

5.1 Regresja liniowa i logistyczna - wprowadzenie

Cytując definicje słowa *regresja* ze słownika PWN

- «powrót do wcześniejszego, gorszego stanu w rozwoju czegoś»
- «zmiany zachodzące w rozwoju jakiegoś organizmu, polegające na redukcji narządów i ograniczeniu jego funkcji w zmieniających się warunkach życia»
- «cofanie się morza z zalanych łądów»
- «zmniejszanie się obszarów pustynnych lub lodowcowych»
- «jeden z mechanizmów obronnych osobowości pojawiający się w sytuacji silnego napięcia emocjonalnego, polegający na powrocie do bardziej prymitywnych form reagowania»

Jak łatwo się domyślić nie będziemy korzystali z żadnego z powyższych znaczeń tego słowa.

Pierwszy terminu regresja w kontekście analizy danych użył Karl Pearson w roku 1908. Dziś (100 lat później) jest to jedno z najbardziej popularnych narzędzi statystycznych. Regresja jest popularna, ponieważ pozwala na opisanie związku pomiędzy zmiennymi objaśniającymi a zmienną objaśnianą.

Zagadnienie regresji to jest bardzo obszerne i z pewnością mogło by być tematem nie jednego semestralnego kursu. My, mając dwie godziny laboratorium zajmiemy się podstawowymi informacjami niezbędnymi by znacząc stosować regresje liniową i logistyczną.

Regresja wykorzystywana jest w wielu różnych dziedzinach począwszy od genomiki przez wszystkie kierunki inżynierskie po kierunki humanistyczne (np. w psychologii), do tego dochodzą zastosowania w medycynie o finansach już nie wspominając. Innymi słowy regresje stosuje się wszędzie. W pewnych dziedzinach rozważa się specyficzne modele wraz z niestandardowymi założeniami, my poniżej powiemy o najprostszych możliwych modelach, a więc prostym modelu regresji liniowej oraz logistycznej (osoby chcące dowiedzieć się więcej powinny sięgnąć do literatury wymienionej na stronie laboratorium).

5.2 Regresja liniowa

Przypuśćmy że obserwujemy n obiektów. Każdy z obiektów możemy scharakteryzować pewnymi cechami. Interesuje nas opisanie zależności wybranej cechy (będziemy ją nazywać zmienną objaśnianą) przez zbiór innych cech (zmiennych objaśniających).

Przyjmijmy, że pomiędzy obserwowanymi zmiennymi istnieje zależność

$$y_i = \beta_0 + \beta_1 * x_{1,i} + \beta_2 * x_{2,i} + \dots + \beta_m * x_{m,i} + \varepsilon_i$$

gdzie y_i to wartość cechy objaśnianej dla obiektu i , $x_{j,i}$ to wartość j tej zmiennej objaśniającej u osobnika i tego, ε_i to niezależne wartości pewnej zmiennej losowej o średniej 0 (w pewnych przypadkach zakłada się dodatkowo, że zmienna ta ma rozkład gaussowski), a β_i to pewne nieznanne wartości.

Interesującym nas zagadnieniem będzie ocena parametrów β_i , a więc siły zależności pomiędzy zmienną x_j a zmienną y . Oceny parametrów β_i wyznacza się najczęściej stosując metodę najmniejszych kwadratów lub największej wiarygodności (jeżeli zmienna ε ma rozkład normalny obie metody prowadzą do tych samych ocen).

Rozważmy jakiś przykład, aby był namacalny niech dotyczy pieniędzy. Interesującą nas zmienną niech będzie cena mieszkania. Skądinąd wiemy, że zależy ona od różnych czynników. Przypuśćmy że mamy dane dotyczące różnych mieszkań, w tym informacje o liczbie pokoi, powierzchni, odległości od ratusza (w bardziej realistycznym scenariuszu informacje o dzielnicy) oraz typu budynku w którym znajduje się to mieszkanie (wielka płyta z lat 70, czteropiętrowy blok, kamienica?). Chcemy ocenić jaki wpływ na cenę mieszkania mają poszczególne czynniki.

Oczywiście zrobimy to w R!

Zacznijmy od importu danych

```
> mieszkania = read.table("http://semestr.pl/cogito/stats/daneMieszkania.csv", header=T, sep=";")
```

Funkcja służąca do budowy modelu to **lm** (od linear model). Zbudujmy model liniowy dla naszych danych

```
> modelPelny <- lm(cena~., mieszkania)
```

Pierwszym argumentem jest formuła opisująca model. W powyższym przykładzie, formuła ta oznacza że w modelu chcemy wykorzystać wszystkie zmienne, poza objaśnianą. Oczywiście możemy też wybrać tylko podzbiór zmiennych. Poniższa formuła buduje model tylko dla dwóch zmiennych objaśniających

```
> modelPokoiTyp <- lm(cena~pokoi+typ.budynku, mieszkania)
```

Zobaczmy jak ten model wygląda „w środku”

```
> summary(modelPelny)

Call:
lm(formula = cena ~ ., data = mieszkania)

Residuals:
    Min       1Q   Median       3Q      Max
-31423.1  -6035.1   374.9   5697.5  28958.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  103715.2    2412.6   42.989 < 2e-16 ***
pokoi         389.1      1915.4    0.203   0.84
powierzchnia  1978.4      103.9   19.050 < 2e-16 ***
dzielnicaKrzyki -21466.8    1641.0  -13.081 < 2e-16 ***
dzielnicaSrodmiescie -14298.3    1792.5   -7.977 1.30e-13 ***
typ.budynkukamienica -11454.7    1764.5   -6.492 7.00e-10 ***
typ.budynkuwieszowiec -11039.3    1685.1   -6.551 5.06e-10 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9719 on 193 degrees of freedom
Multiple R-Squared: 0.9506, Adjusted R-squared: 0.9491
F-statistic: 619.5 on 6 and 193 DF, p-value: < 2.2e-16
```

Szczegółowo te wyniki zostaną omówione na zajęciach.

Ważne jest żeby wiedzieć jak interpretować kolumnę *Estimate* (czemu to odpowiada w modelu?) oraz kolumnę $Pr(> |t|)$ (czemu ona może odpowiadać?).

Czy na podstawie tego modelu można przewidzieć ile średnio kosztuje mieszkanie 3 pokojowe? Jeżeli tak to w jaki sposób?

Czy jest istotna różnica cen pomiędzy dzielnicami? Jeżeli tak to na korzyść której dzielnicy?

Spróbuj zaproponować kilka wniosków podsumowujących otrzymany wynik.

5.3 Regresja logistyczna

Zdarzyć się może, że interesująca nas zmienna (objaśniana) nie ma rozkładu ciągłego ale przyjmuje jedynie kilka wartości, np. dwie. Zmienną „zdrowy/chory” lub „dobry/zły klient” nie można sensownie analizować wykorzystując regresję liniową. W takich przypadkach wykorzystuje się regresję logistyczną.

Ponieważ procedury optymalizacyjne nie znoszą wartości nieciągłych, dlatego w modelu objaśniana będzie nie dwupoziomowa zmienna, a prawdopodobieństwo przyjęcia przez tą zmienną wyróżnionej wartości. Innymi słowy model regresji logistycznej jest postaci

$$Pr(y_i = 1 | x_{i,1}, \dots, x_{i,m}) = \exp(\mu_i) / (1 + \exp(\mu_i)),$$

gdzie

$$\mu_i = \beta_0 + \beta_1 * x_{1,i} + \beta_2 * x_{2,i} + \dots + \beta_m * x_{m,i}.$$

Zapis $Pr(y_i = 1)$ jest symboliczny i oznacza prawdopodobieństwo przyjęcia przez zmienną y wybranej z dwóch wartości.

Mówiąc brzydko „cała reszta jest taka jak dla regresji liniowej”. A więc interesuje nas ocena wartości β_j opisujących siłę związku pomiędzy dwupoziomową zmienną y a objaśniającą zmienną x_j .

Do budowy modelu regresji logistycznej wykorzystać można funkcję **glm** (od generalized linear model). Poniżej przykład konstrukcji modelu z dwoma predyktorami (zmiennymi objaśniającymi)

```
> model <- glm(F3~sex+age, manie, family="binomial")
```

Zobaczmy jak wygląda ten model po dopasowaniu

```
> summary(model)
```

Która płeć jest bardziej narażona na depresję?

Jak $Pr(y = 1)$ zależy od zmiennych x_j ?
Spróbuj narysować wykres!