FOUNDATIONS OF STATISTICAL MODELING

AIRQUALITY: NEW YORK AIRQUALITY MEASUREMENTS IN R

Mohamed Salah Jebali

Matricola: 7078487

Scuola di Ingegneria - Corso di Laurea Magistrale in Intelligenza Artificiale

Università degli Studi di Firenze

Febbraio 2022

Indice

Keywords Regressione lineare multipla · Regressione polinomiale · Regressione logistica · Modelli grafici

1 Introduzione

In questo lavoro si è voluto svolgere l'analisi del dataset **"Airquality"** utilizzando i principali concetti di modellazione statistica appresi durante il corso di *Foundations of Statistical Modeling*.

1.1 Obiettivo dell'analisi

Tra le variabili contenute nel dataset, quella che influenza la qualità dell'aria è la concentrazione di molecole di Ozono contenute nell'atmosfera. L'Ozono è un gas dotato di un elevato potere ossidante, di colore azzurro e dall'odore pungente. L'Ozono è un inquinante molto tossico per l'uomo e, oltre ad essere un irritante per tutte le membrane mucose, un'esposizione critica e prolungata può causare tosse, mal di testa e perfino edema polmonare. Infatti, l'Organizzazione Mondiale per la Sanità (O.M.S.), al fine di ridurre il pericolo di danni acuti e cronici e per assicurare un ulteriore margine di sicurezza, raccomanda come livello limite di esposizione, oltre la quale vi è un rischio per la salute umana, 120 μg /m^3.

Per questo motivo, date le premesse appena fatte, l'analisi del dataset si pone l'obiettivo di studiare gli effetti delle variabili 'Solar.R', 'Wind' e 'Temp' sui livelli di Ozono nell'aria, con lo scopo di trovare il modello più adatto che possa rispondere alle seguenti domande:

- da quali variabili è influenzato il livello di concentrazione di Ozono nell'aria?
- come, queste variabili, influenzano il livello di Ozono?
- qual è la probabilità che, dati certi valori assunti dalle variabili esplicative 'Solar.R', 'Wind' e 'Temp', l'aria sia di scarsa qualità?
- che relazioni ci sono tra le variabili esplicative stesse e come si influenzano?

Prima di procedere con lo studio vero e proprio dei modelli più adatti, prima di regressione e poi di classificazione, abbiamo voluto dare una panoramica sul data set in oggetto.

1.2 Analisi della struttura del data set

Il data set Airquality contiene le misurazioni della qualità dell'aria giornaliere della città di New York raccolte tra Maggio e Settembre del 1973.

Il data frame contiene allo stato iniziale 153 osservazioni e 6 variabili, com'è possibile osservare in Tabella 1

Airquality		
Variabile	Tipologia	Unità di misura
Ozone	int	ppb
Solar.R	int	lang
Wind	numeric	mph
Temp	int	Fahrenheit
Month	numeric	month
Day	numeric	day

Tabella 1: Struttura iniziale del data frame

Per prima cosa vengono omessi i valori NA e si considerano soltanto le prime quattro variabili del data frame, ottenendo così 111 osservazioni e 4 variabili.

```
data(airquality)
aria = na.omit(airquality)
data = aria[c(1,2,3,4)]
str(data)

## 'data.frame': 111 obs. of  4 variables:
```

```
##  $ Ozone  : int  41 36 12 18 23 19 8 16 11 14 ...
##  $ Solar.R: int  190 118 149 313 299 99 19 256 290 274 ...
##  $ Wind   : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
##  $ Temp   : int  67 72 74 62 65 59 61 69 66 68 ...

head(data)

##   Ozone Solar.R Wind Temp
## 1    41     190  7.4   67
## 2    36     118  8.0   72
## 3    12     149 12.6   74
## 4    18     313 11.5   62
## 7    23     299  8.6   65
## 8    19      99 13.8   59
```

Dopodiché procediamo con la conversione delle unità di misura nel Sistema Internazionale di misura, convertendo i ppb in μg /m^3, i lang in J/m^2, le mph a m/s e infine i Fahrenheit in gradi Celsius.

```
data$Ozone <- data$Ozone * 2
data$Solar.R <- data$Solar.R * 41840
data$Wind <- round(data$Wind / 2.24, 3)
data$Temp <- round((data$Temp - 32) / 1.8, 3)

head(data)

##   Ozone  Solar.R  Wind   Temp
## 1    82  7949600 3.304 19.444
## 2    72  4937120 3.571 22.222
## 3    24  6234160 5.625 23.333
## 4    36 13095920 5.134 16.667
## 7    46 12510160 3.839 18.333
## 8    38  4142160 6.161 15.000
```

Dunque, il data frame su cui è stato svolto il lavoro è stato nominato **data** e le variabili hanno le seguenti caratteristiche mostrate in Tabella 2

data		
Variabile	Tipologia	Unità di misura
Ozone	int	μg /m^3
Solar.R	int	J/m^2
Wind	numeric	m/s
Temp	int	Celsius

Tabella 2: struttura del data frame dopo il cleaning

2 Regressione lineare multipla

Lo studio iniziale ha previsto la ricerca di un modello di regressione lineare multipla che potesse adattare correttamente i dati. Le variabili coinvolte sono:

- Y: 'Ozone', ovvero la concentrazione dell'Ozono nell'aria (in μg /m^3)
- X_1: 'Solar.R', ovvero la quantità di energia solare per unità di superficie (in J/m^2)
- X_2: 'Wind', ovvero la velocità del vento (in m/s)
- X_3: 'Temp', ovvero la temperatura dell'atmosfera (in gradi Celsius)

2.1 Analisi preliminare dei plot

Da un'analisi preliminare del plot in Figura 1, possiamo fare delle supposizioni iniziali di natura qualitativa.

```
attach(data)
plot(data)
```

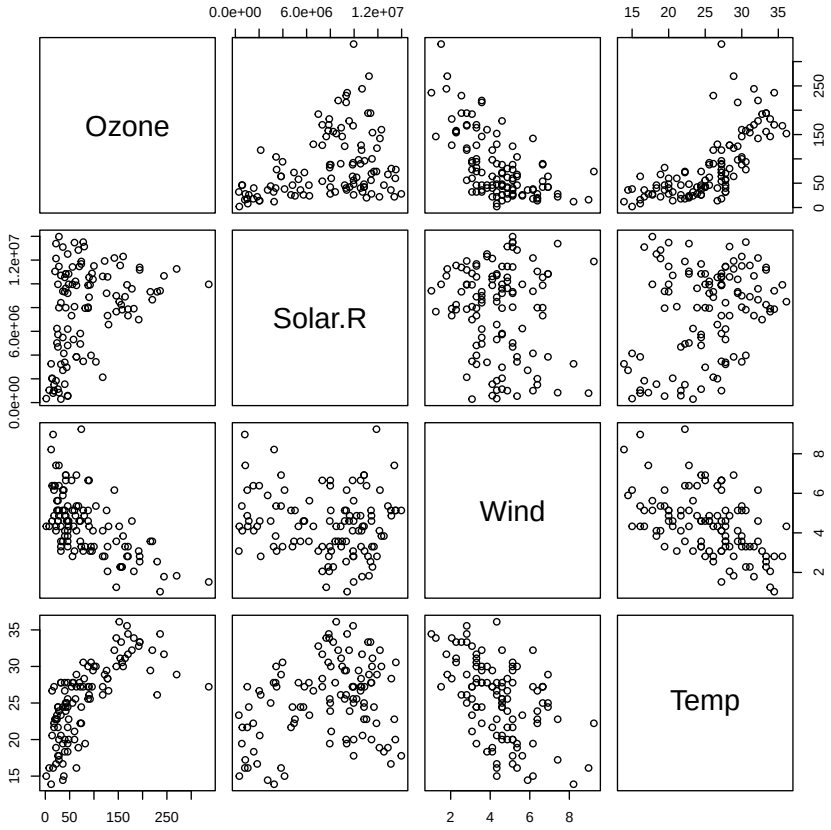


Figura 1: Plot del data frame data

Consideriamo la parte superiore della matrice Plot 1, ovvero quella dove lungo l'asse delle ordinate è presente il valore di 'Ozone', e in quella delle ascisse, rispettivamente, 'Solar.R', 'Wind' e 'Temp':

- Ozone - Solar.R: sembrerebbe esserci una correlazione di proporzionalità diretta tra i livelli di 'Solar.R' e la concentrazione di Ozono nell'aria.

- Ozone - Wind: sembrerebbe esserci una correlazione di proporzionalità indiretta tra la velocità del vento e 'Ozone'.
- Ozone - Temp: sembrerebbe esserci una correlazione di proporzionalità diretta tra la temperatura dell'atmosfera e la concentrazione di Ozono in essa.

Dopo un'analisi preliminare qualitativa, vogliamo verificare analiticamente le supposizioni fatte in prima fase. Nella prossima sezione ci occuperemo di analizzare i modelli di regressione lineare multipla che meglio si adattano ai dati.

2.2 Analisi dei modelli lineari

In questa sezione cercheremo il modello di regressione lineare multipla migliore, procedendo con una strategia "backward", partendo dal modello completo e rimuovendo le variabili che risultano essere meno significative in base al p-value. Testeremo anche i modelli dipendenti da singole variabili.

Modello Completo Consideriamo il modello di regressione lineare completo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{1}$$

```
mq <- lm(Ozone ~ Solar.R + Wind + Temp)
summary(mq)

##
## Call:
## lm(formula = Ozone ~ Solar.R + Wind + Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.976 -28.439  -7.097  20.196 191.245
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.296e+01  3.129e+01  -0.734   0.4647
## Solar.R      2.860e-06  1.108e-06   2.580   0.0112 *
## Wind        -1.493e+01  2.932e+00  -5.094 1.52e-06 ***
## Temp         5.948e+00  9.127e-01   6.516 2.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.36 on 107 degrees of freedom
## Multiple R-squared:  0.6059,Adjusted R-squared:  0.5948
## F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

Commento:

- la variabile 'Solar.R' ha un **effetto positivo** sui livelli di Ozono nell'aria: per ogni unità di 'Solar.R', 'Ozone' aumenta di 0.00000286 μg /m^3 a parità di altre covariate;
- la variabile 'Wind' ha un **effetto negativo** sui livelli di Ozono nell'aria: per ogni unità di 'Wind', 'Ozone' diminuisce di 14.930 μg /m^3 a parità di altre covariate;
- la variabile 'Temp' ha un **effetto positivo** sui livelli di Ozono nell'aria: per ogni unità di 'Temp', 'Ozone' aumenta di 5.948 μg /m^3 a parità di altre covariate;
- le variabili 'Wind' e 'Temp' sono altamente significative;

- la variabile 'Solar.R' risulta essere significativa ma meno delle altre due;
- la variabile 'Wind' ha un effetto, in termini assoluti, molto più grande delle altre due variabili, mentre 'Solar.R' ha decisamente un basso impatto sulla variabile obiettivo;
- la stima della deviazione standard è 42.36;
- L'indice R^2 aggiustato è 0.5948, moderatamente buono;
- La statistica F è altamente significativa, significa che la devianza spiegata dal modello $SS_{\hat{Y}}$ è significativamente superiore al quadrato dei residui SS_e.

Procediamo adesso con l'analisi dei residui del modello di regressione completo, per valutare l'ipotesi di distribuzione normale degli errori. Ne rappresentiamo il qqplot in Figura 2 che confronta la distribuzione empirica dei residui con i quantili della distribuzione normale. Dall'analisi del grafico, l'ipotesi di normalità sembra confermata. Tuttavia, ci sono valori estremi che sembrano allontanarsi dalla retta bisettrice del primo quadrante, facendo pensare ad una possibile regressione polinomiale. Questo particolare caso verrà trattato in una sezione a sé, più avanti.

```
res <- mq$residuals
qqnorm(res)
```

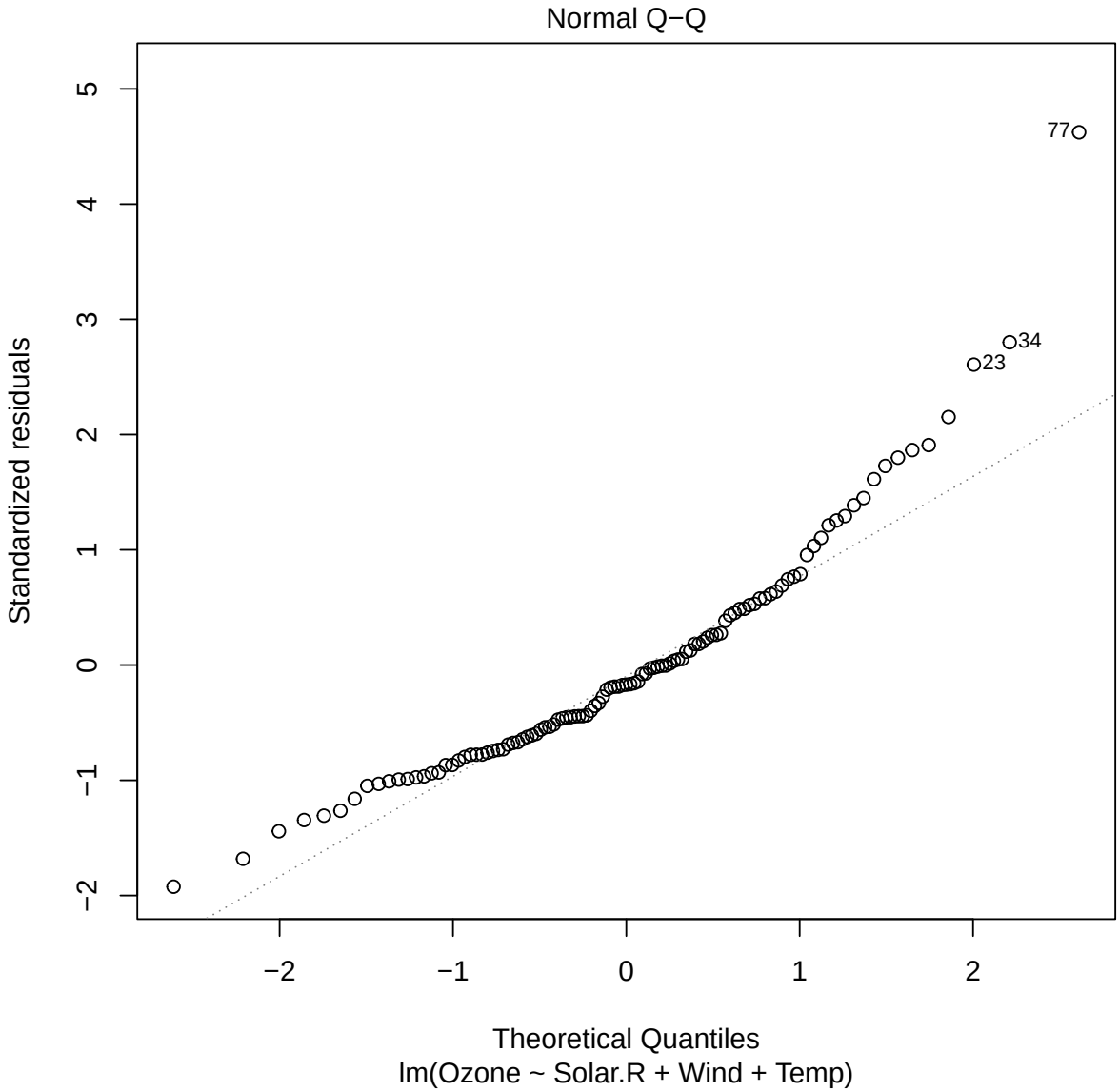


Figura 2: qqplot dei residui del modello di regressione lineare completo

Consideriamo gli intervalli di confidenza separati e congiunti per i coefficienti associati a tutte e tre le variabili esplicative

```
confint(mq)
```

```
##                     2.5 %        97.5 %
## (Intercept) -8.498755e+01  3.907182e+01
## Solar.R      6.624245e-07  5.056814e-06
## Wind        -2.074525e+01 -9.121861e+00
## Temp         4.138262e+00  7.756958e+00
```

Gli intervalli di confidenza al loro interno non presentano il valore zero, seppure quelli di riferimento ai coefficienti di 'Solar.R' sono molto bassi e vicini allo zero. Poiché i test precedenti sono risultati altamente significativi, facendoci rigettare l'ipotesi nulla, e all'interno degli intervalli di confidenza non è presente il valore nullo, possiamo concludere che c'è **coerenza** tra il test di ipotesi e l'intervallo di confidenza.

Modello ridotto 'Wind' + 'Temp' Proviamo il modello senza la variabile 'Solar.R', ovvero X_1, poiché dall'analisi precedente risultava essere la meno significativa in base al $p_{oss} = 0.0112$. Inoltre è quella che in termini assoluti, sembra apportare un contributo minore alla variabile obiettivo. Consideriamo, dunque, il seguente modello:

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{2}$$

```
mq1 <- lm(Ozone ~ Wind + Temp)
summary(mq1)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind + Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.305 -26.435  -6.243  21.203 196.992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.689     32.031  -0.552    0.582
## Wind         -14.760      3.007  -4.909 3.27e-06 ***
## Temp           6.579      0.902   7.294 5.29e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.46 on 108 degrees of freedom
## Multiple R-squared:  0.5814,Adjusted R-squared:  0.5736
## F-statistic: 74.99 on 2 and 108 DF,  p-value: < 2.2e-16
```

Commento:

- la variabile 'Wind' continua ad avere un **effetto negativo** sulla variabile obiettivo e, in termini assoluti, la stima del suo parametro sembra variato di poco, come il suo errore standard.
- la variabile 'Temp' continua ad avere un **effetto positivo** sulla variabile 'Ozone' e la stima del suo parametro sembra variato di poco, come il suo errore standard.
- entrambe le variabili risultano altamente significative, e l'indice del p_{oss} della variabile 'Temp' è addirittura ridotto;

- la stima della deviazione standard è 43.46, simile a quella del modello completo;
- gli indici R^2 sono pressoché invariati;
- la statistica F risulta altamente significativa, e leggermente aumentata di valore.

Passiamo all'analisi dei residui. Analizzando il qqplot in Figura 3 l'ipotesi di normalità sembra confermata per le osservazioni centrali, ma come detto precedentemente, per le osservazioni estreme sembra esserci un discostamento dalla Normale. Effettueremo ulteriori studi nella sezione successiva per valutare l'ipotesi di regressione polinomiale.

```
res1 <- mq1$residuals
qqnorm(res1)
```

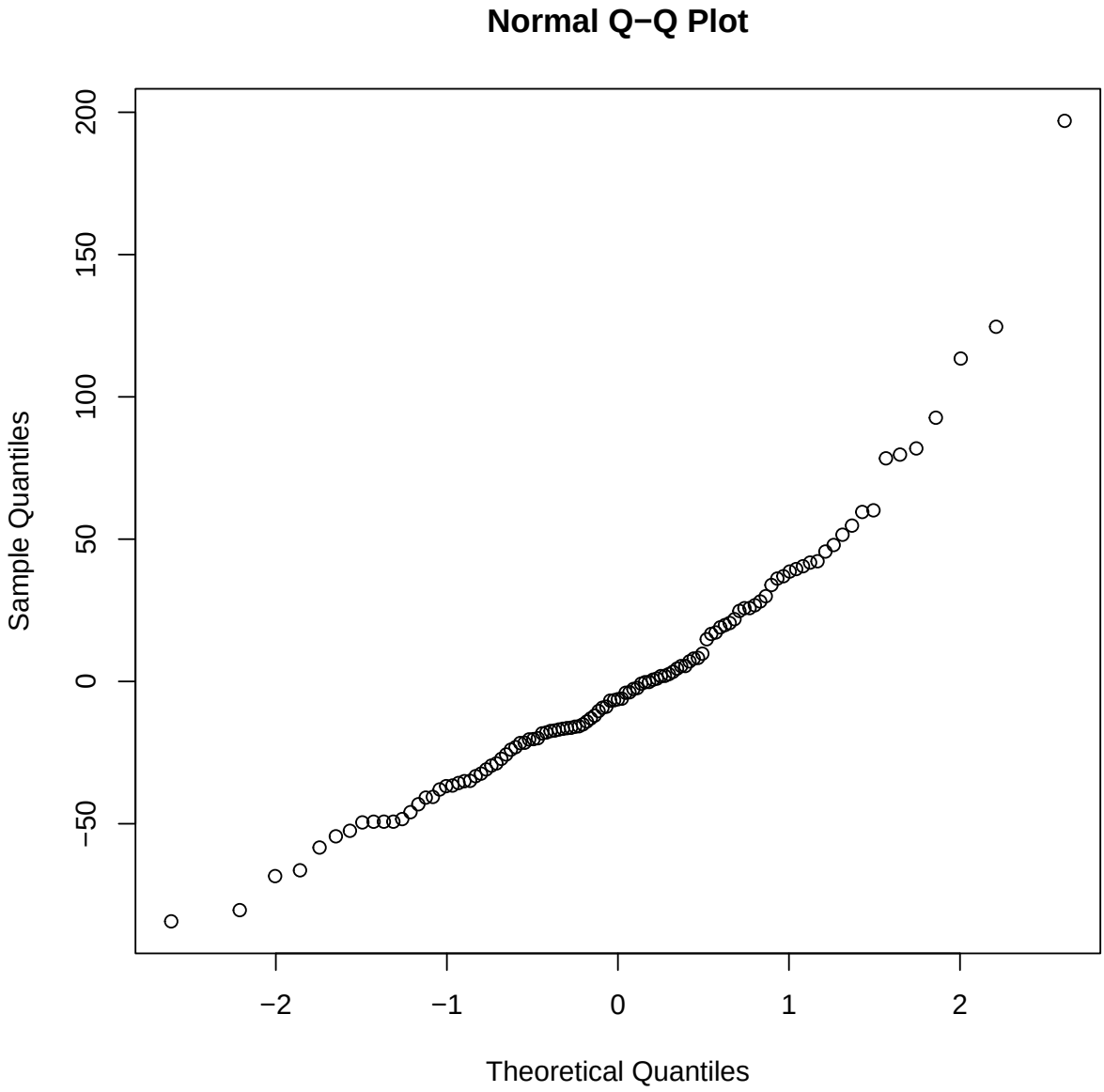


Figura 3: qqplot dei residui del modello di regressione lineare ridotto

Consideriamo gli intervalli di confidenza separati e congiunti per i coefficienti associati a tutte e tre le variabili esplicative.

```
confint(mq1)

##                   2.5 %    97.5 %
## (Intercept) -81.180305 45.801728
## Wind        -20.719523 -8.800034
## Temp          4.791428  8.367232
```

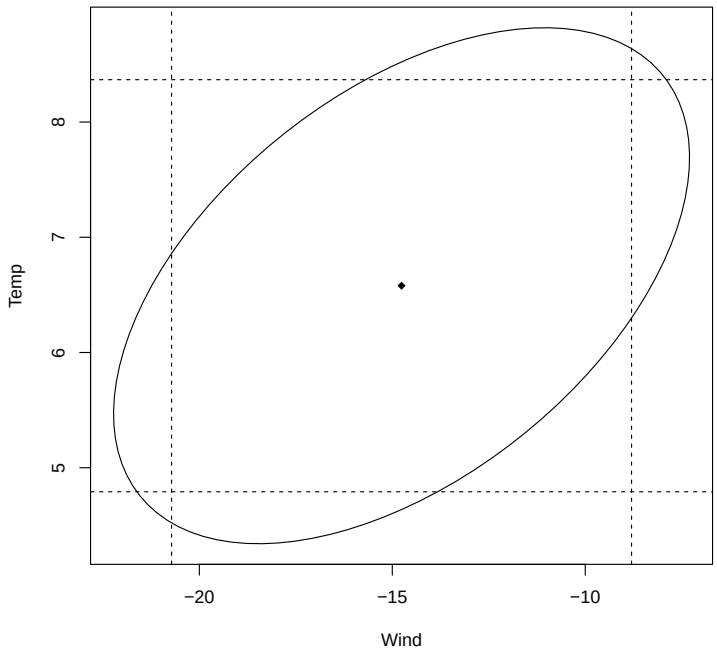
Gli intervalli di confidenza non contengono il valore nullo e dunque possiamo assumere che ci sia un buon livello di **coerenza** tra il test di ipotesi e l'intervallo di confidenza. Tuttavia, l'ampiezza dei singoli intervalli di confidenza è aumentat rispetto a quelli riferenti al modello completo, ciò significa che la stima dei parametri è **meno informativa**.

Costruiamo ora l'ellissoide di confidenza per gli stessi parametri (β_2, β_3)

```
library(ellipse)

t1 <- qt(0.975, mq1$df)
beta1 <- mq1$coeff[c(2,3)]
std.beta1 <- summary(mq1)$coeff[c(2,3),2]
inf1 <- beta1 - t1*std.beta1
sup1 <- beta1 + t1*std.beta1

plot(ellipse(mq1, c(2,3)), type="l")
abline(v=c(inf1[1], sup1[1]), lty=2)
abline(h=c(inf1[2], sup1[2]), lty=2)
points(beta1[1], beta1[2], pch=18)
```



Non contiene il valore nullo e qundi si conferma l'ipotesi di coerenza.

Modello ridotto 'Temp' I modelli testati precendetemente risultano buoni e adattano bene i dati. Tuttavia, per fare una valutazione più accurata, valuto anche i modelli dipendenti dalle singole variabili 'Temp', 'Wind' o 'Solar.R'. Inizio con il modello dipendente da 'Temp', poiché è la variabile più significativa. Consideriamo il seguente modello:

$$Y = \beta_0 + \beta_3 X_3 + \epsilon \tag{3}$$

```
mq2 <- lm(Ozone ~ Temp)
summary(mq2)

##
## Call:
## lm(formula = Ozone ~ Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.842 -34.920  -1.745  20.888 236.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -139.1869    22.3838  -6.218 9.48e-09 ***
## Temp           8.7807     0.8616  10.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.84 on 109 degrees of freedom
## Multiple R-squared:  0.488,Adjusted R-squared:  0.4833
## F-statistic: 103.9 on 1 and 109 DF,  p-value: < 2.2e-16

confint(mq2)

##                   2.5 %    97.5 %
## (Intercept) -183.550985 -94.82289
## Temp           7.073138  10.48830
```

Commento:

- l'effetto della variabile 'Temp' si conferma essere positiva e leggermente maggiore in termini assoluti rispetto ai modelli precedenti;
- si conferma altamente significativa;
- gli intervalli di confidenza sono buoni e non contengono il valore nullo, indice di coerenza;
- gli indici R^2 sono più bassi, mentre la statistica F risulta altamente significatica;
- la statistica F è altamente significativa.

Modello ridotto 'Wind' Adesso consideriamo il modello che dipende soltanto dalla variabile esplicativa 'Wind':

$$Y = \beta_0 + \beta_2 X_2 + \epsilon \tag{4}$$

```
mq3 <- lm(Ozone ~ Wind)
summary(mq3)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87.03 -37.20 -10.07  31.64 176.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  198.074     14.945  13.254  < 2e-16 ***
## Wind         -25.663      3.173  -8.089 9.11e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.85 on 109 degrees of freedom
## Multiple R-squared:  0.3751,Adjusted R-squared:  0.3694
## F-statistic: 65.43 on 1 and 109 DF,  p-value: 9.114e-13
```

```
confint(mq3)
```

```
##                 2.5 %    97.5 %
## (Intercept) 168.45390 227.69335
## Wind        -31.95121 -19.37537
```

Commento:

- l'effetto della variabile 'Wind' si mantiene negativo ma di intensità maggiore;
- la variabile risulta altamente significativa;
- gli intervalli di confidenza sono buoni e non contengono il valore nullo, indice di coerenza;
- gli indici R^2 sono molto bassi;
- la statistica F è altamente significativa.

Modello ridotto 'Solar.R' Come ultima prova consideriamo il modello ridotto dipendente solo dalla variabile esplicativa 'Solar.R':

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \tag{5}$$

```
mq4 <- lm(Ozone ~ Solar.R)
summary(mq4)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -96.58 -42.72 -17.73  32.74 238.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.720e+01  1.350e+01   2.756 0.006856 **
```

```
## Solar.R     6.079e-06  1.567e-06   3.880 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.67 on 109 degrees of freedom
## Multiple R-squared:  0.1213,Adjusted R-squared:  0.1133
## F-statistic: 15.05 on 1 and 109 DF,  p-value: 0.0001793

confint(mq4)

##                   2.5 %       97.5 %
## (Intercept) 1.04492e+01 6.394571e+01
## Solar.R     2.97341e-06 9.183882e-06
```

Commento:

- l'effetto della variabile è positivo;
- la variabile, presa singolarmente, risulta essere altamente significativa per il modello, probabilmente perché c'è una variabile, come ad esempio 'Temp', che "media" 'Solar.R'. Più avanti, con lo studio dei modelli grafici, proveremo a spiegare questo fenomeno;
- gli intervalli di confidenza non contengono il valore nullo, tuttavia gli estremi sono molto vicini allo zero;
- gli indici R^2 sono estremamente bassi;
- la statistica F è significativa, ma non come negli altri modelli.

2.3 Test del rapporto di verosimiglianza per la scelta del modello

Dalle analisi effettuate precedentemente i due modelli più promettenti risultano essere quello completo e quello ridotto dipendente dalle due variabili esplicative più significative: 'Wind' e 'Temp'. Poiché si trattano di due modelli uno innestato all'altro, posso utilizzare il **test del rapporto di verosimiglianza** per prendere una decisione su quale sia il modello migliore da adottare. Per farlo, mi avvalgo della funzione **lrtest** del pacchetto **lmtest**, che prende come primo parametro il modello ridotto e come secondo il modello completo ed esegue il test del rapporto di verosimiglianza, considerando come ipotesi H_0 la scelta di adottare il modello ridotto.

```
library(lmtest)

lrtest(mq1, mq) #nested vs complex

## Likelihood ratio test
##
## Model 1: Ozone ~ Wind + Temp
## Model 2: Ozone ~ Solar.R + Wind + Temp
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -574.65
## 2   5 -571.30  1 6.6992   0.009646 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I risultati ottenuti ci indicano di adottare il **modello completo mq** poiché il valore del p-value suggerisce di rigettare l'ipotesi H_0, ovvero quella di adottare il modello ridotto. Nonostante ciò, dai risultati ottenuti, si osserva che anche quello ridotto, oltre ad essere più semplice, risulta avere ottimi valori metrici ed una statistica F di valore superiore a quella del modello completo. Per questo motivo, si ritiene plausibile

l'eventuale scelta del modello ridotto nel caso si voglia prediligere un modello più semplice, secondo il principio di **parsimonia**.

2.4 Analisi dei plot del modello completo

A questo punto, tenendo in considerazione il modello completo, possiamo fare il plot dei valori predetti dal modello. Per farlo utilizziamo la funzione **ggpredict** del pacchetto **ggeffects**, che prende in input il modello completo **mq** e le tre variabili esplicative 'Solar.R', 'Wind' e 'Temp'. Esegue il plot dei valori predetti, marginalizzando per 'Solar.R' e 'Wind', mantenendo variabile 'Temp'. In Figura 4 è possibile apprezzarne l'output.

```
library(ggeffects)
plot(ggpredict(mq, c("Temp", "Wind", "Solar.R")))
```

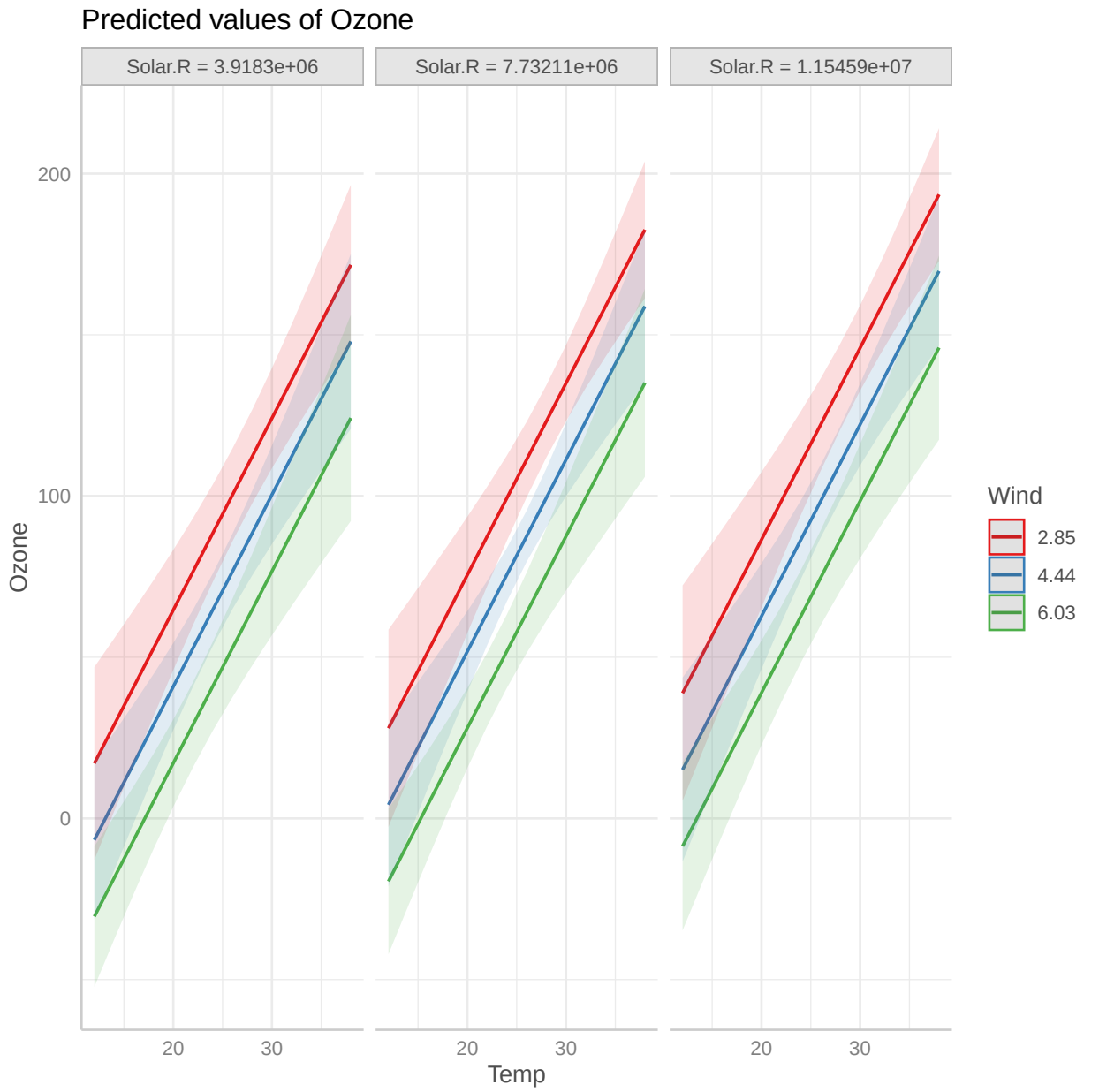


Figura 4: Plot dei valori dell'Ozono predetti dal modello completo

In riferimento alla Figura 4, i tre colori indicano tre velocità del vento diverse: rosso sta per 2.85 m/s, blue sta per 4.44 m/s e verde sta per 6.03 m/s. Le tre colonne indicano 3 valori differenti dell'intensità dei raggi solari, i cui valori sono scritti in alto per ciascuna colonna. Lungo l'asse delle ascisse abbiamo i valori della temperatura mentre lungo l'asse delle ordinate i valori predetti dell'Ozono nell'aria. Ricordando che il valore limite di Ozono, dopo il quale l'aria inizia ad assumere una qualità dannosa per la salute umana, è di 120 μg /m^3, forniamo un breve commento dell'output:

- la velocità del vento, come visto fino ad ora, ha un effetto che mitiga molto i livelli di Ozono nell'aria ed infatti, quando soffia ad almeno 6.03 m/s (parte verde del grafico), i livelli critici di Ozono vengono raggiunti a mala pena in presenza di alte temperature;
- da una certa temperatura in poi, quando il vento soffia a meno di 4.44 m/s, i livelli di Ozono nell'aria diventano di livello critico, toccando punte di circa 200 μg /m^3 in prossimità dei livelli più alti di intensità dei raggi solari;
- come ci aspettavamo, l'effetto della variabile 'Solar.R' non apporta effetti sostanziali ai livelli di Ozono nell'aria, infatti pur aumentando di un ordine di grandezza dal primo grafico al terzo, la variabile **driver**, con effetti positivi sui livelli di Ozono, risulta essere la temperatura.

2.5 Considerazioni finali e conclusioni

Come annunciato precedentemente, i qqplot del modello completo, che risulta essere il migliore tra i modelli di regressione lineare testati fino ad ora, presenta delle anomalie per le osservazioni estreme. In particolare, come mostrato in Figura 5, le osservazioni 23, 34 e 77 si discostano molto dalla retta principale. Inoltre, se consideriamo il grafico **"Residuals vs Fitted"** in Figura 6, possiamo osservare come la **linea rossa** abbia un andamento curvilineo, simile a quello di una parabola, quando invece vorremmo che fosse quanto più possibile parallela alla retta tratteggiata. Queste analisi ci fanno pensare che probabilmente, un modello polinomiale potrebbe adattarsi meglio ai dati in questione e per questo motivo, nella sezione successiva verrà trattata questa ipotesi.

```
plot(mq)
```

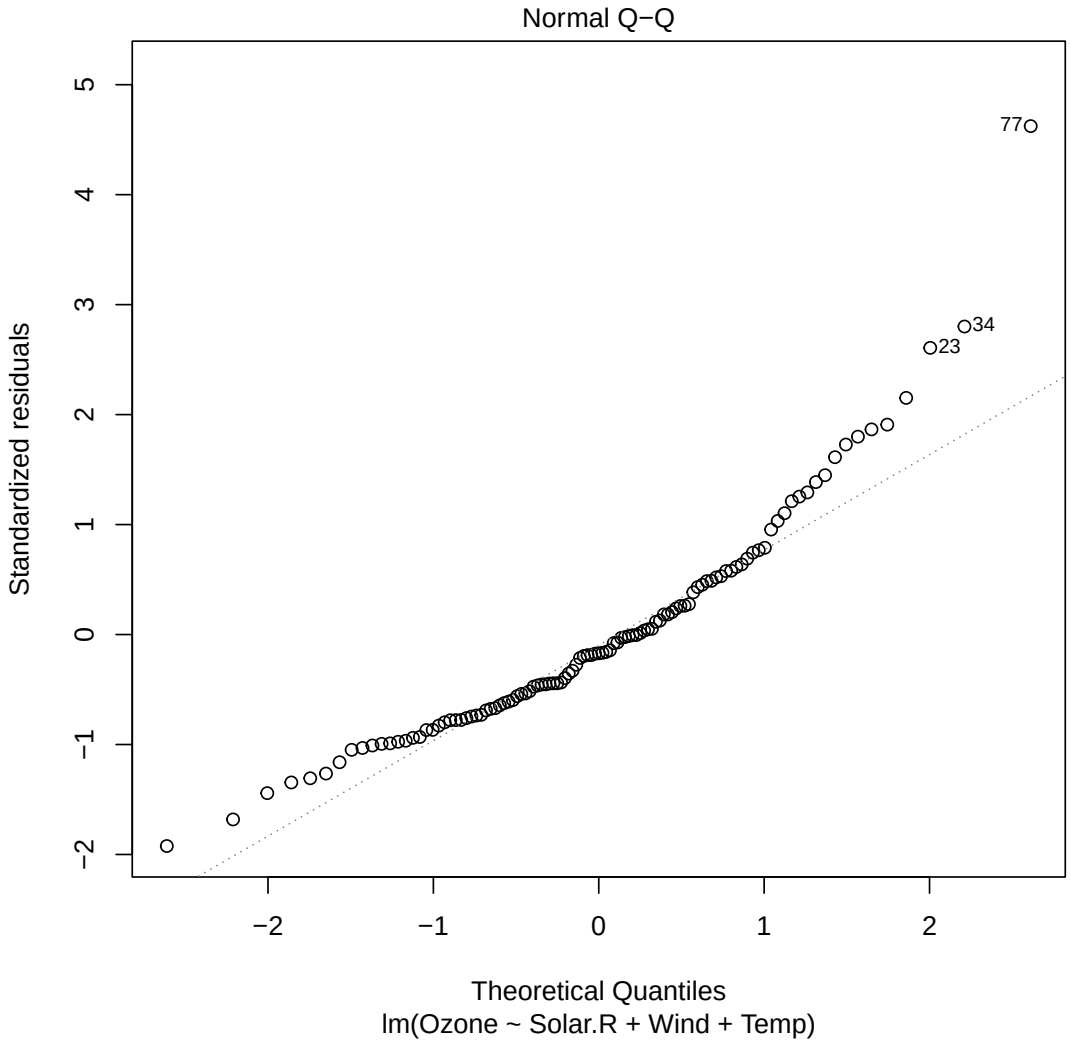


Figura 5: qqplot dei residui del modello di regressione lineare completo

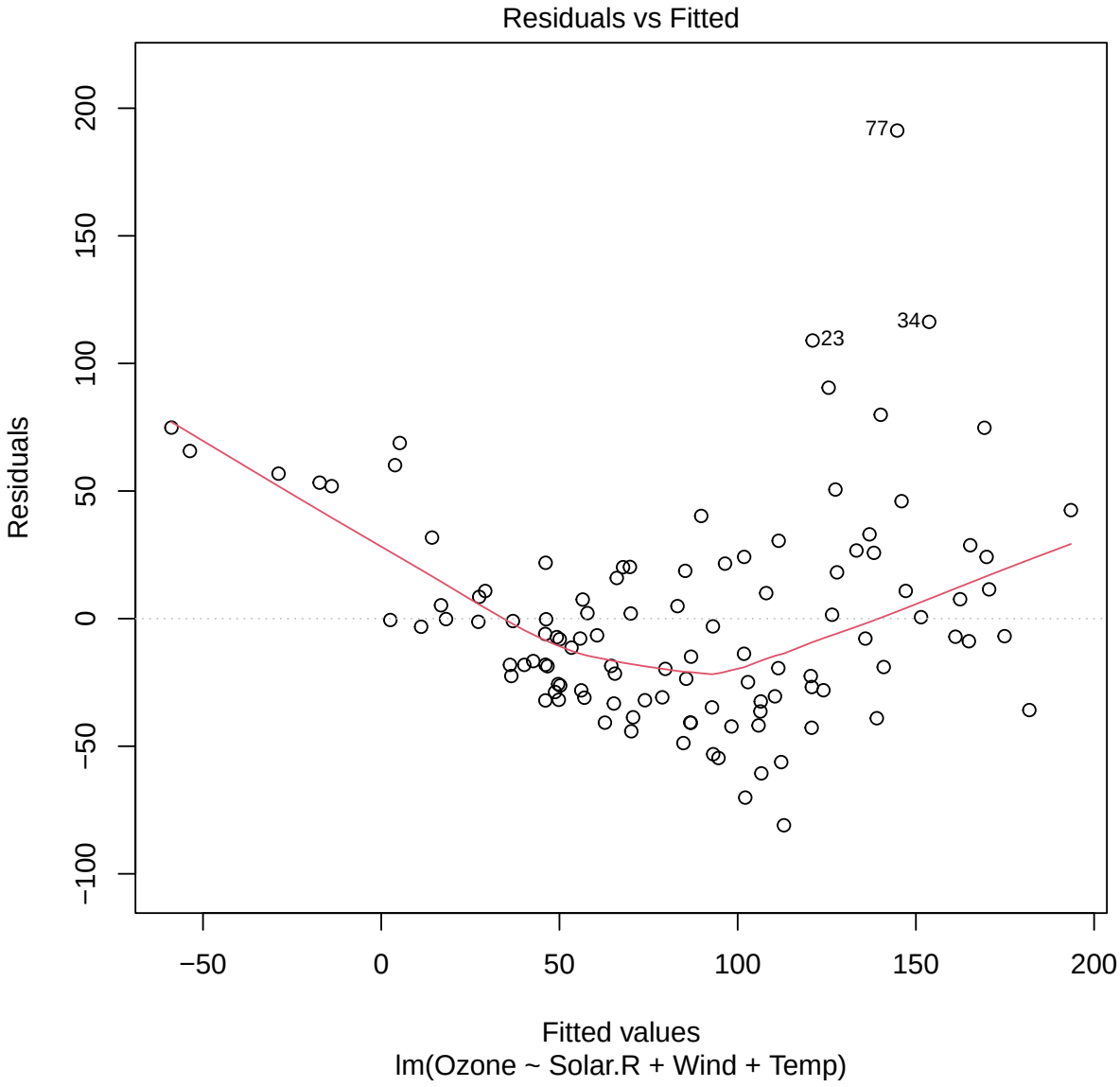


Figura 6: plot del grafico Residuals - Fitted Values del modello completo

3 Regressione polinomiale

I risultati ottenuti nella sezione precedente mostrano che il modello di regressione lineare migliore è quello completo. Tuttavia, un'analisi più approfondita dei qqplot e dei "residuals vs fitted" mostrano che probabilmente un modello polinomiale potrebbe adattarsi meglio ai dati in questione. Per questo motivo, in questa sezione analizzeremo un approccio al problema tramite un modello di regressione polinomiale. Per motivi di praticità tratteremo soltanto il modello che è risultato migliore tra quelli testati.

3.1 Modello di regressione polinomiale multipla

Il modello polinomiale preso in considerazione è un modello completo con la variabile 'Wind' di secondo grado. Il modello in considerazione è il seguente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \beta_4 X_3 + \epsilon \tag{6}$$

```
d=2
pq <- lm(Ozone ~ Solar.R +
           poly(Wind, degree = d, raw = T) + Temp)
summary(pq)

##
## Call:
## lm(formula = Ozone ~ Solar.R + poly(Wind, degree = d, raw = T) +
##     Temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -87.272 -24.491  -7.422  18.167 144.971
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.110e+02  3.748e+01   2.961  0.00378 **
## Solar.R                         2.915e-06  9.882e-07   2.950  0.00391 **
## poly(Wind, degree = d, raw = T)1 -6.803e+01  1.026e+01  -6.629 1.45e-09 ***
## poly(Wind, degree = d, raw = T)2  5.408e+00  1.011e+00   5.351 5.10e-07 ***
## Temp                            5.210e+00  8.253e-01   6.312 6.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.77 on 106 degrees of freedom
## Multiple R-squared:  0.6897,Adjusted R-squared:  0.678
## F-statistic:  58.9 on 4 and 106 DF,  p-value: < 2.2e-16
```

Commento:

- la variabile 'Solar.R' continua ad influenzare positivamente la variabile obiettivo 'Ozone', con simile intensità dei modelli visti precedentemente;
- la variabile 'Wind' continua ad influenzare negativamente la variabile obiettivo e il suo effetto è auemntato di intensità;
- la componente quadratica di 'Wind' ha un effetto positivo, probabilmente indica la curvatura verso l'alto della funzione quadratica;
- la variabile 'Temp' continua ad avere un effetto positivo di intensità simile vista precedentemente;
- tutte le variabili risultano altamente significative, anche 'Solar.R';

- la stima della deviazione standard è 37.77, più bassa delle misurazioni precedenti;
- gli indici R^2 sono aumentati;
- la statistica F è aumentata leggermente di valore e risulta altamente significativa.

Passiamo all'analisi degli intervalli di confidenza

```
confint(pq)
```

```
##                                         2.5 %        97.5 %
## (Intercept)                        3.667822e+01  1.853000e+02
## Solar.R                            9.563028e-07  4.874559e-06
## poly(Wind, degree = d, raw = T)1 -8.837275e+01 -4.768359e+01
## poly(Wind, degree = d, raw = T)2  3.403948e+00  7.411493e+00
## Temp                               3.573293e+00  6.845759e+00
```

Gli intervalli di confidenza risultano buoni e non contengono il valore nullo, indice di coerenza con i test di significatività effettuati precedentemente.

3.2 Analisi dei plot del modello polinomiale

Passiamo all'analisi dei plot dei valori predettii dal modello polinomiale, utilizzando la stessa funzione **ggpredict** utilizzata precedentemente. Questa volta marginalizziamo per le variabili 'Temp' e 'Solar.R' e manteniamo variabile 'Wind' per vedere meglio gli effetti della quadratura. In Figura 7 è possibile apprezzarne l'output.

```
plot(ggpredict(pq, c("Wind", "Temp", "Solar.R")))
```

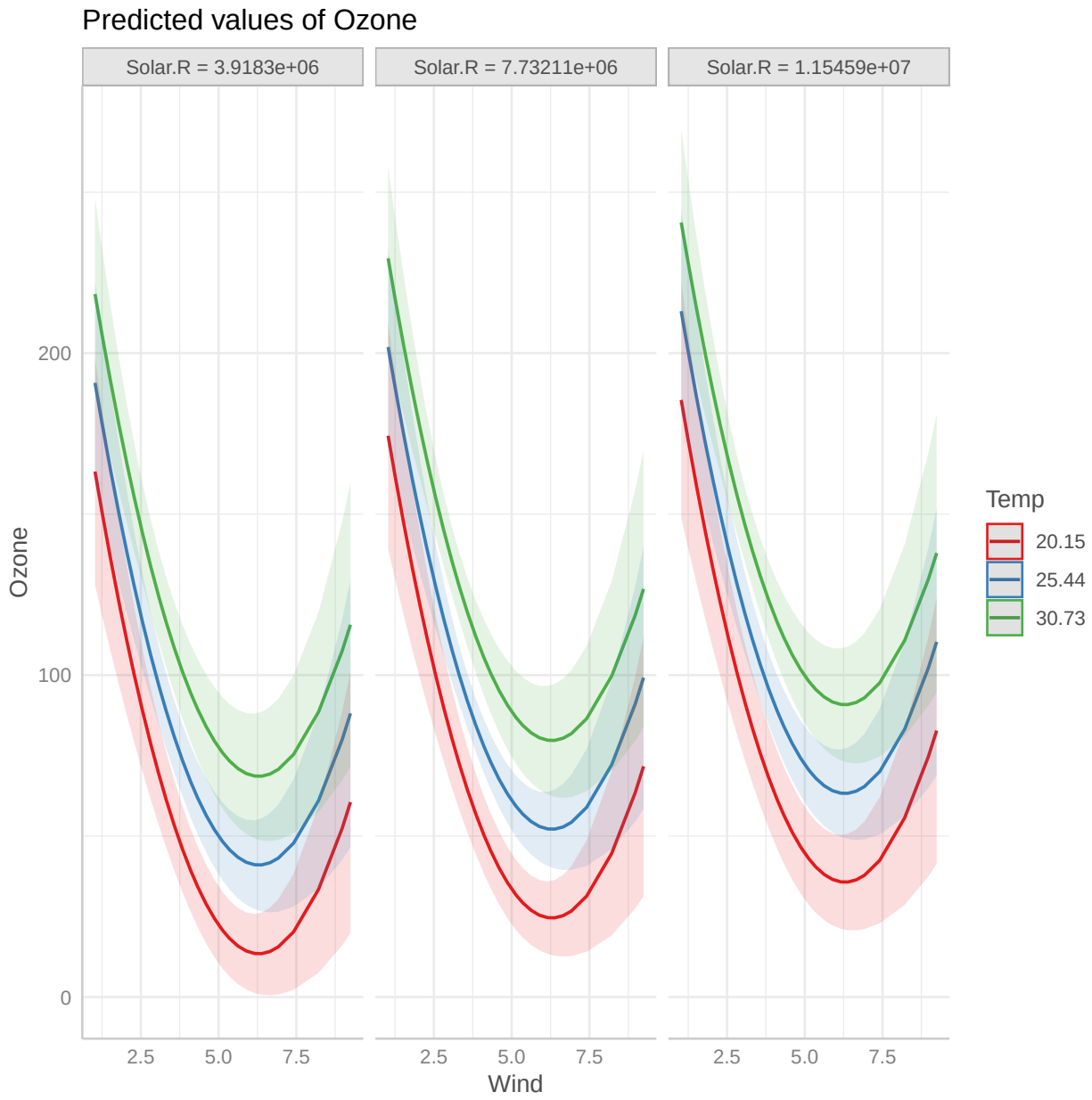


Figura 7: Plot dei valori dell'Ozono predetti dal modello polinomiale

In riferimento alla Figura 7, i tre colori indicano tre valori di temperatura in gradi Celsius diversi: rosso sta per 20.15 C, blu sta per 25.44 C e verde sta per 30.73 C. Le tre colonne indicano 3 valori differenti dell'intensità dei raggi solari, i cui valori sono scritto in alto per ciascuna colonna. Lungo l'asse delle ascisse abbiamo i valori della velocità del vento mentre lungo l'asse delle ordinate abbiamo i valori predetti dell'Ozono nell'aria. Ricordando che il valore limite di Ozono nell'aria, dopo il quale si hanno effetti dannosi per la salute umana, è 120 μg /m^3, forniamo un breve commento del grafico:

- la curvatura del grafico di predizione, come previsto dall'effetto positivo della variabile 'Wind' al secondo grado, è verso l'alto;
- all'aumentare della velocità del vento, a parità di altre condizioni, i livelli di Ozono nell'aria diminuiscono. Tuttavia, questo accade entro un certo limite di velocità che è circa 6.0 m/s, dopo il quale, la velocità del vento assume un effetto positivo, seppur esiguo. Infatti, gli effetti negativi hanno un'intensità maggiore in termini assoluti di quelli positivi;
- gli effetti di 'Temp' e di 'Solar.R' sono in linea con ciò che è stato osservato nell'analisi del modello lineare completo.

3.3 Analisi "residuals vs fitted" e qqplot

Dato il modello polinomiale descritto precedentemente, passiamo all'analisi dei qqplot dei residui e del grafico dei "residuals vs fitted". In Figura 8, possiamo vedere che l'andamento dei residui risulta seguire quello Normale, e i punti, eccetto casi estremi, sono molto vicino alla retta principale, similmente a quanto visto nel modello completo lineare. Invece, in Figura 9, è possibile vedere che il grafico "residuals vs fitted" mostra dei risultati migliori rispetto a quelli del modello lineare, poiché la linea rossa ha un andamento quasi lineare e si discosta poco dalla retta parallela all'asse delle ascisse principale.

```
plot(pq)
```

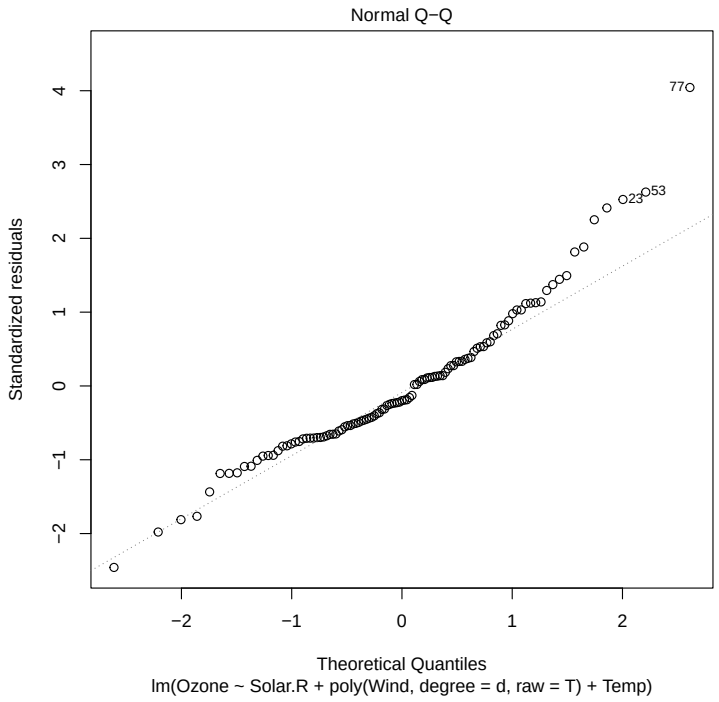


Figura 8: qqplot dei residui del modello di regressione polinomiale

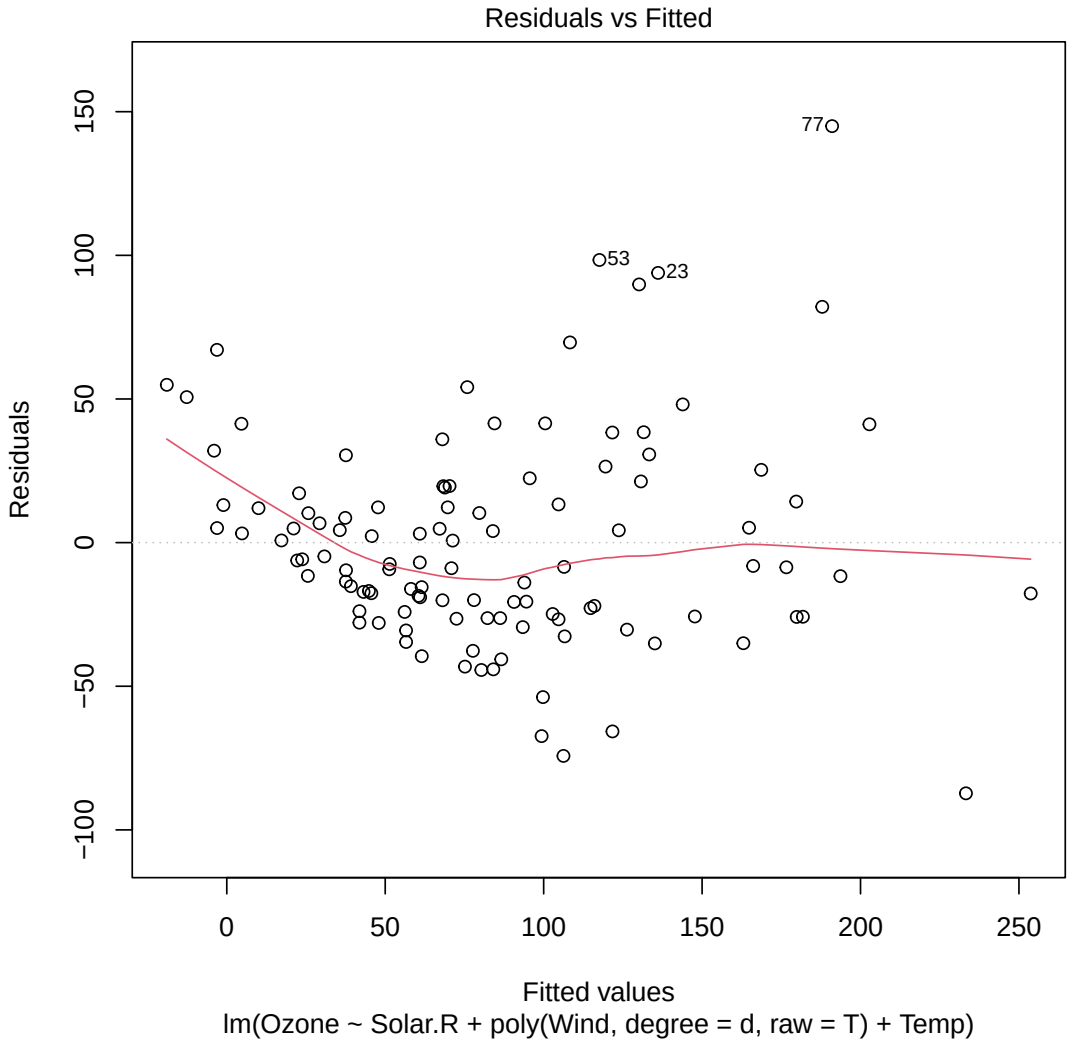


Figura 9: plot del grafico Residuals - Fitted Values del modello polinomiale

3.4 Test F parziale per la scelta del modello

Dalle valutazioni grafiche qualitative sembrerebbe che il modello polinomiale sia quello più adatto e per questo motivo vorremmo testare quest'ipotesi con un metodo quantitativo. Per valutare quale dei due modelli sia il migliore usiamo il **test F parziale** attraverso la funzione **anova**. L'ipotesi nulla H_0 corrisponde a dire che non ci sono differenze significative tra i due modelli e che quindi il modello lineare è il prescelto, mentre l'ipotesi H_1 è a favore del modello polinomiale.

```
mq <- lm(Ozone ~ Solar.R + Wind + Temp)
anova(mq, pq) # H0: mq, H1: pq

## Analysis of Variance Table
##
## Model 1: Ozone ~ Solar.R + Wind + Temp
## Model 2: Ozone ~ Solar.R + poly(Wind, degree = d, raw = T) + Temp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
```

```
## 1    107 192018
## 2    106 151185  1     40832 28.629 5.1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dal risultato del test si evince che l'ipotesi nulla H_0 viene rigettata che dunque il modello migliore tra i due è quello polinomiale.

4 Regressione logistica

Date le premesse fatte all'inizio, riguardo i valori limite proposti dall'Organizzazione Mondiale per la Sanità (O.M.S.) dei livelli di Ozono nell'aria, si è ritenuto più opportuno, ai fini della semplicità interpretativa dei risultati, di proporre un modello che prevedesse quando l'aria è di cattiva qualità piuttosto che no. Per farlo, il modello più adatto per l'obiettivo preposto è quello di **regessione logistica**.

In questa sezione, dunque, ci occuperemo di trovare un modello di regressione logistica, che date le variabili 'Solar.R', 'Wind' e 'Temp' ci possa indicare se l'aria è di cattiva qualità o meno. Prima di tutto, è stata introdotta una nuova variabile binaria '**Quality**' che assume valore **1** quando i livelli di Ozono superano il livello limite di 120 μg /m^3 (indice di cattiva qualità dell'aria) e **0** altrimenti.

```
quality_treshold = 120
Quality <- as.numeric(data$Ozone > quality_treshold)
data$Quality <- Quality
```

4.1 Analisi preliminare qualitativa

Prima di procedere con lo studio dei modelli logistici, effettuiamo alcune analisi qualitative dei dati.

Box plot Per prima cosa, procediamo con un'analisi dei box plot confrontando la variabile obiettivo 'Quality' con le altre variabili esplicative 'Solar.R', 'Wind' e 'Temp'.

```
boxplot(Temp  ~ Quality)
boxplot(Wind  ~ Quality)
boxplot(Solar.R  ~ Quality)
```

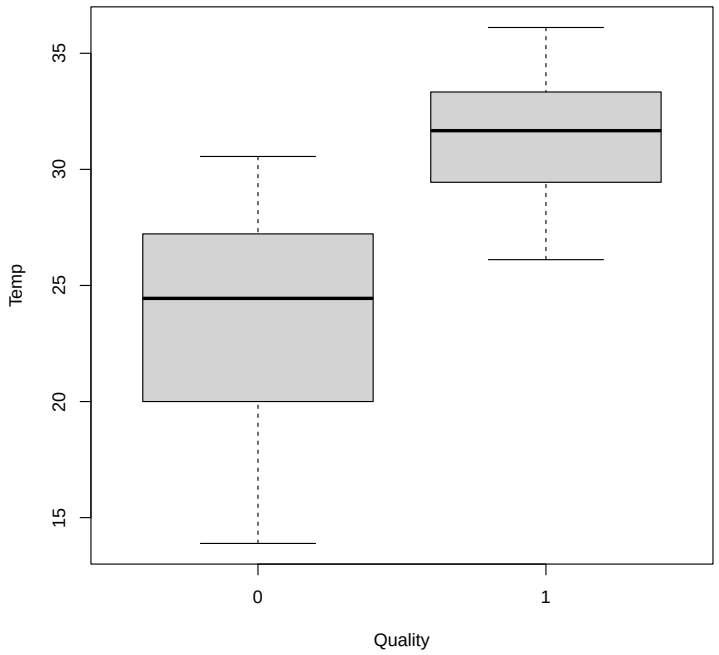


Figura 10: Box plot 'Temp' vs 'Quality'

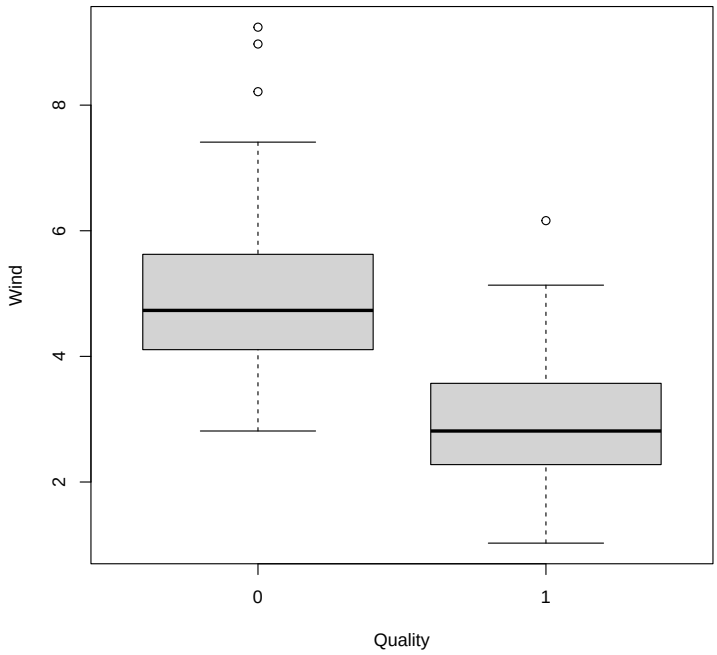


Figura 11: Box plot 'Wind' vs 'Quality'

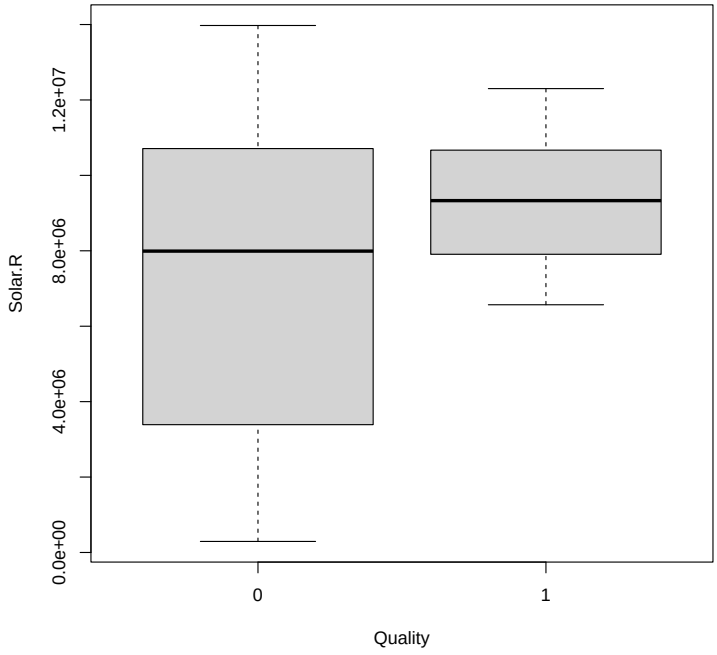


Figura 12: Box plot 'Solar.R' vs 'Quality'

Commento:

- in riferimento alla Figura 10, tra la variabile 'Temp' e 'Quality' sembrerebbe esserci un rapporto di proporzionalità diretta, ovvero con l'aumentare della temperatura la probabilità che l'aria sia di scarsa qualità aumenta;

- in riferimento alla Figura 11, tra la variabile 'Wind' e 'Quality', invece, sembra esserci un rapporto di proporzionalità indiretta, cioè all'aumentare della velocità vento la probabilità che l'aria sia di scarsa qualità diminuisce;

- in riferimento alla Figura 12, tra la variabile 'Solar.R' e 'Quality' sembrerebbe esserci un rapporto di proporzionalità diretta. Tuttavia, le mediane sono molto vicine tra loro, e la box di sinistra si estende molto in verticale, indice di un'alta variabilità. Questo box plot è poco informativo.

Matrice di correlazione Passiamo all'analisi della matrice di correlazione tra le variabili esplicative del modello e la variabile obiettivo 'Quality'.

```
library(corrplot)

## corrplot 0.92 loaded

# mat : matrice dei dati
# ... : argomenti aggiuntivi da passare alla funzione nativa di R cor.test
cor.mtest <- function(mat, ...) {
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat<- matrix(NA, n, n)
  diag(p.mat) <- 0
  for (i in 1:(n - 1)) {
    for (j in (i + 1):n) {
      tmp <- cor.test(mat[, i], mat[, j], ...)
      p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
    }
  }
  colnames(p.mat) <- rownames(p.mat) <- colnames(mat)
  p.mat
}

# matrice dei p-value di correlazione
p.mat <- cor.mtest(data)
M <- cor(data)
col<- colorRampPalette(c("red", "white", "blue"))(20)
corrplot(M, method = "number", type = "upper",
         order = "hclust", col=col,
         p.mat = p.mat,
         sig.level = 0.01, insig = "blank")
```

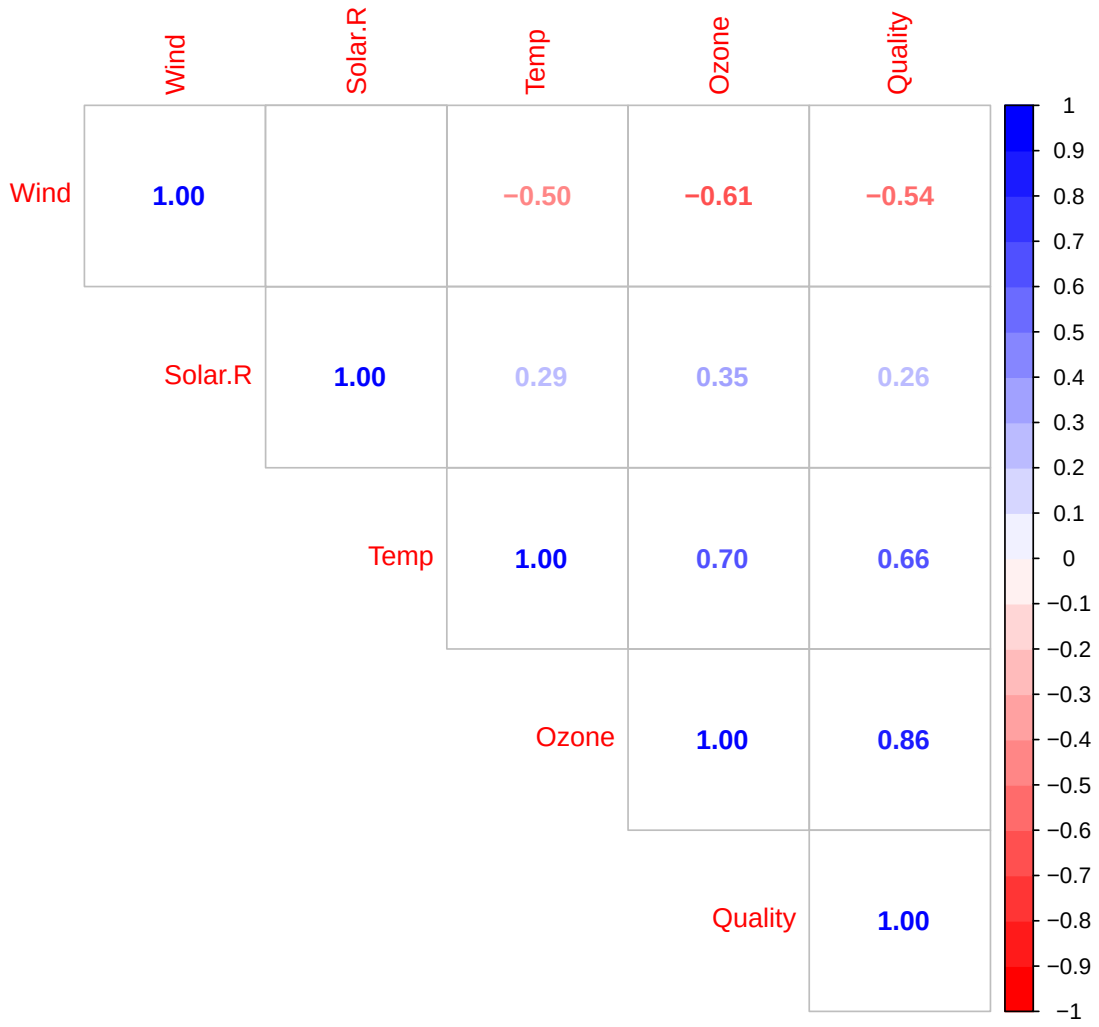


Figura 13: Matrice di correlazione tra le variabili del data frame 'data'

Commento in riferimento alla Figura 13:

- le variabili 'Wind' e 'Solar.R', marginalizzando per le altre variabili, hanno una bassa correlazione. Questo probabilmente spiega perché pur rimuovendo la variabile 'Solar.R' dal modello i valori stimati del parametro per la variabile 'Wind' variano di poco;
- le variabili 'Wind' e 'Temp', al netto delle altre variabili, hanno una correlazione **moderata negativa**. Anche 'Wind' e 'Quality' hanno circa lo stesso livello di correlazione;
- le variabili 'Temp' e 'Quality' hanno un indice di correlazione **forte positivo**, mentre l'indice di correlazione tra 'Solar.R' e 'Temp' è **debole positivo**;
- le variabili 'Solar.R' e 'Quality' hanno un indice di correlazione **debole positivo**.

4.2 Analisi dei modelli logistici

Dopo aver condotto un'analisi qualitativa del problema, procediamo con quella quantitativa, cercando di trovare il modello logistico più adatto alla previsione di nostro interesse. Infatti, il nostro obiettivo è capire come e quali variabili influenzano la probabilità l'aria sia di scarsa qualità. L'approccio utilizzato è quello di partire dal modello più semplice, ovvero quello dipendente dalle singole variabili fino a giungere a quello completo. La scelta del modello viene fatta **favorendo quello più parsimonioso**, in maniera tale da trovare il giusto compromesso tra significatività, numero dei parametri e semplicità interpretativa del modello stesso.

Modello ridotto 'Temp' Consideriamo il modello di regressione logistica dipendente soltanto dalla variabile esplicativa 'Temp'.

```
fit <- glm(Quality ~ Temp, family = binomial)
pseudoRfit <- ((fit$null.deviance/-2) - (fit$deviance/-2)) /
  ((fit$null.deviance/-2))
summary(fit)

##
## Call:
## glm(formula = Quality ~ Temp, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.64273  -0.32522  -0.05670   0.09135   2.34847
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -24.6681     5.3510  -4.610 4.03e-06 ***
## Temp          0.8416     0.1856   4.535 5.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.51  on 110  degrees of freedom
## Residual deviance:  52.15  on 109  degrees of freedom
## AIC: 56.15
##
## Number of Fisher Scoring iterations: 7

pseudoRfit

## [1] 0.5910162

confint(fit)

##                   2.5 %     97.5 %
## (Intercept) -37.1062389 -15.824728
## Temp          0.5346404   1.272777
```

Commento:

- $\hat{\beta}_0 = -24.6681$ e $\hat{\beta}_1 = 0.8416$
- l'aumento della temperatura comporta un aumento della probabilità di avere cattiva qualità dell'aria;
- un aumento unitario della temperatura comporta l'incremento del log-odds di 'Quality' di 0.8416, oppure l'odds di avere cattiva qualità dell'aria di $e^{0.8416} = 2.320$;
- la devianza dei residui è pressoché centrata intorno al valore zero e leggermente asimmetrica, ma dai dati empirici ce lo possiamo aspettare;
- il p-value di $\hat{\beta}_1$ è molto piccolo, quindi l'effetto della temperatura sul rischio di avere una qualità cattiva dell'aria è altamente significativo;
- lo $pseudo - R^2$ è 0.5910, moderatamente buono;
- gli intervalli di confidenza sono buoni e non contengono il valore nullo, coerentemente quanto visto con il test di significatività.

Modello ridotto 'Wind' Consideriamo il modello di regressione logistica dipendente soltanto dalla variabile esplicativa 'Wind'.

```
fit1 <- glm(Quality ~ Wind, family = binomial)
pseudoRfit1 <- ((fit1$null.deviance/-2) - (fit1$deviance/-2)) /
  ((fit1$null.deviance/-2))
summary(fit1)
```

```
##
## Call:
## glm(formula = Quality ~ Wind, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3874  -0.5733  -0.3289   0.3265   2.9643
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.5641     1.1261   4.053 5.06e-05 ***
## Wind         -1.4519     0.3064  -4.739 2.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.51  on 110  degrees of freedom
## Residual deviance:  82.80  on 109  degrees of freedom
## AIC: 86.8
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoRfit1
```

```
## [1] 0.3506418
```

```
confint(fit1)
```

```
##                 2.5 %     97.5 %
## (Intercept)  2.558066  7.0170816
## Wind        -2.130282 -0.9166628
```

Commento:

- $\hat{\beta}_0 = 4.5641$ e $\hat{\beta}_1 = -1.4519$
- l'aumento della velocità del vento comporta una diminuzione della probabilità di avere cattiva qualità dell'aria;
- un aumento unitario della velocità del vento comporta la diminuzione del log-odds di 'Quality' di 1.4519, oppure l'odds di avere cattiva qualità dell'aria di $e^{1.4519} = 4.2712$;
- la devianza dei residui è meno centrata intorno allo zero, rispetto al modello precedente, e risulta meno simmetrica;
- il p-value di $\hat{\beta}_1$ è molto piccolo, quindi l'effetto della velocità del vento sul rischio di avere una qualità cattiva dell'aria è altamente significativo;
- lo $pseudo - R^2$ è 0.3506, molto basso;
- gli intervalli di confidenza sono buoni e non contengono il valore nullo, coerentemente quanto visto con il test di significatività.

Modello ridotto 'Solar.R' Consideriamo il modello di regressione logistica dipendente soltanto dalla variabile esplicativa 'Solar.R'.

```
fit2 <- glm(Quality ~ Solar.R, family = binomial)
pseudoRfit2 <- ((fit2$null.deviance/-2) - (fit2$deviance/-2)) /
  ((fit2$null.deviance/-2))
summary(fit2)

##
## Call:
## glm(formula = Quality ~ Solar.R, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1710  -0.8807  -0.5514   1.3211   1.7845
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.562e+00  6.577e-01  -3.895 9.82e-05 ***
## Solar.R      1.822e-07  6.915e-08   2.635  0.00842 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.51  on 110  degrees of freedom
## Residual deviance: 119.28  on 109  degrees of freedom
## AIC: 123.28
##
## Number of Fisher Scoring iterations: 4

pseudoRfit2

## [1] 0.0645261

confint(fit2)

##                     2.5 %        97.5 %
```

```
## (Intercept) -3.997957e+00 -1.388426e+00
## Solar.R      5.488160e-08  3.289685e-07
```

Commento:

- $\hat{\beta}_0 = -2.5620$ e $\hat{\beta}_1 = 0.0000001822$
- l'aumento dell'intensità dei raggi solari comporta un aumento della probabilità di cattiva qualità dell'aria;
- un aumento unitario dell'intensità dei raggi solari aumenta il log-odds di 'Quality' di 0.0000001822, cioè aumenta l'odds di avere cattiva qualità dell'aria di $e^{0.0000001822} = 1.0000002$;
- la devianza dei residui è meno centrata intorno allo zero, rispetto al modello dipendente da 'Temp', ma risulta abbastanza simmetrica;
- il p-value di $\hat{\beta}_1$ è piccolo, quindi l'effetto di 'Solar.R' sul rischio di avere una qualità cattiva dell'aria è significativo;
- lo $pseudo - R^2$ è 0.06452, decisamente basso;
- gli intervalli di confidenza sono buoni, ma gli estremi di confidenza sono molto piccoli, e dunque molto vicini allo zero. Questo potrebbe essere interpretato come indice di **incoerenza**.

Modello ridotto 'Temp' + 'Wind' Adesso consideriamo il modello di regressione logistica multipla dipendente solo dalle due variabili che sono risultate più significative, ovvero 'Temp' e 'Wind'.

```
fit4 <- glm(Quality ~ Temp + Wind, family = binomial)
pseudoRfit4 <- ((fit4$null.deviance/-2) - (fit4$deviance/-2)) /
  ((fit4$null.deviance/-2))
summary(fit4)

##
## Call:
## glm(formula = Quality ~ Temp + Wind, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.03729  -0.15628  -0.01446   0.02433   2.49240
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.5364     6.7803  -3.324 0.000888 ***
## Temp          0.9421     0.2556   3.685 0.000228 ***
## Wind         -1.3038     0.4315  -3.022 0.002513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.511  on 110  degrees of freedom
## Residual deviance:  37.676  on 108  degrees of freedom
## AIC: 43.676
##
## Number of Fisher Scoring iterations: 8

pseudoRfit4

## [1] 0.7045236
```

```
confint(fit4)

##                   2.5 %     97.5 %
## (Intercept) -38.4171401 -11.269665
## Temp          0.5308523   1.552826
## Wind         -2.2994430  -0.564869
```

Commento:

- $\hat{\beta}_0 = -22.5364$, $\hat{\beta}_1 = 0.9421$ e $\hat{\beta}_2 = -1.3038$
- gli effetti delle variabili 'Temp' e 'Wind' rimangono invariati e gli erorri standard sono molto simili a quelli ottenuti precedentemente;
- la devianza dei residui abbastanza centrata intorno allo zero, e la distribuzione risulta simmetrica, ma non perfettamente, com'è possibile aspettarci da dati empirici;
- i p-value di $\hat{\beta}_1$ e $\hat{\beta}_2$ sono piccoli, quindi gli effetti di 'Temp' e 'Wind', rispettivamente marginalizzati uno rispetto all'altro, sulla probabilità di avere aria di cattiva qualità, sono altamente significativi;
- lo $pseudo - R^2$ è 0.7045, da considerarsi buono.
- gli intervalli di confidenza sono buoni e non contengono il valore nullo. Questo risultato può essere interpretato come indice di **coerenza**.

Modello completo 'Temp' + 'Wind' + 'Solar.R' Infine, consideriamo il modello di regressione logistica multipla completo, cioè dipendente dalle variabili esplicative 'Temp', 'Wind' e 'Solar.R'.

```
fit3 <- glm(Quality ~ Temp + Wind + Solar.R, family = binomial)
pseudoRfit3 <- ((fit3$null.deviance/-2) - (fit3$deviance/-2)) /
  ((fit3$null.deviance/-2))
summary(fit3)

##
## Call:
## glm(formula = Quality ~ Temp + Wind + Solar.R, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.03613  -0.14861  -0.01618   0.02795   2.66469
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.240e+01  6.623e+00  -3.383 0.000717 ***
## Temp         8.781e-01  2.453e-01   3.580 0.000344 ***
## Wind        -1.388e+00  4.644e-01  -2.988 0.002805 **
## Solar.R      2.246e-07  1.626e-07   1.381 0.167213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.511  on 110  degrees of freedom
## Residual deviance:  35.478  on 107  degrees of freedom
## AIC: 43.478
##
## Number of Fisher Scoring iterations: 8
```

```
pseudoRfit3

## [1] 0.7217658

confint(fit3)

##                      2.5 %        97.5 %
## (Intercept) -3.784683e+01 -1.134373e+01
## Temp         4.816114e-01  1.461778e+00
## Wind        -2.462442e+00 -5.964242e-01
## Solar.R     -6.767748e-08  5.895068e-07
```

Commento:

- $\hat{\beta}_0 = -22.400$, $\hat{\beta}_1 = 0.8781$, $\hat{\beta}_2 = -1.3880$ e $\hat{\beta}_3 = 0.000000225$
- gli effetti delle variabili 'Temp' e 'Wind' rimangono più o meno invariati e gli errori standard sono molto simili a quelli ottenuti precedentemente;
- l'effetto di 'Solar.R' rimane positivo e aumenta di intensità, tuttavia la variabile perde di significatività perché probabilmente la sua informazione è mediata da altre variabili, presumibilmente 'Temp' o 'Wind'. Per questo occorrono ulteriori analisi che verranno svolte in seguito;
- la devianza dei residui abbastanza centrata intorno allo zero, e la distribuzione risulta simmetrica, ma non perfettamente, com'è possibile aspettarci da dati empirici;
- i p-value di $\hat{\beta}_1$ e $\hat{\beta}_2$ sono piccoli, quindi gli effetti di 'Temp' e 'Wind', rispettivamente marginalizzati uno rispetto all'altro, sulla probabilità di avere aria di cattiva qualità, sono altamente significativi;
- lo $pseudo - R^2$ è 0.7217, da considerarsi buon e leggermente più elevato del modello precedente;
- gli intervalli di confidenza sono buoni e non contengono il valore nullo per le variabili 'Temp' e 'Wind', mentre quello in riferimento alla variabile 'Solar.R' contiene il valore nullo. Questo risultato ce lo potevamo aspettare poiché 'Solar.R' risulta essere non significativa.

Modello ridotto 'Wind' + 'Solar.R' Dato che nell'analisi precedente abbiamo constato che nel modello completo 'Solar.R' perde di significatività rispetto al modello ridotto dipendente solo da 'Solar.R', l'ipotesi è che l'informazione prodotta da 'Solar.R' venga in qualche modo mediata dalle altre due variabili. Adesso analizzeremo il modello composto solo da 'Wind' e 'Solar.R', commentandolo, e poi faremo la stessa cosa per il modello 'Temp' e 'Solar.R'. Per completare l'analisi, nella sezione successiva, faremo uno studio dei modelli grafici.

```
fit5 <- glm(Quality ~ Wind + Solar.R, family = binomial)
pseudoRfit5 <- ((fit5$null.deviance/-2) - (fit5$deviance/-2)) /
  ((fit5$null.deviance/-2))
summary(fit5)

##
## Call:
## glm(formula = Quality ~ Wind + Solar.R, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5844  -0.5775  -0.1998   0.2888   2.6887
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.685e+00  1.337e+00   2.008   0.0446 *
```

```
## Wind        -1.511e+00  3.257e-01  -4.639  3.5e-06 ***
## Solar.R      2.494e-07  1.010e-07   2.470   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.511  on 110  degrees of freedom
## Residual deviance:  75.178  on 108  degrees of freedom
## AIC: 81.178
##
## Number of Fisher Scoring iterations: 6
```

```
pseudoRfit5
```

```
## [1] 0.4104149
```

```
confint(fit5)
```

```
##                     2.5 %        97.5 %
## (Intercept)  1.461402e-01  5.461247e+00
## Wind        -2.235932e+00 -9.448807e-01
## Solar.R      6.769915e-08  4.702243e-07
```

Come possiamo osservare dall'output, la variabile 'Solar.R' risulta significativa e dunque molto probabilmente non è 'Wind' la variabile che media l'informazione di 'Solar.R'. In generale, i valori parametrici risultano simili a quelli ottenuti nei modelli precedenti mentre l'indice R^2 è basso. Passiamo all'ultima analisi.

Modello ridotto 'Temp' + 'Solar.R' Adesso analizziamo il modello composto da 'Temp' e 'Solar.R' per capire se l'informazione di dei raggi solari viene mediata dalla variabile temperatura.

```
fit6 <- glm(Quality ~ Temp + Solar.R, family = binomial)
pseudoRfit6 <- ((fit6$null.deviance/-2) - (fit6$deviance/-2)) /
  ((fit6$null.deviance/-2))
summary(fit6)
```

```
##
## Call:
## glm(formula = Quality ~ Temp + Solar.R, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01489  -0.29998  -0.04214   0.09459   2.30495
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.563e+01  5.536e+00  -4.630 3.65e-06 ***
## Temp         8.186e-01  1.861e-01   4.400 1.08e-05 ***
## Solar.R      1.856e-07  1.300e-07   1.428    0.153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 127.511  on 110  degrees of freedom
```

```
## Residual deviance:  49.871  on 108  degrees of freedom
## AIC: 55.871
##
## Number of Fisher Scoring iterations: 7

pseudoRfit6

## [1] 0.6088855

confint(fit6)

##                      2.5 %        97.5 %
## (Intercept) -3.861607e+01 -1.652233e+01
## Temp         5.127572e-01  1.253709e+00
## Solar.R     -5.248198e-08  4.687498e-07
```

Com'è possibile notare, la variabile 'Solar.R' **perde totalmente di significatività** se marginalizzata rispetto a 'Temp'. Molto probabilmente, l'informazione di 'Solar.R' è mediata dalla variabile 'Temp'. Questa ipotesi la verificheremo studiando i modelli grafici nella prossima sezione, dove ci aspettiamo che il DAG abbia una serie di nodi 'Solar.R' $\longrightarrow$ 'Temp' $\longrightarrow$ 'Quality' al suo interno.

4.3 Scelta del modello

Dai risultati precedenti, secondo il principio di **parsimonia**, il modello migliore per predire la qualità dell'aria date le variabili esplicative 'Temp', 'Wind' e 'Solar.R' è quello **ridotto composto da 'Wind' e 'Temp'**. Ha ottimi valori di R^2, le variabili sono tutte altamente significative ed informative e risulta semplice da interpretare. Oltre a questo criterio, che può peccare di soggettività, vorrei avvalorare la scelta fatta attraverso un metodo più oggettivo, come ad esempio il **test del rapporto di verosimiglianza**.

Test del rapporto di verosimiglianza Come annunciato precedentemente, faremo un'analisi del test del rapporto di verosimiglianza tra il modello ridotto 'Wind' + 'Temp' e quello completo, che sono i due modelli che hanno presentato i risultati migliori. Per farlo, abbiamo utilizzato sempre il metodo **lrtest** del pacchetto **lmtest** e con ipotesi nulla la scelta del modello ridotto.

```
library(lmtest)
lrtest(fit4, fit3) #nested vs complex

## Likelihood ratio test
##
## Model 1: Quality ~ Temp + Wind
## Model 2: Quality ~ Temp + Wind + Solar.R
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -18.838
## 2   4 -17.739  1 2.1986     0.1381
```

Anche dal test del rapporto di verosimiglianza, possiamo evincere che il modello ridotto sia la scelta migliore, poiché i risultati ottenuti dal test non sono sufficienti per rigettare l'ipotesi nulla.

4.4 Analisi dei plot del modello logistico ridotto 'Temp' + 'Wind'

Adesso, dopo aver selezionato il modello ridotto 'Temp' + 'Wind', è possibile fare il plot dei valori predetti dal modello scelto. Per farlo, come prima, utilizziamo la funzione **ggpredict** del pacchetto **ggeffects**, la quale prende in input il modello in questione, **fit4**, e le due variabili esplicative 'Wind' e 'Temp'. Faremo due grafici, il primo marginalizzando per 3 valori di 'Wind' e tenendo variabile 'Temp', viceversa per il secondo. In Figura 14 e in Figura 15è possibile apprezzarne l'output.

```
library(ggeffects)
plot(ggpredict(fit4,c("Temp", "Wind")))
plot(ggpredict(fit4,c("Wind","Temp")))
```

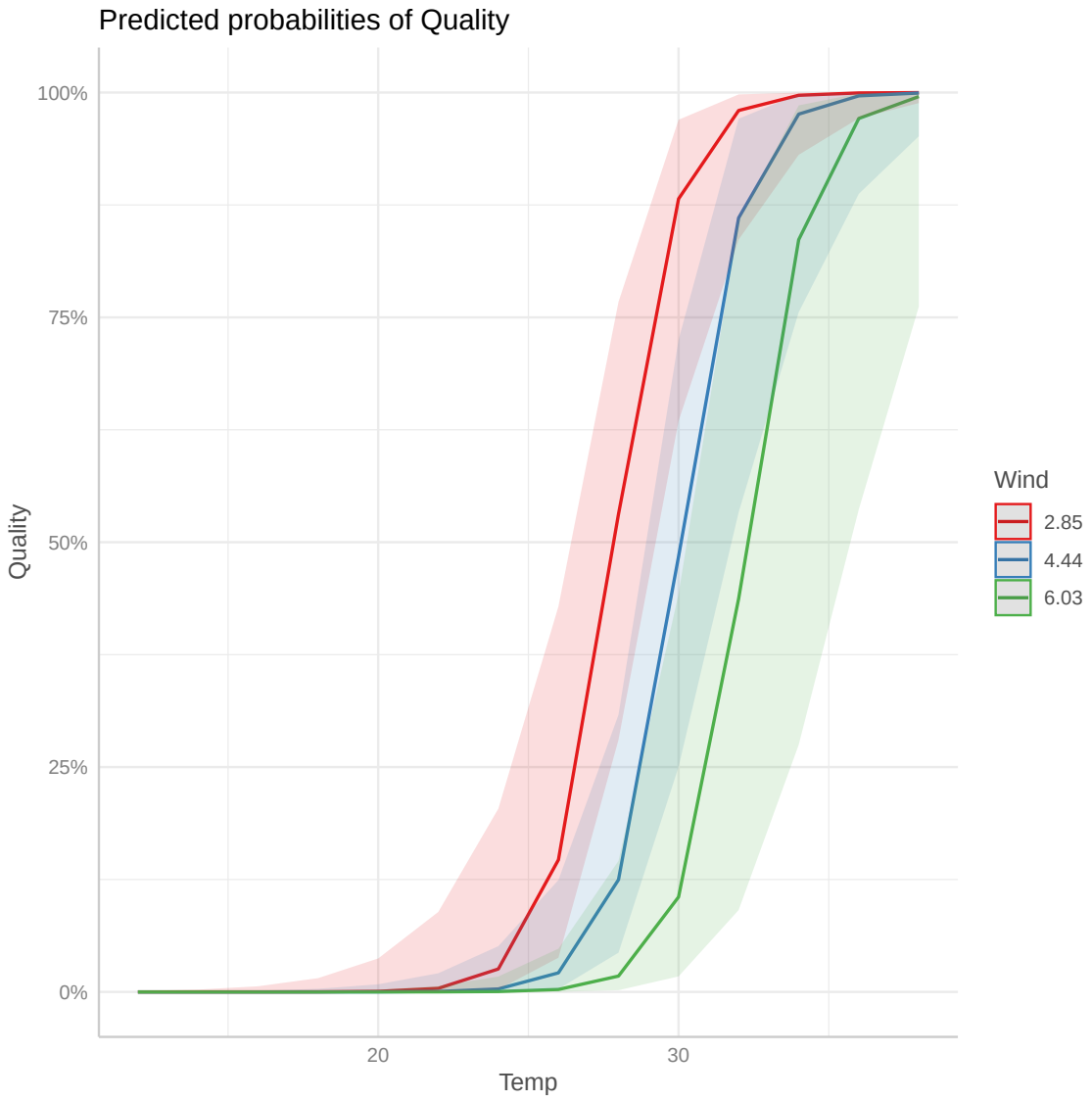


Figura 14: Plot delle probabilità che l'aria sia di cattiva qualità predette dal modello 'Temp' + 'Wind', marginalizzando per la variabile 'Temp'

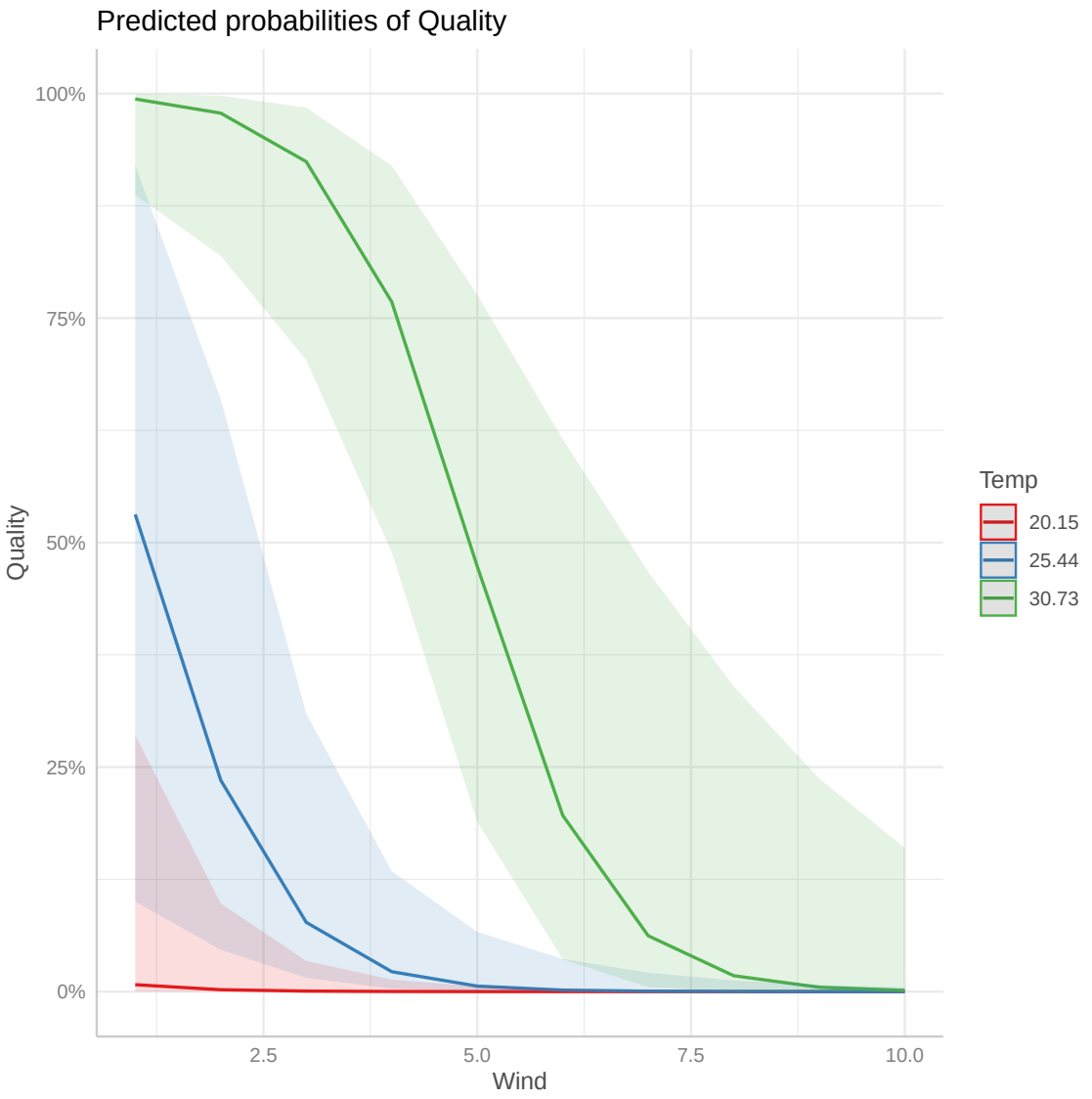


Figura 15: Plot delle probabilità che l'aria sia di cattiva qualità predette dal modello 'Temp' + 'Wind', marginalizzando per la variabile 'Wind'

Commento in riferimento alla Figura 14:

- la prima osservazione che si può fare è che, a circa 25 gradi centigradi, anche a basse velocità del vento, la probabilità che l'aria sia di cattiva qualità è identica a quella che non lo sia. Da quella temperatura in giù, la probabilità che l'aria sia di cattiva qualità inizia ad abbassarsi;
- la velocità del vento gioca un ruolo importante nel mitigare la probabilità che l'aria sia di cattiva qualità, infatti, possiamo notare che, con 'Temp' = 30 gradi centigradi, con la velocità del vento alta (curva verde) la probabilità è molto bassa, mentre con velocità del vento bassa (curva rossa) la probabilità che l'aria di scarsa qualità è decisamente elevata;
- in presenza di velocità del vento forte, occorre che ci sia un'elevata temperatura per aumentare la probabilità di scarsa qualità dell'aria;
- esiste una soglia di temperatura oltre la quale, a qualsiasi velocità del vento, la probabilità di successo è maggiore di quella di insuccesso, ed è circa 33 gradi centigradi.

Commento in riferimento alla Figura 15

- possiamo subito osservare che esiste una temperatura limite, sotto la quale, a qualsiasi velocità del vento, la probabilità di avere scarsa qualità d'aria, è nulla, anche con velocità del vento pari a 0 m/s;
- anche con temperatura di 25.44 gradi centigradi, la probabilità di avere scarsa qualità parte da un livello massimo tale per cui la probabilità di successo è uguale a quella di insuccesso, tale per cui inizia a scendere con l'aumentare della velocità del vento;
- per osservare un fenomeno più simile a quello di una sigmoide, occorre avere una temperatura maggiore o uguale a 30.73, e possiamo notare come inizialmente, fino ad una velocità del vento di circa 5.0 m/s la probabilità di successo supera quella di insuccesso, per poi tendere a zero con velocità del vento maggiori.

5 Selezione del modello attraverso metodi di penalizzazione

Lo studio fatto finora aveva l'obiettivo, nel primo caso, di capire quali variabili tra 'Solar.R', 'Wind' e 'Temp' influenzassero la variabile obiettivo 'Ozone' e come. Studiando il modello di regressione lineare (e poi polinomiale) abbiamo trovato che il modello migliore che adattasse bene i dati è quello completo. Poi, in seconda battuta, abbiamo introdotto una variabile binaria 'Quality' che indicasse quando l'aria fosse di scarsa qualità o meno, rispetto alla soglia limite dell'O.M.S. di 120 $\mu g/m^3$, e abbiamo studiato l'interazione tra le variabili esplicative precedenti e la probabilità che l'aria fosse di scarsa qualità, ovvero che 'Quality' fosse uguale a uno. Abbiamo selezionato il modello secondo il principio di parsimonia, scegliendo il modello ridotto composto dalle variabili esplicative 'Temp' e 'Wind'.

Concluse le precedenti analisi, in questa sezione, vogliamo effettuare la selezione del modello attraverso i metodi ***stepwise*** con diversi criteri e direzioni, per verificare che le scelte fatte sinora siano coerenti con i risultati proposti dagli algoritmi di selezione del modello. In particolare, abbiamo testato il modello di regressione lineare multipla e quello di regressione logistica multipla, per le direzioni **backward**, **forward** e **both** e utilizzando tre criteri di selezione: **likelihood**, **AIC** e **BIC**. Il primo criterio seleziona sempre il modello completo, il secondo favorisce i modelli che prevedono meglio e l'ultimo, invece, favorisce quelli più parsimoniosi.

- **likelihood**: k = 0;
- **AIC**: k = 2;
- **BIC**: k = log(length(variabile_obiettivo)).

Sono state testate tutte le 9 combinazioni possibili per ciascuna tipologia di modello, per un totale di 18 prove.

5.1 Selezione del modello di regressione lineare multipla

Per praticità presenteremo soltanto il punto iniziale e finale della ricerca del modello, omettendo i passaggi intermedi.

```
#Modello Completo
mq <- lm(Ozone ~ Solar.R + Wind + Temp)
#Modello Nullo
mq0 <- lm(Ozone ~ 1)

## FORWARD ##
forw_lik <- step(mq0, scope=formula(mq), direction = "forward", k=0)

## Start:  AIC=930.95
## Ozone ~ 1
##
## Step:  AIC=827.6
## Ozone ~ Temp + Wind + Solar.R

forw_aic <- step(mq0, scope=formula(mq), direction = "forward", k=2)

## Start:  AIC=932.95
## Ozone ~ 1
## Step:  AIC=835.6
## Ozone ~ Temp + Wind + Solar.R

forw_bic <- step(mq0, scope=formula(mq), direction = "forward",
k=log(length(Ozone)))
```

```
## Start:  AIC=935.66
## Ozone ~ 1
## Step:  AIC=846.43
## Ozone ~ Temp + Wind + Solar.R

## MISTO ##
both_lik <- step(mq, scope=formula(mq), direction = "both", k=0)

## Start:  AIC=827.6
## Ozone ~ Solar.R + Wind + Temp

both_aic <- step(mq, scope=formula(mq), direction = "both", k=2)

## Start:  AIC=835.6
## Ozone ~ Solar.R + Wind + Temp

both_bic <- step(mq, scope=formula(mq), direction = "both",
k=log(length(Ozone)))

## Start:  AIC=846.43
## Ozone ~ Solar.R + Wind + Temp

## BACKWARD ##
back_lik <- step(mq, scope=formula(mq0), direction = "backward", k=0)

## Start:  AIC=827.6
## Ozone ~ Solar.R + Wind + Temp

back_aic <- step(mq, scope=formula(mq0), direction = "backward", k=2)

## Start:  AIC=835.6
## Ozone ~ Solar.R + Wind + Temp

back_bic <- step(mq, scope=formula(mq0), direction = "backward",
k=log(length(Ozone)))

## Start:  AIC=846.43
## Ozone ~ Solar.R + Wind + Temp
```

Come possiamo constatare dall'output del codice, tutte le prove restituiscono come miglior scelta il **modello completo**, coerentemente con quanto scelto nel nostro studio iniziale.

5.2 Selezione del modello di regressione logistica

Come fatto precedentemente, riporteremo nel codice soltanto il punto di partenza e quello di arrivo, per ragioni di praticità.

```
#Modello Completo
fit <- glm(Quality ~ Solar.R + Wind + Temp, family = binomial)

#Modello Nullo
fit0 <- glm(Quality ~ 1, family = binomial)

## FORWARD ##
```

```
reg_forw_lik <- step(fit0, scope=formula(fit), direction = "forward", k=0)

## Start:  AIC=127.51
## Quality ~ 1
##
## Step:  AIC=35.48
## Quality ~ Temp + Wind + Solar.R

reg_forw_aic <- step(fit0, scope=formula(fit), direction = "forward", k=2)

## Start:  AIC=129.51
## Quality ~ 1
##
## Step:  AIC=43.48
## Quality ~ Temp + Wind + Solar.R

reg_forw_bic <- step(fit0, scope=formula(fit), direction = "forward",
k=log(length(Quality)))

## Start:  AIC=132.22
## Quality ~ 1
##
## Step:  AIC=51.8
## Quality ~ Temp + Wind

## MISTO ##

reg_both_lik <- step(fit, scope=formula(fit), direction = "both", k=0)

## Start:  AIC=35.48
## Quality ~ Solar.R + Wind + Temp

reg_both_aic <- step(fit, scope=formula(fit), direction = "both", k=2)

## Start:  AIC=43.48
## Quality ~ Solar.R + Wind + Temp

reg_both_bic <- step(fit, scope=formula(fit), direction = "both",
k=log(length(Quality)))

## Start:  AIC=54.32
## Quality ~ Solar.R + Wind + Temp
##
## Step:  AIC=51.8
## Quality ~ Wind + Temp

## BACKWARD ##

reg_back_lik <- step(fit, scope=formula(fit0), direction = "backward", k=0)

## Start:  AIC=35.48
## Quality ~ Solar.R + Wind + Temp

reg_back_aic <- step(fit, scope=formula(fit0), direction = "backward", k=2)

## Start:  AIC=43.48
## Quality ~ Solar.R + Wind + Temp
```

```
reg_back_bic <- step(fit, scope=formula(fit0), direction = "backward",
k=log(length(Quality)))

## Start:  AIC=54.32
## Quality ~ Solar.R + Wind + Temp
##
## Step:  AIC=51.8
## Quality ~ Wind + Temp
```

Dall'output del codice, in tutte le ricerche con criterio **likelihood** e **AIC**, il modello scelto è sempre quello **completo**: nel primo caso, perché il criterio likelihood seleziona sempre il modello completo, mentre nel secondo quello che prevede meglio, che in questo caso risulta essere quello completo. Tuttavia, tutte le ricerche effettuare con il criterio **BIC** restituiscono il modello ridotto 'Temp' + 'Wind', ovvero quello più parsimonioso. Questa scelta è coerente con quella fatta precedentemente durante l'analisi del modello logistico e possiamo ritenerci soddisfatti della scelta fatta.

6 Modelli grafici: undirected graphs e DAG

In questa sezione vogliamo studiare, attraverso i modelli grafici, le relazioni che intercorrono tra le variabili del nostro problema. Durante le analisi precedenti, abbiamo studiato come e quali variabili 'Solar.R', 'Wind' e 'Temp' influenzassero le variabili obiettivo 'Ozone', in caso di studio di regressione, oppure 'Quality' in caso di studio di classificazione. Durante gli studi condotti, abbiamo osservato alcuni fenomeni che potevano farci pensare a delle indipendenze condizionate tra le variabili in gioco: ad esempio, durante lo studio di regressione logistica, abbiamo sollevato l'ipotesi che l'informazione contenuta in 'Solar.R' venisse mediata dalla variabile 'Temp' ed è per questo che nel modello completo 'Solar.R' perdeva di significatività.

Dunque, con l'analisi dei modelli grafici vogliamo provare a rispondere ad alcune domande:

- che relazioni esistono tra le variabili del data set?
- ci sono indipendenze condizionali o marginali?
- alcune variabili mediano le informazioni di altre?

In prima fase, abbiamo provato ad analizzare il problema utilizzando i modelli grafici composti da **grafi non direzionati (undirected graphs)**, mentre in seconda fase, abbiamo utilizzato le **reti Bayesiane** basate su **grafici aciclici direzionati (DAG)**.

6.1 Modelli grafici basati su undirected graphs

Gli strumenti utilizzati per operare con i grafi aciclici non direzionati sono le librerie **gRbase**, **gRain** e **gRim**.

Prima di procedere, creiamo un secondo data frame che al posto di 'Ozone' contiene la variabile 'Quality', che chiamiamo **data_quality**, per poter suddividere l'analisi per la regressione (data set: data) e classificazione (data set: data_quality).

```
str(data)

## 'data.frame': 111 obs. of  5 variables:
##  $ Ozone  : num  82 72 24 36 46 38 16 32 22 28 ...
##  $ Solar.R: num  7949600 4937120 6234160 13095920 12510160 ...
##  $ Wind   : num  3.3 3.57 5.62 5.13 3.84 ...
##  $ Temp   : num  19.4 22.2 23.3 16.7 18.3 ...
##  $ Quality: num  0 0 0 0 0 0 0 0 0 0 ...

data_quality <- data[,-1]
str(data_quality)

## 'data.frame': 111 obs. of  4 variables:
##  $ Solar.R: num  7949600 4937120 6234160 13095920 12510160 ...
##  $ Wind   : num  3.3 3.57 5.62 5.13 3.84 ...
##  $ Temp   : num  19.4 22.2 23.3 16.7 18.3 ...
##  $ Quality: num  0 0 0 0 0 0 0 0 0 0 ...

data <- data[,-5]
```

Undirected graphs per modelli di regressione Per prima cosa abbiamo studiato i modelli grafici basati su grafi non direzionati per i modelli di regressione lineare, utilizzando il data seta **data**, per due diverse direzioni, **backward** e **forward**, e per due diversi criteri di selezione **AIC** e **BIC**. Per entrambe le direzioni scelte, i modelli grafici prodotti con il criterio AIC sono gli stessi e, analogamente, anche quelli prodotti con il criterio BIC sono gli stessi, sia per backward che per forward. Inoltre, nella direzione di ricerca forward, i modelli grafici trovati con **type = "unrestricted"** e **type = "decomposable"** sono i medesimi. Dunque, distinguiamo due modelli grafici, quelli trovati con il criterio di selezione AIC in Figura 16 e quelli trovati

con il criterio di selezione BIC in Figura 17. Di seguito riportiamo il codice in R completo di tutte le ricerche effettuate.

```
library(gRain)
library(gRim)
library(gRbase)

### BACKWARD UNDIRECTED GRAPH MODELS ###
sat.data <- cmod(~.^., data = data)

## AIC ##
m.data <- stepwise(sat.data)
plot(m.data)
```

```
## BIC ##
m.data1 <- stepwise(sat.data, k = log(sum(data)))
plot(m.data1)
```

```
### FORWARD UNDIRECTED GRAPH MODELS ###
ind.data <- cmod(~.^1, data = data)

## AIC ##
m.data2 <- stepwise(ind.data, k=2, type = "unrestricted",
                    direction = "forward", details = 0)
#plot(m.data2)

m.data3 <- stepwise(ind.data, k=2, type = "decomposable",
                    direction = "forward", details = 0)
#plot(m.data3)

## BIC ##
m.data4 <- stepwise(ind.data, k = log(sum(data)),
                    type = "unrestricted",
                    direction = "forward", details = 0)
#plot(m.data4)

m.data5 <- stepwise(ind.data, k = log(sum(data)),
                    type = "decomposable",
                    direction = "forward", details = 0)
#plot(m.data5)
```

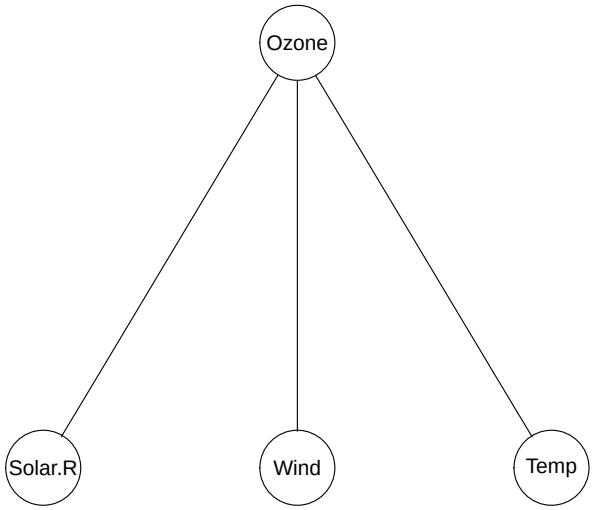


Figura 16: Modello grafico basato su grafo non direzionato trovato con AIC

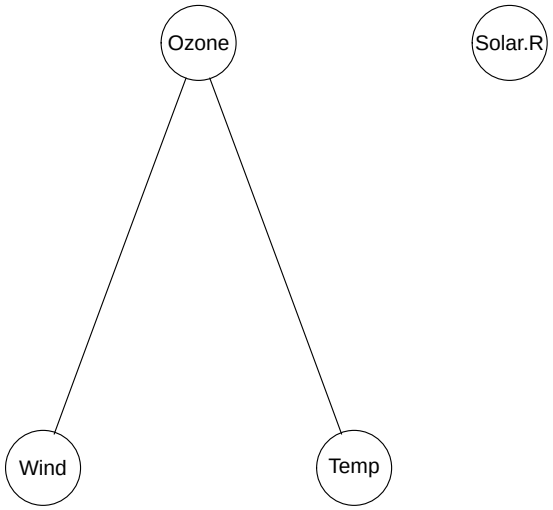


Figura 17: Modello grafico basato su grafo non direzionato trovato con BIC

Riferendoci alla **Figura 16**, possiamo notare che il grafo contiene tutte le variabili del data set, come il modello che abbiamo trovato per la regressione lineare, e lo stesso suggerito dalla selezione dei modelli stepwise con criterio AIC. Notiamo che tutte e tre le variabili esplicative sono condizionate tra di loro dato 'Ozone', poiché 'Ozone' è un nodo separatore che separa tutti i cammini i cui estremi hanno due variabili esplicative scelte tra 'Solar.R', 'Wind' e 'Temp'.

Invece, in riferimento alla **Figura 17**, notiamo che il grafo trovato ha un nodo totalmente separato dagli altri che è 'Solar.R' e dunque risulta essere indipendente marginalmente dalle altre variabili del modello. Il criterio BIC favorisce modelli più parsimoniosi e dunque, probabilmente, per questo motivo, 'Solar.R' è stato escluso dal modello. D'altronde, come abbiamo constato precedentemente, la variabile in questione era la meno significativa tra tutte e quella che apportava meno contributo in termini assoluti al valore dell'Ozono nell'aria. Inoltre, come abbiamo anche osservato durante lo studio del modello di regressione lineare, tutte le volte che toglievamo 'Solar.R' dal modello, le stime dei parametri di 'Wind' e 'Temp' **non variavano**, o la facevano impercettibilmente, e questo modello grafico potrebbe spiegare il fenomeno osservato.

Undirected graphs per modelli di classificazione Dopo aver analizzato i modelli grafici per il problema di regressione, abbiamo fatto lo stesso studio per i modelli di classificazione, utilizzando il data set **data_quality**. Come fatto precedentemente, abbiamo effettuato lo studio per le due diverse direzioni, **backward** e **forward**, e per due diversi criteri di selezione **AIC** e **BIC**. Come prima, i modelli grafici trovati con AIC sono gli stessi sia per le direzioni forward che backward, e stessa cosa per quelli trovati con BIC. Perciò, distinguiamo le due classi di modelli grafici che abbiamo ottenuto con AIC in Figura 18 da quelli trovati con BIC in Figura 19. Di seguito riportiamo il codice in R completo di tutte le ricerche effettuate.

```
### BACKWARD UNDIRECTED GRAPH MODELS ###
sat.data_quality <- cmod(~.^., data = data_quality)

## AIC ##
m.data_quality <- stepwise(sat.data_quality)
plot(m.data_quality)
```

```
## BIC ##
m.data_quality1 <- stepwise(sat.data_quality, k = log(sum(data_quality)))
plot(m.data_quality1)
```

```
### FORWARD UNDIRECTED GRAPH MODELS ###
ind.data_quality <- cmod(~.^1, data = data_quality)

### AIC ##
m.data_quality2 <- stepwise(ind.data_quality, k=2,
                            type = "unrestricted",
                            direction = "forward", details = 0)
#plot(m.data_quality2)

m.data_quality3 <- stepwise(ind.data_quality, k=2,
                            type = "decomposable",
                            direction = "forward", details = 0)
#plot(m.data_quality3)

## BIC ##
m.data_quality4 <- stepwise(ind.data_quality,
                            k = log(sum(data)),
                            type = "unrestricted",
                            direction = "forward", details = 0)
#plot(m.data_quality4)
```

```
m.data_quality5 <- stepwise(ind.data_quality,
                            k = log(sum(data)),
                            type = "decomposable",
                            direction = "forward", details = 0)
#plot(m.data_quality5)
```

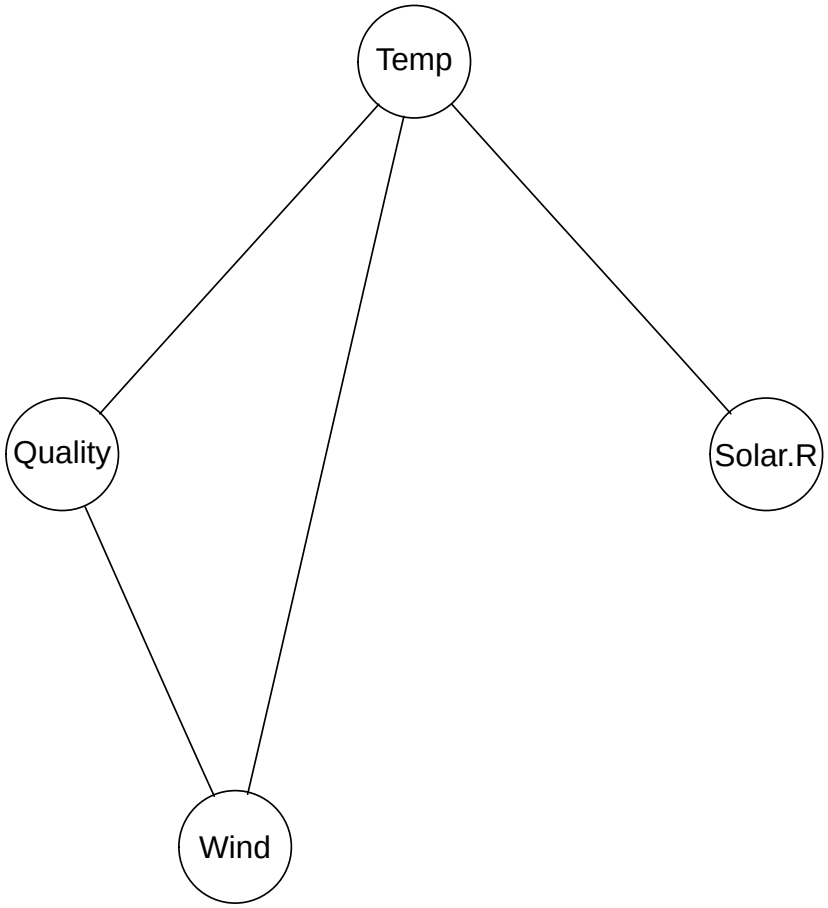


Figura 18: Modello grafico basato su grafo non direzionato trovato con AIC (logistic reg.)

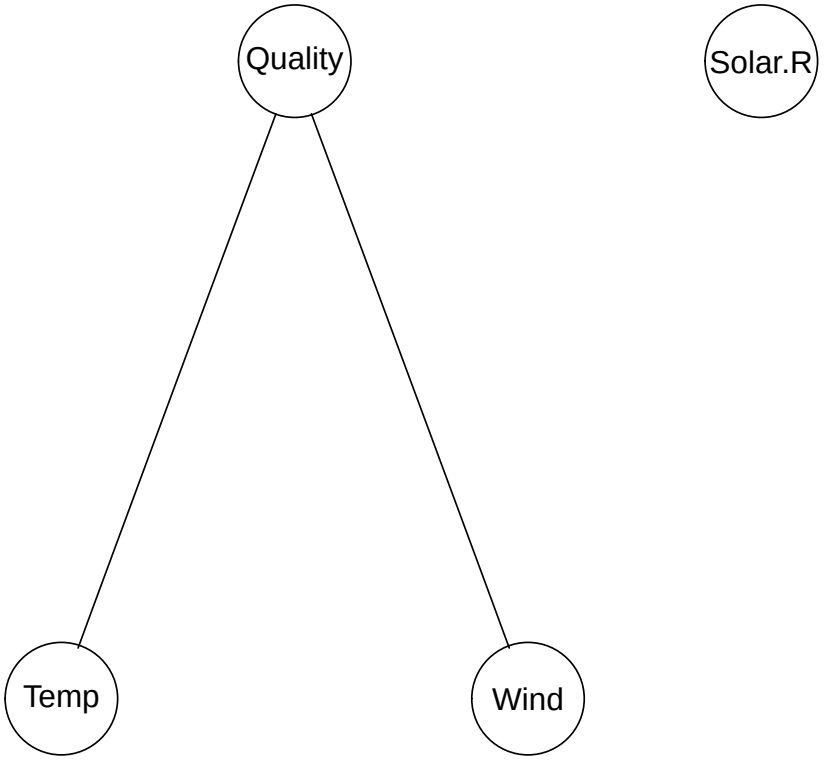


Figura 19: Modello grafico basato su grafo non direzionato trovato con BIC (logistic reg.)

In Figura 18 abbiamo la rappresentazione del grafo trovato attraverso i metodi AIC. Come possiamo notare, tutte le variabili sono comprese nel modello, come previsto dagli studi precedenti, e quindi coerente con i risultati trovati prima. Inoltre, i nodi 'Quality' e 'Solar.R' sono separati dal nodo 'Temp' e dunque, possiamo dire che 'Quality' è condizionatamente indipendente da 'Solar.R' dato 'Temp'. Notiamo, infine, una cricca composta dai nodi 'Quality', 'Temp' e 'Wind'.

Diversamente, in Figura 19, abbiamo il risultato della ricerca con criterio BIC. Il grafo, come visto anche per il modello lineare di regressione, mostra un nodo isolato dagli altri che è 'Solar.R'. Questa rappresentazione del modello è coerente con ciò che è stato trovato negli studi precedenti, dove, con criterio BIC, si favoriva il modello parsimonioso composto soltanto da 'Temp' e 'Wind' in qualità di variabili esplicative. Il fatto che 'Solar.R' sia isolato significa che è marginalmente indipendente dalle altre variabili del sistema, e, come visto anche per il modello lineare di regressione, si spiegherebbe il fatto per cui, togliendola, i parametri stimati delle altre variabili variano di poco o non variano affatto.

6.2 Modelli grafici basati su grafi aciclici direzionati DAG

Utilizzando i grafi non direzionati siamo riusciti a scorgere alcune delle relazioni che esistono tra le variabili del sistema. Tuttavia, per ottenere informazioni più precise ed esplicative, circa le suddette relazioni, occorre capire anche il verso di propagazione dell'informazione all'interno del sistema. Per questo motivo, in questa sotto sezione parleremo degli studi condotti attraverso i grafi aciclici direzionati. Per questa analisi abbiamo costruito una rete Byesiana partendo dal data set (dato che la struttura è nota) utilizzando la funzione **hc()** della libreria **bnlearn**. La funzione utilizza un algoritmo greedy che di default si avvale del criterio BIC.

DAG per modelli di regressione Partiamo cercando il modello grafico più adatto al modello di regressione.

```
library(bnlearn)


data.bn <- hc(data)
plot(as(amat(data.bn), "graphNEL"))
```

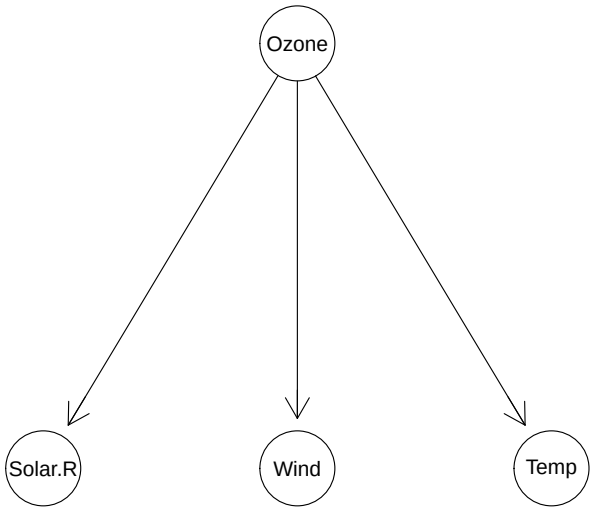


Figura 20: Modello grafico basato su *naive* Bayesian network DAG

Come possiamo vedere dalla Figura 20, il modello trovato è un modello *naive*, che suggerisce che le variabili esplicative 'Solar.R', 'Wind' e 'Temp' siamo direttamente influenzate da 'Ozone'. Occorre integrare il modello con una **conoscenza pregressa** per dargli un ordine gerarchico. Dividiamo il sistema in 3 livelli, dove al terzo livello poniamo la variabile obiettivo:

- livello 1: 'Solar.R';
- livello 2: 'Wind' e 'Temp';
- livello 3: 'Ozone'.

Costruiamo una matrice di adiacenza contenente gli archi non ammessi rispetto all'ordine presunto che abbiamo stabilito. La convertiamo in una data frame attraverso la funzione **get.edgelist()** e dopo, attraverso la funzione **hc()**, troviamo la rete Bayesiana data la conoscenza preliminare.

```
block <- c(3, 1, 2, 2)
blM <- matrix(0, nrow=4, ncol=4)
rownames(blM) <- colnames(blM) <- names(data)
for (b in 2:4) blM[block==b, block<b] <- 1
library(igraph)

blackL <- data.frame(get.edgelist(as(blM, "igraph")))
names(blackL) <- c("from", "to")

data.bn1 <- hc(data, blacklist=blackL)
plot(as(amat(data.bn1), "graphNEL"))
```

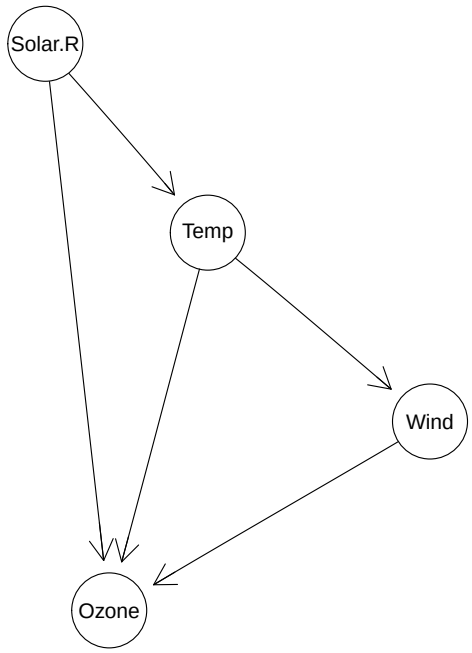


Figura 21: Modello grafico basato su Bayesian network DAG

In Figura 21 è riportato il DAG ottenuto secondo la procedura precedentemente descritta. La prima cosa che osserviamo è che il grafo contiene tutti i nodi, dunque il modello considerato è quello completo. Questo risultato è coerente con la scelta fatta precedentemente e con i modelli trovati mediante i metodi stepwise. Inoltre, osserviamo che 'Solar.R' ha un arco che connette direttamente la variabile ad 'Ozone' e che esiste anche il cammino 'Solar.R' $\longrightarrow$ 'Temp' $\longrightarrow$ 'Ozone'. Ciò significa che parte dell'informazione di 'Solar.R' è mediata dalla variabile 'Temp'. Per questo motivo, nel modello completo di regressione lineare, la stima del parametro dell'effetto dei raggi solari sul livello di Ozono nell'aria perde di significatività. Tuttavia, rimane comunque significativa. Infatti, nel grafo 'Solar.R' è direttamente connessa ad 'Ozone'. Notiamo, poi, che 'Wind' è condizionatamente indipendente da 'Solar.R' dato 'Temp': parte dell'informazione di 'Solar.R' è mediata anche dalla variabile 'Wind'. Infine, le variabili più significative risultavano essere proprio 'Wind' e 'Temp' che, nel grafo, sono direttamente connesse con la variabile 'Ozone'.

DAG per modelli di classificazione Come prima, cerchiamo il modello grafico più adatto al modello di regressione logistica.

```
data_quality.bn <- hc(data_quality)
plot(as(amat(data_quality.bn), "graphNEL"))
```

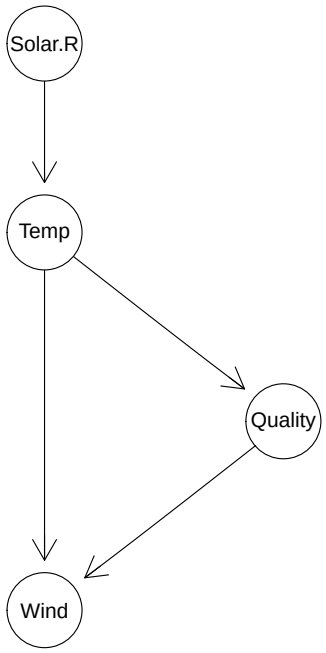


Figura 22: Modello grafico basato su *naive* Bayesian network DAG (logistic reg.)

Similmente a quanto ottenuto per il modello di regressione lineare, in Figura 22 possiamo constatare che il modello grafico è di tipo *naive*. Occorre stabilire un ordine gerarchico e utilizzare la stessa tecnica utilizzata in precedenza per trovare il modello grafico più adatto. Come prima, dividiamo il sistema in 3 livelli ponendo nel livello 3, il più profondo, la variabile obiettivo 'Quality':

- livello 1: 'Solar.R';
- livello 2: 'Wind' e 'Temp';
- livello 3: 'Quality'.

Applicando la stessa procedura di prima, troviamo la rete Bayesiana per il nostro data frame

```
block <- c(1, 2, 2, 3)
blM <- matrix(0, nrow=4, ncol=4)
rownames(blM) <- colnames(blM) <- names(data_quality)
for (b in 2:4) blM[block==b, block<b] <- 1
library(igraph)
blackL <- data.frame(get.edgelist(as(blM, "igraph")))
names(blackL) <- c("from", "to")

data_quality.bn1 <- hc(data_quality, blacklist=blackL)
plot(as(amat(data_quality.bn1), "graphNEL"))
```

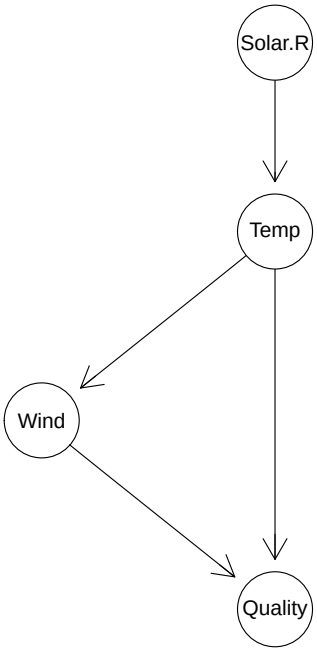


Figura 23: Modello grafico basato su Bayesian network DAG (logistic reg.)

In Figura 23 è riportato il DAG trovato dalla procedura sopra descritta. Innanzitutto, osserviamo che il DAG contiene tutti i nodi, ovvero che viene scelto un modello completo. Questo risultato è coerente con ciò che avevamo trovato e con i modelli trovati attraverso le procedure stepwise con criteri likelihood ed AIC. Inoltre, notiamo che 'Quality' è condizionatamente indipendente da 'Solar.R' data 'Temp', ovvero che tutta l'informazione di 'Solar.R' è mediata dalla variabile 'Temp'. Questo spiegherebbe perché, mentre nel modello ridotto 'Quality' 'Solar.R', la stima del parametro degli effetti di 'Solar.R' è significativa, in quello completo, invece, perda totalmente di significatività. Osserviamo, poi, 'Quality' è direttamente influenzata da 'Wind' e che una parte di informazione di 'Temp' è mediata da 'Wind'. Infatti, 'Temp' è connessa a 'Quality' sia direttamente che attraverso 'Wind' e inoltre, come abbiamo visto nello studio del modello di regressione logistica, nel modello 'Quality' 'Temp', la stima del parametro degli effetti della temperatura è altamente significativa mentre nel modello 'Quality' 'Temp' + 'Wind', è sempre altamente significativa, ma il valore del p-value aumenta.

6.3 Undirected Gaussian graph models

Non abbiamo ritenuto opportuno considerare nello studio i grafi indiretti Gaussiani e quindi l'utilizzo di matrici di concentrazione, poiché l'ipotesi a monte è che il vettore di variabili aleatorie fosse distribuito come una normale multivariata. Infatti, attraverso il Test di Shapiro-Wilk abbiamo constato che le uniche variabili con andamento normale sono 'Wind' e 'Temp'

```
### NORMALITY TEST USING SHAPIRO-WILK TEST ###

## IF P-VALUE > 0.05 WE CAN ASSUME NORMALITY
shapiro.test(data$Ozone)# NO

##
##  Shapiro-Wilk normality test
##
## data:  data$Ozone
## W = 0.87355, p-value = 2.846e-08

shapiro.test(data$Solar.R)# NO

##
##  Shapiro-Wilk normality test
##
## data:  data$Solar.R
## W = 0.93285, p-value = 2.957e-05

shapiro.test(data$Wind)# SI

##
##  Shapiro-Wilk normality test
##
## data:  data$Wind
## W = 0.98076, p-value = 0.1098

shapiro.test(data$Temp)# SI

##
##  Shapiro-Wilk normality test
##
## data:  data$Temp
## W = 0.98007, p-value = 0.0957

shapiro.test(data_quality$Quality)# NO

##
##  Shapiro-Wilk normality test
##
## data:  data_quality$Quality
## W = 0.5476, p-value < 2.2e-16
```

7 Conclusioni

In questo lavoro abbiamo voluto analizzare il dataset **"Airquality"** per studiare che relazioni ci fossero tra le variabili 'Ozone', intesa come variabile obiettivo, e le altre variabili esplicative 'Solar.R', 'Wind' e 'Temp' utilizzando i concetti di modellazione statistica appresi durante il corso di *Foundations of Statistical Modeling*.

In particolare, abbiamo studiato come l'intensità dei raggi solari nell'atmosfera, la velocità del vento e la temperatura potessero influire sui livelli di Ozono nell'aria, determinandone quindi la qualità.

Abbiamo trovato che, il modello di regressione lineare più adatto al modello è il modello di regressione completo. Tuttavia, per studiare un problema di regressione, è emerso che il modello più adatto, rispetto a quello lineare, è quello polinomiale.

Abbiamo poi introdotto una variabile binaria 'Quality' che indicasse la qualità dell'aria rispetto al livello di soglia di $\mu g/m^3$ indicato dall'O.M.S. e abbiamo studiato il problema di classificazione attraverso il modello di regressione logistica. Più precisamente, abbiamo analizzato come le variabili 'Solar.R', 'Wind' e 'Temp' influenzassero la probabilità condizionata che l'aria fosse di scarsa qualità, ovvero che la variabile obiettivo 'Quality' fosse uguale a uno.

Dallo studio condotto è emerso che, secondo il principio di parsimonia, il modello migliore di regressione logistica adatto al problema è quello ridotto della forma 'Quality' 'Wind' + 'Temp'. Anche il modello completo può essere utilizzato, nel caso si prediliga la capacità previsionale.

Per verificare le decisione prese, abbiamo poi condotto la ricerca del modello migliore attraverso gli algoritmi *stepwise* con 3 diversi criteri (likelihoood, AIC e BIC) e abbiamo confrontato i risultati ottenuti con le scelte fatte. Dai confronti è emerso che i risultati ottenuti sono coerenti con i modelli proposti dagli algoritmi di selezione del modello.

Infine, mediante i modelli grafici, abbiamo potuto analizzare e verificare le relazioni tra le variabili che compongono il modello: tra le varie informazioni ottenute, il risultato più interessante riguarda la relazione che c'è tra le variabili 'Temp', 'Solar.R' e 'Quality' nello studio di classificazione, ovvero che l'informazione di 'Solar.R' è totalmente mediata dalla variabile 'Temp' ed è per questo che la stima del parametro degli effetti dell'intensità dei raggi solari perde di totale significatività nel modello di regressione logistica completo.