

# Regresja zmian cen akcji

Karol Oleszek

20 maja 2020

# Spis treści

<b>1</b>	<b>Wstęp</b>	<b>4</b>
<b>2</b>	<b>Cel projektu</b>	<b>5</b>
2.1	Model wyboru akcji do celów inwestycyjnych . . . . .	5
2.2	Zbadanie zależności pomiędzy zmianą cen, a informacjami finansowymi . . . . .	5
<b>3</b>	<b>Opis danych</b>	<b>6</b>
3.1	Zbiór danych . . . . .	6
3.2	Usuwanie braków danych . . . . .	6
3.3	Transformacja zmiennej kategorycznej . . . . .	6
3.4	Zmienne objaśniające . . . . .	6
3.5	Rozkład zmiennej objaśnianej . . . . .	7
3.6	Korelacja . . . . .	8
3.7	Podział zbioru danych . . . . .	9
<b>4</b>	<b>Wybór postaci modelu oraz dobór zmiennych do modelu</b>	<b>10</b>
4.1	Specyfikacja kryteriów . . . . .	10
4.2	Rozważana klasa funkcji . . . . .	10
4.3	Rozważany podzbiór funkcji . . . . .	11
4.4	Kryterium wyboru modelu prognostycznego . . . . .	11
4.5	Kryterium wyboru modelu analitycznego . . . . .	11
<b>5</b>	<b>Weryfikacja poprawności modelu</b>	<b>12</b>
5.1	Współliniowość . . . . .	12
5.2	Koincydencja . . . . .	12
5.3	Efekt katalizy . . . . .	12
5.4	Normalność rozkładu reszt . . . . .	13
5.5	Istotność zmiennych objaśniających . . . . .	13
5.6	Istotność współczynnika determinacji . . . . .	14
5.7	Liniowość postaci modelu . . . . .	14
5.8	Homoskedastyczność . . . . .	14
5.9	Stabilność parametrów modelu . . . . .	15
5.10	Autokorelacja składnika losowego . . . . .	15
<b>6</b>	<b>Korekty</b>	<b>16</b>
6.1	Korekty stabilności . . . . .	16
6.2	Korekta heteroskedastyczności i autokorelacji . . . . .	16
<b>7</b>	<b>Wybrane modele</b>	<b>17</b>
7.1	Wybór modelu . . . . .	17
7.2	Model prognostyczny . . . . .	17
7.3	Postać modelu . . . . .	17
7.4	Wystymowane parametry modelu . . . . .	17
7.5	Wskaźniki jakości modelu . . . . .	17

7.6	Weryfikacja poprawności modelu . . . . .	17
7.6.1	Koincydencja . . . . .	17
7.6.2	Katalizatory . . . . .	18
7.6.3	Współliniowość zmiennych . . . . .	18
7.6.4	Normalność rozkładu reszt . . . . .	18
7.6.5	Istotność zmiennych objaśniających . . . . .	18
7.6.6	Istotność współczynnika determinacji . . . . .	19
7.6.7	Liniiowość postaci modelu . . . . .	19
7.6.8	Stabilność parametrów modelu . . . . .	19
7.6.9	Homoskedastyczność . . . . .	19
7.6.10	Autokorelacja czynnika losowego I rzędu . . . . .	19
7.6.11	Wynik weryfikacji poprawności modelu . . . . .	19
7.7	Model analityczny . . . . .	20
<b>8</b>	<b>Prognoza</b>	<b>20</b>
<b>9</b>	<b>Interpretacja</b>	<b>20</b>
<b>10</b>	<b>Podsumowanie</b>	<b>20</b>
<b>11</b>	<b>Spis tabel</b>	<b>21</b>
<b>12</b>	<b>Spis rysunków</b>	<b>22</b>
<b>13</b>	<b>Literatura</b>	<b>23</b>

# 1 Wstęp

Przewidywanie zmian cen akcji oraz innych instrumentów finansowych znajduje się w centrum zainteresowania inwestorów. Zmiany cen są podstawowym zjawiskiem powodującym bogacenie się lub ubożenie inwestora indywidualnego bądź instytucjonalnego, dlatego też próby zrozumienia i opisanie reguł rządzących tym zjawiskiem są kluczowe dla podejmowania skutecznych decyzji o alokacji kapitału.

Teoria rynków kapitałowych proponuje wiele różnych wyjaśnień zmienności cen: hipoteza rynku efektywnego (Bachelier [1]) zakłada, że ceny rynkowe akcji w danej chwili odzwierciedlają wszystkie dostępne informacje o spółce; autorzy i zwolennicy hipotezy krótkoterminową zmienność cen opisują jako losowy ruch wokół efektywnej wartości. Hipoteza rynku efektywnego znalazła wielu zwolenników, którzy poddawali w wątpliwość samą zasadność przewidywania cen (Cowles [4]), jak również zainspirowała powstanie indeksowych funduszy inwestycyjnych.

Hipoteza rynku efektywnego spotkała się z szeroką krytyką ze strony ekonomistów i inwestorów giełdowych, którzy wskazywali na kontrprzykłady obalające hipotezę. Współcześnie właściwie wszystkie duże organizacje finansowe używają różnego rodzaju systematycznych narzędzi do analizy i prognozy zmian cen na rynkach kapitałowych (Graham Capital Management [5]). Duże oraz wciąż rosnące znaczenie ma też algorytmiczny handel (Capgemini [2]).

Do złożonego problemu jakim jest symulacja i prognozyka zachowania rynków kapitałowych stosuje się bardzo szeroki wachlarz metod statystycznych, algorytmicznych i ekonometrycznych. Duże zastosowanie mają metody uczenia maszynowego (Shunrong Shen [7]), w tym głębokie sieci neuronowe o niekonwencjonalnych architekturach. Ponadto do prognozyki coraz częściej używa się analizy języka naturalnego (Zhaoxia Wang [8]).

Poniższa praca zawiera przekrojową regresję zmian cen akcji na rynku amerykańskim w 2019 z wykorzystaniem standardowych narzędzi ekonometrycznych. Zbiór danych służący do konstrukcji modelu zawiera dane z roku 2018, dotyczące sytuacji finansowych, kapitałowych i operacyjnych spółek, zawarte w formie wskaźników i pozycji ze sprawozdań finansowych.

## 2 Cel projektu

### 2.1 Model wyboru akcji do celów inwestycyjnych

Celem projektu jest wyznaczenie bazowego poziomu efektywności wyboru spółek, których akcje w nadchodzącym roku zyskają na wartości. Za wybór odpowiadał będzie model, który powstanie przy użyciu metody najmniejszych kwadratów i który będzie mógł służyć jako punkt odniesienia do badania efektywności innych metod predykcji.

Efektywność prognostyczna modelu zostanie zbadana przy użyciu *średniego błędu prognozy ex post*, danego wzorem:

$$ME = \frac{1}{s} \sum_{t=1}^s (y_t - y_t^P)$$

,

Gdzie:

$s$  - ilość obserwacji w testowym zbiorze danych

$y_t$  - prawdziwa wartość zmiennej objaśnianej

$y_t^P$  - prognozowana wartość zmiennej objaśnianej

### 2.2 Zbadanie zależności pomiędzy zmianą cen, a informacjami finansowymi

Ponadto model posłuży do oceny wpływu informacji finansowych zawartych w publicznie dostępnych źródłach na przyszłą wartość spółek giełdowych. Ocena ta może być użyteczna przy podejmowaniu decyzji o tym, jakie dane zbierać na temat spółek w celu skutecznego przewidywania ich przyszłej wyceny.

Miarą tej oceny będzie współczynnik determinacji  $R^2$ , dany wzorem:

$$R^2 = \frac{\sum_{t=1}^n (y_t^P - \bar{y})^2}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

,

Gdzie:

$n$  - ilość obserwacji w uczącym zbiorze danych

$y_t$  - prawdziwa wartość zmiennej objaśnianej

$y_t^P$  - prognozowana wartość zmiennej objaśnianej

$\bar{y}$  - średnia arytmetyczna zmiennej objaśnianej

## 3 Opis danych

### 3.1 Zbiór danych

Zbiór danych użyty w projekcie pochodzi z internetowej platformy Kaggle (Carbone [3]). Zawiera on zmienną objaśnianą  $Y$  - procentową zmianę ceny akcji danej spółki w 2019 roku, oraz zmienne objaśniające  $X_i, i = 1 \dots k$  -  $k-1$  wskaźników finansowych i pozycji z formularza  $10-K^1$ , a także zmiennej kategorycznej oznaczającej sektor gospodarki rozważanej spółki.

### 3.2 Usuwanie braków danych

Dane zostały zebrane przy użyciu interfejsu programistycznego *Financial Modeling Prep API* i zawierały pewne braki wynikające z różnic w dokumentach źródłowych. Dla celów analizy usunięte zostały wszystkie obserwacje, w których brakowało więcej niż 50 wartości oraz wszystkie zmienne, w których co najmniej 10% obserwacji nie miało przypisanej wartości. Po tej transformacji, w zbiorze danych pozostało 4122 obserwacji oraz 179 zmiennych (4392 x 222, 9,98% braków przed transformacją). Wciąż brakujące 0,82% wartości zostało zastąpionych średnimi arytmetycznymi odpowiednich zmiennych.

### 3.3 Transformacja zmiennej kategorycznej

Kategoryczna zmienna objaśniająca *Sector*, która przyjmowała 11 różnych wartości (Consumer Cyclical, Energy, Technology, Industrials, Financial Services, Basic Materials, Communication Services, Consumer Defensive, Healthcare, Real Estate, Utilities) została przekształcona na 10 zmiennych zero-jedynkowych. Po tej operacji zbiór danych składał się ze 188 zmiennych.

### 3.4 Zmienne objaśniające

Zbiór danych składa się z grupy zmiennych opisujących realne wielkości ze sprawozdań finansowych wyrażone w dolarach amerykańskich np. zysk brutto, wydatki na badania i rozwój. Na drugą grupę zmiennych składają się wskaźniki finansowe, będące często przekształconymi zmiennymi z pierwszej grupy, wyrażone jako stosunek różnych wielkości np. zysk na akcję, wzrost zysku w ciągu roku. Trzecia grupa zmiennych to przekształcona zmienna *Sector*. Ze względu na liczbę zmiennych pełna lista zmiennych wraz ze statystykami znajduje się w załączniku do pracy.

---

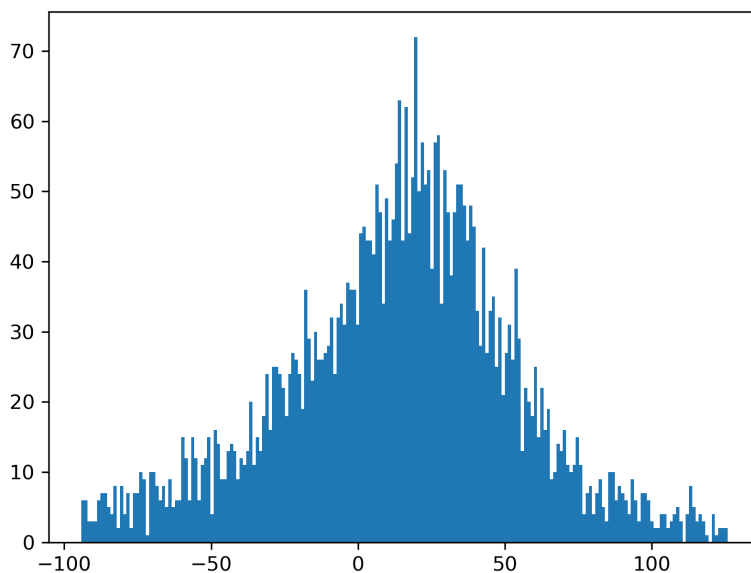
<sup>1</sup>Form 10-K jest to coroczne podsumowanie finansowe składane przez amerykańskie spółki giełdowe do *U.S. Securities and Exchange Commission*, federalnej agencji nadzoru finansowego.

### 3.5 Rozkład zmiennej objaśnianej

Zmienna objaśniana  $Y$  to wyrażona w procentach zmiana ceny akcji spółki w roku 2019. Przyjmuje ona wartości -99,86% do +3756,72%. Większa część wartości jest większa od zera co wskazuje na pozytywną bazową efektywność tzw. strategii *buy and hold*, która polega na zakupie akcji, a następnie oczekiwaniu na wzrost jej wartości.

Statystyka	Wartość
Średnia arytmetyczna	21,15
Odchylenie standardowe	84,93
Kwartyl dolny	-9,30
Mediana	17,83
Kwartyl górny	40,92
Wartość najmniejsza	-99,86
Wartość największa	3756,71

Tabela 1: Statystyki - zmienna objaśniana



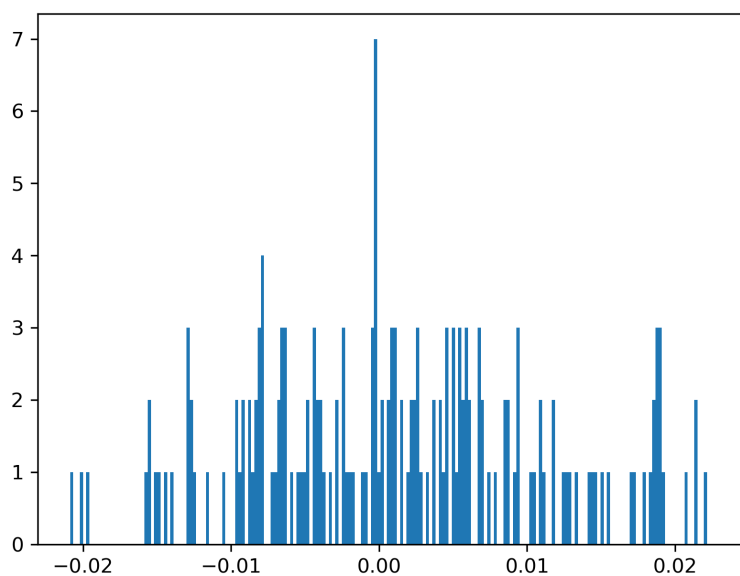
Rysunek 1: Histogram zmiennej objaśnianej

### 3.6 Korelacja

Korelacja zmiennej objaśnianej ze zmiennymi objaśnianymi jest bardzo słaba, co wskazuje na potencjalnie silnie nieliniowy charakter zachodzących zależności.

Statystyka	Wartość
Średnia arytmetyczna	0.0006535276131026212
Odchylenie standardowe	0.01430737685985159
Kwartył dolny	-0.006560508005214499
Mediana	0.0009029552423272379
Kwartył górny	0.008516164727314403
Wartość najmniejsza	-0.07707504123521973
Wartość największa	0.04058496958769639

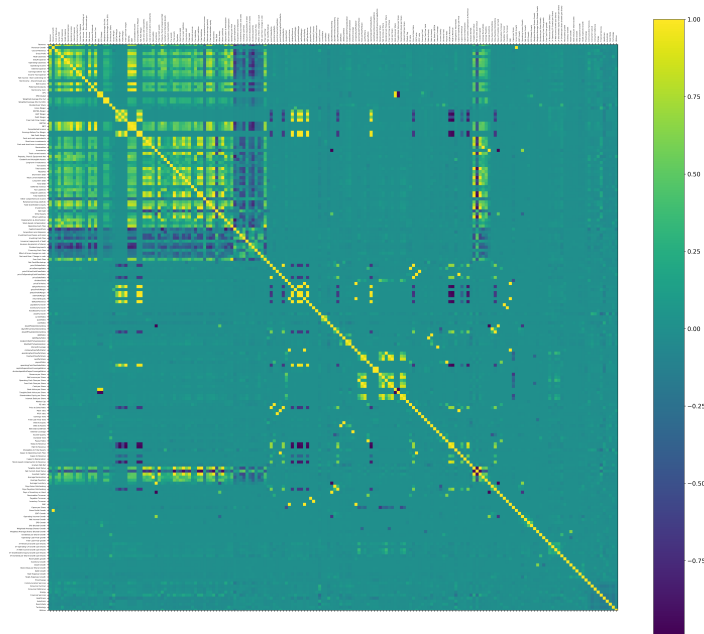
Tabela 2: Korelacja zmiennej objaśnianej ze zmiennymi objaśnianymi



Rysunek 2: Histogram korelacji zmiennej objaśnianej ze zmiennymi objaśnianymi



Korelacja pomiędzy niektórymi zmiennymi objaśniającymi jest silna, co można wyjaśnić zależnościami pomiędzy rozmiarem spółki, a wielkościami w *10-K Form*. Nie wszystkie bazowe zmienne objaśniające mogą znaleźć się w modelu, ponieważ spowodowałoby to wystąpienie zjawiska współliniowości.



Rysunek 3: Macierz korelacji

### 3.7 Podział zbioru danych

Zbiór danych podzielony jest na dane treningowe(90%) użyte do estymacji parametrów modeli oraz dane testowe(10%) służące do oceny prognozy *ex post*.

## 4 Wybór postaci modelu oraz dobór zmiennych do modelu

### 4.1 Specyfikacja kryteriów

Praca odpowiada na dwa pytania:

1. Jaka jest bazowa efektywność predykcji?
2. Jaki jest wpływ informacji zawartych w rozważanych danych na zmiany cen akcji?

Odpowiedzią na pierwsze pytanie jest efektywność modelu o najlepszych właściwościach prognostycznych.

Odpowiedzią na drugie pytanie jest ocena współczynnika determinacji poprawnie zweryfikowanego modelu ekonometrycznego.

### 4.2 Rozważana klasa funkcji

Modele wybrane są z klasy funkcji dopuszczalnych  $\mathcal{F} : \mathbb{R}^K \rightarrow \mathbb{R}$ , gdzie  $K$ : liczba zmiennych w zbiorze danych.

Funkcje przyjmują analityczną postać:

$$\hat{Y} = a_0 + a_1 * X_1 + \dots + a_k * X_k;$$

parametry szacowane są przy użyciu metody najmniejszych kwadratów.

Ze względu na potencjalnie nieliniowy charakter zależności rozważony został

zbiór danych rozszerzony o przetransformowane zmienne:  $e^{\frac{X_i - \overline{X_i}}{\sigma_{X_i}}}, X_i^2, X_i^3$  o łącznym rozmiarze 4392 x 748. Dalsze rozszerzanie zbioru danych osłabiło by stabilność numeryczną modeli i nadmiernie zwiększyło by wymiar Wapnika-Czerwonienkisa, co osłabiło by możliwości generalizacji przez modele (Kamiński [6]).

### 4.3 Rozważany podzbiór funkcji

Ze względu na ilość zmiennych, użycie metody o wykładniczej złożoności obliczeniowej, np. metody Hellwiga, byłoby nieefektywne. W związku z tym rozważany jest K-elementowy zbiór funkcji wyłoniony przy użyciu uproszczonej *metody krokowej wstecz*. Rozważany jest model ze wszystkimi zmiennymi, a następnie z modelu usuwana jest najmniej istotna statystycznie zmienna. Procedura jest powtarzana dopóki w modelu jest więcej niż jedna zmienna. Dla celów stworzenia podzbioru funkcji nie jest sprawdzana normalność rozkładu resztowego.

### 4.4 Kryterium wyboru modelu prognostycznego

Spośród rozważanego podzbioru funkcji wybrana jest funkcja z najmniejszym *absolutnym błędem prognozy ex post* oszacowanym przy użyciu sprawdzianu krzyżowego na danych treningowych.

Sprawdzian krzyżowy polega na podziale zbioru danych na J podzbiorów, a następnie wykonaniu J ocen błędu prognozy (za każdym razem inny podzbiór jest uznawany za testowy, J-1 podzbiorów treningowych) i obliczeniu ich średniej.

$$\hat{ME} = \frac{1}{J} \sum_{t=1}^J (ME_t)$$

### 4.5 Kryterium wyboru modelu analitycznego

Spośród rozważanego podzbioru funkcji wybrana jest funkcja z najwyższym współczynnikiem determinacji  $R^2$ , która została poprawnie zweryfikowana. W przypadku wystąpienia niepożądanych zjawisk w modelu, rozważany podzbiór funkcji powiększony zostaje o model z zastosowanymi stosownymi korektami.

## 5 Weryfikacja poprawności modelu

Poniżej opisane są procedury weryfikacji poprawności modelu oszacowanego metodą najmniejszych kwadratów. Przyjęty poziom istotności  $\alpha = 0,05$ .

### 5.1 Współliniowość

Wyznaczamy k modeli MNK, gdzie zmienną objaśnianą jest jedna ze zmiennych objaśniających, a zmiennymi objaśniającymi pozostałe zmienne. Wyznaczamy k współczynników determinacji; jeżeli którykolwiek z nich większy jest niż 0,9, to w modelu występuje niepożądane zjawisko współliniowości zmiennych.

### 5.2 Koincydencja

Jeżeli dla każdego  $i=1..k$ :

$$\text{sgn}(r_i) = \text{sgn}(a_i)$$

,

gdzie:

$r_i$ : korelacja pomiędzy zmienną objaśnianą, a i-tą zmienną objaśniającą.

$a_i$ : oszacowany i-ty parametr modelu.

to model jest koincydentny. Koincydencja modelu jest pożądaną cechą.

### 5.3 Efekt katalizy

Niech  $(X_i, X_j)$  będzie regularną parą korelacyjną. Wówczas jeżeli  $r_{ij} < 0$  lub  $r_{ij} > \frac{r_i}{r_j}$ , gdzie:

$r_{ij}$ : korelacja między i-tą, a j-tą zmienną objaśniającą.

to zmienna  $X_i$  jest katalizatorem. Poprawny model nie zawiera katalizatorów.

## 5.4 Normalność rozkładu reszt

Normalność rozkładu składnika resztowego modelu jest cechą niezbędną do poprawnej interpretacji m.in. testów istotności parametrów modelu. Normalność rozkładu można zbadać przy użyciu testu Jarque-Bera.

$H_0$ : składnik losowy modelu ma rozkład normalny.

$H_1$ : składnik losowy modelu nie ma rozkładu normalnego.

Statystyka testowa  $JB$  ma rozkład  $\chi^2$  o dwóch stopniach swobody.

$$JB = \frac{n-k}{6} \left( A^2 + \frac{1}{4}(K-3)^2 \right)$$

,

gdzie:

Współczynnik skośności:

$$A = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^3}{\left( \frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^2 \right)^{\frac{3}{2}}}$$

Kurtoza:

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^4}{\left( \frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^2 \right)^2}$$

## 5.5 Istotność zmiennych objaśniających

Zmienne w poprawnym modelu są statystycznie istotne. Do badania istotności zmiennych służy test t-Studenta.

$H_0$ :  $a_i = 0$ .

$H_1$ :  $a_i \neq 0$ .

Statystyka testowa  $t_{a_i}$  ma rozkład t-Studenta o  $n-(k+1)$  stopniach swobody.

$$t_{a_i} = \frac{a_i}{D(a_i)}$$

Dyspersja estymatora i-tego parametru modelu:

$$D(a_i) = \sqrt{d_{ii}}$$

Macierz kowariancji estymatora a:

$$\hat{D}^2(a) = S^2 (X^T X)^{-1}$$

Estymator wariancji  $s^2$  składnika losowego:

$$S^2 = \frac{e^T e}{n - (k + 1)}$$

## 5.6 Istotność współczynnika determinacji

Istotność współczynnika determinacji (inaczej istotność wszystkich zmiennych naraz) jest pożądaną cechą modelu i może być zbadana za pomocą testu  $F$ .

$$H_0: a_1 = a_2 = \dots = a_k = 0.$$

$$H_1: a_1 \neq 0 \vee a_2 \neq 0 \vee \dots \vee a_k \neq 0.$$

Statystyka testowa  $F$  ma rozkład F-Snedecora-Fishera z  $r_1 = k$  i  $r_2 = n - (k + 1)$  stopniami swobody.

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

## 5.7 Liniowość postaci modelu

Do badania liniowości modelu ekonometrycznego służy test serii.

$H_0$ : model hipotetyczny jest liniowy.

$H_1$ : model nie jest liniowy.

Statystyka testowa  $Z$  ma asymptotyczny standardowy rozkład normalny.

$r$  - liczba serii w wektorze resz modelu (uporządkowanych wg. wartości zmiennej objaśnianej).

$N_1$ ,  $N_2$  - liczba dodatnich i ujemnych reszt modelu.

$$Z = \frac{r - \left(\frac{2N_1N_2}{n} + 1\right)}{\sqrt{\frac{2N_1N_2(2N_1N_2 - n)}{(n-1)n^2}}}$$

## 5.8 Homoskedastyczność

Homoskedastyczność jest pożądaną cechą modelu. Ze względu na wysoką proporcję ilości zmiennych objaśniających do ilości obserwacji homoskedastyczność najlepiej sprawdzić jest przy użyciu testu Goldfelda-Quandt.

$H_0: s_1^2 = s_2^2$ . Homoskedastyczność.

$H_1: s_1^2 \neq s_2^2$ . Heteroskedastyczność.

Statystyka testowa  $F$  ma rozkład F-Snedecora-Fishera z  $r_1 = n_1 - (k + 1)$  i  $r_2 = n_2 - (k + 1)$  stopniami swobody.

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2}$$

Próba podzielona jest na dwa zbiory i badana jest równość wariancji w podpróbach.

$$\hat{s}_1^2 = \frac{e_1^T e_1}{n_1 - (k + 1)}$$

$$\hat{s}_2^2 = \frac{e_2^T e_2}{n_2 - (k + 1)}$$

## 5.9 Stabilność parametrów modelu

Stabilność parametrów modelu, która jest pożądaną cechą, może zostać zweryfikowana przy pomocy testu Chowa.

$H_0$ :  $\alpha = \beta = \gamma$ . Parametry modelu są stabilne.

$H_1$ : Parametry modelu nie są stabilne.

Statystyka testowa  $F$  ma rozkład F-Snedecora-Fishera z  $r_1 = k + 1$  i  $r_2 = n - 2(k + 1)$  stopniami swobody.

$$F = \frac{RSK - (RSK_1 + RSK_2)}{RSK_1 + RSK_2} \frac{n - 2(k + 1)}{k + 1}$$

$RSK$  - resztowa suma kwadratów oszacowanego modelu:

$$y_t = \alpha_0 + \alpha_1 x_{1t} + \dots + \alpha_k x_{kt} + \epsilon_t, t = 1, 2, \dots, n$$

$RSK_1$  - resztowa suma kwadratów oszacowanego modelu:

$$y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt} + \epsilon_{1t}, t = 1, 2, \dots, n_1$$

$RSK_2$  - resztowa suma kwadratów oszacowanego modelu:

$$y_t = \gamma_0 + \gamma_1 x_{1t} + \dots + \gamma_k x_{kt} + \epsilon_{2t}, t = n_1 + 1, n_1 + 2, \dots, n$$

## 5.10 Autokorelacja składnika losowego

Autokorelacja składnika losowego jest niepożądaną cechą modelu ekonometrycznego. Występowanie zjawiska autokorelacji pierwszego rzędu można badać przy założeniu, że:

$$e_t = \rho e_{t-1} + \eta_t$$

Służy do tego test mnożnika Lagrange'a autokorelacji składnika losowego.

$H_0$ :  $\rho = 0$ . Zjawisko autokorelacji I rzędu nie występuje.

$H_1$ :  $\rho \neq 0$ . Zjawisko autokorelacji I rzędu występuje.

Szacowany jest model pomocniczy i obliczany jest jego współczynnik determinacji.

$$e_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + \beta_{k+1} e_{t-1} + \mu_t$$

Dla dużej próby ( $n > 30$ ) statystyka testowa  $LM$  ma rozkład  $\chi^2$  z jednym stopniem swobody.

$$LM = (n - 1)R_{e_t}^2$$

## 6 Korekty

### 6.1 Korekty stabilności

W związku z rozważaniem szerokiej klasy funkcji opartych o rozszerzony zbiór danych, a także dużą liczbę zmiennych objaśniających, korekty stabilności postaci modelu i stabilności parametrów nie są używane w projekcie.

### 6.2 Korekta heteroskedastyczności i autokorelacji

Do usunięcia heteroskedastyczności i autokorelacji z modelu ekonometrycznego można użyć uogólnionej metody najmniejszych kwadratów.

Estymator parametrów modelu:

$$a = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

Estymator wariancji składnika losowego:

$$S^2 = \frac{e^T \Omega^{-1} e}{n - (k + 1)}$$

Estymator macierzy kowariancji estymatorów:

$$\hat{D}^2(a) = S^2 (X^T \Omega^{-1} X)^{-1}$$

Macierz  $\Omega$  nie jest znana, dlatego do korekt używa się estymatorów:

Korekta heteroskedastyczności dla  $\sigma^2 = 1$  oraz estymatora  $\sigma_t^2 = e_t^2$ :

$$\Omega_h^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{bmatrix}$$

Korekta autokorelacji I rzędu dla znanego(wyestymowanego)  $\rho$ :

$$\Omega_a^{-1} = \frac{1}{1 - \rho^2} \begin{bmatrix} 1 & -\rho & \dots & 0 \\ -\rho & 1 + \rho^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Korektę łączną zapisać można jako(  $\rho$  wyestymowane po korekcie heteroskedastyczności):

$$\Omega_{ha}^{-1} = \frac{1}{1 - \rho^2} \Omega_h^{-1} \Omega_a^{-1}$$



## 7 Wybrane modele

### 7.1 Wybór modelu

Dla celów wyboru modelu klasa funkcji została rozszerzona o modele z korektą heteroskedastyczności, korektą autokorelacji oraz korektą łączną, co zwiększyło ilość rozważanych modeli do  $4K = 4 * 748 = 2992$ . Poprawność każdego rozważanego modelu została zweryfikowana. Pełne sprawozdanie z estymacji i weryfikacji wszystkich modeli znajduje się w załączniku do pracy.

### 7.2 Model prognostyczny

### 7.3 Postać modelu

$$\hat{Y} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$$

### 7.4 Wyestymowane parametry modelu

$$\alpha_0 = -0.1497101818473765$$

$$\alpha_1 = 0.9388879776939478$$

$$\alpha_2 = 0.08354509891505879$$

$$\alpha_3 = 0.058522677703370385$$

### 7.5 Wskaźniki jakości modelu

Współczynnik determinacji  $R^2 = 0.9961196927344912$

Średni absolutny błąd prognozy *ex post*  $MAE = 0.15393297715304613$

### 7.6 Weryfikacja poprawności modelu

Weryfikacja poprawności modelu przy poziomie istotności  $\alpha = 0.05$

#### 7.6.1 Koincydencja

$$sgn(\alpha_1) = 1$$

$$sgn(r_1) = 1$$

Koincydencja.

$$sgn(\alpha_2) = 1$$

$$sgn(r_2) = 1$$

Koincydencja.

$$sgn(\alpha_3) = 1$$

$$\operatorname{sgn}(r_3) = -1$$

Brak koincydencji.

### 7.6.2 Katalizatory

Zmienna  $X_3$  jest katalizatorem w parze ( $X_3$ ,  $X_1$ )

### 7.6.3 Współliniowość zmiennych

Zmienna  $X_1$  w zależności od reszty zmiennych -  $R^2 = 0.05194183211424597$ .  
Nie występuje współliniowość.

Zmienna  $X_2$  w zależności od reszty zmiennych -  $R^2 = 0.23089677920262042$ .  
Nie występuje współliniowość.

Zmienna  $X_3$  w zależności od reszty zmiennych -  $R^2 = 0.23534537608084938$ .  
Nie występuje współliniowość.

### 7.6.4 Normalność rozkładu reszt

$$JB = 0.14543929209237996$$

$$\chi_{0.05,2}^2 = 5.991464547107983$$

Reszty mają rozkład normalny.

### 7.6.5 Istotność zmiennych objaśniających

$$t_{\alpha_1} = -2.5679436530465547$$

$$t_{0.05,6} = 1.9431802803927818$$

Zmienna  $X_1$  jest statystycznie istotna.

$$t_{\alpha_2} = 1541.529777440913$$

$$t_{0.05,6} = 1.9431802803927818$$

Zmienna  $X_2$  jest statystycznie istotna.

$$t_{\alpha_3} = 54.916346321746126$$

$$t_{0.05,6} = 1.9431802803927818$$

Zmienna  $X_3$  jest statystycznie istotna.

$$t_{\alpha_4} = 38.66532085339449$$

$$t_{0.05,6} = 1.9431802803927818$$

Zmienna  $X_4$  jest statystycznie istotna.

#### 7.6.6 Istotność współczynnika determinacji

$$F = 513.4231000667284$$

$$F_{0.05,3,6} = 4.757062663089414$$

Współczynnik determinacji  $R^2$  jest statystycznie istotny.

#### 7.6.7 Liniowość postaci modelu

$$Z = -0.5619514869490162$$

$$k_{0.05,0,1} = -0.5619514869490162$$

Postać modelu jest liniowa.

#### 7.6.8 Stabilność parametrów modelu

$$F = 1.898250974063121$$

$$F_{0.05,4,2} = 19.246794344808947$$

Parametry modelu są stabilne.

#### 7.6.9 Homoskedastyczność

$$F = 0.19692985596934573$$

$$F_{0.05,1,1} = 161.44763879758827$$

Model jest homoskedastyczny.

#### 7.6.10 Autokorelacja czynnika losowego I rzędu

$$LM = 0.34237529810662015$$

$$\chi^2_{0.05,1} = 3.8414588206941285$$

W modelu nie występuje autokorelacja czynnika losowego I rzędu.

#### 7.6.11 Wynik weryfikacji poprawności modelu

8/10 testów poprawności modelu dało wynik pozytywny. Model nie jest poprawny.

7.7 Model analityczny

8 Prognoza

9 Interpretacja

10 Podsumowanie

## 11 Spis tabel

1	Statystyki - zmienna objaśniana . . . . .	7
2	Korelacja zmiennej objaśnianej ze zmiennymi objaśnianymi . . .	8

## 12 Spis rysunków

1	Histogram zmiennej objaśnianej . . . . .	7
2	Histogram korelacji zmiennej objaśnianej ze zmiennymi objaśnia- nymi . . . . .	8
3	Macierz korelacji . . . . .	9

## 13 Literatura

- [1] Louis Bachelier. *Théorie de la Spéculation*. 1900.
- [2] Capgemini. *High Frequency Trading: Evolution and the Future*. 2012.
- [3] Nicolas Carbone. *200+ Financial Indicators of US stocks (2014-2018)*. 2019.
- [4] Alfred Cowles. *Can stock market forecasters forecast?* 1932.
- [5] L.P. Graham Capital Management. *Systematic Global Macro: Performance, Risk and Correlation Characteristics*. 2013.
- [6] Bogumił Kamiński. *Teoria uczenia statystycznego z perspektywy ekonometriki*. 2017.
- [7] Tongda Zhang Shunrong Shen Haomiao Jiang. *Stock Market Forecasting Using Machine Learning Algorithms*. 2012.
- [8] Zhiping Lin Zhaoxia Wang Seng-Beng Ho. *Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiment*. 2018.