Statystyczna Analiza Danych – laboratorium

Regresja liniowa i intepretacja parametrów

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 7
20/21 kwietnia 2023

Idea zajęć – co i po co będziemy robić?

- Regresja liniowa jest jedną z prostszych a szeroko stosowanych technik modelowania danych
- Służy do badania zależności liniowych
- Zaznajomimy się dziś z działaniem i składnią funkcji `lm` i intepretacją parametrów

Przypomnienie – model liniowy

$$y = X\beta + \varepsilon$$

- y – zmienna objaśniana (endogeniczna, zależna) o rozkładzie (quasi)ciągłym
- X – macierz zmiennych objaśniających (egzogenicznych, niezależnych)
- ε – składnik losowy o rozkładzie normalnym
- β – wektor parametrów

Przypomnienie – oszacowanie

$$y = X\hat{\beta} + e$$

- y – zmienna objaśniana (endogeniczna, zależna) o rozkładzie (quasi)ciągłym, n obserwacji
- X – macierz k zmiennych objaśniających (egzogenicznych, niezależnych)
- e – wektor reszt (residua, oszacowanie składnika losowego)
- $\hat{\beta}$ – wektor oszacowań parametrów, uzyskany MNK: $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta}$ – wartość dopasowana

Zadanie 1

- Wczytaj dane z pliku ais.csv
- Ile zbiór zawiera obserwacji, a ile zmiennych? Które zmienne są ilościowe, a które jakościowe?
- Jaka jest średnia oraz wariancja każdej ze zmiennych ilościowych?
- Które zmienne ilościowe są najbardziej skorelowane, które najsłabiej, a które mają najsilniejszą korelację ujemną?

```
library(GGally)
ggpairs(ais, aes(col=Sex), columns=c(9, 10, 5, 13))
```

Zadanie 2

- Wykorzystując model regresji liniowej zbadaj zależność wagi sportowców (WT) od ich wzrostu (Ht)
- Sprawdź, czy ta zależność jest statystycznie istotna (jeśli jest, to spróbuj skomentować to oszacowanie)
- Wykorzystaj funkcję `predict()`, aby uzyskać przedział ufności na poziomie 95% dla sportowca o wzroście 180cm

Zadanie 3 – regresja wieloraka

- Wykorzystując model regresji liniowej, zbadaj zależność wagi sportowców (WT) od wszystkich pozostałych zmiennych ilościowych
- Sprawdź, które oszacowania parametrów są statystycznie istotne i dokonaj ich interpretacji

lm

```
regresja <- lm(y ~ x1 + x2, dane)
# mozemy przeksztalcac zmienne wewnatrz komendy, nie trzeba ich specjalnie tworzyć
summary(regresja) # zwroci wydruk
# zeby np. wlaczyc do regresji kwadrat zmiennej korzystamy z I():
regresja <- lm(y ~ x1 + I(x1^2) + x2, dane)
```

Co można wyczytać z wydruku? #1

```
> regresja <- lm(Wt ~ RCC + Hc+ Ht + Ferr + SSF + X.Bfat + LBM, w)
> summary(regresja)

Call:
lm(formula = Wt ~ RCC + Hc + Ht + Ferr + SSF + X.Bfat + LBM,
    data = w)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9947 -0.4629 -0.0515  0.3930  3.3417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.7332763  1.5835958  -5.515 1.10e-07 ***
RCC         -0.1431933  0.2966022  -0.483   0.6298
Hc           0.0469937  0.0390209   1.204   0.2299
Ht          -0.0171768  0.0092530  -1.856   0.0649 .
Ferr         0.0002501  0.0011758   0.213   0.8318
SSF          0.0307273  0.0071609   4.291 2.81e-05 ***
X.Bfat       0.7164549  0.0398331  17.986  < 2e-16 ***
LBM          1.1355624  0.0086116 131.864  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.724 on 194 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9973
F-statistic: 1.06e+04 on 7 and 194 DF,  p-value: < 2.2e-16
```

forma funkcyjna modelu

lista zmiennych i oszacowania współczynników przy nich, intercept to stała/wyraz wolny

wartość statystyki testowej w teście istotności współczynnika (test t) i p-wartość

t value = Estimate/Std. Error
UWAGA: H0 o nieistotności

Test t

$$t = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \approx t(n-k)$$

- Istotność statystyczną sprawdzamy, wykorzystując test t
- Hipoteza zerowa o **nieistotności**, hipoteza alternatywna dwustronna
- Wiemy, że $\hat{\beta}$ ma wielowymiarowy rozkład normalny
- Prawidłowe działanie tego testu wymaga (poza spełnieniem założeń KMRL) spełnienia przez reszty założenia o normalności rozkładu
- Sprawdzamy, czy zero należy do przedziału ufności dla oszacowania parametru

Interpretacja oszacowań przy zmiennych ilościowych

- Intepretacja wartości oszacowań zależy od formy funkcyjnej modelu!
- Każdy z parametrów intepretujemy *ceteris paribus* – przy pozostałych wartościach zmiennych niezależnych niezmienionych
- Oszacowania przy stałej nie interpretuje się

Interpretacja oszacowań przy zmiennych ilościowych

- Model na poziomach:
  - Wzrost/spadek wielkości zmiennej x o jednostkę wiąże się z wzrostem/spadkiem y o $\hat{\beta}$ jednostek *ceteris paribus*
- Model na logarytmach:
  - Wzrost/spadek wielkości zmiennej x o 1% wiąże się z wzrostem/spadkiem y o $\hat{\beta}$ % *ceteris paribus*
- Uwaga na zmienne mierzone w odsetkach (zmiana o punkty procentowe!)

Interpretacja – przyczynowość

- Przystępując do modelowania zjawiska za pomocą regresji **zakładamy** kierunek przyczynowości. To x oddziaływuje na y, nie na odwrót
- Dla bezpieczeństwa lepiej unikać słów, które mogą być bezpośrednio zintepretowane jako wskazanie przyczynowości. Korzystajmy ze słów sugerujących współwystępowanie/ korelację
- Badanie przyczynowości jest możliwe tylko w bardzo wąskim znaczeniu, dla szeregów czasowych

Co można wyczytać z wydruku? #2

```
> regresja <- lm(Wt ~ RCC + Hc+ Ht + Ferr + SSF + X.Bfat + LBM, w)
> summary(regresja)

Call:
lm(formula = Wt ~ RCC + Hc + Ht + Ferr + SSF + X.Bfat + LBM,
    data = w)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9947 -0.4629 -0.0515  0.3930  3.3417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.7332763  1.5835958  -5.515 1.10e-07 ***
RCC         -0.1431933  0.2966022  -0.483   0.6298
Hc           0.0469937  0.0390209   1.204   0.2299
Ht          -0.0171768  0.0092530  -1.856   0.0649 .
Ferr         0.0002501  0.0011758   0.213   0.8318
SSF          0.0307273  0.0071609   4.291 2.81e-05 ***
X.Bfat       0.7164549  0.0398331  17.986  < 2e-16 ***
LBM          1.1355624  0.0086116 131.864  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.724 on 194 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9973
F-statistic: 1.06e+04 on 7 and 194 DF,  p-value: < 2.2e-16
```

R^2 i R^2 skorygowane

Wartość statystyki testowej w teście F (łącznej nieistotności) oraz p-wartość.

H0: współczynniki przy zmiennych są łącznie nieistotne

Współczynnik determinacji R^2

- $R^2 = \frac{ESS}{TSS}$
- Stosunek zmienności objaśnionej przez model (ESS) do zmienności całkowitej (TSS)
- W modelach zawierających stałą możliwa dekompozycja: $R^2 = 1 - \frac{RSS}{TSS}$, RSS – zmienność resztowa
- Intepretacja: część zmienności zmiennej zależnej, którą udało się wytłumaczyć zmiennością zmiennych niezależnych w modelu

Współczynnik determinacji R^2 – wady

- R^2 rośnie wraz z dodaniem zmiennych do modelu
- R^2 jest wysokie w modelach z problemem autokorelacji (np. gdy wykonujemy regresję, w której jeden z X jest szeregiem czasowym z trendem)
- R^2 może być też wysokie w modelach, w których występuje problem niedokładnej współliniowości (silna korelacja zmiennych X)
- W modelach bez stałej może przyjąć wartości spoza przedziału [0,1]
- Możliwa korekta: R^2 skorygowane: $R^2_{adj} = 1 - \frac{N-1}{N-k}(1 - R^2)$