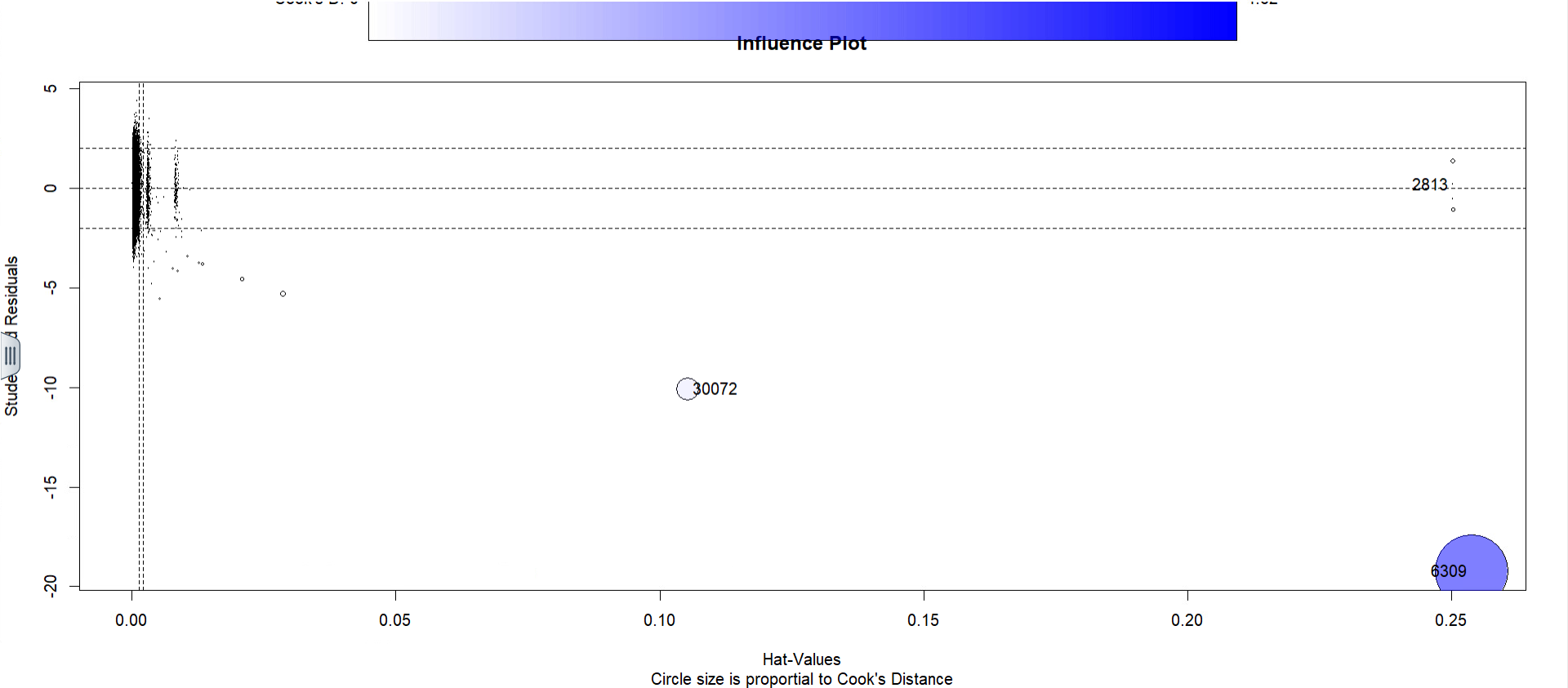


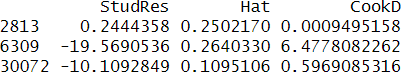
Eseguendo il model selection, AIC suggerisce di non rimuovere nessuna variabile del modello, e che quella attuale è la migliore in termini parsimoniosi e generalizzazione di nuovi dati(idem per SBC)

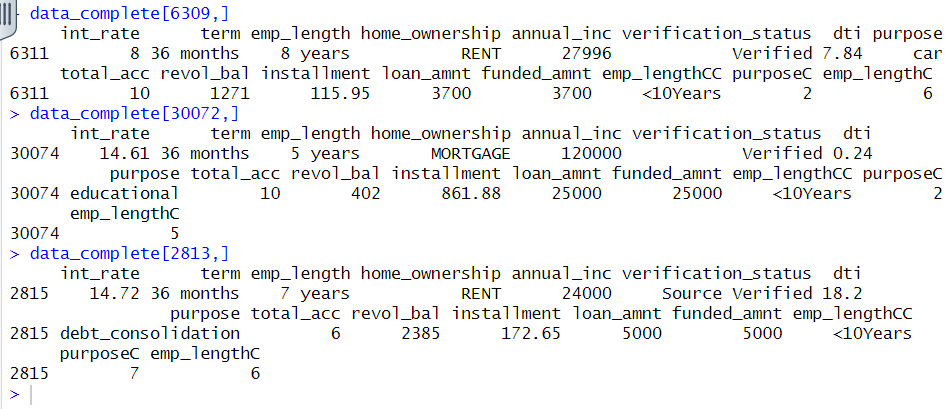
## ETEROSCHEDASTICITA’

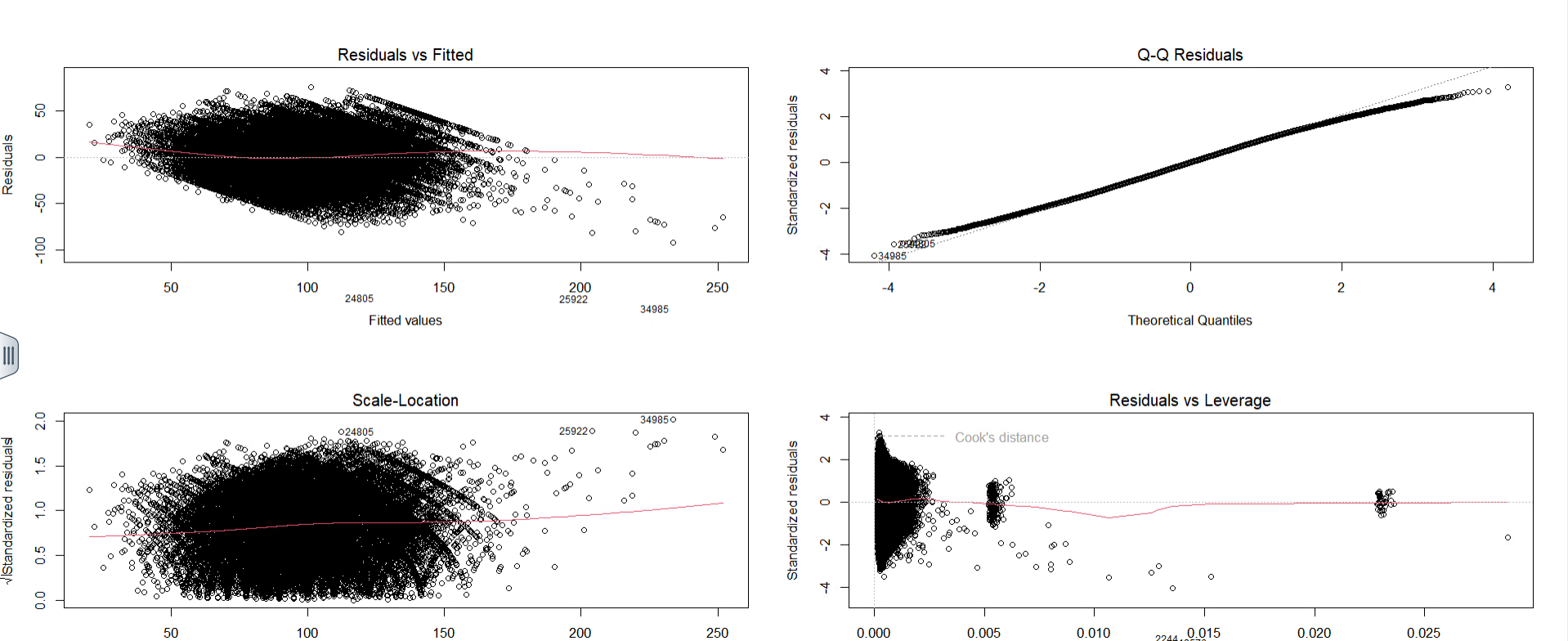


Il test conferma la presenza di eteroschedasticità.





E’ possibile osservare la presenza di un punto fortemente influente da CooK

OUTLIER(POST ELIMINAZIONE PUNTI INFLUENTI COOK E DFITTS DEL MODELLO GAM E BOX COX )

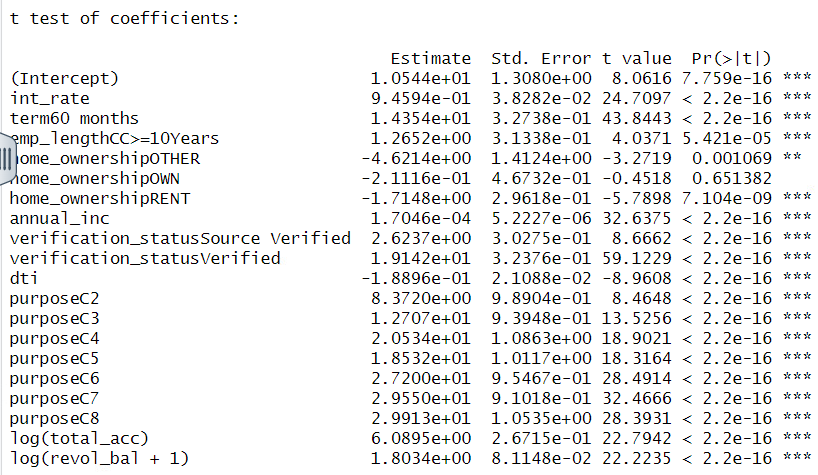
## 

## 

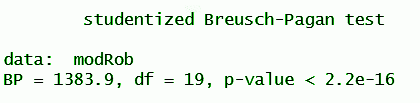
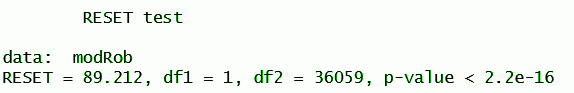
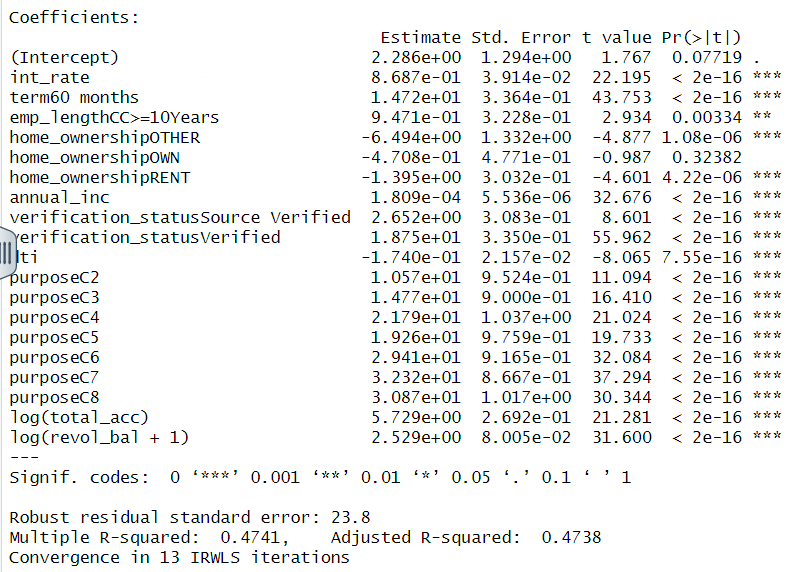
## 

## 

## STANDARD ERRORI CORRETTI CON WHITE

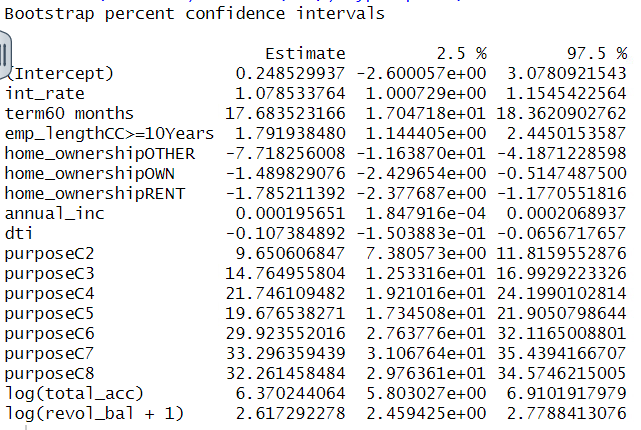
Si nota che c’è una variabile che diventa significativa, e gli standard error sono più alti, ma l’importante è che sono giusti per l’inferenza .

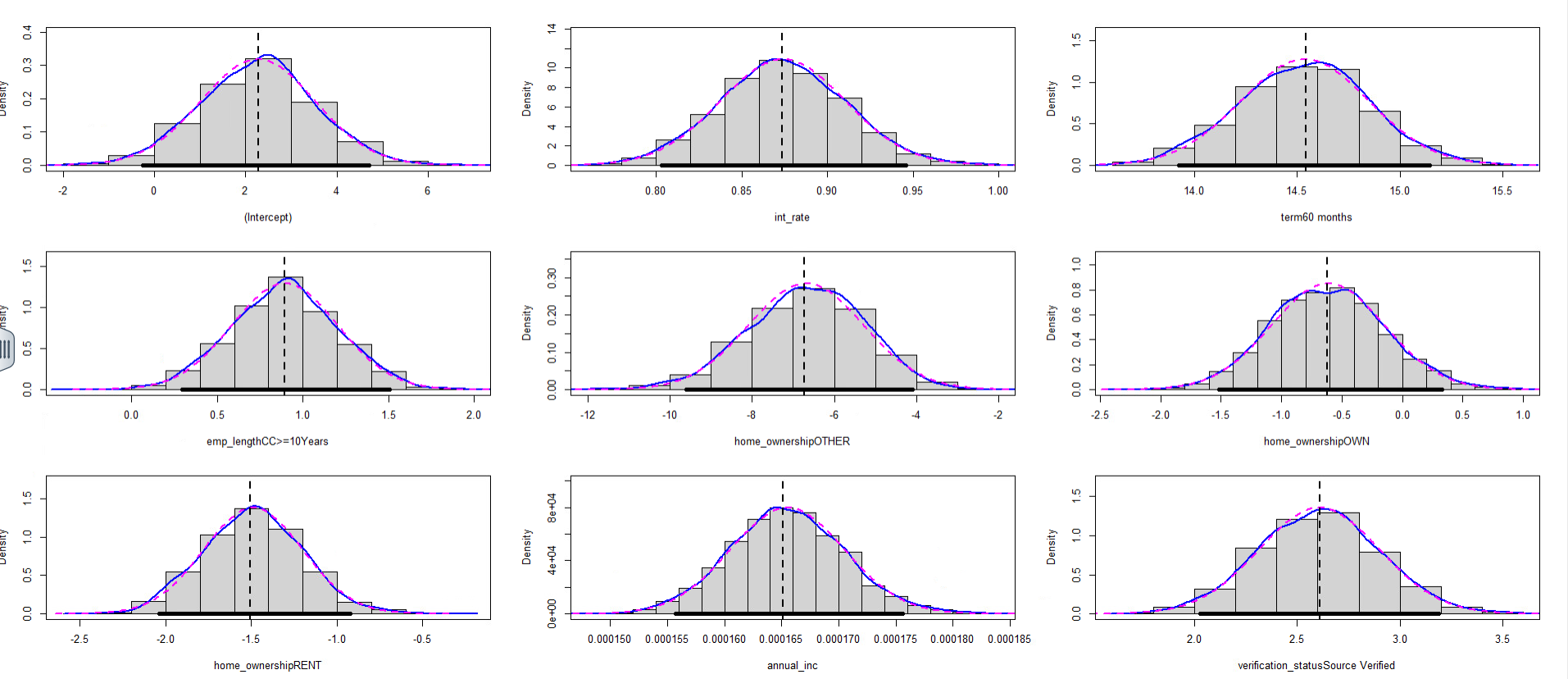
## ROBUST REGRESSION

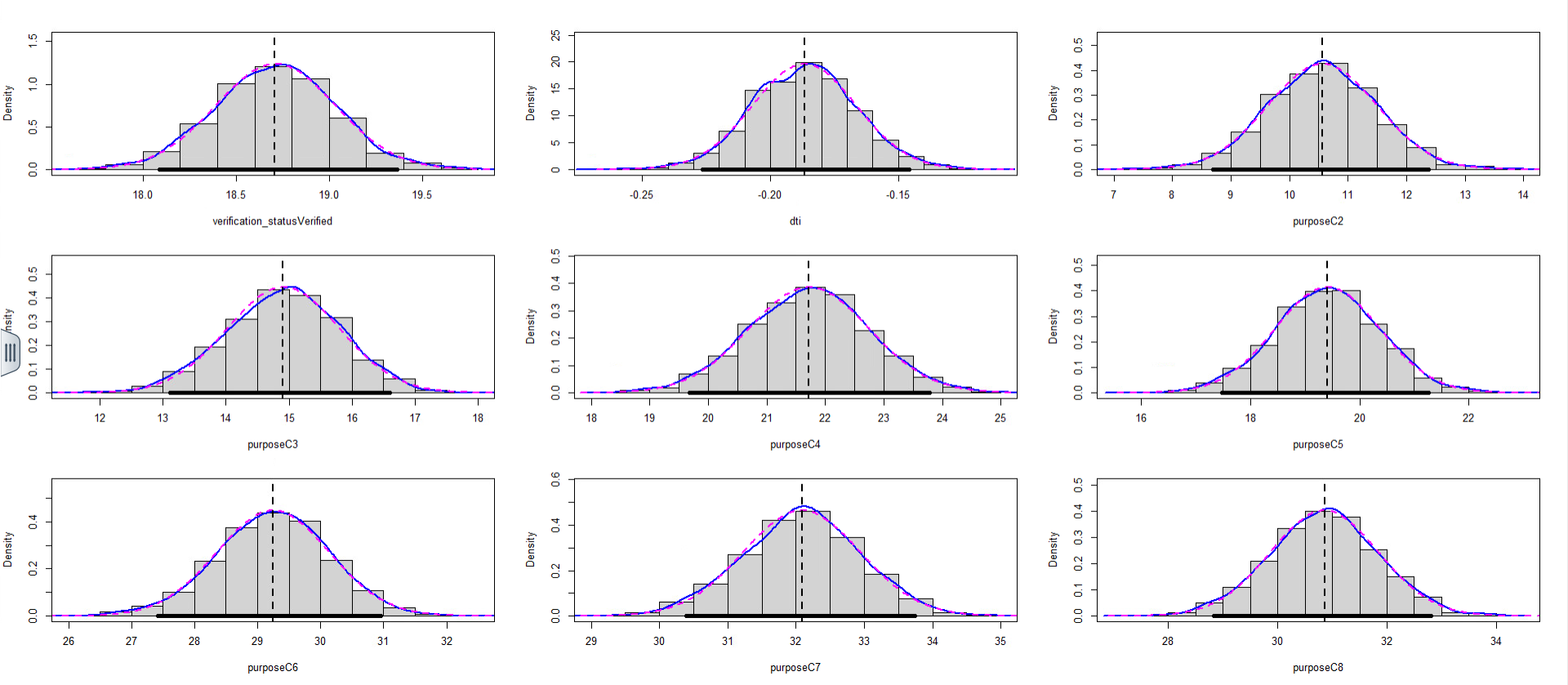


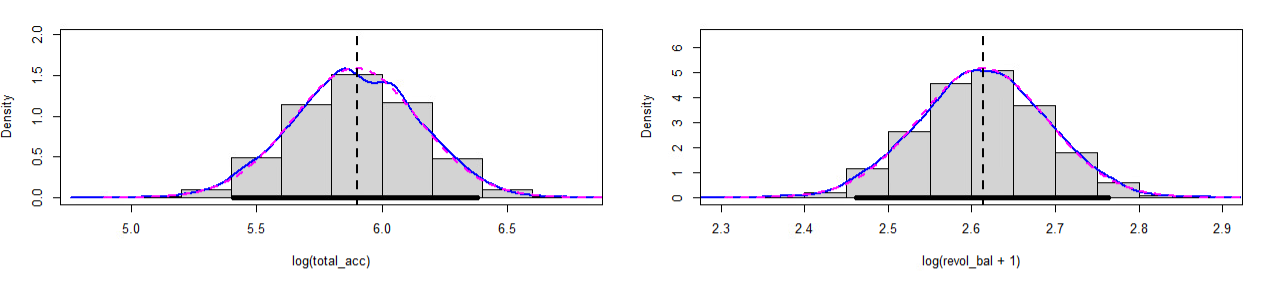
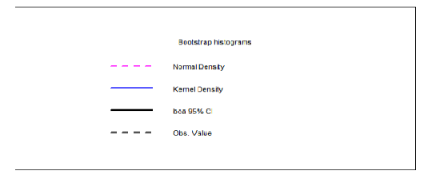
## 

## **BOOSTRAP :**

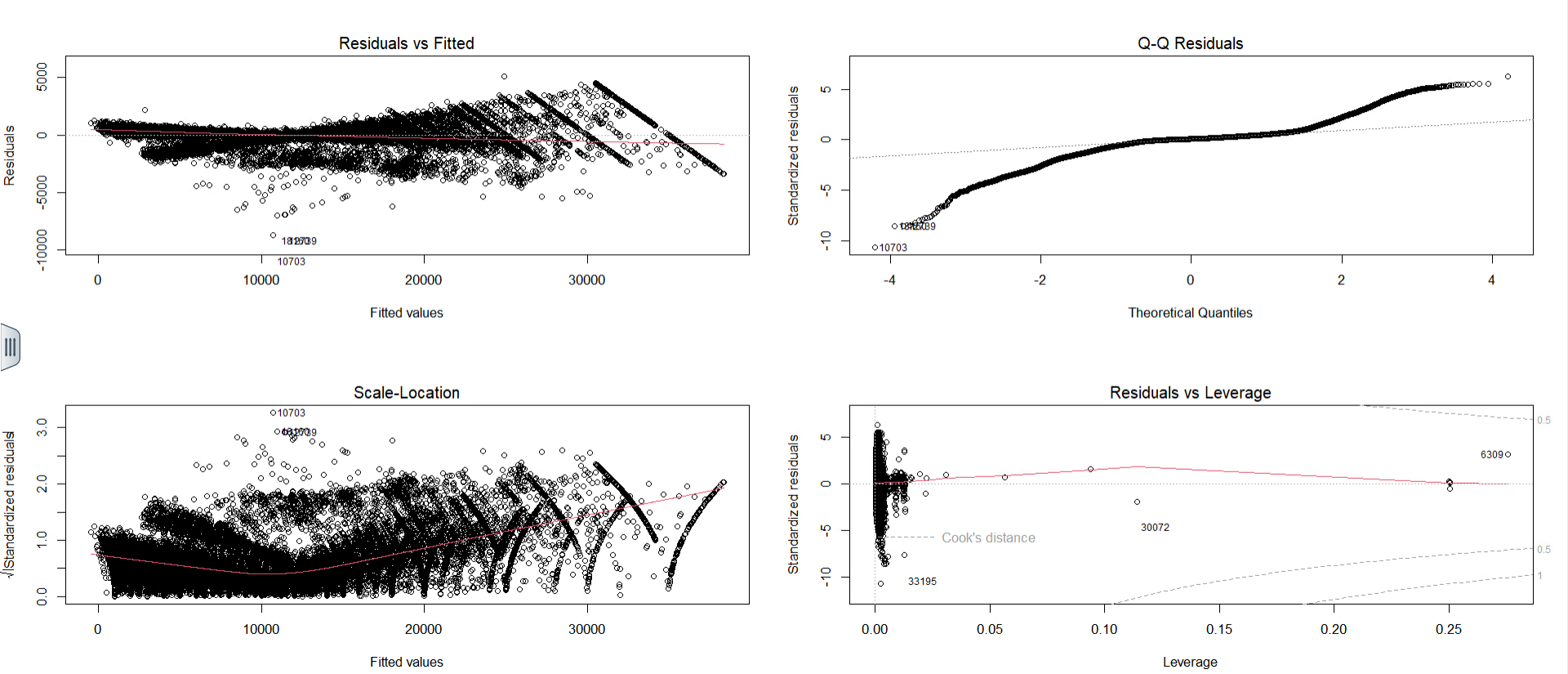


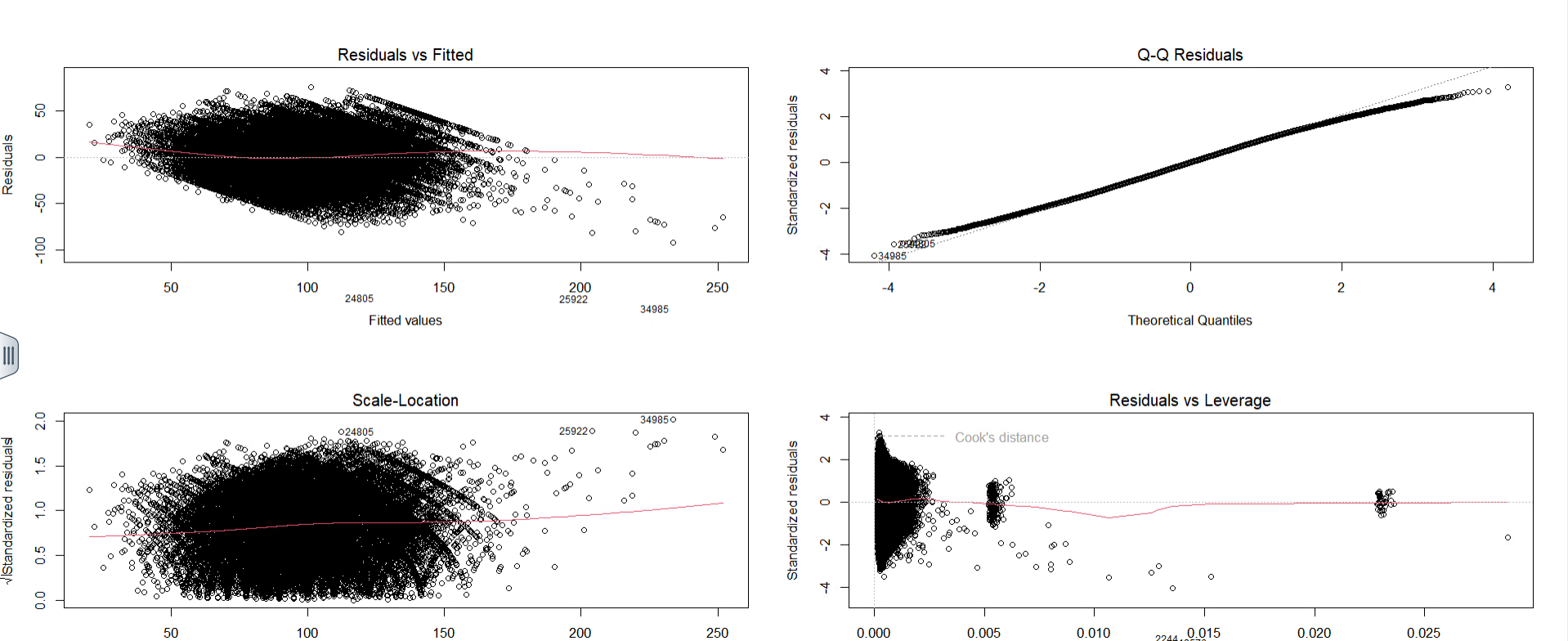




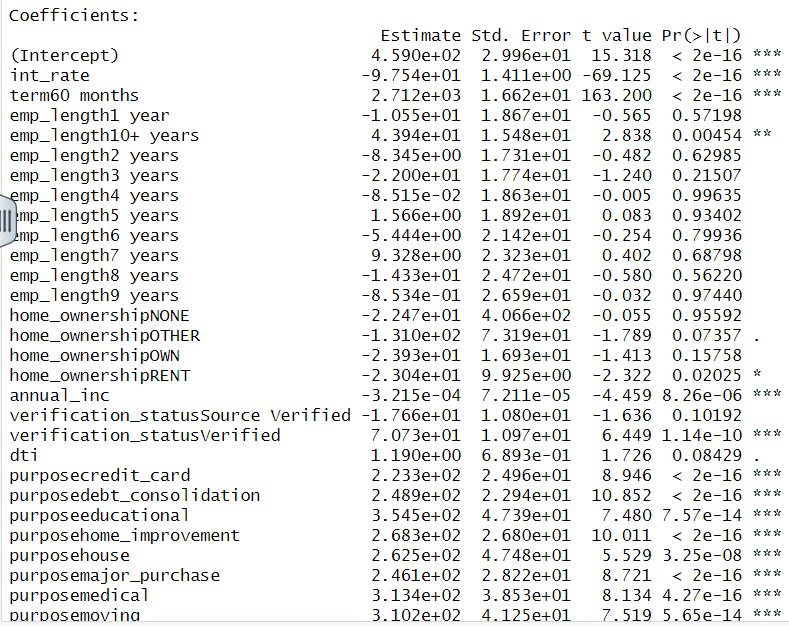
  
Tutte le variabili sono significative , perché non includono lo 0 nell’intervallo di confidenza, tranne l’intercetta.

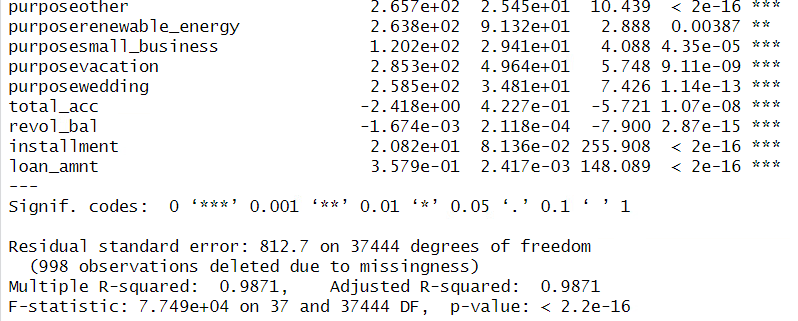
## CONFRONTO



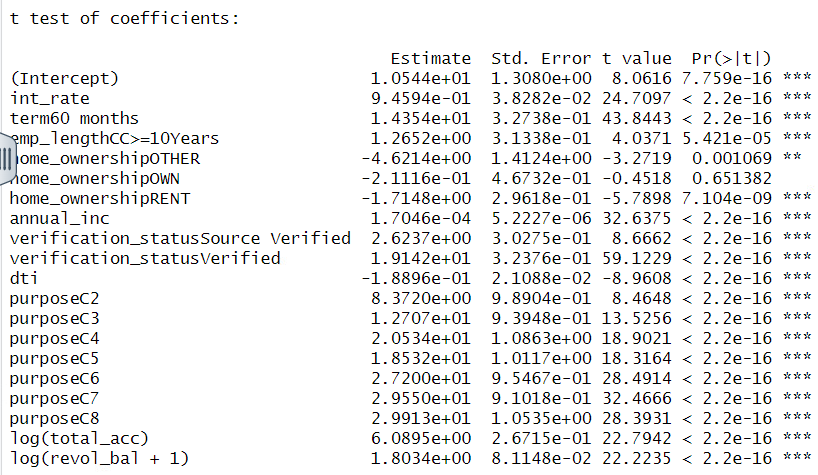


## 





## 



## 

## 

## GLOBAL TEST FOR MODEL ASSUMPTIONS

## 

## 

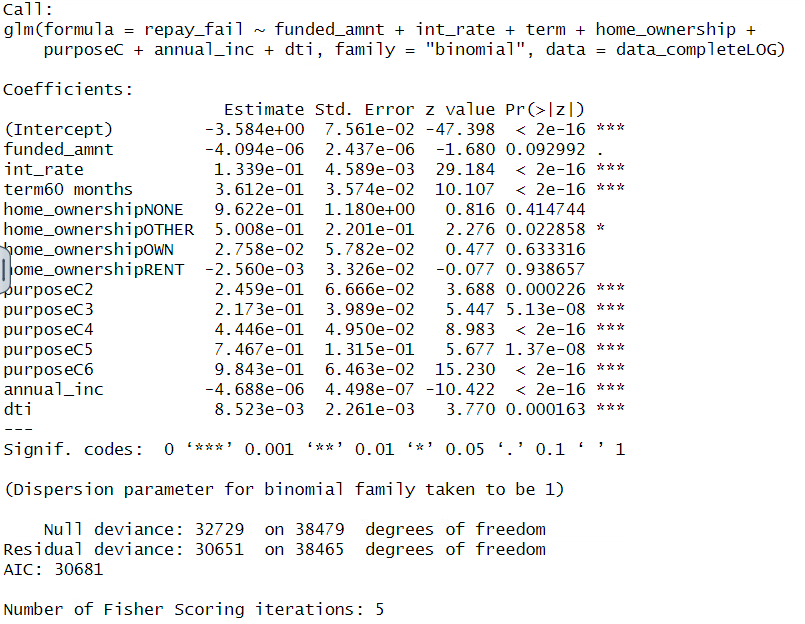
## 

## MODELLO LOGISTICO

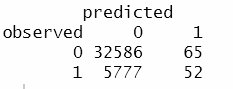
Variabile target dicotomica:

0 = è riuscito a pagare il debito

1 = non è riuscito a pagare il debito







Si confrontano i valore predetti e valori osservati , e si misura la precisione delle previsioni.

**ACCURACY DELLA PREVISIONE DEL MODELLO PARI AL 84%**

**ODDS RATIO:**

