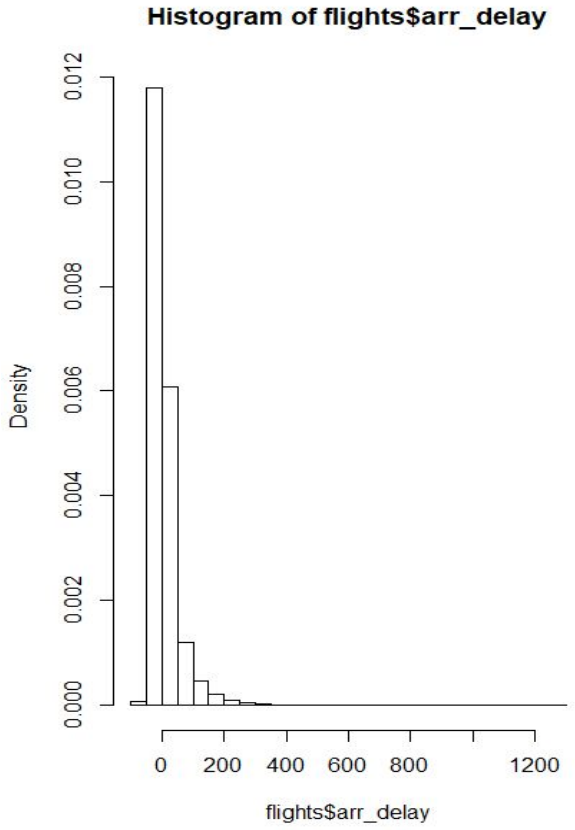
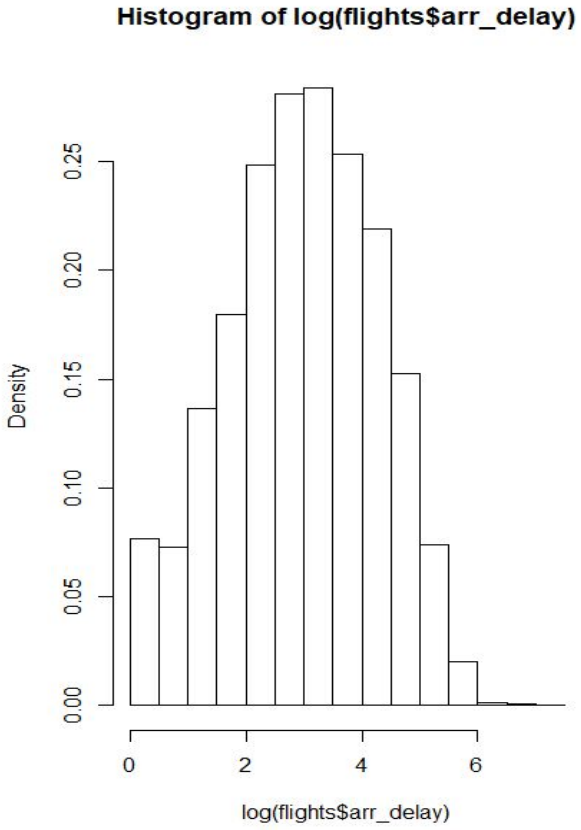
E.M.

Relazione analisi dataset “Flights”

La mia analisi partirà con ispezioni grafiche e controllo di dati mancanti. Cercherò poi di costruire un modello lineare e di valutare l’inserimento di qualche spline.
Sceglierò poi il modello che ha un test error minore per poi rispondere alla domanda.

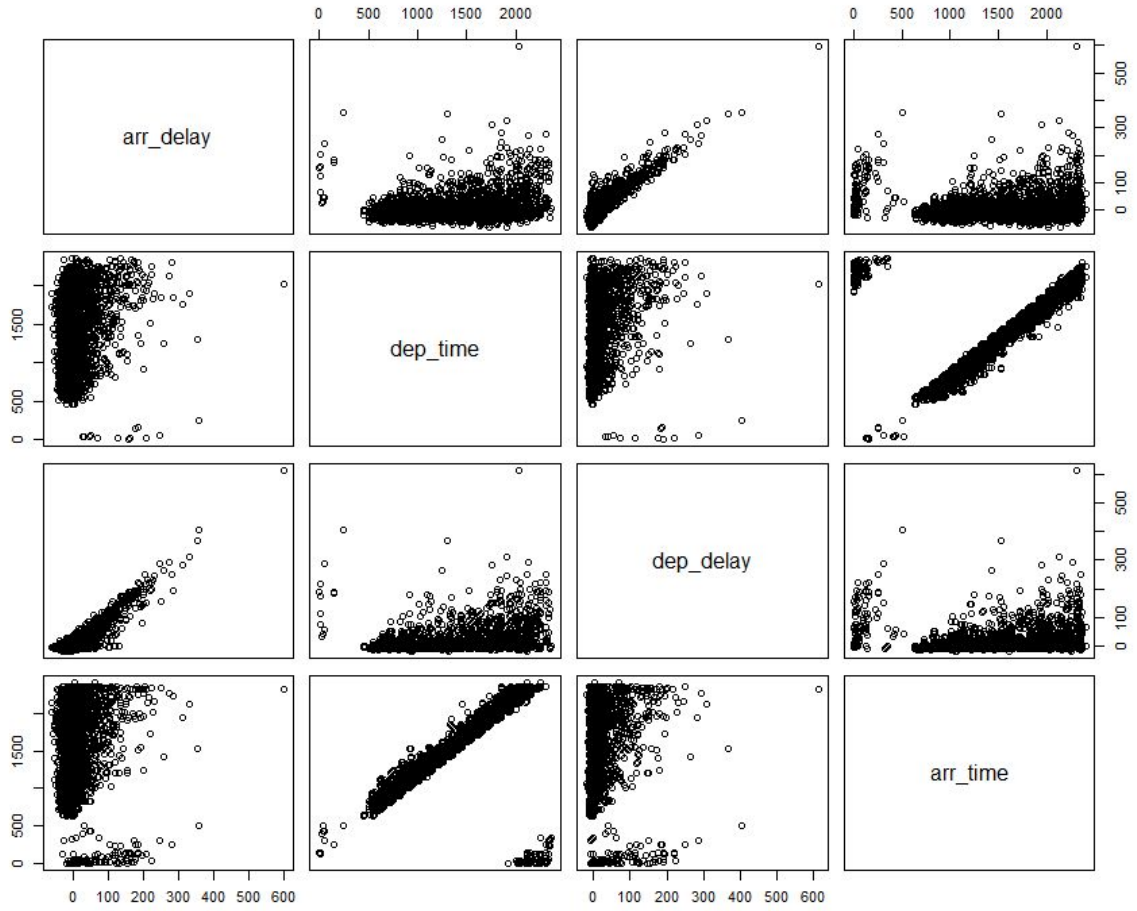
1. Dati mancanti e variabili qualitative considerate come fattori
Il dataset dato non contiene dati mancanti e le variabile qualitative sono interpretate da R come fattori.
2. Ispezioni grafiche





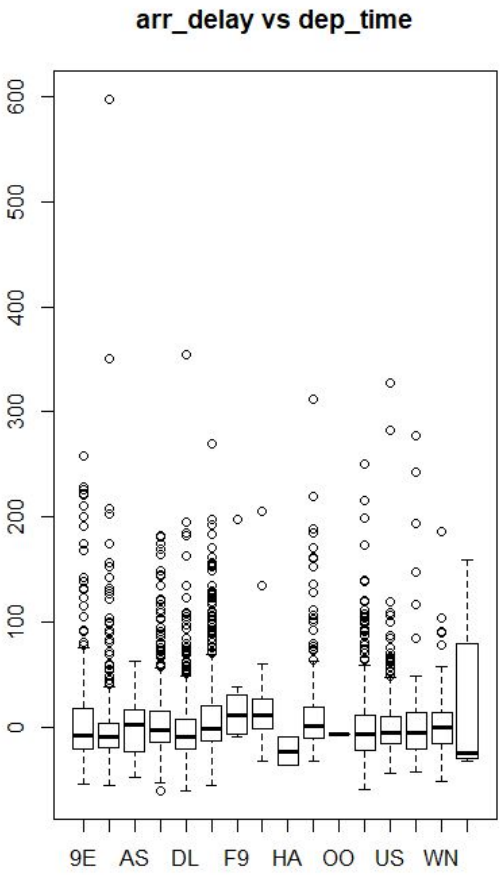
Un logaritmo migliorerebbe la distribuzione ma nel dataset in questione si tiene conto anche dei voli che arrivano in anticipo quindi non si può “normalizzare” la distribuzione, lavorerò con la distribuzione iniziale(grafico a sinistra).

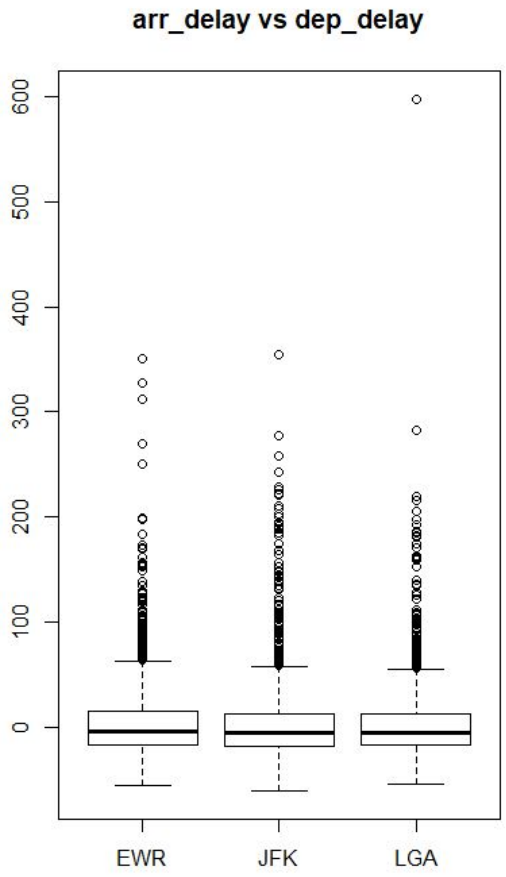
-Relazioni con le esplicative continue



Per semplicità non metto il grafico con la terza esplicativa(richiesta nella domanda 1) del sottoinsieme scelto(molto simile al grafico sopra in altro a sinistra).
Da questo grafico mi aspetto un coefficiente significativo(p-value ~ 0) almeno per la variabile dep_delay,inoltre mi aspetto un comportamento anomalo(ex. L'iterazione tra esse potrebbe essere significativa) vista la forte correlazione dep_time e dep_delay.

-Relazioni con le esplicative qualitative





-Relazione tra esplicative continue e qualitative

Per motivi di tempo non farò' i grafici che rilevano l'eventuale interazione, si evidenzieranno nelle stime dei coefficienti del modello, nell' R^2 ecc..

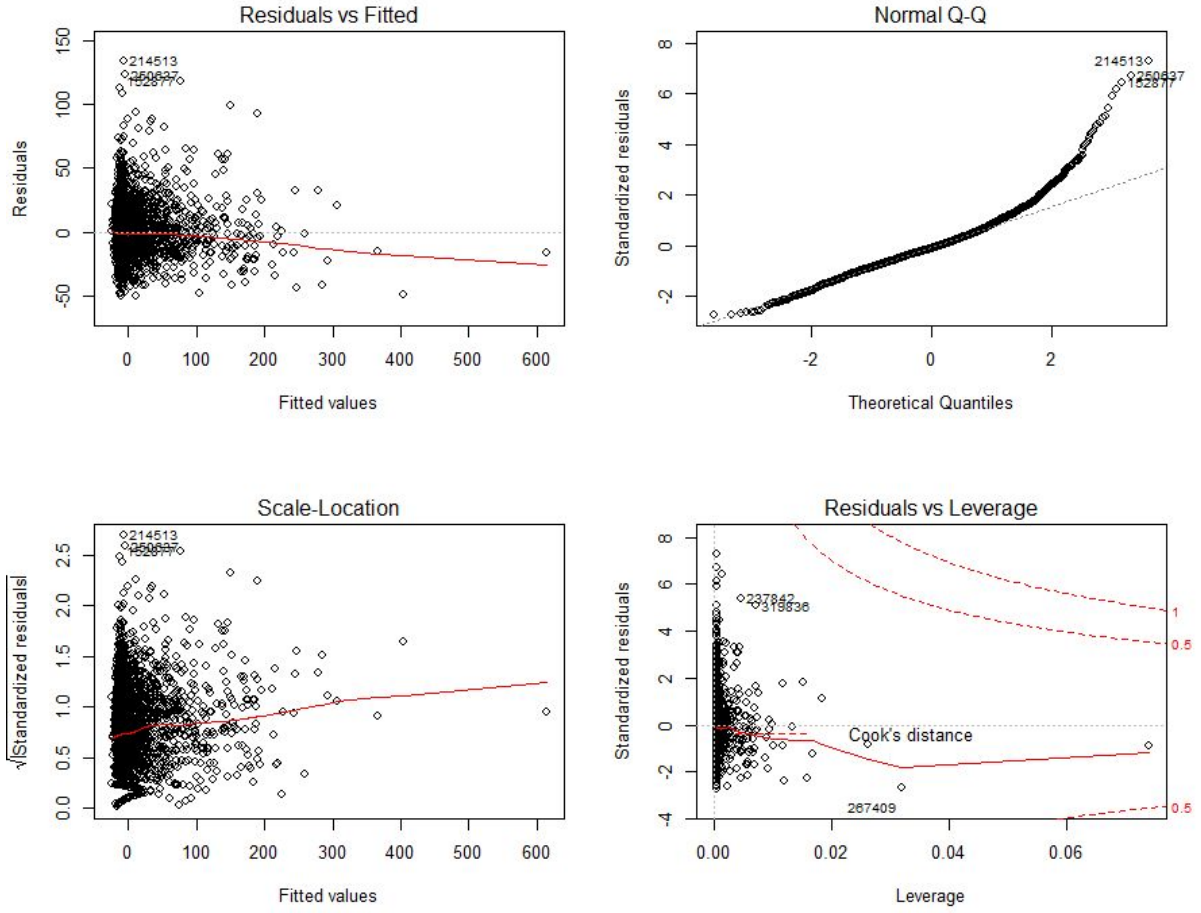
3) Modello Lineare

Dopo aver provato vari modelli lineari, sono giunti ad un modello che spiega la variabile **arr_delay** in funzione di:
dep_delay.
Questo modello non contiene l'interazioni ne altre covariate visto che tali modelli non miglioravano di molto ne l'R^2 (l'aumento nel migliore dei casi era del 0,007%) quindi si è preferito un modello più semplice.

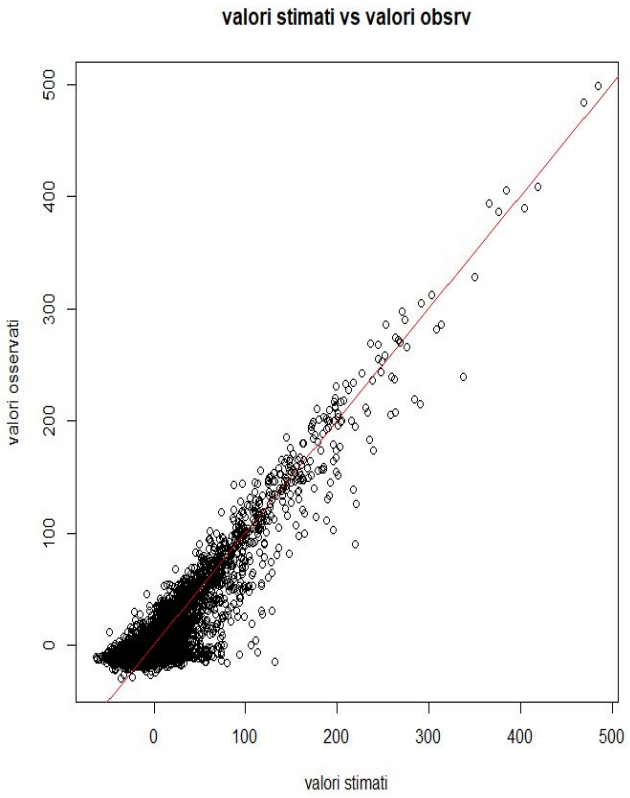
Il modello è:

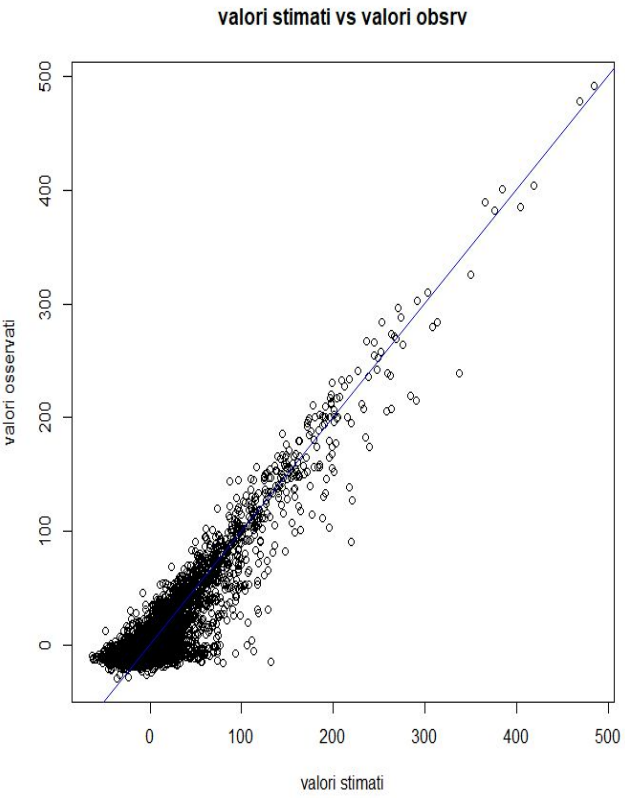
```
Call:
lm(formula = arr_delay ~ dep_delay, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-50.300 -11.146  -2.036   8.909 133.887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.887265   0.335056  -17.57   <2e-16 ***
dep_delay    1.011002   0.008285  122.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.35 on 3271 degrees of freedom
Multiple R-squared:  0.8199,         Adjusted R-squared:  0.8198
F-statistic: 1.489e+04 on 1 and 3271 DF,  p-value: < 2.2e-16
```

L’andamento dei residui è il seguente:



I residui hanno un andamento (non deterministico), la media dei residui non è circa 0 ma comunque è una buona approssimazione(si potrebbe schiacciarli un po più verso lo zero aggiungendo un termine polinomiale),i potenziali outlier vengono catturati nella distanza di Cook (grafico in basso a destra).

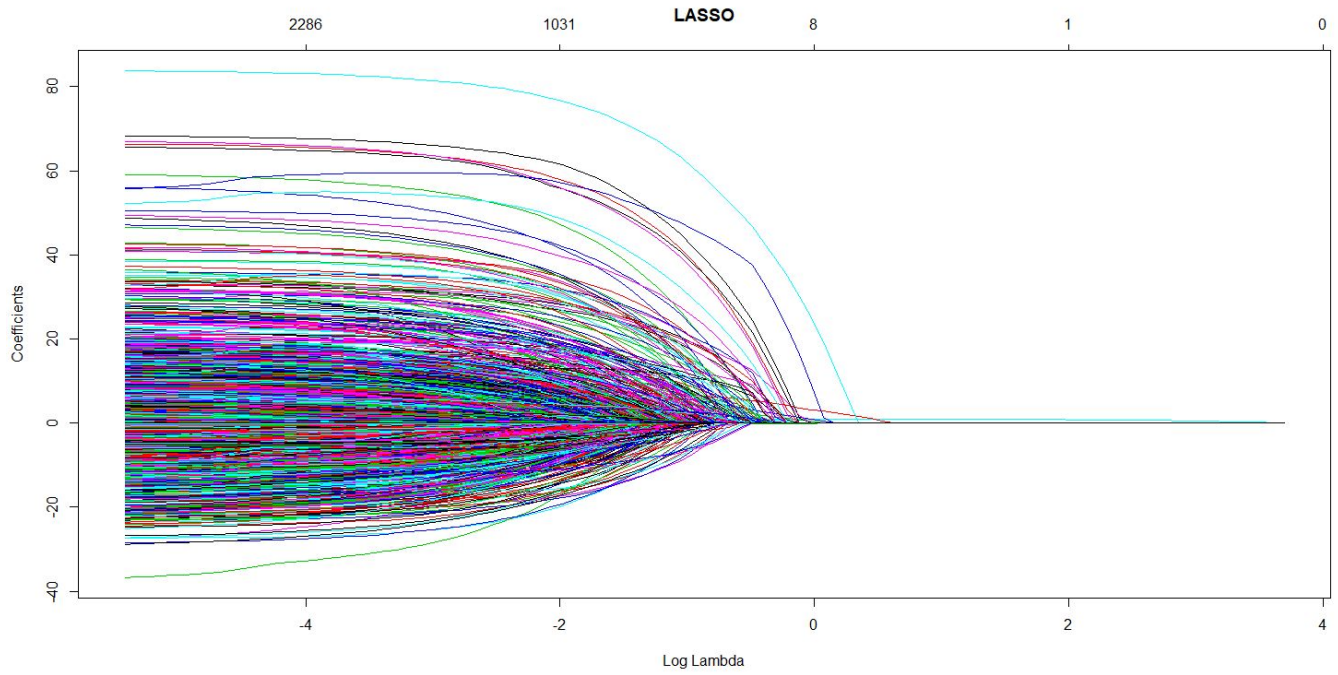




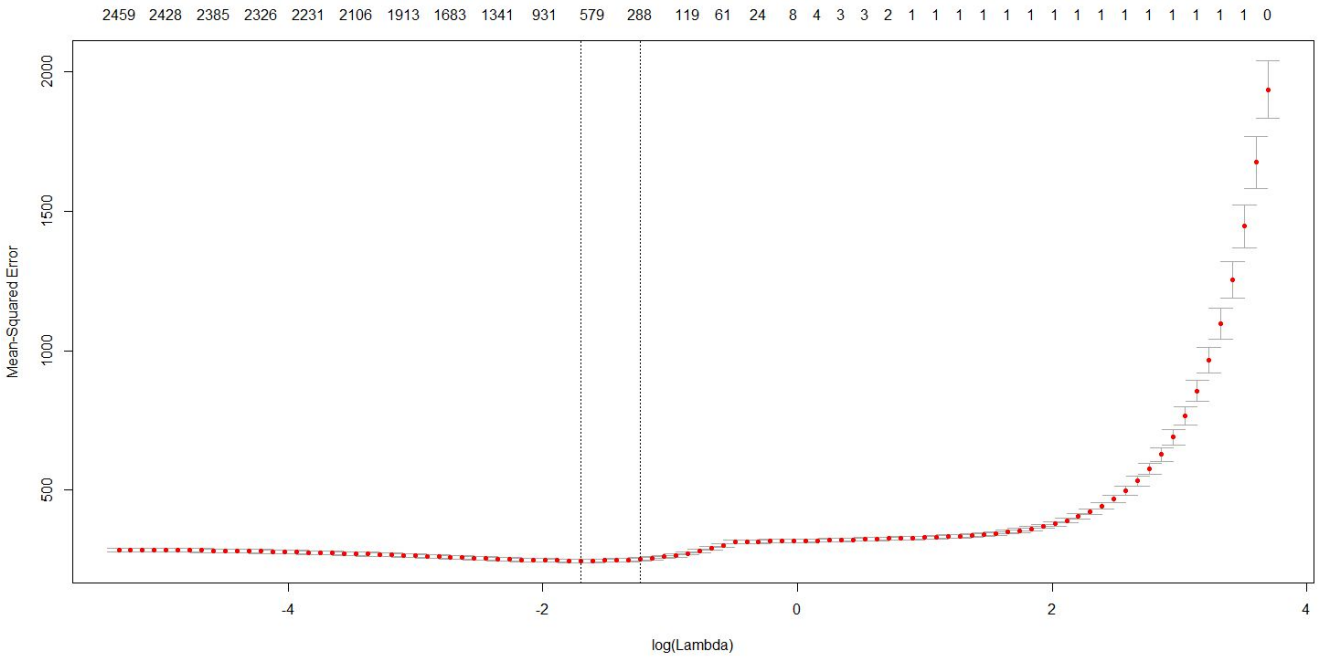
Si è testato un eventuale modello con spline cubica(immagine sopra a destra) ed il risultato è stato molto simile sia graficamente che come valore dell'AIC quindi si è preferito tenere il modello senza spline per semplicità.I punti sono concentrati attorno la retta.

PARTE 2:
Per questa parte saltero' le ispezioni grafiche in quanto farò l'analisi con ridge/lasso per la selezione delle variabili e la stima dei coefficienti.

1) Variazione dei coefficienti rispetto a lambda (Lasso) set.seed (123)



2) Cross validation per trovare il lambda minimo
set.seed(123)



MSE minimo=246.5887
lambda minimo=0.1826737
lambda 1se= 0.2908682

4) Modello lasso con lambda minimo

• Coefficienti //per semplicità riporto solo i primi(più significativi)

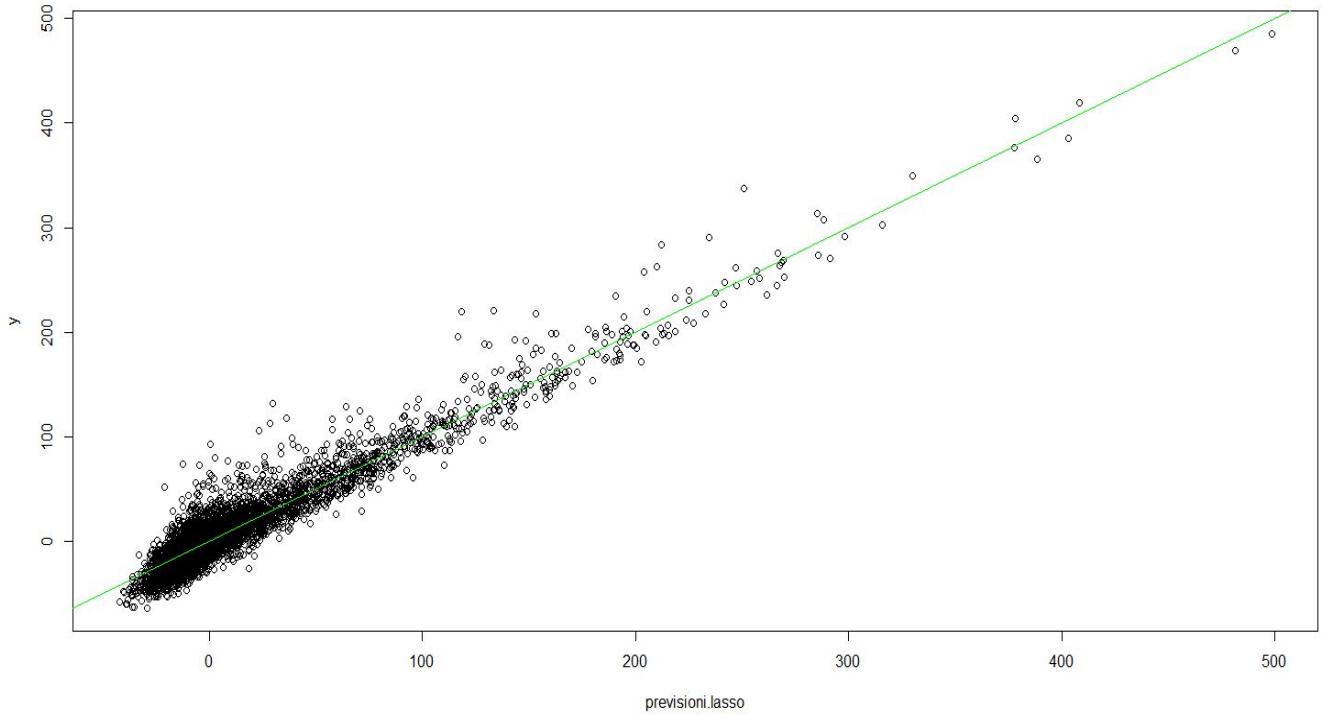
.

2533 x 3 sparse Matrix of class "dgCMatrix"

	s0	s0
(Intercept)	4.328534e+04	-1.661328e+01	-1.661328e+01
year	NA	.	.
month	8.440253e+01	2.402889e-02	2.402889e-02
day	2.779384e+00	.	.
dep_time	4.096502e-03	.	.
sched_dep_time	-1.077721e-02	.	.
dep_delay	1.014350e+00	1.006133e+00	1.006133e+00
arr_time	1.417398e-03	.	.
sched_arr_time	-1.547320e-03	-8.529033e-05	-8.529033e-05
carrierAA	3.120786e+00	.	.
carrierAS	1.738385e+01	-2.590261e+00	-2.590261e+00
carrierB6	1.322389e+01	5.124422e+00	5.124422e+00
carrierDL	2.542901e+01	.	.
carrierEV	5.237942e+00	2.851519e+00	2.851519e+00
carrierF9	6.327549e+00	1.800296e+00	1.800296e+00

carrierFL 3.296359e+01 2.194918e+00 2.194918e+00
carrierHA -1.710847e+00 . .
carrierMQ 2.102275e+01 7.247025e+00 7.247025e+00
carrierUA 6.350391e+00 . .
carrierUS 6.810585e+00 6.198967e+00 6.198967e+00
carrierVX 3.840423e+00 . .
carrierWN 2.021814e+01 -1.374004e+00 -1.374004e+00
carrierYV 2.046352e+01 . .
flight 1.805608e-04 . .
tailnumN10156 1.409265e+01 . .
tailnumN104UW -7.966269e+00 -6.135261e+00 -6.135261e+00
tailnumN10575 -3.285544e+00 -1.325811e+00 -1.325811e+00
tailnumN105UW 2.866202e+01 . .

- **Grafico valori stimati (lambda-min)**



5) Modello con lambda 1se sempre lasso.

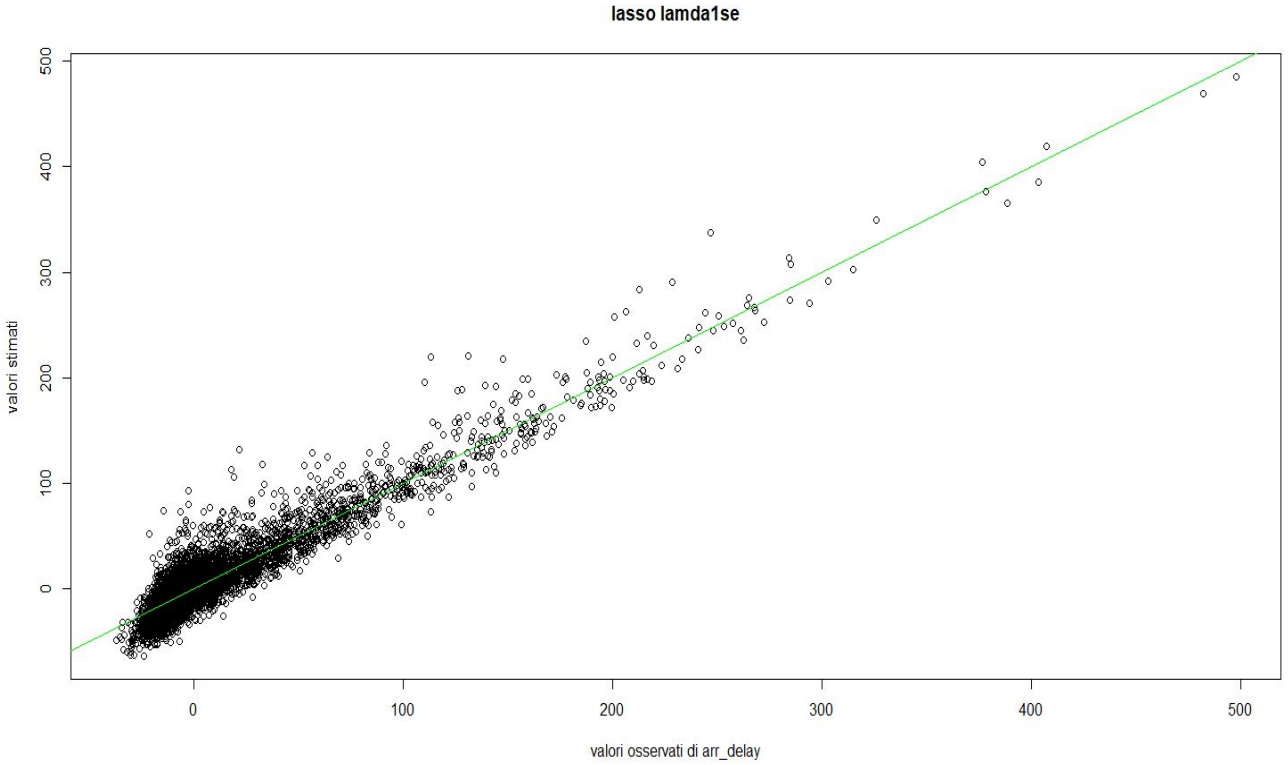
E’ un modello che da’ più regolarità al variare del test error

• **Coefficienti**

//per semplicità riporto solo i primi...come sopra

dep_delay 1.00247108
arr_time .
sched_arr_time .
carrierAA .
carrierAS -2.12893193
carrierB6 4.62465604
carrierDL .
carrierEV 2.71259489
carrierF9 1.06434543
carrierFL 2.57115338
carrierHA .
carrierMQ 7.03850985
carrierUA .
carrierUS 6.02065093
carrierVX .
carrierWN -0.74901962
carrierYV .
flight .
tailnumN10156 .
tailnumN104UW .
tailnumN10575 .
tailnumN105UW .

• **Grafico valori stimati**



Conclusione:

Si è preferito il modello con lasso perché ha un test error inferiore(246) rispetto al modello che utilizza ridge(431), da notare il fatto che il test error è molto alto,ciò ci da solo l'informazione che il modello da preferire è quello che utilizza lasso e nessun' altra.
Bisognerebbe controllare il modello trovato con altri metodi che al momento non conosco, ma sicuramente il modello trovato è un buon modello visto il grafico che "prevede" i valori in modo abbastanza buono.