Analisi dei parametri vocali di pazienti affetti da malattia di Parkinson

Davide Dell'Orto - Matr. 828873

1. Descrizione ed esplorazione del dataset

Table 1: Data head (continued below)

subject	age	sex	test_time	motor_UPDRS	total_UPDRS	Jitter(%)
1	72	0	5.643	28.2	34.4	0.00662
1	72	0	12.67	28.45	34.89	0.003
1	72	0	19.68	28.7	35.39	0.00481

Jitter(Abs)	Jitter.RAP	Jitter.PPQ5	Jitter.DDP	Shimmer	Shimmer(dB)
3.38e-05	0.00401	0.00317	0.01204	0.02565	0.23
1.68e-05	0.00132	0.0015	0.00395	0.02024	0.179
2.462e-05	0.00205	0.00208	0.00616	0.01675	0.181

Shimmer.APQ3	Shimmer.APQ5	Shimmer.APQ11	Shimmer.DDA	NHR	HNR
0.01438	0.01309	0.01662	0.04314	0.01429	21.64
0.00994	0.01072	0.01689	0.02982	0.01111	27.18
0.00734	0.00844	0.01458	0.02202	0.02022	23.05

RPDE	DFA	PPE
0.4189	0.5484	0.1601
0.4349	0.5648	0.1081
0.4622	0.5441	0.2101

Il dataset è composto da 5875 osservazioni e 22 variabili:

- *subject* - Codice dell'individuo (factor)
- *age* - Età dell'individuo (factor)
- *sex* - Sesso dell'individuo (factor)
- *test_time* - Giorni passati dal primo test (num)
- *motor_UPDRS* - Punteggio indicante l'intensità dei disturbi motori (num, da 0 a 108. I disturbi motori, includendo rigidità, tremore ed elasticità dei muscoli facciali, influenzano inevitabilmente il parlato)
- *total_UPDRS* - Punteggio indicante l'intensità della malattia (num, da 0 a 176 dove 176 rappresenta disabilità totale)

- Da *Jitter(%)* a *PPE* - Parametri vocali (num)

Non sono presenti *missing values*, infatti:

```
df[rowSums(is.na(df)) > 0, ]
```

Restituisce un data frame con 0 righe. A questo punto prendo in considerazione solo le variabili numeriche in modo da visualizzarne i principali indici descrittivi e calcolarne la matrice di correlazione:

```
numerics <- unlist(lapply(df, is.numeric))
dfnum <- df[, numerics]

mcor <- round(cor(dfnum), 2)
mcor[upper.tri(mcor)] <- ""
mcor <- as.data.frame(mcor)

pander(summary(dfnum), caption = "Data summary")
```

Table 5: Data summary (continued below)

age	test_time	motor_UPDRS	total_UPDRS
Min. :36.0	Min. : -4.263	Min. : 5.038	Min. : 7.00
1st Qu.:58.0	1st Qu.: 46.847	1st Qu.:15.000	1st Qu.:21.37
Median :65.0	Median : 91.523	Median :20.871	Median :27.58
Mean :64.8	Mean : 92.864	Mean :21.296	Mean :29.02
3rd Qu.:72.0	3rd Qu.:138.445	3rd Qu.:27.596	3rd Qu.:36.40
Max. :85.0	Max. :215.490	Max. :39.511	Max. :54.99

Jitter(%)	Jitter(Abs)	Jitter.RAP	Jitter.PPQ5
Min. :0.000830	Min. :2.250e-06	Min. :0.000330	Min. :0.000430
1st Qu.:0.003580	1st Qu.:2.244e-05	1st Qu.:0.001580	1st Qu.:0.001820
Median :0.004900	Median :3.453e-05	Median :0.002250	Median :0.002490
Mean :0.006154	Mean :4.403e-05	Mean :0.002987	Mean :0.003277
3rd Qu.:0.006800	3rd Qu.:5.333e-05	3rd Qu.:0.003290	3rd Qu.:0.003460
Max. :0.099990	Max. :4.456e-04	Max. :0.057540	Max. :0.069560

Jitter.DDP	Shimmer	Shimmer(dB)	Shimmer.APQ3
Min. :0.000980	Min. :0.00306	Min. :0.026	Min. :0.00161
1st Qu.:0.004730	1st Qu.:0.01912	1st Qu.:0.175	1st Qu.:0.00928
Median :0.006750	Median :0.02751	Median :0.253	Median :0.01370
Mean :0.008962	Mean :0.03404	Mean :0.311	Mean :0.01716
3rd Qu.:0.009870	3rd Qu.:0.03975	3rd Qu.:0.365	3rd Qu.:0.02057
Max. :0.172630	Max. :0.26863	Max. :2.107	Max. :0.16267

Shimmer.APQ5	Shimmer.APQ11	Shimmer.DDA	NHR
Min. :0.00194	Min. :0.00249	Min. :0.00484	Min. :0.000286
1st Qu.:0.01079	1st Qu.:0.01566	1st Qu.:0.02783	1st Qu.:0.010955
Median :0.01594	Median :0.02271	Median :0.04111	Median :0.018448
Mean :0.02014	Mean :0.02748	Mean :0.05147	Mean :0.032120
3rd Qu.:0.02375	3rd Qu.:0.03272	3rd Qu.:0.06173	3rd Qu.:0.031463
Max. :0.16702	Max. :0.27546	Max. :0.48802	Max. :0.748260

HNR	RPDE	DFA	PPE
Min. : 1.659	Min. :0.1510	Min. :0.5140	Min. :0.02198
1st Qu.:19.406	1st Qu.:0.4698	1st Qu.:0.5962	1st Qu.:0.15634
Median :21.920	Median :0.5423	Median :0.6436	Median :0.20550
Mean :21.680	Mean :0.5415	Mean :0.6532	Mean :0.21959
3rd Qu.:24.444	3rd Qu.:0.6140	3rd Qu.:0.7113	3rd Qu.:0.26449
Max. :37.875	Max. :0.9661	Max. :0.8656	Max. :0.73173

```
pander(mcor, caption = "Correlation matrix")
```

Table 10: Correlation matrix (continued below)

	age	test_time	motor_UPDRS	total_UPDRS	Jitter(%)
age	1				
test_time	0.02	1			
motor_UPDRS	0.27	0.07	1		
total_UPDRS	0.31	0.08	0.95	1	
Jitter(%)	0.02	-0.02	0.08	0.07	1
Jitter(Abs)	0.04	-0.01	0.05	0.07	0.87
Jitter.RAP	0.01	-0.03	0.07	0.06	0.98
Jitter.PPQ5	0.01	-0.02	0.08	0.06	0.97
Jitter.DDP	0.01	-0.03	0.07	0.06	0.98
Shimmer	0.1	-0.03	0.1	0.09	0.71
Shimmer(dB)	0.11	-0.03	0.11	0.1	0.72
Shimmer.APQ3	0.1	-0.03	0.08	0.08	0.66
Shimmer.APQ5	0.09	-0.04	0.09	0.08	0.69
Shimmer.APQ11	0.14	-0.04	0.14	0.12	0.65
Shimmer.DDA	0.1	-0.03	0.08	0.08	0.66
NHR	0.01	-0.03	0.07	0.06	0.83
HNR	-0.1	0.04	-0.16	-0.16	-0.68
RPDE	0.09	-0.04	0.13	0.16	0.43
DFA	-0.09	0.02	-0.12	-0.11	0.23
PPE	0.12	0	0.16	0.16	0.72

	Jitter(Abs)	Jitter.RAP	Jitter.PPQ5	Jitter.DDP
age				
test_time				
motor_UPDRS				
total_UPDRS				

	Jitter(Abs)	Jitter.RAP	Jitter.PPQ5	Jitter.DDP
Jitter(%)				
Jitter(Abs)	1			
Jitter.RAP	0.84	1		
Jitter.PPQ5	0.79	0.95	1	
Jitter.DDP	0.84	1	0.95	1
Shimmer	0.65	0.68	0.73	0.68
Shimmer(dB)	0.66	0.69	0.73	0.69
Shimmer.APQ3	0.62	0.65	0.68	0.65
Shimmer.APQ5	0.62	0.66	0.73	0.66
Shimmer.APQ11	0.59	0.6	0.67	0.6
Shimmer.DDA	0.62	0.65	0.68	0.65
NHR	0.7	0.79	0.86	0.79
HNR	-0.71	-0.64	-0.66	-0.64
RPDE	0.55	0.38	0.38	0.38
DFA	0.35	0.21	0.18	0.21
PPE	0.79	0.67	0.66	0.67

	Shimmer	Shimmer(dB)	Shimmer.APQ3	Shimmer.APQ5
age				
test__time				
motor__UPDRS				
total__UPDRS				
Jitter(%)				
Jitter(Abs)				
Jitter.RAP				
Jitter.PPQ5				
Jitter.DDP				
Shimmer	1			
Shimmer(dB)	0.99	1		
Shimmer.APQ3	0.98	0.97	1	
Shimmer.APQ5	0.98	0.98	0.96	1
Shimmer.APQ11	0.94	0.94	0.89	0.94
Shimmer.DDA	0.98	0.97	1	0.96
NHR	0.8	0.8	0.73	0.8
HNR	-0.8	-0.8	-0.78	-0.79
RPDE	0.47	0.47	0.44	0.45
DFA	0.13	0.13	0.13	0.13
PPE	0.62	0.64	0.58	0.59

	Shimmer.APQ11	Shimmer.DDA	NHR	HNR	RPDE	DFA	PPE
age							
test__time							
motor__UPDRS							
total__UPDRS							
Jitter(%)							
Jitter(Abs)							
Jitter.RAP							
Jitter.PPQ5							

	Shimmer.APQ11	Shimmer.DDA	NHR	HNR	RPDE	DFA	PPE
Jitter.DDP							
Shimmer							
Shimmer(dB)							
Shimmer.APQ3							
Shimmer.APQ5							
Shimmer.APQ11	1						
Shimmer.DDA	0.89	1					
NHR	0.71	0.73	1				
HNR	-0.78	-0.78	-0.68	1			
RPDE	0.48	0.44	0.42	-0.66	1		
DFA	0.18	0.13	-0.02	-0.29	0.19	1	
PPE	0.62	0.58	0.56	-0.76	0.57	0.39	1

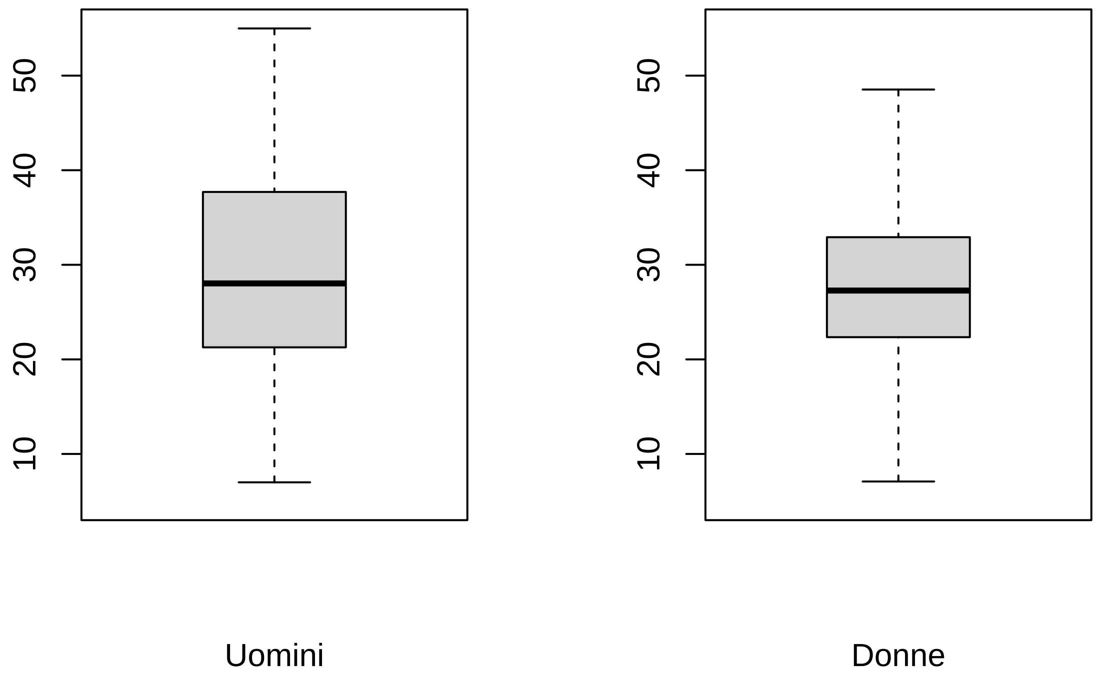
Si notano alcune correlazioni eccessivamente alte che costituiranno un problema di multicollinearità quando si andrà a sviluppare il modello di regressione.

2. Inferenza su medie

Voglio verificare che esista una differenza tra i punteggi medi di uomo e donna. Precisamente, che la media di total_UPDRS, ovvero l'intensità media della malattia, sia significativamente diversa tra i due sessi.

```r
uomini <- df[df$sex == "0", ]
donne <- df[df$sex == "1", ]

par(mfrow = c(1, 2))
boxplot(uomini$total_UPDRS, xlab = "Uomini", ylim = c(5, 55))
boxplot(donne$total_UPDRS, xlab = "Donne", ylim = c(5, 55))
```




```
mean(uomini$total_UPDRS)
```

```
## [1] 29.72406
```

```
mean(donne$total_UPDRS)
```

```
## [1] 27.50523
```

I dati sul campione suggeriscono che vi è una differenza di distribuzione tra gli uomini e le donne così come valori medi di total_UPDRS leggermente diversi. Verifico allora, tramite il test t, che questa differenza in media non sia dovuta al caso ma sia da attribuirsi all'intera popolazione. In particolare, mi accerto prima della significatività della differenza nelle distribuzioni tramite il test chi-quadrato, ovvero mi accerto che la variabile total_UPDRS dipenda dal sesso:

```
UPDRS <- data.frame(c(uomini$total_UPDRS, donne$total_UPDRS), rep(c("M", "W"), c(4008,
    1867)))
colnames(UPDRS) <- c("total_UPDRS", "Sesso")

classes <- c(seq(5, 55, by = 5))
tab <- table(cut(df$total_UPDRS, classes), UPDRS$Sesso)

pander(chisq.test(tab))
```

Table 14: Pearson's Chi-squared test: `tab`

Test statistic	df	P value
1260	9	1.218e-265 * * *

Dato un p-value < 0.05, rifiuto quindi l'ipotesi che le variabili total_UPDRS e sex siano indipendenti. Procedo allora a saggiare la significatività di questa differenza. Nello specifico, dato che le osservazioni a disposizione mi suggeriscono che negli uomini la media è superiore, voglio porre come ipotesi nulla H_0: mean_UPDRS(uomini) $\leq$ mean_UPDRS(donne).

```
pander(t.test(uomini$total_UPDRS, donne$total_UPDRS, var.equal = FALSE, alternative = "greater"),
    caption = "Welch Two Sample t-test")
```

Table 15: Welch Two Sample t-test (continued below)

Test statistic	df	P value	Alternative hypothesis	mean of x
7.74	4032	6.241e-15 * * *	greater	29.72

mean of y
27.51

Con un p-value < 0.05 rifiuto l'ipotesi nulla e posso quindi dire che, in media, valori maggiori negli uomini non sono dovuti al caso, ma sono osservabili su un'ipotetica intera popolazione.

3. Regressione lineare multipla

Prima di procedere con l'individuazione dei regressori è necessario eliminare le variabili eccessivamente correlate tra di loro, al fine di evitare, come accennato all'inizio, il problema della multicollinearità. Dato l'alto numero di variabili e l'importante dimensione della matrice di correlazione, creo una funzione *select* che mi permette di visualizzare rapidamente le correlazioni che superano una determinata soglia, in questo caso quelle maggiori (minori) di 0.8 (-0.8):

```
select <- function(x, value) {
    ind <- which(upper.tri(x), arr.ind = TRUE)
    maxcor <- data.frame(x1 = dimnames(x)[[2]][ind[, 2]], x2 = dimnames(x)[[1]][ind[,
        1]], corr = x[ind])
    return(maxcor[abs(maxcor$corr) >= value, ])
}

pander(select(cor(dfnum), 0.8))
```

	x1	x2	corr
6	Jitter(Abs)	Jitter(%)	0.8656
9	Jitter.RAP	Jitter(%)	0.9842
10	Jitter.RAP	Jitter(Abs)	0.8446
13	Jitter.PPQ5	Jitter(%)	0.9682

	x1	x2	corr
15	Jitter.PPQ5	Jitter.RAP	0.9472
18	Jitter.DDP	Jitter(%)	0.9842
19	Jitter.DDP	Jitter(Abs)	0.8446
20	Jitter.DDP	Jitter.RAP	1
21	Jitter.DDP	Jitter.PPQ5	0.9472
36	Shimmer(dB)	Shimmer	0.9923
44	Shimmer.APQ3	Shimmer	0.9798
45	Shimmer.APQ3	Shimmer(dB)	0.968
53	Shimmer.APQ5	Shimmer	0.9849
54	Shimmer.APQ5	Shimmer(dB)	0.9764
55	Shimmer.APQ5	Shimmer.APQ3	0.9627
63	Shimmer.APQ11	Shimmer	0.9355
64	Shimmer.APQ11	Shimmer(dB)	0.9363
65	Shimmer.APQ11	Shimmer.APQ3	0.8857
66	Shimmer.APQ11	Shimmer.APQ5	0.9389
74	Shimmer.DDA	Shimmer	0.9798
75	Shimmer.DDA	Shimmer(dB)	0.968
76	Shimmer.DDA	Shimmer.APQ3	1
77	Shimmer.DDA	Shimmer.APQ5	0.9627
78	Shimmer.DDA	Shimmer.APQ11	0.8857
81	NHR	Jitter(%)	0.8253
84	NHR	Jitter.PPQ5	0.8649
99	HNR	Shimmer	-0.8014
100	HNR	Shimmer(dB)	-0.8025

Per ogni coppia di variabili viene eliminata quella correlata in misura minore con la variabile target, ritrovandosi al termine con 8 variabili, di cui 7 potenzialmente esplicative:

```
pander(head(dfnum, 3))
```

age	total_UPDRS	Jitter(%)	Shimmer.APQ11	HNR	RPDE	DFA	PPE
72	34.4	0.00662	0.01662	21.64	0.4189	0.5484	0.1601
72	34.89	0.003	0.01689	27.18	0.4349	0.5648	0.1081
72	35.39	0.00481	0.01458	23.05	0.4622	0.5441	0.2101

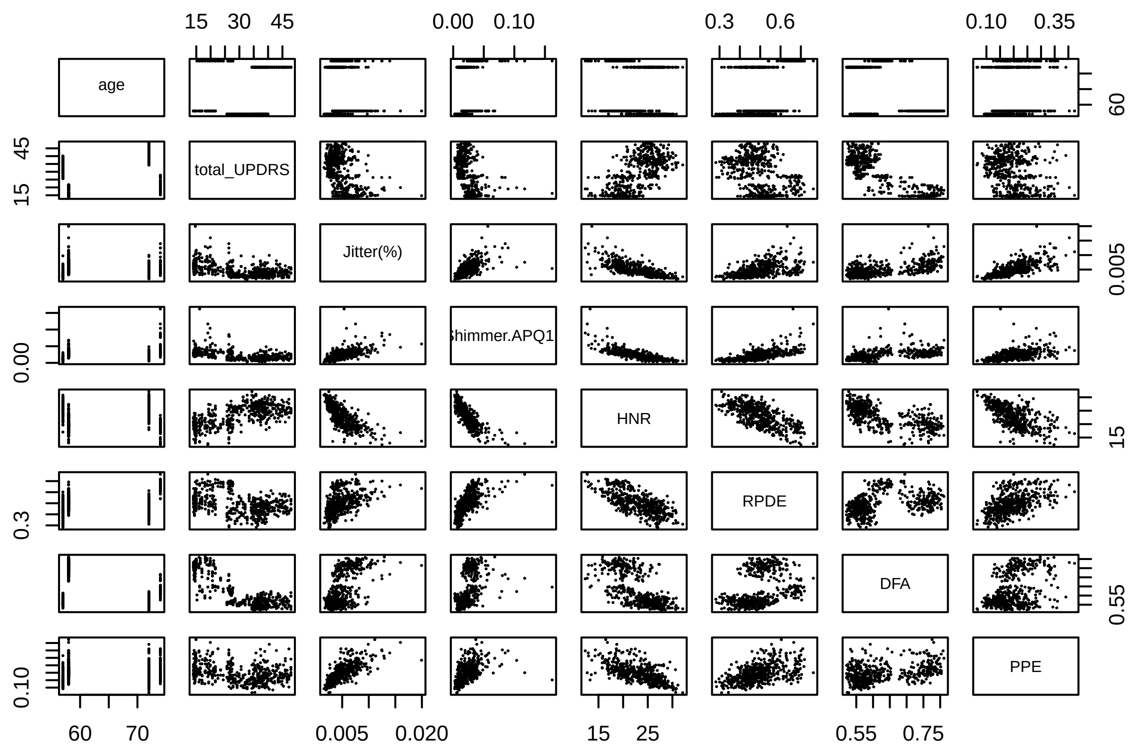
Si procede a questo punto alla costruzione del modello: la prima variabile introdotta è quella con la più alta correlazione con la variabile target total_UPDRS:

```
mcor <- round(cor(dfnum, dfnum$total_UPDRS), 4)
mcor[upper.tri(mcor)] <- ""
mcor <- as.data.frame(mcor[-2, ])
colnames(mcor) <- "total_UPDRS"
pander(mcor)
```

	total_UPDRS
age	0.3103
Jitter(%)	0.0742

	total_UPDRS
Shimmer.APQ11	0.1208
HNR	-0.1621
RPDE	0.1569
DFA	-0.1135
PPE	0.1562

```r
pairs(dfnum[sample(500), ], cex = 0.1)
```



Si inizializza quindi il modello con la variabile age:

```r
model <- lm(df$total_UPDRS ~ df$age)
pander(summary(model), caption = "Fitting model")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.628	0.9841	4.703	2.624e-06
df$age	0.3764	0.01505	25.01	2.623e-131

Table 21: Fitting model

Observations	Residual Std. Error	R^2	Adjusted R^2
5875	10.17	0.09628	0.09613

Entrambi i coefficienti risultano significativamente diversi da 0. L'R^2 è però ancora molto basso (0.096) ed è quindi necessario vedere se e come varia all'introduzione di nuovi regressori. Si sceglie allora di sviluppare nuovi modelli tramite *forward selection* dove ciascun modello include man mano una variabile esplicativa in più, fino da arrivare ad includerle tutte. Tramite il test ANOVA verrà scelto quel modello che avrà la più piccola somma del quadrato degli errori (RSS), per un determinato livello di significatività.

```
model1 <- update(model, . ~ . + df$`Jitter(%)`)
model2 <- update(model1, . ~ . + df$Shimmer.APQ11)
model3 <- update(model2, . ~ . + df$HNR)
model4 <- update(model3, . ~ . + df$RPDE)
model5 <- update(model4, . ~ . + df$DFA)
model6 <- update(model5, . ~ . + df$PPE)

pander(vif(model6))
```

df$age	df$`Jitter(%)`	df$Shimmer.APQ11	df$HNR	df$RPDE	df$DFA	df$PPE
1.065	2.501	2.778	4.429	1.831	1.237	3.345

```
pander(anova(model, model1)[2, ])
```

Table 23: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	5872	604768	1	3029	29.41	6.103e-08

```
pander(anova(model, model2)[2, ])
```

Table 24: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	5871	603258	2	4539	22.09	2.775e-10

```
pander(anova(model, model3)[2, ])
```

Table 25: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	5870	595329	3	12468	40.98	3.365e-26

```r
pander(anova(model, model4)[2, ])
```

Table 26: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	5869	593238	4	14559	36.01	8.991e-30

```r
pander(anova(model, model5)[2, ])
```

Table 27: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	5868	581455	5	26342	53.17	3.816e-54

```r
pander(anova(model, model6)[2, ])
```

Table 28: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	5867	576559	6	31238	52.98	6.924e-64

Il modello che implementa l'RSS minore è quello che include tutte le variabili che, come verificato tramite il VIF, non soffre di multicollinearità: i valori infatti sono tutti sotto la soglia di 10.

```r
pander(summary(model6), caption = "Fitting model6")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.97	2.824	9.197	5.001e-20
df$age	0.3238	0.01513	21.4	7.479e-98
df$Jitter(%)	-148.6	36.37	-4.086	4.454e-05
df$Shimmer.APQ11	-29.1	10.79	-2.697	0.007008
df$HNR	-0.3085	0.06343	-4.863	1.184e-06
df$RPDE	5.821	1.733	3.359	0.000787
df$DFA	-25.55	2.029	-12.59	6.656e-36
df$PPE	18.25	2.585	7.059	1.872e-12

Table 30: Fitting model6

Observations	Residual Std. Error	R^2	Adjusted R^2
5875	9.913	0.1427	0.1417

È evidente come il modello spieghi solamente poca parte della variabilità di total_UPDRS. L'R^2 aggiustato infatti, essendo pari a 0.142, non è migliorato di molto rispetto al modello base. Si è deciso allora di provare a sviluppare un modello quadratico ovvero una regressione lineare multipla polinomiale. In particolare si è adottata una sorta di *backward selection* in quanto, partendo dal modello lineare *model6* sono stati fatti vari tentativi aggiungendo il quadrato di diverse variabili. Il modello scelto è il seguente:

```r
model_b <- lm(df$total_UPDRS ~ df$age + df$`Jitter(%)` + df$Shimmer.APQ11 + df$HNR +
    df$DFA + df$RPDE + df$PPE + I(HNR^2), data = df)

pander(summary(model_b), caption = "Fitting model_b")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.726	3.407	1.094	0.2742
df$age	0.284	0.01537	18.48	3.747e-74
df$Jitter(%)	106.7	42.37	2.517	0.01186
df$Shimmer.APQ11	31.88	11.94	2.671	0.007594
df$HNR	2.329	0.2397	9.715	3.835e-22
df$DFA	-32.9	2.108	-15.61	7.843e-54
df$RPDE	4.268	1.72	2.482	0.0131
df$PPE	12.75	2.603	4.898	9.957e-07
I(HNR^2)	-0.05887	0.005164	-11.4	8.688e-30

Table 32: Fitting model_b

Observations	Residual Std. Error	R^2	Adjusted R^2
5875	9.806	0.1613	0.1602

```r
pander(vif(model_b))
```

df$age	df$Jitter(%)	df$Shimmer.APQ11	df$HNR	df$DFA	df$RPDE
1.123	3.469	3.476	64.63	1.365	1.842

df$PPE	I(HNR^2)
3.464	49.82

Da cui risulta un'intercetta non significativamente diversa da zero. Si vorrebbe accettare l'ipotesi nulla secondo cui l'intercetta è pari a zero e quindi eliminarla, ma per assicurarsi della veridicità del test sui coefficienti bisogna verificare una condizione base di un modello di regressione ovvero la normalità dei residui. Infatti, se questi sono distribuiti normalmente allora anche i coefficienti lo sono ed i test d'ipotesi basati sulla t di Student sono corretti. Inoltre, come mostrato dal VIF l'introduzione del quadrato di una variabile già inserita nel modello porta nuovamente alla collinearità in quanto una è funzione dell'altra, collinearità che in questo caso si risolve centrando HNR^2:

```r
df$HNR_c <- df$HNR - mean(df$HNR)
model_bc <- lm(df$total_UPDRS ~ df$age + df$`Jitter(%)` + df$Shimmer.APQ11 + df$HNR +
    df$DFA + df$RPDE + df$PPE + I(HNR_c^2), data = df)

pander(vif(model_bc))
```

df$age	df$Jitter(%)	df$Shimmer.APQ11	df$HNR	df$DFA	df$RPDE
1.123	3.469	3.476	4.491	1.365	1.842

df$PPE	I(HNR_c^2)
3.464	2.486

```
pander(summary(model_bc), caption = "Fitting model_bc")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	31.39	2.833	11.08	2.986e-28
df$age	0.284	0.01537	18.48	3.747e-74
df$Jitter(%)	106.7	42.37	2.517	0.01186
df$Shimmer.APQ11	31.88	11.94	2.671	0.007594
df$HNR	-0.2237	0.06319	-3.54	0.0004025
df$DFA	-32.9	2.108	-15.61	7.843e-54
df$RPDE	4.268	1.72	2.482	0.0131
df$PPE	12.75	2.603	4.898	9.957e-07
I(HNR_c^2)	-0.05887	0.005164	-11.4	8.688e-30

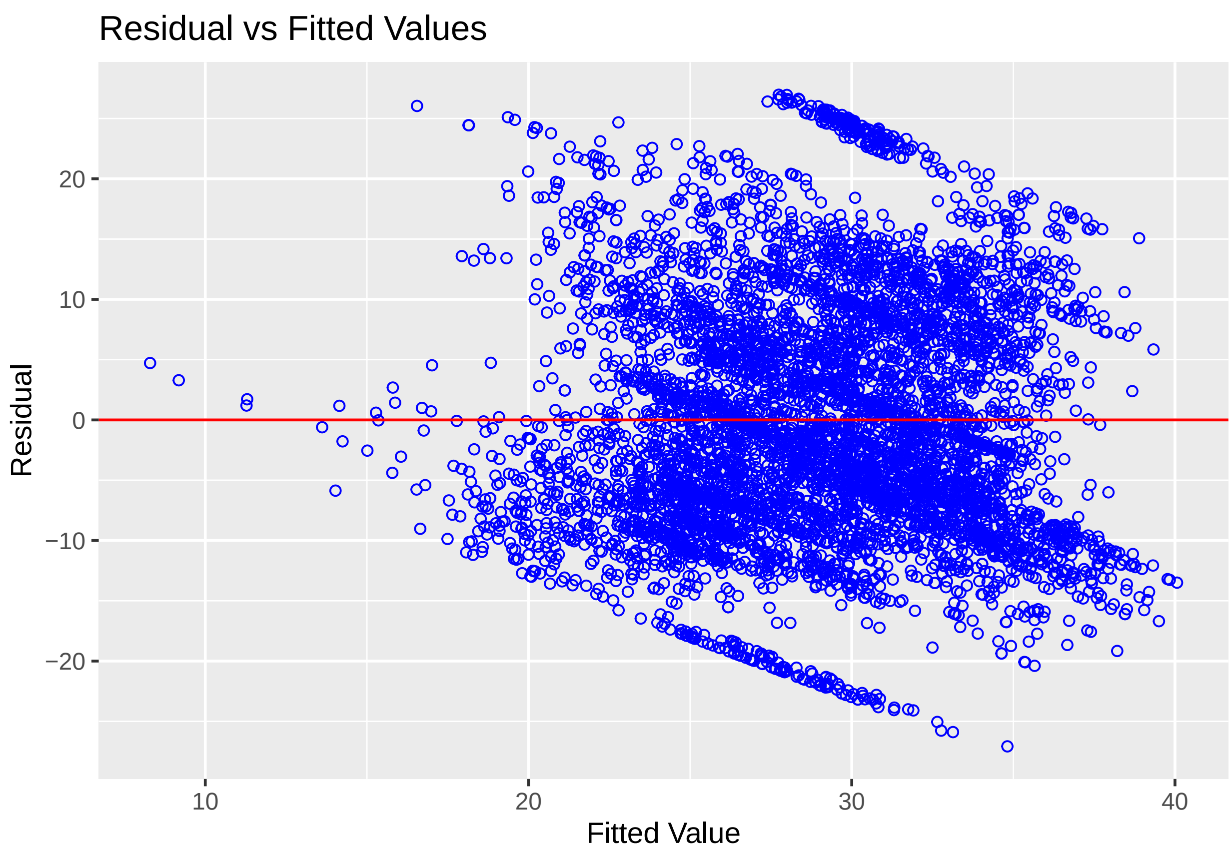
Table 38: Fitting model_bc

Observations	Residual Std. Error	R^2	Adjusted R^2
5875	9.806	0.1613	0.1602

Una volta risolta la multicollinearità l'intercetta risulta significativa e quindi viene mantenuta all'interno del modello.

Si procede a questo punto a verificare un'altra ipotesi generale dei modelli di regressione ovvero l'omoschedasticità dei residui: se questi sono eteroschedasticiti, ovvero non hanno varianza costante, la stima dei coefficienti non risulta corretta.

```
ols_plot_resid_fit(model_bc)
```




```r
ols_test_breusch_pagan(model_bc)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                   Data
##  ------------------------------------------
##  Response : df$total_UPDRS
##  Variables: fitted values of df$total_UPDRS
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    2.502387
##  Prob > Chi2   =    0.1136739
```

La dispersione dei residui mostrata dal grafico fa pensare ad omoschedasticità degli stessi e tale ipotesi è confermata con il test di Breusch-Pagan in quanto per un p-value > 0.05 si accetta l'ipotesi nulla.

Conclusioni

Se si valuta il modello sviluppato fino a questo punto si può concludere dicendo che, considerando un R^2 aggiustato del 16%, non riesce a spiegare la variabilità di total_UPDRS. Questa bassa performance può essere dovuta a due motivi: in primo luogo alla complessità del problema posto in questa sede, che non può certamente essere semplificato in pochi passaggi, ed in secondo luogo all'autocorrelazione dei residui essendo dati panel. Risolvendo quest'ultima potrebbe aumentare la precisione del modello in quanto si andrebbero a stimare i coefficienti con un metodo diverso da quello dei minimi quadrati. Inoltre risultati diversi possono essere raggiunti modificando gli esponenti delle variabili esplicative o introducendo una trasformazione logaritmica.