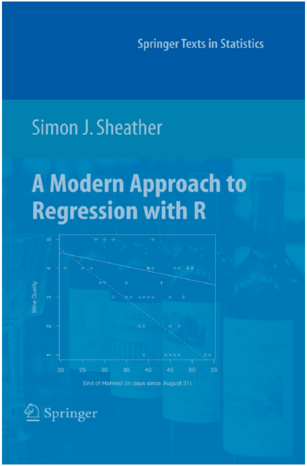
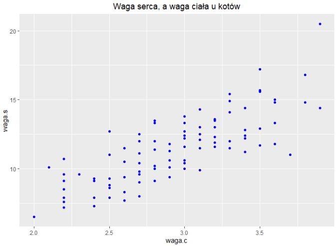

Springer Texts in Statistics
Simon J. Sheather
A Modern Approach to Regression with R
Wine Quality
End of Harvest (in days since August 31)
Springer

Regresja

Na wykresie mamy wagę 97 kotów [kg] i wagę ich serc [g] (dane catsM w R). Zbadamy zależność między wagą serca, a wagą ciała. Oznaczmy przez Y wagę serca, przez x wagę ciała. Wykorzystamy model probabilistyczny

$$Y = f(x) + \varepsilon,$$

gdzie f jest funkcją z ustalonej rodziny funkcji, ε zmienną losową (błędem) taką, że $E(\varepsilon) = 0$. Co oznacza, że $E(Y) = f(x)$.



Powyższy model nazywamy modelem regresji. Jeśli funkcja f ma postać

$$f(x) = b_0 + b_1 x,$$

dla pewnych $b_0, b_1 \in \mathbb{R}$, to model taki nazywamy liniowym – mówimy też o regresji liniowej.

W klasycznym modelu regresji zakładamy, że błąd ε jest zmienną losową o rozkładzie normalnym $N(0, \sigma^2)$. Zauważmy, że wtedy $Y \sim N(f(x), \sigma^2)$.

Estymator najmniejszych kwadratów

Niech $x_1, x_2, \ldots, x_n$ bedą ustalonymi nielosowymi wielkościami i niech $Y_1, Y_2, \ldots, Y_n$ będą odpowiadającymi im wartościami obarczonymi losowym błędem ε_i. Na podstawie próby

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n),$$

chcemy wyznaczyć funkcję f taką, że

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Przy czym zakładamy, że $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ są zmiennymi losowymi niezależnymi o takim samym rozkładzie, o wartości oczekiwanej równej zero i skończonej wariancji.

Jako kryterium dopasowania funkcji f do danych, można przyjąć sumę kwadratów błędów

$$J(f) = \sum_{i=1}^{n} (Y_i - f(x_i))^2. \tag{1}$$

Funkcję, która w danej klasie funkcji, minimalizuje $J(f)$, nazywamy estymatorem najmniejszych kwadratów nieznanej funkcji regresji f.

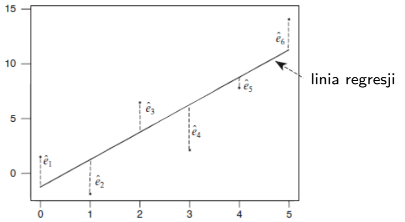
Regresja liniowa. Estymacja parametrów

Dalej zajmiemy się regresją liniową. Zakładamy, że funkcja f ma postać $f(x) = b_0 + b_1 x$, dla pewnych $b_0, b_1 \in \mathbb{R}$.

Załóżmy, że przy danych $x_1, x_2, \ldots, x_n$ obserwowane są wartości $y_1, y_2, \ldots, y_n$. Różnice

$$\epsilon_i = y_i - \underbrace{f(x_i)}_{\hat{y}} = y_i - b_1 x_i - b_0$$

nazywamy resztami i uznajemy za realizacje błędów ε_i, $i = 1, 2, \ldots, n$.



Aby wyznaczyć współczynniki b_0, b_1, minimalizujemy sumę kwadratów reszt

$$\mathrm{RSS}(b_0, b_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \underbrace{(y_i - b_1 x_i - b_0)}_{\epsilon_i}^2. \tag{2}$$

(RSS - Residual Sum of Squares)

Uwaga. $\hat{e}_i = \epsilon_i$

Estymatory parametrów

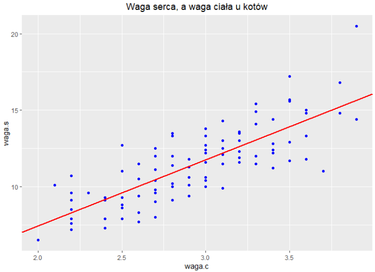
Oznaczmy estymatory najmniejszych kwadratów współczynników b_0 i b_1, otrzymane w wyniku optymalizacji RSS (2), przez β_0 i β_1. Wtedy

$$\boxed{\begin{aligned} \beta_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} \\ \beta_0 &= \overline{Y}_n - \overline{x}_n \beta_1, \qquad (3) \\ &\text{gdzie } \overline{x}_n = \tfrac{1}{n}\sum x_i,\ \overline{Y}_n = \tfrac{1}{n}\sum Y_i. \end{aligned}}$$

Przykład – cd.

Średnie wagi ciała i serc kotów wynoszą odpowiednio

$$\overline{x} = 2{,}9\ [kg], \qquad \overline{y} = 11{,}3\ [g]$$



Podstawiając do (3), otrzymujemy

$$\beta_1 = 4{,}31, \quad \beta_0 = -1{,}18.$$

Skąd linia regresji to

$y = -1{,}18 + 4{,}31x.$

Uwaga. Zobacz Przykład 1.a w skrypcie 10RegresjaLiniowa.Rmd

Regresja w R

Gotowym rozwiązaniem umożliwiającym wykonanie klasycznej regresji w R, jest funkcja `lm()`. Poniżej wyniki z wykorzystaniem tej funkcji, dla danych z Przykładu.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.1841     0.9983  -1.186    0.239
waga.c        4.3127     0.3399  12.688   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.557 on 95 degrees of freedom
Multiple R-squared:  0.6289,    Adjusted R-squared:  0.625
F-statistic:   161 on 1 and 95 DF,  p-value: < 2.2e-16
```

Dalsza część wykładu zawiera teorię niezbędną do zrozumienia powyższych wyników.

1. (Std.Error) Błąd standardowy estymatorów β_0, β_1.
2. (t value, $Pr(> |t|)$) Wynik testu na istotność współczynników b_0, b_1 – wartość statystyki testowej oraz p-value.
3. (F-statistic) Wynik dodatkowego testu na istotność współczynnika b_1.
4. (Residual standard error) Błąd standardowy reszt modelu.
5. (Multiple R-squared) Współczynnik determinacji R.

Uwaga. Zobacz Przykład 1.b w skrypcie 10RegresjaLiniowa.Rmd

Założenia modelu. Warunki Gaussa-Markowa

W klasycznym modelu regresji zakłada się, że błędy $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ są zmiennymi losowymi niezależnymi o rozkładzie $N(0, \sigma^2)$.

1. Przy tych założeniach, estymatory β_0 i β_1 (3) współczynników b_0, b_1, są estymatorami nieobciążonymi oraz mają rozkład normalny

$$\beta_0 \sim N\left(b_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}_n^2}{ns_x^2}\right)\right),$$

$$\beta_1 \sim N\left(b_1, \frac{\sigma^2}{ns_x^2}\right),$$

gdzie $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$.

Istotność współczynników regresji

2. Testujemy istotność współczynników regresji b_0, b_1, czyli hipotezy $(i = 0, 1)$

$$\mathrm{H}_0 : b_i = 0 \quad \text{przeciwko} \quad H_1 : b_i \neq 0.$$

Jako statystyki testowej użyjemy

$$T = \frac{b_i - \beta_i}{se(\beta_i)}, \quad i = 0, 1 \tag{4}$$

która, przy prawdziwości hipotezy H_0 ma rozkład t-Studenta o $n-2$ stopniach swobody.

Wielkości $se(\beta_i)$ $i = 0, 1$, to błędy standardowe estymatorów

$$se(\beta_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}_n^2}{ns_x^2}}, \quad se(\beta_1) = \frac{\hat{\sigma}}{\sqrt{ns_x^2}}$$

Hipotezę zerową H_0 odrzucamy na poziomie istotności α, gdy wartość $p-value = P(T > |t|)$ (gdzie t jest wartością statystyki testowej obliczoną dla danych) jest mniejsza lub równa α.

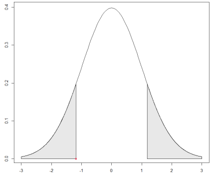
Przyklad cd.

- Testujemy hipotezę zerową: $H_0: \ b_0 = 0$, przeciwko hipotezie alternatywnej: $H_1: \ b_0 \neq 0$.

Jeśli hipoteza zerowa jest prawdziwa, to statystyka (4) ma rozkład t-Studenta o $n-2 = 95$ stopniach swobody. Zarówno małe jak i duże wartości statystyki świadczą przeciwko hipotezie zerowej.

Wartość statystyki testowej (4) obliczona dla danych wynosi:

$$t_0 = \frac{b_0 - \beta_0}{se(\beta_0)} = \frac{-1{,}1841}{0{,}9983} \approx -1{,}186.$$



Stąd $P(|T| > t_0) = 0{,}239$ (p-value) jest większe od poziomiu istotności $\alpha = 5\%$, nie ma więc podstaw do odrzucenia hipotezy zerowej, że współczynnik b_0 jest równy zero.

Uwaga. Zobacz Przykład 1.c w skrypcie 10RegresjaLiniowa.Rmd

Przyklad cd.

- Podobnie jak w przypadku współczynnika b_0, za pomocą tej samej statystyki testowej, testujemy hipotezę zerową: $H_0: \ b_1 = 0$, przeciwko kontrhipotezie: $H_1: \ b_1 \neq 0$.

Tym razem też, jeśli hipoteza zerowa jest prawdziwa, to statystyka (4) ma rozkład t-Studenta o $n-2 = 95$ stopniach swobody oraz zarówno małe jak i duże wartości statystyki świadczą przeciwko hipotezie zerowej.

Wartość statystyki obliczona z próby: $t_1 = 4{,}3127/0{,}3399 = 12{,}688$ jest bardzo duża (zobacz wykres na poprzednim slajdzie), $P(|T| > t_1) < 2.2e - 16$ (p-value). Stąd hipotezę zerową, że współczynnik b_1 jest równy zero odrzucamy (praktycznie na dowolnym poziomie istotności).

Wyniki testu istotności współczynników wskazują na to, że być może powinniśmy rozważyć prostszy model w którym współczynnik b_0 miałby wartość zero: $y = b_1 x$ (regresję należy wykonać wtedy ponownie z wartością współczynnika b_0 ustawioną z góry na zero).

Istotność współczynnika b_1

Funkcja lm() ma zaimplementowany jeszcze jeden test na istotność współczynnika b_1 w rozważanym modelu.

3. Testujemy hipotezę

$$H_0: \; b_1 = 0 \qquad \text{przeciwko hipotezie} \quad H_1: \; b_1 \neq 0.$$

Statystyką testową jest tym razem

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{RSS/(n-2)},$$

która przy prawdziwości hipotezy zerowej ma rozkład $F_{1,n-2}$ (rozkład Snedecora o $\nu_1 = 1$ i $\nu = n{-}2$ stopniach swobody).

Tym razem duże wartości statystyki świadczą przeciwko hipotezie zerowej. Hipotezę zerową odrzucamy na poziomie istotności α, gdy $p{-}value = P(F > F_0)$ (gdzie F_0 jest wartością statystyki obliczoną dla danych), jest mniejsze lub równe α.

Przyklad cd.

W modelu regresji dla kotów mamy: $F = 161$, $p - \text{value} < 2.2e - 16$, zatem ponownie odrzucamy hipotezę o nieistotności ($b_1 = 0$) współczynnika b_1.

Reszty modelu

Po wyestymowaniu współczynników b_0 i b_1 oraz przeprowadzeniu testów na ich istotność, przechodzimy do analizy reszt modelu. Wykonamy test normalności oraz wyestymujemy wariancję σ^2 błędu ε.

Reszty modelu – błąd standardowy reszt (RSE)

4. Analizujemy wariancję $\mathrm{Var}(\varepsilon) = \sigma^2$ reszt modelu.

Można wykazać, że przy przyjętych założeniach, wartość oczekiwana sumy kwadratów reszt wynosi

$$E\left(\sum_{i=1}^{n} \epsilon_i^2\right) = \frac{n-2}{n} E\left(\sum_{i=1}^{n} \varepsilon_i^2\right) = (n-2)\sigma^2.$$

Stąd nieobciążonym estymatorem wariancji błędów jest

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \epsilon^2.$$

Błąd standardowy reszt (*Residual Standard Error* – RSE), to

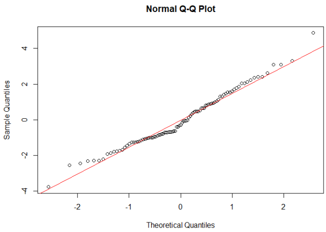
$$RSE = \hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \epsilon^2}.$$

Przyklad cd.

Testujemy hipotezę o normalności rozkładu błędu. W teście Szapiro-Wilka, *p*-value wynosi

$$p = 0{,}1381.$$

Zatem na poziomie istotności 5% nie ma podstaw do odrzucenia hipotezy o normalności rozkładu reszt



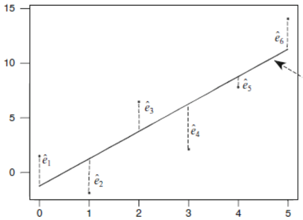
Błąd standardowy reszt: $RSE = \sqrt{230.26/95} \approx 1{,}6$[g] (średnia waga serca: 11.3[g]).

Uwaga. Zobacz Przykład 1.d w skrypcie 10RegresjaLiniowa.Rmd

Współczynnik determinacji – R)

5. Analizujemy też, na ile otrzymany model wyjaśnia zmienność w danych, rozumianą jako

$$\sum_{i=1}^{n}(y_i - \overline{y})^2. \tag{5}$$



W tym celu, wykorzystuje się współczynnik determinacji

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \overset{(6)}{=} 1 - \frac{\sum_{i=1}^{n}\epsilon^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Ostatnia równość wynika z następującej, łatwej do uzasadnienia zależności, że rzeczywista zmienność jest sumą zmienności z modelu oraz kwadratów błędów

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}\epsilon^2 \tag{6}$$

Współczynnik determinacji przyjmuje warości z przedziału $[0, 1]$, im wartości są bliższe 1, tym wyjaśnienie zmienności jest lepsze.

Przykład cd.

W przypadku modelu regresji dla kotów ($y = -1{,}18 + 4{,}31x$), współczynnik determinacji $R \approx 63\%$. Zatem 37% zmienności rozumianej jako (5), nie jest wyjaśniona przez wagę kota. Można więc rozważyć dołączenie do naszego modelu jeszcze innych zmiennych oprócz wagi.

Predykcja z modelu

Predykcja Y^*, gdy $x = x^*$ to wartość

$$\hat{y}^* = \beta_0 + \beta_1 x^*.$$

Uwaga. Zobacz Przykład 1.e,f,g w skrypcie 10RegresjaLiniowa.Rmd