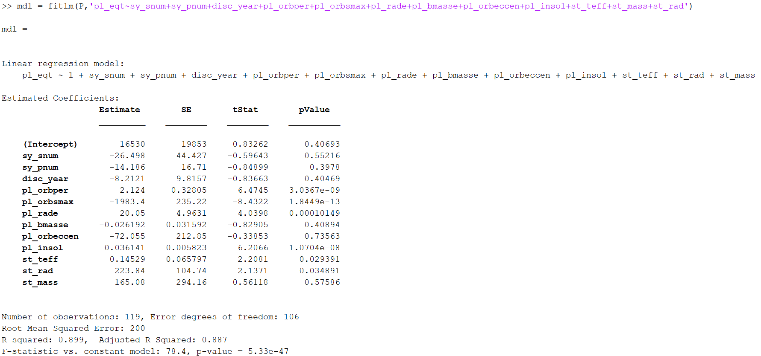
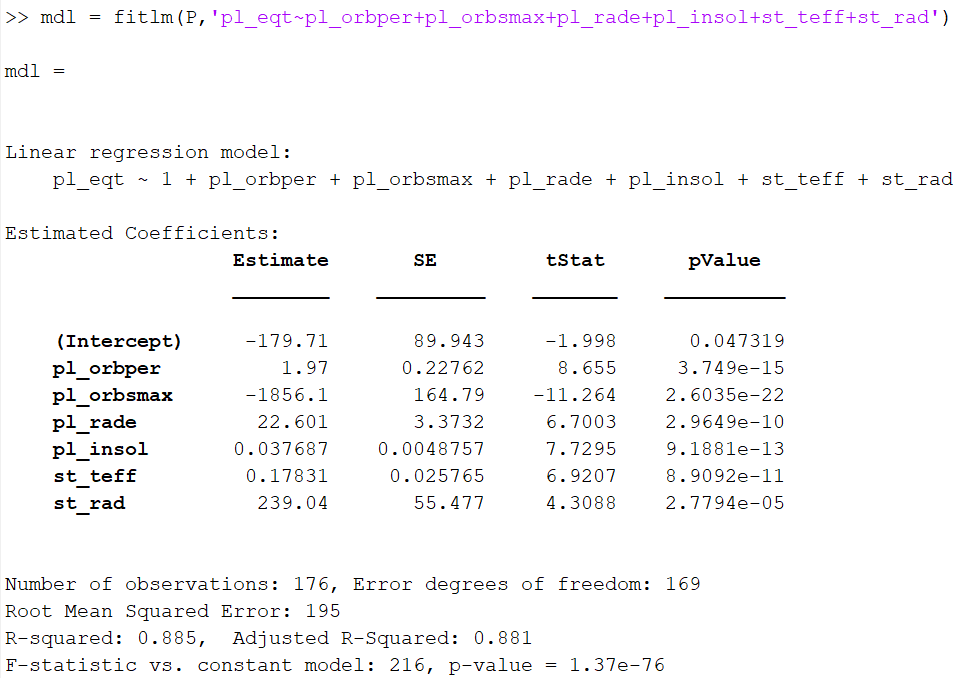
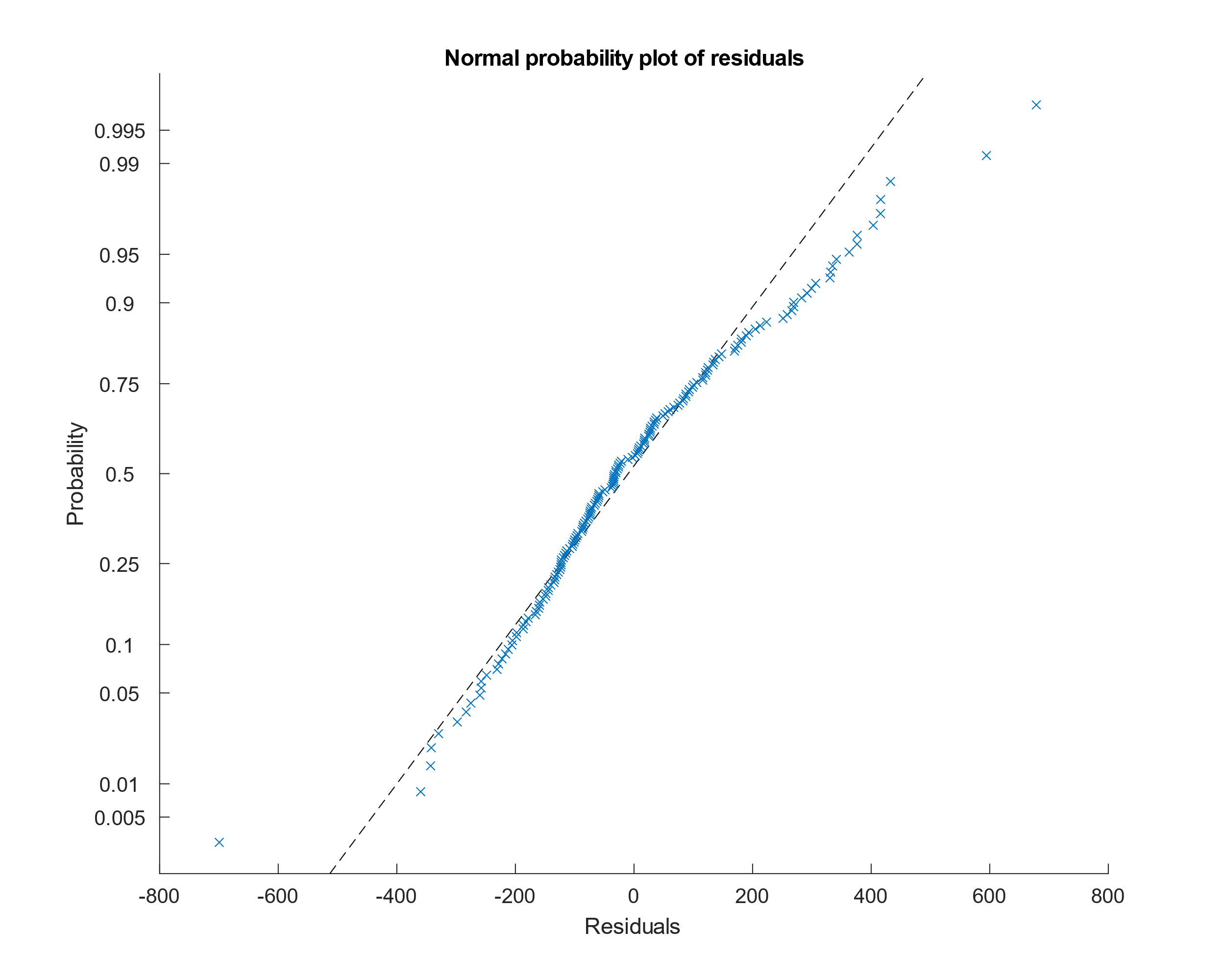
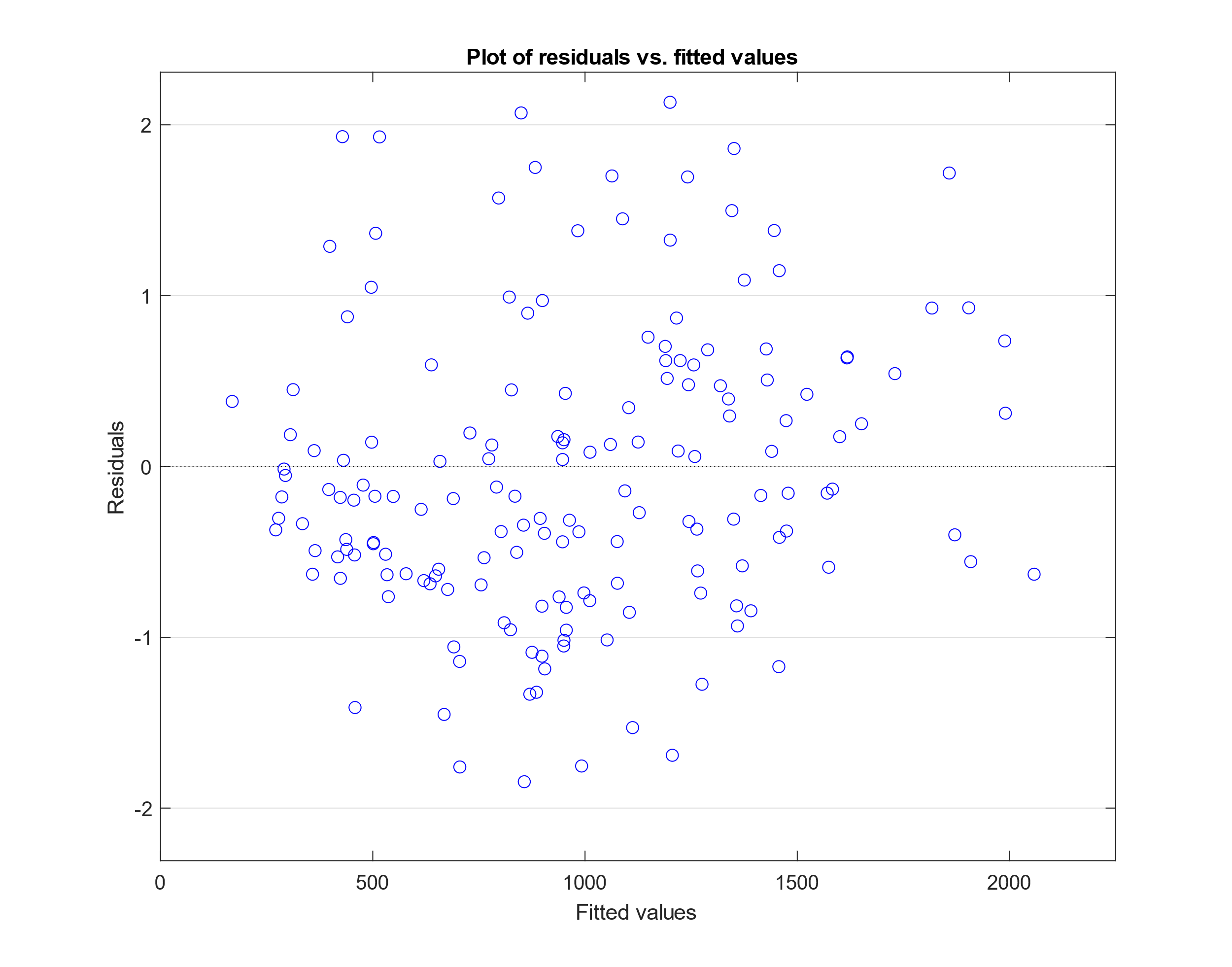
Obbiettivo della nostra regressione sarà quello di determinare un modello per la predizione della temperatura superificiale del pianeta, in base a predittori basati su parametri del sistema planetario d’interesse. Per iniziare sono stati inclusi tutti i predittori che potessere avere una qualche correlazione con la temperatura del pianeta, per questo sono stati esclusi da subito i parametri sulla posizione del pianeta nel cielo.

Procediamo quindi a ed effettuare il fit lineare multiplo e analizziamo i risultati.

Il primo modello che include la totalità dei predittori ha dei problemi evidenti, tutti i predittori categorici riguardanti il metodo di scoperta sono troppo poco numerosi per dare risultati significativi e tutti i loro valori sono N.A. . Procediamo quindi a rimuovere il predittore categorico “disc\_year” ed effettuiamo nuovamente il fit. Il secondo modello è decisamente migliorato ma notiamo che i pValue per molti predittori sono molto alti. I pValue forniti da MATLAB sono infatti relativi a questo test:

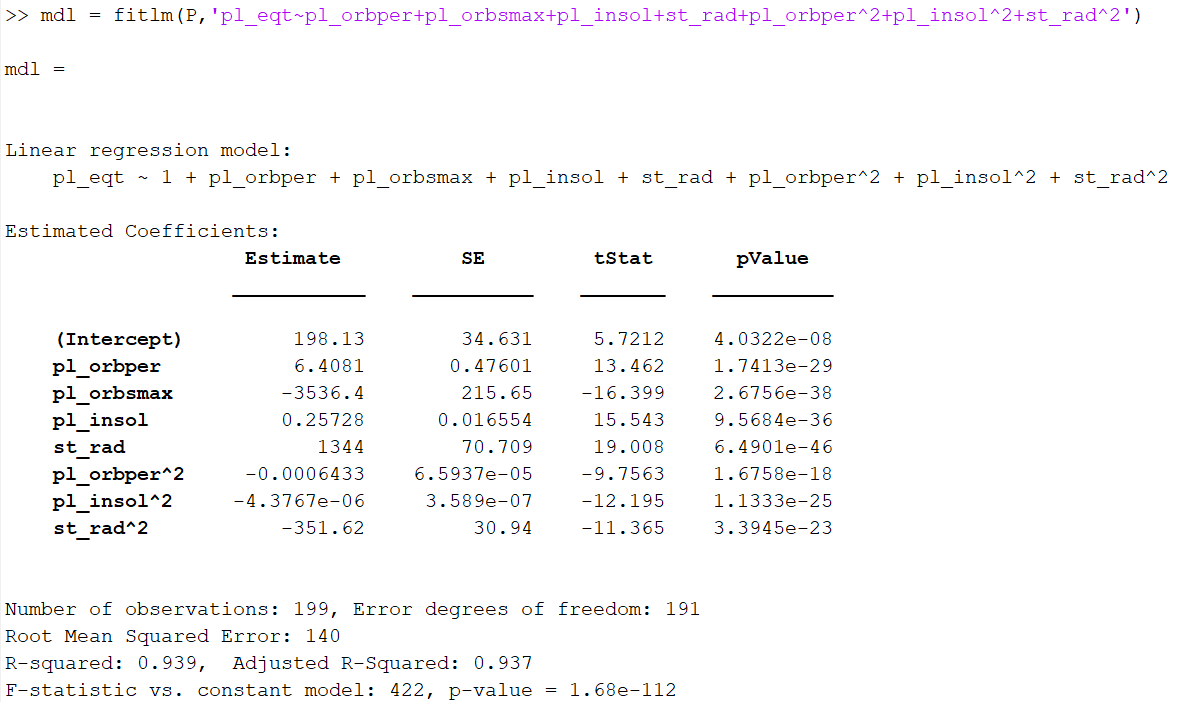
Procediamo quindi iterativamente a rimuovere il predittore con i pValue più alto , il risultato finale è una serie di predittori con pValue molto bassi . Notevolmente con questo processo abbiamo anche rimosso il predittore riferito al numero di stelle, confermando ciò che nel test di ipotesi sembrava probabile. L’unico pValue alto è quello dell’intercetta, vediamo infatti che un test al livello di significatività ci porterebbe a rifiutare per molto poco, non essendo soddisfatti di ciò lasciamo aperta la possibilità che l’intercetta sia nulla.

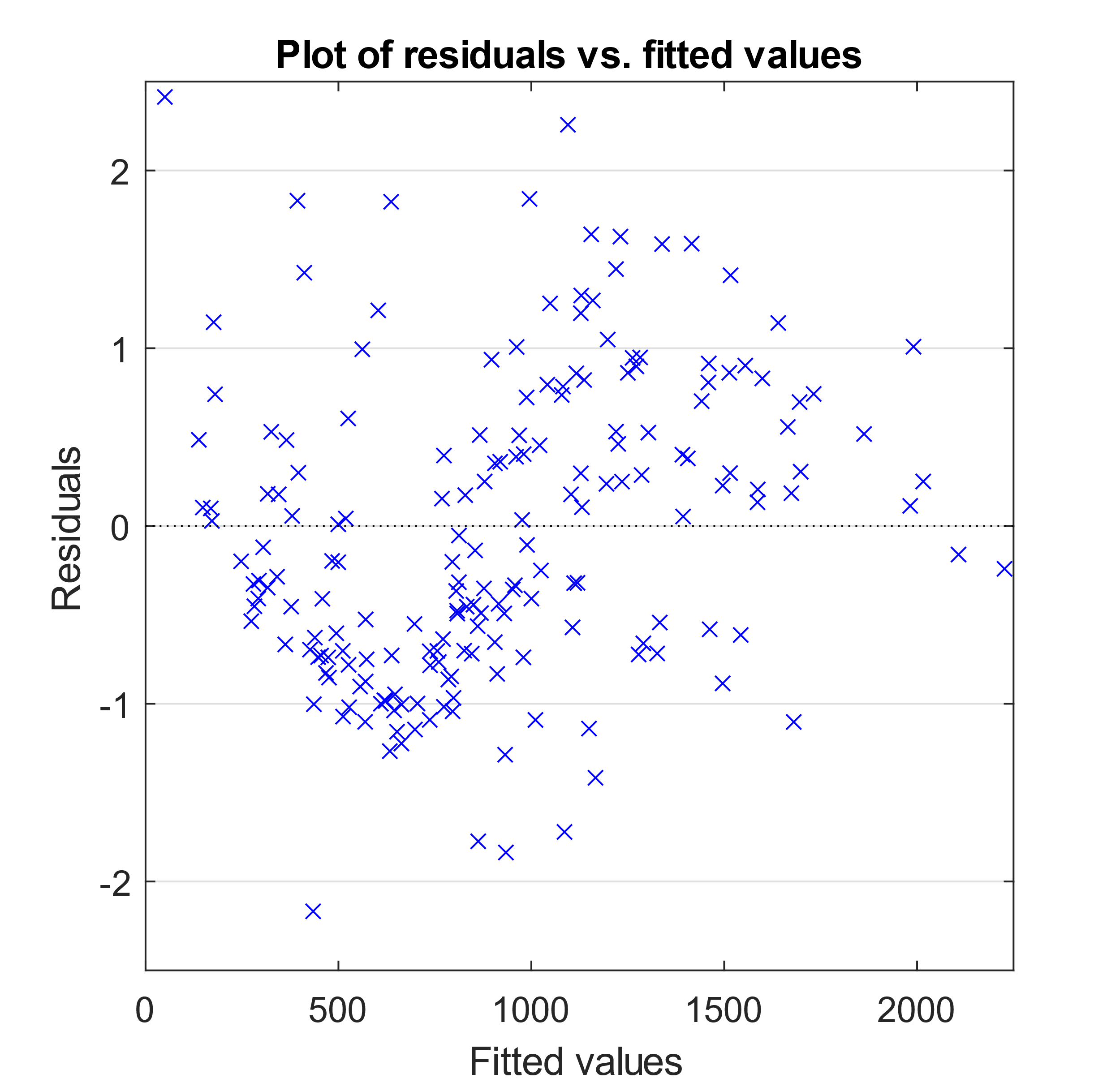
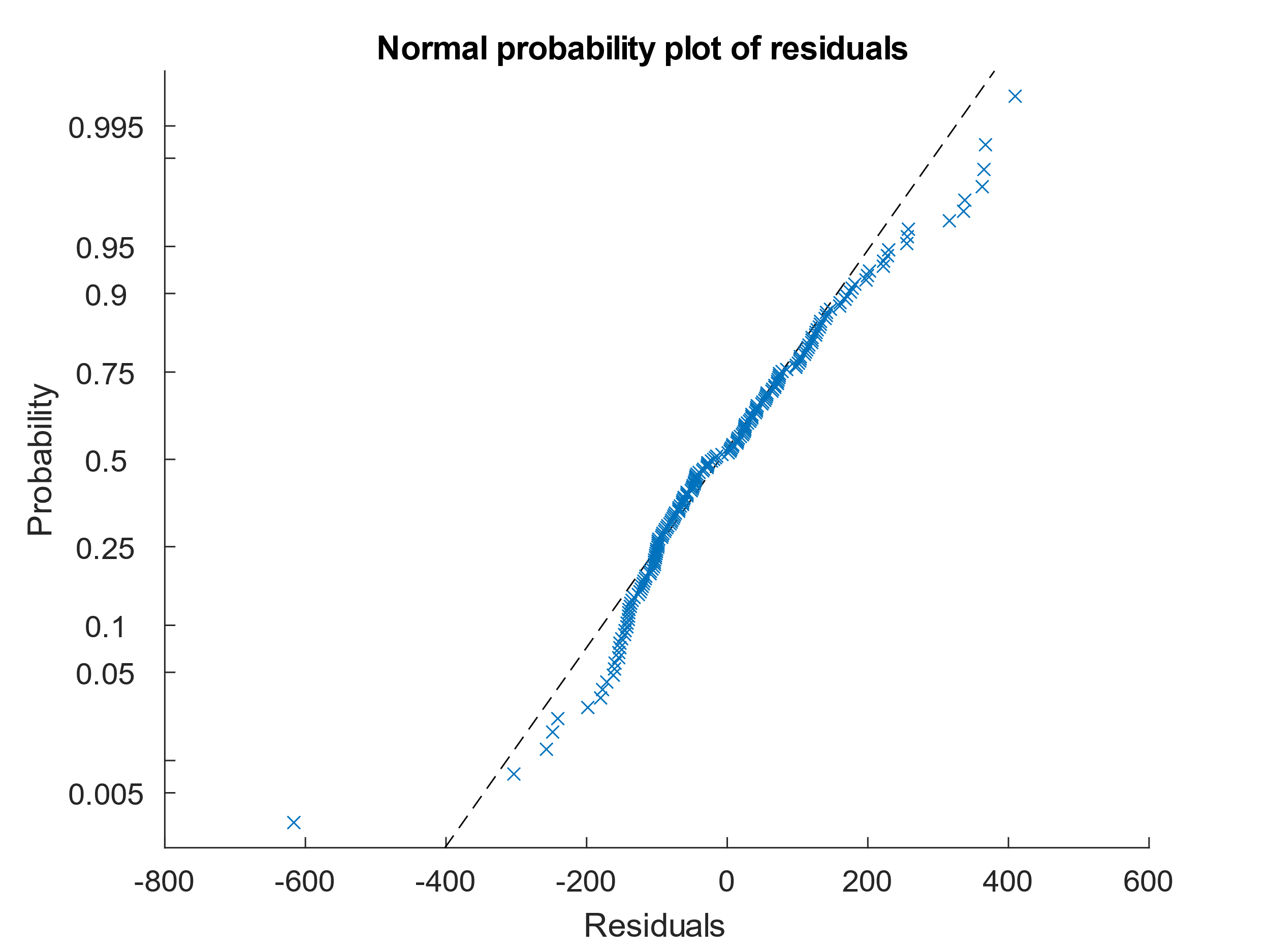
Il coefficiente di determinazione non è variato molto dal primo modello e al momento si attesta ad un buon valore di il numero di predittori è molto basso rispetto al numero di osservazioni quindi siamo sicuri non aver sforato nell’overfitting, ciò è confermato dal valore sempre buono di

Analizziamo quindi la distribuzione dei residui.

Sebbene il grafico dei residui sui fitted values mostri una omoschedasticità al limite dell’accettabile, il QQ-plot dei residui ci obbliga a rivedere il modello.

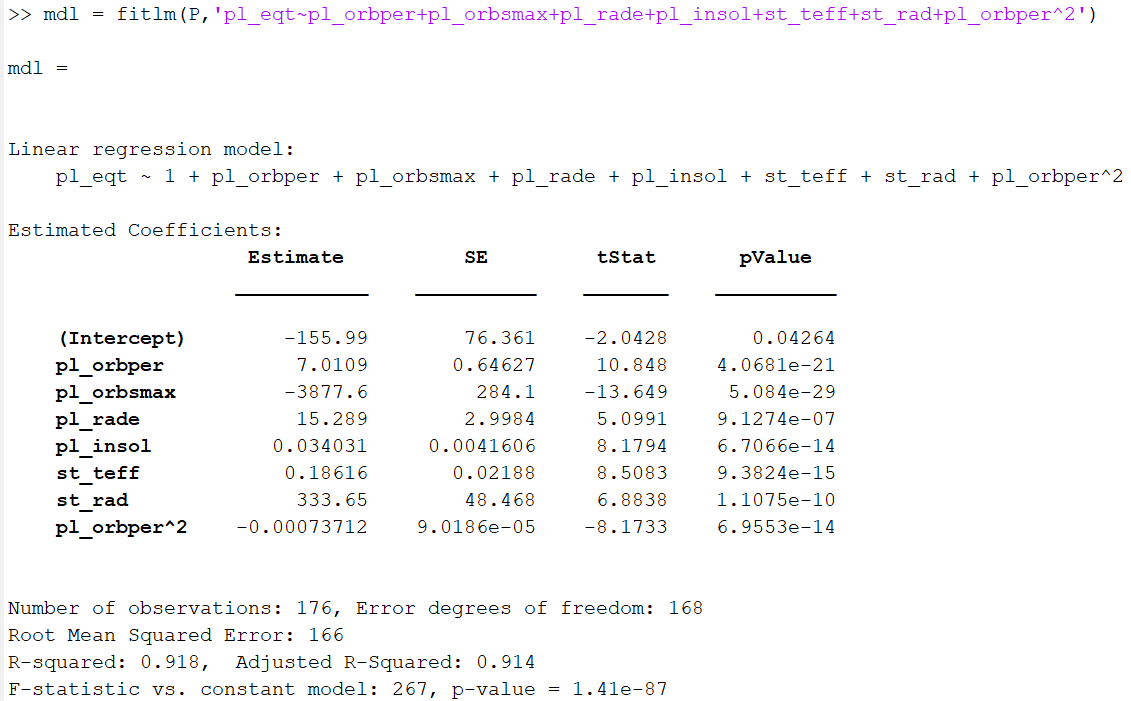
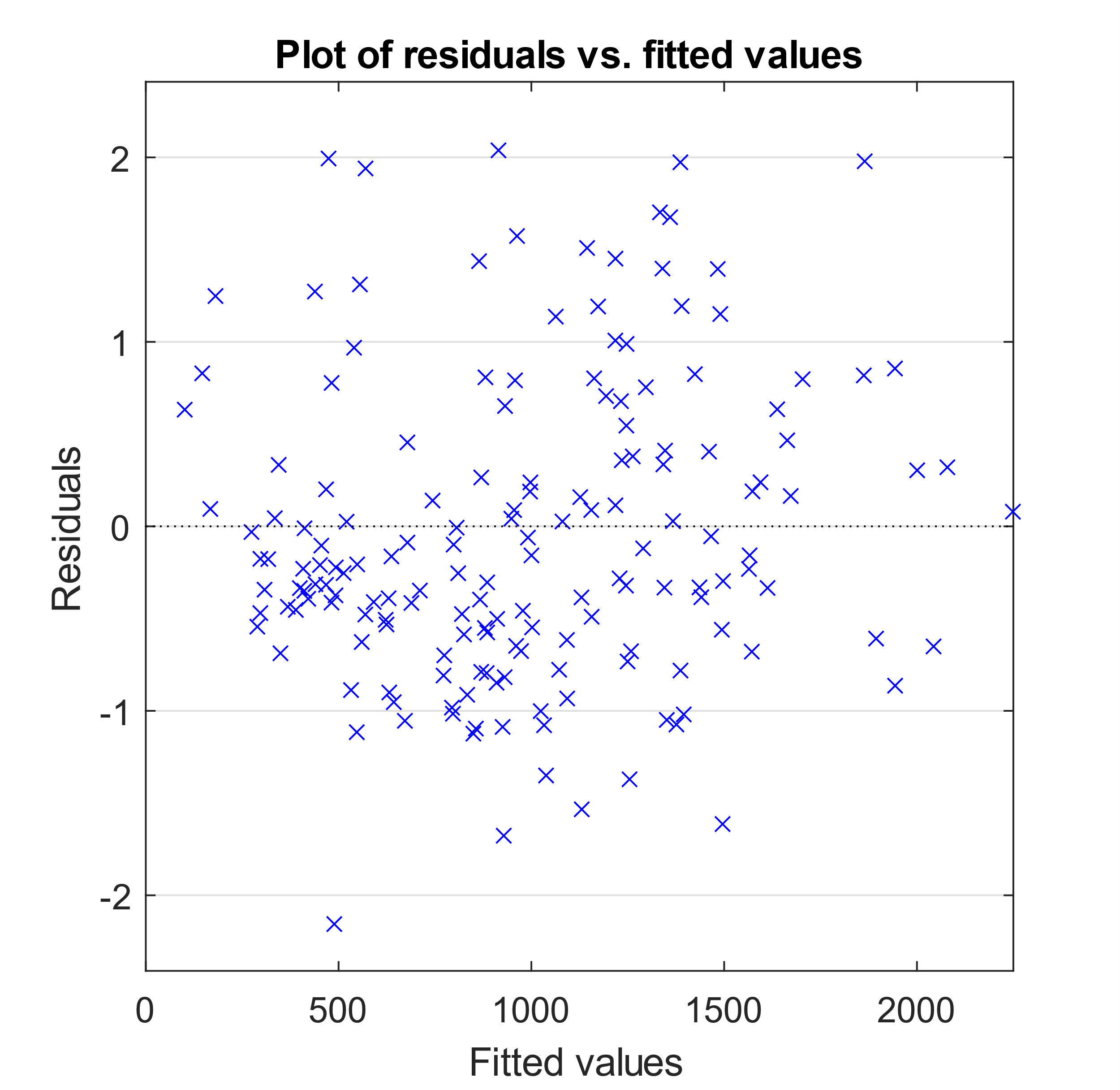
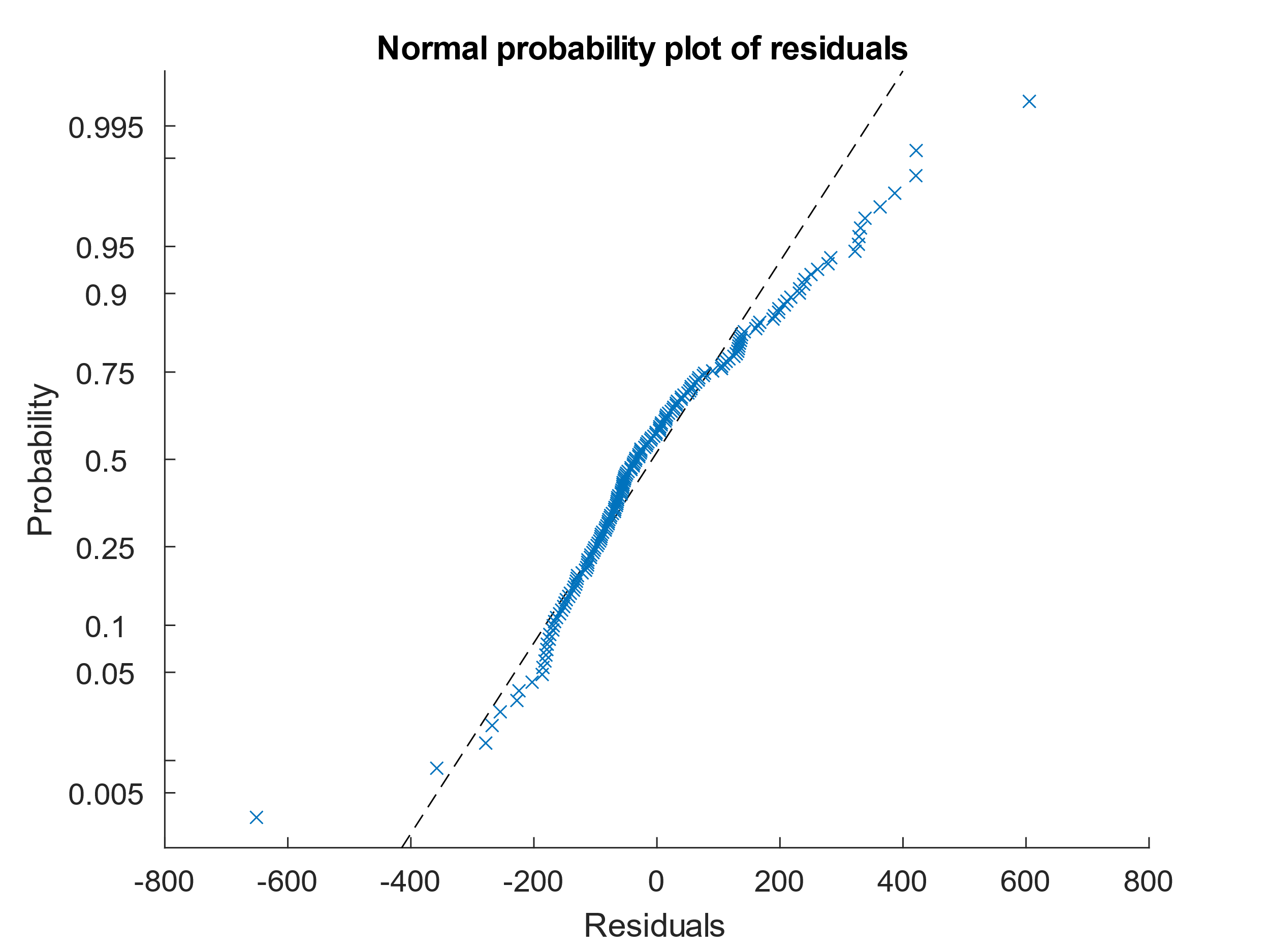
Proponiamo quindi degli altri modelli con predittori elevati al quadrato, la maggior parte contribuiscono ad un aumento del coefficiente di determinazione ma nessuno di essi risolve il problema dell’eteroschedasticità dei dati





Questo modello è stato ottenuto includendo tutti i predittori e i loro quadrati e procedendo come in precedenza con la rimozione dei dei predittori meno significativi in ordine di pValue.

Il valore del coefficiente di determinazione di questo modello è decisamente aumentato, attestandosi a ciononostante l’aderenza dei dati ai quantili nel qq-plot è decisamente peggiorata ed anche il plot dei residui mostra un andamento fortemente eteroschedastico, con un accenno di parabola.

Nuovamente notiamo che il coefficiente di determinazione del modello è molto buono ma i residui sono distintamente eteroschedastici.

Nel tentativo di trovare un modello che rispettasse l’assunzione di omoschedasticità dei residui sono stati svariati i modelli che non hanno dato i risultati sperati, modelli con interazioni semplici, con interazioni quadrate o cubiche, con i logaritmi o gli inversi di alcuni predittori. Sono tutti risultati peggiori di quelli sopra presentati, siamo quindi obbligati a ritenere il primo modello, quello con i residui più omoschedastici. quello definitivo. Sottolineiamo comunque che il modello è ben lontano dall’essere perfetto, tuttavia non abbiamo tutti i parametri che potrebbero influenzare la temperatura, come la composizione del pianeta o la composizione della stella, riteniamo dunque il modello abbastanza soddisfacente ai nostri scopi.