Spis treści

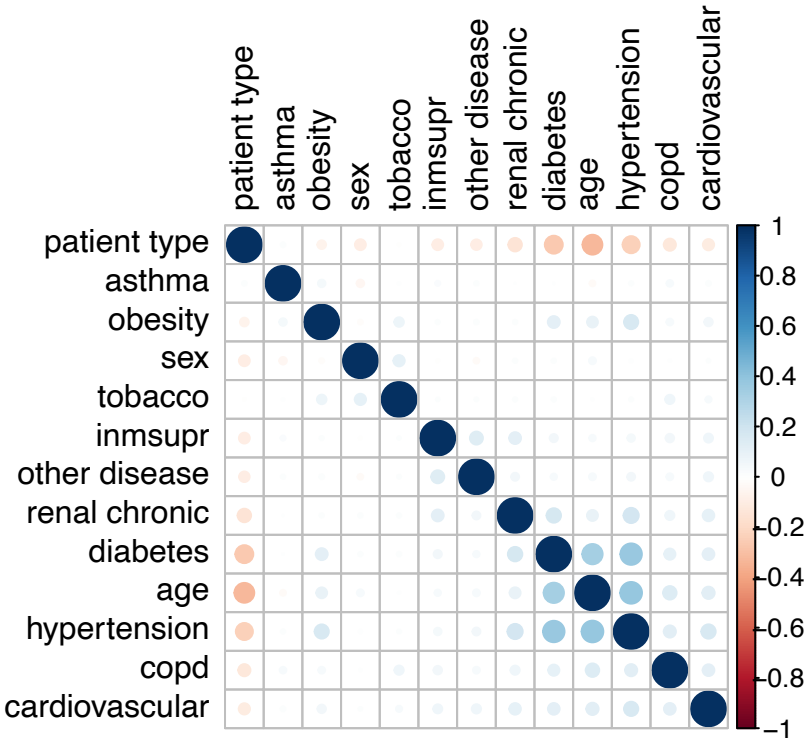
0.1 Dokumentacja modelów

0.1.1 Ocena niezależności zmiennych

Zanim przejdziemy do konstruowania modelu regresji logistycznej, sprawdźmy, czy spełnione jest istotne założenie o niezależności zmiennych objaśniających. W tym celu konstruujemy macierz korelacji. Kolor oraz jego intensywność będzie odpowiadać siły korelacji między zmiennymi. Z racji tego, że zdecydowana większość naszych zmiennych jest kategoryczna, do badania korelacji wykorzystamy rangową metodę Spearman'a.
Współczynnik r_s rangowej korelacji Spearman'a obliczany jest ze wzoru:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{1}$$

gdzie:
d_i oznacza różnicę między rangami zmiennych: $x_{ij} - x_{ik}$,
n oznacza liczbę par (obserwacji).
r_s jest liczbą z przedziału [-1,1], a skrajne jej wartości oznaczają silną korelację między zmiennymi.
Spośród rozważanych przez nas zmiennych objaśniających, nie występuje żadna silna korelacja. Możemy przyjąć, że zmienne rzeczywiście są niezależne.



Rysunek 1: Macierz korelacji Spearman'a zmiennych objaśniających

0.1.2 Wybór zmiennych do modelu

Zajmiemy się teraz doborem zmiennych do modelu regresji logistycznej. Przedstawimy różne podejścia przedstawione w części teoretycznej.

Eliminacja wsteczna

Przypomnijmy, że wykorzystując eliminacje wsteczną, na początku rozważamy model zawierający wszystkie zmienne objaśniające i usuwamy tą, dla której p-wartość testu ilorazu wiarogodności jest największa (chyba że nie przekracza poziomu istotności).
Największa p-wartość jest dla zmiennej copd, ale mieści się w poziomie istotności $\alpha = 0.15$ dlatego nie usuwamy jej z modelu.
Nie kontynuujemy dalej procedury i jako finalny model przyjmujemy ten, który zawiera wszystkie zmienne.

Tabela 1: Krok I eliminacji wstecznej - podsumowanie testu ilorazu wiarogodności

Usunięta zmienna	LogLik	Stopnie swobody	Statystyka testowa	P-wartość
sex	-83192	1	988.71	<2.2e-16
age	-83192	1	9408.9	<2.2e-16
patient type	-83192	1	46520	<2.2e-16
diabetes	-83192	1	320.26	<2.2e-16
obesity	-83192	1	348.51	<2.2e-16
hypertension	-83192	1	51.675	6.55e-13
other disease	-83192	1	51.101	8.771e-13
renal chronic	-83192	1	135.53	<2.2e-16
tobacco	-83192	1	33.509	7.095e-09
asthma	-83192	1	25.559	4.29e-07
inmsupr	-83192	1	26.492	2.646e-07
cardiovascular	-83192	1	35.379	2.714e-09
copd	-83192	1	19.549	9.804e-06

Podsumowanie

Modele stworzone przez procedury selekcji postępujące i eliminacji wstecznej są takie same. Okazuje się, że wszystkie współczynniki są też istotne na podstawie testu Wald'a i poziomie istotności $\alpha = 0.05$.

Tabela 2: Podsumowanie modelu

	Współczynnik	Oszacowanie	Błąd standardowy	Statystyka testowa	P-wartość
1	(Intercept)	3.685	0.028	133.774	<2.2e-16
2	sex	-0.418	0.013	-31.153	<2.2e-16
3	age	-0.039	0.000	-91.609	<2.2e-16
4	patient.type	3.201	0.019	170.227	<2.2e-16
5	diabetes	-0.264	0.015	-17.996	<2.2e-16
6	obesity	-0.293	0.016	-18.866	<2.2e-16
7	hypertension	-0.107	0.015	-7.203	5.90e-13
8	other.disease	-0.202	0.028	-7.221	5.17e-13
9	renal.chronic	-0.316	0.027	-11.764	<2.2e-16
10	tobacco	0.130	0.023	5.746	9.15e-09
11	asthma	0.216	0.043	4.961	7.00e-07
12	inmsupr	-0.184	0.035	-5.199	2.00e-07
13	cardiovascular	0.170	0.029	5.906	3.51e-09
14	copd	0.133	0.030	4.399	1.09e-05

Tabela 3: Realizacje przedziałów ufności dla poszególnych współczynników zmiennych na poziomie ufności 95%

Współczynnik	2.5 %	97.5 %
(Intercept)	3.63	3.74
sex	-0.44	-0.39
age	-0.04	-0.04
patient.type	3.16	3.24
diabetes	-0.29	-0.24
obesity	-0.32	-0.26
hypertension	-0.14	-0.08
other.disease	-0.26	-0.15
renal.chronic	-0.37	-0.26
tobacco	0.09	0.17
asthma	0.13	0.30
inmsupr	-0.25	-0.11
cardiovascular	0.11	0.23
copd	0.07	0.19

0.1.3 Diagnostyka modelu

0.2 Symptomy COVID-19

W tej części rozważymy dane zawierające informacje o symptomach osób chorych (albo nie) na COVID-19 i postaramy się wyróżnić symptomy typowe dla tej choroby.

0.2.1 Opis danych

Zbiór danych zawiera 5434 obserwacje, 19 zmiennych niezależnych oraz jedną zmienną zależną, informującą o tym, czy u pacjenta stwierdzono COVID-19.

1. *breathing problem* - 1 jeśli pacjent ma problemy z oddychaniem i 0 w przeciwnym wypadku,
2. *fever* - 1 jeśli pacjent ma gorączkę i 0 w przeciwnym wypadku,
3. *dry cough* - 1 jeśli pacjent ma suchy kaszel i 0 w przeciwnym wypadku,
4. *sore throat* - 1 jeśli pacjent ma ból gardła i 0 w przeciwnym wypadku,
5. *running nose* - 1 jeśli pacjent ma katar i 0 w przeciwnym wypadku,
6. *asthma* - 1 jeśli pacjent choruje na astmę i 0 w przeciwnym wypadku,
7. *chronic lung disease* - 1 jeśli pacjent ma przewlekłą chorobę płuc i 0 w przeciwnym wypadku,
8. *headache* - 1 jeśli pacjenta boli głowa i 0 w przeciwnym wypadku,
9. *heart disease* - 1 jeśli pacjent cierpi na chorobę serca i 0 w przeciwnym wypadku,

10. *diabetes* - 1 jeśli pacjent ma cukrzycę i 0 w przeciwnym wypadku,

11. *hyper tension* - 1 jeśli pacjent ma nadciśnienie i 0 w przeciwnym wypadku,

12. *fatigue* - 1 jeśli pacjent czuje się stale zmęczony i 0 w przeciwnym wypadku,

13. *gastrointestinal* - 1 jeśli pacjent ma objawy żołądkowo-jelitowe i 0 w przeciwnym wypadku,

14. *abroad travel* - 1 jeśli pacjent był ostatnio za granicą i 0 w przeciwnym wypadku,

15. *contact with COVID Patient* - 1 jeśli pacjent miał styczność z osobą chorą na COVID-19 i 0 w przeciwnym wypadku,

16. *attended large gathering* - 1 jeśli pacjent ostatnio uczestniczył w dużym zgromadzeniu i 0 w przeciwnym wypadku,

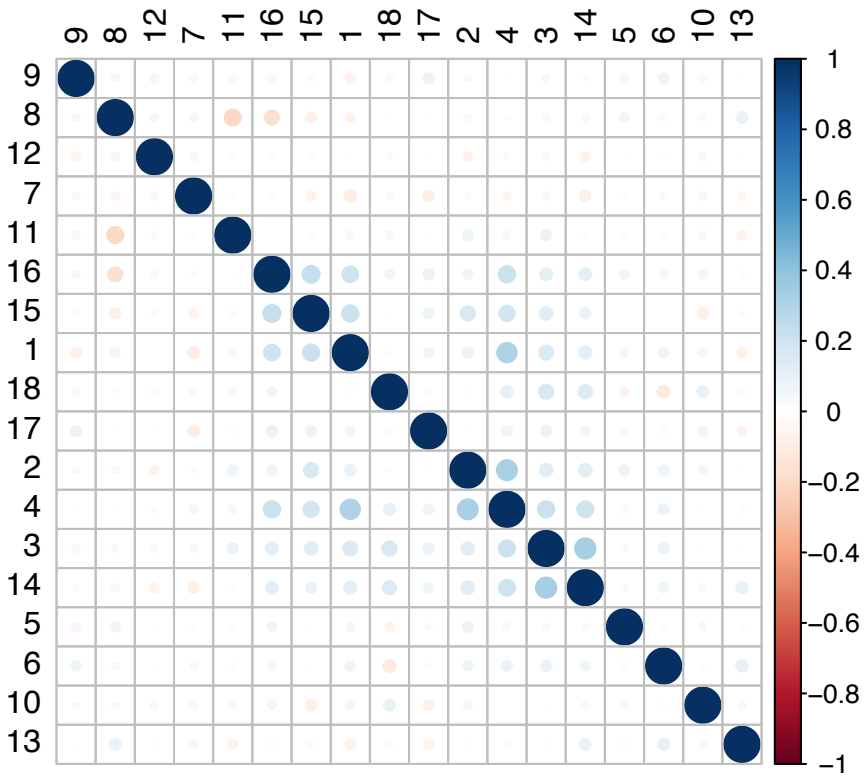
17. *visited public exposed places* - 1 jeśli pacjent odwiedzał miejsca publiczne i 0 w przeciwnym wypadku,

18. *family working in public exposed places* - 1 jeśli bliska rodzina pracuje w miejscach publicznych i 0 w przeciwnym wypadku,

19. *COVID-19* - 1 jeśli u pacjenta zdiagnozowana COVID-19 i 0 w przeciwnym wypadku.

0.2.2 Ocena niezależności zmiennych

Sprawdzimy teraz, czy spełnione jest założenie o niezależności zmiennych objaśniających. Na podstawie macierzy korelacji Spearman'a, stwierdzamy, że nie ma silnych korelacji pomiędzy zmiennymi i przyjmujemy, że zmienne objaśniające są niezależne.



Rysunek 2: Macierz korelacji Spearman'a zmiennych objaśniających

0.2.3 Wybór zmiennych do modelu

Zajmiemy się teraz doborem zmiennych do modelu. Wykorzystamy do tego kryterium AIC oraz selekcję postępującą.

Tabela 4: Podsumowanie modelu zbudowanego w oparciu o selekcję postępująca z kryterium AIC

	Zmienna	Oszacowanie	Błąd st.	Stat. test.	P-war.
1	(Intercept)	-37.10	395.74	-0.09	0.92531
2	abroad travel	22.00	395.74	0.06	0.95567
3	sore throat	4.01	0.29	13.66	<2.2e-16
4	attended large gathering	9.73	0.74	13.16	<2.2e-16
5	breathing problem	3.09	0.23	13.39	<2.2e-16
6	dry cough	4.19	0.31	13.42	<2.2e-16
7	fever	5.02	0.40	12.60	<2.2e-16
8	contact with COVID patient	1.92	0.22	8.61	<2.2e-16
9	running nose	-1.65	0.22	-7.55	4.47e-14
10	family working in public exposed places	1.40	0.23	6.11	9.71e-10
11	visited public exposed places	-0.63	0.21	-3.01	0.00261
12	diabetes	0.40	0.20	2.07	0.03875
13	heart disease	-0.37	0.20	-1.88	0.06026
14	headache	-0.38	0.20	-1.95	0.05132
15	hypertension	-0.37	0.20	-1.80	0.07187

Tabela 5: Realizacje przedziałów ufności dla poszczególnych współczynników zmiennych na poziomie ufności 95%

Zmienna	2.5 %	97.5 %
(Intercept)	-413.14	-363.50
abroad travel	271.78	221.38
sore throat	3.46	4.61
attended large gathering	8.35	11.25
breathing problem	2.65	3.55
dry cough	3.61	4.83
fever	4.27	5.84
contact with COVID patient	1.49	2.37
running nose	-2.09	-1.23
family working in public exposed places	0.96	1.86
visited public.exposed.places	-1.05	-0.22
diabetes	0.02	0.79
heart disease	-0.76	0.01
headache	-0.77	-0.00
hypertension	-0.77	0.03

Podsumowanie

Ostatecznie przyjęty przez nas model zawera zmienne *abroad travel, sore throat, attended large gathering, breathing problem, dry cough, fever, contact with COVID patient, running nose, family working in public exposed places, visited public exposed places, diabetes, heart disease, headache, hypertension*