Efekt 6, Regresja

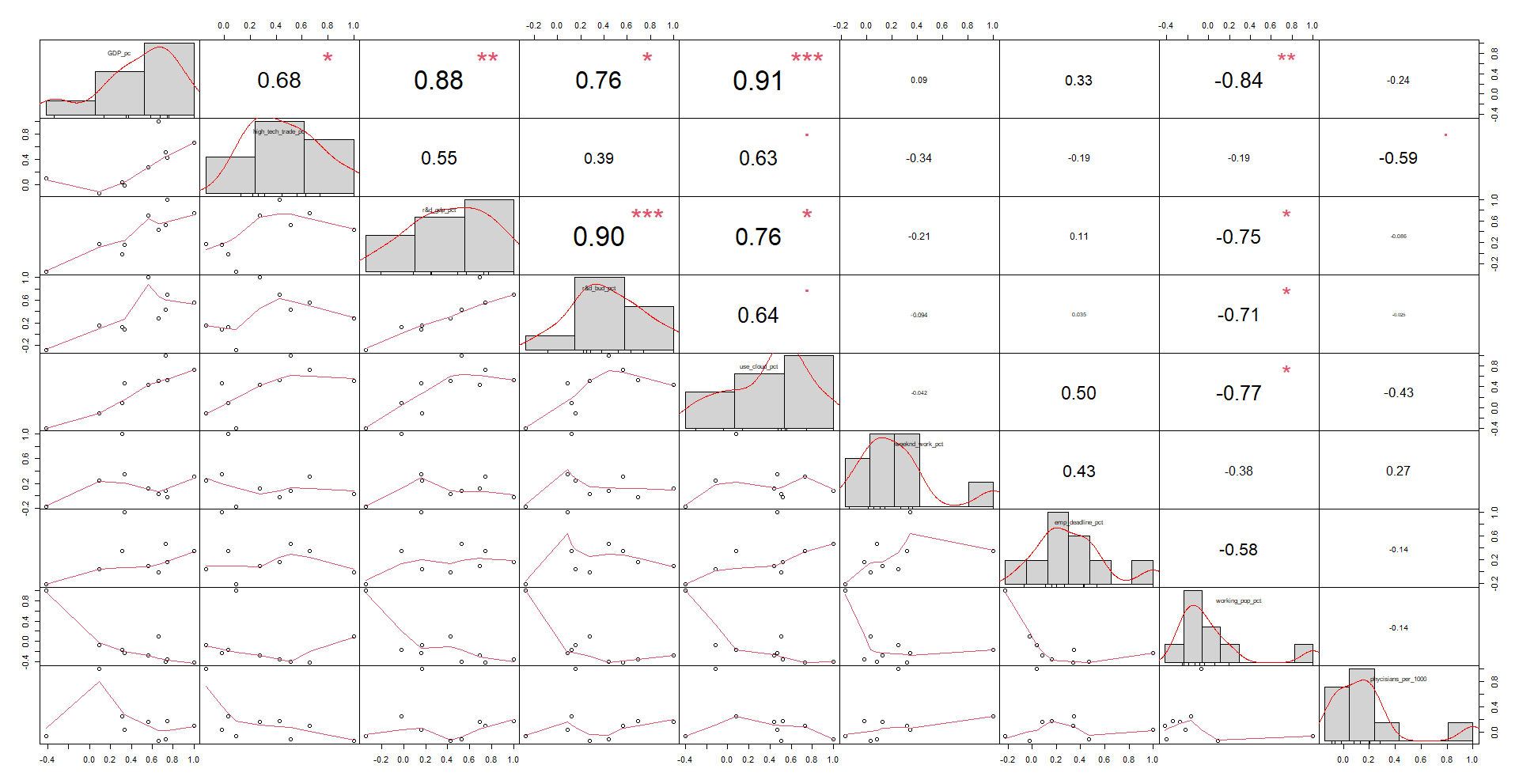
Mateusz Nowak, Damian Okoń, Robert Zamiar

# Pytanie badawcze

Jak wygląda model regresji liniowej wykonany na wszystkich predyktorach? Czy model przeszedł pozytywnie weryfikacje (testy autokorelacji, heteroskedastyczności i normalności reszt)? Jakie jest dopasowanie modelu i jego interpretacja?

# 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

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 2. Korelacje predyktorów ze zmienną objaśnianą - wycinek macierzy korelacji

Zmienna GDP\_pc jest najbardziej skorelowana z zmienną „handel wysokimi technologiami per capita”, wydatkami budżetowymi na badania i rozwój oraz procentem osób korzystających z technologii chmurowych.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek 3. Test istotności korelacji

Test na istotność korelacji wykazał, że istotnie ze skorelowane ze zmienną objaśnianą są między innymi zmienne opisujące wydatki na rozwój (r&d), procent osób korzystających z usług chmurowych, 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.

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Rysunek 4. Model regresji liniowej – wszystkie predyktory liczbowe

Predyktorami istotnie wpływającymi na zmienną objaśnianą GDP\_pc 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.

Obraz zawierający tekst, paragon

Opis wygenerowany automatycznie

Rysunek 5. Model regresji liniowej - wybrane predyktory

W tym modelu wszystkie zmienne łącznie z wyrazem wolnym okazały się istotne, zmalał współczynnik determinacji, co jest efektem zmniejszenie liczby predyktorów. Zmalała również różnica między R2, a skorygowanym R2. Użyto zatem zmiennych istotnie wpływających na dopasowanie modelu.

# Weryfikacja modelu

## Autokorelacja reszt

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 6. Test Durbina-Watsona

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. W naszym przypadku wynik p-value jest wyższy niż 0.05. Oznacza to, że nie ma podstaw do odrzucenia hipotezy zerowej. Autokorelacja reszt nie występuje.

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

Heteroskedastyczność

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 7. Test Breuscha-Pagana

Celem sprawdzenia heteroskedastyczności przeprowadzono test Breuscha-Pagana. Hipoteza H0 tego testu mówi, że heteroskedastyczność reszt nie występuje (wariancja reszt jest stała). H1z kolei stwierdza, że heteroskedastyczność jest obecna. Zgodnie z wartością p-value wykonanego testu (rysunek 7) przyjmujemy hipotezę H0. Heteroskedastyczność reszt nie występuje. Gdyby była ona obecna, model miałby błędne wnioskowanie statystyczne.

## Normalność składnika losowego

Obraz zawierający tekst

Opis wygenerowany automatycznie

Rysunek 8. Test Doornicka-Hansena

Normalność składnika losowego sprawdzono przy pomocy testu Doornika-Hansena przy wykorzystaniu oprogramowania Gretl. Hipoteza H0 mowi, że rozkład składnika losowego jest normalny. Ponieważ p-value jest wyższe od poziomu istnotności alfa=0.05, przyjmujemy hipotezę zerową. Model został właściwie zweryfikowany.

# Odpowiedź na pytanie badawcze

Model przedstawia się równaniem:

**y = -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. Wzrośnie również o 9000,64 euro, jeżeli zmienna związana z procentem wydatków na rozwój państwa zwiększy się o jeden punkt procentowy. Z kolei 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 Pozwala to na wykorzystanie modelu w celach predykcyjnych, przykładowo do przewidywań wartości zmiennej objaśnianej innych krajów dodanych do zbioru (technika wykorzystywana w uczeniu maszynowym).

Model również pozytywnie przeszedł wszystkie weryfikacyjne testy statystyczne, co dokładniej opisano w poprzednim paragrafie.