LINEAR 2 - Data set: AIRQUALITY

INTRODUZIONE

Il data set contiene contiene 154 osservazioni con 6 variabili.

1. OZONO: concentrazioni di Ozono (parti per milione misurata a Roosevelt Island)
2. SOLAR.R: radiazione solare (misurata al Central Park)
3. WIND: velocità media del vento (misurata all'aeroporto LaGuardia)
4. TEMP: temperatura in F (misurata all'aeroporto LaGuardia)
5. MONTH: mese
6. DAY: giorno del mese

Variabile dipendente: TEMP.

Analisi proposte:

1. Statistiche descrittive
2. Regressione lineare e polinomiale

```r
#-- R CODE

library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
panderOptions('knitr.auto.asis', FALSE)

#-- White test function
white.test <- function(lmod,data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM,"P value"=p.value)
}

#-- funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVARION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\F. Esercizi(22) copia\\3.lin(5)\\2.linear\\airquality.txt"),sep="

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- c("Ozone","Solar.R","Wind","Temp")

#-- print delle prime 6 righe del dataset
pander(head(d))
```

id	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7	67	5	1
2	36	118	8	72	5	2
3	12	149	13	74	5	3
4	18	313	12	62	5	4
5	20	178	14	56	5	5
6	28	193	15	66	5	6

STATISTICHE DESCRITTIVE

Si propongono la matrice di correlazione tra le variabili e alcune descrittive di base.

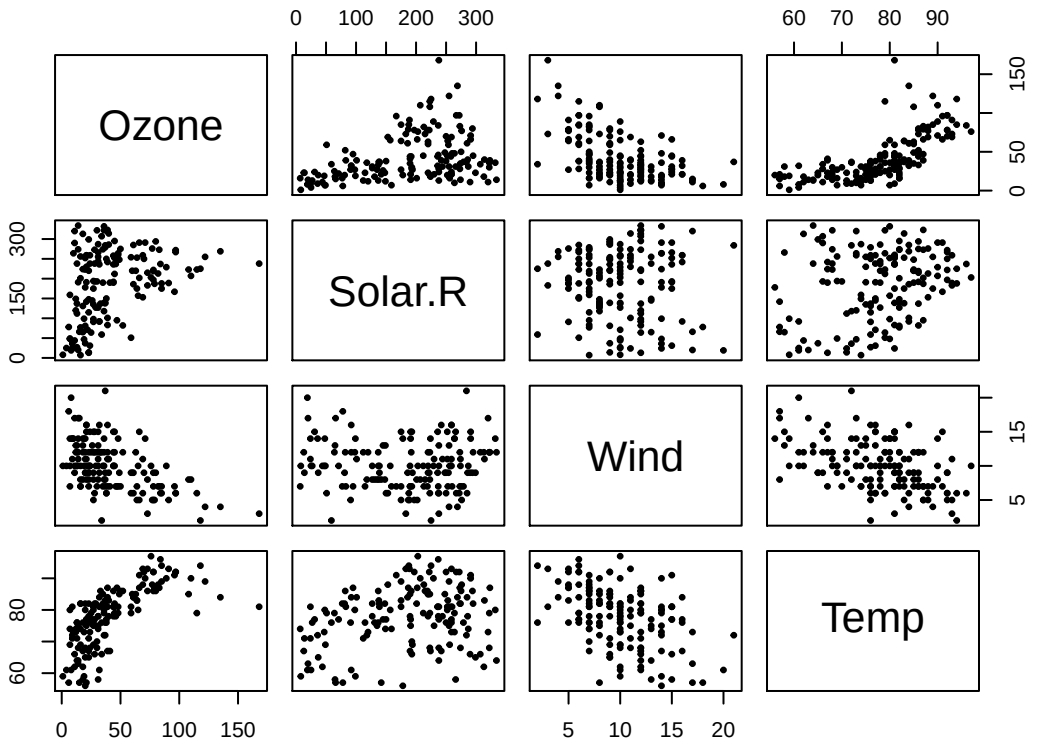
```
#-- R CODE
pander(summary(d[,VAR_NUMERIC])) #-- statistiche descrittive
```

Ozone	Solar.R	Wind	Temp
Min. : 1.00	Min. : 7.0	Min. : 2.00	Min. :56.00
1st Qu.: 20.00	1st Qu.:120.0	1st Qu.: 7.00	1st Qu.:72.00
Median : 33.00	Median :201.0	Median :10.00	Median :79.00
Mean : 41.63	Mean :185.8	Mean :10.02	Mean :77.88
3rd Qu.: 60.00	3rd Qu.:256.0	3rd Qu.:12.00	3rd Qu.:85.00
Max. :168.00	Max. :334.0	Max. :21.00	Max. :97.00

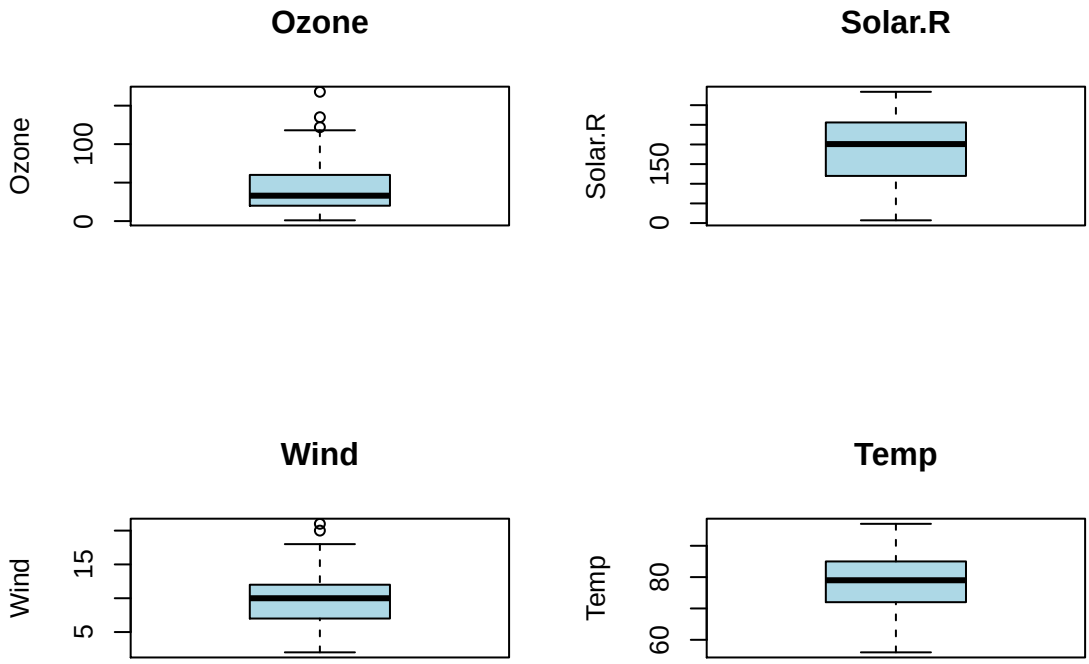
```
pander(cor(d[,VAR_NUMERIC])) #-- matrice di correlazione
```

	Ozone	Solar.R	Wind	Temp
Ozone	1	0.3608	-0.5403	0.6878
Solar.R	0.3608	1	-0.04474	0.2744
Wind	-0.5403	-0.04474	1	-0.4555
Temp	0.6878	0.2744	-0.4555	1

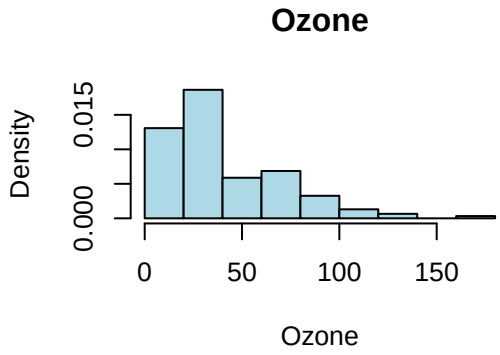
```
plot(d[,VAR_NUMERIC],pch=19,cex=.5) #-- scatter plot multivariato
```

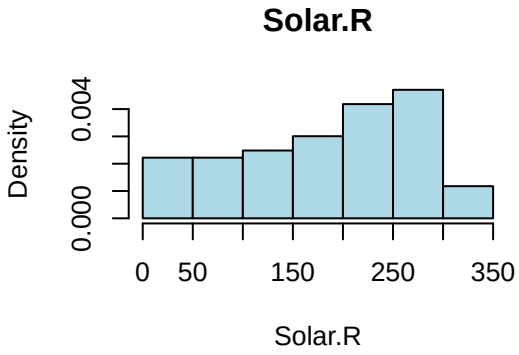


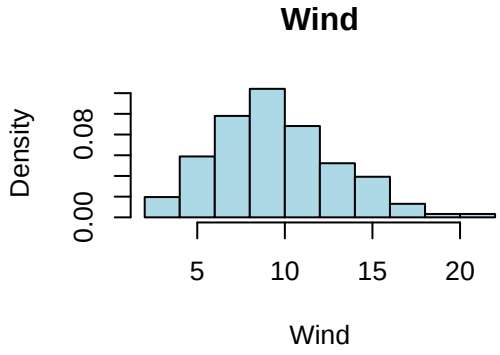

```
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```

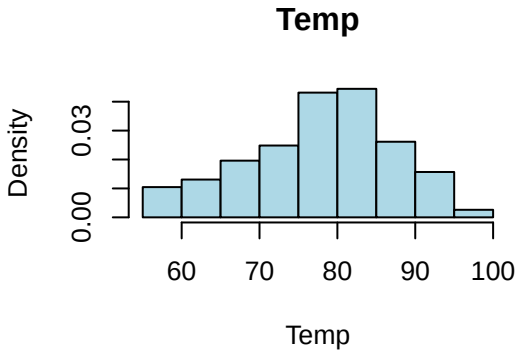



```
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  hist(d[,i],main=i,col="lightblue",xlab=i,freq=F)
}
```









REGRESSIONE

Si analizza la dipendenza di temp da "Ozono" innanzitutto con una regressione lineare.

```
#-- R CODE
mod1 <- lm(Temp~Ozone,d)
pander(summary(mod1),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68.84	0.9561	72	8.637e-119
Ozone	0.2173	0.01866	11.64	9.107e-23

Table 5: Fitting linear model: Temp ~ Ozone

Observations	Residual Std. Error	R^2	Adjusted R^2
153	6.893	0.4731	0.4696

```
pander(anova(mod1),big.mark=",")
```

Table 6: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ozone	1	6,443	6,443	135.6	9.107e-23
Residuals	151	7,175	47.52	NA	NA

```
pander(white.test(mod1),big.mark=",") #-- White test (per dettagli ?bptest)
```

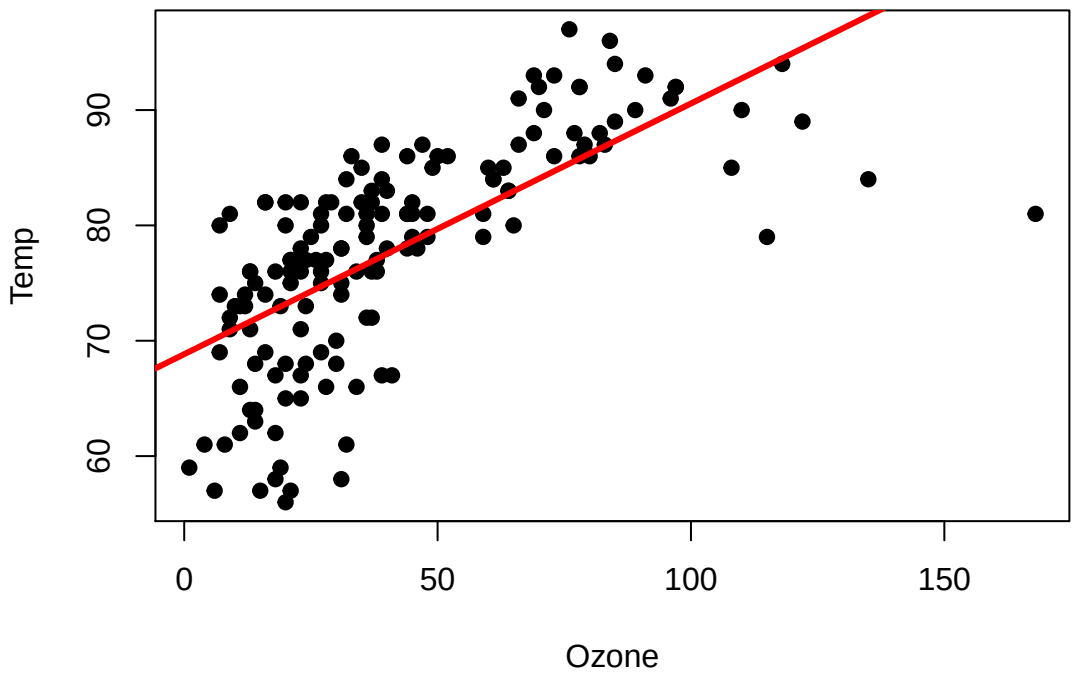
Test.statistic	P.value
47.99	3.793e-11

```
pander(dwtest(mod1),big.mark=",") #-- Durbin-Whatson test
```

Table 8: Durbin-Watson test: `mod1 > >`

Test statistic	P value	Alternative hypothesis
1.04	8.425e-10 * * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(d$Ozone,d$Temp,pch=19,xlab="Ozone",ylab="Temp")
abline(mod1,col=2,lwd=3) #-- abline del modello lineare
```



Il modello ha un fitting buono ma non elevatissimo ($R^2 = 0.47$) e "Ozono" è significativo. Tuttavia l'ipotesi di incorrelazione è respinta così come l'omoschedasticità. Si prova ora con polinomi di grado superiore (2, 3, 4).

```
#-- R CODE
mod2 <- lm(Temp~Ozone+I(Ozone^2),d)
pander(summary(mod2),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.75	1.281	48.98	3.576e-94
Ozone	0.5184	0.05021	10.33	3.249e-19
I(Ozone^2)	-0.002457	0.0003867	-6.355	2.372e-09

Table 10: Fitting linear model: Temp ~ Ozone + I(Ozone^2)

Observations	Residual Std. Error	R^2	Adjusted R^2
153	6.139	0.5849	0.5794

```
pander(anova(mod2),big.mark=",")
```

Table 11: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ozone	1	6,443	6,443	171	1.479e-26
I(Ozone^2)	1	1,522	1,522	40.38	2.372e-09
Residuals	150	5,653	37.69	NA	NA

```
pander(white.test(mod2),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
11.21	0.003674

```
pander(dwtest(mod2),big.mark=",") #-- Durbin-Whatson test
```

Table 13: Durbin-Watson test: `mod2`

Test statistic	P value	Alternative hypothesis
1.083	4.351e-09 * * *	true autocorrelation is greater than 0

Il modello polinomiale di ordine 2 ha un fitting migliore ($R^2 = 0.5849$) e i parametri relativi a "Ozono" e anche a $Ozono^2$ sono significativi. Il valore negativo del parametro segnala che la concavità è verso il basso. Si prova a verificare ora se sia opportuno utilizzare il modello polinomiale di ordine 3 e 4.

```r
#-- R CODE
mod3 <- lm(Temp~Ozone+I(Ozone^2)+I(Ozone^3),d)
pander(summary(mod3),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.94	1.871	34.17	2.239e-72
Ozone	0.429	0.1141	3.76	0.0002439
I(Ozone^2)	-0.0009198	0.001804	-0.5098	0.611
I(Ozone^3)	-6.882e-06	7.889e-06	-0.8724	0.3844

Table 15: Fitting linear model: Temp ~ Ozone + I(Ozone^2) + I(Ozone^3)

Observations	Residual Std. Error	R^2	Adjusted R^2
153	6.144	0.587	0.5787

```r
pander(anova(mod3),big.mark=",")
```

Table 16: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ozone	1	6,443	6,443	170.7	1.771e-26
I(Ozone^2)	1	1,522	1,522	40.32	2.468e-09
I(Ozone^3)	1	28.73	28.73	0.7611	0.3844
Residuals	149	5,624	37.75	NA	NA

```r
pander(white.test(mod3),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
11.27	0.003571

```r
pander(dwtest(mod3),big.mark=",") #-- Durbin-Whatson test
```

Table 18: Durbin-Watson test: `mod3` > >

Test statistic	P value	Alternative hypothesis
1.091	5.359e-09 * * *	true autocorrelation is greater than 0

```r
#-- R CODE
mod4 <- lm(Temp~Ozone+I(Ozone^2)+I(Ozone^3)+I(Ozone^4),d)
pander(summary(mod4),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.59	2.548	26.13	2.626e-57

	Estimate	Std. Error	t value	Pr(>|t|)
Ozone	0.1376	0.2227	0.618	0.5375
I(Ozone^2)	0.007696	0.005941	1.296	0.1972
I(Ozone^3)	-9.549e-05	5.876e-05	-1.625	0.1063
I(Ozone^4)	2.842e-07	1.868e-07	1.522	0.1302

Table 20: Fitting linear model: Temp ~ Ozone + I(Ozone^2) + I(Ozone^3) + I(Ozone^4)

Observations	Residual Std. Error	R^2	Adjusted R^2
153	6.117	0.5934	0.5824

```
pander(anova(mod4),big.mark=",")
```

Table 21: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ozone	1	6,443	6,443	172.2	1.401e-26
I(Ozone^2)	1	1,522	1,522	40.67	2.169e-09
I(Ozone^3)	1	28.73	28.73	0.7678	0.3823
I(Ozone^4)	1	86.63	86.63	2.315	0.1302
Residuals	148	5,538	37.42	NA	NA

```
pander(white.test(mod4),big.mark=",") #-- White test (per dettagli ?bptest)
```

Test.statistic	P.value
14.79	0.0006153

```
pander(dwtest(mod4),big.mark=",") #-- Durbin-Whatson test
```

Table 23: Durbin-Watson test: `mod4`

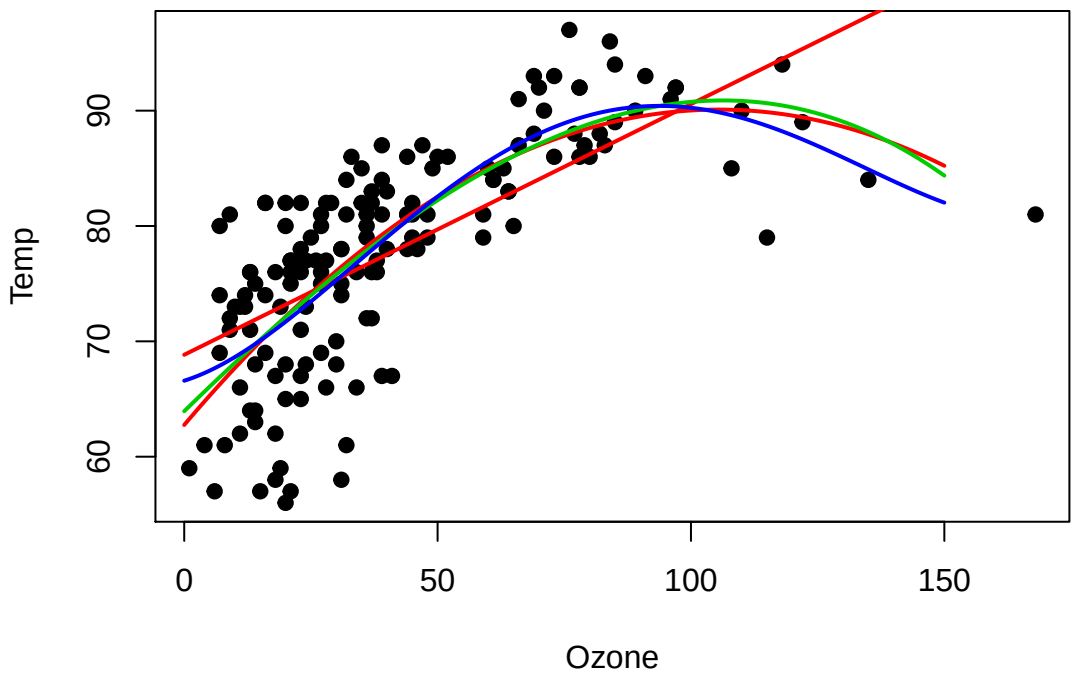
Test statistic	P value	Alternative hypothesis
1.097	6.118e-09 * * *	true autocorrelation is greater than 0

Il fitting del modelli polinomiale di ordine 3 migliora leggermente, ma solo il parametro relativo a Ozono risulta significativo; il modello polinomiale di ordine 4 migliora ancora un po' il fitting ma nessun parametro è significativo.

```
#-- R CODE
plot(d$Ozone,d$Temp,pch=19,xlab="Ozone",ylab="Temp")

lines(seq(0,150,0.1),predict(mod1,data.frame(Ozone=seq(0,150,0.1))),col=2,lwd=2)
#abline(mod1,col=2,lwd=3) #-- abline del modello lineare; graficamente è la stessa cosa della riga sopr
```

```
lines(seq(0,150,0.1),predict(mod2,data.frame(Ozone=seq(0,150,0.1))),col=2,lwd=2)
lines(seq(0,150,0.1),predict(mod3,data.frame(Ozone=seq(0,150,0.1))),col=3,lwd=2)
lines(seq(0,150,0.1),predict(mod4,data.frame(Ozone=seq(0,150,0.1))),col=4,lwd=2)
```



Si prova ora a verificare l'opportunità di usare un modello lin-log che utilizza il logaritmo dell'ozono come variabile esplicativa.

```
#-- R CODE
mod5 <- lm(Temp~I(log(Ozone)),d)
pander(summary(mod5),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	47.79	2.421	19.73	1.179e-43
I(log(Ozone))	8.69	0.6821	12.74	1.036e-25

Table 25: Fitting linear model: Temp ~ I(log(Ozone))

Observations	Residual Std. Error	R^2	Adjusted R^2
153	6.592	0.5181	0.5149

```
pander(anova(mod5),big.mark=",")
```

Table 26: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(Ozone))	1	7,055	7,055	162.3	1.036e-25
Residuals	151	6,563	43.46	NA	NA

```
pander(white.test(mod5),big.mark=",") #-- White test (per dettagli ?bptest)
```

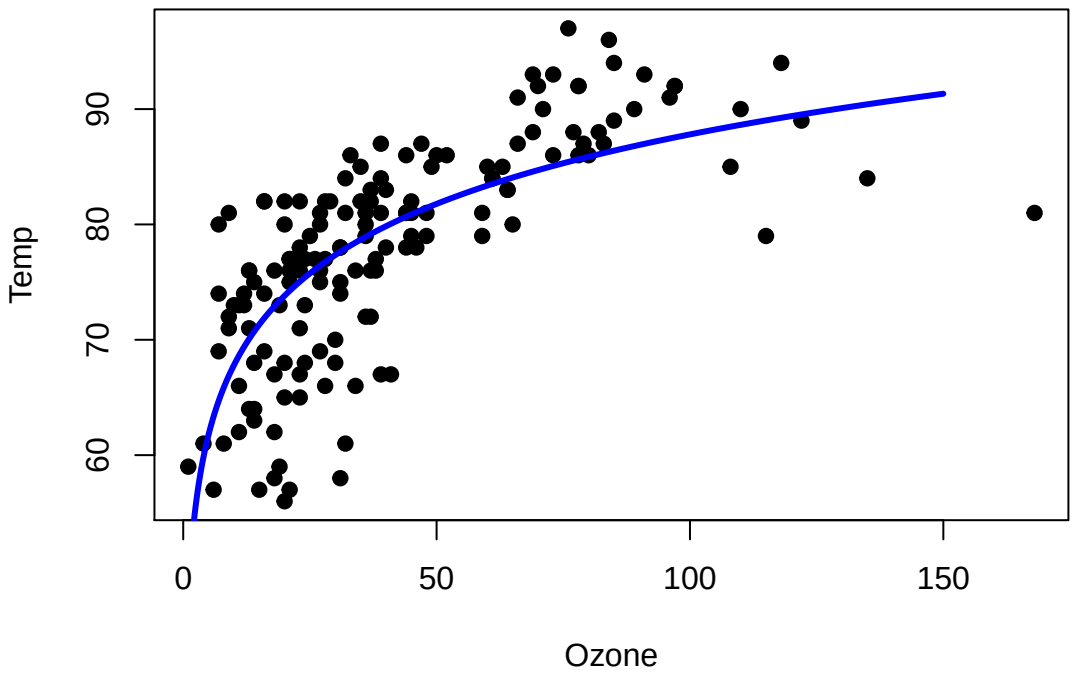
Test.statistic	P.value
5.776	0.05569

```
pander(dwtest(mod5),big.mark=",") #-- Durbin-Whatson test
```

Table 28: Durbin-Watson test: `mod5` > >

Test statistic	P value	Alternative hypothesis
0.9854	9.757e-11 * * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(d$Ozone,d$Temp,pch=19,xlab="Ozone",ylab="Temp",main="")
lines(seq(0,150,0.1),predict(mod5,data.frame(Ozone=seq(0,150,0.1))),col="blue",lwd=3)
```



Il fitting è peggiore ($R^2 = 0.51$), $log(Ozono)$ è significativo, ma è respinta l'ipotesi di non correlazione fra gli errori e anche a riguardo della omoschedaticità. Si propone quindi il modello log-lin in cui la variabile dipendente è $log(Temp)$:

```
#-- R CODE
mod6 <- lm(I(log(Temp))~Ozone,d)
pander(summary(mod6),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.231	0.01313	322.1	3.652e-216
Ozone	0.002804	0.0002564	10.94	7.171e-21

Table 30: Fitting linear model: I(log(Temp)) ~ Ozone

Observations	Residual Std. Error	R^2	Adjusted R^2
153	0.09469	0.442	0.4383

```
pander(anova(mod6),big.mark=",")
```

Table 31: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ozone	1	1.072	1.072	119.6	7.171e-21
Residuals	151	1.354	0.008965	NA	NA

```
pander(white.test(mod6),big.mark=",") #-- White test (per dettagli ?bptest)
```

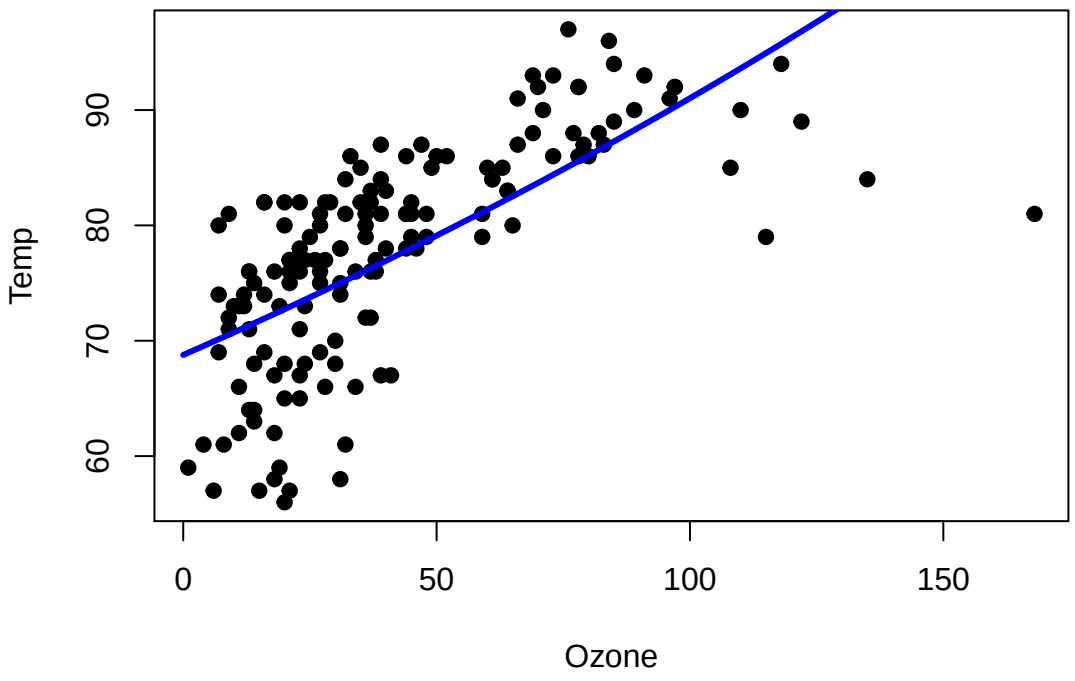
Test.statistic	P.value
37.97	5.688e-09

```
pander(dwtest(mod6),big.mark=",") #-- Durbin-Whatson test
```

Table 33: Durbin-Watson test: `mod6 > >`

Test statistic	P value	Alternative hypothesis
1.005	2.158e-10 * * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(d$Ozone,d$Temp,pch=19,xlab="Ozone",ylab="Temp",main="")
lines(seq(0,150,0.1),exp(predict(mod6,data.frame(Ozone=seq(0,150,0.1)))),col="blue",lwd=3) #-- notare es
```



Il parametro relativo a ozono è significativo ma il fitting peggiora ancora e vale quanto detto per il modello lin log per ciò che concerne la sfericità degli errori. Si propone ora il modello log-log che studia la dipendenza di $log(Temp)$ da $log(Ozono)$.

```
#-- R CODE
mod7 <- lm(I(log(Temp))~I(log(Ozone)),d)
pander(summary(mod7),big.mark=",")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.953	0.03292	120.1	9.876e-152
I(log(Ozone))	0.1139	0.009274	12.29	1.716e-24

Table 35: Fitting linear model: I(log(Temp)) ~ I(log(Ozone))

Observations	Residual Std. Error	R^2	Adjusted R^2
153	0.08963	0.5	0.4966

```
pander(anova(mod7),big.mark=",")
```

Table 36: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(log(Ozone))	1	1.213	1.213	151	1.716e-24
Residuals	151	1.213	0.008034	NA	NA

```
pander(white.test(mod7),big.mark=",") #-- White test (per dettagli ?bptest)
```

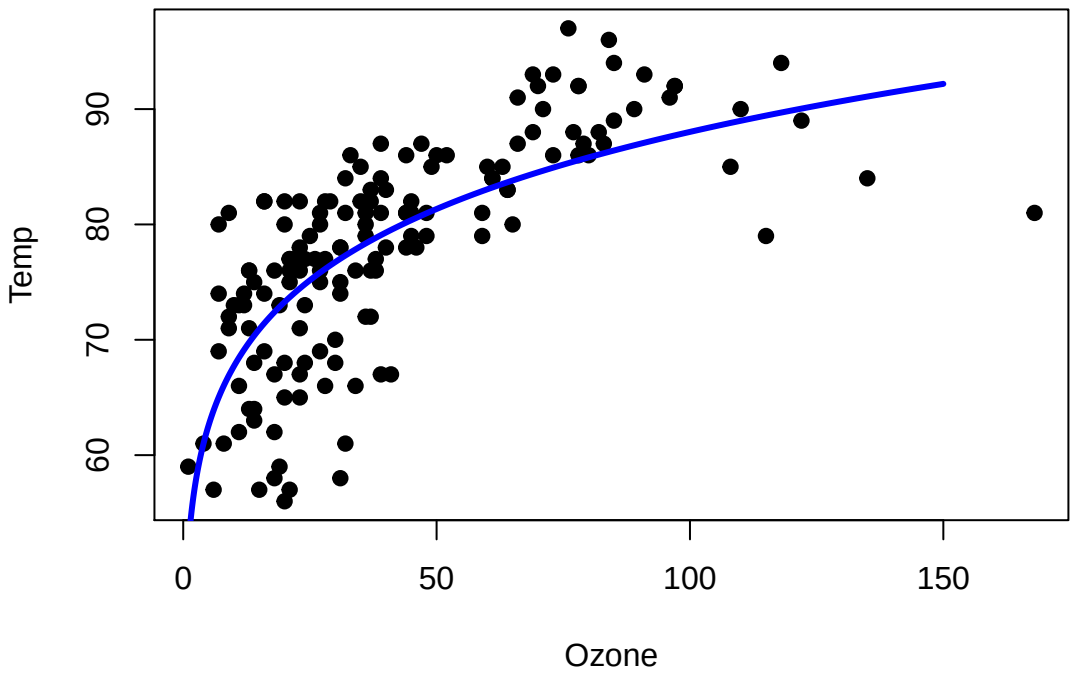
Test.statistic	P.value
7.926	0.01901

```
pander(dwtest(mod7),big.mark=",") #-- Durbin-Whatson test
```

Table 38: Durbin-Watson test: `mod7`

Test statistic	P value	Alternative hypothesis
0.9662	4.448e-11 * * *	true autocorrelation is greater than 0

```
#-- R CODE
plot(d$Ozone,d$Temp,pch=19,xlab="Ozone",ylab="Temp",main="")
lines(seq(0,150,0.1),exp(predict(mod7,data.frame(Ozone=seq(0,150,0.1)))),col="blue",lwd=3) #-- notare es
```



$Log(Ozono)$ è significativo ma il fitting peggiora ancora e inoltre viene respinta sia l'ipotesi di omoschedasticità che quella di non correlazione fra i residui. In definitiva il modello prescelto è il modello quadratico che però necessiterebbe di verifica della sfericità degli errori.