Analiza i przewidywanie kosztów leczenia osób objętych ubezpieczeniem zdrowotnym w Stanach Zjednoczonych

Aleksander Potok
Adam Pojda

2 czerwca 2025

Spis treści

Spis rysunków

Spis tabel

1 Opis danych, statystyki opisowe, źródło danych

1.1 Źródło danych

Dane pochodzą z książki *Machine Learning with R* autorstwa Bretta Lantza. Zawierają dane demograficzne i zdrowotne osób objętych ubezpieczeniem zdrowotnym w Stanach Zjednoczonych, a ich głównym celem jest przewidywanie kosztów leczenia (charges) na podstawie cech danej osoby. Dane są dostępne na portalu Kaggle.com:
https://www.kaggle.com/datasets/mirichoi0218/insurance

1.2 Opis danych

W zbiorze występują następujące zmienne:

- **Charges** - roczny koszt leczenia, w naszym modelu jest to zmienna objaśniana.
- **Age** - wiek pacjenta.
- **Sex** - płeć pacjenta.
- **Bmi** - BMI, Body Mass Index, współczynnik powstały przez podzielenie masy ciała podanej w kilogramach przez kwadrat wysokości podanej w metrach.
- **Children** - liczba dzieci objętych ubezpieczeniem.
- **Smoker** - zmienna binarna zaznaczająca, czy pacjent jest palaczem papierosów.
- **Region** - region USA w jakim mieszka pacjent.

1.3 Statystyki opisowe

Zmienna	Średnia	Mediana	SD	Min	Max	Skośność	Kurtoza
age	39,207	39,000	14,050	18,000	64,000	0,056	1,755
bmi	30,663	30,400	6,098	15,960	53,130	0,284	2,945
children	1,095	1,000	1,205	0,000	5,000	0,937	3,197
charges	13270,422	9382,033	12110,011	1121,874	63770,430	1,514	4,596

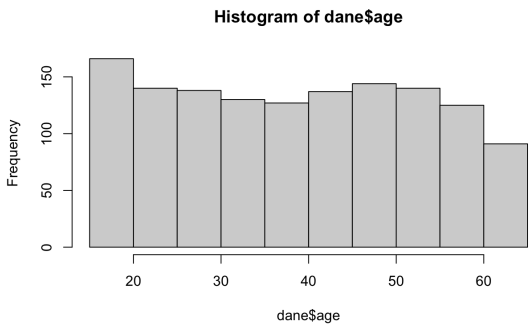
Tabela 1: Statystyki opisowe dla zmiennych numerycznych

Średnia i mediana nie są bliskie wartości maksymalnych i minimalnych.

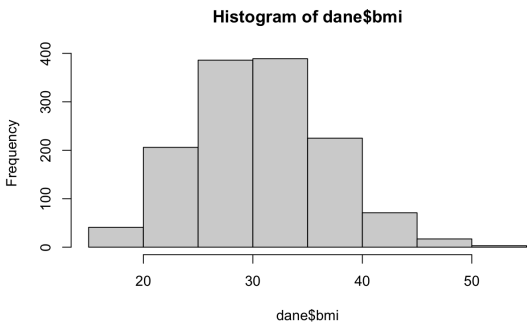
Zmienna `sex` przyjmuje 2 wartości: male (676 osób) i female (662), `smoker` przyjmuje 2 wartości: yes (274) i no (1064), `region` przyjmuje wartości northeast (324), northwest (325), southeast (364) i southwest (325).

1.4 Rozkłady zmiennych liczbowych

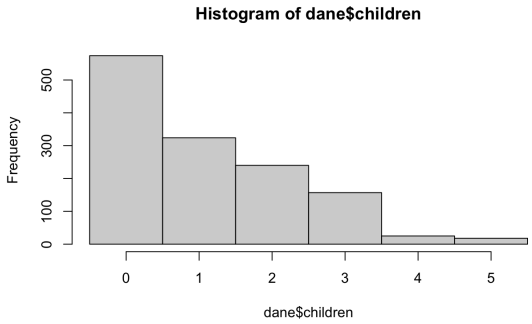
Wiek rozkłada się mniej więcej równo we wszystkich grupach wiekowych. Wartości BMI głównie oscylują w okolicy 30. Dominują osoby bez dzieci lub z małą liczbą dzieci (1-2). Koszty leczenia najczęściej nie przekraczają 15 000 dolarów, ale mimo wszystko jest dużo wartości wyższych.



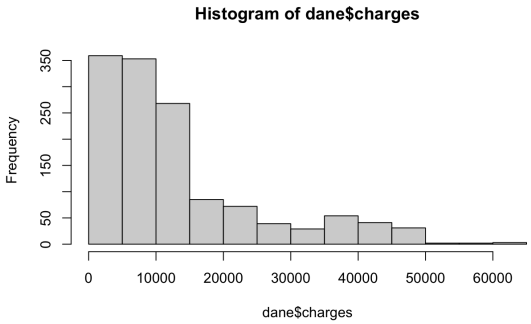
(a) Histogram zmiennej `age`



(b) Histogram zmiennej `bmi`



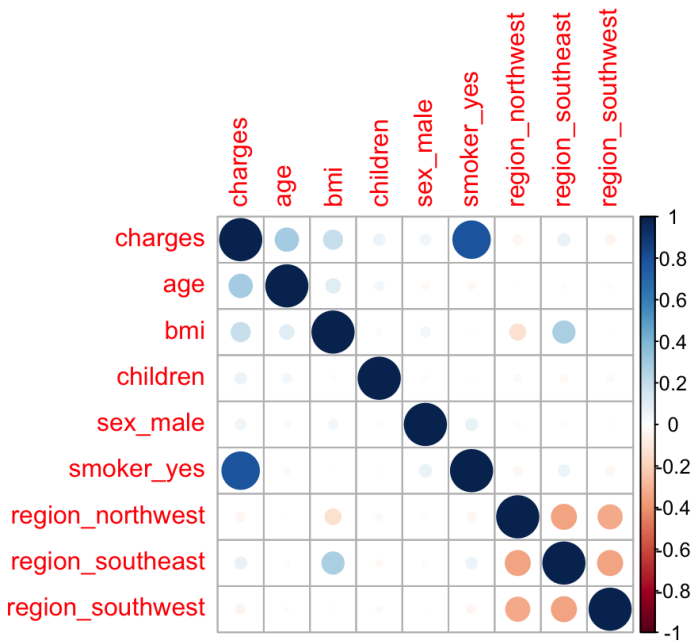
(c) Histogram zmiennej `children`



(d) Histogram zmiennej `charges`

Rysunek 1: Histogramy zmiennych liczbowych

1.5 Korelacja między zmiennymi

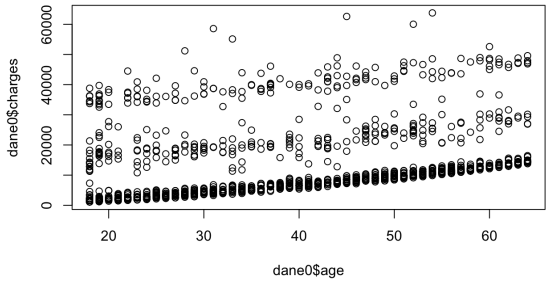


Rysunek 2: Macierz korelacji dla wybranych zmiennych

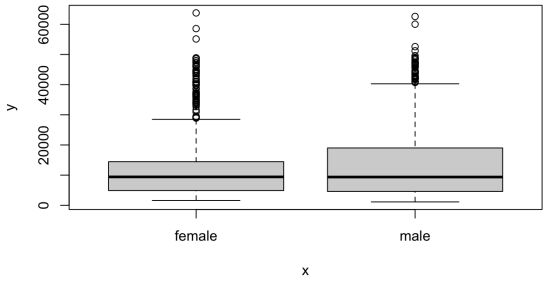
Największą dodatnią korelacją cechują się pary zmiennych: `charges` i `smoker`, `charges` i `age`, a nieco mniejszą – `charges` i `bmi`. Wśród zmiennych objaśniających największa korelacja występuje między `bmi` a `region_southeast`, jednak wynika ona z relacji między zmienną ilościową

a jedną z kategorii zmiennej jakościowej. W przypadku zmiennej `region` widoczne są korelacje ujemne, co jest naturalne, ponieważ są to zmienne binarne wzajemnie się wykluczające.

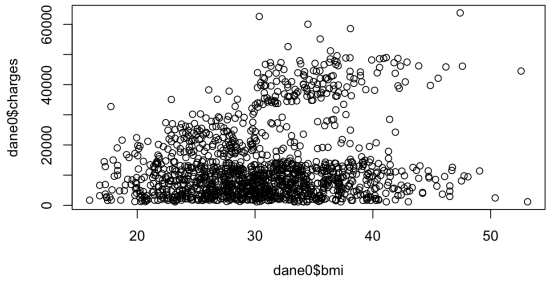
1.6 Rozkłady zmiennych objaśniających względem zmiennej objaśnianej



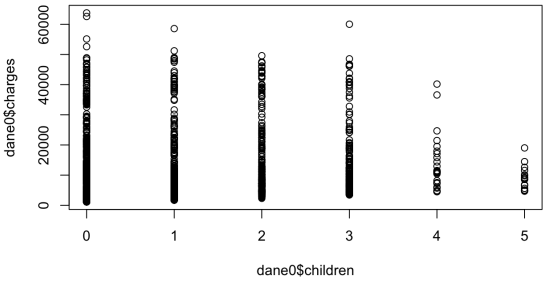
(a) age vs charges



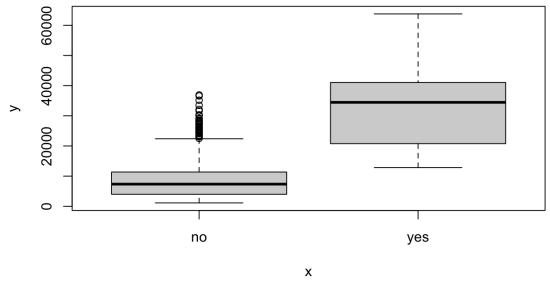
(b) sex vs charges



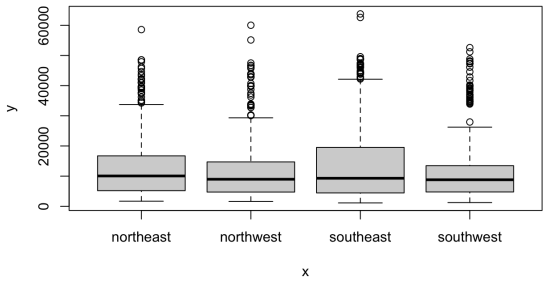
(c) bmi vs charges



(d) children vs charges



(e) smoker vs charges



(f) region vs charges

Rysunek 3: Wykresy zależności pomiędzy kosztami (charges) a zmiennymi objaśniającymi.

Na wykresach zmiennych `sex`, `children` i `region` nie obserwujemy żadnych oczywistych zależności. Na wykresach `age` i `bmi` widzimy kilka grup, ale różnice nie muszą być spowodowane tymi zmiennymi. Na wykresie `smoker` widzimy oczywistą zależność: jeśli ktoś pali to płaci o wiele więcej za leczenie.

2 Przygotowanie danych i konstrukcja modelu

2.1 Przekształcanie danych, podział zmiennych na zbiór treningowy i testowy

Na wykresie 3c widzimy, że zależność między `bmi` a `charges` nie jest liniowa. Osoby z bmi > 30 często mają znacznie wyższe koszty leczenia, a przed tym progiem relacja jest słaba lub nie ma jej wcale. Z tego powodu tworzymy zmienną binarną `obesity`, która pozwoli uchwycić naszemu modelowi punkt przełomowy. Nasza decyzja ma także uzasadnienie medyczne - bmi większe od 30 jest ustalonym i klinicznie uzasadnionym punktem definiującym otyłość.

Zamieniamy również pozostałe zmienne kategoryczne na zmienne binarne (dummy variables). Nasz zbiór danych dzielimy na treningowy i testowy w proporcjach 80:20.

2.2 Dobór zmiennych metodą Hellwiga

Zmienne do naszego modelu dobierzemy za pomocą metody Hellwiga.

`age` + `smoker_yes` + `obesity` = 0,7208844

`age` + `children` + `smoker_yes` + `obesity` = 0,7174251

`age` + `bmi` + `smoker_yes` = 0,7165987

Największe pojemności informacyjne otrzymujemy dla kombinacji ze zmienną binarną `obesity` zamiast `bmi`. Do naszego pierwszego modelu uwzględnimy dodatkowo zmienną `children`, mimo niższej pojemności informacyjnej, ponieważ w dalszym etapach analizy planujemy podział zbioru na palących i niepalących.

2.3 Postać modelu

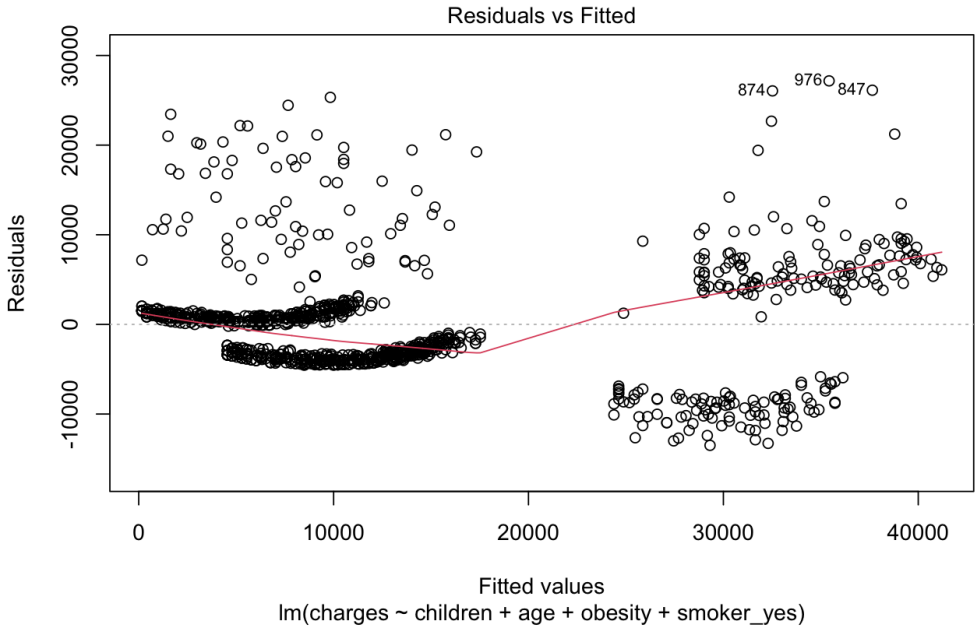
Tworzymy pierwszy model liniowy:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

```
    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4277.09     613.17  -6.975 5.35e-12 ***
children      546.82     155.16   3.524 0.000443 ***
age           246.58      13.52  18.238  < 2e-16 ***
obesity      4384.21     375.83  11.665  < 2e-16 ***
smoker_yes  24222.37     455.35  53.195  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6106 on 1065 degrees of freedom
Multiple R-squared:  0.7555, Adjusted R-squared:  0.7546
F-statistic: 822.7 on 4 and 1065 DF,  p-value: < 2.2e-16
```

Wszystkie zmienne są istotne statycznie. Skorygowana wartość R^2 wynosi 0,7546 . Estymowana wartość parametru dla zmiennej `smoker_yes` wynosi aż 24 222,37 i znacznie przewyższa pozostałe wartości. Może to oznaczać, że występują zmiany strukturalne i należałoby podzielić osoby w zależności od tego czy palą.



Rysunek 4: Wykres dopasowania reszt dla pierwszego modelu

Również na podstawie wykresu reszt widzimy tworzące się podgrupy co wskazuje, że postać modelu nie jest najlepsza.

3 Diagnostyka

3.1 Zmiany strukturalne - test Chowa

Do sprawdzenia czy istnieją u nas zmiany strukturalne wykonamy test Chowa.
H_0 : brak zmian strukturalnych
H_1 : istnieją zmiany strukturalne

Najpierw dzielimy nasz zbiór treningowy na podzbiór palących i niepalących, a następnie obliczamy wartość statystyki θ:

$$\theta = \frac{(S_C - (S_1 + S_2))/k}{(S_1 + S_2)/(N_1 + N_2 - 2k)} = 220,8493$$

Statystyka testowa ma rozkład F o k i $N_1 + N_2 - 2k$ stopniach swobody. Otrzymujemy p-wartość w przybliżeniu zero co oznacza, że lepiej podzielić dane.

3.2 Dobór zmiennych i postać nowych modeli

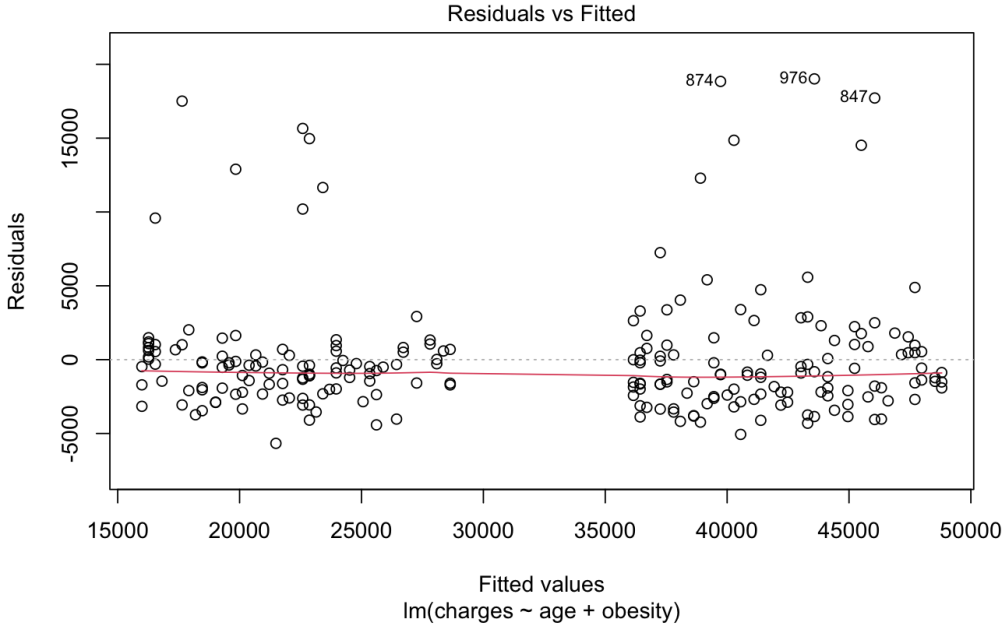
Ze względu na podział danych, musimy ponownie wybrać zmienne objaśniające. Tak jak poprzednio wybierzemy je za pomocą metody Hellwiga. Dla modelu z osobami niepalącymi największą pojemność informacyjną ma kombinacja `age` + `children`, natomiast dla osób palących `age` + `obesity`.

3.2.1 Osoby palące

```
    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11041.7      884.0   12.49   <2e-16 ***
age            274.9       20.3   13.54   <2e-16 ***
obesity      20165.2      557.4   36.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4208 on 226 degrees of freedom
Multiple R-squared:  0.8679, Adjusted R-squared:  0.8667
F-statistic: 742.1 on 2 and 226 DF,  p-value: < 2.2e-16
```

W przypadku modelu dla osób palących wszystkie zmienne są istotne statystycznie, a R^2 wynosi aż 0,8667.



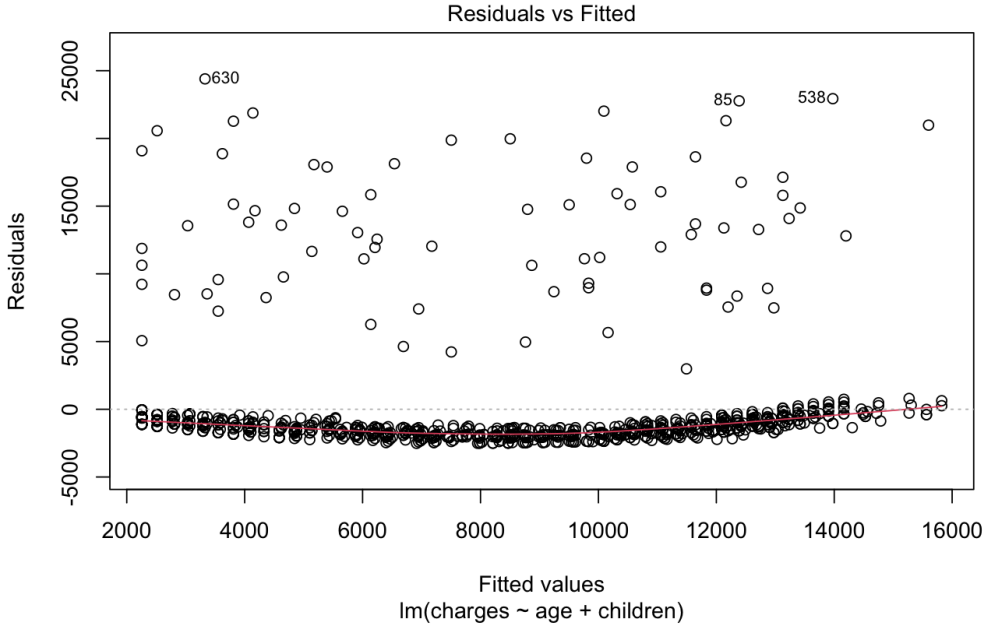
Rysunek 5: Wykres dopasowania reszt dla modelu osób palących.

3.2.2 Osoby niepalące

```
    Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2403.25     495.67  -4.849 1.48e-06 ***
age           258.86      11.39  22.731  < 2e-16 ***
children      553.27     129.56   4.270 2.17e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4593 on 838 degrees of freedom
Multiple R-squared:  0.3919, Adjusted R-squared:  0.3904
F-statistic:   270 on 2 and 838 DF,  p-value: < 2.2e-16
```

Dla osób niepalących również wszystkie zmienne są istotne statystycznie, a R^2 wynosi 0,3904. Widzimy, że współczynnik determinacji nie jest zbyt wysoki. Jest to spowodowane tym, że około 10% naszych obserwacji różni się znacząco od pozostałych. Inne zmienne w dostępne w naszych danych również nie wyjaśniają tego zjawiska. Najprawdopodobniej więc nie jest ono zawarte w zestawie danych, którymi dysponujemy.



Rysunek 6: Wykres dopasowania reszt dla modelu osób niepalących.

3.3 Specyfikacja modelu - test Ramseya

Kolejnym krokiem będzie zbadanie poprawności specyfikacji modelu, a więc określenie, czy przyjęta postać liniowa jest adekwatna. W tym celu wykonamy test Ramseya:

H_0 : Model jest poprawnie wyspecyfikowany

H_1 : Model jest błędnie wyspecyfikowany

3.3.1 Osoby palące

Dla osób palących otrzymujemy p-value równe 0,5341 - przyjęta postać linowa jest poprawna.

3.3.2 Osoby niepalące

Dla osób niepalący otrzymujemy p-wartość = 0,0003929, co sugeruje, że powinniśmy uwzględnić nieliniowe składniki. Po zapisaniu zmiennej `age` za pomocą funkcji `poly(age, 2)`, która generuje wielomianowe przekształcenia zmiennej w sposób ortogonalny przez co składniki nie są skorelowane ze sobą p-wartość rośnie do 0,7223.

Nasz nowy model dla osób niepalących ma $R^2 = 0,4032$ i wszystkie zmienne są istotne statystycznie.

```
    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7634.2      216.8  35.220  < 2e-16 ***
poly(age, 2)1 104270.3     4546.3  22.935  < 2e-16 ***
poly(age, 2)2  20866.2     4800.7   4.347 1.55e-05 ***
children         742.9      135.4   5.486 5.45e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4545 on 837 degrees of freedom
Multiple R-squared:  0.4053, Adjusted R-squared:  0.4032
F-statistic: 190.1 on 3 and 837 DF,  p-value: < 2.2e-16
```

3.4 Współliniowość - VIF

Badając współliniowość skorzystamy ze współczynnika VIF:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

gdzie R_j^2 współczynnik determinacji uzyskany z regresji zmiennej X_j na wszystkie pozostałe zmienne objaśniające. Interpretując wyniki korzystamy z podziału:

- VIF $= 1$ - brak współliniowości.
- VIF $\in (1, 5)$ - niewielka współliniowość.
- VIF > 5 - duża współliniowość, zalecana zmiana modelu.

Dla modelu osób palących współczynnik VIF otrzymuje wartość 1,00007 zarówno dla zmiennej `age`, jak i `obesity`. Dla osób niepalących wartość współczynnika VIF wynosi 1,027923 dla zmiennej `poly(age, 2)1` i `poly(age, 2)2` oraz 1,056625 dla `children`. Wszystkie wartości są bardzo zbliżone do jedności, co wskazuje na brak istotnej współliniowości pomiędzy zmiennymi objaśniającymi.

3.5 Autokorelacja - test Durbina Watsona

W celu weryfikacji obecności autokorelacji składnika losowego skorzystamy z testu Durbina-Watsona, którego statystyka przyjmuje postać:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

gdzie e_t - reszty z modelu regresji dla obserwacji t.

Dla osób niepalących p-value wynosi 0,7379, a dla palących 0,4948. Oznacza to brak podstaw do odrzucenia H_0. Przyjmujemy więc, że nie ma autokorelacji.

Należy jednak zauważyć, że dane mają charakter przekrojowy, a nie czasowy, więc autokorelacja nie powinna występować z definicji. Test przeprowadziliśmy jedynie w celu formalnej weryfikacji jednego z założeń klasycznego modelu regresji liniowej.

3.6 Heteroskedastyczność - test Breuscha-Pagana

Sprawdzając heteroskedastyczność skorzystamy z testu Breuscha-Pagana:

$$LM = n \cdot R^2$$

gdzie R^2 współczynnik determinacji z pomocniczej regresji reszt e_t^2 na zmienne objaśniające z modelu głównego.

Dla osób niepalących p-value wynosi 0,4829, a dla palących 0,823. Oznacza to brak podstaw do odrzucenia H_0, więc modele są homoskedastyczne. Dla porównania, w początkowym modelu p-value wynosiło $8{,}853 \times 10^{-16}$, co świadczyło o obecności heteroskedastyczności. Zatem modyfikacja modelu doprowadziła do usunięcia problemu heteroskedastyczności.

3.7 Normalność reszt

Sprawdzając normalność reszt skorzystamy z testu normalności Shapiro-Wilka:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

gdzie: $x_{(i)}$ to i-ta najmniejsza wartość w próbie.
$\bar{x}$ to średnia arytmetyczna próby.
a_i to wagi obliczane na podstawie kowariancji i wartości oczekiwanych dla próby z rozkładu normalnego.

Dla obu modeli p-value wynosi $2,2 \times 10^{-16}$, co prowadzi do odrzucenia hipotezy zerowej o normalności reszt. W związku z tym dalsze wnioskowanie o istotności parametrów oprzemy na bootstrapowych przedziałach ufności. Dla obu modeli wykonamy próby bootstrapowe na podstawie reszt.

W obu modelach percentylowe przedziały ufności uzyskane metodą bootstrap są bardzo zbliżone do klasycznych przedziałów ufności. Oznacza to, że brak normalności nie zniekształcił istotnie wyników w naszych modelach. W obu modelach dolne i górne granice przedziałów nie obejmują zera, co oznacza, że wszystkie zmienne mają istotny statystycznie wpływ na zmienną zależną.

3.7.1 Osoby palące

	$\hat{\beta}$	$\bar{\beta}^*$	Klasyczny PU 95%		Bootstrapowy PU 95%	
(Intercept)	11041,69	11043,25	9299,66	12783,73	9380,82	12844,11
age	274,95	274,89	234,95	314,95	235,45	315,50
obesity	20165,25	20157,18	19066,93	21263,57	19035,27	21223,29

Tabela 2: Porównanie klasycznych przedziałów ufności oraz uzyskanych metodą bootstrap dla modelu osób palących.

3.7.2 Osoby niepalące

	$\hat{\beta}$	$\bar{\beta}^*$	Klasyczny PU 95%		Bootstrapowy PU 95%	
(Intercept)	7634,20	7635,41	7208,75	8059,65	7228,67	8077,50
poly(age, 2)1	104270,26	104261,73	95346,83	113193,68	95388,01	113233,19
poly(age, 2)2	20866,17	20814,57	11443,38	30288,95	11682,65	30425,85
children	742,88	742,72	477,09	1008,67	483,18	1011,37

Tabela 3: Porównanie klasycznych przedziałów ufności oraz uzyskanych metodą bootstrap dla modelu osób niepalących.

4 Predykcja i błędy

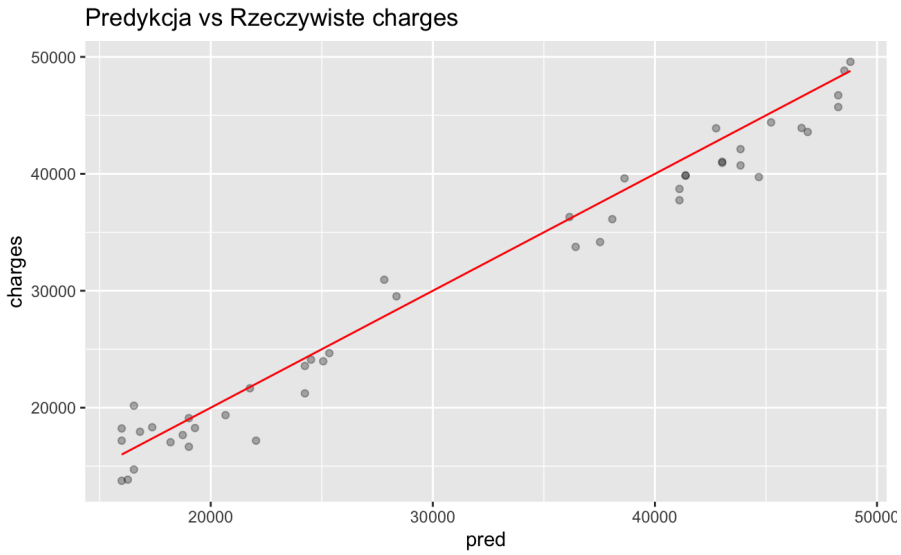
4.1 Miara dopasowania modeli – współczynnik determinacji R^2

Model	**Testowe R^2**
Osoby palące	0,964
Osoby niepalące	0,473
Model łączony (oba modele razem)	0,849

Tabela 4: Testowe wartości współczynnika determinacji R^2 dla poszczególnych modeli.

4.1.1 Osoby palące

Dla modelu dla osób palących otrzymaliśmy bardzo wysoki współczynnik determinacji na zbiorze testowym: $R^2 = 0,964$, co świadczy o niemal idealnym dopasowaniu modelu do danych. Oznacza to, że aż 96,4% zmienności kosztów leczenia (zmiennej `charges`) można wyjaśnić zmiennymi objaśniającymi zawartymi w modelu.

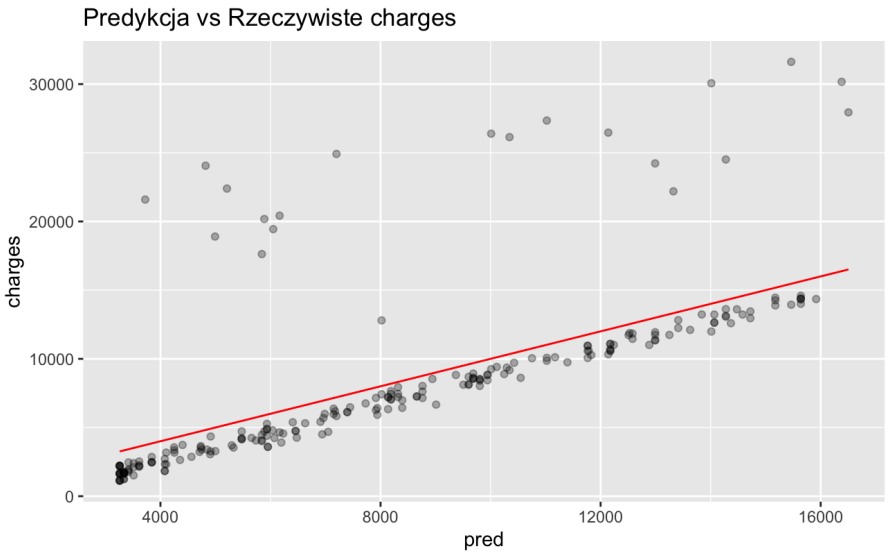


Rysunek 7: Predykcja vs rzeczywiste wartości dla osób palących

4.1.2 Osoby niepalące

Dla modelu osób niepalących wartość testowego R^2 wyniosła jedynie 0,473 co wskazuje na znacznie słabsze dopasowanie modelu – jedynie 47,3% zmienności zmiennej `charges` zostało

wyjaśnione. Tak jak zauważyliśmy wcześniej, około 10% danych z niewyjaśnionych powodów znacząco odstaje od pozostałych danych przez dopasowanie modelu jest niemożliwe.



Rysunek 8: Predykcja vs rzeczywiste wartości dla osób niepalących

4.1.3 Model łączony

Dla całego zbioru danych testowe R^2 wyznaczyliśmy na podstawie prognoz pochodzących z dwóch osobnych modeli (dla palących i niepalących). W tym celu połączyliśmy rzeczywiste wartości zmiennej `charges` oraz odpowiadające im predykcje z obu grup i wyznaczyliśmy R^2 ze wzoru:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Tak uzyskany współczynnik determinacji wyniósł $R^2 = 0,849$, co oznacza, że niemal 85% całkowitej zmienności zmiennej `charges` zostało wyjaśnione przez połączone modele. Dla porównania treningowe R^2 dla całego zbioru wyniosło 0,869. Wyniki są zbliżone, więc nie możemy mówić o nadmiernym dopasowaniu danych.

4.2 Błędy

Dla całego zbioru danych średni błąd bezwzględny (MAE) wynosi 2486.135. Średni błąd procentowy (MAPE) jest równy 0.288539, co oznacza to, że nasz model myli się przeciętnie o 28,85%. RMSE, czyli pierwiastek z średniego kwadratu błędu wynosi 4460,65.

4.2.1 Osoby palące

Błędy dla modelu osób palących są zdecydowanie lepsze. Model średnio myli się zaledwie o 1838,05 (MAE), a średni błąd procentowy (MAPE) wynosi zaledwie 7,02%. Wartość RMSE jest równa 2176,91.

4.2.2 Osoby niepalące

Dla modelu osób niepalących błędy są największe. MAE wynosi 2545,95, MAPE aż 37,81%, a RMSE 4640,67.

5 Podsumowanie

Celem projektu było zbudowanie modelu regresji służącego do analizy i przewidywania kosztów leczenia osób objętych ubezpieczeniem zdrowotnym w Stanach Zjednoczonych. W trakcie pracy udało się znacząco poprawić jakość dopasowania modelu – współczynnik determinacji na zbiorze treningowym wzrósł z 0,7546 do 0,869, co świadczy o skuteczności przyjętych modyfikacji. Wartość R^2 na zbiorze testowym wyniosła 0,849, co oznacza brak przeuczenia modelu.

Wszystkie zmienne uwzględnione w końcowym modelu okazały się istotne statystycznie. Dla osób palących największy wpływ na wzrost kosztów leczenia miała nadwaga (BMI > 30), zwiększając je średnio o 20 165,25. Wiek pacjenta również wpływał na koszty – każdy dodatkowy rok życia podnosił je średnio o 274,95. W przypadku osób niepalących model nie wyjaśniał zmienności zmiennej zależnej tak dobrze, co wynikało z obecności około 10% obserwacji odstających, trudnych do interpretacji. Zmienna liczby dzieci wpływała na koszty – każde dodatkowe dziecko zwiększało je średnio o 742,90. W modelu wykorzystano funkcję poly(age, 2), która tworzy ortogonalne przekształcenia wielomianowe zmiennej wiek. Choć podejście to redukuje problem współliniowości, utrudnia ono bezpośrednią interpretację współczynników regresji.

Ostateczny model cechuje się wysoką trafnością predykcji i dobrą jakością statystyczną, stanowiąc wiarygodne narzędzie do analizy kosztów leczenia w zależności od cech demograficznych pacjentów.