Metody badań w zarządzaniu finansami – WZ – UW – 2019/2020

ZAJĘCIA 2

Ostatnio

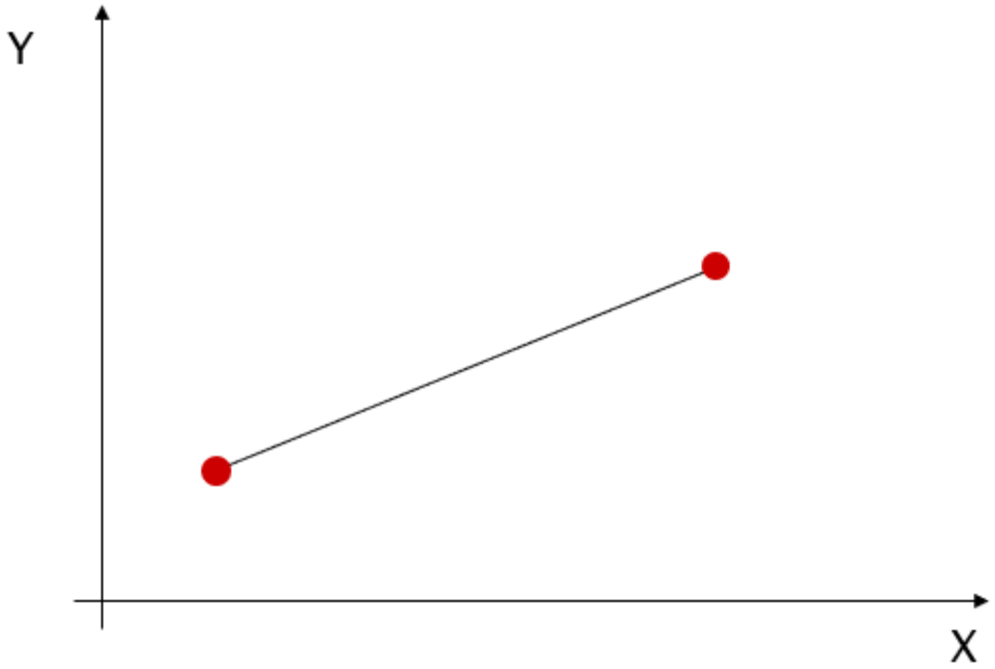
- Analiza danych – biznesowe zrozumienie danych i umiejętność przedstawienia danych (jedno- i wielowymiarowych) na wykresie, tworzenie tabeli przestawnych
- Analiza dynamiki zmian - obliczanie przyrostów bezwzględnych/ względnych danej zmiennej, udziału, kontrybucji
- Wyciąganie wniosków biznesowych
- Prognozowanie zysków
- Błędy prognozy

O czym dziś

- Jak zbudować dobry model ekonometryczny?
- Jakie dane wybrać? Czym się różnią?
- Po co nam ekonometria?
  - Co to jest model?
  - Jak sprawdzić czy model jest poprawny?
  - Jak porównywać modele między sobą

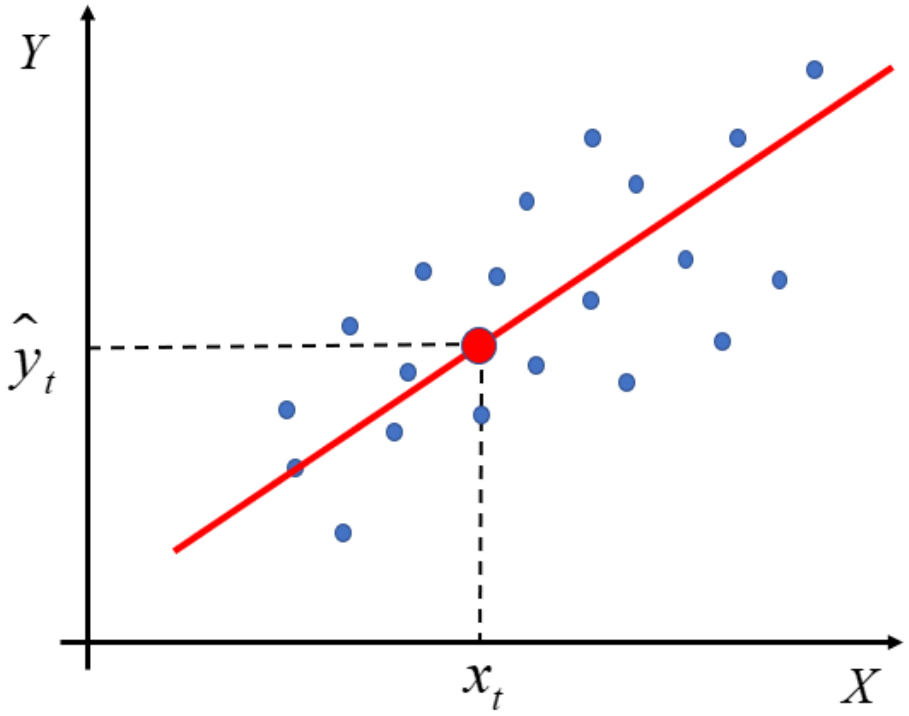
Po co nam model?

- Jak znaleźć optymalną prostą (taką, która najlepiej zobrazuje zależności pomiędzy ilością (X) a kosztem (Y)?

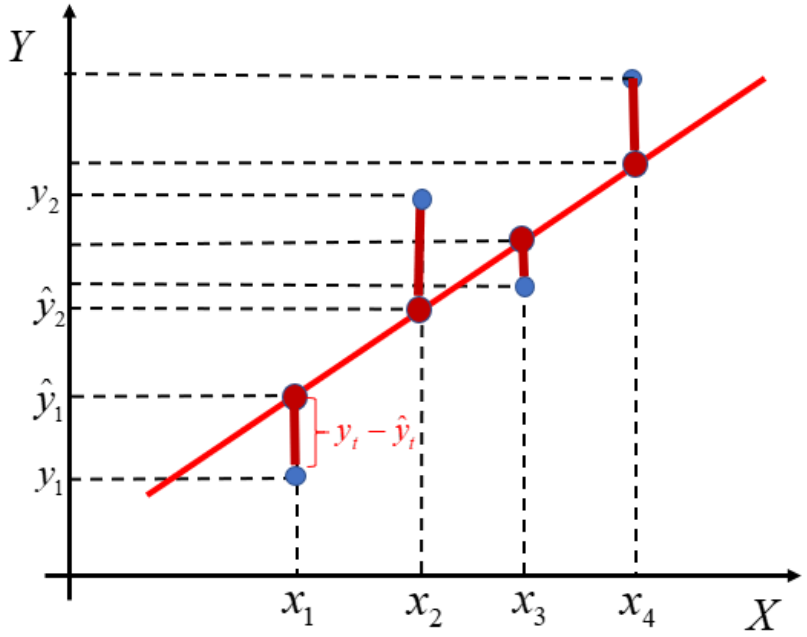


Po co nam ekonometria?

- Jak znaleźć optymalną prostą (taką, która najlepiej zobrazuje zależności pomiędzy ilością (X) a kosztem (Y)?



Optymalnie = jak najmniej się mylić



$\Rightarrow$ Metoda Najmniejszych Kwadratów (MNK)

Co to jest model?

- Sposób opisu zaobserwowanych zjawisk (dane)

- Redukcja rzeczywistości: za pomocą niewielkiej liczby oszacowanych parametrów umożliwia zobrazowanie zależności między zmiennymi

Ćwiczenie: zbiór danych *CASchools*

1. Stwórz zmienne: *TestScore i STR*
2. Przedstaw graficznie zależność między zmiennymi *TestScore i STR (plot)*
3. Czy istnieje korelacja między zmiennymi? (*cor*)
4. Stwórz model regresji liniowej, w którym zmienną objaśnianą będzie *TestScore,* a zmienną objaśniającą *STR* . (*lm*)
5. Przeanalizuj charakterystyki zbudowanego modelu (*summary)*

- https://www.rdocumentation.org/packages/AER/versions/1.2-8/topics/CASchools

Ćwiczenie: 1

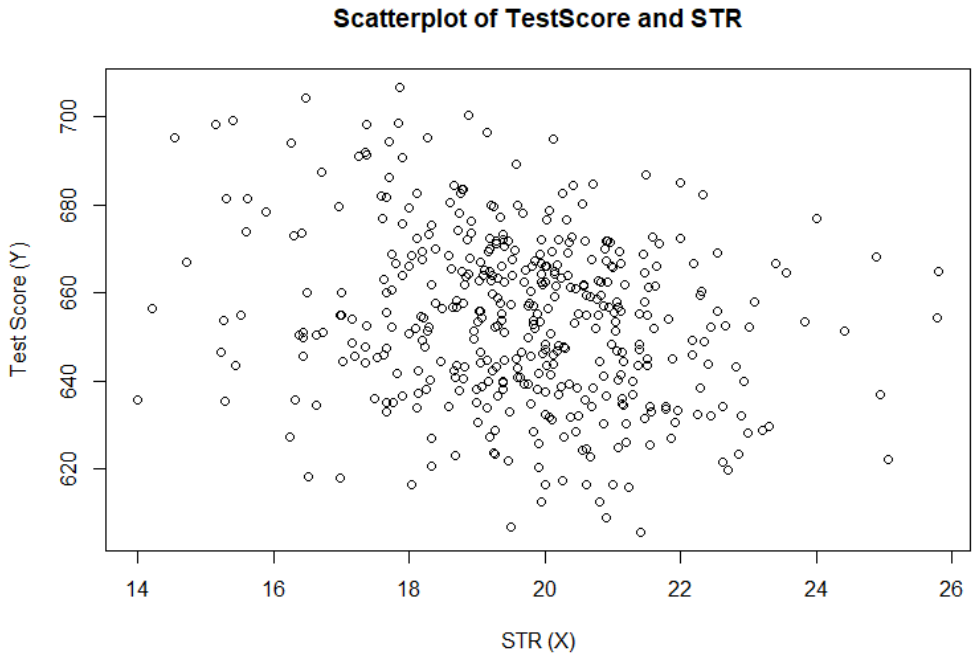
- Korzystając ze zbioru danych *CASchools* dostępnego w R

```
> head(CASchools)
  district                         school  county grades students teachers calworks   lunch computer expenditure    income   english  read  math
1    75119              Sunol Glen Unified Alameda  KK-08      195    10.90   0.5102  2.0408       67    6384.911 22.690001  0.000000 691.6 690.0
2    61499            Manzanita Elementary   Butte  KK-08      240    11.15  15.4167 47.9167      101    5099.381  9.824000  4.583333 660.5 661.9
3    61549      Thermalito Union Elementary   Butte  KK-08     1550    82.90  55.0323 76.3226      169    5501.955  8.978000 30.000002 636.3 650.9
4    61457 Golden Feather Union Elementary   Butte  KK-08      243    14.00  36.4754 77.0492       85    7101.831  8.978000  0.000000 651.9 643.5
5    61523         Palermo Union Elementary   Butte  KK-08     1335    71.50  33.1086 78.4270      171    5235.988  9.080333 13.857677 641.8 639.9
6    62042          Burrel Union Elementary  Fresno  KK-08      137     6.40  12.3188 86.9565       25    5580.147 10.415000 12.408759 605.7 605.4
```

```
> #dodamy dwie ddatkowe zmienne: average test score and the student-teacher ratio
> # define variables
> CASchools$STR <- CASchools$students/CASchools$teachers
> CASchools$score <- (CASchools$read + CASchools$math)/2
>
```

Ćwiczenie: 2

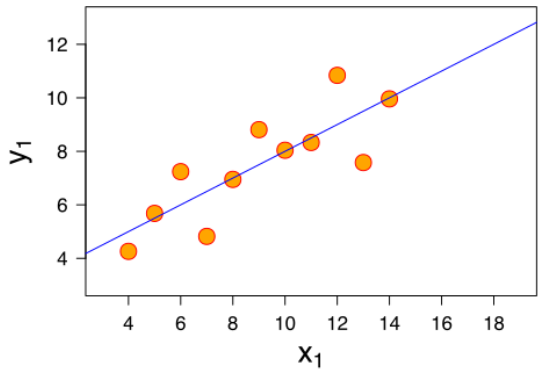
- Przedstaw graficznie zależność między zmiennymi *TestScore i STR (plot)*

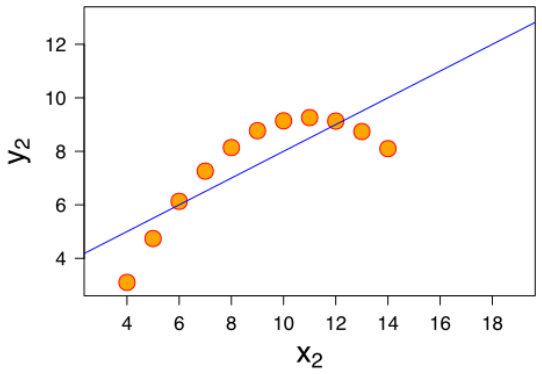


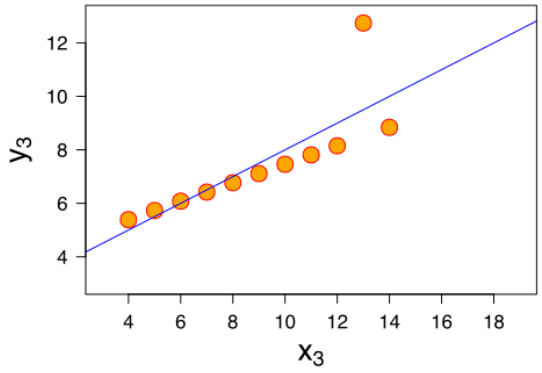
Ćwiczenie: 2

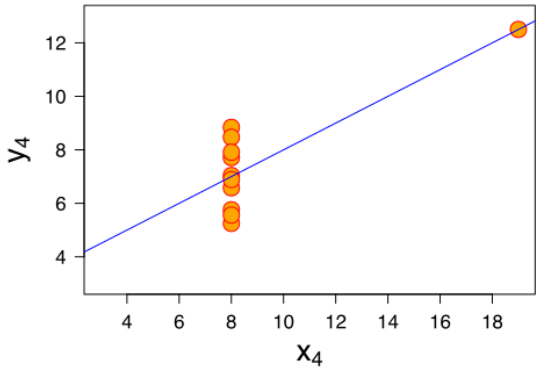
- Po co wykresy? → **Kwartet Anscombe'a**

- *średnia arytmetyczna*
- *wariancja*
- *współczynnik korelacji*
- *równanie regresji liniowej*









Ćwiczenie: 3

- Czy istnieje korelacja między zmiennymi? (cor)
- Trzeba specjalnie poprosić o statystyczną istotność w komendzie (cor)

```
> # compute correlations
> cor(CAschools$STR, CAschools$score)
[1] -0.2263627
>
```

Ćwiczenie: 4

- Stwórz model regresji liniowej, w którym zmienną objaśnianą będzie *TestScore,* a zmienną objaśniającą *STR* . (*lm*)

```
> # estimate the model and assign the result to linear_model
> linear_model <- lm(score ~ STR, data = CASchools)
> # print the standard output of the estimated lm object to the console
> linear_model

Call:
lm(formula = score ~ STR, data = CASchools)

Coefficients:
(Intercept)          STR
     698.93        -2.28
```

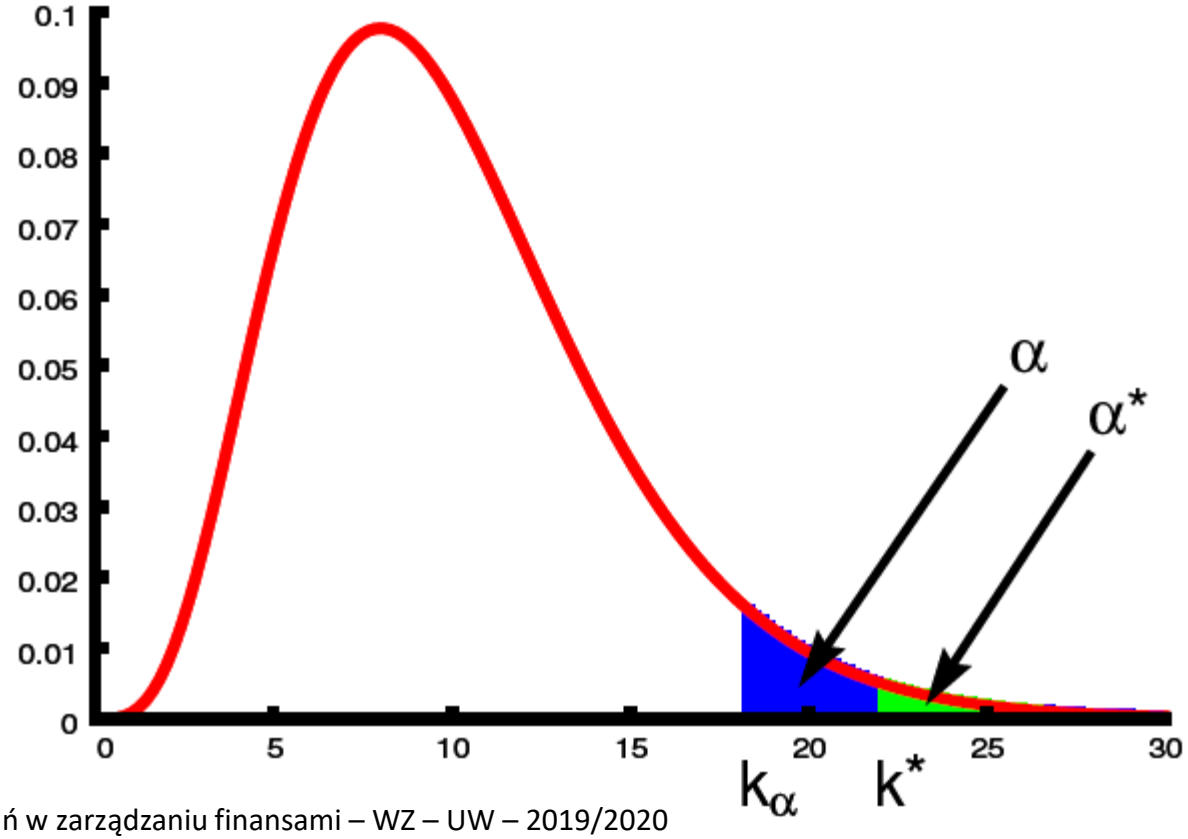
Ćwiczenie: 5

- Przeanalizuj charakterystyki zbudowanego modelu (*summary)*

```
> mod_summary <- summary(linear_model)
> mod_summary

Call:
lm(formula = score ~ STR, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 698.9329     9.4675  73.825  < 2e-16 ***
STR          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Jak dobre są nasze oszacowania

- Możemy przetestować nasze uzyskane wartości w porównaniu do hipotezy mówiącej, że mają one wartość 0



Testowanie hipotez

- α – poziom istotności
  - przyjęte z góry dopuszczalne ryzyko błędnego wnioskowania
  - mówi, jakie jest prawdopodobieństwo odrzucenia prawdziwej hipotezy (błąd I rodzaju)

- Dla każdego α definiujemy przedział ufności

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha$$

Oraz wartości krytyczne rozkładu k_α

Jak zbudowaliśmy model regresji?

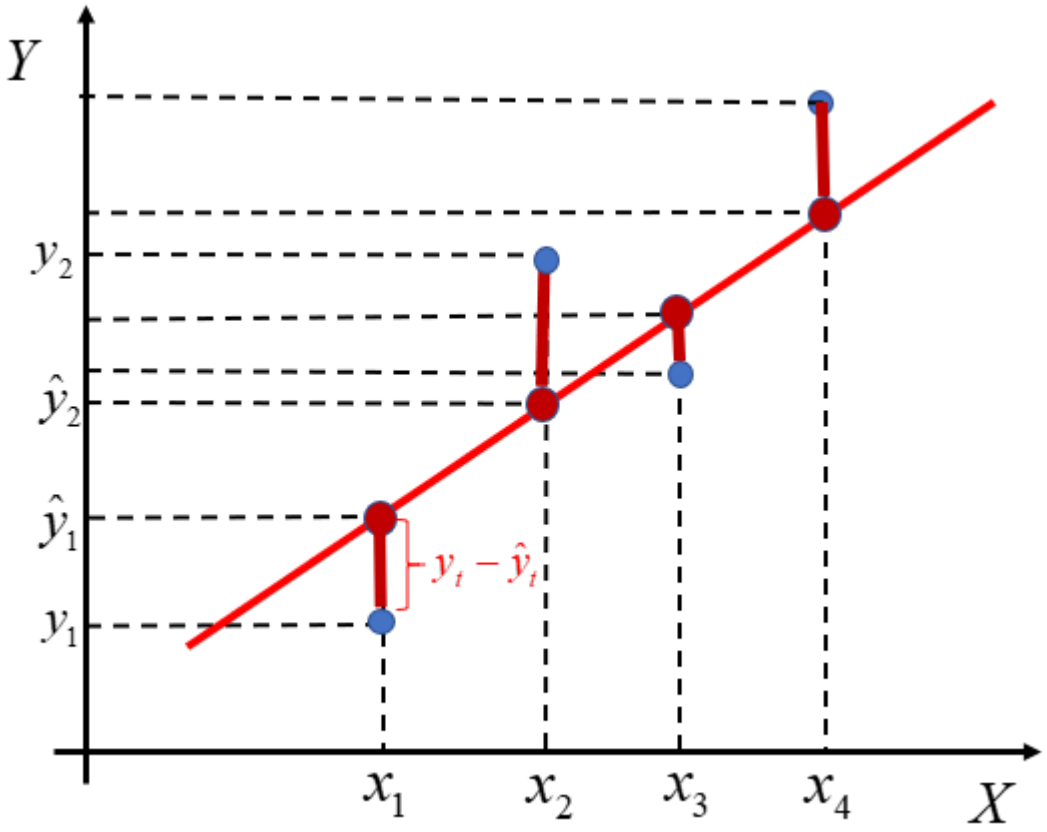
- Model

$$y = \beta X + \varepsilon$$

- Jeśli spełnione są założenia:
  1. liniowości ze względu na parametry
  2. $E(\varepsilon) = 0$
  3. $Var(\varepsilon) = \sigma^2$
  4. $Cov(x, \varepsilon) = 0$

to Carl Friedrich Gauss oraz Andrey Markov gwarantują, że metoda najmniejszych kwadratów daje… najmniejsze kwadraty

Najmniejsze kwadraty to



Metoda Najmniejszych Kwadratów (MNK)

- Dopasowana linia (oszacowany model):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

a reszty określamy jako:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Wartości estymatorów $\hat{\beta}_0$ i $\hat{\beta}_1$ minimalizują sumę kwadratów reszt:

$$SSE = \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Rozwiązanie

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estymatory MNK w prostym modelu regresji

Co wiemy o estymatorach MNK? Twierdzenie Gaussa-Markowa

- W klasie metod dających estymatory **liniowe** i **nieobciążone**

- Jeżeli spełnione są cztery (?) założenia
  1. ~~liniowości ze względu na parametry~~
  2. $E(\varepsilon) = 0$
  3. $Var(\varepsilon) = \sigma^2$
  4. $Cov(x, \varepsilon) = 0$

- To MNK daje **najlepszą** estymację (o najmniejszej wariancji składnika losowego)

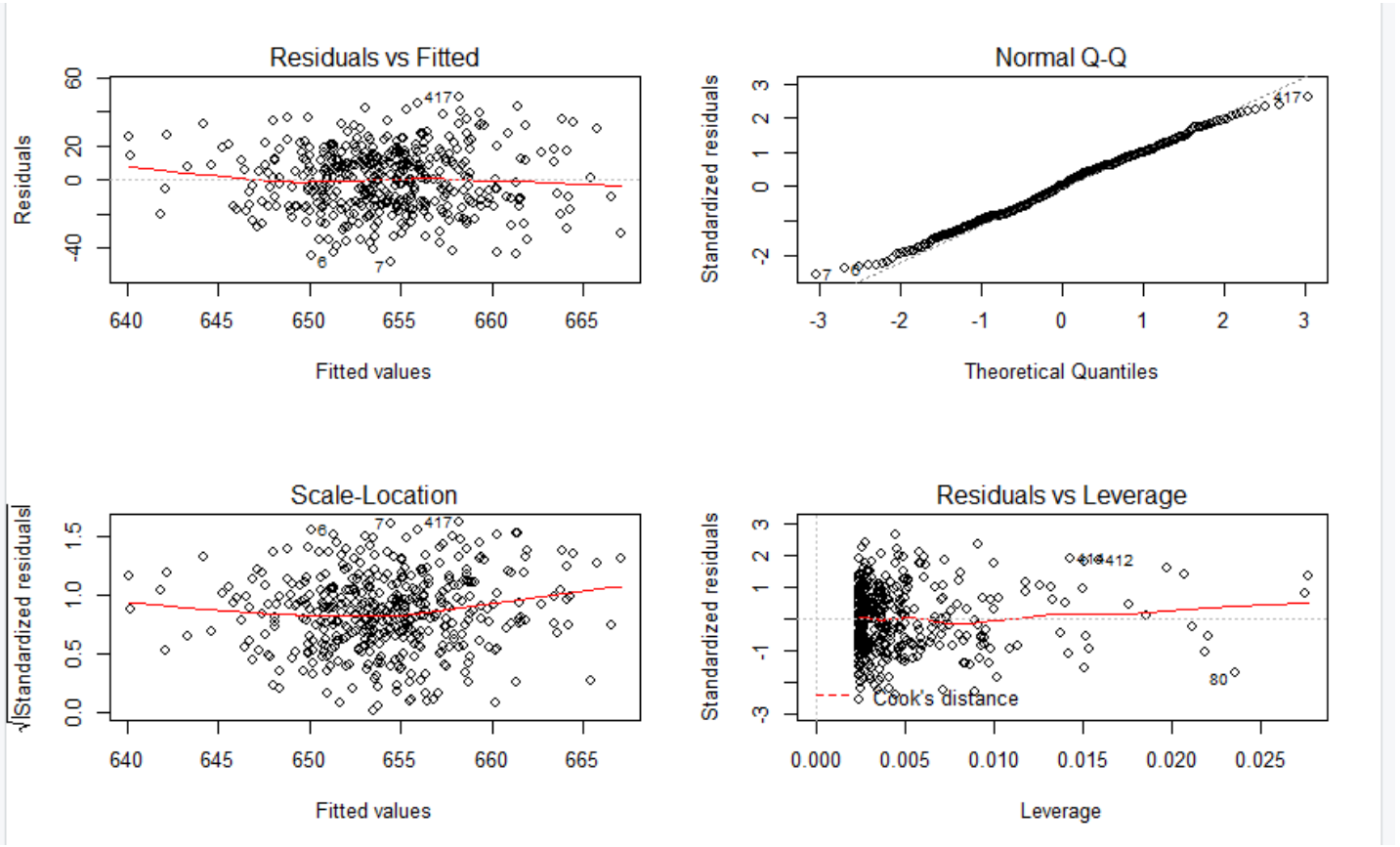
Ćwiczenie w komputerze

- Sprawdź czy spełnione są założenia MNK
- $E(\varepsilon) = 0$

```
#The mean of residuals is zero
mean(linear_model$residuals)
1] -1.636068e-14
```

Ćwiczenie w komputerze

- Sprawdź czy spełnione są założenia MNK



Własności estymatorów MNK

- Nieobciążoność : $E\left[\hat{\beta}\right] = \beta$ (wartość oczekiwana jest równa wartości szacowanego parametru)
- Zgodność: $\underset{n\to\infty}{\text{plim}}\,\hat{\beta} = \beta$
- Efektywność – posiada najmniejszą wariancję w swojej klasie
- Dostateczność

Skąd wiedzieć czy model jest dobry?

- Jak ocenić jakość modelu?
- Możemy rozdzielić y_i na $y_i = E(y_i) + \varepsilon_i$
  - gdzie $E(y_i)$ to część wyjaśniona, a ε_i to część losowa, niewyjaśniona
- Na podstawie modelu empirycznego mamy

$$y_i = \widehat{y_i} + \widehat{\varepsilon_i}$$

- Lub jako odchylenie od średniej

$$y_i - \bar{y} = (\widehat{y_i} - \bar{y}) + \widehat{\varepsilon_i}$$

- *Total sample variation* lub *total sum of squares* (TSS) można rozdzielić na sumę kwadratów objaśnioną w modelu i sumę kwadratów reszt

$$\sum(y_i - \bar{y})^2 = \sum(\widehat{y_i} - \bar{y})^2 + \sum \widehat{\varepsilon_i}^2$$

TSS=SSR+RSS

Skąd wiedzieć czy model jest dobry?

- Współczynnik determinacji R^2 zdefiniowany jest jako proporcja zmienności y wyjaśniona przez model
- $R^2 = \frac{\sum(\widehat{y_i}-\bar{y})^2}{\sum(y_i-\bar{y})^2} = \frac{SSR}{TSS}$
- $R^2 = 1 - \frac{\sum \widehat{\varepsilon_i}^2}{\sum(y_i-\bar{y})^2} = 1 - \frac{RSS}{TSS}$
- Skorygowany R^2
- $R^2 = 1 - \frac{RSS/(N-K)}{TSS/(N-1)}$

Skąd wiedzieć czy model jest dobry?

- Pomoce statystyki do weryfikacji jakości modelu i porównania modeli między sobą:

STATYSTYKA	KRYTERIUM
R-Squared	Im wyższe tym lepsze (>0.70)
Adj R-Squared	Im wyższe tym lepsze
F-Statistic	Im wyższe tym lepsze
Std.Error	Im bliższe zeru tym lepiej
T-statistic	>1.96 (wtedy p-value<0.05)
AIC (Akaike Information Criterion)	Im niższe tym lepsze
BIC (Baysian- Schwartz Information Criterion)	Im niższe tym lepsze
HIC (Hannan-Quinn Information Criterion)	Im niższe tym lepsze

Ćwiczenie: jak dobry był nasz model?

- Przeanalizuj charakterystyki zbudowanego modelu (*summary)*

```
> mod_summary <- summary(linear_model)
> mod_summary

Call:
lm(formula = score ~ STR, data = CASchools)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 698.9329     9.4675  73.825  < 2e-16 ***
STR          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,   Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Co może powiedzieć model?

- Wyestymowane współczynniki są tylko korelacjami
  - Przyczynowość nie istnieje bez teorii i refleksji!
  - Nie istnieje w 100% wiarygodny test przyczynowości

- Jakiegokolwiek testu nie zdałby model – musi mieć sens
  - Regresja pozorna czy dekompozycja
    - Związek między produkcją a kosztami

    vs

    - Sprzedaż ogółem = a sprzedaż1 + b sprzedaż2 + …

  - Problem obserwacji nietypowych
    - Patrz z uwagą na to, co próbują powiedzieć kropki!

JAK ZBUDOWAĆ MODEL?

Etapy budowy modelu ekonometrycznego

- Postawienie hipotezy badawczej
- Wybór postaci funkcyjnej i zbioru zmiennych objaśniających
- Zebranie danych
- Estymacja
- Weryfikacja
- Zastosowanie

Postawienie hipotezy badawczej

- Przykłady pytań badawczych:
  - Centra kosztowe, które nie są bezpośrednio powiązane z produkcją, ale też z pewnością nie są w całości kosztem stałym: jak oszacować funkcję kosztu przedsiębiorstwa?
  - Właściciel lokalnej restauracji Pizza Hit musi zdecydować jaką kwotę wydać na reklamę w lokalnej gazecie, czyli oszacować zależność między wydatkami na reklamę a przychodem. Przyczynowość?
  - Rada miasta zastanawia się, czy i o ile zmniejszy się przestępczość, jeżeli zdecyduje się zwiększyć nakłady finansowe na policję. Na co najlepiej je wydać? Funkcjonariuszy? Monitoring? Edukację prewencyjną?

Czym się różnią dane?

- Zmienne mogą być gromadzone na różnych poziomach agregacji: mikro ….. makro
- Dyskretne vs. Ciągłe
- Jakościowe vs. Ilościowe
- Eksperymentalne vs. Nie-eksperymentalne
- Flow vs. Stock

Źródła danych online

- Dobre miejsca do rozpoczęcia poszukiwań:
- The Economic Network's website: http://www.economicsnetwork.ac.uk/links/data_free.htm
- The Econometrics Journal Links: http://www.feweb.vu.nl/econometriclinks/#data
- The World Bank's website: http://data.worldbank.org/topic Rolnictwo, edukacja, zmiany klimatyczne
- The OECD's website: http://www.oecd-ilibrary.org/statistics - Dla krajów OECD
- The CIC website: https://pwt.sas.upenn.edu Międzynarodowe porównanie produkcji, dochodów i cen
- The American Economic Association's website: http://rfe.org/showCat.php?cat_id=2 Zmienne makroekonomiczne i finansowe, dla USA i Świata
- The FRED's website: http://research.stlouisfed.org/fred2/categories/ Rynek pracy, zmienne demograficzne, finansowe, sektor bankowy, produkcyjny
- The UK Data Archive's website: http://ukdataservice.ac.uk and http://dataarchive.ac.uk/deposit/use/

Zadanie do domu

- Przeczytać artykuł Ross Levine:

FINANCE AND GROWTH: THEORY AND EVIDENCE

- Zapoznać się z danymi Global Finance Development Database (spróbować znaleźć wskaźniki finansowe, które były uwzględnione w artykule FINANCE AND GROWTH: THEORY AND EVIDENCE)

Dziękuję za Państwa czas!

Sylwia Radomska

s.radomska@grape.org.pl

http://grape.org.pl/sradomska