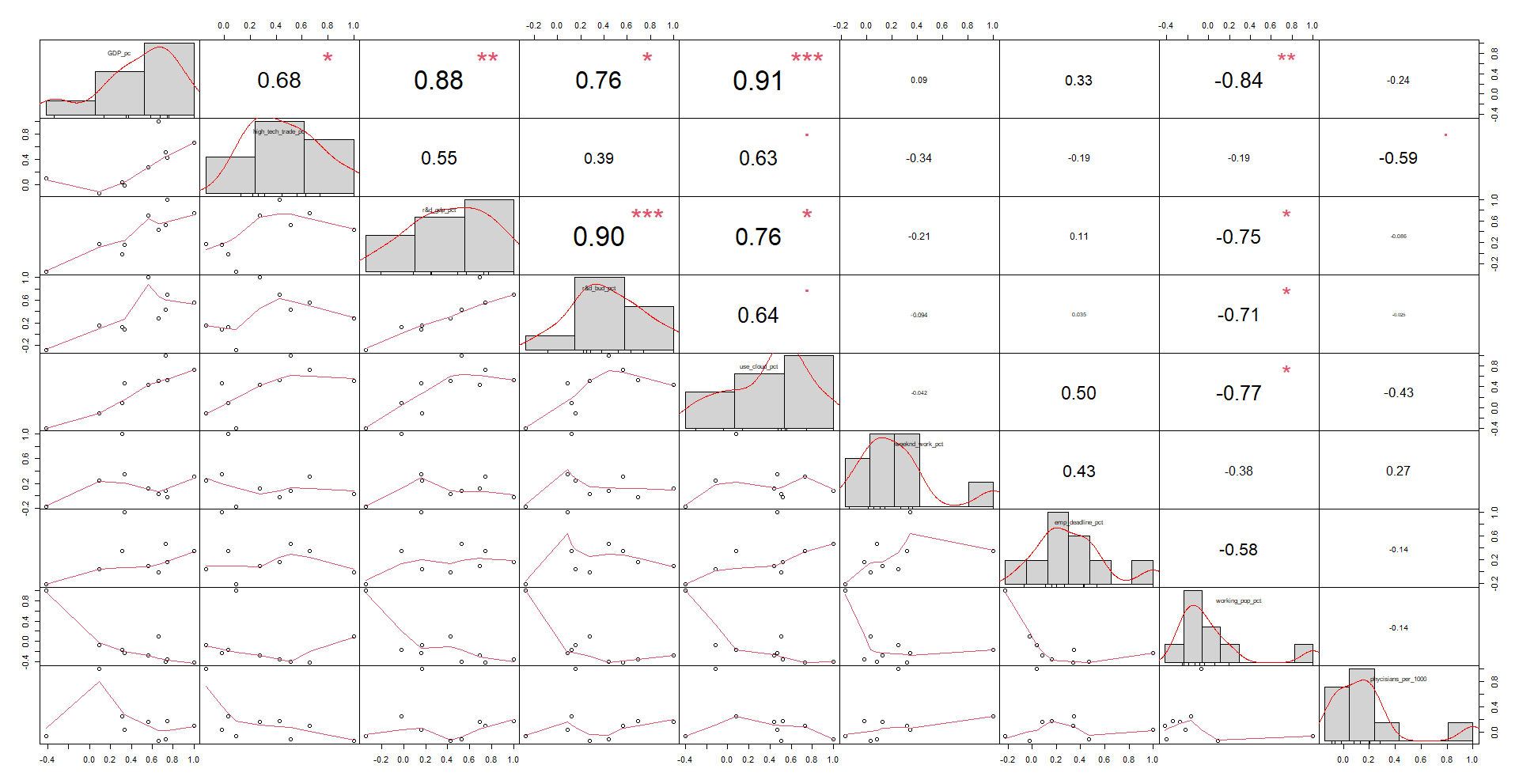
Efekt 6., regresja

Mateusz Nowak, Damian Okoń, Robert Zamiar

# Wstępna analiza zmiennej zależnej – korelacje.

W celu zbadania zależności między zmienną objaśnianą, a obranymi(numerycznymi) predyktorami z utworzonego przez nas zbioru utworzono bazujący na macierzy korelacyjnej – wykres obrazujący zależności.

Rysunek 1 Wykres korelacji między zmiennymi



Rysunek 2Korelacje predyktorów z zmienną objaśnianą - wycinek z macierzy korelacji

Obraz zawierający tekst

Opis wygenerowany automatycznie

Zmienna GDP\_pc jest więc najbardziej skolerowana z zmienną „handel wysokimi technologiami per capita”, wydatkami na badania rozwojowe oraz procentem osób korzystających z technologii cloud’ingowych.

Obraz zawierający stół

Opis wygenerowany automatycznie

Test na istotność korelacji wykazał, że istotnie ze sobą skolerowane ze zmienną objaśnianą są min. Zmienne odpowiedzialne za wydatki rozwojowe(r&d), procent osób korzystających z usług chmurowych czy handel wysokimi technologiami i procent ludzi pracujących. Chcąc jednak bardziej przebadać zależności między zmiennymi, do pierwszego modelu regresji liniowej postanowiono użyć wszystkich predyktorów dostępnych w zbiorze.

# Regresja – budowa modelu

Przy pomocy dodatku RCommander do programu Rstudio stworzono model regresji liniowej.

Rysunek 3Wyniki modelu regresji liniowej

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Predyktorami istotnie wpływającymy na zmienną objaśnianą GDP w utworzonym modelu są: handel wysokimi technologiami, procent PKB przeznaczony na badania rozwojowe oraz procent ludzi pracujących w weekendy – czyli w głównej mierze czynniki gospodarcze. Dopasowanie modelu do danych rzeczywistych wynosi 88%, natomiast duża różnica(prawie 7 punktów procentowych) między współczynnikiem determinacji, a skorygowanym współczynnikiem determinacji wynika z faktu, że do modelu dołożono niepotrzebne predyktory.

Po wstępnej weryfikacji pierwszego modelu regresji liniowej, do kolejnego modelu wybrano 3 istotne zmienne z modelu 1.

Rysunek 4 Wyniki 2. modelu regresji liniowej

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

W tym modelu wszystkie zmienne łącznie z wyrazem wolnym okazały się istotne, zmalał współczynnik determinacji, co jest efektem zmniejszenie liczby predyktorów, zmalała też różnica między R2 a skorygowanym współczynnikiem determinacji – w modelu użyto więc bardziej istotnych zmiennych.

# Weryfikacja modelu

## Autokorelacja reszt

Rysunek 5 Wyniki testu Durbina-Watsona na autokorelację składnika losowego

Obraz zawierający tekst

Opis wygenerowany automatycznie

Pierwszym etapem weryfikacji modelu było sprawdzenie, czy w modelu występuje autokorelacja reszt. Hipotezą zerową w wykonywanym teście jest brak występowania autokorelacji reszt, natomiast alternatywną – występowanie autokorelacji. Ponieważ w naszym przypadku wynik p-value jest wyższy niż 0.05, to nie ma podstaw do odrzucenia hipotezy zerowej. Autokorelacja reszt modelu nie występuje.

Zjawisko autokoleracji w danych makro- lub mikroekonomicznych może być spowodowany niewłaściwym doborem danych(nie względniając np. czynników inflacyjnych) lub kryzysami ekonomicznymi, które mają odzwierciedlenie w danych statystycznych.

## Heteroscedatyczność

Obraz zawierający tekst

Opis wygenerowany automatycznie

Postawiono dwie hipotezy: H0 mówiącą o tym, że heteroskedatyczność reszt nie występuje(wariancja reszt jest stała) oraz H1 o tym, że występuje heteroskedatyczność . Heteroscedatyczność reszt nie występuje – występuje homoskedatyczność. Gdyby heteroskedastyczność występowała nasz model miałby błędne wnioskowanie statystyczne.

## Normalność składnika losowego

Rysunek 6 Wyniki testu Doornika-Hansena

Obraz zawierający tekst

Opis wygenerowany automatycznie

Test na normalność składnika losowego wykonano przy pomocy oprogramowania Gretl – testem tym jest test normalności Doornika-Hansena o hipotezie H0– rozkład składnika losowego posiada rozkład normalny. Ponieważ p-value jest wyższe od poziomu istnotności alfa=0.05, przyjmujemy hipotezę zerową. Rozkład składnika losowego posiada rozkład normalny, a model został właściwie zweryfikowany.

# Podsumowanie

Model przedstawia się równaniem: y(GDP)=-12317+9,33\*high\_tech\_trade\_pc+9004,64\*r.d\_gdp\_pct+516,2\*weekend\_work\_pct.

Równanie modelu oznacza, że prognozowana wartość zmiennej objaśnianej jaką jest GDP\_PC zwiększy się o kolejno: 9.33 euro jeżeli zmienna związana z handlem wysokimi technologiami zwiększy się o jednostkę, 9000,64 euro jeżeli zmienna związana z procentem wydatków rozwojowych państwa zwiększy się o jedną jednostkę(w tym przypadku procent[%]), a wzrost zmiennej związanej z pracą w weekendy o jedną jednostkę zwiększy prognozowaną wartość PKP per capita o 516.2 euro.

Model wyjaśnia rzeczywistość w 79%, co jest wysoką wartością. Wartość ta pozwala na wykorzystanie modelu w celach predykcyjnych – na przykład do przewidywań wartości zmiennej objaśnianej innych krajów dodanych do zbioru(technika wykorzystywana w uczeniu maszynowym).