AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
WYDZIAŁ ZARZĄDZANIA

SAMODZIELNA PRACOWNIA ZASTOSOWAŃ MATEMATYKI W EKONOMII

Praca dyplomowa licencjacka

Budowa modelu prognostycznego dla głównych czynników warunkujących ceny projektów z branży drogowej

Construction of a forecasting model for main factors influencing the prices of road construction projects

Autor: *Justyna Zbiegień*
Kierunek studiów: Informatyka i Ekonometria
Opiekun pracy: *dr Jacek Wolak*

Kraków, 2020

Spis treści

Streszczenie

Celem pracy jest zbudowanie modelu prognozującego liczbę godzin potrzebnych do realizacji danego projektu drogowego. Podjęto próbę budowy i statystycznej weryfikacji modelu liniowego (uwzględnia on m.in. długość drogi, liczbę kolidujących instalacji i poziom biurokracji) oraz modelu zbudowanego za pomocą drzew regresyjnych. Oba podejścia okazały się nie w pełni zadowalające: błąd prognozy MAPE w modelu liniowym wyniósł 25.57%, a dla drzewa regresyjnego otrzymano błąd przekraczający 60%. Są to błędy znacząco przewyższające błędy prognoz szacowanych przez przedsiębiorstwo za pomocą wiedzy eksperckiej, które wynoszą 10-15%.

The aim of the paper is construction of model forecasting number of hours needed for realization a road design. It was made an attempt to build and to make statistical validation of linear model (it contained i.e. length of road, number of colliding installations and level of bureaucracy) and model built with regression trees method. Both approaches turned out to be unsatisfactory: forecasting MAPE error for linear model was equal 25.57% and for regression trees it was obtained MAPE error equal over 60%. There are errors that significantly exceed the forecasting error estimated with professional knowledge by the company, which is equal 10-15%.

Wstęp

Rozwój polskiej gospodarki ma szczególne przełożenie na rozwój drobnej przedsiębiorczości. Wśród wskaźników ekonomicznych można zwrócić uwagę np. na rosnący udział tzw. mikroprzedsiębiorstw w gospodarce – w roku 2017 wyniósł on rekordowe 96,5%[1]. Na przestrzeni dziesięciolecia (lata 2008-2017) widać też wzrost ilościowy mikroprzedsiębiorstw od 1,79 mln do 2,08 mln. W tym samym okresie można zauważyć również stopniowy wzrost udziału sektora budownictwa w gospodarce (z 9,3% do aż 13,6% [1]).

Do sektora budownictwa zalicza się w szczególności przedsiębiorstwa projektowe, które są odpowiedzialne m.in. za tworzenie projektów drogowych na zlecenia zarówno prywatne, jak i publiczne. Zanim jednak przedsiębiorstwo projektowe będzie mogło podjąć się zlecenia publicznego, musi wziąć udział w przetargu publicznym, ogłaszanym przez zleceniodawcę, a jego oferta musi zostać uznana za najbardziej korzystną na podstawie zaproponowanej ceny za dany projekt, zebranego doświadczenia na rynku oraz innych zmiennych, uwzględnionych w danym przetargu.

Istotnym jest, aby przedsiębiorstwo trafnie wyceniło dane zleceni. Na tyle nisko, aby jego oferta była uznana za najbardziej korzystną oraz na tyle wysoko, by jego działalność przynosiła zysk. Branża drogowa jest branżą dość mało dynamiczną, jednak mimo to, na początku danego zlecenia jest wiele niewiadomych, które mogą zmieniać się wraz z tworzeniem projektu, a zatem tworzyć ryzyko niedokładności wyceny jego realizacji. Założono, że wycena danego projektu do przetargu sprowadza się do wyliczenia liczby godzin, które dana jednostka opracowująca spędzi nad jego realizacją. Ograniczenie to wprowadzono ze względu na możliwość przewidzenia kosztów zewnętrznych, tj. zleceń, podatków itd. oraz możliwość subiektywnej oceny kosztu jednej godziny, którą spędzi zespół projektowy nad realizacją projektu. Oszacowanie ceny potencjalnego projektu drogowego do przetargu publicznego jest rzeczywistym problem dla przedsiębiorstw projektowych. W dobie szybkiej zmiany przepisów, wysokiego współczynnika inflacji oraz braku możliwości przewidzenia przyszłego zakresu i wysokości inwestycji sektora publicznego w tym zakresie, problem ten przybiera znacząco na wadze, ponieważ właśnie

[1] P. Chaber i inni (2019, Warszawa) „Raport o stanie sektora małych i średnich przedsiębiorstw w Polsce" PARP s. 5

te czynniki mają istotny wpływ na realizację publicznych inwestycji, a co za tym idzie mają wysoki wpływ na branżę drogową.

Celem niniejszej pracy jest stworzenie modelu, który będzie mógł przewidzieć liczbę godzin potrzebnych do realizacji projektu drogowego na podstawie innych zmiennych znanych przed rozpoczęciem jego realizacji. Model zostanie oszacowany na podstawie danych użyczonych przez mikroprzedsiębiorstwo projektowe Pracownia Projektowa Jadwiga Zbiegień, zajmujące się inwestycjami drogowymi o długości do ok. 3 km. Są to dane o projektach drogowych realizowanych na zamówienia podmiotów publicznych w obszarze południowych województw Polski. Przewiduje się, że model potwierdzi podejrzenia o istnieniu dużego wpływu decyzji urzędów, zmieniających się przepisów prawnych oraz rosnącej biurokracji na ceny projektów drogowych.

Praca składa się ze wstępu, trzech rozdziałów merytorycznych oraz zakończenia. Rozdział pierwszy opisuje wybrane dane do analizy oraz opisuje użyte metody do skonstruowania modelu. W rozdziale drugim znajdują się przeprowadzone badania empiryczne, które zostały przeprowadzone w języku programowania R oraz w programie MS Excel. Rozdział trzeci zawiera wnioski na podstawie wyników badań empirycznych oraz podsumowanie wszystkich działań zawartych w pracy. Na ostatnich stronach można również znaleźć spis literatury, z której zaczerpnięto wiedzę niezbędną do przeprowadzenia badań i koniecznych analiz.

Rozdział 1: Opis zmiennych i metod

1.1 Wprowadzenie do danych – opis problemu

W celu lepszego zrozumienia występowania trudności związanych z budową modelu prognostycznego oraz zrozumienia zmiennych do niego przyjętych, poniżej opisany został zwykły proces realizacji projektu drogowego.

Po wybraniu w przetargu oferty przedsiębiorstwa projektowego, zobowiązuje się ono do realizacji projektu planowanej inwestycji. W ramach pierwszego etapu prac zleca ono stworzenie mapy geodecie, która następnie zostanie wykorzystana do celów projektowych oraz badania geologiczne w celu określenia nośności terenu. Na podstawie tych dokumentów tworzona jest koncepcja planowanej inwestycji, która musi zostać zaakceptowana przez inwestora, tj. zlecający podmiot publiczny.

Na początku kreowania koncepcji istotnym czynnikiem jest to, jakiego typu programu użyje się do stworzenia projektu technicznego. W praktyce działania takich przedsiębiorstw wykorzystywane są dwa najbardziej popularne programy: AutoCAD oraz AutoCAD Civil 3D. AutoCAD Civil 3D jest oprogramowaniem służącym do projektowania przestrzennego, dlatego sprawdza się on do projektowania dłuższych odcinków drogi. Automatycznie dopasowuje on narysowane poprzeczne obrazy drogi (tzw. poprzeczki) do właściwego jej projektu obejmującego cały zakres planowanej inwestycji. Jednak wadą tego programu jest wymagany wysoki poziom szczegółowości rysunku, przez co krótsze odcinki drogi są szybciej rysowane w mniej zaawansowanym programie – AutoCAD-zie.

Po zaakceptowaniu koncepcji projektu przechodzi się jednocześnie do przygotowania dokumentów zezwalających na budowę inwestycji. Kluczowym jest otrzymanie pozytywnej decyzji o środowiskowych uwarunkowaniach (DUŚ), w ramach wydania której Inspektorat Środowiska sprawdza, czy planowana inwestycja nie ma negatywnego wpływu na środowisko. Decyzja ta może zostać wydana z szeregiem wymogów, według których dana inwestycja ma być zaprojektowana. Do katalogu takich wymogów mogą wchodzić kategorie, takie jak użyte materiały, czy przebieg sieci. Po otrzymaniu pozytywnej DUŚ wysyła się zapytania do gestorów sieci, takich jak sieć gazowa, kanalizacyjna, elektroenergetyczna, kanalizacyjno-sanitarna, o ich przebudowę w celu realizacji inwestycji. Gestorzy mogą zadecydować o warunkach według których sieci zostaną przebudowane, m.in. o rodzajach kabli, odległości, wymaganych materiałach.

Jednocześnie czekając na decyzje gestorów sieci przedsiębiorstwo projektowe wysyła dokumentację dotyczącą odwodnienia drogi do Państwowego Gospodarstwa Wody Polskie i czeka na akceptację przedstawionego rozwiązania. Czas oczekiwania na decyzję wydaną przez ten organ administracji może obecnie przekraczać nawet rok. Państwowe Gospodarstwo Wody Polskie zostało powołane w roku 2018 na mocy ustawy Prawo Wodne z dnia 20 lipca 2017 roku. Przed wejściem w życie tej ustawy decyzję o odwodnieniu drogi podejmowało starostwo powiatowe, a czas oczekiwania na nią był znacznie krótszy - około 2 miesiące. Wraz z decyzją o odwodnieniu konieczne jest uzyskanie pozwolenia wodnoprawnego dotyczącego głównie wpuszczenia wód opadowych, zabudowania rowu, budowy przepustów oraz budowy mostów. Ustawa Prawo Wodne wprowadziła wymóg uzyskania pozwolenia wodnoprawnego w szerszym katalogu przypadków, a w tym dla małych inwestycji, dla których wcześniej nie było ono wymagane.

Po otrzymaniu pozwoleń oraz wydaniu niezbędnych decyzji, przedsiębiorstwo projektowe uszczegóławia koncepcję tworząc w ten sposób projekt budowlany. Wskazuje się w nim strony postępowania administracyjnego, którego przedmiotem jest uzyskanie zgody na przeprowadzenie prac budowlanych, a więc właścicieli działek, na których ma być realizowana inwestycja. Jeżeli inwestycja mieści się w pasie drogowym, czyli na działce należącej do gminy, oraz nie zawierają się w niej skomplikowane obiekty inżynierskie, postępowanie administracyjne w sprawie zgody na wykonanie prac budowlanych przeprowadza się w trybie zgłoszenia robót budowlanych. Jest to uproszczona procedura, w której składa się tylko dwa egzemplarze projektu, nie jest wymagane złożenie podpisów przez dwóch projektantów, a czas oczekiwania na akceptację zgłoszenia to 21 dni.

Następnie wykonywany jest szczegółowy projekt wykonawczy i kosztorys oraz przygotowuje się m.in. odpowiednią dokumentację do przetargu na wykonanie inwestycji. Natomiast, jeżeli w projekcie występują skomplikowane obiekty inżynierskie, takie jak mury oporowe, przepusty lub mosty, postępowanie administracyjne przeprowadza się w trybie uzyskania pozwolenia na budowę. Jest to procedura bardziej skomplikowana, do której wymagane jest złożenie czterech egzemplarzy projektu oraz zgoda wszystkich właścicieli działek sąsiadujących z inwestycją.

Uzyskanie pozwolenia na budowę jest również wymagane w przypadku procedury pozwalającej na odstąpienie od warunków technicznych, w ramach której wymagana jest zgoda ministra infrastruktury. Procedura ta wykorzystywana jest np. w przypadku braku możliwości zrobienia chodnika o wymaganej szerokości 2 metrów. W dokumentacji składanej w celu uzyskania pozwolenia na budowę wymagane są również podpisy dwóch projektantów posiadających uprawnienia oraz podpisy wszystkich projektantów na stronie tytułowej, co może być problematyczne w przypadku udziału w projekcie projektantów spoza przedsiębiorstwa.

Decyzja o pozwoleniu na budowę wydawana jest w czasie dwóch miesięcy z możliwością odroczenia w przypadku braku zgody na inwestycję co najmniej jednego właściciela działki sąsiadującej. Jeżeli inwestycja nie mieści się w pasie drogowym wówczas wymagane jest uzyskanie decyzji ZRID – Zgody na Realizację Inwestycji Drogowej. Jest to najbardziej czasochłonne i kosztowne postępowanie, ponieważ w jego ramach wymagany jest podział działek prywatnych, który musi zostać zlecony geodecie na koszt przedsiębiorstwa projektowego. Wydzielone części działek prywatnych są włączane do pasa drogowego, a co za tym idzie wywłaszczani są z nich właściciele, co powoduje zazwyczaj liczne protesty, które utrudniają i przedłużają proces realizacji inwestycji. Czas oczekiwania na decyzję o zgodzie na realizację inwestycji drogowej wynosi zazwyczaj ok. 3 miesiące. Istotnym czynnikiem wpływającym na przedłużenie procesu realizacji takiego projektu jest częsta zmiana uregulowań, których przedmiotem są inwestycje drogowe, a wraz z nią zwiększenie biurokracji oraz wymaganych dokumentów i pozwoleń.

1.2 Opis zmiennych wykorzystanych w modelowaniu

Dane dotyczące projektów drogowych zostały zaczerpnięte z danych udostępnionych przez przedsiębiorstwo Pracownia Projektowa Jadwiga Zbiegień. Są to dane dotyczące projektów drogowych pochodzących z okresu od I kwartału 2015 roku do drugiego kwartał 2019 roku. Dane podzielono według następujących zmiennych:

Y – liczba godzin poświęconych na realizację projektu;

X_1 – długość drogi w metrach;

X_2 – liczba kolidujących instalacji (np. instalacja gazowa, kanalizacja);

X_3 – zmienna binarna wskazująca na istnienie wymogu uzyskania pozwolenia wodnoprawnego (1 – wymagane, 0 – niewymagane);

X_4 – zmienna binarna wskazująca na okres realizacji projektu przypadający przed utworzeniem Państwowego Gospodarstwa Wody Polskie (1 – projekt po utworzeniu organu, 0 – projekt przed utworzeniem organu);

X_5 – skala trudności urzędów (w zależności od gminy) podana od 1-3;

X_6 – zmienna kategoryczna wskazujące na tryb uzyskiwania odpowiednich pozwoleń i decyzji:

$X_{6.1}$ – zgłoszenie;

$X_{6.2}$ – PNB (Pozwolenie na budowę);

$X_{6.3}$ – ZRID (Zgoda na Realizację Inwestycji Drogowej);

X_7 – zmienna binarna informująca o tym czy projekt przebiega przez teren zabudowany (1- teren zabudowany, 0 – teren niezabudowany);

X_8 – zmienna binarna informująca o tym, czy w projekcie znajdują się skomplikowane obiekty inżynierskie, np. mosty, przepusty (1 – znajdują się, 0 – nie znajdują się);

X_9 – zmienna kategoryczna informująca o dodatkowych elementach drogi:

$X_{9.1}$ – sama droga;

$X_{9.2}$ – droga z chodnikiem;

$X_{9.3}$ – droga ze ścieżką rowerową;

X_{10} – skala doświadczenia zespołu od 1 do 4;

X_{11} – wykorzystane oprogramowanie (1- AutoCAD Civil 3D, 0 -AutoCAD);

X_{12} – zmienna binarna mówiąca o tym, czy wymagana jest decyzja o uwarunkowaniach środowiskowych DUŚ (1 – wymóg decyzji, 0 – brak wymogu decyzji);

X_{13} – zmienna binarna informująca, czy projektowane jest więcej niż 300 m drogi (1- tak, 0 -nie);

X_{14} – zmienna binarna informująca, czy nastąpiła zmiana co najmniej jednego członka zespołu podczas trwania projektu (1- zmiana nastąpiła, 0 -zmiana nie nastąpiła);

X_{15} – zmienna binarna informująca, czy nastąpiła zmiana przepisów podczas trwania projektu (1 – zmiana nastąpiła, 0 – zmiana nie nastąpiła);

X_{16} – liczba wymaganych dokumentów w skali od 1 do 3.

Zmienne dotyczące skali są subiektywne i określone ekspercko przez dyrektora firmy na podstawie wieloletniego doświadczenia. Poniższa tabela przedstawia przykładowe obserwacje:

Tabela 1: Przykładowe obserwacje dotyczące wykonywanych projektów drogowych

Y	X_1	X_2	X_3	X_4	X_5	$X_{6.1}$	$X_{6.2}$	$X_{6.3}$	X_7	X_8	$X_{9.1}$	$X_{9.2}$	$X_{9.3}$	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
545	1000	1	1	0	1	0	0	1	1	0	1	0	0	3	0	1	1	0	0	1
186	400	1	0	0	2	0	0	1	1	0	1	0	0	3	0	0	1	0	0	1
178	150	1	1	0	2	1	0	0	0	0	0	0	1	3	0	0	0	0	0	1
137	800	0	0	0	2	1	0	0	1	0	1	0	0	4	0	0	1	0	0	1
290	400	0	1	0	3	1	0	0	0	1	0	0	1	3	0	0	1	0	0	2
212	400	0	1	0	3	1	0	0	0	1	0	0	1	3	0	0	1	0	0	2

Źródło: Opracowanie własne

1.3 Jednorównaniowy liniowy model ekonometryczny

Metoda najmniejszych kwadratów (MNK)

Niniejsza praca ma na celu – w szczególności – konstrukcję modelu ekonometrycznego za pomocą klasycznego modelu regresji liniowej. Model będzie skonstruowany na podstawie opisanych wyżej danych, a celem jego budowy jest próba określenia liczby godzin spędzonych nad danym projektem na podstawie zmiennych charakteryzujących ów projekt. Klasyczny model regresji przyjmuje postać:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \varepsilon \qquad (1)$$

gdzie:

y – zmienna zależna (objaśniana),

$x_1,\ldots,x_n$ – zmienne niezależne (objaśniające),

$\alpha_1,\ldots, \alpha_n$ – parametry,

ε - składnik losowy (czynnik stochastyczny równania).[2]

Aby model można nazwać regresją liniową, musi on spełnić szereg założeń zwanych założeniami Gausa-Markowa:

[2] P. Strawiński „Notatki do ćwiczeń z ekonometrii" s. 1

1. Zmienne niezależne są nielosowe i nieskorelowane ze zmienną losową (czyli nie są współliniowe);
2. Liczebność próby jest większa od liczby szacowanych parametrów, tj. spełniony jest warunek $n \geq k + 1$;
3. Egzogeniczność zmiennych niezależnych, tj. $E(\varepsilon_i) = 0$, gdzie i=1,…,n;
4. Występuje stałość wariancji: $D^2(\varepsilon_i)=I\delta^2$.

Celem estymacji jest dopasowanie linii regresji do badanych danych. Jedną z metod estymacji parametrów jest Metoda Najmniejszych Kwadratów, która dopasowuje parametry modelu w taki sposób, by suma kwadratów odległości punktów teoretycznych i empirycznych była możliwie najmniejsza.

$$\min \sum (y_i - \hat{y})^2 \qquad (2)$$

Estymowane parametry w myśl metody najmniejszych kwadratów otrzymuje się poprzez mnożenie macierzy według wzoru:

$$\hat{a} = (X^T X)^{-1} \cdot X^T Y \qquad (3)$$

gdzie:

â – wektor wyestymowanych parametrów,
X – macierz zmiennych objaśniających,
Y – macierz zmiennej objaśnianej.

Stosując się do zasad Gaussa-Markowa otrzymane estymatory parametrów w modelu liniowym można przedstawić za pomocą niższych wzorów[3]:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4)$$

$$\hat{\varepsilon} = \bar{Y} - \hat{\alpha}\bar{x} \qquad (5)$$

Estymatory otrzymane Metodą Najmniejszych Kwadratów są tzw. estymatorami BLUE – best, linear, unbiased estimator. Oznacza to, że są liniowe, zgodne, nieobciążone i efektywne[4].

[3] J. Jóźwiak, J. Podgórski „Statystyka od podstaw", Warszawa 2012, s. 361-365
[4] J. Mućk „Metody ekonometryczne. Metoda najmniejszych kwadratów" s.26

Podstawową miarą dopasowania modelu jest tzw. współczynnik determinacji R^2 (zwany też współczynnikiem korelacji wielorakiej w modelach z więcej niż jedną zmienną niezależną). Mówi on nam, jaki procent zmienności zmiennej objaśnianej (zależnej) jest wyjaśniona przez zbudowany model ekonometryczny. Ten współczynnik liczy się na podstawie wzoru na równość wariancji, która wynika z założeń Gausa-Markowa. Jest on następujący:

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2 \qquad (6)$$

Suma po lewej stronie to całkowita suma kwadratów (Total Sum of Squares). Natomiast po prawej stronie znajdują się składniki: estymowana suma kwadratów (Estimated Sum of Squares) oraz resztowa suma kwadratów (Residual Sum of Squares). Powyższe równanie można zatem zapisać jako:

$$TSS = ESS + RSS \qquad (7)$$

Dokonując następujących przekształceń otrzymany zostanie współczynnik R^2:

$$TSS = ESS + RSS \qquad /:TSS$$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = R^2$$

Zatem – jak widać – ten współczynnik jest ilorazem estymowanej sumy kwadratów oraz całkowitej sumy kwadratów i jest liczbą z przedziału od 0 do 1.

Należy pamiętać, że R^2 jest jedynie dobrą miarą dopasowania modelu dla modeli liniowych, ponieważ jest on miarą dokładności dopasowania prostej. Współczynnik ten może być sztucznie podwyższany lub fałszowany przez występowanie efektu katalizy lub autokorelacji w modelu, dlatego w dalszych częściach pracy zostaną opisane testy wykrywające te niepożądane zjawiska.

Gdy liczba szacowanych parametrów jest niewiele mniejsza od liczby dostępnych obserwacji, to współczynnik determinacji również może być zawyżony. Dlatego w takich modelach pod uwagę bierze się tzw. skorygowane $\overline{R^2}$, które jest zawsze niższe od zwykłego współczynnika, ale za to eliminuje zafałszowany wynik. W dobrze dobranym modelu liniowym spodziewane jest R^2 na jak najwyższym poziomie. Za dopuszczalne uznaje się R^2 powyżej 60%, natomiast współczynniki na poziomie wyższym od 80% uznaje się za dobrze dopasowane.

Kolejną miarą dopasowania modelu są kryteria informacyjne. W przeciwieństwie do współczynnika R^2, mogą one być miara nie tylko w modelach MNK. Obecnie wyróżnia się kilka podstawowych kryteriów informacyjnych, tj. AIC, BIC czy Hannana-Quinna. W tych badaniach użyto jedynie pierwszych dwóch i to one zostaną poniżej zdefiniowane.

Kryterium informacyjne Akaike AIC (z ang. **A**kaike **I**nformation **C**riterion) oraz Bayesowskie kryterium informacyjne Shwartza BIC (z ang. **B**ayes **I**nformation **C**riterion) są ze sobą ściśle powiązane i opierają się na funkcji prawdopodobieństwa – do modelu można dodawać zmienne powiększając prawdopodobieństwo dopasowania, jednak może to prowadzić do jego sztucznego podwyższania, aby do tego nie dopuścić kontroluje się właśnie AIC oraz BIC. W modelu najbardziej jest pożądane, aby kryterium BIC było jak najmniejsze, natomiast kryterium AIC jak największe.

W przypadku modelu MNK można te kryteria opisać danymi wzorami:

$$\text{AIC} = \log(\frac{e'e}{2}) + \frac{2K}{N} \tag{8}$$

$$\text{BIC} = \log(\frac{e'e}{2}) + \frac{K\,log(N)}{N} \tag{9}$$

gdzie:

$e'e$ - suma kwadratów reszt,

K – liczba zmiennych,

N – liczba obserwacji[5].

W dalszych podrozdziałach zostanie opisany szereg testów i metod, według których zostanie skonstruowany model ekonometryczny o możliwie najlepszym dopasowaniu (współczynnika R^2 oraz kryteriów informacyjnych) i jednocześnie spełniającym wszystkie założenia modelu liniowego. Jednak przed wykonywaniem wszelakich testów oraz przed stworzeniem pierwszego modelu należy usunąć dane odstające z badanej populacji. Do ich usuwania zostanie użyta zasady trzech sigm.

[5] J. Mycielski (2010) „Skrypt. Rozdział 10 Metodologia testowania hipotez", UW, s. 6

Istotność parametrów

W regresji liniowej nie wszystkie parametry są statystycznie istotne i zatem nie powinny należeć do zbudowanego modelu ekonometrycznego. Do zbadania danego zagadnienia zostanie wykorzystany jeden z testów statystycznych. Pozwoli on zweryfikować przyjętą hipotezę dotyczącą populacji na podstawie informacji zawartej w próbie (czyli dostępnych obserwacjach).

Każdy test statystyczny zbudowany jest z dwóch hipotez: zerowej oraz alternatywnej. Hipoteza zerowa (H_0) zwykle dotyczy wartości parametru, natomiast hipoteza alternatywna (H_1) jest jej dopełnieniem (najczęściej zaprzeczeniem hipotezy H_0). Hipotezy te wyglądają następująco:

$$H_0: \alpha_k = c$$

$$H_1: \alpha_k \neq c$$

gdzie:
c – stała określona dla danego modelu,
α_k – k-ty parametr w modelu (k = 1,…,n).

Decyzję o przyjęciu lub odrzuceniu hipotezy zerowej określa się na podstawie wartości statystyki testowej, czyli wartości obliczonej na podstawie obserwacji w próbie. Należy pamiętać, że przy założeniu prawdziwości hipotezy zerowej wartość statystyki testowej jest znana. W testach statystycznych istnieje również pojęcie obszaru odrzuceń. Jest on obszarem wartości odstających (nietypowych) dla znanej statystyki testowej (przy założeniu prawdziwości hipotezy zerowej), występujących z bardzo niskim prawdopodobieństwem[6]. Przy testach statystycznych mogą jednak występować błędy. Błędy te dzieli się na błędy I oraz II-rodzaju. Poniższa tabela przedstawia kiedy oba rodzaje błędów zachodzą:

[6] M. Topolewski „Ekonometria I, Temat 3., Weryfikacja istotności oszacowań parametrów" s.4

Tabela 2: Rodzaje błędów w testach statystycznych

	H_0 prawdziwa	H_0 fałszywa
Brak odrzucenia H_0	Decyzja prawidłowa z prawdopodobieństwem 1-α	**Błąd II-rodzaju** z prawdopodobieństwem β
Odrzucenie H_0	**Błąd I-rodzaju** Z prawdopodobieństwem α	Decyzja prawidłowa Z prawdopodobieństwem 1- β

Źródło: Opracowanie własne

Jak wynika z powyższej tabeli błąd I-rodzaju zachodzi, gdy odrzucana jest prawdziwa hipoteza zerowa, natomiast błąd II-rodzaju zachodzi gdy przyjmowana jest fałszywa hipoteza zerowa. Z tabeli również można odczytać, że prawdopodobieństwo popełnienia błędu I-rodzaju wynosi α, czyli poziom istotności. W testach statystycznych przyjmowany on jest najczęściej na poziomie 0.01, 0.05 lub 0.1. W niniejszej pracy wszystkie testy statystyczne będą przyjmowały $\alpha = 0.05$.

Wyżej wspomniano, że decyzję o przyjęciu lub odrzuceniu hipotezy zerowej można podjąć na podstawie wartości statystyki testowej, jednak najczęściej decyzje podejmuje się na podstawie wartości-*p* (*p-value*), empirycznego poziomu istotności. Wartość-p to najniższy poziom istotności jaki musi zostać przyjęty, aby odrzucić hipotezę zerową, przy danej wartości statystyki testowej. Inaczej mówiąc wartość-*p* to prawdopodobieństwo popełnienia błędu I-go rodzaju. Z tego wynika, że jeśli pożądanym jest przyjęcie hipotezy alternatywnej, *p-value* jest spodziewane na bardzo niskim poziomie, tj. poniżej poziomu α, czyli w tym wypadku poniżej 0.05. Natomiast, jeżeli pożądane jest przyjęcie hipotezy zerowej *p-value* jest spodziewane na poziomie większym od α, czyli większym od 0.05.

Do badania istotności oszacowanych parametrów w regresji liniowej jest wykorzystywany test t-studenta. Test ten sprawdza, czy parametr jest statystycznie różny od zera (czyli istotny). Jego hipotezy wyglądają następująco:

$$H_0: \alpha_k = 0$$

$$H_1: \alpha_k \neq 0$$

gdzie α_k to k-ty parametr w modelu ($k = 1,\ldots,n$)[7].

[7] T. Kufel (2013, Warszawa),„Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL" s.57

Przy założeniu, że $\bar{x}$ posiada rozkład normalny N(μ_0, $\sigma_{\bar{x}}$), wartość statystyki testowej t-studenta opisuje się wzorem:

$$t = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \qquad (10)$$

gdzie:

$\bar{x}$ - średnia z próby,

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$,

n – liczba obserwacji (n-1 to liczba stopni swobody).

Test t-studenta mierzy jedynie indywidualną istotność poszczególnych zmiennych. Do sprawdzenia istotności wszystkich zmiennych jednocześnie stosuje się test Walda. Polega on na „ucinaniu" podstawowego modelu i sprawdzaniu istotności parametrów za pomocą współczynnika determinacji R^2 (jest to jedna z wersji tego testu).

Hipotezy tego testu są następujące:

$$H_0: \alpha_1 = \ldots = \alpha_k = 0$$

$$H_1: \text{co najmniej jeden parametr } \alpha_i, i=1,2,\ldots,k, \text{ jest różny od zera.}$$

Po oszacowaniu modelu podstawowego i obliczeniu współczynnika R^2 wyznacza się wartość statystyki testowej:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-(k+1)}{k} \qquad (11)$$

gdzie n to liczba obserwacji oraz k to liczba zmiennych. Statystyka *F* ma rozkład F-Snedecora z $r_1 = k$ oraz $r_2 = n-(k+1)$ stopniami swobody.

Badanie normalności rozkładu

Jedno z podstawowych założeń dotyczących modelu liniowego mówi o normalności rozkładu reszt. Jest to bardzo ważne założenie konieczne do spełnienia. Jest tak, gdyż o ile składnik losowy jednorównaniowego modelu ekonometrycznego ma rozkład normlany, to estymator uzyskany za pomocą MNK ma własności użyteczne w konstruowaniu testów w celu sprawdzenie różnych cech modelu. Pozytywna ocena tego założenia ma więc zasadnicze znaczenie w procesie weryfikacji modelu.

Najpopularniejszym testem normalności rozkładu jest test Shapiro-Wilka, choć oczywiście istnieją również inne testy służącego do weryfikacji pytania o rozkład (np. test Jarque-Bera, test Andersona-Darlinga).

Test Shapiro-Wilka przyjmuje następujące hipotezy:

H_0: Próba pochodzi z populacji o rozkładzie normalnym

H_1: Próba nie pochodzi z populacji o rozkładzie normalnym

W celu uzyskania statystyki testowej *W* należy wykonać poniższe obliczenia:

$$S^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\text{dla n nieparzystego: } m = \frac{n-1}{2}$$

$$\text{dla n parzystego: } m = \frac{n}{2} \qquad (12)$$

$$b = \sum_{m}^{n}(a_m(y_{n+1-m} - y_m))$$

$$W = \frac{b^2}{S^2}$$

gdzie:
n – liczba obserwacji,
y – obserwacje zmiennej objaśnianej,
a_i – wartości odczytywane z tablicy, i = m,…,n.

Jak wynika z powyższych wzorów oraz hipotez w tym teście, pożądanym jest uzyskanie wartości *p-value* na wysokim poziomie (powyżej poziomu istotności).

Korelacja zmiennych

W modelu regresji kolejnym podstawowym wyznacznikiem przyjmowania zmiennych do modelu jest korelacja zmiennych. Korelacje opisuje się jako liniowa współzależność zmiennych i obiektów[8]. Korelacja może przyjmować wartość między -1 a 1, przy czym im korelacji jest bliżej 0, tym związek między obserwowanymi zmiennymi jest mniejszy.

Jak wcześniej opisano we wcześniejszym podrozdziale, pożądanym jest uzyskanie jak najwyższej korelacji zmiennych niezależnych ze zmienną zależną i zarazem jak

[8] Wydawnictwo Naukowe PWN „Nowa encyklopedia powszechna PWN. Tom 3” s.492

najmniejszej korelacji pomiędzy zmiennymi niezależnymi. Do zbadania korelacji pomiędzy zmiennymi stosuje się najczęściej jeden z dwóch wzorów: współczynnik korelacji rang Spearmana oraz współczynnik korelacji Pearsona.

Użyty w pracy współczynnik korelacji Pearsona wylicza się ze wzoru:

$$r_{xy} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{13}$$

Oznacza to, że liczy on iloraz kowariancji zmiennych i iloczyn ich odchyleń standardowych. Wzór ten można rozpisać i przedstawić w następującej postaci:

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{14}$$

Jest to wówczas estymator współczynnika korelacji Pearsona.

Metoda Hellwiga a metoda krokowa

W ekonometrii istnieją dwie podstawowe metody doboru zmiennych do modelu regresji liniowej. Pierwszą z nich opracował profesor Z. Hellwig w 1969 roku. Jest ona znana jako metoda optymalnego wyboru predykant albo metoda wskaźników pojemności informacji. Metoda ta wykorzystuje zasadę według, której zmienne niezależne powinny być jak najmniej współzależne, ale za to wysoce skorelowane ze zmienną objaśnianą Y[9]. Na podstawie macierzy korelacji:

$$R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{bmatrix}$$

liczona jest możliwe największa pojemność integralna i na jej podstawie wybierany jest zestaw zmiennych do modelu. Wzór na integralną pojemność informacyjną jest następujący:

$$H = \Sigma h_{ij} \tag{15}$$

gdzie:

h_{ij} – indywidualna pojemność informacyjna,

[9] A. Kopiński, D. Porębski (2014) „Zastosowanie metody Hellwiga do konstrukcji modelu ekonometrycznego dla stóp zwrotu funduszy inwestycyjnych" s. 3

i – numer kombinacji (i = 1,2,..,2^m-1),

j – numer zmiennej rozpatrywanej kombinacji oraz j ≠ i.

Natomiast wzór na indywidualną pojemność informacyjną jest następujący:

$$h_{ij} = \frac{r_j^2}{1+\sum_{i=1}^{n}|r_{ij}|} \qquad (16)$$

gdzie:

r_j -współczynnik korelacji zmiennej objaśniającej o numerze j ze zmienną objaśnianą,

r_{ij} - współczynnik korelacji między i-tą i j-tą zmienną objaśniającą.

Drugą bardzo znaną metodą do wyboru zmiennych do modelu jest tzw. metoda krokowa. W przeciwieństwie do metody Hellwiga, metoda krokowa wybiera zmienne na podstawie ich istotności wynikającej z testu t-studenta. Metodę krokową można przeprowadzić na dwa sposoby „wstecz" oraz „wprzód". W metodzie „wstecz" wychodzi się z modelu z wszystkim zmiennymi i po kolei się je eliminuje, natomiast w metodzie „wprzód" wychodzi się od najbardziej istotnej zmiennej i po kolei dodaje następne. W metodzie „wstecz" wylicza się w pierwszej kolejność bezwzględną wartość statystyki testowej t-studenta. Następnie oblicza się minimalną wartość statystyki z podanych statystyk testowych dla zmiennych.

Kolejnym krokiem wyznaczenie jest krytycznej wartości t* dla rozkładu t-studenta wyznaczaną ze wzoru zawartego w poprzednim podrozdziale *Istotność parametrów*. Ostatecznie porównujemy statystykę minimalną z wartością krytyczną t* i podejmowana jest odpowiednia decyzja:

Jeżeli:

$Z_{min} < t^*$

To:

Jest usuwana z modelu zmienna realizująca minimum, następnie ponownie jest szacowany model poprzez powtórzenie operacji.

Jeżeli:

$Z_{min} > t^*$

To:

Za ostateczny model przyjmuje się model z ostatnim zestawem zmiennych objaśniających[10].

[10] B. Jasiulis-Gołdyn (2014) „Ekonometria - wykład 8. 3.1 Specyfikacja i weryfikacja modelu liniowego dobór zmiennych objaśniających - część 1" s. 13

gdzie:

Z_{min} – najmniejsza statystyka testowa.

Metoda krokowa „wprzód" jest bardziej skomplikowana, jednak zaczyna się tym samym krokiem – policzeniem istotności zmiennych ze wzoru t-studenta. Następnie wybierana jest zmienna dla której wartość statystyki jest największa. Kontrolnie sprawdzany jest czy spełniona jest nierówność:

$$Z_{max} < t^*$$

gdzie:

Z_{max} – największa statystyka testowa.

Jeżeli powyższa nierówność została spełniona, oznacza to, że zmienna została źle dobra i należy powtórzyć krok lub przyjąć model skonstruowany ze zmiennych przed dodaniem ostatniej, która spełnia powyższą równość. Należy spostrzec, że metoda „wprzód" odznacza się większym stopniem komplikacji, ponieważ na każdym kroku jest do wyboru k-1 możliwości (k – liczba zmiennych), co znacząco utrudnia i opóźnia obliczenia.

W niniejszej pracy zostanie zastosowana metoda Hellwiga i metoda krokowa „wstecz", a następnie wyniki obu tych metod zostaną ze sobą skonfrontowane w celu dokonania stosownej analizy i finalnie wyboru właściwego zestawu zmiennych objaśniających.

Badanie koincydencji i współliniowości

Na tym etapie tworzenia modelu liniowego bada się nie tylko istotność parametrów ale również ich sensowność. W tym celu bada się koincydentność modelu. Mówimy, że model jest koincydentny, jeśli dla każdej zmiennej objaśniającej spełniony jest warunek:

$$sgn(r_i) = sgn(\hat{\alpha}_i)$$

gdzie:

$i = 1, 2, \ldots, k,$

$\hat{\alpha}_i$ - jest oszacowaniem parametru strukturalnego α_i występującego przy zmiennej objaśniającej X_i,

r_i – współczynnik korelacji między zmienną X_i a zmienną zależną Y.

Oznacza to, że model można nazwać koincydentnym, jeżeli zachodzi zgodność dla wszystkich znaków przy współczynnikach korelacji między zmiennymi niezależnymi

a zmienną zależną ze znakami przy oszacowanych parametrach. Jeżeli okaże się, że zachodzi niezgodność chociaż dla jednej zmiennej należy dobrać inny zestaw zmiennych objaśniających lub usunąć zmienną niekoincydentną.

Przyczynami braku koicydentności może być np. niewłaściwa postać analityczna modelu ekonometrycznego, bądź występująca współliniowość, tj. silna zależność między zmiennymi objaśniającymi[11]. Współliniowość między zmiennymi może być ścisła lub przybliżona. Tą drugą można wykryć ze wzoru na parametr VIF_k (z ang. Variance Inflation Factors[12]) zwany też inaczej CIW – czynnik inflacji wariancji. Wzór na VIF_k jest następujący:

$$VIF_k = \frac{1}{1-R_k^2} \qquad (17)$$

gdzie:

R_k^2 - współczynnik determinacji pomiędzy zmienną *k* a pozostałymi zmiennymi niezależnymi.

Parametr VIF może osiągać minimalną wartość 1.0. Przyjmuje się, że gdy przekroczy on wartość 10.0 dla danej zmiennej, to uznaje się ją za zmienną współliniową i należy ją usunąć z modelu.

Jeżeli jednak pomimo współliniowości nie chcemy usuwać zmiennej z modelu można spróbować usunąć inne zmienne z najniższą korelacją ze zmienną Y, zmienić zakres próby, rozszerzyć model o dodatkowe równanie, dokonać transformacji zmiennej odznaczającej się współliniowością lub użyć metody głównych składowych.

Jeżeli zachodzi współliniowość ścisła (dokładna), to model nie zostanie oszacowany, ponieważ wyznacznik macierz X^TX jest równy zeru, a przez to oszacowane błędy standardowe ocen parametrów powstałe z macierzy wariancji i kowariancji mają relatywnie duże wartości, co prowadzi do zaniżenia wartości statystyki t-Studenta w ocenie istotności parametru[13].

[11] M. Sobczyk (2013) „Ekonometria" s. 14 C.H. Beck
[12] G.S. Maddala (2006) „Ekonometria" s. 318-321 Wydawnictwo Naukowe PWN
[13] T. Kufel (2013) „Ekonometria Rozwiązywanie problemów z wykorzystaniem programu GRETL" s. 64 Wydawnictwo Naukowe PWN

Efekt katalizy

Jak wspomniano w podrozdziale o Metodzie Najmniejszych Kwadratów, główną miarą dopasowania modelu do danych empirycznych jest współczynnik determinacji R^2. Okazuje się jednak, że informacja, którą niesie może być zafałszowana, jeśli w modelu występują zmienne nazywane katalizatorami. Jednak zanim zacznie się szukać zmiennych odpowiedzialnych za wystąpienie niepożądanego zjawiska katalizy, należy przekształcić macierze korelacji w regularną parę korelacyjną.

$$R0 = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_k \end{bmatrix} \qquad R = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{bmatrix} \tag{18}$$

gdzie:

R0 – macierz korelacji między zmiennymi niezależnymi a zmienną zależną,

R – macierz korelacji między zmiennymi niezależnymi,

k – liczba zmiennych.

R0 oraz R tworzą parę korelacyjną, aby przekształcić ją w regularną parę korelacyjną należy dla każdego r_i (i=1,…,k) < 0 podmienić zmienną X_i poprzez przekształcenie:

$$X_i' = -X_i$$

Dzięki temu przekształceniu obie macierze korelacji ulegną zmianie i zostanie otrzymana regularna para korelacyjna. Działając na otrzymanych macierzach można rozpocząć szukanie katalizatorów. Zmienna X_j jest nazywana katalizatorem z pary zmiennych (X_i, X_j), gdzie i<j, gdy spełniony jest jeden z poniższych warunków:

$$r_{ij} < 0 \qquad \text{lub} \qquad r_{ij} > \frac{r_i}{r_j}$$

Oznacza to, że w modelu występuje co najmniej jeden katalizator, gdy w macierzy korelacji R znajduje się co najmniej jedna ujemna wartość. Po wykryciu katalizatorów należy sprawdzić natężenie efektu katalizy – może się okazać, że zmienna będąca katalizatorem nieznacznie zafałszowuje współczynnik determinacji R^2.

Natężenie efektu katalizy liczy się z danego wzoru:

$$\eta = \overline{R^2} - H \tag{19}$$

gdzie:

$\overline{R^2}$ – skorygowany współczynnik determinacji,

H – integralna pojemność informacyjna według Hellwiga.

Po policzeniu natężenia katalizy należy policzyć względne natężenie katalizy, które daje obiektywny pogląd na moc katalizatorów:

$$W_\eta = \frac{\eta}{R^2} \cdot 100\% \tag{20}$$

Dopiero względne natężenie katalizy daje racjonalny pogląd na zafałszowanie współczynnika determinacji. Jeżeli względne natężenie efektu katalizy jest bliskie 0, można uznać, że współczynnik determinacji został trafnie określony, natomiast jeżeli względne natężenie katalizy jest istotnie większe należy usunąć zmienne będące katalizatorami z modelu.

Stabilność parametrów

Bardzo ważną cechą w modelu ekonometrycznym jest stabilność parametrów. Jest to warunek konieczny do tworzenia prognoz oraz do analizy strukturalnej. W celu zbadania stabilności parametrów wykorzystuje się Test Chowa[14]. Polega on na rozdarciu danych w pewnym punkcie zwanym punktem załamania strukturalnego τ[15]. Dla szeregów czasowych jest on momentem w czasie, po którym nastąpiła zmiana strukturalna. Istnieje kilka sposobów dzielenia zmiennych nad podgrupy. Można podzielić je na pół według ilości lub na wybrane grupy czy też według zmiennej binarnej. W tej pracy Test Chowa zostanie przeprowadzony dla dwóch równych podgrup podzielonych według ilości. Podzielnie zmiennych wygląda następująco:

[14] R. Ramanathan (1995) „Introductory Econometrics with Applications" s. 360
[15] Davidson, MacKinnon (2004) „Econometric Theory and Methods" s.145-146

model liniowy:

$$y_t = \alpha_0 + \alpha_1 x_{1t} + \ldots + \alpha_k x_{kt} + \varepsilon_t, \quad t=1,2,\ldots,n$$

dzielimy na:

$$y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \varepsilon_t, \quad t=1,2,\ldots,n_1$$

$$y_t = \gamma_0 + \gamma_1 x_{1t} + \ldots + \gamma_k x_{kt} + \varepsilon_t, \quad t=n_1, n_1+1,\ldots,n \qquad (21)$$

gdzie:

α_i – parametr przy zmiennej x_i,

ε_t – składnik losowy.

Po tak podzielonym modelu następuje przeprowadzenie Testu Chowa o następujących hipotezach:

$$H_0: \alpha_i = \beta_i = \gamma_i$$

$$H_1: \alpha_i \neq \beta_i \neq \gamma_i$$

Hipoteza zerowa mówi o stabilności parametrów, natomiast hipoteza alternatywna mówi o ich niestabilności. Pożądanym jest uzyskać w testach *p-value* jak najwyższe (tj. co najmniej powyżej poziomu istotności, w tym wypadku 0.05), aby przyjąć hipotezę zerową i odrzucić alternatywną.

Statystyka testowa wygląda następująco:

$$F = \frac{SSE - SSE_1 - \frac{SSE_2}{k+1}}{\frac{SSE_1 + SSE_2}{n - 2(k+1)}} \qquad (22)$$

gdzie:

SSE – suma kwadratów reszt w całej próbie,

SSE_1 – suma kwadratów reszt w 1 próbie,

SSE_2 – suma kwadratów reszt w 2 próbie,

n – liczba obserwacji,

k – liczba zmiennych[16].

Statystyka F ma rozkład F-snedecora z k+1 oraz n-2(k+1) stopniami swobody.

[16] J. Mućk „Ekonometria. Prognozowanie ekonometryczne, ocena stabilności oszacowań parametrów strukturalnych" s. 10

Losowość próby i liniowość modelu

W wielu zagadnieniach wnioskowania statystycznego istotnym założeniem jest losowość próby, która może rzutować na liniowość modelu ekonometrycznego. Prostym testem do weryfikacji tej własności jest nieparametryczny test zwany Testem Serii. W teście badane są serie, czyli ciągi kolejnych reszt o tym samym znaku. Założenia testu są następujące:

H_0: Reszty charakteryzują się losowością

H_1: Reszty nie wykazują losowości

W celu zweryfikowania prawdziwości hipotezy H_0 o losowych resztach należy skupić się na ciągu reszt i wyznaczyć ich serie. Serie to nic innego, jak serie dodatnich i ujemnych reszt.

Liczbę serii oznaczamy jako S, aby zbadać, czy model odznacza się losowością (prawdziwość hipotezy H_0) S musi spełniać daną nierówność:

$$S_1^* < S < S_2^*$$

S_1^* oraz S_2^* odczytuje się z tablic z rozkładu liczby serii dla $\frac{1}{2}\alpha$ oraz dla 1-$\frac{1}{2}\alpha$ (gdzie α to poziom istotności). Należy zauważyć, że tablice zawierają wartości do 20, więc są one przeznaczone jedynie dla testu przeprowadzonego na małej próbie. Dla dużych prób stosuje się inny wzór, którego statystyka U dana jest wzorem:

$$U = \frac{R - E(R)}{\sqrt{D^2(R)}} \qquad (23)$$

gdzie:

R – liczba serii,

E(R) – wartość oczekiwana R,

$D^2(R)$ – wariancja R.

Wzory na wartość oczekiwaną i odchylenie standardowe (pierwiastek z wariancji) są następujące:

$$E(R) = \frac{2n_1 n_2}{n} + 1 \quad (24)$$

$$\sqrt{D^2(R)} = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} \quad (25)$$

gdzie:

n_1 – liczba serii dodatnich,

n_2 – liczba serii ujemnych[17].

Kolejnym testem służącym do oceny poprawnego dopasowania modelu jest Test Ramsey'a RESET. Ma on za zadanie sprawdzić, czy liniowa postać modelu jest najlepszym wyborem dla danego modelu względem funkcji kwadratowej oraz sześciennej.

Nazwa testu RESET pochodzi z j. angielskiego *regression specification error test.* Test Ramsey'a powstał w 1968 roku i w przeciwieństwie do Testu Serii, który również w sposób pośredni bada liniowość modelu, jest testem parametrycznym. Test ten polega na dodawaniu do modelu kolejnych potęg $\hat{y}$ jako zmienne objaśniające i sprawdzaniu współczynnika determinacji R^2. Współczynnik ten powinien się zwiększyć i wykonywany jest test na jego istotność (test F). Jeżeli wzrost współczynnika okaże się istotny oznacza to złe dobranie modelu, co skutkuje zmianą jego struktury z liniowej na nieliniową. Poniższy przykład zobrazuje w jaki sposób dokładane są do modelu kolejne potęgi $\hat{y}$:

estymowany model:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \ldots + \hat{\alpha}_k x_k + \hat{\varepsilon}_t \quad (26)$$

model po dodaniu kolejnych potęg $\hat{y}$:

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \ldots + \hat{\alpha}_k x_k + \hat{\alpha}_{k+1} \hat{y}^k + \hat{\varepsilon}_t$$

gdzie:

$\hat{\alpha}_i$ – oszacowany parametr dla zmiennej x_i (i=1, 2, …, k),

[17] L. Kowalski „Testy losowości" s. 4-14

k – liczba zmiennych,

$\hat{\varepsilon}_t$ – oszacowany składnik losowy.

Badanie heteroskedastyczności

Jednym z podstawowych założeń modelu stworzonego metodą MNK jest równość wariancji, czyli tzw. warunek homoskedastyczności. Jest to warunek konieczny do kontynuowania badania przy modelach liniowych. Homoskedastyczność oznacza, że zaburzenia losowe są jednakowo rozproszone wokół zerowej wartości oczekiwanej. Jeśli wariancje nie byłyby jednakowe, to sytuację taką nazywa się heteroskedastycznością. Poniżej przedstawione są macierze kowariancji dwóch modeli: jednego odznaczającego się homoskedastycznością, a drugiego heteroskedastycznością:

Rysunek 1: Homoskedastyczność - macierz kowariancji

$$D^2(\epsilon) = \sigma^2 \mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Rysunek 2: Heteroskedastyczność - macierz kowariancji

$$D^2(\epsilon) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Źródło: Z. Waśko (2018) „Ekonometria I, Heteroskedastyczność" s.5, http://web.sgh.waw.pl/~mrubas/Econometrics/pdf/EI_T6PL.pdf [dostęp 08.05.2020r.]

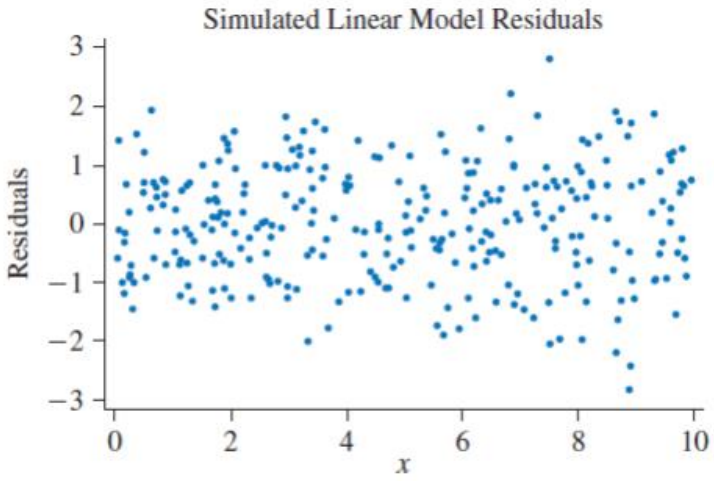
homoskedastyczność heteroskedastyczność

Jak widać powyżej w przypadku modelu z nierówną wariancją nie spełniony jest jedno z założeń Gaussa-Markowa mówiące, że:

$$var\ (\varepsilon_i) = \sigma^2 \qquad (27)$$

Dodatkowo poniższe okresy obrazują jak wpływa heteroskedastyczność na położenie reszt w modelu:

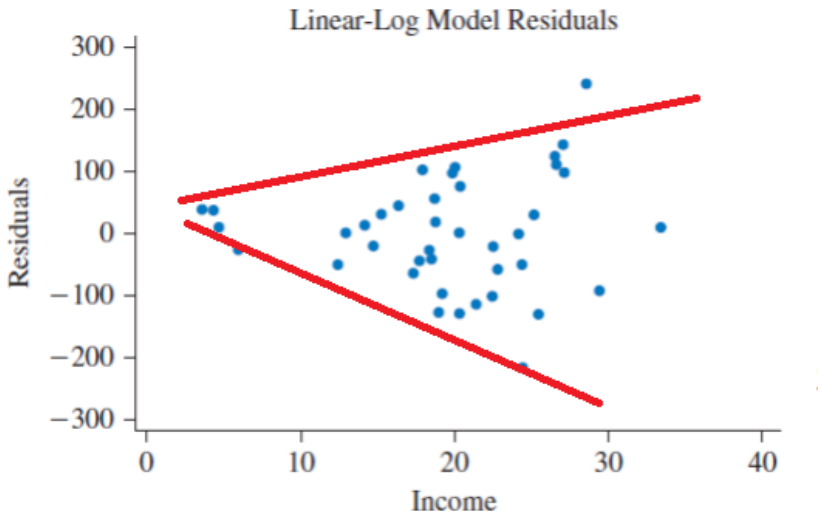
Rysunek 3: Homoskedastyczność - wykres reszt



Źródło: Z. Waśko (2018) „Ekonometria I, Heteroskedastyczność" s.9,
http://web.sgh.waw.pl/~mrubas/Econometrics/pdf/EI_T6PL.pdf [dostęp 08.05.2020r.]

Homoskedastyczność

Rysunek 4: Heteroskedastyczność - wykres reszt



Źródło: Z. Waśko (2018) „Ekonometria I, Heteroskedastyczność" s.9,
http://web.sgh.waw.pl/~mrubas/Econometrics/pdf/EI_T6PL.pdf [dostęp 08.05.2020r.]

Heteroskedastyczność

W wykresie z heteroskedastycznością od razu widać, że reszty są najpierw skupione a potem się rozszerzają według czerwonych linii tworząc niczym lejek – to klasyczna

oznaka występowania heteroskedastyczności w modelu. Na wykresie z homoskedastycznością widać, że reszty są rozproszone całkowicie losowo.

Istnieją dwa podstawowe testy statystyczne badające równość wariancji w modelu – test White'a oraz Breuscha-Pagana. Oba te testy mają identyczne hipotezy oraz w obu pożądane jest otrzymanie *p-value* wyższym od poziomu istotności. Ich hipotezy są następujące:

H_0: homoskedastyczność

H_1: heteroskedastyczność

Test Breuscha-Pagana sprawdza, czy wariancja błędów z regresji jest zależna od wartości zmiennych niezależnych. Jeśli odpowiedź na to pytanie jest pozytywna, to heteroskedastyczność jest obecna. Test ten polega na wykryciu heteroskedastyczności dla konkretnej zmiennej niezależnej, a jego statystka testowa znajduje się w rozkładzie χ^2 o k stopniach swobody, gdzie k to liczba zmiennych[18].

Test White'a to szczególny przypadek testu Breuscha-Pagana, który nie bada od której zmiennej objaśniającej zależy wariancja błędów jedynie bada czy sama nierówność wariancji występuje w modelu. Test ten polega na stworzeniu modelu zależnego od kwadratów i iloczynów zmiennych objaśniających.

Badanie autokorelacji

Badanie autokorelacji składnika losowego jest kolejnym etapem tworzenia modelu regresji liniowej. W ekonometrii autokorelacją nazywamy skorelowanie zmiennej z tą samą zmienną z innego okresu. Niestety w dużych modelach jest to częste zjawisko i wręcz niemożliwe do ominięcia[19]. W przypadku występowania zjawiska autokorelacji składnika losowego macierz wariancji-kowariancji nie jest macierzą diagonalną. Dzieje się tak, ponieważ składniki losowe dla obserwacji pochodzących z różnych okresów nie są niezależne - są skorelowane. Jednym z testów statystycznych stosowanych do wykrycia autokorelacji pierwszego rzędu jest tzw. Test Durbina-Watsona. Istnieje również Test mnożnika Lagrange'a (LM) dla autokorelacji wyższych rzędów. Polega on na zbudowaniu modelu uwzględniającego element losowy opóźniony oraz

[18] S. Cichocki, N. Nahrebecka „Testy diagnostyczne" s. 9-20
[19] P. Strawiński „Notatki do ćwiczeń z ekonometrii" s. 88

przeprowadzenie Testu Walda dla tych zmiennych. Jednak ze względu na charakter danych (dane panelowe – nie szeregi czasowe) test na wykrywanie autokorelacji wyższych rzędów nie jest konieczny do przeprowadzenia, dlatego skupiono się na teście Durbina-Watsona. Jednak aby zrozumieć założenia i hipotezy postawione w tym teście, należy zwrócić uwagę co się dzieje, gdy w modelu pojawia się autokorelacja.

Pierwszym założeniem jest, że składniki losowe ε_t związane są zależnością:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t \tag{28}$$

gdzie:
ρ – nieznany parametr zwany współczynnikiem autokorelacji, $|\rho| < 1$,
η_t – zmienna losowa spełniająca następujące warunki:

- $E(\eta) = 0$,
- $D^2(\eta) = \sigma_0^2 I$, gdzie: $\sigma_0^2 I < +\infty$.

Wektor η jest liniowym wektorem o elementach η_t. Dzięki temu założeniu można wykazać, że:

$$\sigma^2 = \frac{1}{1-\rho^2}\sigma_0^2$$

oraz

$$D^2(\varepsilon) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \dots & 1 \end{bmatrix} \tag{29}$$

W tym wypadku nieobciążonym estymatorem ρ jest:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{(\sum_{t=2}^{n} e_t^2)^2 (\sum_{t=2}^{n} e_{t-1}^2)^2} \tag{30}$$

Aby sprawdzić czy składniki losowe rozpatrywanego modelu pochodzą z procesu autokorelacji pierwszego rzędu należy sprawdzić hipotezy:

H_0: $\rho = 0$ (brak skorelowania składnika losowego)

H_1: $\sim H_0$

Do weryfikacji tych hipotez stosuje się właśnie Test Durbina-Watsona. Założenia tego testu są następujące:

1. Model ekonometryczny posiada wyraz wolny;
2. Składnik losowy ma rozkład normalny;
3. W modelu nie wstępuje opóźniona zmienna objaśniana jako zmienna objaśniająca (np. $y_t = \alpha_0 + \alpha_1 x_1 + \alpha_2 y_{t-1} + \varepsilon_t$)

Statystyka testowa $d \in [0,4]$ i wygląda następująco:

$$d = \frac{\sum_{t=1}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad (31)$$

Można zauważyć, że $d \approx 2(1-\hat{\rho})$. Stąd wiadomo, że jeżeli $\hat{\rho} = 0$, to $d = 2$. Z tablic dla testu Durbina-Watsona odczytuje się wartości krytyczny d_L oraz d_U, których rozkład zależy od liczby obserwacji n i liczby zmiennych k oraz od poziomu istotności.

Kryteria podejmowania decyzji o przyjęciu lub odrzuceniu hipotezy zerowej zależą od rozpatrywanych hipotez. Możliwe są dwa warianty:

Hipotezy:		Hipotezy:	
	H_0: $\rho = 0$		H_0: $\rho = 0$
	H_1: $\rho > 0$		H_1: $\rho < 0$
Kryteria:		Kryteria:	
$d \leq d_L$	odrzucamy H_0	$d \geq 4-d_L$	odrzucamy H_0
$d \geq d_U$	nie podejmujemy decyzji	$4-d_U<d<4-d_L$	nie podejmujemy decyzji
$d_L<d<d_U$	nie ma podstaw do odrzucenia H_0	$d \leq 4-d_U$	nie ma podstaw do odrzucenia H_0

Przyczyny występowania składnika autokorelacji w modelu są różne. Mogą to być przykładowo: natura pewnych procesów gospodarczych (skutki pewnych procesów gospodarczych, skutki pewnych zdarzeń rozciągają się na wiele okresów), niepoprawna postać analityczna modelu oraz niepełny zestaw zmiennych objaśniających.

Prognoza

Głównym celem tej pracy jest zbudowanie modelu prognostycznego, oznacza to, że model nie tylko ma spełniać szereg testów i sprawdzić wpływ oraz istotność zmiennych objaśniających na zmienną objaśnianą. Przeprowadzenie powyższych testów jest jednak niezbędne, ponieważ, aby model jednorównaniowy mógł być modelem prognostycznym musi spełnić szereg warunków, a są to m.in.:

- model musi mieć rozkład normlany reszt,
- postać analityczna modelu jest zgodna,
- zachodzi stabilność oraz istotność parametrów,
- brak autokorelacji oraz heteroskedastyczności,
- reszty muszą cechować się losowością,
- brak współliniowości między zmiennymi,
- spełnienie wszystkich założeń Gaussa-Markowa.

Po spełnieniu wszystkich warunków można przejść do prognozowania. Wyznacza się podział na prognozę punktową oraz przedziałową, jednak ta druga nie zostanie omówiona, ponieważ nie zostanie użyta w tej pracy do prognozowania wartości Y.

Następnie liczony jest błąd prognozy. Pożądanym jest, aby był on jak najmniejszy, ponieważ świadczy to o dobrze zbudowanym modelu. Prognozę punktową można podzielić na prognozę *ex ante* oraz *ex post*. Różnią się one tym, że pierwszą przeprowadza się w oparciu o szacowanie przyszłych wartości zmiennej Y, natomiast drugą na podstawie już znanych nam wartości zmiennej Y, ale nie wziętych pod uwagę do liczenia modelu. Przyjmuje się, że z dostępnych danych odrzuca się około 10% obserwacji, które będą właśnie tą próbą kontrolną w prognozie *ex-post*.

Ocenę prognozy *ex-ante* rozumie się jako próbę predykcji w momencie, gdy nie znana jest prawdziwa wartość zmiennej objaśnianej w przewidywanym okresie. Błąd tej prognozy wyrażany jest jako:

$$e_{\tau}^{P} = y_{\tau} - \hat{y}_{\tau}^{P} \qquad (32)$$

gdzie:

τ – okres prognozy $\tau > n$, dla $\tau = 1,2,\ldots,s$.

W momencie konstruowania prognozy wartość rzeczywista y_τ zmiennej objaśnianej jest zazwyczaj nieznana. W celu wyznaczenia prognozy najpierw należy policzyć nieobciążony i zgodny estymator wariancji σ^2 składnika losowego w modelu za pomocą wzoru:

$$S^2 = \frac{e^T e}{n-(k+1)} = \frac{y^T y - a^T X^T y}{n-(k+1)} \quad (33)$$

gdzie:

e – macierz reszt,

n – liczba obserwacji,

k – liczba zmiennych,

y – macierz wartości zmiennej Y,

a – macierz parametrów,

X – macierz wartości zmiennych X.

Po wyznaczeniu S^2 należy wyznaczyć wariancję błędu prognozy jako:

$$(S_\tau^P)^2 = S^2(x_\tau^T (X^T X)^{-1} x_\tau + 1) \quad (34)$$

oraz średni błąd predykcji prognozy *ex ante*:

$$S_\tau^P = \sqrt{S^2 (x_\tau^T (X^T X)^{-1} x\tau + 1)} \quad (35)$$

Błąd ten informuje, o ile oszacowana wartość zmiennej prognozowanej średnio odchyla się od rzeczywistej wartości zmiennej prognozowanej. Natomiast średni względny błąd predykcji wyznacza się jako:

$$\left| \frac{S_\tau^P}{y_\tau^P} \right| \quad (36)$$

W prognozie *ex-post* jest dużo łatwiej w ocenie względnych i bezwzględnych błędów prognozy, ponieważ rzeczywista wartość y_τ zmiennej objaśnianej jest znana, dlatego błąd prognozy wyrażony jest wzorem:

$$y_\tau - y_\tau^P \quad , \text{dla } \tau=1,2,\ldots,s \quad (37)$$

Natomiast względny błąd prognozy jest równy:

$$\frac{y_\tau - y_\tau^P}{y_\tau} \quad (38)$$

Jednak dla całej prognozy ex post przyjęło się wyznaczać nie tyle względne błędy co błędy średnie. Wyznacza się kilka podstawowych błędów prognostycznych:

Średni błąd prognozy – pomaga on ocenić przeciążenie prognozy

$$\mathbf{ME} = \frac{1}{s}\sum_{\tau=1}^{s}(y_\tau - y_\tau^P) \tag{39}$$

Średni absolutny błąd – podaje w jednostkach bezwzględnych o ile średnio prognoza różni się od wartości rzeczywistej

$$\mathbf{MAE} = \frac{1}{s}\sum_{\tau=1}^{s}|y_\tau - y_\tau^P| \tag{40}$$

Pierwiastek błędu średniookresowego – interpretuje się go podobnie jak błąd MAE (jest jedynie bardziej czuły na wartości skrajne)

$$\mathbf{RMSE} = \sqrt{\frac{1}{s}\sum_{\tau=1}^{s}(y_\tau - y_\tau^P)} \tag{41}$$

Średni absolutny błąd procentowy – podaje o ile procent średnio prognoza różni się od wartości rzeczywistej; jest najbardziej realną wartością jeśli chodzi o błąd prognozy i najczęściej się go podaje

$$\mathbf{MAPE} = \frac{1}{s}\sum_{\tau=1}^{s}\left|\frac{y_\tau - y_\tau^P}{y_\tau}\right| \cdot 100\% \tag{42}$$

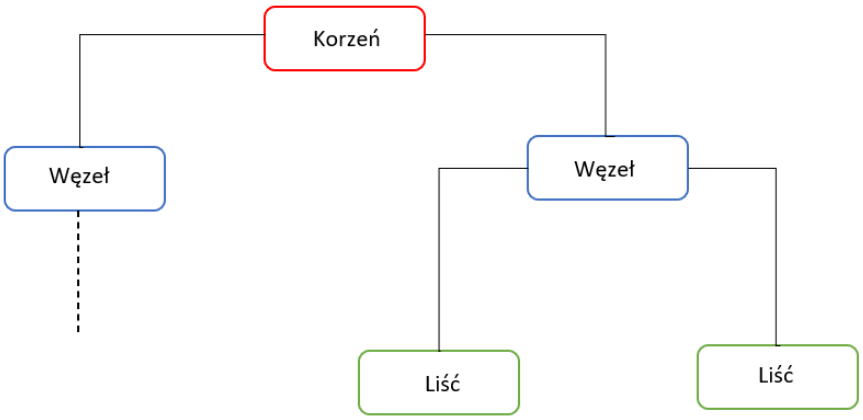
Z racji dostępności do stosunkowo dużej liczby danych (powyżej 50 obserwacji) w pracy liczy się prognozę na podstawie prognozy *ex-post* (10% obserwacji wyklucza się do badań) oraz błąd prognozy będzie rozumiany jako średni błąd procentowy MAPE.

1.4. Drzewa decyzyjne

Jedną z najczęściej używanych metod statystycznych do prognozowania wartości zmiennej objaśnianej są drzewa decyzyjne. Ich popularność wynika z łatwości interpretacji otrzymywanych wyników oraz z ich prostej i intuicyjnej budowy. Konstrukcja drzew wymaga mniej skomplikowanych metod niż tych, stosowanych w celu zbudowania modelu ekonometrycznego, co jest dużą zaletą w świecie statystyki.

Drzewa decyzyjne występują w postaci grafów skierowanych. Ich wierzchołki noszą nazwę węzłów (z ang. *nodes*), natomiast krawędzie gałęzi (z ang. *branches*). Węzły są inaczej rozwidleniami, z których wychodzą krawędzie, jednak istnieją węzły bez potomków i są to tzw. liście (z ang. *leaves*) oraz występują węzły bez rodzica, czyli tzw. korzenie (z ang. *roots*). Poniższy rysunek przedstawia klasyczny wygląd drzewa decyzyjnego:

Rysunek 5: Konstrukcja drzewa decyzyjnego



Źródło: Opracowanie własne

Wszystkie węzły w drzewie posiadają ściśle określone testy na atrybutach (zmiennych objaśniających), które determinują dany podział atrybutu na podzbiory. Jeśli dana obserwacja z rozpatrywanego atrybuty spełnia określony w węźle warunek, zostaje przypisana do lewej gałęzi wychodzącej z tego węzła i wzięta pod uwagę do dalszych rozważań w kolejnym, niższym węźle, natomiast w przeciwnej sytuacji zostaje ona przypisana do gałęzi prawej. Zdarza się, że istnieje więcej gałęzi, a co za tym idzie, podzbiorów. W ten sposób drzewa decyzyjne znajdują podzbiór obserwacji do siebie jak najbardziej podobnych, aby na ich podstawie określić szukaną wartość zmiennej objaśnianej, opisanej w liściu (węźle bez potomków). Dzięki tej technice łatwo jest zapisać algorytm zawarty w drzewie jako zbiór reguł decyzyjnych, który jest łatwy do zinterpretowania nawet przez osobę nieposiadającą zaawansowanej wiedzy z zakresu statystyki[20].

Drzewa decyzyjne można ogólnie podzielić na dwa rodzaje według zmiennej objaśnianej Y - drzewa regresyjne, gdzie Y jest zmienną ilościową (czyli można ją przedstawić na skali przedziałowej) oraz drzewa klasyfikacyjne, w których zmienna Y występuje w postaci dyskretnej (często dwuwartościowej).

Konstrukcja drzew klasyfikacyjnych i regresyjnych nie różni się znacząco, jednak ze względu na charakter wybranych danych, w badaniach empirycznych zostanie skonstruowane drzewo regresyjne tak, aby móc zachować charakter ilościowy zmiennej

[20] J. Kozak, P. Juszczuk (2016) „Algorytmy do konstruowania drzew decyzyjnych w przewidywaniu skuteczności kampanii telemarketingowej banku", Studio Informatica Pomerania nr 1 (39), s. 49-59

Y. W dalszych rozważaniach drzewa decyzyjne zostaną opisane w kontekście drzew regresyjnych.

Konstrukcja drzewa regresyjnego składa się z dwóch etapów:

1. Przedzieleniu zmiennej objaśniającej według jakiegoś punktu, by móc zakwalifikować obserwacje do jednego z dwóch podzbiorów;
2. Wybrania zmiennej objaśniającej (atrybutu), według której będą dzielone dane w węźle lub decyzji o zakończeniu podziału danych i utworzeniu liścia.

Ogólnie celem drzewa regresyjnego jest zminimalizowanie resztkowej sumy kwadratów RSS (z ang. *Residual Sum of Squares*) danej wzorem:

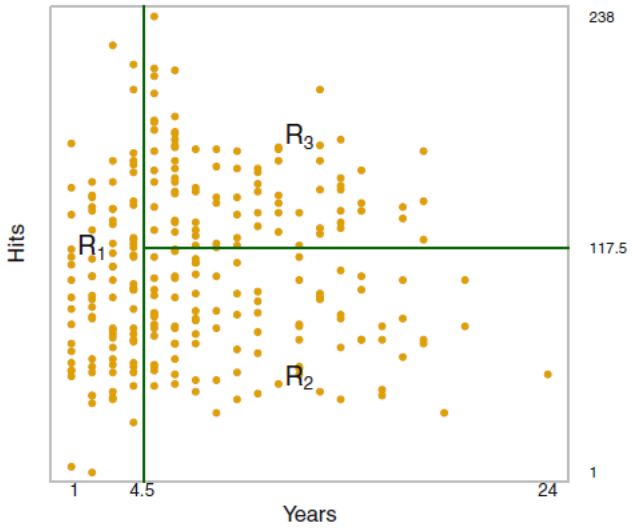
$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} (y - \hat{y}_{R_j})^2 \qquad (43)$$

gdzie:

R_j – j-ty obszar (z ang. *box*) podobnych obserwacji.

Poniższy rysunek obrazuje przykładowy podział danych na trzy obszary $\{R_1, R_2, R_3\}$:

Rysunek 6: Podział danych według obszarów w drzewie decyzyjnym



Źródło: James G., Witten D., Hastie T., Tibshirani R. (2017) " An Introduction to Statistical Learning with Applications in R" Nowy Jork: Springer, s. 305

Jest to podział danych (na rysunku są to pomarańczowe kropki) między trzy obszary, które dzielą je na podobne sobie wartości według danych kryteriów.

Jednak ze względu na dużą liczbę kombinacji jest to wręcz niemożliwe do oszacowania przy ograniczonej mocy obliczeniowej. W takim przypadku stosuje się schodzący algorytm zachłanny (z ang. *top-down greedy algorithm*) zwany z ang. *recursive binary splitting*. Jest on algorytmem schodzącym, ponieważ stosuję się go z góry do dołu, czyli od najwyższych węzłów do najniższych, a zachłanny oznacza, że wybiera najlepszą dla niego decyzję bez rozważania, jakie będzie miała ona skutki w dalszych etapach algorytmu[21]. Chcąc przeprowadzić ten algorytm należy wykonać działanie opisane wyżej w punkcie 1. Trzeba znaleźć taki punkt *s* dla danego *j*-tego atrybutu, dla którego spełnione są poniższe równania:

$$R_1(j,s) = \{X|X_j < s\} \text{ oraz } R_2(j,s) = \{X|X_j \geq s\} \tag{44}$$

$$min\,[\sum_{i:\, x_i \epsilon R_1(j,s)} (y - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \epsilon R_2(j,s)} (y - \hat{y}_{R_2})^2\,] \tag{45}$$

gdzie:

R_1 i R_2 – są to dwa podzbiory utworzone poprzez rozdzielenie obserwacji przez j-tą zmienną według punktu *s*.

W ten sposób minimalizowany jest błąd RSS w znacznie szybszy sposób, zwłaszcza w sytuacji, gdy rozpatrywany jest zbiór o dużej liczbie danych.

Po znalezieniu punktów *s* dla każdej zmiennej objaśniającej w celu konstrukcji drzewa należy spełnić punkt 2. Jest to wybór atrybutu według, którego będą dzielone dane w konkretnym węźle. Im wyżej (płycej) znajduje się węzeł z warunkiem dotyczącym zmiennej X_j, tym ta zmienna jest istotniejsza i ma większy wpływ na zmienna objaśnianą Y. Najbardziej istotne atrybuty przy konstrukcji drzew znajduje się za pomocą tzw. entropii. Dosłownie tłumacząc entropia jest średnią ilością informacji jaka przypada na pojedynczą wiadomość. Algorytm entropii pozwala zatem określić ile informacji zyskamy przy podziale obserwacji według danego atrybutu. Matematyczny wzór na entropię przedstawia się następująco:

$$E(S) = \sum_{i=1}^{j} -p_i log_2 p_i \tag{46}$$

[21] James G., Witten D., Hastie T., Tibshirani R. (2017) " An Introduction to Statistical Learning with Applications in R" Nowy Jork: Springer, s. 305-307

gdzie:

p_i – jest prawdopodobieństwem przyporządkowania obserwacji do *i*-tej gałęzi.

Dla drzew regresyjnych w większości przypadków znajdują się tylko dwie gałęzie, zatem wzór na entropię można zapisać w następujący sposób:

$$E(S) = -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus} \quad (47)$$

gdzie:

$p_{\oplus}$ - prawdopodobieństwo z jakim obserwacje spełnią warunek zawarty w węźle,

$p_{\ominus}$ - prawdopodobieństwo z jakim obserwacje nie spełnią warunku zawartego w węźle.

Należy zwrócić uwagę, że podstawa logarytmu w tym wzorze jest dowolna i nie ma większego znaczenia tak długo, jak nie będzie ona zmieniana w trakcie obliczeń. Często przyjmuje się podstawę równą 2, wówczas jednostką entropii jest bit lub używa się też logarytmów naturalnych, gdzie jednostka informacji zwana jest nitem. Pożądanym jest uzyskanie jak najniższej entropii, ponieważ w sytuacji gdy entropia jest bliska zeru, obserwacje dzielą się mniej więcej równomiernie pomiędzy dwoma gałęziami. W sytuacji, gdy obserwacje są skoncentrowane w jednym obszarze, np. wszystkie nie spełniają warunku postawionego w węźle, entropia przyjmuje wartości bliskie 1, co nie jest pożądaną cechą. Dzieje się tak, ponieważ najistotniejsze atrybuty posiadają własność, dzięki której po ich podziale zyskuje się jak najwięcej informacji o zbiorze danych. Zysk informacyjny (z ang. *Information Gain*) wylicza się z następującego wzoru:

$$IG(Y,X) = E(Y) - E(Y|X) \quad (48)$$

Oznacza to, że zysk informacyjny to różnica informacji jaką niesie ze sobą zmienna Y w danym węźle i informacji, jaką niesie ze sobą Y po podziale jej według zmiennej X. W momencie, gdy zmienna X dzieli się na dwa lub więcej podzbiorów, liczy się średnią arytmetyczną ich entropii i po tym odejmuje od entropii zmiennej Y według wyższego węzła. Pożądanym jest uzyskanie jak największego wyniku IG, a zatem jak najniższego wyniku entropii po podziale według zmiennej X. Wybierana jest do podziału zmienna o najwyższym IG. Cały proces jest powtarzany dla każdego kolejnego węzła, aż kolejny podział (rozwidlenie) nie przyniesie żadnego zysku informacyjnego lub mniejszy niż określona wcześniej wartość. Jest to sposób, dzięki któremu drzewo nie będzie rosło bez końca, co jest jednym z zalet entropii. Kolejnym prostym sposobem jest wyznaczenie przed rozpoczęciem algorytmu pożądanej głębokości drzewa (z ang. *maxdepth*), wtedy

algorytm się zatrzyma gdy dojdzie do następnego poziomu, nawet gdy kolejny węzeł przyniósłby większy zysk informacyjny.

Po zbudowaniu drzewa może się ono wydać zbyt skomplikowane, tzn. może zawierać zbyt dużo węzłów, czy zbyt dużą głębokość. Prowadzi to w niektórych przypadkach do przetrenowania zbioru danych i skutkuje za dużą wariancją, ciężką interpretacją wyników a nawet zwiększeniem RSS. W tym wypadku by uniknąć przetrenowania używa się metody zwanej przycinaniem drzewa (z ang. *pruning*). Jest kilka metod, według których można przyciąć drzewo, najprostszą i najczęściej używaną jest wybranie takiej głębokości drzewa dla której błąd RSS się zmniejsza[22], nie oznacza to jednak, że średni względny błąd predykcji MAPE również zmaleje.

[22] James G., Witten D., Hastie T., Tibshirani R. (2017) " An Introduction to Statistical Learning with Applications in R" Nowy Jork: Springer, s. 307-308

Rozdział 2 : Badania empiryczne

W pierwszej kolejności zdecydowano się na zbudowaniu jednorównaniowego liniowego modelu ekonometrycznego. Punktem wyjścia do rozpoczęcia badań empirycznych jest przygotowanie danych w taki sposób, aby mogły one zostać przyjęte do dalszych badań. Dane zawierają 52 obserwacje, tj. 52 różne projekty zrealizowane przez firmę. Ze względu na występowanie dwóch zmiennych kategorycznych: X_6 oraz X_9, usunięto po jednej kategorii z każdej z nich, tj. $X_{6.3}$ oraz $X_{9.3}$ w celu stworzenia modelu liniowego. Oznacza to, że pozostałe kategorie zmiennych odnosić się będą do zmiennych usuniętych. Przed skonstruowaniem pierwszego modelu użyto również zasady trzech sigm w celu usunięcia wartości odstających mogących zaburzyć wyniki przy tworzeniu modelu. Zważając na fakt, że wśród danych przeważają zmienne binarne oraz zmienne dyskretne (np. 1, 2, 3), zasady tej użyto jedynie na zmiennych ilościowych – Y oraz X_1. Po usunięciu obserwacji odstających otrzymano 48 obserwacji co oznacza, że aż cztery obserwacje były znacznie odstające od pozostałych pod względem liczby godzin lub pod względem długości drogi.

Następnie sprawdzono zmienność zmiennych, aby wykluczyć z modelu zmienne relatywnie stałe, które nie wniosą żadnych istotnych wartości do modelu liniowego. Zmienność poszczególnych zmiennych przedstawiona została poniżej i została wyliczona na podstawie ilorazu odchylenia standardowego i średniej arytmetycznej:

Tabela 3: Współczynnik zmienności

Zmienna	Wsp. zmienności
X_1	0.606
X_2	1.371
X_3	0.583
X_4	1.658
X_5	0.418
$X_{6.1}$	1.054
$X_{6.2}$	3.914
X_7	0.748
X_8	1.658
$X_{9.1}$	1.750
$X_{9.2}$	2.796

X_{10}	0.392
X_{11}	0.583
X_{12}	1.658
X_{13}	0.485
X_{14}	1.658
X_{15}	1.97
X_{16}	0.343

Źródło: Opracowanie własne

Z danych przedstawionych w powyższej tabeli wynika, że żadna zmienna nie charakteryzuje się współczynnikiem zmienności poniżej 10%, co oznacza, że żadna zmienna nie jest relatywnie stała i została przyjęta do dalszych badań.

Kolejnym etapem jest stworzenie pierwszego modelu liniowego za pomocą metody najmniejszych kwadratów (model bazowy), na podstawie którego będą przeprowadzone wszystkie następne operacje. Model ten prezentuje się następująco:

Rysunek 7: Model bazowy

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  113.65358  224.24773   0.507 0.616112
x1             0.37982    0.08756   4.338 0.000159 ***
x2            20.57934   24.03228   0.856 0.398841
x3           101.34749   56.69350   1.788 0.084290 .
x4           -30.75254   76.30556  -0.403 0.689891
x5            84.71267   46.53963   1.820 0.079060 .
x6.1        -215.61229   66.60933  -3.237 0.003019 **
x6.2        -120.61835   96.66843  -1.248 0.222103
x7            32.75547   53.46899   0.613 0.544911
x8            -7.79145   53.80980  -0.145 0.885873
x9.1          -4.45183   62.57593  -0.071 0.943772
x9.2         -66.94988   72.79097  -0.920 0.365288
x10          -95.18394   33.50650  -2.841 0.008148 **
x11          -33.47753   62.34438  -0.537 0.595379
x12           35.79510   72.56266   0.493 0.625516
x13          -96.80522   67.76703  -1.429 0.163831
x14          -30.19285   72.61448  -0.416 0.680620
x15           -3.82051   70.28059  -0.054 0.957020
x16          129.62918   59.55471   2.177 0.037794 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.8 on 29 degrees of freedom
Multiple R-squared:  0.8807,    Adjusted R-squared:  0.8066
F-statistic: 11.89 on 18 and 29 DF,  p-value: 6.311e-09
```

Źródło: Opracowanie własne

Powstały model ma postać:

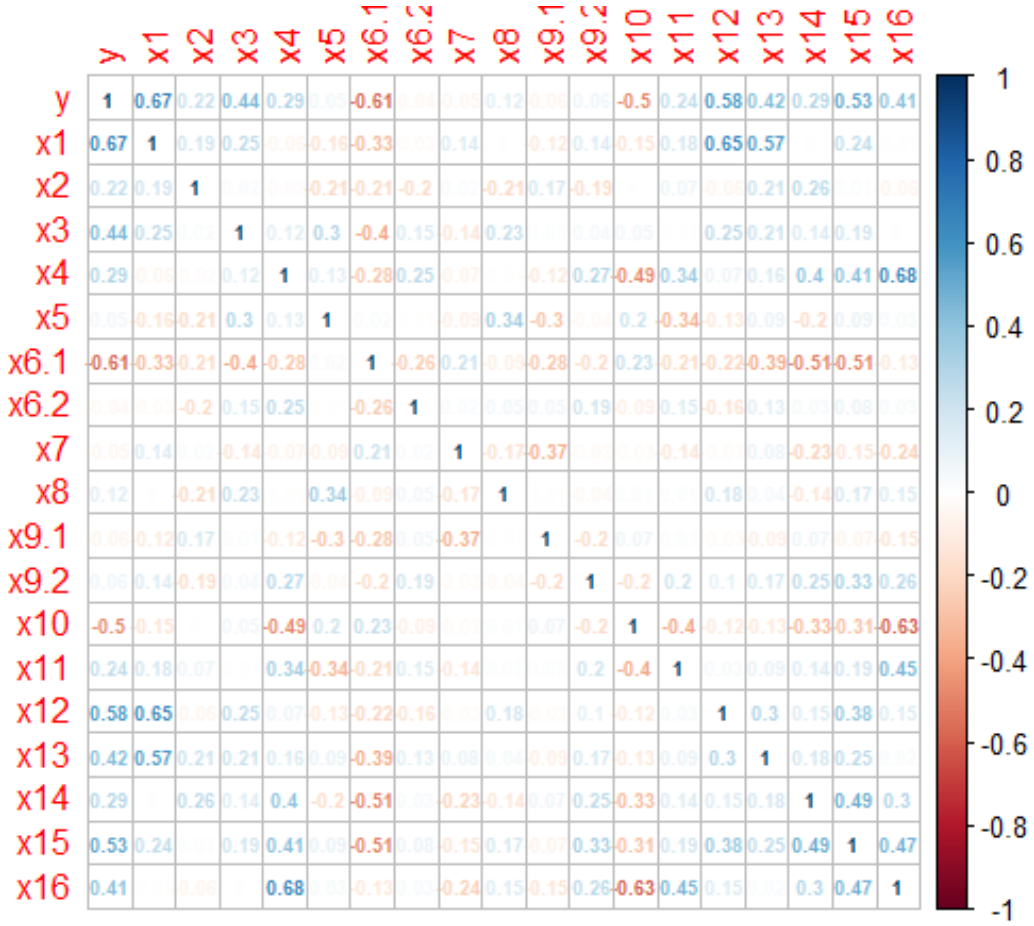
$$\mathbf{y} = 113.65 + 0.38\mathbf{X_1} + 20.58\mathbf{X_2} + 101.35\mathbf{X_3} - 30.75\mathbf{X_4} + 84.71\mathbf{X_5} - 215.61\mathbf{X_{6.1}} - 120.62\mathbf{X_{6.2}} + 32.76\mathbf{X_7} - 7.79\mathbf{X_8} - 4.45\mathbf{X_{9.1}} - 66.95\mathbf{X_{9.2}} - 95.18\mathbf{X_{10}} - 33.48\mathbf{X_{11}} + 35.795\mathbf{X_{13}} - 30.19\mathbf{X_{14}} - 3.82\mathbf{X_{15}} + 129.63\mathbf{X_{16}}$$

Z rysunku 7. można również wyczytać, że skorygowane R^2 wynosi 80.66%, co jest wysoką wartością i satysfakcjonującym wynikiem. Model wskazał, że najistotniejsze zmienne według testu t-studenta to X_1, X_{10}, $X_{6.1}$, X_{16}, X_5 oraz X_3.

Sprawdzono również, czy w utworzonym modelu bazowym występuje normalny rozkład reszt. Przeprowadzono test Shapiro-Wilka (podrozdział 1.3.) i jego wartość *p-value* wyniosła 0.46, co oznacza, że nie ma podstaw do odrzucenia hipotezy H_0 i potwierdza to normalność rozkładu.

Dalszą częścią jest usunięcie reszt odstających, zgodnie z już wcześniej zastosowaną zasadą trzech sigm. Spośród 48 obserwacji usunięto tylko jedną, która posiadała reszty odstające. Następnie sprawdzono korelacje między zmiennymi za pomocą wzoru na współczynnik korelacji Pearsona opisany w podrozdziale 1.3. Macierz korelacji wygląda następująco:

Rysunek 8: Macierz korelacji



Źródło: Opracowanie własne

Niestety powyższa macierz wskazuje na brak występowania wysokich korelacji między zmienną objaśnianą a zmiennymi objaśniającymi. Najwyższy poziom korelacji posiadają odpowiednio zmienne: X_1 na poziomie 67%, $X_{6.1}$ na poziomie -61% oraz X_{12} na poziomie 58%, co nie jest w pełni zadowalającym wynikiem.

Biorąc pod uwagę istotność zmiennych oraz macierz korelacji zdecydowano się na usunięcie zmiennych: X_7 oraz $X_{9.1}$, które posiadają niską korelację ze zmienną Y oraz są nieistotne według testu t-studenta.

W kolejnym kroku zostaną przeprowadzone pozostałe testy statystyczne opisane w poprzednim rozdziale w celu znalezienia możliwie najbardziej dopasowanego modelu. Przed dokonaniem tej operacji usunięto losowe 10% obserwacji, które zostaną użyte jako grupa kontrolna w prognozie *ex post*. Są to cztery obserwacje, ponieważ 47*10% = 4.7 i zdecydowano się zaokrąglić liczbę obserwacji w dół, by móc wykorzystać jak najwięcej

obserwacji do utworzenia modelu. Po zastosowaniu tego zabiegu skonstruowano model nr 2:

Rysunek 9: Model po usunięciu X_7 oraz $X_{9.2}$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  216.4437   139.5329   1.551 0.131340
x1             0.2479     0.0751   3.301 0.002491 **
x2            28.9156    18.9439   1.526 0.137391
x3           113.8072    44.7564   2.543 0.016393 *
x4           -65.6996    59.0667  -1.112 0.274846
x5            68.5844    33.9647   2.019 0.052470 .
x6.1        -183.1009    47.2622  -3.874 0.000539 ***
x6.2         -66.6441    77.1724  -0.864 0.394674
x8           -55.0071    43.6399  -1.260 0.217213
x9.2        -127.3610    58.8877  -2.163 0.038656 *
x10         -101.8193    25.6964  -3.962 0.000423 ***
x11          -13.7026    48.2759  -0.284 0.778484
x12          114.1928    60.2646   1.895 0.067785 .
x13          -43.8076    54.7804  -0.800 0.430174
x14          -53.3314    54.4652  -0.979 0.335320
x15           65.9828    57.0348   1.157 0.256448
x16          102.4630    44.6131   2.297 0.028793 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 107 on 30 degrees of freedom
Multiple R-squared:  0.8985,    Adjusted R-squared:  0.8444
F-statistic:  16.6 on 16 and 30 DF,  p-value: 1.061e-10
```

Źródło: Opracowanie własne

Na rysunku 9. widać, że skorygowane R^2 w drugim modelu wzrosło do 84.4%, co wskazuje, że zmienne X_7 oraz $X_{9.1}$ zostały słusznie usunięte. Drugi model również odznacza się większą istotnością zmiennych, jako że teraz aż 8 zmiennych zostało oznaczonych jako względnie istotnych a 3 jako bardzo istotne: - X_1, $X_{6.1}$ oraz X_{10}.

Jednak mimo lepszego dopasowania modelu do obserwacji nadal wiele zmiennych nie przeszło testu t-studenta na istotność. Po ponownym wykonaniu testu na normalność rozkładu otrzymano *p-value* na poziomie 0.341, co świadczy o nieodrzuceniu hipotezy H_0 mówiącej o normalności rozkładu reszt. Sprawdzono również pozostałe dwa wyznaczniki poziomu dopasowania modelu: AIC oraz BIC, które zostały opisane w podrozdziale 1.3:

Tabela 4: Model 2 - AIC oraz BIC

AIC	BIC
587.5283	620.8309

Źródło: Opracowanie własne

Jak wynika z tabeli 4. wskaźniki te nie są na zadowalającym poziomie, jako że BIC jest większe od AIC, a pożądana jest sytuacja odwrotna.

Następnie użyto metody Hellwiga (podrozdział 1.3) do określenia najlepszego zestawu doboru zmiennych. Obliczono integralną pojemność informacyjną i przeprowadzono według niej odpowiedni dobór zmiennych. Wynikiem tej funkcji jest największa integralna pojemność wynosząca 0.791 dla zestawu zmiennych: X_1, X_3, $X_{6.1}$, X_{10}, X_{12}, X_{16}. Model nr 3 składający się z tego zestawu prezentuje się następująco:

Rysunek 10: Model wg Hellwiga

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  221.90811  149.67345   1.483  0.14688
x1             0.26261    0.06461   4.064  0.00025 ***
x3           146.82967   49.55789   2.963  0.00537 **
x6.1        -136.85259   45.09455  -3.035  0.00445 **
x10          -87.30938   29.24718  -2.985  0.00507 **
x12           76.61516   57.70523   1.328  0.19264
x16           83.16584   40.26424   2.066  0.04613 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 124.9 on 36 degrees of freedom
Multiple R-squared:  0.8304,    Adjusted R-squared:  0.8021
F-statistic: 29.38 on 6 and 36 DF,  p-value: 1.803e-12
```

Źródło: Opracowanie własne

Niestety model skonstruowany metodą Hellwiga posiada niższe skorygowane R^2 w stosunku do modelu nr 2, tj. równe 80.21%, jednak pomimo tego jest to bardzo wysoka wartość dla współczynnika determinacji. W modelu tym, aż 5 na 6 zmiennych odznacza się wysoką istotnością, jedynie X_{12} posiada *p-value* dla testu t-studenta na poziomie 0.19, co jest znacząco większe od przyjętego w badaniu poziomu istotności. *P-value* dla testu Shapiro-Wilka wyniosło dla modelu 0.08636, co jest niewiele wyższe od poziomu istotności, ale przyjmuje się, że rozkład normalny reszt został zachowany. Wskaźniki AIC oraz BIC są kolejno na poziomie 545.5239 oraz 559.6135, co jest lepszym wynikiem niż w poprzednim modelu, mimo że wskaźnik AIC jest nieznacznie niższy.

Dla porównania utworzono również model metodą krokową wstecz, a jej wynikiem jest następujący dobór zmiennych: X_1, X_2, X_3, X_4, X_5, $X_{6.1}$, $X_{6.2}$, X_8, $X_{9.2}$, X_{10}, X_{12}, X_{14}, X_{16}. Jak widać, znacznie większa liczba zmiennych została przyjęta w modelu krokowym, przynajmniej w porównaniu do modelu otrzymanego metodą Hellwiga, co jednak może świadczyć o lepszym dopasowaniu. Tak skonstruowany model nr 4 przedstawia rysunek 11.:

Rysunek 11: Model otrzymany metodą krokową

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   93.3058   137.5234   0.678 0.502851
x1             0.1837     0.0661   2.779 0.009463 **
x2            33.6375    18.9111   1.779 0.085773 .
x3           123.4926    45.4441   2.717 0.010981 *
x4           -92.4401    57.6338  -1.604 0.119569
x5            84.8351    29.8912   2.838 0.008201 **
x6.1        -194.6294    47.2525  -4.119 0.000289 ***
x6.2        -126.2796    87.8345  -1.438 0.161224
x8           -77.4388    43.7712  -1.769 0.087382 .
x9.2        -154.4472    60.7621  -2.542 0.016631 *
x10          -94.7641    25.3641  -3.736 0.000815 ***
x12          157.4663    56.6587   2.779 0.009463 **
x14          -73.1126    53.4543  -1.368 0.181894
x16          154.1524    42.3694   3.638 0.001058 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 103.7 on 29 degrees of freedom
Multiple R-squared:  0.9057,    Adjusted R-squared:  0.8634
F-statistic: 21.42 on 13 and 29 DF,  p-value: 2.057e-11
```

Źródło: Opracowanie własne

Dla modelu nr 4 skorygowany współczynnik determinacji zwiększył się do 86.34%, co jest dobrym wynikiem, natomiast istotność zmiennych zmniejszyła się, ponieważ aż trzy zmienne uznano za nieistotne: X_4, $X_{6.2}$ oraz X_{14}. Dla testu Shapiro-Wilka otrzymano *p-value* na poziomie 0.5175, co jest dużo lepszym wynikiem niż w poprzednim modelu. Wskaźniki AIC oraz BIC są równe kolejno 534.294 oraz 560.712, więc pozostały niemal niezmienne w stosunku do modelu otrzymanego metodą Hellwiga. Poniższa tabela konfrontuje oba te modele w celu wybrania jednego z nich:

Tabela 5: Porównanie modeli

Czynniki	Model wg Hellwiga	Model z metody krokowej
Zmienne	X_1, X_3, $X_{6.1}$, X_{10}, X_{12}, X_{16}	X_1, X_2, X_3, X_4, X_5, $X_{6.1}$, $X_{6.2}$, X_8, $X_{9.2}$, X_{10}, X_{12}, X_{14}, X_{16}
p-value dla testu Shapiro-Wilka	0.08636	0.5175
R^2	83.04%	90.57%
Skorygowane R^2	80.21%	86.34%
AIC	545.5239	534.2941
BIC	559.6135	560.7121

Źródło: Opracowanie własne

Podjęto decyzję o kontynuacji badań na modelu otrzymanym metodą krokową, ze względu na wyższy współczynnik determinacji, lepszy rozkład reszt (normalny), większą liczbę zmiennych oraz ze względu na większą celowość z punktu widzenia dziedziny drogownictwa.

Po wybraniu modelu należy zbadać jego koincydentność, co zostało opisane w podrozdziale 1.3. Poniższa tabela 9. przedstawia *sgn()* otrzymanych wartości z macierzy korelacji oraz parametrów oszacowanych dla poszczególnych zmiennych:

Tabela 6: Badanie koincydencji

Zmienna	X_1	X_2	X_3	X_4	X_5	$X_{6.1}$	$X_{6.2}$	X_8	$X_{9.2}$	X_{10}	X_{12}	X_{14}	X_{16}
Parametr	+	+	+	-	+	-	-	-	-	-	+	-	+
Korelacja	+	+	+	+	+	-	-	+	+	-	+	+	+

Źródło: Opracowanie własne

W tabeli podkreślono zmienne, które nie są koincydentne, jednak przed podjęciem decyzji o ich usunięciu postanowiono zbadać występowanie katalizatorów oraz natężenie efektu katalizy według zasad opisanych w podrozdziale 1.3. Po przeprowadzeniu stosownych obliczeń otrzymano następujące katalizatory: X_2, X_3, X_4, X_5, $X_{6.2}$, X_8, $X_{9.2}$, X_{14}, X_{16}, przy natężenie efektu katalizy wynoszącego 24.47% i względnym natężeniu efektu katalizy równym 28.7%, co jest niedopuszczalnym wynikiem.

Na podstawie braku koincydentności oraz efektu katalizy zdecydowano się na usunięcie zmiennych: X_4, X_5, $X_{6.2}$, X_8, $X_{9.2}$ oraz X_{14}. Oznacza to, że w modelu pozostawiono zmienne X_1, X_2, X_3, $X_{6.1}$, X_{10}, X_{12} oraz X_{16}. Nie usunięto wszystkich zmiennych będących katalizatorami, dlatego podjęto próbę policzenia względnego i bezwzględnego natężenia katalizy po raz kolejny. Wynik względnego natężenia katalizy jest równy 4.77% a bezwzględnego 3.95%, co jest wynikiem akceptowalnym i na jego podstawie postanowiono nie usuwać pozostałych katalizatorów.

Po przeprowadzonych powyższych zabiegów skonstruowano model nr 5 i jest on następujący:

Rysunek 12: Finalny model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  189.89530  148.33730   1.280  0.20891
x1             0.23458    0.06596   3.556  0.00110 **
x2            29.39616   19.04554   1.543  0.13171
x3           150.94352   48.70599   3.099  0.00382 **
x6.1        -120.06984   45.56924  -2.635  0.01245 *
x10          -87.46448   28.70152  -3.047  0.00437 **
x12          101.46999   58.87340   1.724  0.09362 .
x16           87.71444   39.62255   2.214  0.03346 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 122.5 on 35 degrees of freedom
Multiple R-squared:  0.8412,    Adjusted R-squared:  0.8095
F-statistic: 26.49 on 7 and 35 DF,  p-value: 3.361e-12
```

Źródło: Opracowanie własne

Skorygowany współczynnik determinacji wynosi 80.95%, a istotność zmiennych jest na wysokim poziomie, poza zmienną X_2, dla której *p-value* w teście t-studenta wynosi 0.1317, co jest niestety istotnie większe od 0.05. *P-value* w teście Shapiro-Wilka utrzymało się na poziomie 0.3633, co świadczy o normalności rozkładu reszt, a współczynnik AIC orz BIC są równe 544.6924 i 560.5432. Na podstawie tych danych zdecydowano się na przeprowadzenie dalszych testów statystycznych, uznając tym samym model nr 5 za model składający się z finalnego zestawu zmiennych.

Pierwszym przeprowadzonym testem jest test Chowa, opisany w podrozdziale 1.3., sprawdzający stabilność parametrów. Jest to pierwszy i jeden z najważniejszych testów, świadczący o stopniu dopasowania modelu. *P-value* dla tego testu wyniosło 0.2015, co jest większe od przyjętego poziomu istotności i oznacza przyjęcie hipotezy zerowej

mówiącej o stabilności parametrów. Jest to kluczowe w budowaniu modelu prognostycznego, aby model nie posiadał załamania strukturalnego oraz aby parametry nie zmieniały się w czasie (dotyczy to głównie szeregów czasowych), czyli były stabilne.

Kolejnym przeprowadzonym testem jest test serii badający losowość reszt oraz liniowość modelu. Aby go przeprowadzić należy zbadać reszty modelu oraz ich znaki w celu wyznaczenia serii. Do przeprowadzenia testu przygotowano dane w sposób omówiony we wcześniejszym podrozdziale 1.3. Po zbadaniu reszt oszacowano, że liczba serii dodatnich wynosi 10 a ujemnych 9. Z tablic odczytano, że S1* oraz S2* są kolejno równe 5 oraz 15. Suma wszystkich serii wynosi 19, zatem dana nierówność nie jest spełniona, co świadczy o braku losowości reszt oraz braku liniowości w modelu.

Zanim zdecydowano się na zmianę struktury modelu, postanowiono skonfrontować wynik testu serii z testem RESET Ramseya, który został opisany wraz z testem serii w podrozdziale 1.3. Test RESET przeprowadzono według wzoru przedstawionego we wcześniejszym podrozdziale i otrzymano według niego *p-value* równe dla testu 0.1294, co świadczy o liniowości modelu. Jak widać otrzymane wyniki w obu testach są sprzeczne, dlatego postanowiono kontynuować badania i sprawdzić czy, mimo niepożądanego wyniku testu serii, osiąga on zamierzony cel – dobre przewidywanie wartości Y, czyli relatywnie niski błąd prognozy.

Następnie przeprowadzono test Walda na sprawdzenie istotności wszystkich parametrów, czyli zarazem istotności całego modelu. Test przeprowadzono według zasad opisanych w podrozdziale 1.3. Na rysunku 13. przedstawiono wyniki tego testu:

Rysunek 13: Wynik testu Walda

```
Model 1: y ~ x1 + x2 + x3 + x6.1 + x10 + x12 + x16
Model 2: y ~ 1
  Res.Df Df      F   Pr(>F)
1     39
2     46 -7 26.759 5.14e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Źródło: Opracowanie własne

Z powyższego rysunku wynika, że *p-value* dla testu wynosi $5.14e^{-13}$, co jest istotnie mniejsze od $\alpha = 0.05$ i świadczy o wysokiej istotności parametrów. Jest to jak najbardziej pożądana wartość, która umożliwia kontunuowanie badań nad autokorelacją oraz heteroskedastycznością.

Przedostatnim badaniem wykonanym na stworzonym modelu jest wykrywanie heteroskedastyczności. Do przeprowadzenia tego badania użyto dwóch testów: Breuscha-Pagana oraz White'a. Oba testy zostały opisane w podrozdziale 1.3, gdzie zostały przedstawione wszystkie wzory użyte do obliczeń. Dla obu tych testów otrzymano kolejno *p-value* wynoszące 0.1647 oraz 0.651, co świadczy o braku heteroskedastyczności w modelu.

Oznacza to, że wszystkie wariancje zmiennych losowych są równe, a co za tym idzie uzyskane estymatory są zgodne. W przypadku wystąpienia różnicy w wariancjach zmiennych losowych model mógłby się odznaczać niedoszacowanymi średnimi błędami estymatorów, co mogłoby prowadzić do niepoprawnej i zafałszowanej prognozy.

Ostatnim etapem testów statystycznych jest zbadanie autokorelacji. W celu wykrycia autokorelacji pierwszego rzędu przeprowadzono test Durbina-Watsona. Został on opisany w podrozdziale 1.3. wraz przedstawieniem stosownych wzorów matematycznych. W teście otrzymano *p-value* wynoszące 0.3658 a statystykę testową równą 1.9606, co świadczy o przyjęciu hipotezy zerowej oraz o braku autokorelacji pierwszego rzędu składnika losowego. Oznacza to, że żadna zmienna niezależna nie jest w postaci dowolnego przekształcenia zmiennej niezależnej. Dzięki temu prognozowana wartość zmiennej Y w żadnym stopniu nie będzie zależna od poprzednich obserwacji.

Po przeprowadzeniu wszystkich testów statystycznych, obliczono średni absolutny błąd procentowy prognozy *ex post* MAPE, by sprawdzić czy tak zbudowany model spełnia swój cel. Średni błąd prognozy dla wcześniej usuniętych 4 obserwacji (grupy kontrolnej) wyniósł 25.57%, co nie jest pożądanym wynikiem. Poniższa tabela przedstawia wyniki błędu predykcyjnego dla wszystkich skonstruowanych modeli liniowych:

Tabela 7: Porównanie model pod względem błędu prognozy

Numer modelu	Błąd MAPE	Mediana błędu
Model nr 1	25.68%	20.90%
Model nr 2	23.12%	18.90%
Model nr 3	24.84%	17.41%
Model nr 4	23.80%	18.00%
Model nr 5	24.31%	16.81%

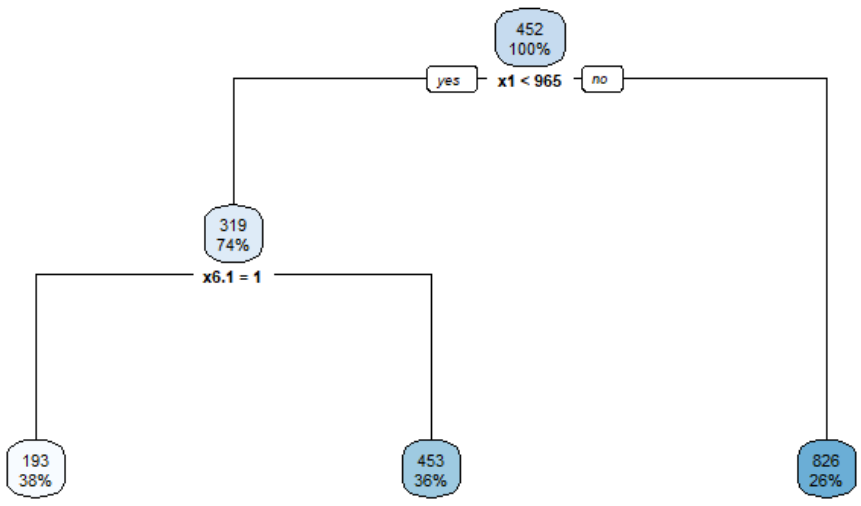
Źródło: Opracowanie własne

Błąd MAPE obliczony został dla wszystkich obserwacji (nie tylko dla 4 wylosowanych jako grupa kontrolna projektów), w tym dla obserwacji odstających.

Finalny model nr 5 nie spełnia swojego celu, mimo spełnienia wszystkich założeń dla modelu liniowego oraz dobrych wyników w przeprowadzonych testach statystycznych. Z tego względu oraz ze względu na wynik testu serii świadczący o złym dobraniu modelu, zdecydowano się na konstrukcję drzewa regresyjnego w celu porównania wyników i próby osiągnięcia lepszego modelu predykcyjnego.

Przed rozpoczęciem tworzenia drzewa decyzyjnego podzielono obserwacje na dwa zestawy – treningowy oraz testowy. Do zestawu treningowego wybrano 80% obserwacji, a pozostałe 20% pozostawiono w celu sprawdzenia możliwości predykcyjnych zbudowanego drzewa jako zestaw testowy. Następnie zbudowano pierwsze drzewo regresyjne używając algorytmu entropii, opisanego w podrozdziale 1.4. oraz według opisanych tam kryteriów mówiących o minimalizacji błędu RSS. Powstało następujące drzewo:

Rysunek 14: Drzewo regresyjne nr 1



Źródło: Opracowanie własne

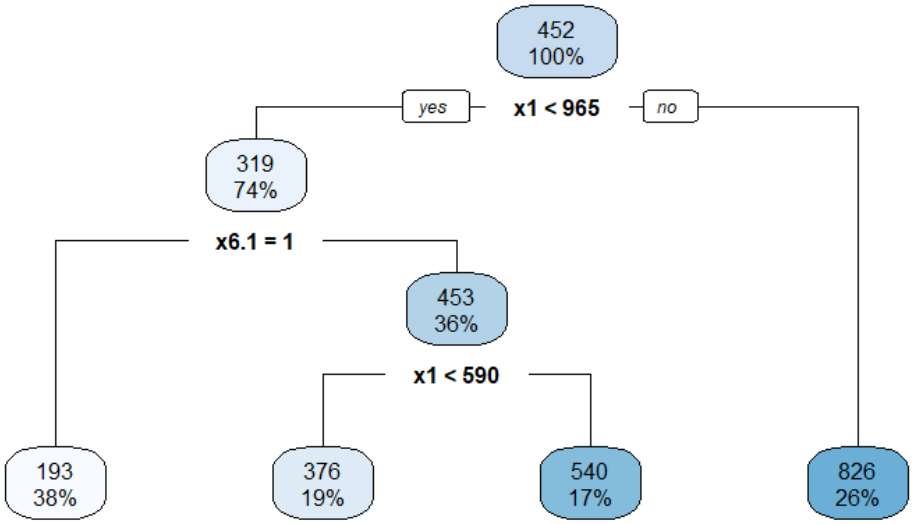
Rys. 14 pokazuje, że do drzewa regresyjnego przyjęto jedynie zmienne X_1 oraz $X_{6.1}$. Algorytm uznał, że tylko te zmienne mają znaczący wpływ na zmienną objaśnianą Y i tylko dwa węzły generują zysk informacyjny. Może się tak dziać, ponieważ wzięto pod uwagę bardzo małą liczbę danych oraz niektóre zmienne objaśniające mogą posiadać

niskie zdolności dyskryminacyjne, przez co nie generują zysku informacyjnego i zostają odrzucone z drzewa.

Z analizy uzyskanego drzewa wynika, że obserwacje podzielono na trzy podzbiory, które uznano za wystarczająco podobne do tworzenia prognozy. W pierwszym podzbiorze znalazło się 38% obserwacji, w drugim 36%, natomiast w ostatnim najmniej, bo 26%. Błąd PRESS (z ang. Predicted Residual Error Sum of Squares), który jest odzwierciedleniem błędu RSS wyniósł 78.69% .Spróbowano zbudować kolejne drzewo, jednak z określonymi wcześniej parametrami – maksymalną głębokością równą 12, minimalną liczbą obserwacji jaka musi znaleźć się w podzbiorze aby dokonać rozwidlenia równą 11 oraz wymogiem, aby przy każdym podziale danych na podzbiory zysk informacyjny zwiększał się o co najmniej 2%. Po zdefiniowaniu takich wymagań otrzymano wyniki identyczne jak dla drzewa pierwszego.

Po nieudanej próbie znalezienia lepszego drzewa zdecydowano się na utworzenie pętli, w której tworzone są drzewa w zależności od głębokości i liczby obserwacji wymaganych w podzbiorze. Ustalono te parametry kolejno od 1 do 16 oraz od 5 do 20, zwiększając je dokładnie o jeden w każdej następnej iteracji pętli. W ten sposób stworzono 256 drzew o różnych cechach. Następnie zdefiniowano funkcje, które mają na celu wybranie najlepszego drzewa z wszystkich wygenerowanych według najniższego błędu RSS oraz największego zysku informacyjnego. Najlepsze okazało się drzewo, dla którego ustalono minimalną liczbę obserwacji w podzbiorze równą 14, a głębokość drzewa równą 11. Dla tak wybranego drzewa błąd PRESS wyniósł 77.44%, co jest nieznacznie lepszym wynikiem niż PRESS dla drzewa pierwszego. Przez niewielkie różnice, ostatecznie wybrane drzewo wygląda bardzo zbliżenie do drzewa skonstruowanego bez żadnych ograniczeń na samym początku badań. Poniższy rysunek przedstawia jego konstrukcję:

Rysunek 15: Drzewo regresyjne nr 3

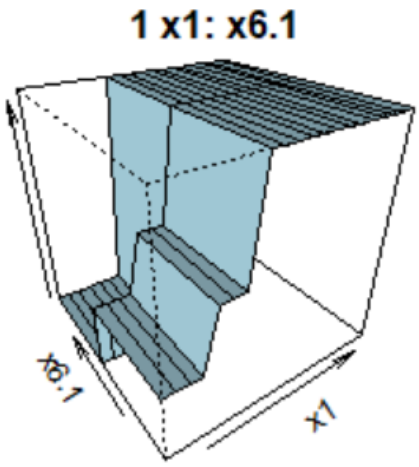


Źródło: Opracowanie własne

Przycinanie drzewa nie zmieniło jego konstrukcji, ponieważ jest ono drzewem mało skomplikowanym, o małej liczbie węzłów oraz charakteryzującego się bardzo niską głębokością. Z rys.15 można zobaczyć, że drzewo nadal bierze pod uwagę tylko dwie, te same, zmienne ze wszystkich osiemnastu. W tym modelu podzielono obserwacje na cztery różne podzbiory, zamiast trzech (tak jak to było w drzewie nr 1), przez co zwiększono wariancję, jednak przy tym zmniejszono błąd RSS. Dla tak skonstruowanego drzewa błąd predykcji MAPE wyniósł 63.08%, co niestety jest jeszcze gorszym wynikiem niż ten otrzymany, w liniowym modelu ekonometrycznym.

Dzieje się tak prawdopodobnie dlatego, że ponieważ mimo iż drzewa decyzyjne są wygodnym narzędziem statystycznym, mają one przewagę jedynie przy dużych zestawach danych. Można spostrzec, iż drzewo sugeruje, że dla najmniej skomplikowanych projektów przewidywany jest czas realizacji równy 193 godziny, dla projektów najdłuższych czas realizacji wyniesie 826 godzin, a dla projektów o krótkiej długości, ale większym skomplikowaniu czas trwania realizacji ma wynieść odpowiednio 376 lub 540 godzin. Podział zmiennej Y według zmiennych X_1 oraz $X_{6.1}$ przedstawia poniższy rysunek:

Rysunek 16: Podział danych według drzewa nr 3



Źródło: Opracowanie własne

Ten wykres jednoznacznie upewnia w przekonaniu, że największy wpływ na zmienną Y ma zmienna X_1, tj. długość drogi. Poniższy rysunek przedstawia natomiast reguły decyzyjne, według których skonstruowane jest ostatnie drzewo regresyjne:

Rysunek 17: Reguły decyzyjne w drzewie nr 3

```
  y                                                cover
193 when x1 <  965          & x6.1 is 1            38%
376 when x1 <  590          & x6.1 is 0            19%
540 when x1 is 590 to 965 & x6.1 is 0              17%
826 when x1 >=        965                          26%
```

Źródło: Opracowanie własne

Kolumna *cover* pokazuje ile procent obserwacji jest zawartych w danej regule decyzyjnej. Jak widać najwięcej obserwacji zostało ujętych w pierwszej regule (czyli ją spełnia) sprawdzającej czy długość drogi jest poniżej 965 metrów, co jest zrozumiałe, ponieważ jest to reguła zawarta w pierwszym węźle – na wejściu rozpatruje on 100% obserwacji.

Rozdział 3 : Wnioski i ich interpretacja

Badania empiryczne wykazały, że zarówno drzewa regresyjne, jak i liniowe modele ekonometryczne, nie zawsze dostatecznie dokładnie opisują rzeczywistość. Próby predykcji liczby godzin spędzonych nad realizacją projektu drogowego okazały się być pracochłonne, ze względu na poziom komplikacji przedmiotu. Duża liczba zmiennych binarnych oraz mała liczba zmiennych ilościowych zdecydowanie utrudniła próby konstrukcji zwłaszcza modelu liniowego, dlatego w dalszych etapach badań zdecydowano się również na utworzenie drzewa regresyjnego.

Finalny liniowy model ekonometryczny bierze pod uwagę zmienne: X_1, X_2, X_3, $X_{6.1}$, X_{10}, X_{12} oraz X_{16}. Oznacza to, że model jest budowany na podstawie:

- długości drogi;
- liczby kolidujących instalacji;
- wymagania pozwolenia wodnoprawnego;
- postępowania według przepisów omówionych w zgłoszeniu robót budowlanych;
- doświadczenia zespołu projektowego;
- wymagania decyzji o uwarunkowaniach środowiskowych DUŚ;
- wymaganej ilości dokumentacji.

Tak skonstruowany model można zinterpretować w następujący sposób:

Tabela 8: Interpretacja skonstruowanego modelu

Zmienna	Współczynnik	Interpretacja
Stała	189.9	Bez względu na cechy projektu, czas realizacji każdego z nich jest równy 189.9 h.
X_1	0.23	Wraz z każdym metrem drogi, czas realizacji projektu wydłuża się o 0.23 h.
X_2	29.4	Wraz z każdą kolidującą instalacją, czas realizacji projektu wydłuża się o 29.4 h.

X_3	150.94	Wraz z wymaganiem pozwolenia wodnoprawnego, czas realizacji projektu wydłuża się o 150.94 h.
$X_{6.1}$	-120.07	Wraz z postępowaniem według przepisów omówionych w zgłoszeniu robót budowlanych, czas realizacji projektu skraca się o 120.07 h.
X_{10}	-87.47	Wraz z każdym poziomem doświadczenia zespołu projektowego, czas realizacji projektu skraca się o 87.47 h.
X_{12}	101.47	Wraz z wymaganiem decyzji o uwarunkowaniach środowiskowych DUŚ, czas realizacji projektu wydłuża się o 101.47 h.
X_{16}	87.71	Wraz z każdym stopniem wymagania ilości dokumentów, czas realizacji projektu wydłuża się o 87.71 h.

Źródło: Opracowanie własne

Finalny model nie bierze pod uwagę innych, z punktu widzenia wiedzy eksperckiej relatywnie istotnych zmiennych, tj. zmiany przepisów, typu drogi, czy obiektów inżynierskich.

Niestety, mimo iż liczba obiektów inżynierskich jest znana przed rozpoczęciem realizacji projektu, nie została ona uwzględniona w modelu. Oznacza to, że jednostka opracowująca dany projekt przed zaprognozowaniem liczby godzin potrzebnych do jego realizacji, powinna subiektywnie ocenić koszt lub czas, który poświęci na realizację tych obiektów. Do obiektów inżynierskich zalicza się między innymi mosty oraz przepusty, które są istotną częścią projektów drogowych i mogą całkowicie zmienić wynik szacowania ich czasochłonności oraz ich kosztu realizacji. Należy również zwrócić uwagę, że dane obejmują projekty o długości drogi nie większej niż 2.5 km. Można z tego wywnioskować, że jeżeli zespół projektowy ma zamiar zrealizować projekt o większej długości drogi, skonstruowany model może okazać się nieprzydatny, ponieważ zaprognozuje za dużą, lub co bardziej prawdopodobne, za małą liczbę godzin.

Skonstruowane drzewo regresyjne okazało się być jeszcze gorszym wyborem od liniowego modelu ekonometrycznego, ponieważ jego błąd prognozy MAPE wyniósł aż ponad 60%, co jest absolutnie niedopuszczalnym z praktycznego punktu widzenia wynikiem. Drzewo to również bierze pod uwagę tylko dwie zmienne:

X_1 – długość drogi oraz

$X_{6.1}$ - postępowania według przepisów omówionych w zgłoszeniu robót budowlanych.

Poniższa tabela przedstawia zestawienie ostatecznie przyjętego modelu liniowego z ostatecznie przyjętym drzewem regresyjnym:

Tabela 9: Porównanie błędu prognozy - model liniowy wraz z drzewem regresyjnym

	Model liniowy	Drzewo regresyjne
Błąd prognozy MAPE	25.57%	63.08%

Źródło: Opracowanie własne

Z tabeli nr 9 można zobaczyć, że błąd prognozy MAPE dla drzewa regresyjnego znacznie odbiega od normy. W poprzednim rozdziale stwierdzono, że jest to spowodowane analizą zbyt małej liczby obserwacji, przez co zestaw treningowy podczas konstruowania drzewa nie miał wystarczająco dużo danych, aby zbudować dobrze dopasowany model.

Oba modele nie spełniają swojej roli i nie nadają się do użytku w realnym życiu, ponieważ generują za duże ryzyko przy wycenie projektu drogowego. W porównaniu do prognoz tworzonych za pomocą wiedzy eksperckiej, bez stosowania żadnych narzędzi analitycznych i statystycznych, błąd prognostyczny jest znacznie większy, ponieważ dotychczasowy średnio procentowy błąd prognozy wynosił 10-15%. Niepewność tej branży jest bardzo duża, co może być spowodowane ciągłymi zmianami, z których nie wszystkie zostały ujęte w przyjętych zmiennych.

Wycena projektu drogowego za pomocą modelu analitycznego jest realnym problemem, który nadal pozostaje nierozwiązany ze względu na swoją złożoność. Kolejnym powodem, dla którego zbudowane modele predykcyjne nie spełniają swojej funkcji może być za mały zestaw danych. Należałoby zebrać znacznie więcej projektów do analizy z większej liczby przedsiębiorstw projektowych, niestety z oczywistych powodów nie było to możliwe dla tego badania. Ostatnią możliwością jest użycie o wiele bardziej skomplikowanych metod w celu zbudowania lepiej dopasowanego modelu

predykcyjnego, jedną z nich może być użycie regresji logistycznej czy stworzenie nieliniowego modelu ekonometrycznego.

Zakończenie

Celem pracy było stworzenie modeli, które zaprognozują liczbę godzin potrzebnych do opracowania projektu drogowego na podstawie najistotniejszych zmiennych niezależnych. Podstawowym założeniem było użycie regresji liniowej oraz drzew regresyjnych. Przy budowie modelu liniowego uznano, że model nie spełnił swojego celu ze względu na wysokie wskazanie błędu prognozy MAPE oraz ze względu na brak losowości reszt według testu serii. Następnie skonstruowano drzewo regresyjne w celu porównania wyników ze zbudowanym modelem liniowym oraz próby konstrukcji lepszego modelu prognozującego. Niestety drzewo regresyjne również nie osiągnęło zamierzonych wyników, ponieważ błąd prognozy MAPE wyniósł 63%. Zdecydowano się na wybranie liniowego modelu ekonometrycznego, dla którego wyniki są bez porównania lepsze, lecz nadal nie wystarczające. Może być to spowodowane posiadaniem zbyt małej liczby obserwacji (tj. projektów drogowych). Jednym z rozwiązań byłoby powiększenie zestawu danych o nowe obserwacje lub użycie bardziej zaawansowanych metod prognostycznych, takich jak nieliniowy model ekonometryczny. Przyjęty model liniowy przewiduje niestety dla projektów o długości powyżej 2.5 km za małą liczbę godzin, co potwierdziły już pierwsze próby wykorzystania modelu. Jednak mimo to, praca ta zbliżyła się do rozwiązania jednego z rzeczywistych problemów znajdujących się w branży inwestycji drogowych – oszacowanie kosztu projektu na etapie przetargu za pomocą narzędzi analitycznych. Niestety skonstruowany model posiada medianę błędu procentowego w wysokości 16.8%, co nadal jest wysokim poziomem ryzyka w tej branży, który może skutkować złym oszacowaniem projektu drogowego na poziomie kilkudziesięciu tysięcy złotych. Do tej pory w przedsiębiorstwie używano prognoz tworzonych za podstawie wiedzy eksperckiej i ich średni względny błąd wahał się między 10-15%. Oszacowane modele niestety nie obniżyły tego błędu, więc przedsiębiorstwo projektowe zdecydowało się na pozostaniu przy swoich metodach wyceny projektów drogowych.

Bibliografia

Chaber P. i inni. (2019). *Raport o stanie sektora małych i średnich przedsiębiorstw w Poslce.* Warszawa: PARP.

Davidson, R. (2004). *Econometric Theory and Methods.* Nowy Jork: Oxford University Press.

Gajewska-Jaszczuk L. i inni. (1996). *Nowa encyklopedia powszechna PWN. Tom 3.* Warszawa: Wydawnictwo Naukowe PWN.

Gray-Steinhauer, L. K. (2013). *Non-linear & Logistic Regression.* University of Alberta.

James G., Witten D., Hastie T., Tibshirani R. (2017). *An Introduction to Statistical Learning with Applications in R.* Nowy Jork: Springer.

Jóźwiak J., Podgórski J. (2012). *Statystyka od podstaw*. Warszawa: Polskie Wydawnictwo Ekonomiczne.

Kopiński A., Porębski D. (2014). Zastosowanie metody Hellwiga do konstrukcji modelu ekonometrycznego dla stóp zwrotu funduszy inwestycyjnych. *Annales Universitatis Mariae Curie-Skłodowska. Sectio H. Oeconomia*, 147-156.

Kozak J., Juszczuk P. (2016). Algorytmy do konstruowania drzew decyzyjnych w przewidywaniu skuteczności kampanii telemarketingowej banku. *Studio Informatica Pomerania*, 49-59.

Kufel, T. (2013). *Ekonometria. Rozwiązywanie problemów z wykorzystaniem programu GRETL*. Warszawa: Wydawnictwo Naukowe PWN.

Maddala, G. (2006). *Ekonometria.* Warszawa: Wydawnictwo Naukowe PWN.

Mycielski, J. (2010). *Skrypt. Rozdział 10 Metodologia testowania hipotez.* Pobrano z lokalizacji http://www.ekonometria.wne.uw.edu.pl/uploads/Main/Met_testowanie.pdf, Dnia (2020, 05, 22)

Ramanathan, R. (1995). *Introductory Econometrics with Applications.* Cincinnati, Ohio: South-Western College Pub.

Sobczyk, M. (2013). *Ekonometria.* Warszawa: C.H. Beck.

Stanisz, A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, Tom. 2 Modele liniowe i nieliniowe.* Kraków: StatSoft Polska Sp z o.o.

Zagdański A., Suchwałko A. (2016). *Analiza i prognozowanie szeregów czasowych. Praktyczne wprowadzenie na podsatwie środowiska R.* Warszawa: Wydawnictwo Naukowe PWN.

Spis tabel

Spis rysunków