Uniwersytet Ekonomiczny
we Wrocławiu

Wykorzystanie modelu ekonometrycznego do oszacowania wpływu parametrów technicznych samolotów jednosilnikowych na ich cenę

Projekt zaliczeniowy z przedmiotu Narzędzia ekonometryczne w analizie danych

Rok akademicki 2021/2022, Semestr letni

Analityka Gospodarcza, 1 rok, II Stopień

Oskar Wenerowicz (173469)

Robert Zając (173474)

Spis treści

1. Określenie problemu badawczego

Przeprowadzona w projekcie analiza miała na celu zbadanie zależności pomiędzy parametrami technicznymi, a ceną samolotów jednosilnikowych produkowanych w latach 1947 – 1979 r.

2. Opis wykorzystanych danych

Wykorzystane w analizie dane pochodzą z pakietu „robustbase" dostępnego w R-Studio. Dotyczą one 23 jednosilnikowych samolotów zbudowanych w latach 1947-1979 oraz pochodzą z Office of Naval Research. Źródło: P. J. Rousseeuw and A. M. Leroy (1987) Robust Regression and Outlier Detection; Wiley, page 154, table 22.

Do analizy wykorzystano zbiór danych zawierający 1 zmienną objaśnianą i 4 zmienne objaśniające:

- Y – Cena samolotu (w 100 000$);
- X1 – Aspect ratio (pl. wydłużenie płata – oblicza się je dzieląc kwadrat rozpiętości przez powierzchnię nośną);
- X2 – Lift-to-Drag Ratio (pl. doskonałość aerodynamiczna - stosunek współczynnika siły nośnej do współczynnika oporu);
- X3 – waga samolotu (w funtach);
- X4 – maksymalny ciąg.

3. Analiza regresji

3.1. Oszacowanie parametrów modelu

W ramach analizy regresji został stworzony pierwszy model ekonometryczny, zawierający zmienną objaśnianą i wszystkie zmienne objaśniające.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = aircraft)

Residuals:
    Min      1Q  Median      3Q     Max
-14.891  -3.955  -1.233   5.753  17.594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.7913892 10.1157023  -0.375  0.71219
X1          -3.8529189  1.7630016  -2.185  0.04232 *
X2           2.4882665  1.1867538   2.097  0.05042 .
X3           0.0034988  0.0004790   7.305 8.72e-07 ***
X4          -0.0019537  0.0004986  -3.918  0.00101 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.406 on 18 degrees of freedom
Multiple R-squared:  0.8836,    Adjusted R-squared:  0.8578
F-statistic: 34.17 on 4 and 18 DF,  p-value: 3.501e-08
```

W przypadku tego modelu wyraz wolny okazał się nieistotny, a zmienna X2 okazała się istotna dopiero na poziomie istotności równym 0,1. W celu znalezienia najlepszego modelu, został stworzony drugi model, bez zmiennej X2, jednakże model 1 nie został ostatecznie odrzucony.

```
Call:
lm(formula = Y ~ X1 + X3 + X4, data = aircraft)

Residuals:
    Min      1Q  Median      3Q     Max
-14.818  -5.673  -1.435   4.419  19.448

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.4741002 10.9813903  -0.316  0.75518
X1          -2.9905703  1.8612765  -1.607  0.12461
X3           0.0032615  0.0005053   6.455 3.46e-06 ***
X4          -0.0014778  0.0004820  -3.066  0.00636 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.126 on 19 degrees of freedom
Multiple R-squared:  0.8552,    Adjusted R-squared:  0.8324
F-statistic: 37.41 on 3 and 19 DF,  p-value: 3.581e-08
```

Na podstawie testu T-Studenta przeprowadzonym dla zmiennych w drugim modelu, okazało się, że zmienna X1 również jest nieistotna, wyraz wolny również. W celu dalszego poszukiwania najlepszego modelu, stworzony został model 3.

```
Call:
lm(formula = Y ~ X3 + X4, data = aircraft)

Residuals:
     Min       1Q   Median       3Q      Max
-18.9947  -3.2270   0.4554   4.2111  20.0868

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.973e+01  4.438e+00  -4.445 0.000249 ***
X3           3.314e-03  5.238e-04   6.328 3.55e-06 ***
X4          -1.251e-03  4.787e-04  -2.613 0.016642 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.48 on 20 degrees of freedom
Multiple R-squared:  0.8356,    Adjusted R-squared:  0.8191
F-statistic: 50.81 on 2 and 20 DF,  p-value: 1.447e-08
```

W modelu 3 wszystkie zmienne oraz wyraz wolny okazały się istotne, jednakże model ten zawierał tylko 2 zmienne.

Ostateczny wybór padł jednak na model 1. Postanowiono go przyjąć na poziomie istotności α=0,1, dzięki czemu wszystkie zmienne okazały się istotne i trafiły do modelu. Skorygowany R^2 dla modelu 1 wyniósł 85%, co jest wynikiem lepszym niż dla modelu 3 (81%). Większa liczba zmiennych, a co za tym idzie większa ilość zawartych informacji sprawiły przeważyły szalę zwycięstwa na stronę modelu 1. Współczynnik R^2 wyniósł 88% co świadczy o bardzo dobrym dopasowaniu danych do modelu. Ostateczny model prezentuje się następująco:

$$Y = -3{,}853\,\mathrm{X1} + 2{,}488\,\mathrm{X2} + 0{,}003\,\mathrm{X3} - 0{,}002\,\mathrm{X4}$$

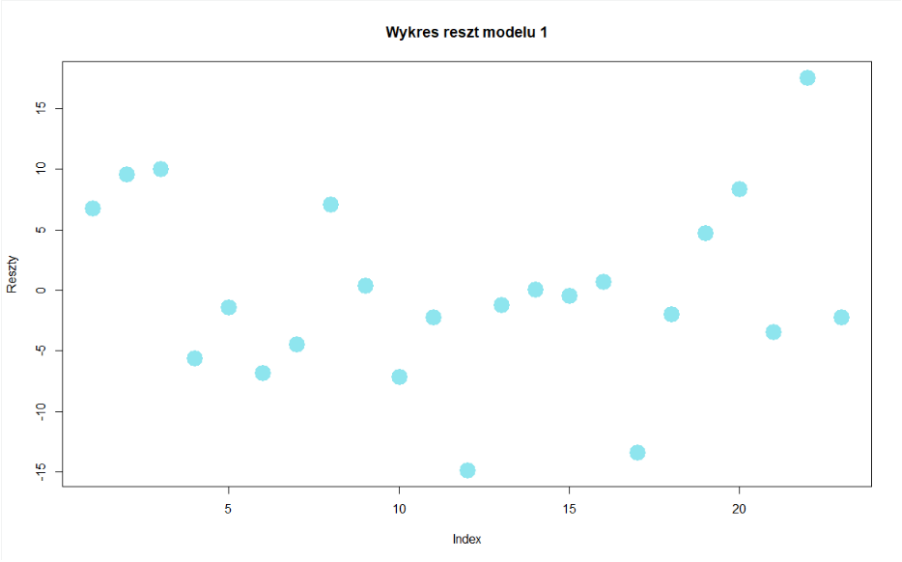
3.2. Interpretacja parametrów modelu:

- Wraz ze wzrostem *Aspect ratio* o 1 jednostkę, cena samolotu maleje o 385 000$ *ceteris paribus;*
- Wraz ze wzrostem *Lift-to-Drag Ratio* o 1 jednostkę, cena samolotu wzrasta o 248 800$ *ceteris paribus;*

- Wraz ze wzrostem wagi samolotu o 1 funt, cena samolotu wzrasta o 300$ *ceteris paribus;*
- Wraz ze wzrostem maksymalnego ciągu o 1 jednostkę, cena samolotu wzrasta o 200$ *ceteris paribus;*

3.3. Badanie normalności reszt modelu

Ułożenie punktów na wykresie reszt modelu wskazuje na normalność ich rozkładu.



W celu statystycznego potwierdzenia normalności reszt rozkładu, został przeprowadzony test Shapiro-wilka.

```
        Shapiro-Wilk normality test

data:  model_1$residuals
W = 0.97007, p-value = 0.6908
```

Wartość *p* jest zdecydowanie wyższa od przyjętego poziomu istotności α=0,1. Potwierdza to przypuszczenia płynące z wykresu – rozkład reszt jest normalny.

3.4. Badanie heteroskedastyczności

W celu sprawdzenia czy model jest heteroskedastyczny został przeprowadzony test White'a (dostępny w pakiecie „skedastic").

```
> white_lm(model_1)
# A tibble: 1 x 5
  statistic p.value parameter method       alternative
      <dbl>   <dbl>     <dbl> <chr>        <chr>
1      13.9  0.0836         8 White's Test greater
```

Wartość *p* jest większa niż 0,05, co sprawia, że nie występują podstawy do odrzucenia hipotezy zerowej H0. Występuję homoskedastyczność.

4. Weryfikacja istnienia problemu współliniowości

	Eigenvalue	Condition Index	intercept	X1	X2	X3	X4
1	4.34724160	1.000000	0.001492489	0.002377524	0.008934543	0.001555902	0.001588325
2	0.42605520	3.194287	0.006342241	0.066585006	0.002825651	0.010252378	0.022364760
3	0.18946152	4.790120	0.007955830	0.004244203	0.823526993	0.020756835	0.005470407
4	0.01964006	14.877690	0.981504002	0.833682340	0.002117784	0.110608991	0.010269280
5	0.01760163	15.715583	0.002705439	0.093110927	0.162595029	0.856825895	0.960307229

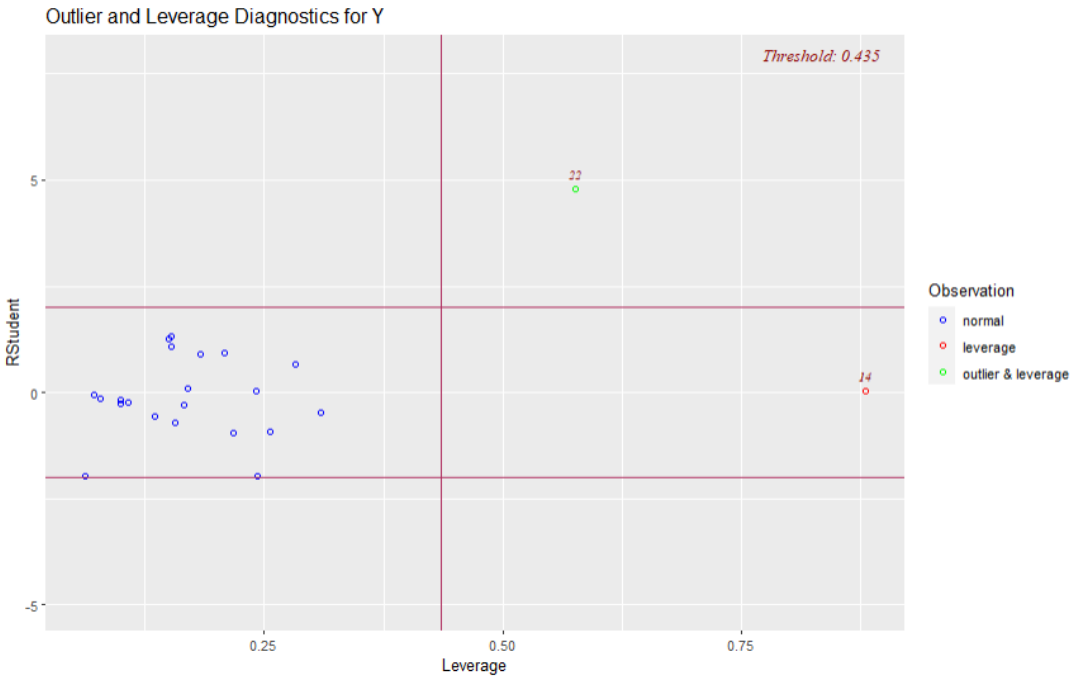
Dla wyrazu wolnego oraz zmiennych X1, X2 *Condition Index* wynosi mniej niż 10, co jest uznawane za dobry znak. Natomiast dla zmiennych X3, X4 *Condition Index* znajduje się w przedziale <10; 30>, a co za tym idzie należy się bliżej przyjrzeć tym zmiennym. Możliwe wystąpienie współliniowości.

```
> ols_vif_tol(model_1)
  Variables Tolerance      VIF
1        X1 0.5188004 1.927524
2        X2 0.6985447 1.431548
3        X3 0.1538220 6.501019
4        X4 0.1185881 8.432547
```

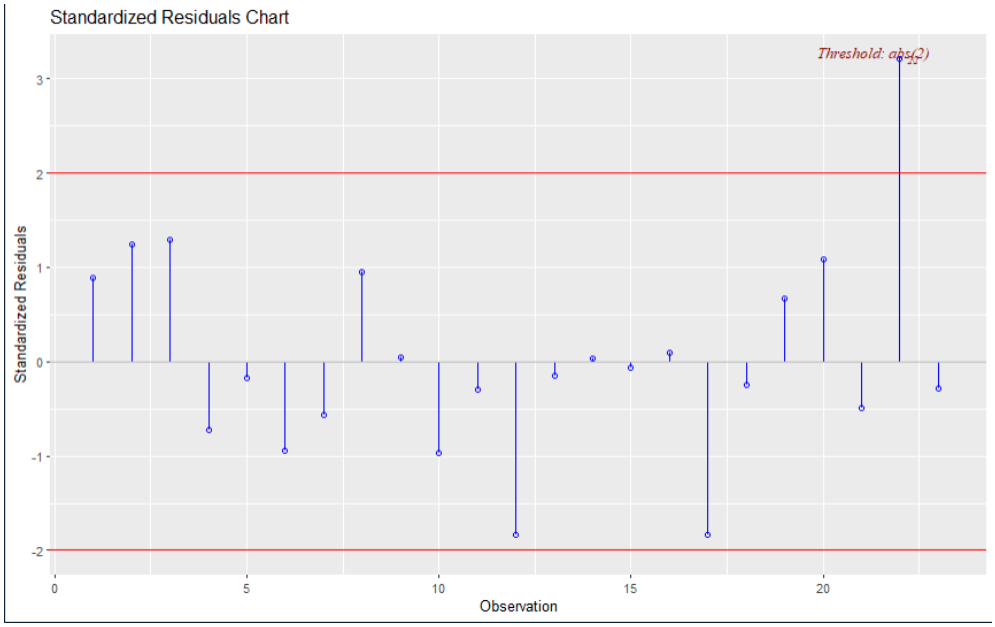
Dla żadnej ze zmiennych tolerancja nie jest mniejsza niż 0,1, a więc nie występuje problem z współliniowością. Również dla wszystkich zmiennych wartości statystyki VIF są poniżej 10.

Wartości *Tolerance* i *VIF* pozwalają odrzucić poprzednie przypuszczenie dotyczące współliniowości.
W przypadku modelu 1 nie występuje współliniowość.

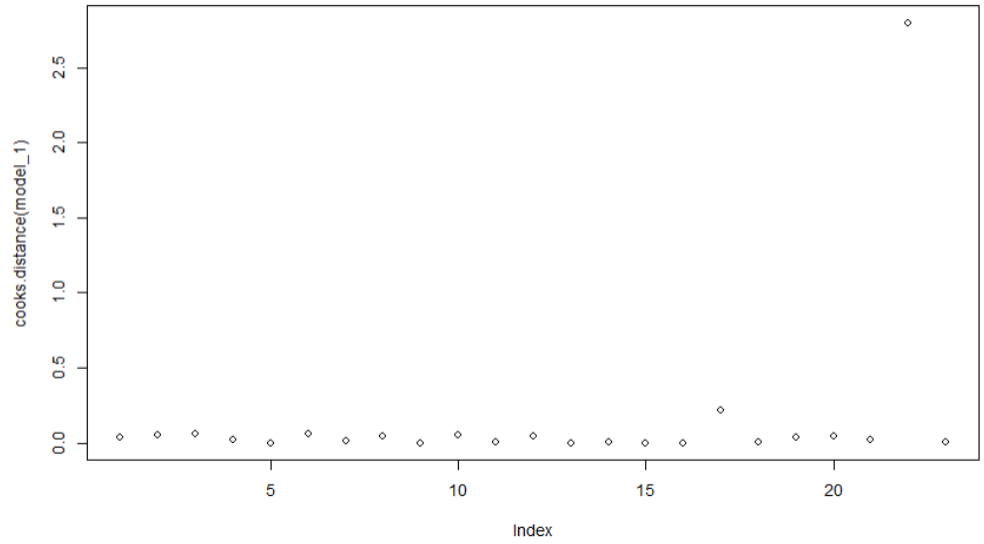
5. Identyfikacja obserwacji odstających, o wysokiej dźwigni, wpływowych



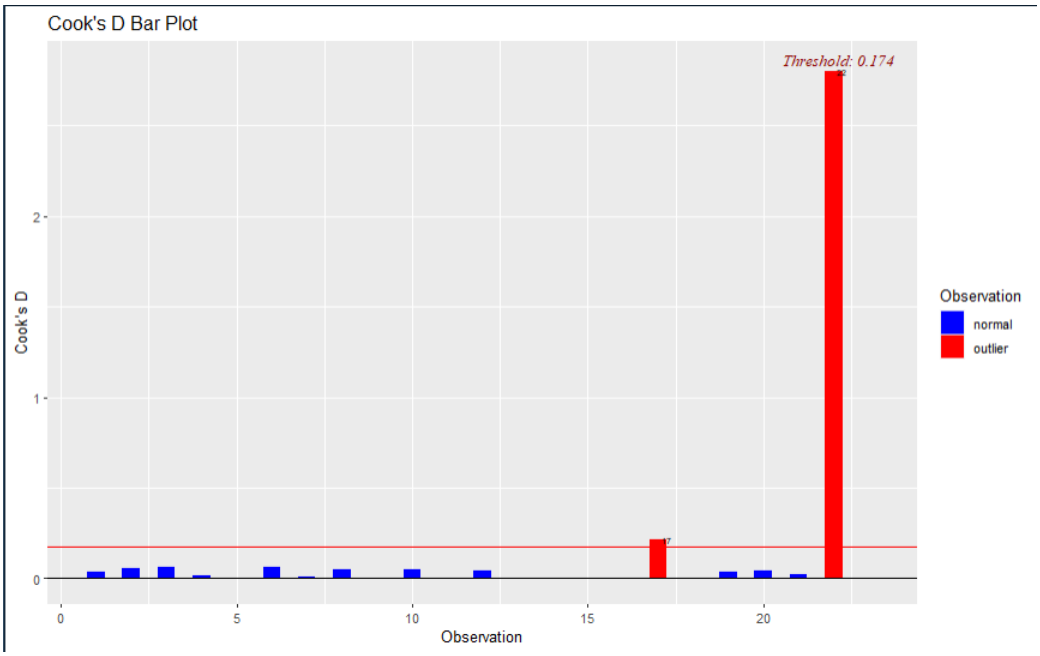
Powyższy wykres pozwala zaobserwować, że pośród danych są dwie obserwacje o wysokiej dźwigni (nr 14 i nr 22) oraz, że jedna z nich jest również obserwacją odstająca (nr 22).



Na wykresie standaryzowanych reszt można zaobserwować jedną obserwację odstającą. Jest to obserwacja nr 22.



Na wykresie odległości Cooke'a można zauważyć, że po raz kolejny obserwacja nr 22 jest obserwacją odstającą oraz obserwacja nr 17 wyróżnia się bardziej niż reszta. W celu zweryfikowania czy jest odstająca został wykonany kolejny wykres.



Powyższy wykres obserwację nr 17 jako odstającą, co nie pokrywa się z poprzednimi wykresami, a więc ciężko jednoznacznie określić ją mianem outlier'a.

6. Wnioski końcowe

Analiza regresji oraz stworzenie modelu ekonometrycznego pozwoliło stwierdzić, że największy wpływ na cenę jednosilnikowych samolotów produkowanych w latach 1947 – 1979 miały zmienne: *Aspect ratio* (pl. wydłużenie płata) i *Lift-to-Drag Ratio* (pl. doskonałość aerodynamiczna). Wzrost pierwszej z nich sprawia, że cena samolotu spada. Jest to spowodowane faktem, że najwyższe wartości *Aspect ratio* odnotowywane są w samolotach typu: szybowiec, które są relatywnie tańsze. Druga ze zmiennych wpływa natomiast na zmienną objaśnianą proporcjonalnie. Wraz ze wzrostem *Lift-to-Drag Ratio* wzrasta również cena samolotu ceteris paribus. Zmiennymi, które również wpływają na cenę jednosilnikowych samolotów, lecz nie w tak znacznym stopniu są jego waga oraz maksymalny ciąg.

Wykonany model ekonometryczny można uznać za poprawny, ponieważ jego reszty posiadają rozkład normalny oraz nie występuje heteroskedastyczność. Dalsza analiza wykazała również, że nie występuje współliniowość.

Identyfikacja obserwacji odstających, o wysokiej dźwigni, wpływowych wykazała, że w zbiorze danych występują dwie obserwacje o wysokiej dźwigni (nr 14 i nr 22). Oznacza to, że mają one silny wpływ na resztę obserwacji. Dodatkowo obserwacja nr 22 jest również outlier'em (obserwacją odstającą). Podejrzaną o bycie obserwacją odstającą jest również obserwacja nr 17, jednakże znajduje się ona na granicy i raz jest klasyfikowana jako outlier, a raz nie, dlatego ciężko podjąć ostateczną decyzję, gdzie ją przypisać.