**Regresja - zadania i przykłady.**

W5 e0

# **Zadanie 1.** Poniżej zamieszczono fragmenty wydruków dotyczących dopasowania modelu regresji do zmien- nej **ozone** w oparciu o promieniowanie (**radiation**), oraz w oparciu o promieniowanie i temperaturę (**tempe- rature**). Zbiór zawiera 111 obserwacji.

1. Podaj przybliżoną liczbę wartości resztowych w pierwszym modelu większych od *−*0*,*5895.
2. Podaj procent zmienności dodatkowo wyjaśniony przez wprowadzenie zmiennej **temperature** do modelu ozone *∼* radiation.
3. Na podstawie wyniku przeprowadzonego testu stwierdź, czy wprowadzenie zmiennej **temperature** jest wskazane. Uzasadnij.
4. Oblicz brakującą wartość na wydruku (miejsce zaznaczone kropkami “.......”) i wytłumacz, jak otrzymano odpowiadającą p-wartość 0,0007.

W5 e1

---------- Model 1. Call: lm(formula = ozone ~ radiation, data = ozonedata) Residuals:

Min 1Q Median 3Q Max

-1.5811 -0.5895 -0.1162 0.5986 2.0508

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.4859713 0.1746316 14.24 < 2e-16

radiation 0.0041223 0.0008482 4.86 3.96e-06

Residual standard error: 0.8109 on 109 degrees of freedom Multiple R-Squared: 0.1781,

F-statistic: 23.62 on 1 and 109 DF, p-value: 3.964e-06

---------- Model 2. Call: lm(formula = ozone ~ temperature + radiation) Residuals:

Min 1Q Median 3Q Max

-1.183 -0.4025 -0.03355 0.2965 1.95

|  |  |
| --- | --- |
| Coefficients: |  |
|  | Value | Std. Error | t value | Pr(>|t|) |
| (Intercept) | -2.1530 | 0.4398 | -4.8951 | 0.0000 |
| temperature | 0.0643 | 0.0059 | 10.9681 | 0.0000 |
| radiation | 0.0021 | ...... | 3.4968 | 0.0007 |

Residual standard error: 0.5603 on 108 degrees of freedom Multiple R-Squared: 0.6112

F-statistic: 84.88 on 2 and 108 degrees of freedom, the p-value is 0

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------

a) -0,5895 to wartość dla 1 kwartyla. Stąd ¾ wartości resztowych będzie większych (ze 111 obserwacji) -> 3/4 \* 111 ~= 83

Odp. Przybliżona liczba wartości resztowych w pierwszym modelu większych od -0,5895 to 83.
b) W pierwszym modelu R2= 0,1781 (współczynnik determinacji – na tyle model wyjaśnia zależność).
W drugim przypadku R2=0,6112, tj. model wyjaśnia zależność w ok. 61%. Tak więc dodatnie zmiennej temperature zwiększyło wyjasnianie do 61%. Stąd zmiana procentu wyjaśniania to 0,6112 - 0,1781 = 0,4331 (czyli 43,31%)
Odp. Procent zmienności dodatkowo wyjaśniony przez wprowadzenie zmiennej temperature do modelu ozone~radiation to 43,31%.
c) Własności zmiennej temperature:

- zmienna jest istotna (p-wartość <0,01)
- po jej wprowadzeniu wzrasta współczynnik determinacji R2
- zmalał błąd standardowy wartości resztowych

Odp. Wprowadzenie zmiennej temperature jest wskazane na podstawie wyżej przytoczonych wniosków.
d) Brakująca wartość na wydruku: Std. Error (błąd standardowy wartości resztowych) = value(wartość parametru)/ t value(wartość statystyki testowej), czyli 0,0021/3,4968 ~= 6,0055e-4 = 0,0006055

Otrzymanie odpowiadającej p-wartości:

0,0007 jest to empiryczny poziom istotności dla testu i jest najniższym poziomem istotności dla którego należy odrzucić H0 (ponieważ H0 jest takie, że zmienna temperature jest nieistotna). Szukam w tablicy rozkładu t-Studenta wartości statystyki testowej (3.4968) dla danej liczby stopni swobody (111-2=109) i odczytuję obszar krytyczny (poziom istotności) α, który odpowiada tej wartości statystyki, czyli naszą p-wartość 0,0007.

Odp. Brakująca wartość na wydruku to 0,0006055. P-wartość otrzymano na podstawie tablicy rozkładu t-Studenta w opisany sposób.

W5 e2

**Zadanie 2.** Zbiór **cheese** zawiera dane dotyczące smaku sera (zmienna **Taste**, miara subiektywna) oraz zmiennych

# **Acetic** – logarytm zawartości kwasu octowego;

**H2S** – logarytm zawartości siarkowodoru;

**Lactic** – zawartość kwasu mlekowego.

Rozpatrzono dwa modele regresji dla zmiennej objaśnianej Taste. W pierwszym zmienną objaśniającą jest jedynie zmienna Acetic, w drugim dodatkowo zmienne H2S i Lactic. Na podstawie załączonego wydruku odpowiedz na następujące pytania:

1. Wnioski dla zmiennej Acetic są inne w pierwszym i drugim modelu. Sprecyzuj na czym polega różnica i wytłumacz czym jest spowodowana.
2. Oblicz brakującą wartość dla zmiennej H2S w drugim modelu.
3. O ile wzrósł procent wyjaśnionej zmienności zmiennej Taste po dodaniu do pierwszego modelu zmiennych Lactic i H2S?

W5 e3

---------- Model 1: lm(formula = Taste ~ Acetic, data = cheese) Residuals:

Min 1Q Median 3Q Max

-29.642 -7.443 2.082 6.597 26.581

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) -61.499 24.846 -2.475 0.01964

Acetic 15.648 4.496 3.481 0.00166

Residual standard error: 13.82 on 28 degrees of freedom Multiple R-Squared: 0.302, Adjusted R-squared: 0.2771

F-statistic: 12.11 on 1 and 28 DF, p-value: 0.001658

---------- Model 2: lm(formula = Taste ~ Acetic + H2S + Lactic, data = cheese) Residuals:

Min 1Q Median 3Q Max

-17.391 -6.612 -1.009 4.908 25.449

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | -28.8768 | 19.7354 | -1.463 | 0.15540 |
| Acetic | 0.3277 | 4.4598 | 0.073 | 0.94198 |
| H2S | ........ | 1.2484 | 3.133 | 0.00425 |
| Lactic | 19.6705 | 8.6291 | 2.280 | 0.03108 |

Residual standard error: 10.13 on 26 degrees of freedom Multiple R-Squared: 0.6518, Adjusted R-squared: 0.6116

F-statistic: 16.22 on 3 and 26 DF, p-value: 3.81e-06 W5e4

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------

a) Odp. Pomimo ze kierunek oddziaływania (+ albo -) jest ten sam, wartość zmniejszyła się, ponieważ lepiej zmienność Taste wyjaśnia Lactic i H2S, przez co zmienna Acetic okazała się nieistotna (wprowadzenie jej do tego modelu nie poprawia istotnie modelu w przypadku obecności zmiennych Lactic i H2S).

b) Brakującą wartość obliczamy w sposób analogiczny do poprzedniego zadania, tj. Estimate (wartość parametru) = t value(wartość statystyki testowej t) \* Std. Error(błąd standardowy wartości resztowych) czyli 3,133\*1,2484 ~= 3,9112

Odp. Brakująca wartość Estimate (tj. wartość parametru) wynosi 3,9112.

c) Porównuję wskaźniki determinacji R2 i Adjusted R2. Wskaźnik Adjusted R2 bierze dodatkowo pod uwagę liczbę zmiennych uwzględnionych w modelu. Obliczam różnice odpowiednio R2 i Adjusted R2.

Dla R2: 0,6518-0,302 = 0,3498 (34,98% różnicy)

Dla Adjusted R2: 0,6116-0,2771 = 0,3345 (33,45% różnicy)

Odp. Procent wyjaśnionej zmienności zmiennej Taste po dodaniu do pierwszego modelu zmiennych Lactic i H2S wzrósł odpowiednio o 34,98% i 33,45% dla wskaźnika uwzględniającego ilość zmiennych w modelu.

**Zadanie 3.** Poniżej zamieszczona jest część wydruku dotycząca dopasowania modelu regresji do danych doty- czących liczby gatunków żółwi (zmienna zależna **Species**) na 30 wyspach archipelagu Galapagos. Rozpatrzono następujące zmienne niezależne:

**Area** - powierzchnia wyspy (km2),

**Elevation**- wysokość najwyższego punktu (m), **Nearest** - odległość do najbliższej wyspy (km), **Scruz** - odległość do wyspy Santa Cruz,

**Adjacent** - powierzchnia najbliższej sąsiedniej wyspy.

W5 e5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Baltra | Species58 | Endemics Area Elevation 23 25.09 346 | Nearest0.6 | Scruz 0.6 | Adjacent1.84 |
| Bartolome | 31 | 21 1.24 109 | 0.6 | 26.3 | 572.33 |
| Caldwell | 3 | 3 0.21 114 | 2.8 | 58.7 | 0.78 |
| Champion | 25 | 9 0.10 46 | 1.9 | 47.4 | 0.18 |
| Coamano | 2 | 1 0.05 77 | 1.9 | 1.9 | 903.82 |
| Daphne.Major | 18 | 11 0.34 119 | 8.0 | 8.0 | 1.84 |
| Daphne.Minor | 24 | 0 0.08 93 | 6.0 | 12.0 | 0.34 |
| Darwin | 10 | 7 2. | 33 168 | 34.1 | 290.2 | 2.85 |
| Eden | 8 | 4 0. | 03 71 | 0.4 | 0.4 | 17.95 |
| Enderby | 2 | 2 0.18 112 | 2.6 | 50.2 | 0.10 |
| Espanola | 97 | 26 58.27 198 | 1.1 | 88.3 | 0.57 |
| Fernandina | 93 | 35 634.49 1494 | 4.3 | 95.3 | 4669.32 |
| Gardner1 | 58 | 17 0.57 49 | 1.1 | 93.1 | 58.27 |
| Gardner2 | 5 | 4 0.78 227 | 4.6 | 62.2 | 0.21 |
| Genovesa | 40 | 19 17.35 76 | 47.4 | 92.2 | 129.49 |
| Isabela | 347 | 89 4669.32 1707 | 0.7 | 28.1 | 634.49 |
| Marchena | 51 | 23 129.49 343 | 29.1 | 85.9 | 59.56 |
| Onslow | 2 | 2 0. | 01 25 | 3.3 | 45.9 | 0.10 |
| Pinta | 104 | 37 59.56 777 | 29.1 | 119.6 | 129.49 |
| Pinzon | 108 | 33 17.95 458 | 10.7 | 10.7 | 0.03 |
| Las.Plazas | 12 | 9 0.23 94 | 0.5 | 0.6 | 25.09 |
| Rabida | 70 | 30 4.89 367 | 4.4 | 24.4 | 572.33 |
| SanCristobal | 280 | 65 551.62 716 | 45.2 | 66.6 | 0.57 |
| SanSalvador | 237 | 81 572.33 906 | 0.2 | 19.8 | 4.89 |
| SantaCruz | 444 | 95 903.82 864 | 0.6 | 0.0 | 0.52 |
| SantaFe | 62 | 28 24.08 259 | 16.5 | 16.5 | 0.52 |
| SantaMaria | 285 | 73 170.92 640 | 2.6 | 49.2 | 0.10 |
| Seymour | 44 | 16 1.84 147 | 0.6 | 9.6 | 25.09 |
| Tortuga | 16 | 8 1.24 186 | 6.8 | 50.9 | 17.95 |
| Wolf | 21 | 12 2. | 85 253 | 34.1 | 254.7 | 2.33 |

W5 e6

> summary(lm(Species~Area+Elevation+Nearest+Scruz+Adjacent)) Call:

lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent) Residuals:

|  |  |  |
| --- | --- | --- |
| Min 1Q | Median | 3Q Max |
| -111.679 -34.898 | -7.862 | 33.460 182.584 |

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 7.068221 | 19.154198 | 0.369 | 0.715351 |
| Area | -0.023938 | 0.022422 | -1.068 | 0.296318 |
| Elevation | 0.319465 | 0.053663 | 5.953 | 3.82e-06 |
| Nearest | 0.009144 | 1.054136 | 0.009 | 0.993151 |
| Scruz | -0.240524 | 0.215402 | -1.117 | 0.275208 |
| Adjacent | -0.074805 | 0.017700 | -4.226 | 0.000297 |

Residual standard error: 60.98 on 24 degrees of freedom Multiple R-Squared: 0.7658, Adjusted R-squared: 0.7171

F-statistic: 15.7 on 5 and 24 DF, p-value: 6.838e-007

# (1p.) Podaj procent zmienności liczby gatunków niewyjaśnionej przez zaproponowany model.

1. (2p.) Sformułuj hipotezę zerową i alternatywną, której odpowiada liczba 0.296318. Jaką decyzję podejmiesz w tym przypadku ?

W5 e7

1. (3p.) Sformułuj hipotezę zerową i alternatywną, której odpowiada liczba 0.000275 w prostszym modelu poniżej. Jaką decyzję podejmiesz w tym przypadku? Porównaj z (b) i skomentuj ewentualne różnice.

> summary(lm(Species~Area)) Call:

lm(formula = Species ~ Area)

Residuals:

Min 1Q Median 3Q Max

-99.495 -53.431 -29.045 3.423 306.137

Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) 63.78286 17.52442 3.640 0.001094 \*\*

Area 0.08196 0.01971 4.158 0.000275 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.73 on 28 degrees of freedom Multiple R-Squared: 0.3817, Adjusted R-squared: 0.3596

F-statistic: 17.29 on 1 and 28 DF, p-value: 0.0002748

a) Procent zmienności niewyjaśnionej przez model: 1- R2 (czyli 1-0,7658 = 23,42 tj. 23,42%) lub 1-Adj.R2 (czyli 1-0,7171 = 0,2829 tj. 28,29%).

Odp. Procent zmienności niewyjaśnionej przez model wynosi 23,42%.

b) H0: βA =0, H1: βA /= 0. βA – wskaźnik regresji dla zmiennej Area

0,296318 = p-wartość dla testu istotności zmiennej Area, czyli prawdopodobieństwo zawierania się wartości statystyki testowej w zbiorze krytycznym tj. |t value| > |− 1,068| jeśli spełnione jest H0, gdzie t value to wartość statystyki testowej t

Decyzja: brak podstaw do odrzucenia H0 na każdym standardowo przyjmowanym poziomie istotności (p-value tj. p-wartość>0,1>0,05>0,01). Wniosek: zmienna jest nieistotna w modelu

Odp. Hipotezy: H0: βA =0, H1: βA /= 0. Decyzja: pozostawienie H0.

c) H0: βA =0, H1: βA /= 0

Zmienna jest istotna, ponieważ posiada niski wskaźnik p-value (p-wartość) = 0,000275. P-wartość = P(|t value|>=4,158 jeśli spełnione jest H0). t value oznacza tutaj wartość statystyki testowej t.

Decyzja: Hipoteza H0 odrzucona.

Zmienna Area samodzielnie wnosi informacje do wyjaśniania zmienności "Species", ale umieszona w modelu razem z innymi traci na znaczeniu (inne zmienne wyjaśniają tę samą zmienność)

Odp. Hipotezy: H0: βA =0, H1: βA /= 0. Decyzja: odrzucenie H0 i przyjęcie H1. Zmienna Area straciła na znaczeniu w modelu c) ponieważ dodatkowe zmienne dodane w tym modelu a nieobecne w b) lepiej tłumaczą zmienność liczby gatunków żółwi Species.

W5 e8

# **Zadanie 4.** Na podstawie danych **ﬁsh** dotyczących 159 ryb złowionych w jeziorze Laengelmavesi koło Tampere starano się znaleźć zależność między ich wagą (**Weight**) a wysokością (**Height**), szerokością (**Width**) i długościami **L1**, **L2**, **L3** (patrz rys. 2). W pierwszym modelu uwzględniono wszystkie zmienne niezależne, w drugim usunięto zmienną Height. Przyjęto *α* = 0*,*05.

1. (1 p.) Które ze zmiennych w pierwszym modelu są istotne? Uzasadnij, sformułuj odpowiednie hipotezy zerowe dla zmiennych istotnych.
2. (2 p.) Czy zmienna *L*3 jest istotna w obu modelach? Dlaczego tak się dzieje?
3. (1 p.) Co oznacza liczba 0,9907 dla trzeciego modelu i jakiej zmiennej dotyczy?
4. (2 p.) Na podstawie załączonych rysunków oceń dopasowanie modelu pierwszego i trzeciego.

W5 e9

lm(formula = Weight ~ L1 + L2 + L3 + Height + Width, data = fish) Coefficients:

Estimate Std. Error t value Pr(>|t|)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (Intercept) | -504.084 | 30.370 | -16.598 | < 2e-16 |
| L1 | 52.829 | 40.694 | 1.298 | 0.19632 |
| L2 | 3.997 | 42.030 | 0.095 | 0.92438 |
| L3 | -29.292 | 17.648 | -1.660 | 0.09915 |
| Height | 30.043 | 8.883 | 3.382 | 0.00093 |
| Width | 10.638 | 21.029 | 0.506 | 0.61374 |
| --- |  |  |  |  |

Residual standard error: 120.4 on 142 degrees of freedom Multiple R-Squared: 0.8909, Adjusted R-squared: 0.8871 F-statistic: 232 on 5 and 142 DF, p-value: < 2.2e-16

lm(formula = Weight ~ L1 + L2 + L3 + Width, data = fish)

 ///// \_

Coefficients: / \ |

Estimate Std. Error t value Pr(>|t|) /\ \\_ / / Height

(Intercept) -523.502 30.892 -16.946 < 2e-16 < )

 ) \ |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| L1 | 11.544 | 40.212 | 0.287 | 0.7745 | \/\_\\ / | \ \ | \_ |
| L2 | -13.082 | 43.222 | -0.303 | 0.7626 |  |  |  |
| L3 | 22.430 | 9.123 | 2.459 | 0.0151 | |------- | L1 | -------| |
| Width | 65.719 | 13.781 | 4.769 | 4.52e-06 | |------- | L2 | ----------| |
| --- |  |  |  |  | |------- | L3 | ------------| |

Residual standard error: 124.7 on 143 degrees of freedom Multiple R-Squared: 0.8821, Adjusted R-squared: 0.8788 F-statistic: 267.6 on 4 and 143 DF, p-value: < 2.2e-16

W5 e10

> fish3.lm <- lm(Weight^0.3 ~ L1 + L2 + L3 + Height + Width, data=fish)

> print(summary(fish3.lm))

lm(formula = Weight^0.3 ~ L1 + L2 + L3 + Height + Width, data = fish) Coefficients:

Estimate Std. Error t value Pr(>|t|) (Intercept) 0.54870 0.04462 12.298 < 2e-16

L1 0.01622 0.05978 0.271 0.787

L2 0.08231 0.06174 1.333 0.185

L3 -0.01549 0.02593 -0.597 0.551

Height 0.11443 0.01305 8.768 5.06e-15

Width 0.35494 0.03089 11.489 < 2e-16

Residual standard error: 0.1769 on 142 degrees of freedom Multiple R-Squared: 0.9907, Adjusted R-squared: 0.9904 F-statistic: 3022 on 5 and 142 DF, p-value: < 2.2e-16

W5 e11

a) Odp. W pierwszym modelu istotna jest jedynie zmienna Height. Warunkiem istotności zmiennej jest p-value(p-wartość) dla niej <0,05. Hipoteza zerowa jak w poprzednim zadaniu: H0: βH =0, H1: βH /=0 gdzie βH – wskaźnik regresji zmiennej Height.

b) Odp. W pierwszym modelu zmienna L3 nie jest istotna, w drugim tak. Dzieje się tak, ponieważ wyjaśnia ona część zmienności, którą wcześniej wyjaśniała usunięta w modelu drugim zmienna Height (zmienna ta była istotna).

c) Odp. Jest to współczynnik determinacji –nie dotyczy on żadnej zmiennej, a mówi jaka część zmienności zmiennej Weight^0,3 jest wyjaśniana przez predykatory (L1, L2, L3, Height i Width).

d) Odp. Model pierwszy posiada wyższe współczynniki determinacji R2 i Adjusted R2, co wskazuje na wyjaśnianie więcej zmienności Weight^0,3 niż Weight poprzez zadane zmienne. Szacowanie wagi ryby na podstawie sumy jej wymiarów kłóci się z jakimkolwiek fizycznym sposobem wyznaczania wagi ciał na podstawie ich rozmiarów, w związku z tym uważam oba modele za nieoddające istoty problemu. Nie potrafię wyjaśnić pochodzenia współczynnika 0,3 w trzecim modelu ani powodu, dla którego zmienia on poziom wyjaśniania o blisko 10%.

wykonał

Sławomir Jabłoński,

s14736