Diagnostyka modelu liniowego

Filip Wichrowski

2023-11-08

Zbiór danych *marketing*

Przeprowadźmy diagnostykę modelu liniowego w oparciu o zbiór danych *marketing*. Zawiera on informacje dotyczące wydatków na kampanie reklamowe pewnego ubezpieczenia zdrowotnego w różnych mediach oraz liczbę członków, którzy w tych kampaniach marketingowych się zarejestrowali.

```
marketing = read.csv("marketing.csv", header = T)
head(marketing, 5)
```

```
##      TV Internet Mailing Members
## 1 460.2     75.6   138.4    44.2
## 2  89.0     78.6    90.2    20.8
## 3  34.4     91.8   138.6    18.6
## 4 303.0     82.6   117.0    37.0
## 5 361.6     21.6   116.8    25.8
```

Mamy następujące zmienne:

- TV – liczba dolarów (w tysiącach) przeznaczona na reklamy w TV,
- Internet – liczba dolarów (w tysiącach) przeznaczona na reklamy online,
- Mailing – liczba dolarów (w tysiącach) przeznaczona na reklamy mailowe,
- Members – liczba osób (w tysiącach), które zapisały się na ubezpieczenie zdrowotne.

Zmienną objaśnianą jest zmienna *Members*, natomiast za predyktory przyjmujemy pozostałe zmienne. Dopasujmy model liniowy

```
mod <- lm(Members ~ ., data = marketing)
```

Sprawdźmy, jak wygląda output funkcji `summary()`:

```
summary(mod)
```

```
##
## Call:
## lm(formula = Members ~ ., data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6554  -1.7816   0.4836   2.3786   5.6584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.877779   0.623816   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Internet     0.188530   0.008611  21.893   <2e-16 ***
## Mailing     -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.371 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

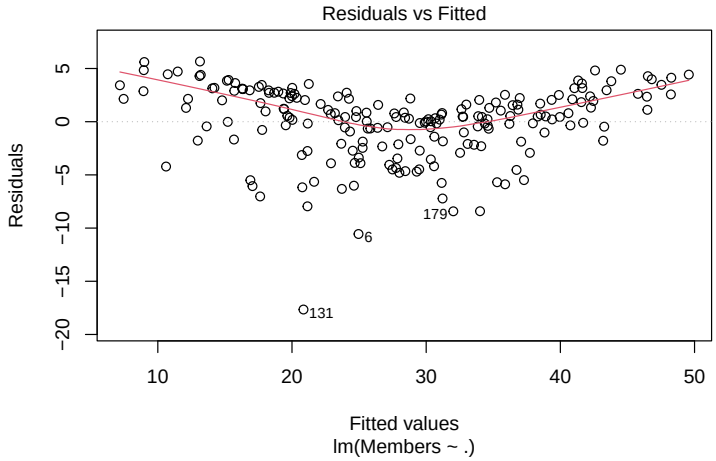
Co możemy zauważyć? Po pierwsze, warto spojrzeć na *Residuals*: idealnie, chcielibyśmy oczywiście, aby pochodziły z rozkładu normalnego, który jest symetryczny względem 0, o medianie równej wartości średniej i równej 0 (dla modelu z interceptem). W naszym wypadku mediana jest niezerowa, a rozkład wydaje się być lewostronnie skośny ($Q_3 - M < M - Q_1$ oraz $Q_3 - Q_1 > 0$). Następnie

- zmienne *TV, Internet* są istotne statystycznie na poziomie $\alpha = 0.05$, natomiast zmienna *Mailing* nie,
- nasz model jest istotnie lepszy, niż model wyłącznie z interceptem (p-wartość testu F poniżej ustalonego poziomu istotności)

Przypomnijmy jednak, że istotność statystyczna parametrów **nie** oznacza, iż nasz model jest dobrze dopasowany (por. Lab3, zad. 1a). Otrzymaliśmy współczynnik $R^2 = 0.8972$; pomimo, że jest to duża wartość, to dopóki wszystkie założenia modelu liniowego nie są spełnione oraz zależność rzeczywiście nie jest liniowa, to nie możemy interpretować go jako miary dopasowania modelu. Przejdźmy zatem to diagnostyki naszego modelu. W tym celu wykorzystamy funkcję `plot()`.

Residuals vs. fitted values

```
plot(mod, 1)
```



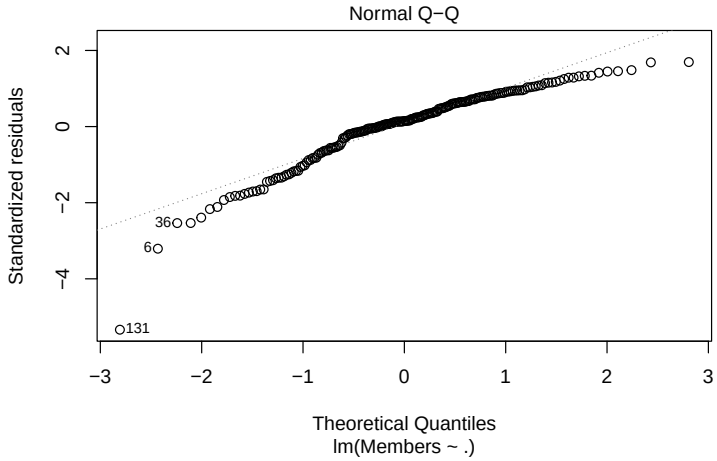
Wykres powyżej to wykres rezyduów e_i od wartości dopasowanych $\hat{y}_i$. Czerwona linia to LOWESS (locally weighted scatterplot smoothing) dopasowana do punktów. Wykres ten pozwala nam: zidentyfikować heteroskedastyczność, czyli niestałość wariancji rezyduów, ocenić ich przybliżoną normalność, a trend wyznaczany przez czerwoną linię może zasugerować potencjalną transformację zmiennej odpowiedzi/predyktoraów. Widoczne również mogą być obserwacje odstające.

Dla dobrze dopasowanego modelu chcielibyśmy, aby ten wykres maksymalnie przypominał chmurę punktów bez żadnej struktury, symetryczną względem prostej $y = 0$, bez żadnego widocznego trendu. Jeśli obecny jest trend, to być może zmienna odpowiedzi lub predyktory mogą wymagać transformacji, np. transformacji potęgowej. Jeśli rozrzut rezyduów ewidentnie zmienia się wraz ze zmianą wartości dopasowanych, to jest możliwe, że mamy do czynienia z heteroskedastycznością. Zazwyczaj skuteczniej możemy zidentyfikować heteroskedastyczność przy pomocy (omówionego w dalszej części) wykresu scale-location. Warto odnotować, że regresja liniowa jest całkiem odporna na heteroskedastyczność i w praktyce nie ma na nią dużego wpływu, gdy stosunek wariancji maksymalnej do wariancji minimalnej nie przekracza 4 (ponownie, to tylko heurystyka).

W naszym przypadku mamy widoczną umiarkowaną zależność rezyduów od wartości dopasowanych. Może to zasugerować konieczność wykonania przekształcenia zmiennej odpowiedzi (np. w tym wypadku pierwiastek) albo predyktora(ów), czyli dołączenie np. zależności kwadratowej. Wydaje się, że nie mamy istotnej heteroskedastyczności: rozrzut rezyduów jest z grubsza stały, może z wyjątkiej wartości obszaru dla $\hat{y} \geq 40$. Interesująca wydaje się obserwacja 131, która może być potencjalnie odstająca.

QQ plot

```
plot(mod, 2)
```

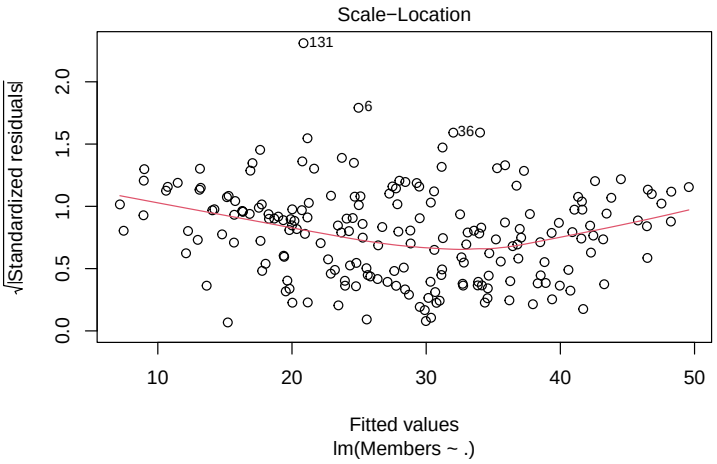


Wykres powyżej to wykres kwantyl-kwantyl rezyduów studentyzowanych r_i od kwantyli teoretycznych rozkładu $N(0,1)$. Może posłużyć nam do oceny założenia o normalności rezyduów (pożądane jest, aby punkty układały się możliwie najbliżej linii przerywanej), a także poszukiwania obserwacji odstających.

W naszym wypadku rozkład rezudyów studentyzowanych wydaje się być lewostronnie skośny (por. analiza QQ plotów na ostatnich zajęciach), zatem transformacja np. zmiennej odpowiedzi może być konieczna. Ponownie jednak interesująca wydaje się być obserwacja 131, dla której rezyduum ma istotnie dużą wartość. Ponadto, mamy kilka obserwacji (np. $6, 36$), dla których $|r_i| > 2$.

Scale-location

```
plot(mod, 3)
```



Wykres ten również pozwala nam ocenić naruszenia założenia o homoskedastyczności. Dla dobrze dopasowanego modelu chcemy mieć chmurę punktów bez żadnej struktury. Jeśli obecny jest trend (wyznaczany przez czerwoną linię LOWESS), to możemy mieć naruszenie założenia o homoskedastyczności. W naszym wypadku raczej nie ma istotnego problemu z heteroskedastycznością, natomiast (ponownie) może być konieczna transformacja. Problematyczna w tym kontekście może być niewystarczająca liczba punktów w skrajnych częściach wykresu.

W oparciu o ten wykres, możemy również zidentyfikować potencjalne obserwacje odstające. W naszym przypadku zdecydowanie wybija się obserwacja 131. Przypomnijmy, że heurystykami do poszukiwania obserwacji odstających są

$$|r_i| > 2 \tag{1}$$
$$|t_i| > 2 \tag{2}$$

Ważna obserwacja dotyczy tego, że oś pionowa jest *pierwiastkiem* z $|r_i|$. Spróbujmy znaleźć obserwacje potencjalnie odstające:

```
as.numeric(which(abs(rstandard(mod)) > 2))
```

```
## [1]   6  26  36  79 127 131 179
```

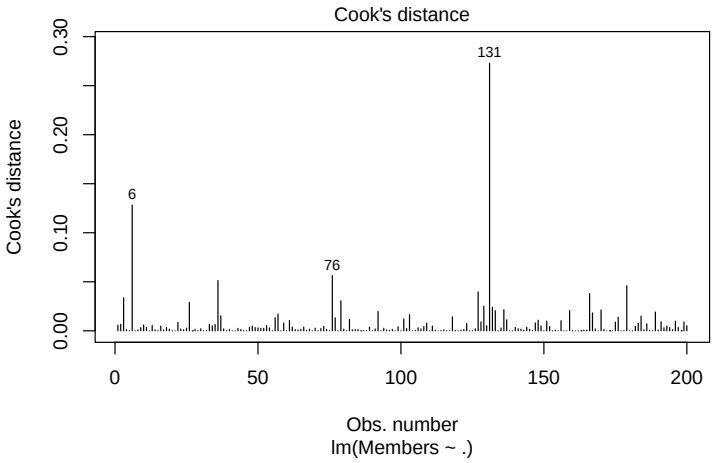
```
as.numeric(which(abs(rstudent(mod)) > 2))
```

```
## [1]   6  26  36  79 127 131 179
```

Zarówno rezydua studentyzowane (`rstandard()`), jak i studentyzowane modyfikowane (`rstudent()`) zidentyfikowały obserwację: $6, 26, 36, 79, 127, 131, 179$ jako potencjalnie odstające.

Cook's distance

```
plot(mod, 4)
```



Wykres ten nazywany jest potocznie diagramem Cooka – przedstawia on odległości Cooka (czyli jak średnio przewidywane wartości $\hat{y}$ przesuną się, jeśli i-ta obserwacja zostanie usunięta ze zbioru danych) w funkcji indeksów obserwacji. Przypomnijmy heurystyki służące do określania, czy dana obserwacja jest wpływowa

$$D_i > 1 \tag{3}$$
$$D_i > 4/n \tag{4}$$
$$D_i > 4/(n-p) \tag{5}$$

W naszym przypadku widzimy, że szczególnie obserwacja o indeksie 131 wybija się ponad pozostałe (obserwacja o indeksie 6 również wizualnie odstaje od reszty obserwacji). Sprawdźmy, jaka jest wartość Cooka z nią związana

```
D = cooks.distance(mod)
as.numeric(which(D == max(D))) # indeks obserwacji o najwiekszej odl. Cooka
```

```
## [1] 131
```

```
max(D)
```

```
## [1] 0.2729561
```

Otrzymujemy $D_{131} = 0.27$. Sprawdźmy, czy któraś heurystyka wykryje ją jako wpływową

```
X = model.matrix(mod)

n = nrow(X)
p = ncol(X)

max(D) > 1
```

```
## [1] FALSE
```

```
max(D) > 4/n
```

```
## [1] TRUE
```

```
max(D) > 4/(n-p)
```

```
## [1] TRUE
```

Widzimy, że ostatnie dwie heurystyki wskazują na to, iż taka obserwacja rzeczywiście jest wpływowa. Możemy również spojrzeć które obserwacje mają odległość Cooka większą, niż wartości $4/n$ oraz $4/(n-p)$

```
as.numeric(which(D > 4/n))
```

```
##  [1]   3   6  26  36  76  79 127 129 131 132 133 136 159 166 170 179
```

```
as.numeric(which(D > 4/(n-p)))
```

```
##  [1]   3   6  26  36  76  79 127 129 131 132 133 136 159 166 170 179
```

Jest ich całkiem sporo. Warto przypomnieć w tym miejscu, że heurystyki są pewnymi **sugestiami**, a nie formalnymi testami. Dlatego tak ważna jest wizualizacja wyników, którą przeprowadzamy.

Następnie, możemy skorzystać z dźwigni h_{ii}. Przypomnijmy, że heurystyka tutaj to

$$h_{ii} > 2p/n \tag{6}$$

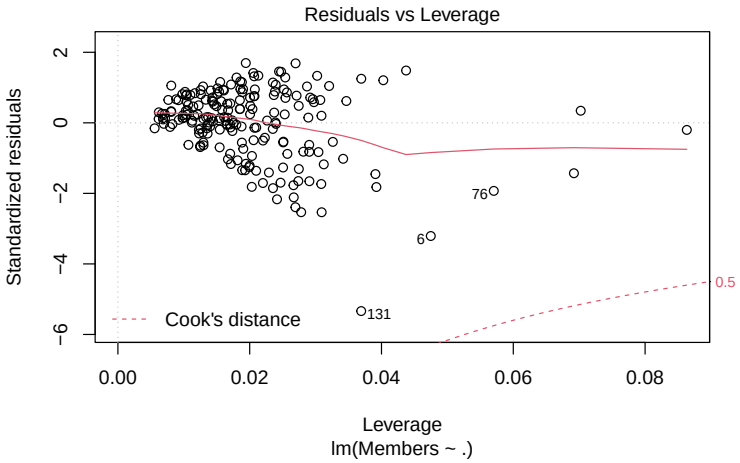
```
h = hatvalues(mod)

as.numeric(which(h > 2*p/n))
```

```
## [1]   6  17  37  76 102 129 166
```

Otrzymaliśmy w ten sposób obserwacje o *dużej* (większej, niż dwukrotność wartości średniej) dźwigni.

Residuals vs Leverage

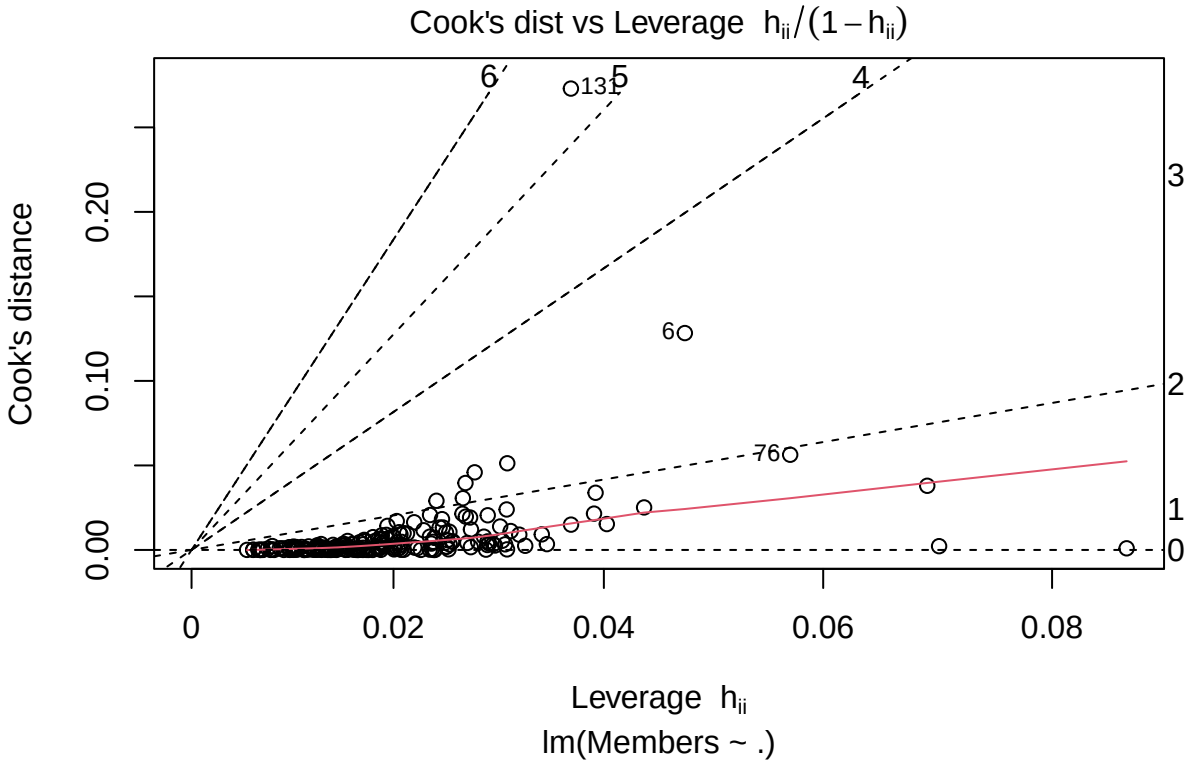
```
plot(mod, 5)
```



Na tym wykresie mamy zależność rezyduów studentyzowanych r_i od dźwigni h_{ii}. Czerwone linie przerywane, które widzimy, to linie dla których odległość Cooka jest równa 0.5 (niewidoczna w tym wypadku, bo znajdująca się powyżej górnej linii przerywanej oraz poniżej dolnej linii przerywanej, jest linia przerywana dla odległości Cooka równej 1). W tym wypadku interesują nas głównie punkty, które znajdują się powyżej górnej linii przerywanej oraz poniżej dolnej linii przerywanej – w tym wypadku nie mam żadnego. Tym razem to, jak punkty są ułożone, czy tworzą chmurę, czy jest jakaś zależność, nie jest ważne. W dalszym ciągu możemy jednak wyczytać z tego wykresu obserwacje o dużych rezyduach studentyzowanych, a także dużej dźwigni.

Cook's dist vs Leverage

```
plot(mod, 6)
```



To dodatkowy wykres, ale warto o nim wspomnieć. Przypomnijmy, że odległość Cooka może być wyrażona jako

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

Nieprzypadkowo, na osi poziomej mamy właśnie $\frac{h_{ii}}{1-h_{ii}}$. Wykres ten pozwala nam zatem ocenić, czy odległość Cooka jest duża dlatego, że mamy duże rezyduum studentyzowane r_i, czy dlatego, że mamy dużą dźwignię (czy jest może i tak, i tak). Linie przerywane to wartości rezyduów studentyzowanych (oetykietowane, w tym wypadku, jako $0, 1, ..., 6$).