WŁASNOŚCI ESTYMATORA MNK.
Rozwiązania.

I. Nieobciążoność

1. Mamy estymator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'y$ gdzie λ jest skalarem.

   (i) Czy estymator ten jest liniowy?
   Rozwiązanie: Tak, ponieważ estymator jest liniową funkcją zmiennej zależnej:

$$\hat{\boldsymbol{\beta}} = \underbrace{(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'}_{\boldsymbol{A}\text{ - stała}}y$$

   (ii) Udowodnij że dla każdego $\lambda > 0$ ten estymator przy spełnieniu założeń KMRL jest obciążony.
   Rozwiązanie:

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}] \\
&= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}]}_{\boldsymbol{X\beta}} \\
&= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{\beta} \neq \boldsymbol{\beta}
\end{aligned}$$

   jeżeli $\lambda > 0$.

2. Mamy prosty model liniowy:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

   Pokazaliśmy, że estymator parametru β_1 ma następującą postać:

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

   Pokaż, że ten estymator jest nieobciążony.
   Rozwiązanie: Mamy pokazać że poniższa równość jest spełniona

$$\begin{aligned}
\mathbb{E}[\beta_1|x] &= \mathbb{E}[\hat{\beta}_1|x_i] \\
&= \mathbb{E}[\frac{\sum_{i=1}^n x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}|x_i] = *
\end{aligned}$$

   Najpierw skorzystajmy z faktu: $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$

$$\begin{aligned}
* &= \mathbb{E}[\frac{\sum_{i=1}^n x_i y_i - n\frac{1}{n}\sum_{i=1}^n y_i \frac{1}{n}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - n(\frac{1}{n}\sum_{i=1}^n x_i)^2}|x_i] \\
* &= \mathbb{E}[\frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n y_i \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}|x_i] = *
\end{aligned}$$

   Zauważmy, że w wartości oczekiwanej pozostanie tylko y_i

$$* = \frac{\sum_{i=1}^n x_i \, \mathbb{E}[y_i|x_i] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[y_i \mid x_i] \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2}$$

Z założeń KMRL wynika że $\mathbb{E}[y_i|x_i] = \beta_0 + \beta_1 x_i$:

$$
\begin{aligned}
* &= \frac{\sum_{i=1}^n x_i(\beta_0 + \beta_1 x_i) - \frac{1}{n}\sum_{i=1}^n(\beta_0 + \beta_1 x_i)\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2} \\
* &= \frac{\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 - \frac{1}{n}\beta_0 n \sum_{i=1}^n x_i + \beta_1 \frac{1}{n}(\sum_{i=1}^n x_i)^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2} \\
* &= \frac{\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 - \beta_0 \sum_{i=1}^n x_i - \frac{1}{n}\beta_1 (\sum_{i=1}^n x_i)^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2} \\
* &= \frac{\beta_1 \sum_{i=1}^n x_i^2 - \frac{1}{n}\beta_1 (\sum_{i=1}^n x_i)^2}{\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2} \\
* &= \beta_1
\end{aligned}
$$

Co należało pokazać.

3. Dany jest model regresji liniowej, spełniający założenia KMRL:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Jesteśmy zainteresowani szacowaniem sumy parametrów przy x_1 oraz x_2, niech $\theta = \beta_1 + \beta_2$. Pokaż, że estymator $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2$ jest nieobciążony

Rozwiązanie: Pokażmy że $\theta = \mathbb{E}[\hat{\theta}]$. Ponadto wiemy, że model podany w zadaniu spełnia założenia KMRL, więc $\hat{\theta}_1$ oraz $\hat{\theta}_2$ są nieobciążone. Zatem:

$$
\begin{aligned}
\theta &= \mathbb{E}[\hat{\theta}] \\
&= \mathbb{E}[\hat{\beta}_1 + \hat{\beta}_2] \\
&= \mathbb{E}[\hat{\beta}_1] + \mathbb{E}[\hat{\beta}_2] \\
&= \beta_1 + \beta_2 \\
&= \theta
\end{aligned}
$$

II. Założenia KMRL.

1. Które z poniższych przyczyn mogą spowodować, że estymator MNK będzie obciążony?
   (i) Heteroskedastyczność
   (ii) Pominięcie istotnej zmiennej
   (iii) Korelacja z próby wynosząca 0.95 między dwoma zmiennymi objaśniającymi, uwzględnionymi w modelu

   Rozwiązanie: Heteroskedastyczność jest sprzeczna z założeniem KMRL o sferyczności wariancji składnika losowego, co nie wpływa na obciążoność estymatora MNK. Korelacja z próby równa 0.95 między dwoma zmiennymi objaśniającymi to współliniowość, co również nie wpływa na obciążoność estymatora MNK. Natomiast w przypadku pominięcia istotnej zmiennej, złamane zostaje założenia KMRL o warunkowej wartości oczekiwanej równej zero ($\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0$), co sprawia że estymator MNK będzie obciążony.

2. Które z poniższych warunków są niezbędne, aby pokazać że estymator MNK jest nieobciążony i efektywny?
   (i) $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$
   (ii) $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2$
   (iii) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall\, j \neq i$
   (iv) $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

Rozwiązanie: Innymi słowy, które założenia są potrzebne aby udowodnić twierdzenie Gaussa-Markowa? Warunkowa wartość oczekiwana równa zero (i), stałość wariancji-homoskedastyczny składnik losowy (ii) oraz brak autokorealcji (iii) z wyżej podanych. Założenie o normalności składnika losowego nie jest potrzebne aby pokazać nieobciążoność oraz efektywność estymatora MNK w Klasycznym Modelu Regresji Liniowej.

3. Jakie jest znaczenie terminu *heteroskedastyczność*?

   (i) Wariancja błędów nie jest stała.
   (ii) Wariancja zmiennej zależnej nie jest stała.
   (iii) Błędy nie są od siebie liniowo niezależne.
   (iv) Błędy mają średnią niezerową.

   Rozwiązanie: Heteroskedastyczność to sytuacji w której wariancja błędów nie jest stała.

4. Które z poniższych sytuacji mogą być konsekwencją naruszenia jednego lub większej liczby założeń KMRL?

   (i) Oszacowania współczynników są obciążone.
   (ii) Oszacowania błędu standardowego nie są optymalne.
   (iii) Rozkłady przyjęte dla statystyk testowych są niewłaściwe.
   (iv) Wnioski dotyczące siły relacji pomiędzy osobą zależną a zmienną niezależną mogą być nieprawidłowe.

   Rozwiązanie: (i) - obciążenie estymatora głównie może być konsekwencją niespełnienia założenia o warunkowej wartości oczekiwanej równej zero. (ii) - niewłaściwe oszacowanie błędów standardowych może być konsekwencją heteroskedastyczności. (iv) - może być konsekwencją obu powyższych - zarówno obciążone oszacowanie jaki niewłaściwe oszacowanie błędów std. O (iii) będziemy jeszcze mówić.

5. Jakie byłyby konsekwencje dla estymatora MNK, gdyby heteroskedastyczność była obecna w modelu regresji, ale została zignorowana?

   (i) Estymator będzie obciążony.
   (ii) Estymator nie będzie zgodny.
   (iii) Estymator będzie nieefektywny.
   (iv) Wszystkie powyższe punkty będą prawdziwe.

   Rozwiązanie: Estymator będzie nieefektywny - estymator MNK nie będzie już najbardziej efektywny w klasie liniowych nieobciążonych estymatorów. Natomiast estymator nadal będzie nieobciążony. Zgodność jest asymptotyczną własnością estymatorów i nie będzie ona przedmiotem naszych zajęć.

6. Dany jest model z 5 zmiennymi objaśniającymi szacowany na 100 obserwacjach

   (i) Jaki jest rozmiar macierzy wariancji estymatora modelu?
   Rozwiązanie: Mamy 6 parametrów do oszacowania, zatem macierz wariacji-kowariancji estymatora będzie miała rozmiar 6×6
   (ii) Zapisz postać macierzy z poprzedniego podpunktu w KMRL.
   Rozwiązanie:

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \begin{bmatrix} \mathrm{Var}[\beta_0] & \mathrm{Cov}[\beta_0, \beta_1] & \dots & \mathrm{Cov}[\beta_0, \beta_6] \\ \mathrm{Cov}[\beta_1, \beta_0] & \mathrm{Var}[\beta_1] & \dots & \mathrm{Cov}[\beta_1, \beta_6] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[\beta_6, \beta_0] & \mathrm{Cov}[\beta_6, \beta_1] & \dots & \mathrm{Var}[\beta_6] \end{bmatrix}_{6\times 6}$$

   (iii) Jaki jest rozmiar macierzy wariancji wariancji składnika losowego?

Rozwiązanie: Model jest szacowany na 100 obserwacjach, czy wektor składnika losowego $\boldsymbol{\varepsilon}$ ma rozmiar 100. Zatem macierz wariancji składnika losowego $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$ będzie miała rozmiar 100×100

(iv) Zapisz postać macierzy z poprzedniego podpunktu w KMRL.

Rozwiązanie:

$$\mathrm{Var}[\boldsymbol{\varepsilon}] = \begin{bmatrix} \mathrm{Var}[\varepsilon_0] & \mathrm{Cov}[\varepsilon_0, \varepsilon_1] & \dots & \mathrm{Cov}[\varepsilon_0, \varepsilon_{100}] \\ \mathrm{Cov}[\varepsilon_1, \varepsilon_0] & \mathrm{Var}[\varepsilon_1] & \dots & \mathrm{Cov}[\varepsilon_1, \varepsilon_{100}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[\varepsilon_{100}, \varepsilon_0] & \mathrm{Cov}[\varepsilon_{100}, \varepsilon_1] & \dots & \mathrm{Var}[\varepsilon_{100}] \end{bmatrix}_{100\times 100}$$