UNIWERSYTET JAGIELLOŃSKI
WYDZIAŁ MATEMATYKI I INFORMATYKI
INSTYTUT MATEMATYKI

Daniel Biernat

Wyjaśnienie zmian cen akcji spółki KGHM za pomocą głównych czynników ryzyka

Kraków 2022

Spis treści

1. Wstęp

Celem modelu będzie wyjaśnienie zmienności cen akcji spółki KGHM. Ideą stojącą za wyborem tematu było uogólnienie modelu $CAPM$ poprzez uwzględnienie czynników ryzyka charakterystycznych dla spółki.

2. Dobór oraz obróbka zmiennych

2.1. Przedstawienie zmiennych źródłowych

Dane, na których budowany będzie model są cenami zamknięcia w poszczególnych dniach z okresu od 13 maja 2021 roku do 12 maja 2022 roku. Zbiór danych składa się z 261 obserwacji. Wszystkie dane pochodzą ze strony agregującej dane giełdowe $https://stooq.pl/$.

Zbiór danych składa się z następujących zmiennych:

- **Cena akcji KGHM** – cena akcji skorygowana o wypłatę dywidendy. Zgodnie z celem modelu, zmienna objaśniana będzie przekształceniem tej zmiennej.
- **Wartość indeksu WIG20** – zgodnie z modelem $CAPM$, zmienna ta odpowiedzialna będzie za powiązanie zmian cen spółki KGHM ze zmianami rynku.
- **Kurs dolara** – zmienna ta pozwoli uwzględnić ryzyko walutowe.
- **Cena miedzi w dolarach za tonę** – miedź jest głównym surowcem wydobywanym przez spółkę. Celem zmiennej będzie uwzględnienie ryzyka związanego ze zmianą ceny surowca będącego głównym źródłem przychodu spółki.
- **Cena srebra w dolarach za uncję** – srebro, zaraz obok miedzi, jest głównym surowcem wydobywanym przez spółkę. Celem zmiennej będzie uzupełnienie zmiennej opisującej ceny miedzi, aby uzyskać lepszy wgląd w zależność ceny akcji spółki od rynku surowców.
- **Zmienna logiczna opisująca czy dany dzień był poniedziałkiem** – celem zmiennej jest uwzględnienie w modelu zwiększonej zmienności cen w dniach, które poprzedzał dzień bez notowań. Jest to zmienna kategoryczna.

2.2. Przekształcenie danych oraz operacje związane z brakami w danych

Zbiór danych zawiera 8 braków notowań cen akcji spółki. Są to dni świąteczne, w których giełda jest zamknięta. Obserwacje te usuwam, po uprzednim oznaczeniu dni następujących jako dni wznowienia notowań.

Po powyższej operacji w zbiorze danych znajduje się 6 obserwacji z brakiem notowań cen miedzi. Braki te uzupełniam stosując interpolację liniową.

Ceny miedzi oraz srebra przekształcam, przemnażając je przez kurs dolara, aby uzyskać zgodność waluty notowań surowców z cenami akcji spółki.

Wszystkie zmienne numeryczne przekształcam obliczając proste stopy zwrotu w celu redukcji autokorelacji obserwacji oraz uzyskania analogii do modelu $CAPM$.

Po powyższych operacjach otrzymujemy zbiór danych składa się z 252 obserwacji, o następujących zmiennych:

- **KGH** – dzienne proste stopy zwrotu z akcji spółki KGHM. Będzie to zmienna objaśniana. Zaobserwowane wartości należą do przedziału od -0.067 do 0.093, a ich średnia wynosi -0.002.
- **W20** – dzienne proste stopy zwrotu z indeksu WIG20. Zaobserwowane wartości należą do przedziału od -0.108 do 0.084, a ich średnia wynosi -0.0007.
- **USD** – dzienne proste stopy zwrotu dolara. Zaobserwowane wartości należą do przedziału od -0.042 do 0.029, a ich średnia wynosi 0.0007.
- **COPP** – dzienne proste stopy zwrotu miedzi w złotówkach. Zaobserwowane wartości należą do przedziału od -0.061 do 0.052, a ich średnia wynosi 0.0002.
- **SLVR** – dzienne proste stopy zwrotu srebra w złotówkach. Zaobserwowane wartości należą do przedziału od -0.065 do 0.055, a ich średnia wynosi -0.0002.
- **modnday** – zmienna kategoryczna o poziomach $TRUE$ oraz $FALSE$ identyfikująca, czy dzień poprzedzający daną obserwacje był wolny od notowań giełdowych na Giełdzie Papierów Wartościowych w Warszawie. Zmienna ta przyjmuje poziom $TRUE$ dla 56 obserwacji oraz poziom $FALSE$ dla 196 obserwacji.

3. Modele pośrednie

3.1. Model jednoczynnikowy

3.1.1. Przedstawienie modelu

W pierwszym z modeli rozważać będziemy zależność

$$KGH_t = \beta_{W20} \cdot W20_t + \varepsilon_t.$$

Model ten służył będzie za punkt odniesienia, względem którego porównywane będą późniejsze modele.

3.1.2. Podstawowe statystyki modelu

Estymator współczynnika β_{W20} wynosi 1.04, z 95− procentowym przedziałem ufności $(0.87, 1.21)$.

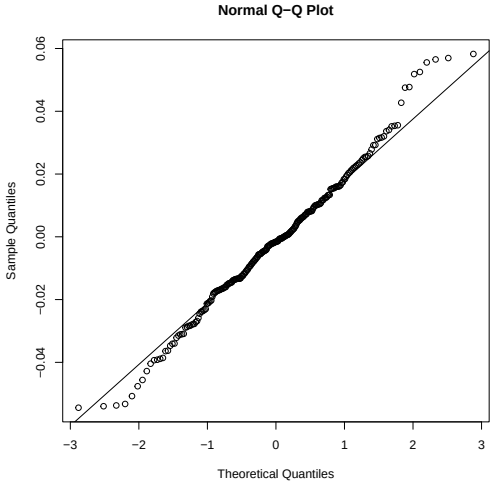
Skorygowany współczynnik R^2 modelu wynosi 0.37, a błąd standardowy residuów wynosi 0.022. Celem porównania modelu z objaśnieniem zmiennej KGH za pomocą jedynie wyrazu wolnego wykonałem test F otrzymując p-value mniejsze niż 10^{-16}, a zatem model ten jest istotnie lepszy od objaśniania zmiennej KGH przez wartość średnią. Wykonując test t o hipotezie zerowej $\beta_{W20} = 0$ otrzymałem p-value mniejsze niż 10^{-16}, a zatem współczynnik ten jest istotnie różny od 0.

Badając model rozszerzony o wyraz wolny, test t sprawdzający hipotezę o niezerowości współczynnika przy wyrazie wolnym otrzymałem p-value wynoszące 0.29, a zatem współczynnik przy wyrazie wolnym nie jest istotnie

różny od zera. Analogiczny test dla zmiennej β_{W20} uzyskuje p-value mniejsze niż 10^{-16}. Z tego powodu odrzucam rozszerzenie modelu.

3.1.3. Weryfikacja założeń

W celu zbadania normalności residuów modelu wykonałem testy $Shapiro-Wilka$, $Jarque-Bera$ oraz $Andersona-Darlinga$ otrzymując kolejno p-value : 0.11, 0.39 oraz 0.22. Test $Shapiro-Wilka$ oparty jest na ważonej sumie przyrostów ciągu posortowanych obserwacji unormowanych przez ich wariancję. Test $Jarqe-Bera$ oparty jest na empirycznym czwartym momencie rozkładu, a test $Andersona-Darlinga$ na scałkowanym kwadracie różnic dystrybuant. Wszystkie testy badają hipotezę zerową o normalności rozkładu z zaprzeczeniem hipotezy zerowej jako hipotezą alternatywną. Żaden z testów nie identyfikuje braku spełnienia założenia o normalności reszt.

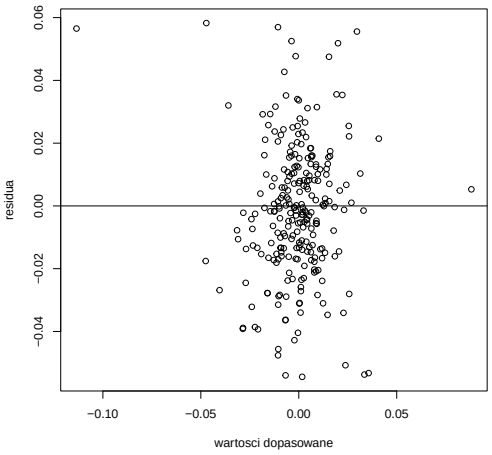


Rysunek 1. Wykres kwantylowy

Na rysunku nr 1 naniesiono kwantyle rozkładu residuów w zależności od teoretycznych kwantyli rozkładu normalnego. Analiza graficzna wykresu pozwala zauważyć odstępstwo od założenia o normalności w prawym ogonie rozkładu, jednakże nie jest ono duże.

W celu zbadania autokorelacji reszt modelu przeprowadziłem test $Durbina-Watsona$, w którym hipoteza zerowa zakłada, że współczynnik autokorelacji rzędu 1 wynosi 0, z hipotezą alternatywną będącą zaprzeczeniem hipotezy zerowej. Statystyka testowa oparta jest na unormowanej sumie kwadratów przyrostów residuów. Otrzymane p-value wynosi 0.20, co nie daje podstaw do odrzucenia hipotezy o braku autokorelacji rzędu 1.

Badając homoskedastyczność reszt za pomocą testu $Goldfelda-Quandta$, w którym w hipotezie zerowej zakłada się równość wariancji na pierwszej i drugiej połowie danych, wobec przeciwnej hipotezy alternatywnej, uzyskałem p-value wynoszące 0.02, co może sugerować problem z założeniem o stałej wariancji residuów.



Rysunek 2. Residua vs wartości dopasowane

Analiza zależności residuów i wartości dopasowanych na rysunku nr 2, wskazuje na zwiększoną wariancję residuów dla wartości dopasowanych odbiegającej od średniej. W celu przetestowania tej hipotezy wykonałem test F, w którym hipoteza zerowa zakłada równość wariancji na dwóch zbiorach, wobec przeciwnej hipotezy alternatywnej. Zbiór danych podzieliłem na środkowe 60% obserwacji ze względu na wartości dopasowane, oraz pozostałe 40% obserwacji. Otrzymane p-value wynosi 0.003 co potwierdza graficzną analizę wykresu.

W celu zbadania zależności liniowej w modelu wykonałem test $rainbow$, polegający na zbudowaniu modelu na środkowej części danych ze względu na porządek chronologiczny i porównaniu go z modelem pełnym. Test ten osiąga p-value 0.014, co może wskazywać na nieliniową zależność między zmienną objaśnianą a zmiennymi objaśniającymi. Nie podejmuję jednak próby przekształcania zmiennych, aby zachować analogię do modelu $CAPM$.

3.1.4. Podsumowanie

Model jednoczynnikowy nie spełnia wszystkich założeń standardowego modelu regresji liniowej. Model ten jest jednak dobrym punktem odniesienia dla bardziej skomplikowanych modeli. Celem budowy dalszych modeli będzie uzyskanie lepszego objaśnienia zmiennej KGH oraz uzyskanie modelu bliższego założeniom standardowej regresji liniowej.

3.2. Model Prosty

3.2.1. Przedstawienie modelu

Analiza modelu ze wszystkimi zmiennymi objaśniającymi $KGH_t = \beta_0 + \beta_1 \cdot USD_t + \beta_2 \cdot SLVR_t + \beta_3 \cdot COPP_t + \beta_4 \cdot W20_t + \beta_4 \cdot \chi_{monday_t=TRUE} + \varepsilon_t$, prowadzi do odrzucenia zmiennej $monday$ ze względu na największą wartość p-value w t teście. Otrzymane p-value testów t wynoszą kolejno 0.31, $8 \cdot 10^{-4}$, $6 \cdot 10^{-6}$, $7 \cdot 10^{-11}$, $2 \cdot 10^{-16}$, 0.64.

Po odrzuceniu zmiennej $monday$ i ponownym wykonaniu testów t otrzymujemy wartości p-value : 0.36, $8 \cdot 10^{-04}$, $5 \cdot 10^{-06}$, $7 \cdot 10^{-11}$, $8 \cdot 10^{-22}$. Wynik ten prowadzi do odrzucenia wyrazu wolnego ze względu na brak podstaw do odrzucenia hipotezy o jego zerowości.

Otrzymujemy w ten sposób model, który w dalszej części nazywać będziemy modelem prostym. W modelu tym badana jest następująca zależność liniowa:

$$KGH_t = \beta_1 \cdot USD_t + \beta_2 \cdot SLVR_t + \beta_3 \cdot COPP_t + \beta_4 \cdot W20_t + \varepsilon_t.$$

3.2.2. Podstawowe statystyki modelu

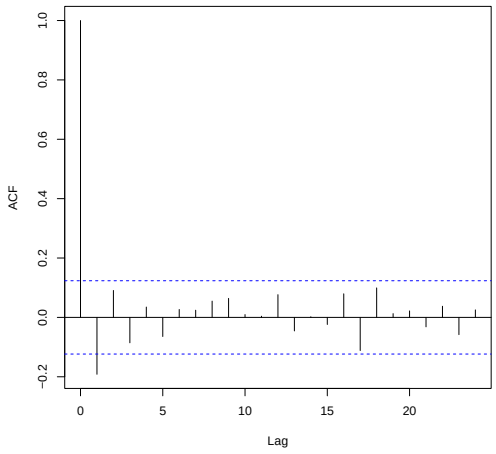
Współczynniki modelu wraz z 95−procentowymi przedziałami ufności wynoszą odpowiednio: $\beta_1 = -0.73 \pm 0.41$,$\beta_2 = 0.41 \pm 0.17$, $\beta_3 = 0.63 \pm 0.18$, $\beta_4 = 0.89 \pm 0.16$.

Odchylenie standardowe residuów wynosi 0.01875, a skorygowany współczynnik R^2 modelu wynosi 0.52.

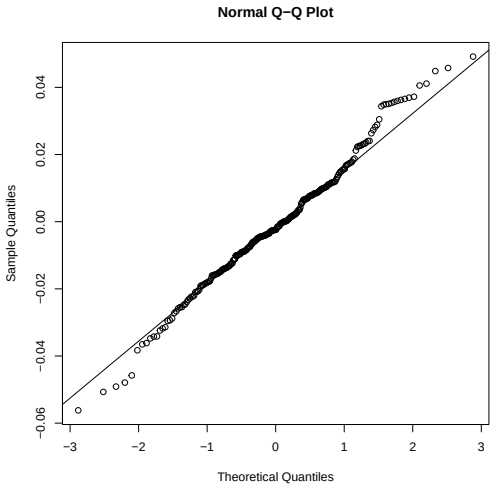
Test F dla modelu daje p-value mniejsze niż 10^{-16}, a zatem model ten istotnie lepiej wyjaśnia zmienność KGH niż objaśnianie jej przez wartość średnią.

3.2.3. Weryfikacja założeń

Analizując rysunek nr 3, na którym znajdują się estymatory autokorelacji kolejnych rzędów, można zauważyć istotną autokorelacje rzędu 1. W celu weryfikacji, czy autokorelacja ta jest istotnie różna od zera przeprowadzam test $Durbina - Watsona$, otrzymując p-value 0.0032. Wynik ten identyfikuje istotne odstępstwo od założenia klasycznego modelu regresji liniowej o warunkowej niezależności reszt. Estymator współczynnika autokorelacji rzędu 1 wynosi -0.19. Ze względu na jego małą wartość nie odrzucam modelu. Rozwiązanie problemu autokorelacji polegające na uwzględnieniu obserwacji z dnia poprzedniego istotnie komplikuje model, przez co staje się on trudny w interpretacji.



Rysunek 3. Autokorelacja residuów

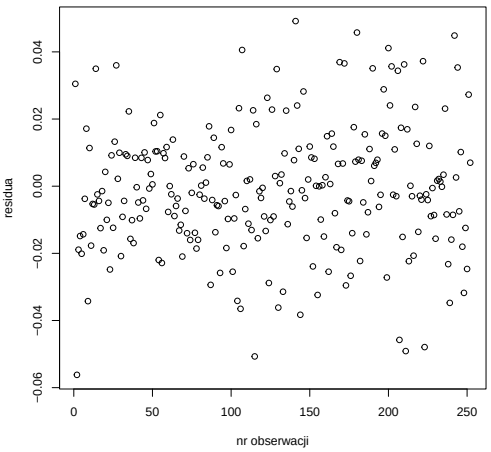


Rysunek 4. Wykres kwantylowy

Analiza graficzna rysunku nr 4, na którym znajduje się wykres kwantylowy reszt modelu nie identyfikuje znaczących problemów z normalnością residuów. Zauważmy, że pomimo występowania odstępstwa w prawym ogonie rozkładu, problem ten jest istotnie mniejszy niż w modelu jednoczynnikowym (Rysunek nr 1).

Testy $Shapiro - Wilka$, $Andersona - Darlinga$ oraz $Jarque - Bera$ dają p-value odpowiednio: $0.09, 0.04, 0.49$ co może wskazywać na odstępstwo od założenia o normalności residuów.

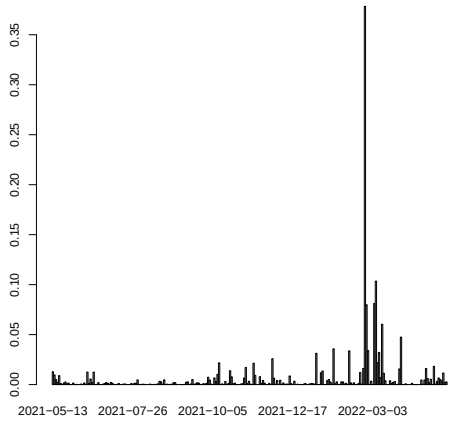
Analiza rysunku nr 5, na którym znajdują się wartości residuów kolejnych obserwacji pozwala zauważyć zmniejszona wariancję residuów dla początkowych obserwacji.



Rysunek 5. Residua w czasie

Potwierdza to test $Goldfelda - Quandta$, w którym hipoteza zerowa zakłada równość wariancji na połowach danych dobranych chronologicznie. Test daje p-value wynoszące 0.0086. Przeprowadzenie testu na zbiorze danych z pominięciem pierwszych stu obserwacji daje p-value na poziomie 0.20, co sugeruje, że problem występuje dla obserwacji początkowych.

W celu zbadania wpływowości obserwacji na model, dla każdej z obserwacji wyliczam odległość Cooka zdefiniowaną jako sumę kwadratów różnic wartości dopasowanych oraz wartości dopasowanych po usunięciu obserwacji, unormowaną przez estymator wariancji reszt modelu przemnożony przez liczbę zmiennych objaśniających. Analizując rysunek nr 6, na którym znajduje się odległość Cooka dla poszczególnych obserwacji, istotnie wyróżnia się jedna z nich. Są to dane z dnia 24 lutego 2022, w którym to zaszły gwałtowne zmiany wycen aktywów spowodowane rozpoczęciem wojny.



Rysunek 6. Odległość Cooka

3.2.4. Podsumowanie

Odstępstwa od założeń standardowej regresji liniowej opisane w podrozdziale 3.2.3 oraz dominacja modelu przez pojedynczą obserwacje są istotnymi wadami, jednakże na korzyść modelu świadczy poprawa skorygowanej statystyki R^2 oraz odchylenia standardowego residuów względem modelu jednoczynnikowego.

4. Modele na zawężonym zbiorze danych

Problemy z normalnością residuów oraz homoskedastycznością modeli opisanych w rozdziale 3 sugerują, że badana zależność liniowa może być

zmienna w czasie. Z tego powodu konstruuje nowe, analogiczne modele na zawężonym zbiorze obserwacji.

Ze zbioru danych odrzucam dane z dnia 24 lutego 2022. Zmienność wyceny akcji spółki KGHM w tym dniu niekoniecznie była spowodowana czynnikami branymi pod uwagę w modelu. Dźwignia tej obserwacji wynosi 0.1903 wobec średniej 0.0079, a zatem przewyższa średnią ponad 20-krotnie.

Ponadto, konstruuje nowy zbiór danych składający się z ostatnich 150 obserwacji. Taka operacja pozwoli zredukować wpływ potencjalnych zmian zależności między wycenami w czasie.

4.1. Model Jednoczynnikowy na zawężonym zbiorze danych

4.1.1. Przedstawienie modelu

Model ten bada zależność

$$KGH_t = \beta_{W20} \cdot W20_t + \varepsilon_t$$

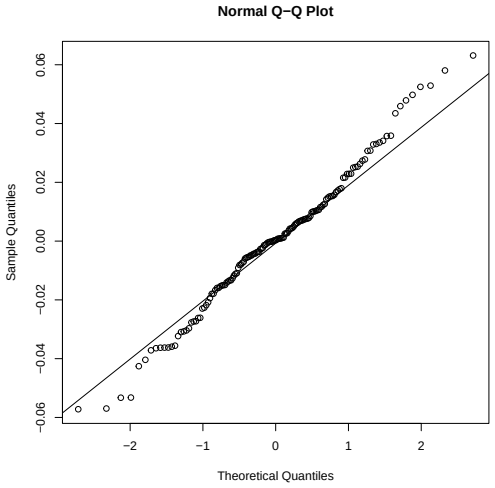
na ostatnich 150 obserwacjach po uprzednim pominięciu danych z dnia 24 lutego 2022.

4.1.2. Podstawowe statystyki modelu

Estymator współczynnika β_{W20} wraz z 95−procentowym przedziałem ufności wynosi 1.15 ± 0.23, a test t daje p-value mniejsze niż 10^{-16}. Test F względem modelu objaśniającego jedynie za pomocą wyrazu wolnego daje p-value również mniejsze niż 10^{-16}. Test t dla wyrazu wolnego w modelu $KGH_t = \beta_0 + \beta_{W20} \cdot W20_t + \varepsilon_t$ daje p-value wynoszące 0.89, a zatem odrzucam rozszerzenie modelu.

Skorygowany współczynnik R^2 modelu wynosi 0.40, a odchylenie standardowe residuów 0.023.

4.1.3. Weryfikacja założeń



Rysunek 7. Wykres kwantylowy

Testy normalności dla residuów $Shapiro - Wilka$, $Jarque - Bera$ oraz $Andersona - Darlinga$ dają kolejno p-value : 0.30, 0.86 oraz 0.12. Wynik ten nie sugeruje problemów z założeniem o normalności residuów. Analiza wykresu kwantylowego znajdującego się na rysunku nr 7 pozwala zauważyć pewne odstępstwa w ogonach, jednakże nie są one na tyle duże, aby podjąć decyzję o odrzuceniu modelu.

W celu zbadania autokorelacji residuów przeprowadzam test $Durbina - Watsona$, otrzymując p-value wynoszące 0.07. Estyma-

tor współczynnika autokorelacji rzędu jeden wynosi -0.14. Wynik ten nie identyfikuje istotnych problemów z autokorelacją residuów.

Testując homoskedastyczność residuów za pomocą testu $Goldfelda - Quandta$ otrzymałem p-value wynoszące 0.15. Test F badający stałość wariancji na zbiorze 60% środkowych obserwacji ze względu na wartości dopasowane oraz na jego dopełnieniu otrzymałem p-value wynoszące 0.03, co może sugerować problemy ze stałością wariancji oraz nie stanowi znaczącej poprawy względem modelu budowanego na pełnym zbiorze danych.

Test *rainbow* dla modelu daje p-value wynoszące 0.51 wobec 0.014 analogicznego modelu na pełnym zbiorze danych. Wynik ten stanowi znaczącą poprawę oraz świadczy na korzyść przypuszczenia, że badana zależność jest liniowa, ale zmienna w czasie.

W celu zbadania poprawności założenia o relacji liniowej, wykonałem test *reset*, badający model rozszerzony o kolejne potęgi zmiennych objaśniających, uzyskując p-value na poziomie 0.87, co nie daje podstaw do odrzucenia hipotezy o zależności liniowej.

Dynamika cen instrumentów finansowych charakteryzuje się wzmożoną zmiennością w dniach wznowienia notowań. Przeprowadziłem test F równości wariancji dla podzbiorów wyznaczanych przez zmienną *monday* w celu sprawdzenia odporności modelu na to zjawisko. Uzyskałem w ten sposób p-value wynoszące 0.34, co nie sugeruje odstępstwa od homoskedastyczności ze względu na opisany efekt.

4.1.4. Podsumowanie

Pomimo odstępstw od założeń standardowego modelu regresji liniowej, model wykazuje istotną poprawę względem analogicznego modelu przed redukcją danych. Ze względu na swoją prostotę model ten jest dobrym punktem odniesienia, względem którego porównywane będą bardziej rozbudowane modele budowane na zawężonym zbiorze danych.

4.2. Model właściwy

4.2.1. Przedstawienie modelu

Model ten badał będzie zależność liniową

$$KGH_t = \beta_1 \cdot USD_t + \beta_2 \cdot SLVR_t + \beta_3 \cdot COPP_t + \beta_4 \cdot W20_t + \varepsilon_t,$$

na ostatnich 150 obserwacjach po uprzednim pominięciu danych z dnia 24 lutego 2022.

4.2.2. Podstawowe statystyki modelu

Współczynniki modelu wraz z $95-$procentowymi przedziałami ufności wynoszą odpowiednio: $\beta_1 = -0.81 \pm 0.59, \beta_2 = 0.46 \pm 0.25, \beta_3 = 0.62 \pm 0.26, \beta_4 = 1.01 \pm 0.23$.

Odchylenie standardowe residuów wynosi 0.02059, a skorygowany współczynnik R^2 modelu wynosi 0.54.

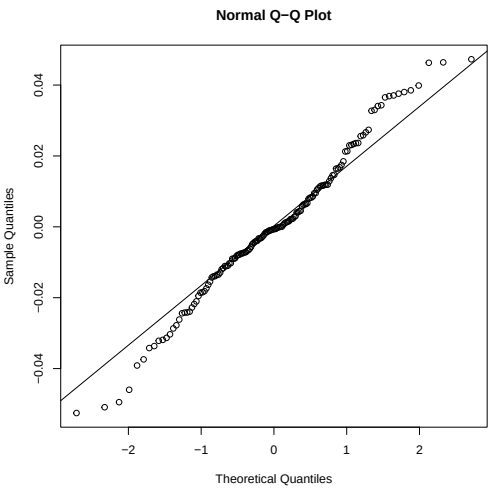
Test F dla modelu daje p-value mniejsze niż 10^{-16}, a zatem model ten istotnie lepiej wyjaśnia zmienność KGH niż objaśnianie jej przez wartość średnią.

Wykonując testy t dla kolejnych zmiennych objaśniających otrzymuje p-value kolejno: $0.008, 0.0004, 6 \cdot 10^{06}, 3 \cdot 10^{-15}$, a zatem wszystkie zmienne objaśniające w modelu są istotne statystycznie.

Rozszerzenie modelu kolejno o wyraz wolny oraz zmienną *monday* nie identyfikuje tych zmiennych jako istotne statystycznie. Testy t dają p-value odpowiednio 0.997 i 0.40.

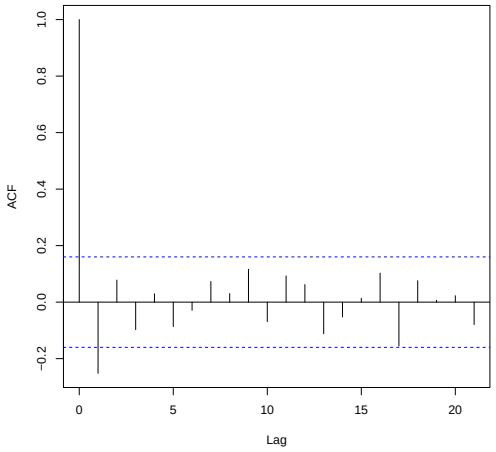
4.2.3. Weryfikacja założeń

W celu zbadania normalności residuów wykonałem testy $Jarque-Bera$, $Shapiro-Wilka$ oraz $Andersona-Darlinga$ otrzymując p-value kolejno: $0.98, 0.16$ oraz 0.07, co nie wskazuje na brak spełnienia założenia o normalności. Analiza wykresu kwantylowego znajdującego się na rysunku nr 8 wskazuje na pewne odstępstwa od normalności w ogonach, jednakże nie są one na tyle alarmujące, aby odrzucić hipotezę o normalności residuów.



Rysunek 8. Wykres kwantylowy

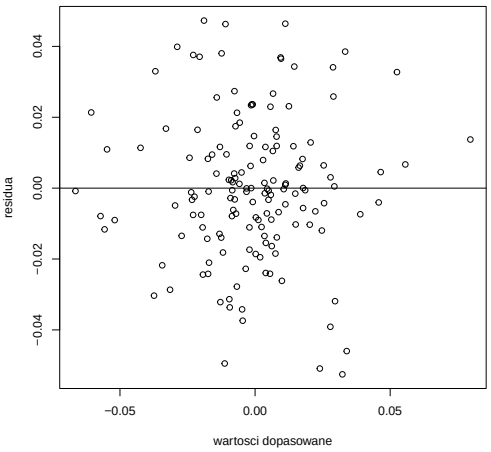
Analiza wykresu autokorelacji znajdującego się na rysunku nr 9 wskazuje na istotnie różną od zera autokorelacje rzędu 1, co potwierdza test $Durbina - Watsona$, w którym p-value wynosi 0.0017. Ze względu na niską wartość estymatora współczynnika autokorelacji rzędu 1, który wynosi -0.25, nie odrzucam modelu, jednakże należy zaznaczyć, że to odstępstwo od założeń może wpłynąć na model.



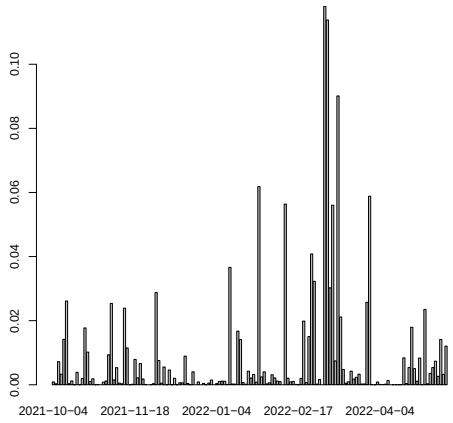
Rysunek 9. Wykres autokorelacji

W celu zbadania słuszności założenia o istnieniu zależności liniowej wykonałem testy $reset$, $Breuscha - Pagana$ oraz $rainbow$ otrzymując kolejno p-value: 0.67, 0.16 oraz 0.64 co nie sugeruje problemów ze strukturą modelu.

Analiza rysunku nr 10, na którym znajduje się wykres residuów w zależności od wartości dopasowanych nie przeczy homoskedastyczności residuów. Podobnie jak w modelu jednoczynnikowym dla pełnego zbioru danych można zauważyć zwiększoną wariancję dla środkowych obserwacji, jednakże wykonanie testu F równości wariancji na podzbiorze środkowych 60% obserwacji oraz jego dopełnieniu daje p-value 0.22 co nie daje podstaw do odrzucenia hipotezy o stałości wariancji. W celu zbadania poprawy względem analogicznego modelu na pełnym zbiorze danych wykonałem test $Goldfelda - Quandta$ otrzymując p-value 0.21 wobec 0.0086 przed redukcją zbioru danych. Test F równości wariancji dla podzbiorów wyznaczonych przez zmienną *monday* daje p-value wynoszące 0.33 co nie identyfikuje negatywnego wpływu efektu poniedziałku na model.



Rysunek 10. Residua vs wartości dopasowane

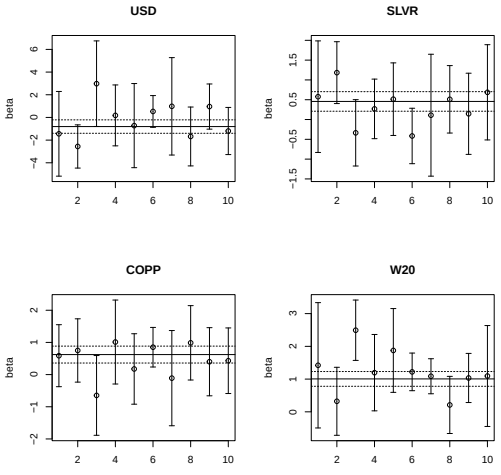


Rysunek 11. Odległość Cooka

Analiza rysunku nr 11, na którym znajduje się odległość Cooka dla poszczególnych obserwacji pozwala zauważyć, że pomimo zróżnicowania wpływu na model pomiędzy obserwacjami żadna z nich nie dominuje modelu.

W celu zbadania założenia o liniowej niezależności zmiennych objaśniających obliczam pierwiastek wartości bezwzględnej stosunku największej oraz najmniejszej wartości własnej macierzy zmiennych objaśniających. Wartość ta wynosi 4.05 co nie identyfikuje problemu z liniową niezależnością.

4.2.4. Analiza stabilności



Rysunek 12. Stabilność estymacji współczynników

W celu zbadania stabilności modelu właściwego podzieliłem jego zbiór danych na 10 rozłącznych 15−to elementowych podzbiorów w sposób niezaburzający porządku czasowego. Na każdym z podzbiorów zbudowałem model analogiczny do właściwego oraz naniosłem na wykresy na rysunku nr 12 wyestymowane wartości współczynników modeli wraz z ich przedziałami ufności. Na wykresach liniami poziomymi ciągłymi zaznaczyłem wyestymowane współczynniki modelu właściwego oraz ich przedziały ufności liniami poziomymi przerywanymi. Większość estymacji na podzbiorach nie odbiega w sposób istotny od estymacji parametrów na całym zbiorze, jednakże parametry wyestymowane na podzbiorze nr 3 wyraźnie odbiegają od parametrów wyestymowanych dla modelu właściwego, co może sugerować problem ze stabilnością modelu.

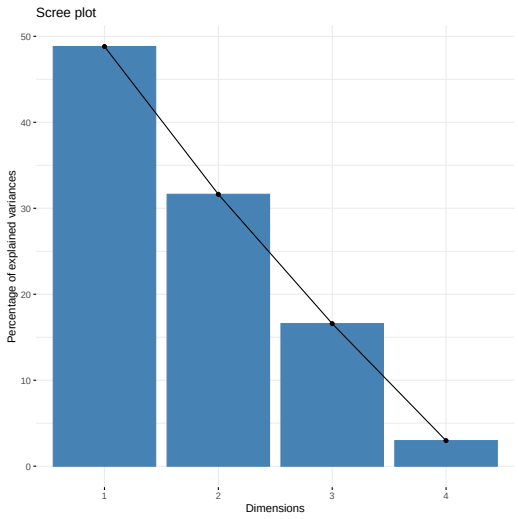
4.2.5. Podsumowanie

Analiza modelu nie sugeruje znaczących odstępstw od założeń standardowego modelu regresji liniowej za wyjątkiem problemu z autokorelacją residuów. Model ten można uznać za rozsądną alternatywę dla modelu jednoczynnikowego na zawężonym zbiorze danych.

4.3. Model PCA

4.3.1. Przedstawienie modelu

W celu zredukowania liczby zmiennych objaśniających w modelu właściwym dokonuje ortogonalizacji PCA otrzymując następujące komponenty składowe:
$PC1 = 0.32 \cdot USD + 0.66 \cdot SLVR + 0.67 \cdot COPP - 0.077 \cdot W20,$
$PC2 = -0.13 \cdot USD - 0.10 \cdot SLVR + 0.27 \cdot COPP + 0.95 \cdot W20,$
$PC3 = 0.08 \cdot USD - 0.72 \cdot SLVR + 0.64 \cdot COPP - 0.24 \cdot W20,$
$PC4 = 0.93 \cdot USD - 0.18 \cdot SLVR - 0.26 \cdot COPP + 0.18 \cdot W20.$



Rysunek 13. Scree plot

Wykres znajdujący się na rysunku nr 13 przedstawia procentowy udział informacji o zmienności danych zawarty w poszczególnych komponentach składowych. Ze względu na fakt, że pierwsze dwa komponenty objaśniają ponad 80% wariancji decyduje się na model $KGH_t = \beta_{PC1} \cdot PC1_t + \beta_{PC2} \cdot PC2_t + \varepsilon_t$, jako alternatywę dla modelu właściwego.

4.3.2. Podstawowe statystyki modelu

Współczynniki modelu wraz z 95−procentowymi przedziałami ufności wynoszą $\beta_{PC1} = 0.38 \pm 0.15$, $\beta_{PC2} = 1.18 \pm 00.20$. Model ten jest równoważny modelowi w bazie kanonicznej postaci

$$KGH_t = -0.025 \cdot USD_t + 0.13 \cdot SLVR_t + 0.57 \cdot COPP_t + 1.09 \cdot W20_t + \varepsilon_t.$$

Współczynniki przy zmiennych USD_t oraz $SLVR_t$ są istotnie różne od współczynników w modelu właściwym. Parametr R^2 modelu wynosi 0.52, a odchylenie standardowe residuów 0.021.

4.3.3. Weryfikacja założeń

W celu zbadania normalności residuów modelu wykonałem testy $Jarque-Bera$, $Shapiro-Wilka$ oraz $Andersona-Darlinga$ otrzymując p-value odpowiednio 0.80, 0.08 oraz 0.04 co może wskazywać na istotne odstępstwa od założeń.

Wykonując test $Durbina-Watsona$ w celu weryfikacji autokorelacji residuów otrzymałem p-value wynoszące 0.0002. Estymator autokorelacji rzędu 1 wynosi -0.30 co stanowi istotnie większe odstępstwo od założeń w porównaniu z poprzednimi modelami.

4.3.4. Podsumowanie

Ze względu na większe odstępstwa od założeń skłaniam się ku odrzuceniu modelu jako istotnie lepszego od modelu właściwego. W celu sprawdzenia słuszności tej decyzji wezmę go pod uwagę w porównaniu modeli.

5. Walidacja krzyżowa

Celem porównania modeli wyznaczam ostatnie 30 obserwacji jako zbiór testowy oraz buduje modele walidacyjne na zbiorze danych z pominięciem

zbioru testowego. Dla każdego z modeli wyznaczam odchylenie standardowe błędów na zbiorze testowym dane wzorem

$$RMSE_{test} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2},$$

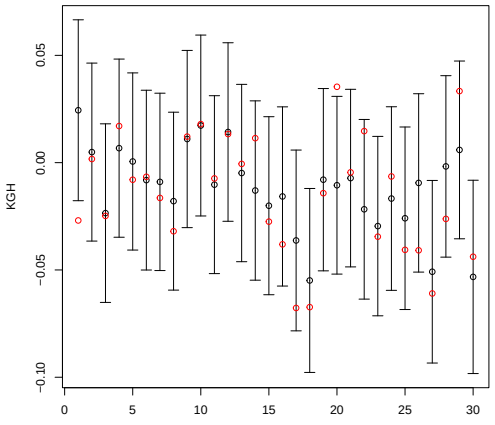
gdzie m oznacza liczność zbioru testowego, y_i oznacza wartość zmiennej objaśnianej w $i-$tej obserwacji zbioru testowego, a $\hat{y}_i$ oznacza predykcję modelu walidacyjnego dla $i-$tej obserwacji zbioru testowego.

model	R^2_{adj}	RMSE	$RMSE_{test}$
jednoczynnikowy na pełnym zbiorze danych	0.37	0.0216	0.0237
prosty na pełnym zbiorze danych	0.52	**0.0188**	0.0201
model jednoczynnikowy na zawężonych danych	0.40	0.0234	0.0233
model właściwy	**0.53**	0.0206	**0.0198**
model PCA	0.51	0.0211	0.0213

Tabela 1. Statystyki modeli

Analiza danych w tabeli 1 pozwala zauważyć istotne pogorszenie parametrów modelu PCA względem modelu właściwego co w połączeniu z większymi odstępstwami od założeń daje podstawy do jego odrzucenia.

Kolejnym wnioskiem płynącym z analizy tabeli 1 jest dominacja modelu prostego oraz właściwego nad modelami jednoczynnikowymi na tych samych zbiorach danych.



Rysunek 14. Wykres predykcji

Na wykresie nr 14 kolorem czarnym zaznaczono predykcje modelu właściwego z walidacji krzyżowej wraz z 95−procentowymi przedziałami ufności. Kolorem czerwonym oznaczono rzeczywiste wartości zmiennej objaśnianej. Analiza wykresu pozwala zauważyć, że wartości predykcji w większości nie odbiegają znacząco od wartości rzeczywistych. Alarmujące natomiast są dwie predykcje, których przedziały ufności nie zawierają wartości zaobserwowanej. Obserwacje te powinny zdarzać się raz na 20 predykcji. Aby zweryfikować istotność tego problemu, przy pomocy komendy *binom.test*, zweryfikowałem hipotezę zerową, że prawdopodobieństwo błędnej predykcji wynosi 5% wobec przeciwnej hipotezy alternatywnej. Otrzymane p-value wynosi 0.0197 co może sugerować słabą moc prognostyczną modelu. Ze względu na objaśniający charakter modelu nie identyfikuje tej wady jako istotnej.

6. Komentarz końcowy

6.1. Uzasadnienie wyboru modelu właściwego

Redukcja problemu homoskedastyczności poprzez zawężenie zbioru obserwacji sugeruje brak stałych zależności liniowych pomiędzy cenami akcji $KGHM$ a ceną miedzi, srebra oraz kursu dolara w długim okresie czasu, jednakże modele budowane na zawężonym zbiorze danych nie wykazują znaczących odstępstw od liniowości. Na korzyść redukcji danych świadczy również poprawa testowego $RMSE$ w walidacji krzyżowej po zawężeniu zbioru obserwacji. Z tych powodów uznaję decyzję o zmniejszeniu zbioru danych za słuszną.

Porównanie modelu właściwego z modelem jednoczynnikowym na zawężonym zbiorze danych świadczy na korzyść modelu właściwego. Model właściwy zachowuje się znacząco lepiej względem wszystkich przeprowadzonych testów statystycznych, za wyjątkiem testów dotyczących autokorelacji. Rozbudowany model uzyskuje estymator współczynnika autokorelacji rzędu 1 wynoszący -0.25 wobec -0.14 dla modelu jednoczynnikowego. Analiza podstawowych statystyk modeli oraz walidacja krzyżowa dokonana w rozdziale 5 wskazuje na dominacje modelu właściwego nad modelem jednoczynnikowym na zawężonym zbiorze danych. Z tych powodów uznaje model właściwy za model lepiej wyjaśniający zmienność ceny akcji spółki niż model jednoczynnikowy.

Model PCA wykazuje większe odstępstwa od założeń standardowego modelu regresji oraz uzyskuje gorsze wyniki w walidacji krzyżowej. Z tego powodu odrzucam go na rzecz modelu właściwego.

6.2. Interpretacja modelu

Współczynnik liniowy zmiennej USD wynoszący 0.81 można interpretować jako średnią zmianę ceny akcji spółki $KGHM$ wyrażoną w procentach, jeśli kurs dolara wzrośnie o 1% przy braku zmian cen miedzi, złota oraz wartości indeksu WIG20. W analogiczny sposób można interpretować pozostałe współczynniki.

Analiza znaków współczynników identyfikuje pozytywny wpływ wzrostu cen miedzi, srebra oraz indeksu WIG20 na cenę akcji spółki oraz negatywny wpływ wzrostu kursu dolara.

6.3. Możliwości modyfikacji modelu

Ze względu na postawioną hipotezę o zmienności współczynników modelu w długim okresie czasu można podjąć próbę budowy modelu na danych z interwałem mniejszym niż dzienny, jednakże dostęp do takich danych jest problematyczny. Wartym uwagi pomysłem modyfikacji modelu byłoby zastosowanie wag dla obserwacji tak, aby obserwacje najnowsze miały największy wpływ na model.