Modele linowe raport nr 4

Dominik Mika

28 stycznia 2021

1 Zadanie 1

W tym zadaniu mamy daną regresję linową wieloraką postaci:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Mamy dane następujące estymatory: $b_0 = 1$, $b_1 = 4$, $b_2 = 3$, $s = 3$.

Stąd estymowane równanie regresji ma postać:

$$\hat{Y} = 1 + 4X_1 + 3X_2$$

1.1 a)

W pierwszym podpunkcie chcemy przewidzieć wartość Y, dla $X_1 = 2$, $X_2 = 6$.

Korzystamy z postaci naszego wyestymowanego równania regresji. Podstawiamy wartości zmiennych objaśniających do tego równania:

$$\hat{Y} = 1 + 4 \cdot 2 + 3 \cdot 6 = 27$$

1.2 b)

W następnym podpunkcie mamy dany estymator odchylenia standardowego wartości oczekiwanej Y, dla $X_1 = 2$, $X_2 = 6$, który jest równy 2. Przy pomocy tych danych chcemy obliczyć wartość estymatora wariancji predykcji $\sigma^2(pred)$.

Wiemy, że $s(\mu_h) = \sqrt{s^2 \mathbb{X}_h' (\mathbb{X}'\mathbb{X})^{-1} \mathbb{X}_h} = 2$. Znamy również $s = 3$.

Natomiast $\sigma^2(pred) = s^2(1 + \mathbb{X}_h'(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}_h)$, czyli $s^2(pred) = s^2 + s^2(\mu_h) = 4 + 9 = 13$.

1.3 c)

W tym podpunkcie chcemy skonstruować 95% przedział ufności dla β_1, wiedząc, że $s(b_1) = 1$, a $n = 20$. Wiemy, że przedział jest postaci:

$$b_1 \pm t_c s(b_1),$$

gdzie $b_1 = 4$, $s(b_1) = 1$, a t_c to kwantyl z rozkładu studenta na poziomie $1 - \frac{\alpha}{2}$ z $n - 3$ stopniami swobody. Podstawiając te wartości do powyższego wzoru otrzymujemy przedział ufności:

$$[4 - 2.11 \cdot 1,\ 4 + 2.11 \cdot 1] = [1.89,\ 6.11]$$

.

2 Zadanie 2

W tym zadaniu mamy daną regresję linową wieloraką postaci:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

Dodatkowo mamy tabelę sum I oraz II typu:

	Typ I	Typ II
X_1	300	30
X_2	40	25
X_3	20	?

Wiemy też, że $SST = 760$ oraz $n = 24$.

2.1 a)

W tym podpunkcie pytamy się ile wynosi suma II typu dla X_3.

Wiemy, że sumy typu I i II są równe dla ostatniej zmiennej objaśniającej($= SSM(X_3|X_1, X_2)$). Stąd suma typu II dla X_3 jest równa 20.

2.2 b)

Chcemy przetestować istotność parametru β_1 w pełnym modelu. Dlatego formułujemy hipotezę, że $\beta_1 = 0$. Statystyka testowa ma wtedy postać:

$$F = \frac{(SSE(R) - SSE(F))/(dfE(R) - dfE(F))}{MSE(F)} = \frac{SSE(X_1|X_2, X_3)}{MSE} = \frac{30}{40/20} = 1.5$$

Natomiast kwantyl wynosi $F_c = F(1 - \alpha, 1, 20) = 4.35$.

Widzimy, więc że $F < F_c$, czyli nie mamy podstaw do odrzucenia hipotezy zerowej.

2.3 c)

Tym razem testujemy hipotezę, że $\beta_2 = \beta_3 = 0$.

Statystyka testowa ma postać:

$$F = \frac{(SSE(R) - SSE(F))/(dfE(R) - dfE(F))}{MSE(F)} = \frac{(SSM(F) - SSM(R))/(dfE(R) - dfE(F))}{MSE(F)}$$
$$= \frac{(SSM - SSM(X_1))/(dfE(R) - dfE(F))}{MSE(F)} = \frac{(360 - 300)/2}{20} = 1.5 < 3.49 = F^*(0.95, 2, 20)$$

Znów nie mamy podstaw do odrzucenia hipotezy zerowej.

2.4 d)

Testujemy hipotezę, czy $\beta_1 = \beta_2 = \beta_3 = 0$. Wtedy statystyka testowa ma postać:

$$F = \frac{MSM}{MSE} = \frac{360/30}{400/20} = 6 > 3.098 = F^*(0.95, 3, 20),$$

czyli możemy odrzucić H_0.

2.5 e)

W tym podpunkcie tworzymy model liniowy, w którym mamy jedną zmienną objaśniającą X_1. Chcemy, wtedy przetestować, czy $\beta_1 = 0$. Wtedy statystyka testowa ma postać:

$$F = \frac{MSM(X_1)}{MSE(X_1)} = \frac{SSM(X_1)/dfM(X_1)}{(SSE(X_1|X2, X_3) + SSE(X_1, X_2, X_3))/dfE(X_1)} = \frac{300}{460/22} \approx 14.35 > 4.3 = F^*(0.95, 1, 22),$$

czyli odrzucamy H_0.

Widzimy, że wynik zmienił się w porównaniu do podpunktu **a)**.

2.6 f)

Chcemy obliczyć współczynnik korelacji między Y oraz X_1. Wiemy, że $\rho^2_{X_1,Y} = R^2 = \frac{SSM(X_1)}{SST} = \frac{300}{760}$, czyli $\rho_{X_1,Y} \approx \pm 0.628$.

3 Zadanie 3

W tym zadaniu chcemy zbadać wpływ korelacji zmiennych na diagnostykę modelu.

3.1 a)

W tym celu generujemy macierz planu $X_{100\times 2}$, taką że jej wierszami są niezależne wektory z rozkładu normalnego $N(0, \Sigma/100)$, gdzie

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

Następnie generujemy wektor odpowiedzi $Y = \beta_1 X_1 + \epsilon$, gdzie $\beta_1 = 3$, X_1 pierwszą kolumną macierzy X, a $\epsilon \sim N(0, I)$.

3.2 b)

Chcemy skonstruować 95% przedział ufności dla parametru β_1 i przeprowadzić test t na poziomie istotności $\alpha = 0.05$ dla hipotezy $\beta_1 = 0$ w modelu zredukowanym- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ oraz w modelu pełnym- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$.

Przedział ufności w przypadku modelu zredukowanego wygląda następująco:

$$[1.041,\ 5.639].$$

Statystyka testowa wynosi $T_R = 2.883 > 1.984 = t_{c_1}$, czyli odrzucamy hipotezę zerową.

W przypadku modelu pełnego przedział ufności przedstawia się następująco:

$$[-0.697,\ 9.305].$$

Statystyka testowa wynosi $T_F = 1.708 < 1.985 = t_{c_2}$, czyli nie mamy podstaw do odrzucenia hipotezy zerowej.

3.3 c)

W tym podpunkcie w obu modelach chcemy obliczyć estymator odchylenia standardowego β_1 oraz moc testu.

Skorzystamy ze wzoru $s^2(\beta_1) = s^2(\mathbb{X}'\mathbb{X})^{-1}_{2,2}$, gdzie $\mathbb{X}$ to macierz planu, a $s^2 = \frac{||Y-\hat{Y}||^2}{n-p}$.

W przypadku modelu zredukowanego estymator odchylenia β_1 jest równy $s(\hat{\beta}_1) = 1.158$, natomiast moc testu wynosi 0.727. Dla modelu pełnego mamy $s(\hat{\beta}_1) = 2.52$, a moc testu wynosi 0.218.

3.4 d)

Tym razem wygenerujemy 1000 niezależnych kopii ϵ i 1000 kopii zmiennej odpowiedzi. W każdym powtórzeniu będziemy estymować β_1 i przeprowadzimy test istotności dla β_1. Następnie obliczymy odchylenie standardowe wektora bet oraz moc testu.

W modelu zredukowanym otrzymaliśmy, że $s(\hat{\beta_1}) = 1.126$, a moc testu wyniosi 0.722. Natomiast w modelu pełnym $s(\hat{\beta_1}) = 2.454$, a moc testu wyniosi 0.237.

Widzimy, że w obu przypadkach wyniki te są do siebie zbliżone.

4 Zadanie 4

To zadanie będzie miało na celu pokazanie wpływu rozpatrywanej ilości zmiennych objaśniacających na diagnostykę modelu.

Tworzymy macierz planu $X_{1000\times950}$, której elementami są zmienne niezależne o rozkładzie $N(0, \sigma = 0.1)$. Następnie stworzymy wektor odpowiedzi

$$Y = X\beta + \epsilon,$$

gdzie $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)$, a $\epsilon \sim N(0, 1)$.

4.1 a)

W tym podpunkcie będziemy tworzyć model liniowy w oparciu o pierwsze 1, 2, 5, 10, 50, 100, 500, 950 kolumn macierzy planu X.

Dla każdego z tych modeli obliczymy sumę kwadratów residuów, MSE=$\frac{||X(\hat{\beta}-\beta)||^2}{n-p}$, gdzie p to liczba zmiennych objaśniających plus jeden. Dodatkowo obliczymy wartość kryterium AIC, p-wartości odpowiadające testom istotności dla dwóch pierwszych zmiennych w modelu oraz liczbę błędnych diagnoz testu istotności dla parametrów regresji.

Wyniki przedstawiają się w tabeli następująco:

Ilość zmiennych	SSE	MSE	p-wartości	AIC	Błędne diagnozy
1	1434.699	3×10^{-4}	0, -	3204.8	0
2	1318.225	0.0013	0, 0	3122.2	0
5	1057.508	0.0062	0, 0	2907.8	0
10	1056.246	0.0075	0, 0	2916.6	0
50	1008.506	0.0581	0, 0	2950.3	1
100	964.074	0.1106	0, 0	3005.3	4
500	531.861	1.0618	0, 0	3210.5	28
950	60.958	19.9525	0.032, 0	1944.3	14

Widzimy, że na podstawie kryterium AIC wybralibyśmy model uwzględniający 950 zmiennych.

4.2 b)

W tym podpunkcie powtórzymy eksperyment z poprzedniego, tylko tym razem będziemy brać zmienne, dla których estymowane współczynniki były największe.

Wyniki dla tego eksperymentu przedstawiają się następująco:

Ilość zmiennych	SSE	MSE	p-wartości	AIC	Błędne diagnozy
1	1402.159	0.001	0, -	3181.9	0
2	1304.604	7×10^{-4}	0, 0	3111.8	0
5	1301.08	0.2376	0, 0	3115.1	0
10	1289.302	0.2529	0, 0	3116	0
50	1078.302	0.4874	0, 0	3017.3	9
100	1014.272	0.6447	0, 0	3056	14
500	318.358	10.5992	0, 0	2697.3	295
950	60.958	375.967	0, 0.002	1944.3	14

Widzimy, że wyniki są podobne do tych z poprzedniego podpunktu. Duża różnica wystąpiła, w MSE dla modelu z 950 zmiennymi, ponieważ wzrosło niemal dwudziestokrotnie. Znacznie zwiększyła się liczba błędnych diagnoz testu w przypadku modelu z 500 zmiennymi. Dodatkowo w tym przypadku uwzgledniając kryterium AIC wybralibyśmy ostatni model, czyli ten sam co poprzednio.

W następnych zadaniach korzystać będziemy ze zbioru danych **CH06PR15.txt**. Zawiera ona informacje o wieku pacjenta, ciężkość choroby, poziom lęku i poziom zadowolenia.

5 Zadanie 5

W tym zadaniu tworzymy model liniowy, w którym wiek, ciężkość choroby, lęk są zmiennymi objaśniającymi, a satysfakcja jest zmienną odpowiedzi. Następnie podsumujemy ten model.

Estymowane równanie regresji ma postać:

$$\hat{Y} = 1.053 - 0.006X_1 + 0.002X_2 + 0.03X_3$$

Współczynnik determinacji wynosi $R^2 = \frac{SSM}{SST} = 0.542$.

Następnie przeprowadzamy test istotności na poziomie 95%, gdzie:

$$H_0: \quad \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1: \quad \exists_i \beta_i \neq 0$$

Statystyka testowa wynosi $F = \frac{MSM}{MSE} = 16.538$ i przy hipotezie zerowej ma rozkład Fishera-Snedecora z 3 i 42 stopniami swobody. Natomiast p-wartość wynosi 3.043×10^{-7}, stąd widzimy, że odrzucamy hipotezę zerową i zmienne objaśniające wpływają w jakiś sposób na satysfakcję.

6 Zadanie 6

W tym zadaniu chcemy skonstruować przedziały ufności dla parametrów β_i oraz przetestować, czy któryś z nich jest nieistotny.

Przedziały ufności zadają się wzorem:

$$\hat{\beta}_i \pm t_c s(\hat{\beta}_i),$$

gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n-4)$, $s^2(\hat{\beta}_i) = s^2(\mathbb{X}'\mathbb{X})^{-1}_{i+1,i+1}$.

Natomiast test wygląda następująco:

$$H_0: \quad \beta_i = 0 \quad \text{vs} \quad H_1: \quad \beta_i \neq 0$$

Statystyka testowa ma postać $T = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$.

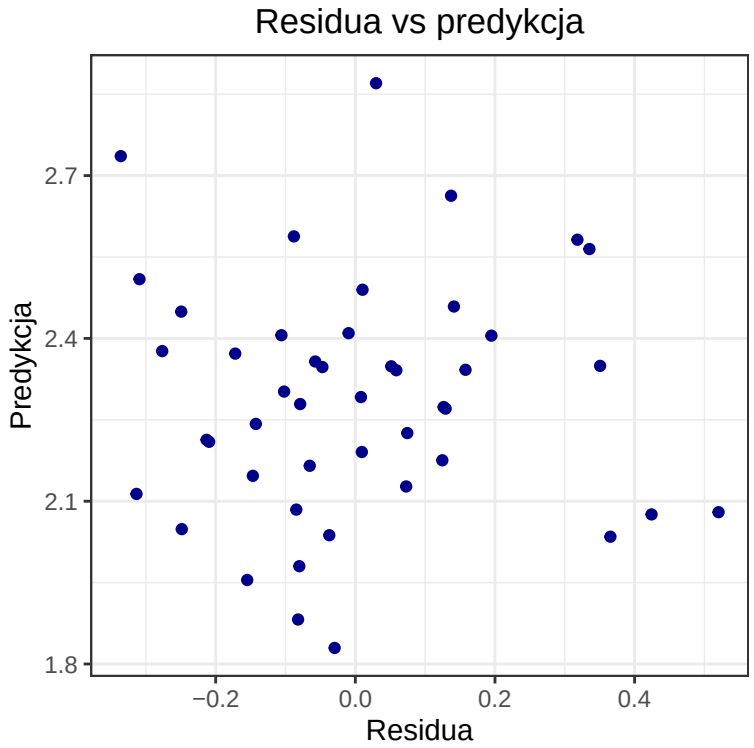
	Przedziały ufności	Statystyka testowa	p-wartość
β_1	$[-0.0121,\ 4 \times 10^{-4}]$	-1.897	0.065
β_2	$[-0.0097,\ 0.0136]$	0.333	0.741
β_3	$[0.0115,\ 0.0488]$	3.257	0.002

Widzimy, że dla tych przedziałów, które zawierają 0 p-wartość jest większa.

7 Zadanie 7

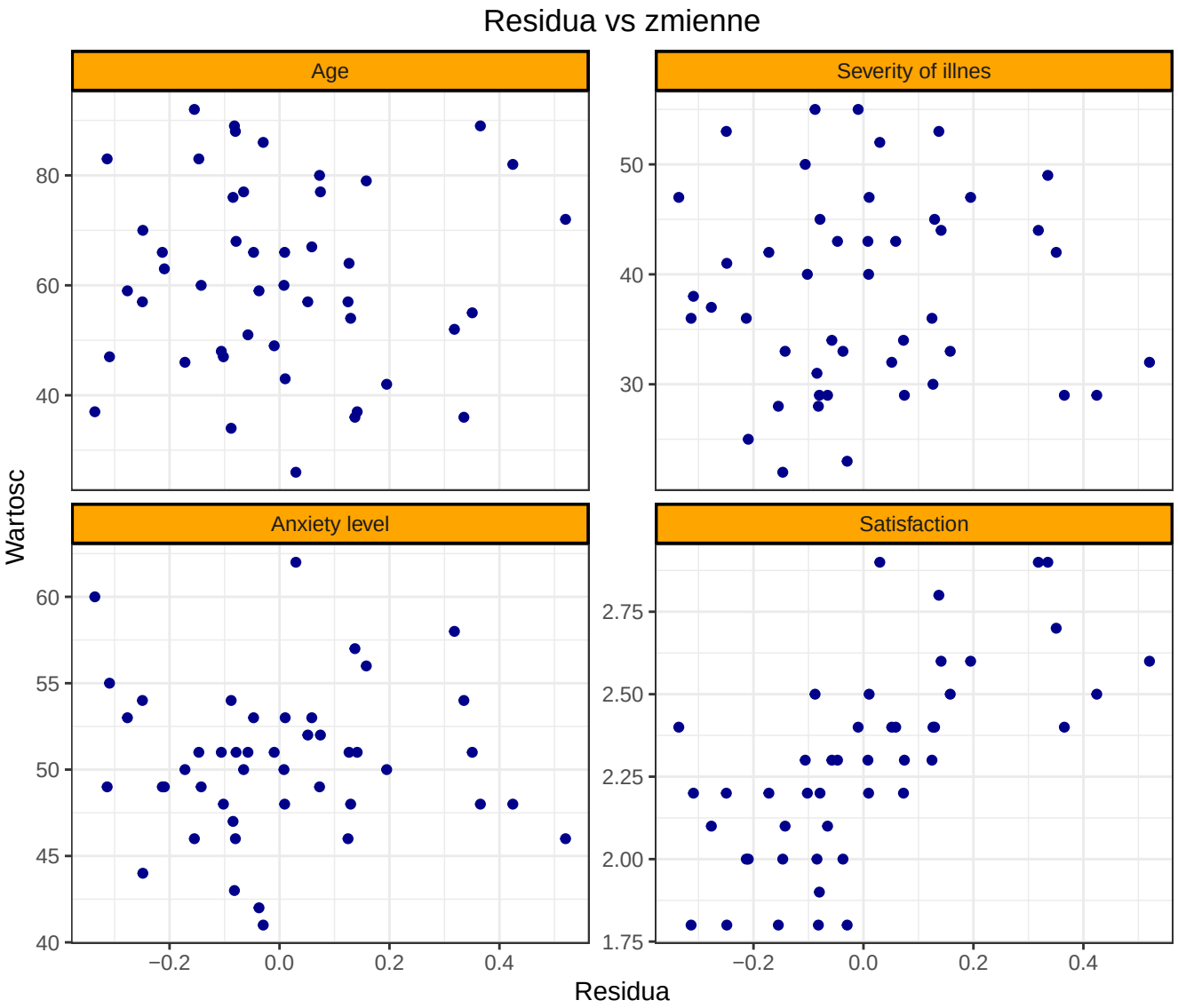
W tym zadaniu narysujemy wykres rozrzutu residuów między zmienną predykcją satysfakcji oraz zmiennymi objaśniającymi.

Poniżej przedstawia się wykres rozrzutu między residuami, a predykcją:



Na wykresie nie widzimy żadnych trendów.

Poniżej znajdują się wykresy residuów vs wszystkie zmienne w zbiorze:

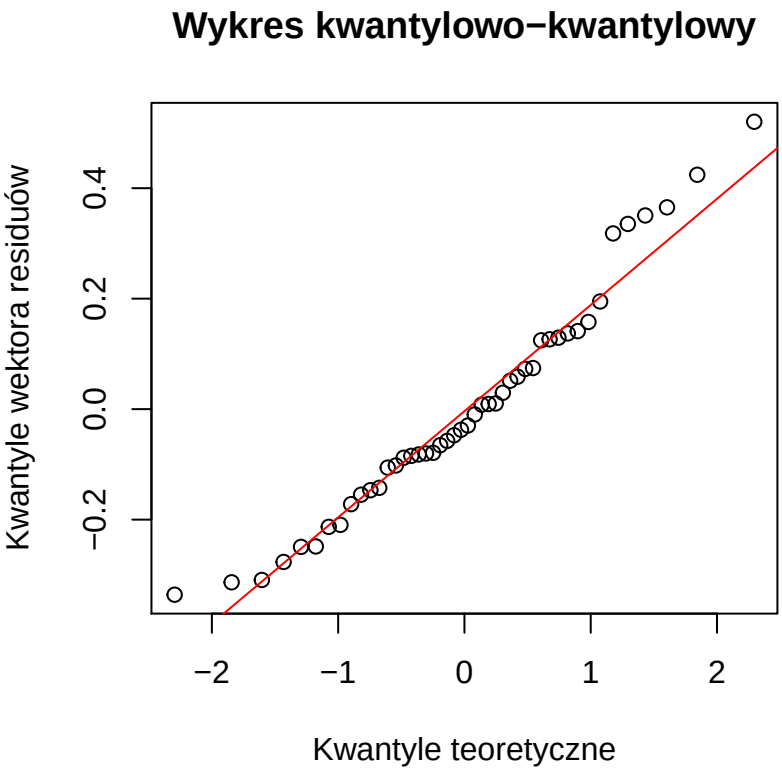


Widzimy, że na wszystkich wykresach punkty są rozrzucone w miarę losowo, wyjątkiem jest oczywiście zmienna objaśniana, czyli poziom satysfakcji. Chociaż i tam nie prezentują się jakieś szczególne wzorce. Możemy, więc stwierdzić, że te residua i te zmiennę są niezależne.

8 Zadanie 8

W ostatnim zadaniu chcemy sprawdzić, czy residua w naszym modelu mają rozkład normalny. W tym celu wykorzystamy test Shapiro-Wilka(funkcję w R **shapiro.test**). Dodatkowo narysujemy wykres kwantylowo-kwantylowy.

Poniżej prezentuje się wykres kwantylowo-kwantylowy:



Widzimy, że punkty układają się wzdłuż prostej „normalnej". Dodatkowo statystyka testowa z testu Shapiro-Wilka wynosi 0.962864, a p-wartość 0.1481, czyli nie mamy podstaw, żeby odrzucić hipotezę zerową, co w teście Shapiro-Wilka mówi nam, że prawdopodobnie zmienna ta ma rozkład normalny.
Podsumowując, możemy domyślać się, że residua mają rozkład normalny.