TESTOWANIE HIPOTEZ

BŁĄD ZMIENNEJ POMINIĘTEJ

EKONOMETRIA WNE 2022/23

Sebastian Zalas

12 stycznia 2022

WPROWADZENIE

- Nauczymy się jak testować hipotezy dotyczące tego czy wartości poszczególnych oszacowań są zbieżne z tym co sądzimy o relacjach zachodzących w świecie.
  - przykład: czy oszacowana elastyczność z danych jest równa wartości przewidywanej przez teorię ekonomii?
  - przykład: czy oszacowany związek zmiennych y i x nie jest jednorazowy?

- oraz przedyskutujemy:
  - błąd zmiennej pominiętej (*Ommited Variable Bias*)
  - współliniowość
  - uwzględnianie zmiennej nieistotnej

PLAN

1. Test t
2. Test F
3. Błąd zmiennej pominiętej (*Omitted Variable Bias*)
4. Współliniowość
5. Zmienne nieistotne
6. Przykład w `R`

TEST T

ROZKŁAD ESTYMATORA MNK

- Rozkład estymatora jest kluczowy dla testowania hipotez.
- W KMRL mamy założenie o normalności składnika losowego:

$$\boldsymbol{e} \mid \boldsymbol{X} \sim N(0, \boldsymbol{I}\sigma^2)$$

własność MNK

$$\hat{\beta} - \beta = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{e}$$

ponieważ $\hat{\beta} - \beta$ jest liniową funkcją $\boldsymbol{e}$, to $\hat{\beta} - \beta$ również ma rozkład normalny:

$$\begin{aligned} \hat{\beta} - \beta \mid \boldsymbol{X} &\sim (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top N(0, \boldsymbol{I}\sigma^2) \\ &\sim N(0, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \\ &= N(0, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \end{aligned}$$

co oznacza, że:

$$\hat{\beta} \sim N(\beta, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \tag{1}$$

STATYSTYKA T

- Zapiszmy równanie (1) w inny sposób:

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}}} \sim N(0, 1)$$

otrzymujemy statystykę t o standardowym rozkładzie normalnym

- Nie obserwujemy wariancji i należy skorzystać z oszacowania:

$$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}}} \sim t(n - k - 1)$$

Wtedy statystyka t ma rozkład t-studenta z $n - k - 1$ stopniami swobody

TEST T: HIPOTEZA

- Mamy dany model:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

- chcemy sprawdzić, czy hipoteza $\beta = d$ jest prawdziwa, gdzie d przyjmuje jakąś znaną wartość (np. zero). Formalnie zapisujemy to w poniższy sposób:

$$H_0 : \beta = d$$

$$H_1 : \beta \neq d$$

TESTOWANIE HIPOTEZY

- Przy założeniu, że hipoteza zerowa jest prawdziwa, obliczamy wartość **statystyki testowej**:

$$\frac{\hat{\beta} - d}{se(\hat{\beta})} \sim t(n - k - 1)$$

- następnie ustalamy **poziom istotności**, standardowo przyjmuje się $\alpha = 0.05$
- znajdujemy **wartości krytyczne**: $t^*_{\frac{\alpha}{2},n-k-1}$ oraz $t^*_{\frac{1-\alpha}{2},n-k-1}$
- przyjmujemy hipotezę zerową, jeżeli statystyka testowa znajduje się pomiędzy wartościami krytycznymi
- w przeciwnym przypadku, odrzucamy hipotezę zerową.

TESTOWANIE ISTOTNOŚCI

- Oszacowanie modelu:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k + \varepsilon_i$$

- Hipoteza zerowa i alternatywna:

$$H_0 : \hat{\beta}_k = 0$$

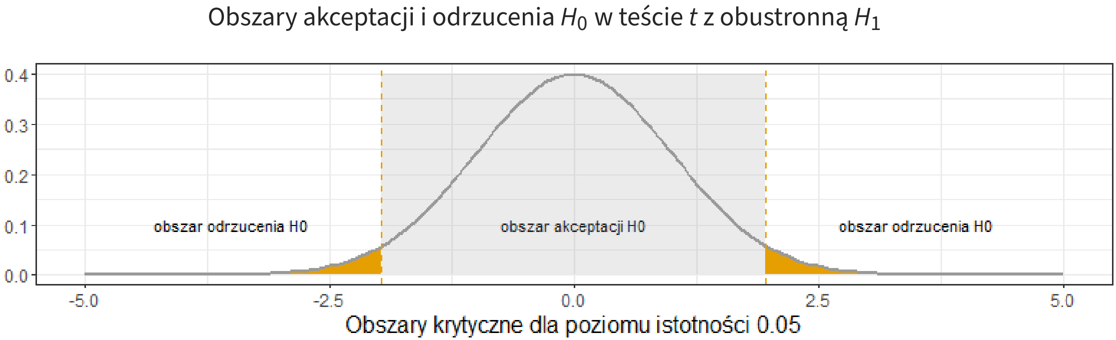
$$H_1 : \hat{\beta}_k \neq 0$$

- obliczmy wartość statystyki t:

$$t = \frac{\hat{\beta}_k}{se(\hat{\beta})} \sim t(n-k-1)$$

- znajdujemy **wartości krytyczne**: $t^*_{\frac{\alpha}{2},n-k-1}$ oraz $t^*_{\frac{1-\alpha}{2},n-k-1}$
- przyjmujemy hipotezę zerową, jeżeli statystyka testowa znajduje się pomiędzy wartościami krytycznymi $\Rightarrow$ testowane oszacowanie nie jest istotne statystycznie

TESTOWANIE ISTOTNOŚCI



PRZEDZIAŁY UFNOŚCI (I)

- Możemy skonstruować taki przedział, że będzie on zawierał prawdziwy parametr β z ustalonym prawdopodobieństwem równym $1 - \alpha$.
- dla 95% przedziałów ufności: $1 - \alpha = 0.95$
- z rozkładu t – *studenta* możemy uzyskać takie wartości krytyczne, że jakakolwiek zm. losowa mająca taki rozkład będzie zawierać się w przedziale $(t^*_{\frac{\alpha}{2},n-k-1}, t, t^*_{\frac{1-\alpha}{2},n-k-1})$ z prawdopodobieństwem $1 - \alpha$:

$$P(t^*_{\frac{\alpha}{2},n-k-1} \leq t \leq t^*_{\frac{1-\alpha}{2},n-k-1}) = 1 - \alpha$$

$$P(t^*_{\frac{\alpha}{2},n-k-1} \leq \frac{\beta_k - \hat{\beta}_k}{se(\hat{\beta})} \leq t^*_{\frac{1-\alpha}{2},n-k-1}) = 1 - \alpha$$

$$P(se(\hat{\beta})t^*_{\frac{\alpha}{2},n-k-1} \leq \beta_k - \hat{\beta}_k \leq se(\hat{\beta})t^*_{\frac{1-\alpha}{2},n-k-1}) = 1 - \alpha$$

PRZEDZIAŁY UFNOŚCI (II)

- Po przekształceniach otrzymujemy:

$$P(\hat{\beta}_k + se(\hat{\beta})t^*_{\frac{\alpha}{2},n-k-1} \leq \beta_k \leq \hat{\beta}_k + se(\hat{\beta})t^*_{\frac{1-\alpha}{2},n-k-1}) = 1 - \alpha$$

- rozkład t – *studenta* jest symetryczny, więc $-t^*_{\frac{\alpha}{2},n-k-1} = t^*_{\frac{1-\alpha}{2},n-k-1}$, więc przedział ufności ma postać:

$$P(\hat{\beta}_k - se(\hat{\beta})t^*_{\frac{1-\alpha}{2},n-k-1} \leq \beta_k \leq \hat{\beta}_k + se(\hat{\beta})t^*_{\frac{1-\alpha}{2},n-k-1}) = 1 - \alpha$$

- jeśli $\alpha = 0.95$, to wtedy z prawdopodobieństwem 0.95

$$\hat{\beta}_k - se(\hat{\beta})t^*_{\frac{1-\alpha}{2},n-k-1} \leq \beta_k \leq \hat{\beta}_k + se(\hat{\beta})t^*_{\frac{1-\alpha}{2},n-k-1}$$

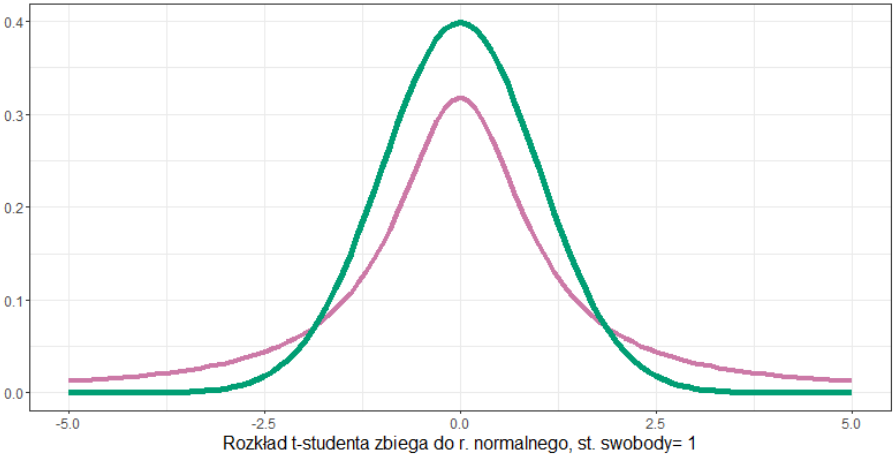
CO SIĘ DZIEJE GDY SKŁADNIK LOSOWY NIE MA ROZKŁADU NORMALNEGO?

- jeśli $\varepsilon \sim N(0, \sigma^2)$ to wtedy statystyka testowa t ma dokładnie rozkład t – *studenta*
- jeśli zaś składnik losowy nie ma rozkładu normalnego nadal możemy przeprowadzić test jesli mamy wystarczająco dużą próbę
- Na podstawie Centralnego Twierdzenia Granicznego:

$$t = \frac{\hat{\beta}_k}{se(\hat{\beta})} \overset{as.}{\sim} N(0, 1)$$

- wtedy odczytujemy wartości krytyczne ze standardowego rozkładu normalnego: $\mid t^* \mid \pm 1.96$

ROZKŁAD T-STUDENTA ZBIEGA DO ROZKŁADU NORMALNEGO STANDARDOWEGO.



TEST F

TESTOWANIE WIELU RESTRYKCJI:

- Korzystając z testu *t* możemy przetestować hipotezę dotyczącą jednego parametru, $\hat{\beta}_k$. Test F pozwala na przetestowanie wielu restrykcji (*multiple restrictions*), które chcemy nałożyć na model.

- Przykład: mamy dany model:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon_i$$

- chcemy sprawdzić czy $\beta_2 = \beta_3 = 0$, zatem:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_2 \neq 0 \text{ lub } \beta_3 \neq 0$$

TEST F (I)

- szacujemy podstawowy model, bez restrykcji (oznaczmy go literą U -*unrestricted*):

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \ldots + \hat{\beta}_k x_k + \varepsilon_i$$

- następnie szacujemy z restrykcjami (oznaczmy go literą R -*restricted*):

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k + \varepsilon_i$$

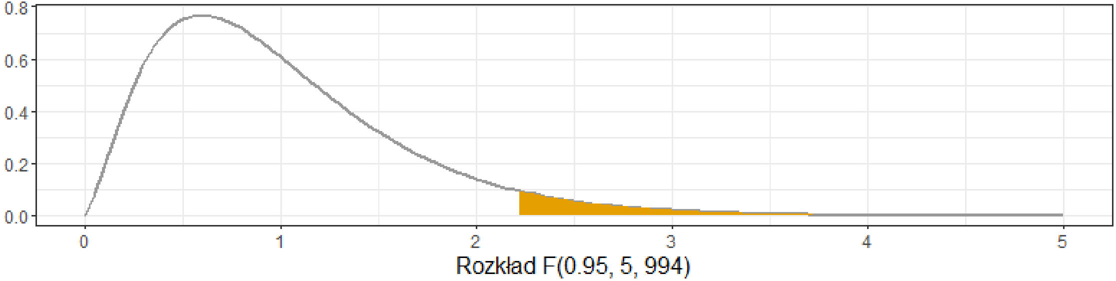
- budujemy statystykę testową

$$F = \frac{\frac{R_U^2 - R_R^2}{q}}{\frac{1 - R_U^2}{n - k_U - 1}} \sim F(1 - \alpha, q, n - k - 1)$$

  gdzie q to liczba restrykcji

TEST F (II)

Obszary akceptacji i odrzucenia H_0 w teście F



TEST F (III)

- Korzystamy z rozkładu F z $(q, b - k - 1)$ stopniami swobody aby wyznaczyć wartość krytyczną F^*. Przyjmujemy H_0 jeśli statystyka testowa jest niższa niż wartość krytyczna: $F < F^*$
- Im większa różnica między $R_U^2 - R_R^2 \Rightarrow$ tym wyższa statystyka $F \Rightarrow$ lepsze dopasowanie do danych dzięki dodaniu zmiennych do modelu.
- Przykład: powszechnie stosowany **test łącznej istotności**. Model podstawowy, bez restrykcji:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \ldots + \hat{\beta}_k x_k$$

- w tym przypadku, model z restrykcjami ma postać:

$$y_i = \hat{\beta}_0$$

- Hipotezy

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$$
$$H_1 : \text{chociaż jeden parametr \underline{nie} jest równy zero}$$

- statystyka testowa, w tym przypadku jest nieco uproszczona:

$$F = \frac{\frac{SST^2 - SSE^2}{q}}{\frac{1 - SSE^2}{n-k-1}} \sim F(1 - \alpha, q, n - k - 1)$$

statystyka *F* oraz odpowiadające jej *p-value* są raportowane w pakietach statystycznych, np. w R.

RESTRYKCJE W ZAPISIE MACIERZOWYM (I)

- Hipotezy dotyczące kilku parametrów możemy zapisać w formie macierzowej:

$$\boldsymbol{H}\beta = \boldsymbol{h}$$

gdzie $\boldsymbol{H}$ to macierz $(q \times (k+1))$ opisująca restrykcje, q to liczba restrykcji, $\boldsymbol{h}$ to wektor stałych z każdej restrykcji.

- Przykład – test łącznej istotności:

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{q \times k+1} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{k+1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_q$$

RESTRYKCJE W ZAPISIE MACIERZOWYM (II)

- Przykład – test następujących restrykcji:
  1. $\beta_1 = \beta_3$
  2. $\beta_2 = a$
  3. $\beta_1 + \beta_4 = b$

$$\begin{bmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ a \\ b \end{bmatrix}$$

PRZYKŁAD RESTRYKCJI: STAŁE KORZYŚCI SKALI

- Możemy narzucić restrykcje na model, korzystając z teorii ekonomii.
- Przykład – funkcja produkcji:

$$y = \beta_0 + \beta_k k + \beta_l l + \varepsilon$$

- ekonomiści często zakładają że charakteryzują stałe korzyści skali, co można przetestować formułując hipotezę:

$$\beta_k + \beta_l = 1$$

BŁĄD ZMIENNEJ POMINIĘTEJ (*omitted variable bias*)

BŁĄD ZMIENNEJ POMINIĘTEJ (I)

- Problem ten występuje gdy w regresji brakuje zmiennej opisującej zmienną zależną y
- ma istotne konsekwencje: gdy pominiemy zmienną, oszacowanie otrzymane metodą MNK będzie <u>obciążone</u>, gdyż $\mathbb{E}[\varepsilon \mid \boldsymbol{x}] \neq 0$

BŁĄD ZMIENNEJ POMINIĘTEJ (II)

- Przyjmijmy, że chcemy oszacować liniowy model:

$$y_i = \beta_0 + \beta_1 x_i + \gamma w_i + \varepsilon_i \tag{2}$$

  gdzie: x_i - wykształcenie, w_i - wrodzone zdolności

- nie obserwujemy (y_i, x_i), więc możemy oszacować jedynie poniższy model:

$$y_i = \beta_0 + \beta_1 x + e_i = \boldsymbol{x\beta} + \boldsymbol{e} \tag{3}$$

  gdzie: $e_i = (\gamma w_i + \varepsilon_i)$, co implikuje:

$$\mathbb{E}[e_i \mid x_i] = \mathbb{E}[\gamma w_i + \varepsilon_i \mid x_i] = \mathbb{E}[\varepsilon_i \mid x_i] + \mathbb{E}[\gamma w_i \mid x_i] \neq 0$$

- oszacowanie β_1 z modelu (2) będzie obciążone.

BŁĄD ZMIENNEJ POMINIĘTEJ (III)

- Obciążenie:

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}_1 \mid \boldsymbol{X}] &= \beta_1 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{e} \\
&= \beta_1 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{w}\gamma + \mathbb{E}[\varepsilon \mid \boldsymbol{x}]) \\
\mathbb{E}[\hat{\beta}_1 \mid \boldsymbol{X}] - \beta_1 &= \beta_1 + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{w}\gamma
\end{aligned}
$$

- jeżeli zachodzi $\gamma = 0$ lub $\mathbb{E}[x \mid w] = 0$, to wtedy oczekujemy że $\hat{\boldsymbol{\beta}}$ nie jest zbliżone do wektora parametrów $\boldsymbol{\beta}$

BŁĄD ZMIENNEJ POMINIĘTEJ (IV)

- Kiedy możemy spodziewać się, że $\hat{\boldsymbol{\beta}}$ będzie dobrym oszacowaniem $\boldsymbol{\beta}$?
  1. $\gamma = 0$ wtedy w_i oraz y_i są nieskorelowane, po uwzględnieniu x_i
     przykład: wrodzone zdolności nie są skorelowane z zarobkami pw. wykształcenia.
  2. $\mathbb{E}[x_i \mid w_i] = 0$: jeśli skupimy się tylko na β_1 to wystarczy aby x_i oraz w_i były nieskorelowane.

- Korzystając z założeń lub teorii ekonomii możemy przewidzieć kierunek i rozmiar obciążenia.

BŁĄD ZMIENNEJ POMINIĘTEJ (V)

- Przykład:
  - wrodzone zdolności w_i oraz zarobki y_i są dodatnio skorelowane $\Rightarrow \gamma > 0$.
  - zdolności w_i oraz wykształcenie x_i są dodatnio skorelowane $\Rightarrow \text{Cov}(x_i, w_i) > 0$

$$\mathbb{E}[\hat{\beta}_1] - \beta_1 = \gamma \frac{\text{Cov}(x_i, w_i)}{\text{Var}(x_i)} > 0$$

- Taki szacunek może dać pojęcie o kierunku obciążenia, którego powinniśmy się spodziewać jeśli spróbujemy oszacować β_1. W przykładzie z zarobkami i wykształceniem, przeszacowujemy wpływ wykształcenia poprzez pominięcie wrodzonych zdolności.

WSPÓŁLINIOWOŚĆ

WSPÓŁLINIOWOŚĆ (I)

- **Współliniowość** pojawia się gdy występuje silna korelacja między dwiema zmiennymi objaśniającymi.
- Dokładna współliniowość $\Rightarrow \hat{\beta}^{MNK}$ jest niezdefiniowany.
- Niedokładna współliniowość może pojawić się w modelu gdy:
  - umieszczenie transformacji x np. x^2
  - cecha danych: zmienne są ze sobą silnie skorelowane

WSPÓŁLINIOWOŚĆ (II)

- Szacujemy model:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

- Wariancja estymatora $\hat{\beta}_1$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(1 - R_1^2)\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2}$$

gdzie R_1^2 to R^2 z regresji x_1 na pozostałe zm. objaśniające.

- Gdy R_1^2 jest wysokie to:
  - błędy standardowe są wysokie $\Rightarrow$ statystyki t są niskie $\Rightarrow$ możemy błędnie ocenić niektóre zmienne jako nieistotne
  - oszacowania mogą być bardzo wrażliwe na niektóre wartości obserwacji lub na usunięcie nieistotnych zmiennych

WSPÓŁLINIOWOŚĆ (III)

- Do wykrywania współliniowości służy wskaźnik inflacji wariancji VIF, obliczany dla każdej zmiennej objaśniającej:

$$VIF_i = \frac{1}{1 - R_i^2}$$

  $VIF_i > 10$ sugeruje istnienie silniej współliniowości.
- można także analizować korelację między zmiennymi objaśniającymi.
- Współliniowość W modelu można zmniejszyć poprzez usunięcie zmiennej z wysokim VIF.

ZMIENNE NIEISTOTNE

UWZGLĘDNIENIE ZMIENNEJ NIEISTOTNEJ

- Przyjmijmy że specyfikacja modelu ma postać:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

  lecz tak naprawdę x_2 nie ma wpływu na y, wtedy $\beta_2 = 0$.
- Jakie są konsekwencje umieszczenia w regresji x_2?
  - nieobciążoność pozostaje nienaruszona: $\mathbb{E}[\hat{\beta}_2] = 0$
  - wariancja $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(1-R_1^2)\sum_{i=1}^n (x_{i,1}-\bar{x}_1)^2}$
  - dla porównania, wariancja estymatora $\tilde{\beta}_1$ bez uwzględniania x_2: $\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_{i,1}-\bar{x}_1)^2}$
  - $\mathrm{Var}(\tilde{\beta}_1) < \mathrm{Var}(\hat{\beta}_1) \Rightarrow$ spadek wariancji
- usunięcie nieistotnych zmiennych powinno poprawić precyzję oszacowań

PRZYKŁAD W R

`R` (I)

- Wydruk z komendy `lm()` podaje oszacowania, błędy standardowe, oblicza statystyki t dla każdego oszacowania oraz statystykę F dla całej regresji. Podaje także odpowiednie $p-value$ dla każdej statystyki. Łatwo możemy je obejrzeć (`summary()`) i ocenić istotność statystyczną oszacowań lub łączną istotność regresji
- Uwaga! Komenda `lm()` podaje właściwe statystyki w przypadku homoskedastyczności sk. losowego. Jeśli mamy podstawy aby twierdzić że występuje problem heteroskeadstyczności, musimy policzyć błędy standardowe etc. komendą `coeftest()`
- Skorzystamy ze zbioru danych `firmy` i oszacujemy funkcję produkcji:

$$y = \beta_0 + \beta_k k + \beta_l l + \varepsilon_i$$

R (II)

Wydruk z funkcji `summary()`

oszacowanie | błąd std. | statystyka t | p-value

```
Call:
lm(formula = lnY ~ lnK + lnL, data = firmy)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0351 -0.3296  0.0594  0.4029  3.4112

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.852151   0.039033   21.83   <2e-16 ***
lnK         0.549811   0.005276  104.21   <2e-16 ***
lnL         0.365328   0.005460   66.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7177 on 13146 degrees of freedom
Multiple R-squared:  0.7011,    Adjusted R-squared:  0.701
F-statistic: 1.541e+04 on 2 and 13146 DF,  p-value: < 2.2e-16
```

statystyka F - łączna istotność, DF - liczba st. swobody

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl